J Comb Optim (2011) 22:145–152
DOI 10.1007/s10878-009-9279-9

Pooling designs for clone library screening in the inhibitor complex model

**Fei-huang Chang · Huilan Chang ·
Frank K. Hwang**

Published online: 9 February 2010

Abstract In this paper we introduce inhibitors into the complex model and we give a lower bound and an upper bound of tests in a nonadaptive pooling design for some inhibitor complex model. We propose a very efficient pooling design for the general inhibitor complex model and extend it to the error-tolerant case.

Keywords Pooling design · Group testing · Inhibitor complex model · Error-tolerance

1 Introduction

In the clone library screening problem, the goal is to identify a small set of *positive clones* from a large set of *negative clones*. Group testing is often the tool to accomplish this. A *group test* is performed on an arbitrary subset of clones with two possible outcomes: a *negative outcome* if all clones in the subset are negative, and a *positive outcome* if otherwise. A sequence of tests which can identify a fixed but unknown set of positive clones is a solution, usually called a *design*, to the problem. Since a test is time-consuming, it is much preferred to perform all tests simultaneously. In this case, we will refer to the design as a *nonadaptive pooling design*.

This research is partially supported by a Republic of China National Science grant NSC 93-2115-M-009-013.

F. Chang
Department of Mathematics and Science, National Taiwan Normal University, Taipei County 24449, Taiwan, ROC
e-mail: cfh@ntnu.edu.tw

H. Chang (✉) · F.K. Hwang
Department of Applied Mathematics, National Chiao Tung University, Hsinchu 300, Taiwan, ROC
e-mail: huilan0102@gmail.com

To have an efficient design, we need to assume some knowledge about the positive set. The most innocent assumption is an upper bound d of the number of positive clones in the test population. A pooling design is usually represented by the incidence matrix M where rows are indexed by tests and columns by clones. A binary matrix is called *d-disjunct* if no column is covered by the union of any other d columns. In the clone model, it is well known (Du and Hwang 2000) that a d-disjunct matrix can identify all p positive clones ($p \le d$). In certain applications, there is a third type of clones called *inhibitors* whose existence may cancel the effect of positive clones. Farach et al. (1997) first proposed the 1-inhibitor model in which a single inhibitor clone dictates the test outcome to be negative regardless of how many positive clones are in the test. De Bonis and Vaccaro (2003) extended the above model to k-inhibitor model in which k inhibitor clones dictate the test outcome to be negative. However, only sequential designs were given in De Bonis and Vaccaro (1998), De Bonis and Vaccaro (2003), and Farach et al. (1997). Recently, nonadaptive pooling designs have been proposed for the inhibitor model (Dyachkov et al. 2001; Hwang and Liu 2003; Du and Hwang 2005), and extended to the general inhibitor model (Chang and Hwang 2007) in which the exact cancellation effect of inhibitors on positive clones is not specified.

In some DNA screening environment, it takes a subset of clones, called a *complex*, to induce a positive outcome. We call such a model the *complex model*, as versus the *clone model* discussed before. Thus in the complex model, a fixed but unknown set of complexes are designated *positive*, while other candidates of positive complexes are called *negative complexes*. It is usually assumed that two positive complexes can overlap, but neither contains the other. Torney (1999) first introduced the complex model and gave the complexes of eukaryotic DNA transcription and RNA translation as an example.

In this paper we introduce inhibitors into the complex model and call it the *inhibitor complex model*. In this model, an *inhibitor* is a third type of complex, other than positive and negative, whose presence may cancel the effect of positive complexes. In the simplest inhibitor complex model, the *1-inhibitor complex model*, the mere existence of a single inhibitor dictates the test outcome to be negative, regardless of the presence of positive complexes. If the requirement is changed from a single inhibitor to k inhibitors, then it is the *k-inhibitor complex model*. In general, in a *(k, g)-inhibitor complex model*, k inhibitors cancel the effect of g positive complexes. Usually, we don't know the two parameters k and g for sure. We will refer to a model without such specification the *general inhibitor complex model*. In this paper, we propose a very efficient nonadaptive pooling design for the general inhibitor complex model, i.e., it works against any (k, g)-inhibitor complex model, and extend it to the error-tolerant case.

2 The general inhibitor complex model

In the general inhibitor clone model, the $(d + s)$-disjunct matrix is the main tool to identify the positive clones from n clones including at most d positive clones and at most s inhibitor clones. We extend this idea to the general inhibitor complex model.

We attach the parameters (r, d, s) to a general inhibitor complex model to denote the fact that among the complexes, which are subsets of the n clones, there are at most d positive complexes, at most s inhibitors, and a complex has at most r clones. For a complex X, we say a row *covers* X if it intersects (has a 1-entry in) every column of X while $\bigcap X$ denotes the set of rows covering X. Let H denote the given set of complexes in the considered problem. Then H can be viewed as a hypergraph with clones as vertices and complexes as edges. Following the terminology of Gao et al. (2006), a binary matrix is $(H : d)$-*disjunct* if for any $d+1$ complexes $X_0, X_1, \dots, X_d$ there always exists a row covering X_0 but none of $X_1, \dots, X_d$, i.e.,

$$\bigcap X_0 \not\subseteq \bigcup_{i=1}^{d} \bigcap X_i.$$

Denote the minimum number of rows in an $(H : d)$-disjunct matrix with n columns by $t(H : d, n)$.

It is generally assumed that no edge is a subset of another edge for otherwise the requirement of $(H : d)$-disjunctness cannot be fulfilled when X_0 is contained in one of the X_i. We write H_r for H to denote the fact that each edge in H has maximum rank (number of vertices) r.

By the definition given above, clearly, we have

Lemma 1 *$(H_r : d)$-disjunct implies $(H_{r'} : d')$-disjunct for $d \geq d'$ and $r \geq r'$.*

We give a lower bound for the (r, d, s) general inhibitor complex model, which is an extension of a similar result in De Bonis and Vaccaro (1998) for the general inhibitor clone model.

Theorem 2 *The number of rows in a nonadaptive pooling design for the (r, d, s) general inhibitor complex model is at least $t(H_r : s, n)$.*

Proof Since a lower bound of the 1-inhibitor complex model is clearly a lower bound of the general inhibitor complex model, it suffices to prove for the 1-inhibitor case.

Let M be the testing matrix of a nonadaptive pooling design. Suppose M is not $(H_r : s)$-disjunct. Then there exists a set of $s+1$ complexes $X_0, \dots, X_s$, $\bigcap X_0 \subseteq \bigcup_{i=1}^{s} \bigcap X_i$. Consider the sample that X_0 is a positive complex and $\{X_1, \dots, X_s\}$ is the set of inhibitors. Then outcomes of the tests covering X_0 are negative and we can't identify X_0 from such outcomes. □

Theorem 3 *An $(H_r : d+s)$-disjunct matrix can identify all positive complexes under the (r, d, s) general inhibitor complex model.*

Proof Let Q denote a negative complex, P a positive complex and I an inhibitor complex. Let $t(X)$ denote the number of negative pools complex X appears in. For an s-set S of complexes, let $t^S(X)$ denote the same except that a 0-outcome is changed to 1 if that test covers a complex contained in S. Define $\mathcal{H}$ as the set of all choices of

s edges from H and $t^{\mathcal{H}}(X) = \min_{S\in\mathcal{H}} t^S(X)$. Then

$$t^{\mathcal{H}}(P) = t^{S'}(P) = 0,$$

where S' is an s-set containing all inhibitors.

By the definition of an $(H : d + s)$-disjunct matrix, for any complex X, $\bigcap X$ is not contained in $\bigcup_{i=1}^{d+s} \bigcap X_i$ for any other complexes $X_1, \ldots, X_{d+s}$, i.e., there exists a row covering X but none of $X_1, \ldots, X_{d+s}$. In particular, when $\{X_1, \ldots, X_d\}$ covers the set of positive complexes, and $\{X_{d+1}, \ldots, X_{d+s}\}$ is a given set S, we have $t^S(X) \geq 1$, for $X \in \{Q, I\}$ for all S. Consequently, $t^{\mathcal{H}}(X) \geq 1$, for $X \in \{Q, I\}$. Thus we conclude $\{X; t^{\mathcal{H}}(X) = 0\}$ is the set of all positive complexes. □

Corollary 4 *The number of rows in a nonadaptive pooling design for the* (r, d, s) *general inhibitor complex model is at most* $t(H : d + s, n)$.

Next, we discuss the error-tolerant case. We consider two types of errors: the 10-type, changing 1-outcome to 0, and the 01-type, changing 0-outcome to 1. Let e_{10}^* and e_{01}^* denote the unknown numbers of the 10-type errors and the 01-type errors, respectively, and denote upper bounds of e_{10}^* and e_{01}^* as e_{10} and e_{01}, either known or unknown. We assume e, an upper bound of the total number of errors, is known. We extend the $(H : d)$-disjunct matrix to the error-tolerant case. A binary matrix is called $(H : d; z)$*-disjunct* if for any $d + 1$ complexes $X_0, X_1, \ldots, X_d$, there exist at least z rows which cover X_0, but none of $X_1, \ldots, X_d$. Construction of $(H : d; z)$-disjunct matrices was studied in Gao et al. (2006). For n clones and $z \ll n$, the construction yields a matrix with $O((d \log n)^{r+1})$ rows.

Theorem 5 *An* $(H : d + s; c + e + 1)$*-disjunct matrix can identify all positive complexes under the* (r, d, s) *general inhibitor complex model with at most* e *errors, where*

$$c = \begin{cases} \text{(i) } e_{10} + e_{01} - e, & \textit{if } e_{10} \textit{ and } e_{01} \textit{ are known,} \\ \text{(ii) } e, & \textit{if there are no estimates of } e_{10} \textit{ and } e_{01}, \\ \text{(iii) } 0, & \textit{if the number of positive complexes is } d. \end{cases}$$

Proof Ignoring the inhibitors for the moment, then a positive complex P can appear in a negative pool only if its outcome is one of the 10-type errors. So when S contains all inhibitors,

$$t^{\mathcal{H}}(P) = t^S(P) \leq e_{10}^*.$$

On the other hand, for $X \in \{Q, I\}$, then by the definition of $(H : d + s; c + e + 1)$-disjunct, X appears in at least $c + e + 1$ rows each covering none of the up-to-d positive complexes, nor the s complexes in S; hence the corresponding tests have negative outcomes. Errors of the 01-type may reduce the number of such negative pools. But still,

$$t^{\mathcal{H}}(X) = \min_{S\in\mathcal{H}} t^S(X) \geq \min_S \{c + e + 1 - e_{01}^*\} = c + e + 1 - e_{01}^*.$$

Since $e \geq e^*_{10} + e^*_{01}$, $t^{\mathcal{H}}(X) \geq c + e^*_{10} + 1 > t^{\mathcal{H}}(P)$. The problem is we do not know e^*_{10} and hence not knowing where to draw the line to separate P from I and Q.

We consider three cases:

(i) We know e_{10} and e_{01}. Set $c = e_{01} + e_{10} - e$. Then for $X \in \{Q, I\}$,

$$t^{\mathcal{H}}(X) \geq (e_{01} + e_{10} - e) + e + 1 - e^*_{01} = e_{10} + 1.$$

Thus $\{X' : t^{\mathcal{H}}(X') \leq e_{10}\}$ is the set of all positive complexes.

(ii) If we have no estimates of e_{10} and e_{01}, set $c = e$. Then

$$t^{\mathcal{H}}(X) \geq e + e + 1 - e^*_{01} \geq e + 1.$$

Thus $\{X' : t^{\mathcal{H}}(X') \leq e\}$ is the set of all positive complexes.

(iii) If the number of positive complexes is known to be exactly d, then set $c = 0$ and we have $\{X' : t^{\mathcal{H}}(X')$ is among the d smallest $t^{\mathcal{H}}$ values$\}$ is the set of all positive complexes. □

3 A faster procedure

Suppose H has n vertices and h edges. Then h can be much larger than n. For example, if H is the complete r-graph, i.e., the edge-set consists of all r-sets of vertices, then $h = \binom{n}{r}$. The computation of $t^{\mathcal{H}}(X)$ requires to go through all s-sets of edges and there are $\binom{h}{s}$ of them. This could be a very large number. We now show that there is a way to reduce this work in order of magnitude.

A seemingly unrelated notion, the $(d, r; z)$-disjunct matrix, has been well studied (see Dyachkov et al. 2001; Stinson and Wei 2004 for recent development) under the contexts of superimposed codes and secure key distribution methods. A binary matrix is $(d, r; z)$-disjunct if for any $d + r$ columns $C_1, \ldots, C_{d+r}$, there always exist z rows with 1-entries in $C_1, \ldots, C_r$ and 0-entries in $C_{r+1}, \ldots, C_{d+r}$. We now show the relevance of this matrix to our problem.

Theorem 6 *A* $(d + s, r; 2e + 1)$*-disjunct matrix can identify all positive complexes under the* (r, d, s) *general inhibitor complex model with at most e errors.*

Proof We use the same terminology defined in Theorem 3 except that S now stands for an s-set of vertices, instead of an s-set of edges. Consider a positive complex P and let $\{X_1, \ldots, X_s\}$ denote a set of other complexes containing all inhibitors. Since no edge is contained in another edge, there exists vertex $v_i \in X_i \backslash P$ for $1 \leq i \leq s$. Let V denote the set of these v_i. If $|V| < s$, add $s - |V|$ arbitrary other vertices to it to become V'. Let N denote the set of the n vertices.

$$t^{\mathcal{N}}(P) = t^{V'}(P) \leq e,$$

since P can be in a negative pool only by the occurrence of error.

Next, consider $X \in \{Q, I\}$ and let $\{X_1, \ldots, X_d\}$ denote a set of other complexes containing all positive complexes. By a similar reason as given in the first paragraph, we can define $w_i \in X_i \backslash X$ and $W = \{w_i\}$. Again, if $|W| < d$, add $d - |W|$ arbitrary other vertices to become W'. By the definition of $(d+s, r; 2e+1)$-disjunct matrix, there exist at least $2e+1$ rows in which each of the columns in X has a 1-entry but none of the columns in W' nor the s columns in an arbitrary s-set S does. Hence the outcomes of these $2e+1$ pools should be negative under S except for the occurrence of errors. Hence

$$t^{\mathcal{N}}(X) \geq 2e + 1 - e = e + 1.$$

Thus $\{X : t^{\mathcal{N}}(X) \leq e\}$ is the set of positive complexes. □

To compute $t^{\mathcal{N}}(X)$, we need only to go through $\binom{n}{s}$ s-sets of columns, a big deduction from $\binom{h}{s}$.

4 The *k*-inhibitor complex model

In the *k-inhibitor complex model*, the outcome of a test is positive if and only if it contains at least one positive complex and at most $k-1$ inhibitors. While Sect. 2 provided a nonadaptive pooling design for this model, we now give a more efficient one, following the approach of De Bonis and Vaccaro (2003) for the k-inhibitor clone model. Call a binary matrix M an $(H : d, m$ out of $s)$-*disjunct* matrix if for any $d+s+1$ complexes, $X_0, X_1, \ldots, X_d, X_{d+1}, \ldots, X_{d+s}$, there exists a test which covers X_0 but none of $X_1, \ldots, X_d$ and does not cover at least m of $X_{d+1}, \ldots, X_{d+s}$. Clearly,

Theorem 7 *An* $(H : d, s-k+1$ *out of* $s)$*-disjunct matrix can identify all positive complexes under the* (r, d, s) *k-inhibitor complex model.*

Proof For an s-set S of complexes and a fixed k, let $t^S(X)$ denote the number of negative pools complex X appears in except that a 0-outcome is changed to 1 if that test covers k complexes from S. Then

$$t^{\mathcal{H}}(P) = \min_{S \in \mathcal{H}} t^S(P) = t^{S'}(P) = 0,$$

where S' is an s-set containing all inhibitors.

On the other hand, by the definition of an $(H : d, s-k+1$ out of $s)$-disjunct matrix, for any complex X, there exists a test which covers X but none of $X_1, \ldots, X_d$ and does not cover at least $s-k+1$ of $X_{d+1}, \ldots, X_{d+s}$. In particular, when $\{X_1, \ldots, X_d\}$ covers the set of positive complexes, and $\{X_{d+1}, \ldots, X_{d+s}\}$ is the given set S, we have $t^S(X) \geq 1$, for $X \in \{Q, I\}$ for all S. Consequently, $t^{\mathcal{H}}(X) \geq 1$, for $X \in \{Q, I\}$. Thus we conclude $\{X; t^{\mathcal{H}}(X) = 0\}$ is the set of all positive complexes. □

Corollary 8 *An* $(H : d, s-k+1$ *out of* $s; c+e+1)$*-disjunct matrix can identify all positive complexes under the* (r, d, s) *k-inhibitor complex model with at most e errors, where*

$$c = \begin{cases} \text{(i) } e_{10} + e_{01} - e, & \textit{if } e_{10} \textit{ and } e_{01} \textit{ are known}, \\ \text{(ii) } e, & \textit{if there are no estimates of } e_{10} \textit{ and } e_{01}, \\ \text{(iii) } 0, & \textit{if } p = d. \end{cases}$$

There is also a fast procedure for this result. A binary matrix is $((d, m$ out of $s)$, $r; z)$-disjunct if for any $d+s+r$ columns $C_1, \ldots, C_{d+s+r}$, there always exist z rows with 1-entries in $C_1, \ldots, C_r$, 0-entries in $C_{r+1}, \ldots, C_{d+r}$, and at least m 0-entries in $C_{d+r+1}, \ldots, C_{d+r+s}$. A $((d, s-k+1$ out of $s), r; 2e+1)$-disjunct matrix can identify all positive complexes under the (r, d, s) k-inhibitor complex model with at most e errors. This procedure also results in a big deduction in computation, namely, from computing $t^{\mathcal{H}}(X)$ to computing $t^{\mathcal{N}}(X)$.

For the 1-inhibitor model, a further reduction in computation is possible. Note that any pool containing an inhibitor complex induces a negative outcome unless an error occurs. We use this property to confine all inhibitor complexes to a small set O of complexes. Let $t_1(X)$ denote the number of positive pools X appears in. Define $O = \{X : t_1(X) \leq e\}$. Then O contains all inhibitor complexes. So instead of computing $t^{\mathcal{H}}(X)$ over all s-set $S \in \mathcal{H}$, we need only to compute $t^{O}(X)$ over all $S \subseteq O$. Further, $(H : d, s$ out of $s; c+e+1)$-disjunct is just $(H : d+s; c+e+1)$-disjunct. Hence for any $s+1$ complexes $X_0, X_1, \ldots, X_s$, there exist $c+e+1$ rows covering X_0 but none of $X_1, \ldots, X_s$. In particular, this is true when X_0 is positive and $\{X_1, \ldots, X_s\}$ contains all inhibitor complexes. Thus $t_1(P) \geq c+e+1$ and we need to compute $t^{O}(X)$ only for those X satisfying $t_1(X) \geq c+e+1$ to identify positive complexes.

A similar reduction is possible for the faster procedure. Define $O' = \{C : C \in X \in O\}$. Then we need to compute $t^{O'}(X)$ instead of $t^{\mathcal{N}}(X)$.

5 Conclusions

We introduced a new pooling design environment by allowing the coexistence of inhibitors and complexes which, separately, have been well studied in the literature. We gave a nonadaptive pooling design, with error-tolerance ability, to the most general model in such an environment with no need to know the exact relation between inhibitors and positive complexes. Thus the design is applicable to many practical situations. This design is an $(H : d; z)$-type disjunct matrix whose construction has been studied in Dyachkov et al. (2001), Gao et al. (2006), and Stinson and Wei (2004).

References

Chang FH, Hwang FK (2007) The identification of positive clones in a general inhibitor model. J Comput Syst Sci 73:1090–1094

De Bonis A, Vaccaro U (1998) Improved algorithms for group testing with inhibitors. Inf Process Lett 67:57–64

De Bonis A, Vaccaro U (2003) Constructions of generalized superimposed codes with applications to group testing and conflict resolution in multiple access channels. Theor Comput Sci A 356:223–243

Dyachkov AG, Macula AJ, Torney DC, Villenkin PA (2001) Two models of nonadaptive group testing for designing screening experiments. In: Atkinson AC, Hackl P, Muller WG (eds) Proc 6th int workshop on model-oriented design and analysis. Physica-Verlag, Heidelberg, pp 63–75

Du DZ, Hwang FK (2000) Combinational group testing and its applications, 2nd edn. Would Scientific, Singapore

Du DZ, Hwang FK (2005) Identifying d positive clones in the presence of inhibitors. Int J Bioinform Res Appl 1:162–168

Farach M, Kannan S, Knill E, Muthvkrishnan S (1997) Group testing problems with sequences in experimental molecular biology. In: Carpentieri B et al (eds) Proc compression and complexity of sequences. IEEE Press, New York, pp 357–367

Gao H, Hwang FK, Thai M, Wu W, Znati T (2006) Construction of $d(H)$-disjunct matrix for group testing in hypergraphs. J Comb Optim 12:297–301

Hwang FK, Liu YC (2003) Error-tolerant pooling designs with inhibitors. J Comput Biol 10:231–236

Stinson DR, Wei R (2004) Generalized cover-free families. Discrete Math 279:463–477

Torney DC (1999) Sets pooling designs. Ann Comb 3:95–101