GT Lecture 8

Nov. 23

Complex Model: Positives are subsets, not only of size 1.

(••) Tests are positive, if it contains one of the subsets (positive complex).

- 我们必需假设当 X, X' 皆为 complexes 时，彼此间不互相包含的关係，亦即 $X \not\subseteq X'$ 且 $X' \not\subseteq X$。

- Or the model must be monotone on positives, i.e., if X is positive and $X \subseteq X'$, then X' is also positive. But, this is not quite true in applications.

- If not, and X contains a proper subset X^+ which is positive, then X can only appear in positive pools no matter it is positive or negative. Thus, X cannot be identified.

- Let H denote the given set of complexes. Then H can be viewed as a hypergraph with clones as vertices and complexes as edges. Accordingly, complex model is dealing with searching a hidden graph (subgraph) P in a given hypergraph H. Clearly, P is induced by the set of positive edges.

- The forbidable complexes of eukaryotic DNA transcription and RNA translation could involve hundred of molecules.

- Application: protein-to-protein interaction (Many) (Lappe and Holm, 2004).

(*) (Creating a protein-protein interaction network.)

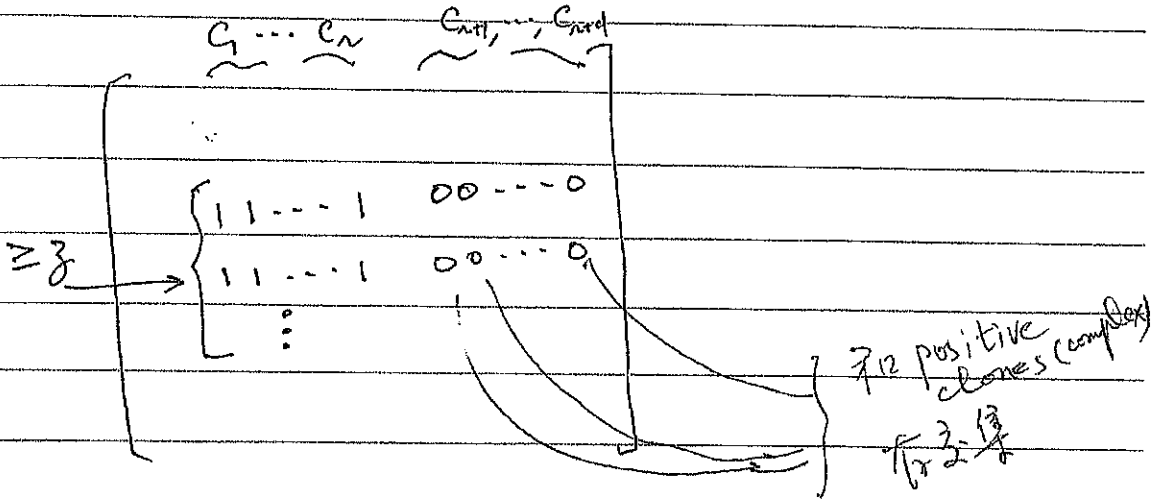
- A complex contains at most r clones! (Assumption)

(Generalized cover-free family)

Definition ($(d, r; z]$-disjunct) Stinson and Wei, '99

A matrix is $(d, r; z]$-disjunct if for any $r+d$ columns $C_1, C_2, \cdots, C_{r+d}$,

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



Note that if $r=1$, then we use $(d; z]$-disjunct for short.

→ Using a $(d, r; z]$-disjunct matrix

(*) **Proposition** For a negative complex X, there are at least z rows containing X but no positive complexes.

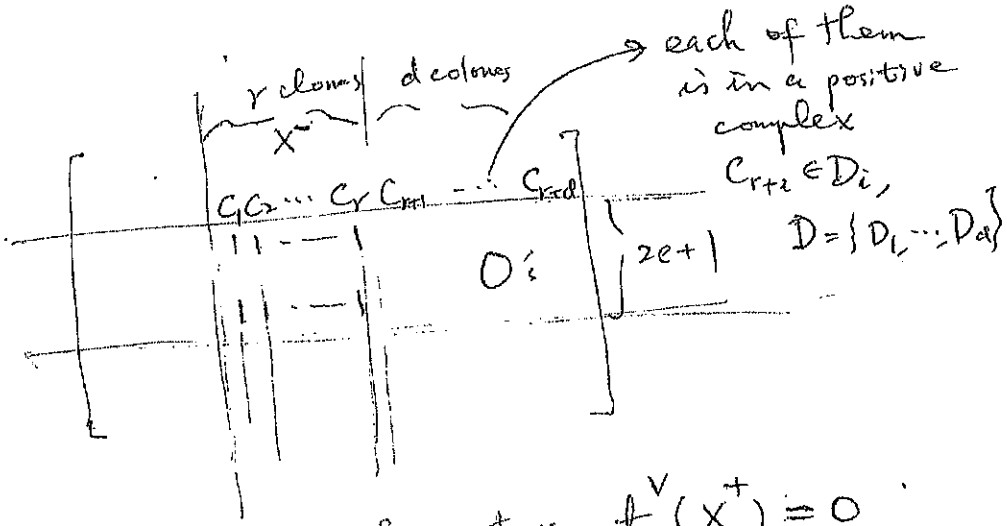
Proof For a complex X, $t_0^V(X) =_{def} |\cap X \setminus V|$ and $t_1^V(X) =_{def} |(\cap X) \cap V|$.

- We can use $(d, r; 2e+1]$-disjunct matrix to identify up-to-d positives with at most e errors.

Theorem

A $(d, r; 2e+1]$-disjunct matrix can be applied to identify all d (at most) positive complexes with at most e errors.

Proof.



By the definition of the matrix $t_0^V(X^+) = 0$ (with no errors). So, if there are at most e errors, $t_0^V(X^+) \le e$.

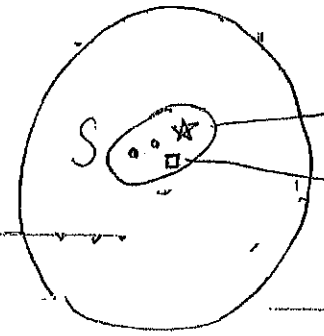
On the other hand $t_0^V(X^-)$ will be at least $2e+1$ if we let $\{C_1, \cdots, C_r\}$ be a negative complex (no errors) and $C_{r+i} \in D_i$ where D_i is a positive complex in $\mathcal{D}$.

Now, even there are e errors, $t_0^V(X^-) \ge e+1$.

(*) If there are h inhibitors, then we use a (See next page!)

$(d+h, r; 2e+1]$-disjunct matrix.

(Ideal 和前面不是 complex model 相似)



The test for S is negative.

Definition (Inhibitor)

A clone is an inhibitor if the clone neutralize positive clones, that is, the presence of an inhibitor in a pool dictates a negative outcome, regardless of the presence of positive clones in the pool.

(註) Inhibitor 出現時，該 Pool 是陰性反應。

Theorem (Error-tolerant inhibitor model) (不是 Complex model!)

A $(d+h; 2e+1)$-disjunct matrix can be applied to identify all positive clones if there are at most d positives and h inhibitors.

Proof. Observe that a clone C, which appears in at most e positive pools, i.e., $t^V(C) \le e$, cannot be positive due to the $(d+h; 2e+1)$-disjunctness property. (If C is positive, then $t^V(C) \ge 2e+1$ if there are no errors.) For the inhibitors, each inhibitor can occur in at most e positive pools (with e errors), i.e., $t^V(I) \le e$. $t^V(C) \le e$.

Now, let O be the set of columns C such that ~~$t^V(C)$~~ This implies that O contains all inhibitors but no positives. Hence, D, the set of positive clones is a subset of $N \setminus O$.

Let $C^- \in N \setminus O$. By disjunctness property, we have $t_o^V(C^-) \geq e+1$. On the other hand, if $C^+ \in N \setminus O$, then $t_o^V(C^+) \leq e$. The proof follows. ∎

(•) Notice here, we are aimming at identifying all the positive items. We may have the idea where the inhibitors are hidding, but it is not our goal to classify them. Indeed, if we would like to do it, then a "higher power" matrix is needed.

(••) The model with inhibitors are important in real experiment. This is why these papers about this model will appear in J. Computational Molecular Biology.

(•••) For more information, please refer to the Ph. D. thesis of Hong-Bin Chen (陳宏賓) and Hui-lan Chang (張惠蘭), both were graduates of our department.

在加入 Inhibitors 之後，我们並不需要去找到那些 Clones 是 Inhibitors。

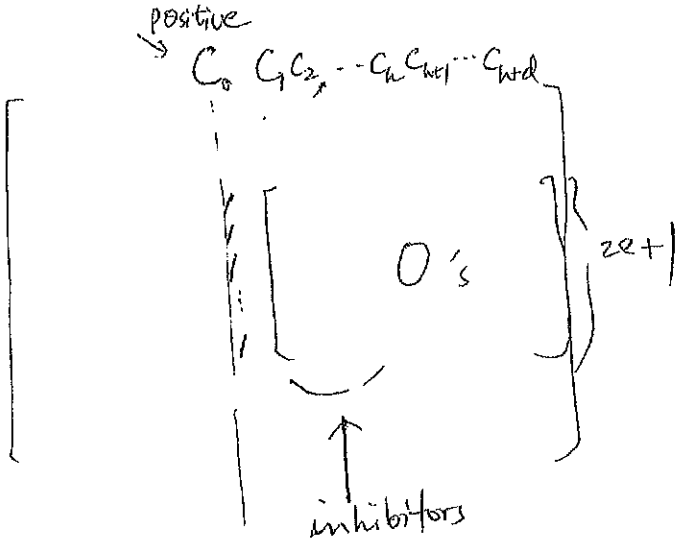
概念：先把不可能是 positive clones 的 clones 找出來，形成集合 O。這个集合中可能還有一些 negative clones。

這集合 O 可以由 positive pools 組成。

Reason

由於用了 $(d+h; 2e+1)$-disjunct 矩陣，

$t_1^V(C) \geq 2e+1$ (沒有 Errors)。



由於 Inhibitors 不加入這些 pools，所以 $t_1^V(C) \geq 2e+1$。

$\Rightarrow$ 如果有 e 个錯，$t_1^V(C) \geq e+1$。

JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 17, Number 7, 2010
© Mary Ann Liebert, Inc.
Pp. 927–941
DOI: 10.1089/cmb.2009.0138

Identification and Classification Problems on Pooling Designs for Inhibitor Models

HUILAN CHANG, HONG-BIN CHEN, and HUNG-LIN FU

ABSTRACT

Pooling designs are common tools to efficiently distinguish positive clones from negative clones in clone library screening. In some applications, there is a third type of clones called "inhibitors" whose effect is in a sense to obscure the positive clones in pools. Various inhibitor models have been proposed in the literature. We address the inhibitor problems of designing efficient nonadaptive procedures for both identification and classification problems, and improve previous results in three aspects: (1) The algorithm that is used to identify the positive clones works on a more general inhibitor model and has a polynomial-time decoding procedure that recovers the set of positives from the knowledge of the outcomes. (2) The algorithm that is used to classify all clones works in one-stage, i.e., all tests are arranged in advance without knowing the outcomes of other tests, along with a polynomial-time decoding procedure. (3) We extend our results to pooling designs on complexes where the property to be screened is defined on subsets of biological objects, instead of on individual ones.

Key words: complex model, group testing, inhibitor, nonadaptive algorithm, pooling design.

1. INTRODUCTION

POOLING DESIGNS ARE USED IN HIGH-THROUGHPUT screening projects where the goal is the identification of low-frequency events in a large collection of samples. In general, pooling designs deal with binary types (positive or negative) responses, and the task is to identify the positive ones in a large collection of samples by testing samples in groups.

In most applications, one standard assumption is that a pool with a positive response contains at least one positive whereas a pool with a negative response contains no positives. In chemical and biological experiments, however, there are complications.

(a) High-throughput biological assays are usually somewhat unreliable, and thus both false positive and false negative observations are to be expected in experiments. An intuitive way to deal with the issue consists in repeating all tests several times, but this is usually not efficient. In practice, the error-tolerance ability is an important added benefit to pooling designs.

Department of Applied Mathematics, National Chiao Tung University, Hsinchu, Taiwan.

(b) Sometimes, a small percentage of positive pools contain no single positive individuals. The contributing reason is that the positive responses may be caused when a pair or some set of clones are placed together.
(c) In addition to positive and negative clones, there is a third type of clones called *inhibitors* whose effect is in a sense to obscure positive clones so that a positive individual may be falsely declared as a negative.

Farach et al. (1997) were the first to introduce the complication (c) for applications in the field of molecular biology. An example in molecular biology is the so-called enzyme inhibitors. Enzyme inhibitors are molecules that interact in some way with the enzyme to prevent it from working in the normal manner. Similar phenomena were noted in blood testing applications by Phatarfod and Sudbury (1994), and in drug discovery applications by Xie et al. (2001).

Various models can be formulated with inhibitors in the pooling design, depending on how many inhibitors can interfere with how many positive clones. The usual assumption is that one inhibitor dictates the pool to be negative regardless of how many positive clones are in the pool. We refer this model as the 1-*inhibitor model*. This model of group testing has been widely studied in the literature (De Bonis, 2008; De Bonis et al., 1998, 2005; Farach et al., 1997; Hwang et al., 2003). Another variant proposed by De Bonis and Vaccaro (2003) is the *k-inhibitor model* where a pool has a positive response if and only if it contains at least one positive clone and at most $k-1$ inhibitors. Hwang and Chang (2007) extended these inhibitor models to the *general inhibitor model* which is a more general version including all above variations of interference between inhibitors and positive clones. We now give a formal definition to this model in the following.

Definition 1. *Given a set $\mathcal{N}$ of n clones consisting of three types of clones: a set $\mathcal{P}$ of positive clones, a set $\mathcal{I}$ of inhibitors and the others being negative clones with $|\mathcal{P}| \leq d$ and $|\mathcal{I}| \leq h$, the general inhibitor model is a model of group testing in which the allowed questions are of the following form: "Is $Q \cap \mathcal{P} \neq \emptyset$", where $Q \subseteq \mathcal{N}$, and answers are correctly given if either $Q \cap \mathcal{P} \neq \emptyset$ and $Q \cap \mathcal{I} = \emptyset$ both hold, or $Q \cap \mathcal{P} = \emptyset$ occurs. Answers might be incorrectly given if $Q \cap \mathcal{P} \neq \emptyset$ and $Q \cap \mathcal{I} = \emptyset$ both hold.*

Obviously, there are many other models depending on how positive clones and inhibitors interfere with each other. In reality, however, rarely do we have exact information beforehand. The interference effect can be very different from that have been investigated in the literature. We may face the situation that a test Q yields a positive response both in the case when $|Q \cap \mathcal{P}| \geq \ell$ and in the case when $|Q \cap \mathcal{P}| \geq 1$ and $|Q \cap \mathcal{I}| \leq u$, with the thresholds ℓ and u being two unknown positive integers. More complicated cases can be that the interference is specific to some certain pairs (x_i, y_j) for some $x_i \in \mathcal{P}$ and $y_j \in \mathcal{I}$, or even that there is a malicious adversary that makes the responses to those tests each of which containing both inhibitors and positive clones potentially different in different tests. Besides the mathematical complexity of dealing with all these variation models, there are also the practical questions of determining which model fits the real need. Accordingly, we make the unpredictability assumption on the general inhibitor model that the response is unpredictable when the test contains both inhibitors and positive clones.

1.1. Previous results and our contributions

Farach et al. first introduced the 1-inhibitor model and gave a randomized algorithm to identify all positives in $O((d+h)\log n)$ tests. Later, De Bonis and Vaccaro (1998) connected the 1-inhibitor model to a certain generalization of superimposed codes (D'yachkov et al., 1983), and provided a lower bound $\Omega(\frac{h^2}{d\log h}\log n)$ on the number of required tests for any algorithm that can identify exactly d positives in the presence of h inhibitors. Further, De Bonis et al. (2005) gave an asymptotically optimal 4-stage algorithm for the 1-inhibitor model under the assumption that the exact number of positives and an upper bound on the number of inhibitors are known beforehand. Recently, De Bonis (2008) proposed an almost optimal algorithm using $O(\frac{h^2}{d}\log(n/h))$ tests under the hypothesis that the exact number d of positives is given. It is remarkable that this algorithm is a trivial two-stage algorithm, where potential candidates narrowed down by the first stage are tested individually in the second stage. However, all those algorithms mentioned above are sequential; specifically, tests cannot be performed in parallel.

In the first part of this paper, we deal with the problems involving both complications (a) and (c). We focus on designing efficient nonadaptive algorithms for the sorts of pooling design problems with the

presence of inhibitors and errors. D'yachkov et al. (2001) were the first to give a non-trivial nonadaptive algorithm (i.e., all tests are specified in advance) for the 1-inhibitor model. They exploited a $(d+h)$-disjunct matrix as a pooling design whose decoding complexity is asymptotically $O\left(\sum_{h'\le h}\left(t(n-h')\binom{n}{h'}\right)\right)$, where t is the number of tests needed. Notice that a well-known upper bound on the number of rows of $(d+h)$-disjunct matrices with n columns is $O((d+h)^2 \log n)$ (Du et al., 2006). Further, Hwang and Liu (2003) gave an error-tolerant nonadaptive algorithm that can correct up to e erroneous outcomes by using a $(d+h+2e)$-disjunct matrix and reduced the decoding complexity down to $O\left(t(n-|I|)\binom{|I|}{h}\right)$, where I is a set containing all inhibitors but no positives. On the other hand, De Bonis (2008) gave a lower bound by proving that the number of tests used by any nonadaptive algorithm for the 1-inhibitor model is bounded below by the minimum number of rows of $(d+h-1)$-disjunct matrices with n columns. This lower bound on the number of tests required for the 1-inhibitor model asymptotically differs by a $1/\log(d+h)$ factor from the best known upper bound $O((d+h)^2 \log n)$ (D'yachkov et al., 2001). Later, Hwang and Chang (2007) observed that the previous results (D'yachkov et al., 2001; Hwang et al., 2003) on the decoding procedure and analysis for the 1-inhibitor model also work for the general inhibitor model, suggesting that the applicability is widely extended while no extra tests are required. However, the decoding complexity of previous results are all growing with an exponential rate, $O(n^h)$. In this paper, we develop a new idea for the general inhibitor model with an efficient decoding procedure, taking $O((d+h)^2 n \log n)$ time, that recovers the set of positives $\mathcal{P}$ from the knowledge of the tests outcomes. Notice that our algorithm has a significant improvement in decoding complexity and the number of tests required is asymptotically as well as compared to the best known results.

In recent years, a great deal of effort has been made on the inhibitor models, especially for identifying all positive clones. What seems to be lacking, however, is to classify all clones, that is, not just identify all positive clones but also identify all inhibitors and negative clones. Of particular note is that one cannot simply test every individual and thus classify all clones. The reason for this is that one cannot distinguish a negative clone from an inhibitor in pools without any positive clone. Although identifying the inhibitors can be very important, little is known about constructing pooling designs to accomplish this. So far, known results on classifying all clones are sequential. There remains an open problem whether there exists a nonadaptive algorithm for the classification problem of the inhibitor model. In this article, we answer this problem by providing an efficient nonadaptive algorithm for classifying all clones in the 1-inhibitor model. It is remarkable that the pooling design we propose has a polynomial-time decoding procedure that recovers the three types of clones from the knowledge of the tests outcomes. Furthermore, we extend our results to the k-inhibitor model.

In the second part of this article, we turn our attention to the problems concerning all the three complications mentioned. In molecular biology, the biological objects (e.g., clones, molecules, proteins) are being identified while it remains a challenge to learn how they cooperate to produce various attributes. One modern application is the reconstruction of protein interaction networks (Lappe et al., 2003) by experiments that signal the presence of interaction in a pool of proteins. In this setting, the property to be screened can be defined on subsets of clones, instead of on individual ones. Such a model is usually called the *complex model* in the group testing literature. To deal with such a complication along with the presence of inhibitors, we study a synthetic model where the inhibitor model and the complex model are combined together. This model is referred to as the *inhibitor complex model* introduced by Chang et al. We show that the idea of designing an efficient decoding procedure used in the inhibitor models still works in their complex version.

The remainder of this article is organized as follows. In Section 2, we introduce some notations and definitions. In Section 3, we study the identification problem on the general inhibitor model. We provide an efficient nonadaptive algorithm based on a new structure introduced in this article, and then analyze upper and lower bounds on the number of tests required for identifying all positives on the general inhibitor model. In Section 4, we deal with the classification problem on the 1-inhibitor model. We first give an efficient nonadaptive algorithm that identifies all inhibitors, and then propose an nonadaptive algorithm that can classify all clones on the 1-inhibitor model. We also provide a lower bound on the number of tests required for the classification problem of the 1-inhibitor model. Moreover, we extend our results to the k-inhibitor model. In Section 5, we study inhibitor models defined on complexes by extending the results in previous sections to their complex versions and give some constructions of pooling designs whereby we solve the problems on the inhibitor complex model. Finally, Section 6 provides our concluding remarks.

2. PRELIMINARIES

We start with some notations and definitions. Throughout this article, a pooling design is represented by a 0-1 matrix where columns correspond to objects, rows correspond to tests, and cell $(i, j) = 1$ signifies that the j-th object is in the i-th test whereas $(i, j) = 0$ for otherwise. An example of the representation along with the complications is shown in Figure 1.

For convenience, a column (row) can be treated as the set of row (column) indices where the column (row) has a 1-entry. Viewing a column as a set of row indices, $C \cup C'$ and $C \cap C'$ are plainly to be defined as set union and set intersection for columns C and C'. We say that a set of columns X appears (or is contained) in a row means all columns in X have a 1-entry in that row. A pool with a negative (positive) outcome is called a negative (positive) pool. For a column C, denote $t_0(C)$ ($t_1(C)$) as the number of negative (positive) pools in which the column C appears. The followings are our basic preparations for pooling designs.

Definition 2. *A binary matrix is $(d; z)$-disjunct if for any $d+1$ columns $C_0, C_1, \cdots, C_d$,*

$$\left| C_0 \setminus \bigcup_{i=1}^{d} C_i \right| \geq z.$$

Definition 3. *A binary matrix is $(h; y)$-inclusive if for any $h+1$ columns $C_0, \cdots, C_h$,*

$$\left| C_0 \cap \left(\bigcup_{i=1}^{h} C_i \right) \right| \leq y.$$

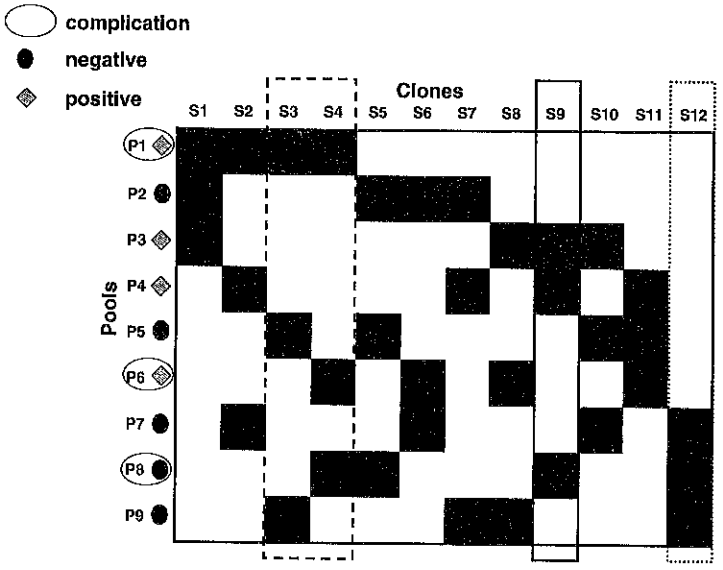


FIG. 1. An example of a pooling design for 12 clones with $S9$ being the unique positive clone. A false-positive observation, complication (a), occurs in pool $P6$. Pool $P1$ occurs complication (b), a positive response caused by the pair $\{S3, S4\}$ of clones. Pool $P8$ yields a negative response because of the inhibition from $S12$, complication (c). Notice that this example is only to demonstrate a situation where the complications occur. In reality, we have no idea about where exactly the complications occur as soon as the test outcomes are obtained.

3. IDENTIFICATION PROBLEMS ON THE GENERAL INHIBITOR MODEL

3.1. Methods

In this section, we deal with the problem of identifying all positives on the general inhibitor model. For convenience, we attach parameters (n, d, h) to inhibitor models with each parameter corresponding to the number referred in Definition 1. We now present a pooling design based on a $(d; z)$-disjunct and $(h; y)$-inclusive matrix with $z - e > y + e$ and then demonstrate how to recover the set $\mathcal{P}$ from the outcomes.

Theorem 3.1. *For the (n, d, h) general inhibitor model with at most e erroneous outcomes, a $(d; z)$-disjunct and $(h; y)$-inclusive matrix with $z - e > y + e$ can identify all positives in $\mathcal{N}$.*

Proof. Let M be a $(d; z)$-disjunct and $(h; y)$-inclusive matrix of n columns, $z - e > y + e$, corresponding to our pooling design. Consider a positive item C^+ and a set I of inhibitors, no more than h. By the $(h; y)$-inclusiveness property, there are at most y rows each intersecting C^+ and some of I. Therefore, C^+ appears in at most y negative pools because of inhibitors. Even for the worst case that e pools are erroneous, C^+ appears in at most $y + e$ negative pools, i.e., $t_0(C^+) \leq y + e$.

To successfully identify all positives, it is sufficient to prove that $t_0(C) > y + e$ for all non-positive item C. Consider a non-positive item C and a set P of positives, no more than d. By the $(d; z)$-disjunctness property, there are at least z rows each intersecting C but none of P. The outcomes of the pools corresponding to these z rows should be negative if no errors occur. Therefore, we conclude that $t_0(C) \geq z - e$, which implies $t_0(C) > y + e$.

From the above discussion, we can determine through the function $t_0(C)$ whether an item C is positive or not. ■

In the following, we propose a polynomial-time decoding procedure, called the *cut-off method*, that recovers the set of all positives from the knowledge of the tests outcomes by using a $(d; z)$-disjunct and $(h; y)$-inclusive matrix with $z - e > y + e$. Figure 2 is an example of Algorithm 1.

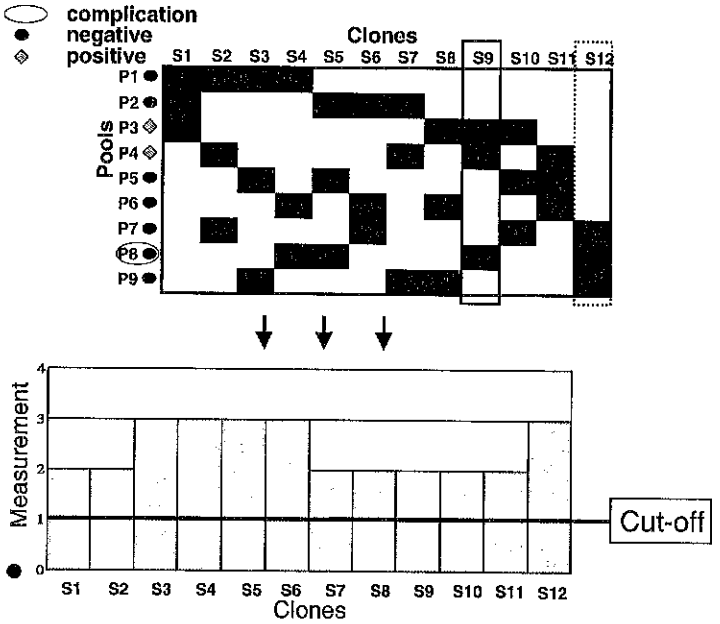


FIG. 2. An example of the cut-off method. **(Top)** We use a (1; 2)-disjunct and (1; 1)-inclusive matrix to be a pooling design for 12 clones. Suppose that no erroneous outcome occurs and that there exist at most a positive clone, say $S9$, and at most an inhibitor, say $S12$. From the assumption, we can easily determine the cut-off value 1 in advance. **(Bottom)** The measurement value, determined by the function t_0, of each clone whereby we can determine that $S9$ is the positive clone because of $t_0(S9) \leq 1$. In fact, this example also shows that the pooling design is not able to determine which one is the inhibitor.

Algorithm 1 FIND-P

```
Use a (d; z)-disjunct and (h; y)-inclusive matrix with z − e > y + e
P ← ∅
for each item C ∈ 𝒩 do
    Compute t_0(C)
    if t_0(C) ≤ y + e then
        P ← P ∪ {C}
    end if
end for
Return P.
```

We now estimate the time complexity required for the decoding algorithm. It is easily seen that the decoding procedure is to compute $t_0(C)$ once for every item C. Moreover, the cost of each single computation takes $O(t)$ time where t is the number of tests required for the pooling design. Hence, we conclude that the total cost of the decoding complexity is $O(tn)$. In the following, we will derive upper and lower bounds on the number t.

3.2. *Upper and lower bounds on the number of tests*

One way to construct such a matrix with the property mentioned in Theorem 3.1 is to have each column having weight at least w and each pair of columns containing at most λ elements in common. Using the column-intersection rule as mentioned above, we have that any h columns intersect another column at no more than $h\lambda$ rows and that there are at least $w - d\lambda$ rows in which any column does not be covered by a set of any d other columns. By requesting $w - d\lambda - e > h\lambda + e$, which implies $w > (d+h)\lambda + 2e$, we have that the constructed matrix is $(d; z)$-disjunct and $(h; y)$-inclusive with $z - e > y + e$ where $z = w - d\lambda$ and $y = h\lambda$. Thus, the following result is obtained.

Lemma 3.2. *Let M be a matrix such that every column has weight at least w, and every pair of two columns intersects at no more than λ rows. If $w > (d+h)\lambda + 2e$, then M is $(d; z)$-disjunct and $(h; y)$-inclusive with $z - e > y + e$ where $z = w - d\lambda$ and $y = h\lambda$.*

It is worth pointing out that Hwang and Sós's (1987) construction of disjunct matrices satisfies the above conditions and provides us a way to construct the desired matrix. We now exploit their construction to analyze and estimate the number of tests required for our algorithm. Given integers t and k, Hwang and Sós (1987) construct a $t \times n$ matrix with $w = 4kl$ and $\lambda = 4l - 1$, where $n \geq (2/3)3^{t/16k^2}$ and $l = \lceil t/16k^2 \rceil$. Accordingly, we obtain the following result immediately.

Lemma 3.3. *Given integers d, h, e and t, there exists a $t \times n$ $(d; z)$-disjunct and $(h; y)$-inclusive matrix with $z - e > y + e$ such that $n \geq (2/3)3^{t/16k^2}$, where $k > ((4l-1)(d+h) + 2e)/4l$.*

By setting $k = d + h + 2e$, we have the following results.

Theorem 3.4. *For any integers d, h, e and n, there exists a $(d; z)$-disjunct and $(h; y)$-inclusive matrix of n columns satisfying $z - e > y + e$ and the number of rows $t \leq 16(d+h+2e)^2 \log(3n/2)/\log 3$.*

Theorem 3.5. *For the (n, d, h) general inhibitor model with at most e erroneous outcomes, there exists a nonadaptive algorithm that recovers the set of positives using $O((d+h+e)^2 \log n)$ tests and $O((d+h+e)^2 n \log n)$ decoding time.*

Proof. The result follows immediately from Theorems 3.4 and 3.1. ■

The following result uncovers a relation between our design and disjunct matrices, and consequently we conclude that the number of tests required is lower bounded by the number of rows of $(d+h; 2e+1)$-disjunct matrices.

Theorem 3.6. *A matrix M which is $(d; z)$-disjunct and $(h; y)$-inclusive with $z - e > y + e$ is $(d+h; 2e+1)$-disjunct.*

Proof. For any $d+h+1$ columns $C_0, C_1, \cdots, C_{d+h}$ in M, we have

$$\begin{cases} \left| C_0 \cap \left(\bigcup_{i=1}^{h} C_i \right) \right| \leq y, \\ \left| C_0 \setminus \bigcup_{i=h+1}^{d+h+1} C_i \right| \geq z. \end{cases}$$

From the above two inequalities, we obtain the result that $\left| C_0 \setminus \bigcup_{i=1}^{d+h} C_i \right| \geq z - y > 2e$, for any $d+h+1$ columns $C_0, C_1, \cdots, C_{d+h}$ in M. The theorem follows directly from the definition of $(d+h;\ 2e+1)$-disjunctness. ■

4. CLASSIFICATION PROBLEMS ON THE INHIBITOR MODELS

The problem we consider in this section is to classify all items on the inhibitor models, instead of identifying only positives. In order to distinguish inhibitors from negatives, we need to make an additional assumption that among the given n items there exists at least one positive item. The reason for this is that one cannot distinguish negative items from inhibitors in the pools without any positive items.

The instrument of our design is a generalized disjunct matrix defined as follows.

Definition 4. *A binary matrix is said to be $(d, r; z]$-disjunct if for any $d+r$ columns $C_1, C_2, \cdots, C_{d+r}$,*

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

It is easy to see that a $(d; z)$-disjunct matrix is equivalent to a $(d, 1; z]$-disjunct matrix. Let $t(n, d, r; z]$ denote the minimum number of rows among all $(d, r; z]$-disjunct matrices with n columns. Stinson and Wei (2004) gave a lower bound

$$t(n, d, r; z] \geq 0.7c \frac{(d+r)\binom{d+r}{r}}{\log \binom{d+r}{r}} \log n + \frac{c(z-1)}{2} \binom{d+r}{r} \tag{1}$$

when n is sufficient large, where c is a constant. Recently, Chen et al. (2006) provided an upper bound

$$t(n, d, r; z] < z(k/r)^r (k/d)^d [1 + k(1 + \ln(n/k+1))], \tag{2}$$

where $k = d + r$.

4.1. The 1-inhibitor model

We start with an algorithm for the 1-inhibitor model of identifying only inhibitors. An interesting feature is that a trivial strategy does not work on the 1-inhibitor model of identifying only inhibitors. Obviously, one cannot simply test every single individual and then identify all inhibitors. The nonadaptive algorithm we propose here improves previous results in the number of stages required to perform experiments.

Theorem 4.1. *Assume that there is at least one positive item in $\mathcal{N}$. For the (n, d, h) 1-inhibitor model with at most e erroneous outcomes, there exists a nonadaptive algorithm, corresponding to an $(h, 2; 2e+1]$-disjunct matrix of n columns, that can identify all inhibitors in $\mathcal{N}$ using $O(eh^3 \log n)$ tests and $O(eh^3 n \log n)$ decoding time.*

Proof. Let M be an $(h, 2; 2e+1]$-disjunct matrix of n columns corresponding to our pooling design. Consider a positive item C^+ and an h-subset I which contains all inhibitors. By the $(h, 2; 2e+1]$-disjunctness property of M, there exist at least $2e+1$ rows each intersecting C^+ but none of I. The pools corresponding to these rows should be positive except erroneous pools. Even for the worst case that e pools are erroneous, C^+ still appears in at least $e+1$ positive pools, i.e., $t_I(C^+) \geq e+1$.

Consider a negative item C^-, a positive item C^+ and an h-subset I which contains all inhibitors. By the $(h, 2; 2e+1]$-disjunctness property, there exist at least $2e+1$ rows each intersecting C^- and C^+, but none of I. A similar argument implies that $t_I(C^-) \geq e+1$. On the other hand, even for the worst case that e outcomes are erroneous, $t_I(C) \leq e$ for every inhibitor C. Therefore, we can separate all inhibitors from the others.

Trivially, this reasoning also yields a decoding procedure using $O(tn)$ time that recovers the set of inhibitors, where t is the number of tests required. ■

Another interesting feature of the problem of identifying only inhibitors is that the number of tests required does not depend on the number of positive items. This is quite different from what happens in the problem of identifying all positives in the presence of inhibitors for which the number of inhibitors is significant to the number of tests required for identifying all positives.

From Theorem 4.1, we directly obtain a method to design a two-stage algorithm for the 1-inhibitor model of identifying all the positives and inhibitors. In the first stage we use an $(h, 2; 2e+1]$-disjunct matrix to identify and eliminate all inhibitors, and then exploit a $(d; 2e+1)$-disjunct matrix to identify all positives in the second stage. An idea of combining these two stages provides us a one-stage approach to classify all items. It is quite nature to consider the construction of a matrix which is $(h, 2; 2e+1]$-disjunct and also $(d; 2e+1)$-disjunct after deleting any h columns and all rows intersecting these columns. The pooling design corresponding to such a matrix can be used to identify all positives and inhibitors in a similar way to the two-stage algorithm. By definition, it is easily seen that a $(d+h, 2; 2e+1]$-disjunct matrix is $(h, 2; 2e+1]$-disjunct. Moreover, it preserves the $(d; 2e+1)$-disjunctness property after deleting any h columns and all rows intersecting these columns. For otherwise, there exists a column C and a $(d+h)$-set R of columns such that there are at most $2e$ rows intersecting C but none of R, violating the *$(d+h, 2; 2e+1]$-disjunctness property. Accordingly, we have the following result.*

Theorem 4.2. *Assume that there is at least one positive item in $\mathcal{N}$. For the (n, d, h) 1-inhibitor model with at most e erroneous outcomes, there exists a nonadaptive algorithm, corresponding to a $(d+h, 2; 2e+1]$-disjunct matrix of n columns, that can classify all items in $\mathcal{N}$ using $O(e(d+h)^3 \log n)$ tests and $O(e(d+h)^3 n \log n)$ decoding time.*

Proof. Let M be a $(d+h, 2; 2e+1]$-disjunct matrix of n columns corresponding to our pooling design. The $(d+h, 2; 2e+1]$-disjunctness property, which implies $(h, 2; 2e+1]$-disjunctness, can identify all inhibitors according to Theorem 4.1. After eliminating the up-to-h columns which represent the inhibitors and all rows which intersect these columns, the resulting matrix remains $(d; 2e+1)$-disjunct due to the $(d+h, 2; 2e+1]$-disjunctness property. Therefore, the up-to-d positives can be identified. Trivially, the reasoning above yields a decoding procedure using $O(tn)$ time that classify all items, where t is the number of tests required. ■

Algorithm 2 is the decoding procedure corresponding to Theorem 4.2.

Algorithm 2 CLASSIFY-ALL-ITEMS

```
1:  Use a (d+h, 2; 2e+1]-disjunct matrix M.
2:  P ← ∅, I ← ∅
3:  for each item C ∈ N do
4:     Compute t_I(C)
5:     if t_I(C) ≤ e then
6:        I ← I ∪ {C}
7:     end if
8:  end for
9:  Delete all columns corresponding to I and all rows intersecting these columns from M.
10: for each item C ∈ N\I do
11:    Compute t_0(C)
12:    if t_0(C) ≤ e then
13:       P ← P ∪ {C}
14:    end if
15:    end for
16: Return P and I.
```

In the following, we derive a lower bound on the number of tests required for the 1-inhibitor model of classifying all items.

Theorem 4.3. *Assume that there is at least one positive item in $\mathcal{N}$. For the (n, d, h) 1-inhibitor model, any nonadaptive algorithm that successfully classifies all items in $\mathcal{N}$ uses at least $\{t(n, d+h-1, 1; 1], t(n, h-1, 2; 1]\} = \max\{\Omega(\frac{(d+h)^2}{\log(d+h)}\log n), \Omega(\frac{h^3}{\log h}\log n)\}$ tests.*

Proof. Let M be a matrix corresponding to a nonadaptive algorithm that classifies all items in $\mathcal{N}$ for the 1-inhibitor model. The bound $t(n, d+h-1, 1; 1]$ follows directly by the bound (De Bonis, 2008) for identifying only positives.

Suppose that M is not $(h-1, 2; 1]$-disjunct. By definition, there are $h+1$ columns $C_0, C_1, \cdots, C_h$ such that $(C_0 \cap C_1) \subseteq \bigcup_{i=2}^{h} C_i$. Consider the case that C_1 is corresponding to the unique positive item and $C_2, C_3, \cdots, C_h$ are inhibitors. In this case one cannot determine whether C_0 is an inhibitor or a negative item since C_0 always appears in negative pools. Thus, we have the lower bound $t(n, h-1, 2; 1]$. Combining the above discussion and the inequality (1), the proof is complete. ■

4.2. The k-inhibitor model

In this section we consider the k-inhibitor model where a pool yields a positive response if and only if it contains at least one positive clone and less than k inhibitors. Assume that the threshold k is known in advance.

Theorem 4.4. *Assume that there is at least one positive item in $\mathcal{N}$. For the (n, d, h) k-inhibitor model with at most e erroneous outcomes, there exists a nonadaptive algorithm, corresponding to an $(h-k+1, k+1; 2e+1]$-disjunct matrix of n columns, that can identify all inhibitors in $\mathcal{N}$.*

Proof. Let M be an $(h-k+1, k+1; 2e+1]$-disjunct matrix of n columns corresponding to our pooling design. Consider a positive item C, a k-subset Y not a subset of the set of inhibitors and an $(h-k+1)$-subset Z which contains either all inhibitors not in Y or $h-k+1$ inhibitors. By the $(h-k+1, k+1; 2e+1]$-disjunctness property, there exist at least $2e+1$ rows each intersecting C and all of Y, but none of Z. The pools corresponding to these rows should be positive except erroneous pools. Even for the worst case that e pools are erroneous, the k-subset Y still appears in at least $e+1$ positive pools, i.e., $t_1(Y) \geq e+1$. On the other hand, $t_1(X) \leq e$ for every k-subset X consisting of inhibitors. Let $\mathcal{O} = \{C \in X : t_1(X) \leq e \text{ for each } k\text{-subset } X\}$. From the discussion above, we conclude that $\mathcal{O}$ is the set consisting of inhibitors. ■

In the 1-inhibitor model, we exploited two disjunct matrices separately to design a two stage algorithm for classifying all items in $\mathcal{N}$. Similarly, there is a two-stage algorithm for the k-inhibitor model by replacing an $(h, 2; 2e+1]$-disjunct matrix in the first stage with an $(h-k+1, k+1; 2e+1]$-disjunct matrix. A slight modification in the k-inhibitor model is that we only need to eliminate at most $h-k+1$ inhibitors from $\mathcal{N}$ so that the remaining inhibitors, at most $k-1$, do not obscure the positives.

Theorem 4.5. *Assume that there is at least one positive item in $\mathcal{N}$. For the (n, d, h) k-inhibitor model with at most e erroneous outcomes, there exists a nonadaptive algorithm, corresponding to a $(d+h-k+1, k+1; 2e+1]$-disjunct matrix of n columns, that can classify all items in $\mathcal{N}$.*

Proof. It follows by the fact that a $(d+h-k+1, k+1; 2e+1]$-disjunct matrix is $(h-k+1, k+1; 2e+1]$-disjunct, and the resulting matrix obtained by deleting any $h-k+1$ columns and all rows intersecting these columns remains $(d, k+1; 2e+1]$-disjunct. ■

This decoding algorithm is similar to that of Theorem 4.2 except replacing the concept "item" with "k-subset" in Algorithm 2. An analogous argument shows that the decoding complexity of this algorithm is $O\left(\binom{n}{k}kt\right)$ in the worst case, since each operation of computing $t_1(X)$ takes $O(kt)$ time where t is the number of tests needed.

5. THE INHIBITOR COMPLEX MODEL

As versus the clone model discussed so far, we consider its natural generalization—the complex model, where the property to be screened is defined on a subset of clones, called *complex*. From an application's point of view, the problem of group testing on complexes is worthy of being studied since the collective appearance of some molecules rather than a single molecule would cause a certain given biological feature in some biological phenomenon. Torney (1999) introduced the concept of the complex model and gave some substances in eukaryotic DNA transcription and RNA translation as examples of complexes. In the complex model, a fixed but unknown set of complexes are designated *positive complexes* whereas the other complexes are called *negative complexes*. A group test is executed on a subset of the given collection of clones and yields a positive outcome only when it contains at least one positive complex. Therefore, specifically, the theme of this section are pooling designs for identifying an unknown family of positive complexes from a given collection of clones by group tests.

Group testing on complexes is widely applied in modern molecular and cellular biology. A prominent example is its application on identification of protein-to-protein interactions (Li et al., 2005; Lappe et al., 2003). The interactions between proteins are important for many biological functions. For example, in signal transduction process, the conveyance of signals from the exterior of a cell to the inside of that cell is by protein-protein interactions of the signaling molecules. This process plays a fundamental role in many biological projects. Thus identifying all protein-to-protein interactions is an important task for many biological processes in living cells; furthermore, information about these interactions improves our understanding of diseases and provides the basis for new therapeutic approaches. The development of some laboratory approaches (Lappe et al., 2003) enables the application of group testing to this problem. Li et al. (2005) formulated this identification problem as a group testing problem in bipartite graphs which can be regarded as a special case of group testing on complexes. Besides the protein-protein interactions problem, some other problems such as graph testing, superimposed codes and secure key distribution (Chen et al., 2007) are also highly related to the complex model. Recent developments in this topic can be found elsewhere (Chen et al., 2008; Du et al., 2006; Gao et al., 2006; Macula et al., 2000, 2004).

In this section, we focus on *inhibitor complex model* introduced by Chang et al., where an inhibitor is a third type of complexes. As mentioned in the clone model, the presence of an inhibitor may cancel the effect of positive complexes; in other words, a group test executed on a set of clones containing an inhibitor may yield a negative outcome even if that set contains a positive complex. Furthermore, we can subdivide the inhibitor complex model into the 1-inhibitor, k-inhibitor and general inhibitor models based on the interference effect between positive complexes and inhibitors. For instance, in k-inhibitor model a pool of clones inducing more than k inhibitors would yield a negative response. Indeed, the concept "inclusiveness" used for inhibitor clone models still works on inhibitor complex models and identification of inhibitors can also be done under some natural assumptions.

Throughout the rest of this article, we consider the inhibitor complex problems on a given family $\mathcal{C}$ of subsets of a collection $\mathcal{N}$ of n items. It is known beforehand that the family $\mathcal{C}$ consists of at most d positive complexes, at most h inhibitors and some negative complexes, where every complex in $\mathcal{C}$ consists of at most r items. We extend the parameters used in clone model to (n, d, h, r) to include an additional information of maximum size of a complex. Some natural assumptions on complexes will be added in accordance with the goals of these problems. Of particular note is the standard assumption that members of $\mathcal{C}$ are subject to non-inclusion. The reason for this is that no positive complex can include any other positive complex or inhibitor.

5.1. *Identification problems*

In this subsection we study the problem of identifying all positive complexes for the general inhibitor complex model. Recall that a binary matrix is said to be $(d, r; z)$-*disjunct* if for any $d+r$ columns $C_1, \cdots, C_{r+d}$, $\left|\bigcap_{i=1}^{r} C_i \setminus \bigcup_{i=r+1}^{d+r} C_i\right| \geq z$. Chang et al. proved that a $(d+h, r; 2e+1)$-disjunct matrix can identify all positive complexes under the general inhibitor complex model with at most e erroneous outcomes. However, it is a little bit expensive that the corresponding decoding procedure counts $t_0^S(X)$ for each $S \in \binom{\mathcal{N}}{h}$ for each candidate complex $X \in \mathcal{C}$, where $t_0^S(X)$ is the number of negative pools containing X

but none of S and $\binom{\mathcal{N}}{h}$ denote the family consisting of all h-subsets of $\mathcal{N}$. Here, we provide an efficient design for this model and its decoding procedure is to compute $t_0(X)$ for each candidate complex $X \in \mathcal{C}$, where $t_0(X)$ denotes the number of negative pools all columns in X appears in. The improvement on decoding ability is attributed to the adoption of inclusiveness property in our designs.

Definition 5 *A matrix is* $(h, r; y)$-inclusive *if for any* $h+r$ *columns* $C_1, \cdots, C_{r+h}$,

$$\left|\left(\bigcap_{i=1}^{r} C_i\right) \bigcap \left(\bigcup_{i=r+1}^{r+h} C_i\right)\right| \leq y.$$

Furthermore, in Section 5.3 we will show that some known disjunct matrices retain the generalized inclusiveness property well.

Lemma 5.1. *A matrix which is* $(d, r; z)$*-disjunct and also* $(h, r; y)$*-inclusive with* $z - e > y + e$ *is* $(d+h, r; 2e+1)$*-disjunct.*

Proof. In this proof we will use an argument similar to the one exploited in Theorem 3.6. For any $r+d+h$ columns $C_1, \cdots, C_{r+d+h}$, there exist z rows with 1-entries in $C_1, \cdots, C_r$ and 0-entries in $C_{r+1}, \cdots, C_{r+d}$ and at most y rows with 1-entries in $C_1, \cdots, C_r$ and at least one 1-entry in $C_{r+d+1}, \cdots, C_{r+d+h}$, so the number of rows with 1-entries in $C_1, \cdots, C_r$ and 0-entries in $C_{r+1}, \cdots, C_{r+d+h}$ is at least $z - y > 2e$. The result follows directly. ■

5.1.1. The general inhibitor complex model. By Lemma 5.1, we have the following theorem immediately. To show the decoding ability of such a matrix, we give an alternative proof here.

Theorem 5.2. *For the* (n, d, h, r) *general inhibitor complex model with at most* e *erroneous outcomes, a* $(d, r; z)$*-disjunct and* $(h, r; y)$*-inclusive matrix with* $z - e > y + e$ *can identify all positive complexes.*

Proof. Consider a positive complex P and let $\{X_1, \cdots, X_h\}$ denote a set of other complexes containing all inhibitors. Under the hypothesis that no complex is contained in another, there exist $v_i \in X_i \setminus P$ for $1 \leq i \leq h$. By $(h, r; y)$-inclusiveness property, the number of pools containing P and at least one of v_i's is at most y. Hence, P can only appear in at most y negative pools if there is no error. This implies $t_0(P) \leq y + e$.

Conversely, consider a complex $X \in \mathcal{C}$ which is not positive. Similarly, there exists an item $v \in P \setminus X$ for each positive complex P. Let B be a set of these v's. By $(d, r; z)$-disjunctness property, there are at least z rows containing X and none of B. This shows that the pools corresponding to these rows each yields a negative outcome if there is no error. Even in the worst case that all errors occur in these pools, we have that $t_0(X) \geq z - e > y + e$. Hence, we conclude that $\{X: t_0(X) \leq y + e\}$ is the set of positive complexes. ■

5.1.2. The k*-inhibitor complex model.* The $(d+h-k+1, 1; 2e+1)$-disjunct matrix has been used to solve the k-inhibitor clone model where at most e erroneous outcomes are allowed. It is easily extended to the complex model as follows.

Corollary 5.3. *For the* (n, d, h, r) *k-inhibitor complex model with at most* e *erroneous outcomes, a* $(d+h-k+1, r; 2e+1)$*-disjunct matrix can identify all positive complexes.*

According to Theorem 5.1 and Corollary 5.3, we can derive the following 1-stage algorithm that identifies all positive complexes on the k-inhibitor complex model. Again, we give another proof that indicates the decoding process.

Theorem 5.4. *For the* (n, d, h, r) *k-inhibitor complex model with at most* e *erroneous outcomes, a matrix which is* $(d, r; z)$*-disjunct and also* $(h-k+1, r; y)$*-inclusive with* $z - e > y + e$ *can identify all positive complexes.*

Proof. By a similar proof as in Theorem 5.1, we have $t_0(X) \geq z - e$ for each X being a negative complex or an inhibitor. On the other hand, let P be a positive complex and $\{X_1, \cdots, X_{h-k+1}\}$ be a set of other complexes containing as many inhibitors as possible. Since no complex is included in another, we have $v_i \in X_i \setminus P$ for $1 \leq i \leq h-k+1$. By $(h-k+1, r; y)$-inclusiveness property, the number of pools containing both P and at least one of v_i's is no more than y. Since a pool containing P and none of these v_i's would be tested positive, P can only appear in at most y negative pools if there is no error. Thus $t_0(P) \leq y + e$. We conclude that $\{X: t_0(X) \leq y + e\}$ is the set of positive complexes. ■

The decoding procedure shown in Theorem 5.4 is to count $t_0(X)$ once for each complex $X \in \mathcal{C}$, whereas Corollary 5.3 suggested a decoding algorithm of counting $t_0^S(X)$ for each $S \in \binom{\mathcal{N}}{h-k+1}$ for each complex $X \in \mathcal{C}$.

Notice that plugging $r = 1$ into Theorem 5.4 leads to a 1-stage algorithm for the k-inhibitor clone model.

5.2. *Classification problems on 1-inhibitor complex model*

In order to distinguish inhibitors from other complexes, we need some essential assumptions on complexes. It is naturally assumed that for each negative complex, there is always a positive complex such that no inhibitor is included in their union; otherwise, the recognition of such negative complex will be ambiguous, i.e., it can be recognized as either negative or inhibitory due to the fact that it appears in negative pools only.

Theorem 5.5. *Assume that there is at least one positive complex. For the (n, d, h, r) 1-inhibitor complex model with at most e erroneous outcomes, an $(h, 2r; 2e+1]$-disjunct matrix can identify all inhibitors.*

Proof. Consider a positive complex P and let $\{X_1, \cdots, X_h\}$ be a set of other complexes containing all inhibitors. Since no complex is contained in another, there exist $v_i \in X_i \setminus P$ for $1 \leq i \leq h$. By $(h, 2r; 2e+1]$-disjunctness property, there exist at least $2e+1$ rows each containing P but none of v_i's. The pools corresponding to these rows should be tested positive if no erroneous outcome occurs. Hence, we have that $t_1(P) > e$ even in the worst case that e erroneous outcomes occur.

Next, consider a negative complex R. According to the assumption on complexes, there exists a positive complex P such that there is an item $v \in I \setminus (P \cup R)$ for each inhibitor I. Let B denote the set of these v's. By $(h, 2r; 2e+1]$-disjunctness property, there exist at least $2e+1$ rows each containing P and R, but none of B. Hence, we have that $t_1(R) \geq e+1$ despite e erroneous outcomes.

On the other hand, $t_1(X) \leq e$ for any inhibitor X since an inhibitor appears in a positive pool only when an erroneous outcome occurs. Thus, it is easily seen that $\{X: t_1(X) \leq e\}$ is the set of inhibitors. ■

Similarly, Theorem 5.5 also provides a two-stage approach to identify all positive complexes and inhibitors. First, use an $(h, 2r; 2e+1]$-disjunct matrix to identify all inhibitors and then use a $(d, r; 2e+1]$-disjunct matrix to identify positive complexes from the unidentified complexes. Combining these two disjunctness properties together, a one-stage algorithm can be proposed.

Theorem 5.6. *Assume that there is at least one positive complex. For the (n, d, h, r) 1-inhibitor complex model with at most e erroneous outcomes, a $(d+h, 2r; 2e+1]$-disjunct matrix can classify all complexes in $\mathcal{C}$.*

Proof. First, since a $(d+h, 2r; 2e+1]$-disjunct matrix is $(h, 2r; 2e+1]$-disjunct, according to Theorem 5.5, we can identify all inhibitors by computing $t_1(X)$ for each complex X. Let $\{X_1, X_2, \cdots, X_{h'}\}$ be the set of all inhibitors that have been found and A_X be a set of v_i's, where $v_i \in X_i \setminus X$, for a complex X. Therefore, $t_0^{A_P}(P) \leq e$ for any positive complex P since the appearance of a positive complex in a negative pool can occur only when the pool contains an inhibitor or its testing result is fault. On the contrary, let R be a negative complex and B be a set consisting of at most d items chosen by taking an item $v \in X \setminus R$ for each positive complex X. It is easy to see that a $(d+h, 2r; 2e+1]$-disjunct matrix is $(d+h, r; 2e+1]$-disjunct. This implies that there are at least $2e+1$ rows containing R but none of $B \cup A_R$. Hence, we have $t_0^{A_R}(R) > e$ despite e erroneous outcomes. ■

Indeed, the decoding procedure of above design is to distinguish inhibitors from other complexes by the value $t_1(X)$ for each complex $X \in C$ and then find $t_0^{A_X}(X)$ for each non-inhibitor complex X.

5.3. Constructions

As mentioned in the previous section, a matrix with disjunctness and inclusiveness property has a great contribution to simplifying the decoding process; however, constructions of such matrices were rare. Some constructed disjunct matrices are potentially inclusive, especially when the number of rows covering any designated r columns is greater than a fixed constant, as shown in the following.

Lemma 5.7. *Let M be a binary matrix in which the number of rows covering any designated r columns is at least w. Then M is $(h, r; z_h)$-disjunct if and only if M is $(h, r; w - z_h)$-inclusive.*

D'yachkov et al. (2002) gave a simple construction of $(d, r]$-disjunct matrices by taking all k-subsets of N as the rows and then it is further extended to the error-tolerant case (Du et al., 2006). We observe its inclusiveness property as follows.

Theorem 5.8. *The $\binom{n}{k} \times n$ binary matrix where the rows consist of all k-subsets of the set $[n]$, $r \le k \le \min(n-d, n-h)$, is $(d, r; z_d)$-disjunct and $(h, r; y_h)$-inclusive, where*

$$z_d = \binom{n-d-r}{k-r}, \quad y_h = \binom{n-r}{k-r} - z_h.$$

Moreover, $z_d - y_h > 0$ for $h, d << n$.

Proof. It is easily seen that this matrix is $(d, r; \binom{n-d-r}{k-r})$-disjunct for $r \le k \le n-d$. Given an r-set R, the number of rows covering R is $\binom{n-r}{k-r}$. By Lemma 5.7, we have the result. Furthermore, $z_d - y_h = \binom{n-r-d}{k-r} + \binom{n-r-h}{k-r} - \binom{n-r}{k-r} > 0$ for $h, d << n$. ∎

A $T-(v, k, \lambda)$ design is a collection of k-subsets, called blocks, of a set of v points such that for any T points there exist exactly λ blocks containing those T points (Anderson, 1990). Typically, the incidence matrix of a T-design with blocks as rows is not good to be a pooling design for clone models since the number of rows is not smaller than the number of columns by the Fisher inequality. However, T-designs become feasible for a pooling design of complex models since the number of pools we use only need to beat the number $\binom{n}{r}$, all potential candidates of positive complexes. Mitchell and Piper (1988) gave a construction of $(d, r]$-disjunct matrix based on T-designs. We extend their results to an error-tolerant version and extract the inclusiveness property of T-designs.

Theorem 5.9. *A $T-(v, k, \lambda)$ design yields a $t \times v$ $(d, r; \lambda^* - d\lambda]$-disjunct and $(h, r; h\lambda]$-inclusive matrix for $d, h < \min(\lambda^*/\lambda, v - T + 1)$ where*

$$t = \frac{\binom{v}{T}\lambda}{\binom{k}{T}}, r = T - 1, \text{ and } \lambda^* = \lambda\frac{v-T+1}{k-T+1}.$$

Moreover, its error tolerance achieves $\left\lceil \frac{\lambda^* - \lambda(d+h)}{2} \right\rceil - 1$.

Proof. First of all, we consider the inclusiveness property. For any set S of $T-1$ columns, any column not in S can cover all columns of S in at most λ rows. Thus any h columns other than those in S cover all columns of S in at most $h\lambda$ rows for any $1 \le h \le v - T + 1$. This shows that the matrix is $(h, T-1; h\lambda]$-inclusive. Additionally, for any set S of $T-1$ columns, we consider the cardinality of the set $\{(x, B) : B \text{ is a block}, S \subset B \text{ and } x \in B \setminus S\}$, say w. For each point x not in S, there are exactly λ blocks containing $S \cup \{x\}$. Thus $w = (v - T + 1)\lambda$. Without loss of generality, assume that there are λ^* blocks containing S. Obviously, $\lambda^* = \lambda\frac{v-T+1}{k-T+1}$ since $w = \lambda^*(k - T + 1)$. Therefore, by Lemma 5.7, we conclude that this matrix is $(d, T-1; \lambda^* - d\lambda]$-disjunct. ∎

Example. A $3-(q^2+1, q+1, 1)$ design exists for prime power q and its incidence matrix of size $q(q^2+1) \times (q^2+1)$ is $(d, 2; q+1-d]$-disjunct and $(h, 2; h]$-inclusive.

6. CONCLUSION

In this paper, we deal with three common complications in high-throughput screening to make pooling designs appropriate in practice. We present a novel concept "inclusiveness" on pooling designs which

leads to a significant improvement in the decoding procedure. As shown in Algorithm 1 or 2, we can determine whether a clone is positive or not by comparing the measurement value of the clone under functions t_0 or t_1 with a cut-off value which can be calculated in advance. The crucial point is that in our decoding procedure the measurement value is only calculated once for each potential candidate, leading to a considerable reduction in decoding complexity.

This paper developed models and methods for pooling clones in situations when inhibitors and synergy effects exist. A great advantage of our pooling designs is that we treated inhibitors and complexes with synergy effects as features and provided a way to identify them, instead of treating them as bugs. Our methods play an important role not only in DNA sequencing but also in drug discovery, where a large collections of chemical compounds are screened to find highly active compounds. According to pilot experiments in drug discovery, synergy effects are exceedingly common. We believe that a combination of compounds with a synergy effect also has strong drug potential, as does a highly active compound. However, in most research such situations at the early stage of the drug discovery are usually considered a source of contamination. A primary reason for this is that tracking down compounds acting synergistically can be expensive and time-consuming. The results of this article suggest an efficient nonadaptive strategy so that the time required to perform experiments and analyze outcomes can be substantially reduced. Although our pooling strategies are lacking in experimental support, the value is in calling awareness to such a direction for further research.

ACKNOWLEDGMENTS

This research was partially supported by the NSC (97-2115-M-009-011-MY3).

DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Anderson, I. 1990. *Combinatorial Designs: Construction Methods.* Ellis Horwood, New York.

De Bonis, A. 2008. New combinatorial structures with applications to efficient group testing with inhibitors. *J. Comb. Optim.* 15, 77–94.

De Bonis, A., Gasieniec, L., and Vaccaro, U. 2005. Optimal two-stage algorithms for group testing problems. *SIAM J. Comput.* 34, 1253–1270.

De Bonis, A., and Vaccaro, U. 1998. Improved algorithms for group testing with inhibitors. *Inform. Process Lett.* 67, 57–64.

De Bonis, A., and Vaccaro, U. 2003. Constructions of generalized superimposed codes with applications to group testing and conflict resolution in multiple access channels. *Theoret. Comput. Sci.* 306, 223–243.

Chang, F.H., Chang, H., and Hwang, F.K. Pooling designs for clone library screening in the inhibitor complex model. *J. Comb. Optim.* (to appear).

Chen, H.-B., Du, D.Z., and Hwang, F.K. 2007. An unexpected meeting of four seemingly unrelated problems: graph testing, DNA complex screening, superimposed codes and secure key distribution. *J. Comb. Optim.* 14, 121–129.

Chen, H.-B., Fu, H.L., and Hwang, F.K. 2008. An upper bound of the number of tests in pooling designs for the error-tolerant complex model. *Optim. Lett.* 2, 425–431.

Du, D.Z., and Hwang, F.K. 2006. *Pooling Designs and Nonadaptive Group Testing—Important Tools for DNA Sequencing*. World Scientific, Singapore.

D'yachkov, A.G., Macula, A.J., Torney, D.C., et al. 2001. Two models of nonadaptive group testing for designing screening experiments, 63–75. *In* Attkinson, A.C., Hackl, P., and Muller, W.G., eds. *Proc. 6th Int. Workshop in Model Oriented Design and Analysis.* Physica-Verlag, Berlin.

D'yachkov, A.G., Vilenkin, P.A., Macula, A.J., et al. 2002. 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.

D'yachkov, A.G., and Rykov, V.V. 1983. A survey of superimposed code theory. *Probl. Control Inform. Theory* 12, 229–242.

Farach, M., Kannan, S., Knill, E., et al. 1997. Group testing with sequences in experimental molecular biology. *Proc. Compress. Complex. Seq.*, 357–367.

Gao, H., Hwang, F.K., Thai, M., et al. 2006. Construction of $d(H)$-disjunct matrix for group testing in hypergraphs. *J. Comb. Optim.* 12, 297–301.

Hwang, F.K., and Liu, Y.C. 2003. Error-tolerant pooling designs with inhibitors. *J. Comput. Biol.* 10, 231–236.

Hwang, F.K., and Chang, F.H. 2007. The identification of positive clones in a general inhibitor model. *J. Comput. Syst. Sci.* 73, 1090–1094.

Hwang, F.K., and Sós, V.T. 1987. Nonadaptive hypergeometric group testing. *Studia Sci. Math. Hungar.* 22.

Li, Y., Thai, M., Liu, Z., et al. 2005. Protein-to-protein interactions and group testing in bipartite graphs. *Int. J. Bioinform. Res. Appl.* 1, 414–419.

Lappe, M., and Holm, L. 2003. Unraveling protein interaction networks with near-optimal efficiency. *Nat. Biotechnol.* 22, 98–103.

Mitchell, C.J., and Piper, F.C. 1988. Key storage in secure networks. *Discr. Appl. Math.* 21, 215–228.

Macula, A.J., Rykov, V.V., and Yekhanin, S. 2004. Trivial two-stage group testing for complexes using almost disjunct matrices. *Discr. Appl. Math.* 137, 97–107.

Macula, A.J., Torney, D.C., and Villenkin, P.A. 2000. Two-stage group testing for complexes in the presence of errors. *DIMACS Ser. Discr. Math. Theoret. Comput. Sci.* 55, 145–157.

Phatarfod, R.M., and Sudbury, A. 1994. The use of a square array scheme in blood testing. *Statist. Med.* 13, 2337–2343.

Stinson, D.R., and Wei, R. 2004. Generalized cover-free families. *Discr. Math.* 279, 463–477.

Torney, D.C. 1999. Sets pooling designs. *Ann. Comb.* 3, 95–101.

Xie, M., Tatsuoka, K., Sacks, J., et al. 2001. Group testing with blockers and synergism. *J. Am. Statist. Assoc.* 96, 92–102.

Address correspondence to:
Dr. Hong-Bin Chen
Department of Applied Mathematics
National Chiao Tung University
Hsinchu, Taiwan 30010

E-mail: andan.am92g@nctu.edu.tw