Available online at www.sciencedirect.com

ScienceDirect

Journal of Computer and System Sciences 73 (2007) 1090–1094

JOURNAL OF COMPUTER AND SYSTEM SCIENCES

www.elsevier.com/locate/jcss

The identification of positive clones in a general inhibitor model ☆

F.K. Hwang *, F.H. Chang

Department of Applied Mathematics, National Chiao Tung University, Hsinchu 300, Taiwan, ROC

Received 13 April 2006; received in revised form 29 January 2007

Available online 13 March 2007

Abstract

In using pooling designs to identify clones containing a specific subsequence called positive clones, sometimes there exist nonpositive clones which can cancel the effect of positive clones. Various models have been studied which differ in the power of cancellation. Although the various models pose interesting mathematical problems, and ingenious constructions of pooling designs have been proposed, in practice we rarely are sure about the true model and thus about which pooling design to use. In this paper we give a pooling design which fits all inhibitor models, and does not use more tests than in the more specific models. In particular, we obtain a 1-round pooling design for the k-inhibitor model for which only sequential designs are currently known.

Keywords: Pooling design; Inhibitor

1. Introduction

In many DNA experiments, we want to know whether a clone (or a contig, a gene) contains a specific subsequence. If it does, we call the clone *positive*, otherwise, it is *negative*. Group testing is often used to identify the positive clones. A group test is applied to an arbitrary subset of the clones and yields a positive outcome if and only if that subset contains a positive clone (the test does not reveal which or how many), otherwise, the outcome is negative. The goal is to use a minimum number of tests to identify all positive clones. In biological applications, more important than the number of tests is the number of rounds these tests can be performed (all test in the same round are performed parallelly). Note that the choice of clones to be included in a test can only use information from test outcomes of previous rounds. Thus the fewer the rounds, the less information is available in the test design. A 1-round testing scheme is usually referred to as a *pooling design* in biological literature.

In some applications, some negative clones are special in the sense that they can cancel the effect of positive clones in deciding the test outcome. Such clones are called *inhibitors*. Different models can be formulated by considering different cancellation effect. The simplest model was first proposed by Farach et al. [7] in which the presence of a single inhibitor dictates the test outcome to be negative regardless of how many positive clones are in the test. Later, De Bonis and Vaccaro [3] generalized the above model to the k-inhibitor model in which a test has a positive outcome

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

* Corresponding author.
E-mail address: fkhwang@gmail.com (F.K. Hwang).

0022-0000/$ – see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.jcss.2007.03.005

if and only if it contains a positive clone and at most $k-1$ inhibitors (the first model is the special case $k=1$). Pooling designs have been proposed for the 1-inhibitor model [4], and extended to the error-tolerant version [8].

Obviously, there are many other models one can propose as to how many inhibitors can dominate how many positive clones. Besides the cumbersomeness in constructing pooling designs for all these variety models, the main problem is in reality, rarely do we have exact information on the model, thus at a loss of which pooling design to use. In this paper we consider the general inhibitor model which includes all variations of cancellation effect and show that the pooling designs proposed for the 1-inhibitor model and its error-tolerant version are applicable to the general inhibitor model. Thus we can use the same pooling design in the presence of inhibitors without worrying whether the model is correct.

2. The error free version

Consider a set S of n clones including p positive clones and q inhibitors, p and q are unknown except $p \leqslant d$ and $q \leqslant n$. We also do not know the exact cancellation effect between the positive clones and the inhibitors.

A pooling design is usually represented by the incidence matrix M where rows are indexed by tests and columns by clones, i.e., $m_{ij}=1$ if clone j is contained in test i, and $m_{ij}=0$ otherwise. It is convenient to treat a column vector as a subset of the row indices $\{i\colon m_{ij}=1\}$. Then we can talk about a union of columns. A binary matrix is called *d-disjunct* if no column is covered by the union of any other d columns. It is well known [5] that a d-disjunct matrix can identify all positive clones if $p \leqslant d$. In fact, the d-disjunct matrix has become the main tool in constructing pooling designs.

An isolated row is a row containing a single 1-entry. Suppose a d-disjunct matrix M has an isolated row. Then this row can be deleted along with the column it is incident to without affecting the identification of other clones, and the reduced matrix is easily seen to be still d-disjunct. Therefore we assume that the disjunct matrices considered in this paper has no isolated row.

Lemma 1. *In a $(d+x)$-disjunct matrix, a column has at least $x+1$ 1-entries not covered by the union of any other d columns.*

Proof. Suppose to the contrary there exist a column C and a set D of d columns such that C has at most x 1-entries not covered by D. Then these 1-entries can be covered by at most x additional columns since each 1-entry incident to C is also incident to another column (no isolated row). Thus C is covered by at most $d+x$ other columns, violating the assumption of $(d+x)$-disjunctness. □

Theorem 2. *A $(d+r)$-disjunct matrix can identify all p positive clones in S.*

Proof. Let N denote a negative clone, P a positive clone and I an inhibitor. Let $t(C)$ denote the number of negative pools clone C appears in. For a given r-set R of columns, let $t^R(C)$ denote the same except that a 0-outcome is changed to 1 if that row contains a 1-entry from R. R is treated as a candidate set of the inhibitors (or a set to cover the inhibitors if their number is less then r). Define

$$t^*(C)=\min_R t^R(C),$$

where the minimum is over all $\binom{n}{r}$ choices of r-sets of columns. Then

$$t^*(P)=\min_R t^R(P)=t^{R'}(P)=0,$$

where R' is an r-set covering all inhibitors.

On the other hand, by the definition of a $(d+r)$-disjunct matrix, any column C has an 1-entry not covered by any $d+r$ columns. In particular, for $C \in \{N,I\}$, the set of positive clones and R, forming a set of at most $d+r$ columns, must leave at least one 1-entry of C uncovered. Clearly, this uncovered 1-entry is in a negative pool. Thus we have

$$t^*(C)\geqslant 1, \quad \text{for } C \in \{N,I\}.$$

Consequently the set $\{C\colon t^*(C)=0\}$ is the set of positive clones. □

The procedure takes $O(n^{r+1})$ time since for each C we have to try $\binom{n}{r}$ R. We can reduce the number of choice of R by noting

$$t(N) \geqslant r+1 \quad \text{and} \quad t(I) \geqslant r+1.$$

Thus any column C with $t(C) \leqslant r$ must be positive and can be excluded from R.

3. The error-tolerant version

The assumption of the error-tolerant version are same as the error-free version except that $e_{01} + e_{10}$ errors may occur to the outcomes, say, e_{01} errors change a 0-outcome to 1, and e_{10} change a 1-outcome to 0. We assume an upper bound e of $e_{01} + e_{10}$ is known.

The notion of d^z-disjunctness was first raised by Macula [9]. We use the notion (d, z) replaced d^z. A binary matrix is *(d, z)-disjunct* if each column has at least z 1-entries not covered by any other d columns. Note that (d, z)-disjunct is just d-disjunct.

Let M be a $(d + r, c + e + 1)$-disjunct matrix where c is a constant to be fixed later.

Ignoring the inhibitors for the moment, then a positive clone P can appear in a negative pool only if its outcome is one of the e_{10} errors. So when R contains all inhibitors, then

$$t^*(P) = t^R(P) \leqslant e_{10}.$$

On the other hand, for $C \in \{N, I\}$, then C has at least $c+e+1$ 1-entries not covered by the at most d positive columns and the r columns in R for all R before errors. Clearly, these uncovered 1-entries all appear in negative pools. Errors of the e_{01}-type may reduce the number of such negative pools. But still,

$$t^*(C) = \min_R t^R(C) \geqslant \min_R\{c + e + 1 - e_{01}\} = c + e + 1 - e_{01}.$$

Since

$$e \geqslant e_{10} + e_{01},$$

$$t^*(C) \geqslant c + e_{10} + 1 > t^*(P).$$

The problem is we do not know e_{10} to separate P from N and I in general, we consider some special cases:

(i) We know an upper bound $\overline{e_{10}}$ of e_{10} and an upper bound $\overline{e_{01}}$ of e_{01}. Set $c = \overline{e_{01}} + \overline{e_{10}} - e$. Then for $C \in \{N, I\}$, $t^*(C) \geqslant (\overline{e_{01}} + \overline{e_{10}} - e) + e + 1 - e_{01} = \overline{e_{10}} + 1$. Thus $\{C'\colon t^*(C') \leqslant \overline{e_{10}}\}$ is the set of positive clones. In particular, when e is obtained by adding up $\overline{e_{01}}$ and $\overline{e_{10}}$, then $c = 0$.

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

$$t^*(C) \geqslant e + e + 1 - e_{01} \geqslant e + 1 > e_{10} \geqslant t^*(P).$$

Thus $\{C'\colon t^*(C') \leqslant e\}$ is the set of positive clones.

(iii) If $p = d$, then set $c = 0$ and the set $\{C'\colon t^*(C')$ is among the d smallest$\}$ is the set of positive clones.

Note that case 2 is the solution given in [8], and case 3 the solution given in [6] for the 1-inhibitor model.

4. The *k*-inhibitor model

De Bonis and Vaccaro [1,2] proposed a 4-stage scheme for the k-inhibitor model:

stage 1. Find a positive pool.
stage 2. Find a positive pool containing exactly $k - 1$ inhibitors.
stage 3. Identify all inhibitor and remove them.
stage 4. Identify all positive clones.

They gave pooling designs for stages 1, 3 and 4, but stage 2 is sequential. The total number of tests required is $O([(r/k)^2 + r + d]\log n)$. Using the results in Section 2, we solve the k-inhibitor model with a pooling design in $O((d + r)^2 \log n)$ tests. Moreover, there exists a pooling design with fewer tests. Called a binary matrix M a $(x + (m$ out of $y))$-*disjunct* matrix if for any $x + y + 1$ columns, $C_0, C_1, \dots, C_x, C_{x+1}, \dots, C_{x+y}$ there exists a row which intersects C_0 but none of $C_1, \dots, C_x$, and does not intersect at least m of $C_{x+1}, \dots, C_{x+y}$. Clearly, $(x + (0$ out of $y))$-disjunct is simply x-disjunct, $(x + (y$ out of $y))$-disjunct is simply $(x + y)$-disjunct, and $(x + m)$-disjunct implies $(x + (m$ out of $y))$-disjunct for any $y \geqslant m$.

We now use such a matrix to identify all positive clones. While we can use the general procedure given in Sections 2 and 3 to obtain a 1-round pooling design with or without errors, we can take advantage of the special feature of the k-inhibitor model to obtain a more efficient 1-round pooling design.

Theorem 3. *A $(d + (k$ out of $r))$-disjunct matrix can identify all p positive clones under the k-inhibitor model.*

Proof. For a given r-set R of columns, let $t^{R_k}(C)$ denote the number of negative pools clone C appears in except that a 0-outcome is changed to 1 if the row contains k 1-entry from R. Define $t^{*_k}(C) = \min_R t^{R_k}(C)$, where the minimum is over all $\binom{n}{r}$ choices of r-sets of columns. Then

$$t^{*_k}(P) = \min_R t^{R_k}(P) = t^{R'_k}(P) = 0,$$

where R' is an r-set covering all inhibitors.

On the other hand, for $C \in \{N, I\}$, the set of positive clones and R must leave at least one 1-entry of C uncovered by the definition of a $(d + (k$ out of $r))$-disjunct matrix. Thus we have

$$t^{*_k}(C) \geqslant 1, \quad \text{for } C \in \{N, I\}. \qquad \square$$

Next, we extend to the error-tolerant version. Call a binary matrix M a $(d + (m$ out of $r), z)$-*disjunct* matrix if for any $d + r + 1$ columns, $C_0, C_1, \dots, C_d, C_{d+1}, \dots, C_{d+r}$ there exists at least z rows which intersect C_0 but none of $C_1, \dots, C_d$, and does not intersect at least m of $C_{d+1}, \dots, C_{d+r}$. For the error-tolerant case, similar to Section 3, can identify all positive clones under the k-inhibitor model with e-error correcting property.

Corollary 4. *A $(d + (k$ out of $r), 2e + 1)$-disjunct matrix can identify all p positive clones under the general k-inhibitor model with at most e errors.*

Corollary 5. *Suppose a $(d + r - k + 1, 2e + 1)$-disjunct matrix is used. Then R can be taken from all $(r - k + 1)$-subsets of the column sets.*

5. Conclusions

We show that a $(d + r)$-disjunct matrix not only works for the 1-inhibitor model, but any inhibitor models and a $(d + r, 2e + 1)$-disjunct matrix works the same for the e-error-tolerant version. Thus the applicability is widely extended while no extra test is required.

In particular we can apply our results to the k-inhibitor model to obtain a pooling design while only a sequential procedure is available in the literature. The number of required tests drops from $(d + r)^2 \log n$ for the general model to $(d + r - k + 1)^2 \log n$ for the k-inhibitor model. Moreover, the number of tests can be reduced by using a $(d + (k$ out of $r))$-disjunct matrix. Obviously, the $(d + (k$ out of $r))$-disjunct matrix is a d-disjunct matrix which is known [5] to have a lower bound of $O(d^2 \log n / \log d)$ tests. We are unable to obtain a better bound specifically for the $(d + (k$ our of $r))$-disjunct matrix.

A small price is paid in decoding. Namely, in the 1-inhibitor model or other special cases, we can restrict the candidates of inhibitor to a small set and thus fewer sets R need to be run through. However, even in the 1-inhibitor model, one cannot estimate the amount of reduction to yield a better time complexity than $O(n^{r+1})$.

References

[1] A. De Bonis, L. Gasieniec, U. Vaccaro, Optimal two-stage algorithms for group testing problems, SIAM J. Comput. 34 (2005) 1253–1270.

[2] A. De Bonis, U. Vaccaro, Improved algorithms for group testing with inhibitors, Inform. Process. Lett. 67 (1998) 57–64.

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

[4] A.G. Dyachkov, A.J. Macula, D.C. Torney, P.A. Villenkin, Two models of nonadaptive group testing for designing screening experiments, in: A.C. Atkinson, P. Hackl, W.G. Muller (Eds.), Proc. 6th Intern. Workshop on Model-Oriented Design and Analysis, Physica-Verlag, 2001, pp. 63–75.

[5] D.Z. Du, F.K. Hwang, Combinational Group Testing and Its Applications, second ed., World Scientific, Singapore, 2000.

[6] D.Z. Du, F.K. Hwang, Identifying d positive clones in the presence of inhibitors, Int. J. Bioinform. Res. Appl. 1 (2005) 162–168.

[7] M. Farach, S. Kannan, E. Knill, S. Muthukrishnan, Group testing problems with sequences in experimental molecular biology, in: B. Carpentieri, et al. (Eds.), Proc. Compression and Complexity of Sequences, IEEE Press, 1997, pp. 357–367.

[8] F.K. Hwang, Y.C. Liu, Error-tolerant pooling designs with inhibitors, J. Comput. Biology (2003).

[9] A.J. Macula, Error-correcting nonadaptive group testing with d^e-disjunct matrices, Discrete Appl. Math. 80 (1996) 217–220.