

Group Testing

Lecture 12 ①
1.2 ~ 1.9

Date back: 1943, Blood testing during World War II
(Dorfman and Rosenblatt)

Group Testing Problem

Classical

Given a set N of n clones, each of which is either positive or negative but not both. The goal of group testing is to identify all positive ones (positives in short) by using as few number of "group tests" as possible. Here, a group test is an experiment on a subset of clones and the outcome of a group test on a subset $S \subseteq N$ is "1" if S contains at least one positive clone and "0" otherwise.

Example $N = \{1, 2, 3, 4, 5, 6\}$ and 3 is the only positive clone.
Then $t(\{1, 3, 3\}) = 1$ and $t(\{2, 4, 5\}) = 0$.

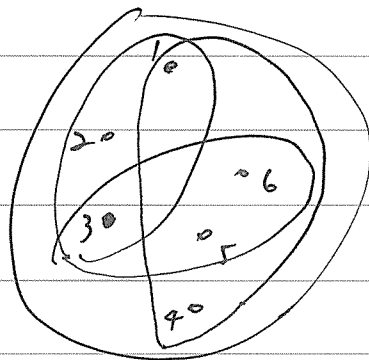
- Assumption on combinatorial group testing.

Let d be the number of positives in a set N of size n .
Then, we assume that $d \ll n$, i.e., d/n is very close to "0".

(Note) If d is too large comparing to n , then "group testing" is not working well in general.

Algorithms

1. Adaptive : An adaptive algorithm conducts tests one by one (Sequential)
 2. Non-adaptive : A non-adaptive algorithm specifies all tests simultaneously.
 3. k -stage algorithm : All tests in the "first" stage are specified simultaneously.
- The non-adaptive algorithm is also known as pooling design which can be depicted by using a matrix M .
Let ^{the} number of clones be n and there are t tests.
The the matrix M corresponding to a pooling design has the following properties :
 - (a) M is a t by n matrix ; and
 - (b) $M(i, j) = \begin{cases} 1 & \text{if the } i^{\text{th}} \text{ test including clone "j"} \\ 0 & \text{otherwise.} \end{cases}$



$$\begin{array}{c}
 \begin{matrix} t_1 \\ t_2 \\ t_3 \end{matrix} \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} \text{outcome vector} \end{bmatrix}
 \end{array}$$

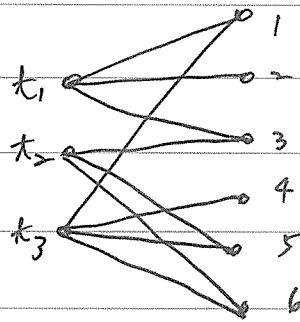
If (3) is positive, then the outcome vector is $(1, 1, 0)^t$.

③

How much do we know about $(0,1)$ -matrix?

(Fact 1) If M is a $t \times n$ $(0,1)$ -matrix, then we can find a labeled bipartite graph whose incidence matrix is M . Here $|A| = t$ and $|B| = n$. $G = (A, B)$

So, the above matrix is corresponding to the following bipartite graph



(Fact 2) If we let A be the set of elements and B is the collection of blocks, then we have a variety-block graph.

Again, the above example shows that there are six blocks $\{ \{1,3\}, \{1\}, \{1,2\}, \{3\}, \{2,3\}, \{2\} \}$.

- A design can be described by using a $(0,1)$ -matrix, so is a hypergraph.

(Fact 3) A $t \times n$ $(0,1)$ -matrix ^{M} can also be seen as a code where codewords are columns of M .

$$\begin{array}{r} 101 \\ 100 \\ 110 \\ 001 \\ \hline \end{array}$$

...entire

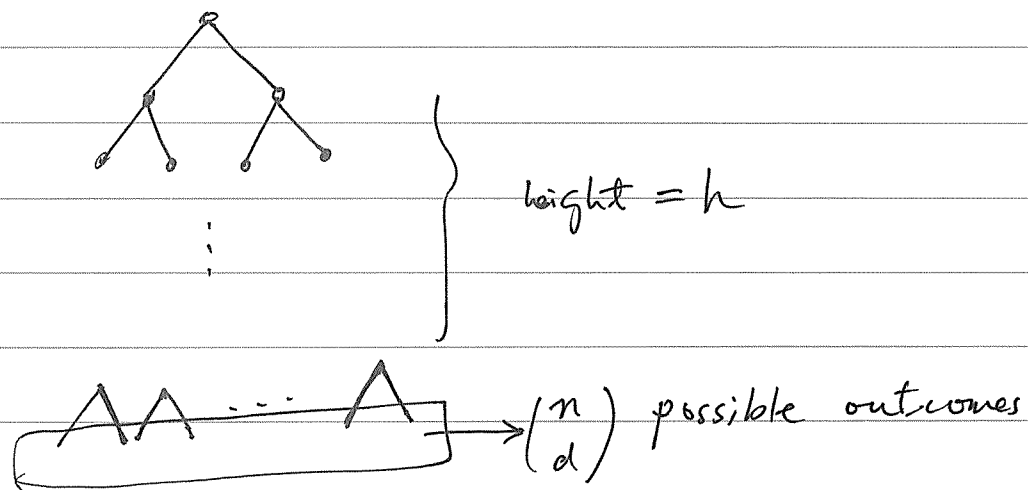
(Note) For practical use, we don't consider codewords from zeros!

(*) The idea of finding positives

If we have d positives, then there are $\binom{n}{d}$ possible combinations to have d positives. In order to find these positives, we may use different algorithms. In general, adaptive algorithms take less tests than non-adaptive algorithm. But, the later one "save time".

Information lower bound

We can use the idea of computational tree^{to}, find how many tests are needed (at least).



$$2^h = \binom{n}{d} \Rightarrow h \approx \left\lceil \log_2 \binom{n}{d} \right\rceil.$$

(*) We need at least $\left\lceil \log_2 \binom{n}{d} \right\rceil$ tests!

Another point of view

Consider all possible outcomes if t tests are applied. Then, there are at most 2^t outcome vectors which are distinct. So, if we use t tests to find d positives out of n , then $\binom{n}{d} \leq 2^t$ and thus

$$t \geq \left\lceil \log_2 \binom{n}{d} \right\rceil \approx d \log_2 \frac{n}{d}.$$

(Note) This idea is good for both adaptive and non-adaptive algorithms.

(Fact 4) With the idea of ^{checking the} outcome vectors, we may extend the study of error-correcting codes to construct pooling designs which can also solve the group testing problem with the occurrence of errors in experiments.

In what follows, we shall focus on non-adaptive algorithms or k -stage algorithms due to the fact that apply them in computational molecular biology.

Definition (pool)

Each test is also said to be a pool. A positive pool is a test (corresponding to the pool) which is positive, i.e., some of the clones in the experiment is positive and a negative pool is defined correspondingly, i.e., every clone in this pool is negative.

Separable matrices

$$M = [M(i,j)]_{m \times n} \quad (t \text{ tests and } n \text{ clones})$$

- We shall index the columns of M by $1, 2, \dots, n$ and the rows of M by $t_1, t_2, \dots, t_m$.

- The clone (j^{th}) can be represented by a set C_j where $C_j = \{i \mid M(i,j) = 1\}$.

- The union of s columns $C_{j_1}, C_{j_2}, \dots, C_{j_s}$ is $\bigcup_{k=1}^s C_{j_k}$ which is the Boolean sum of these columns.

e.g.
$$\begin{pmatrix} 0 \\ 1 \\ 0 \\ 1 \end{pmatrix} \cup \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 1 \\ 1 \end{pmatrix}.$$

$$\begin{aligned} &\downarrow \\ &U(\{j_1, j_2, \dots, j_s\}) \\ &= U(K) \end{aligned}$$

- The intersection of s' columns $C_{j_1}, C_{j_2}, \dots, C_{j_{s'}}$ is $\bigcap_{k=1}^{s'} C_{j_k}$.

e.g.
$$\begin{pmatrix} 0 \\ 1 \\ 0 \\ 1 \end{pmatrix} \cap \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}.$$

Definition (d -separable).

A matrix is d -separable if for any two ^{distinct} sets of d columns D and D' , $U(D) \neq U(D')$.

Definition ($\bar{d}$ -separable)

A matrix is $\bar{d}$ -separable if for any two distinct sets of at most d columns D and D' , i.e., $|D| \leq d, |D'| \leq d, U(D) \neq U(D')$.

Definition (Disjunct matrices)

M is called d -disjunct if no column is contained in the union of any other d columns.

M is d -separating if ~~one of~~ M is either a d -separable, $\bar{d}$ -separable or d -disjunct matrix.

$\bar{d}+1$ -separable

Lemma $\Rightarrow d$ -disjunct $\Rightarrow \bar{d}$ -separable $\Rightarrow d$ -separable. (easy)

Proof. If M is not $\bar{d}$ -separable, then there D and D' with $1 \leq |D| \leq d$, $1 \leq |D'| \leq d$, and $U(D) = U(D')$. This implies that there exists a column from D which is contained in $U(D')$ and there $U(D'')$ where $D' \subseteq D''$ and $|D''| = d$. Hence M is not d -disjunct.

(Note: 直接否亦可。)

Proposition A d -separable matrix can be applied to find "exactly d positives." Proof. (See next page.)

Proposition $\bar{d}$ -separable $\Rightarrow$ to find at most d positives

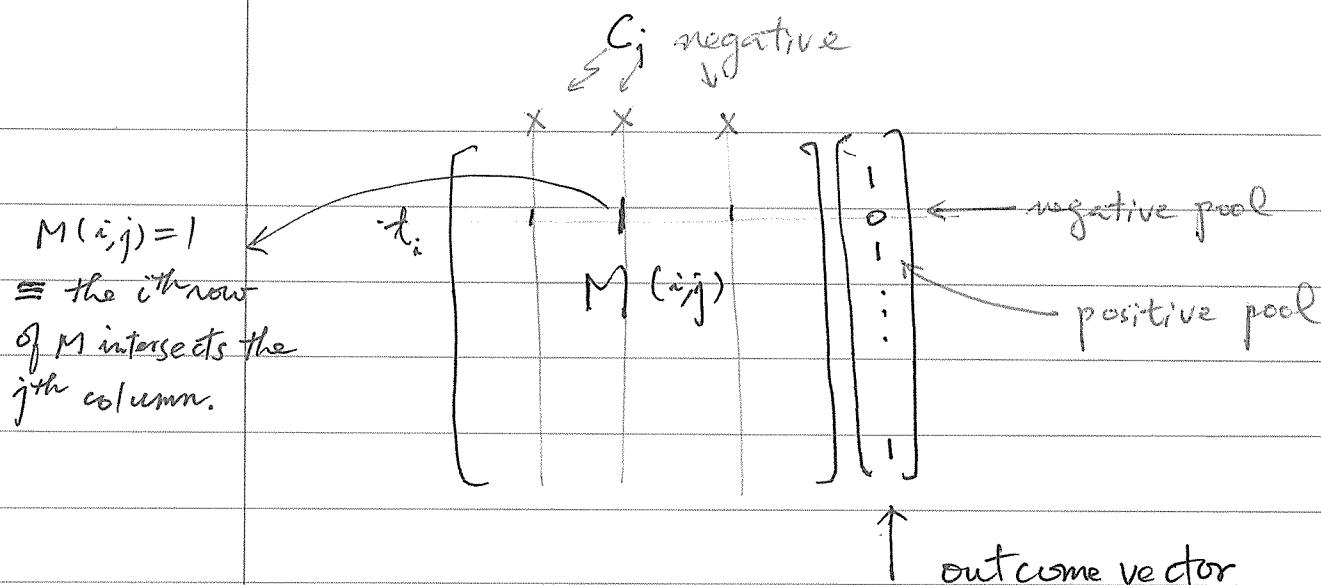
(*) To locate the set of positives we need to check all the outcome vectors which is quite complicate.

(Decoding Algorithm) is needed.

Here is an observation which is important!

$\downarrow$ next page

②'



- A test is ^(positive) negative if the corresponding component in the outcome vector is "0".
- negative (positive) test is also called negative (positive) pool.

觀察: Negative pool 中包含的 clones 全部是 "Negative"!

結果: 先利用 "negative" pools 把一些 clones 消去, 再從剩下的 clones 中找 d 個即可。

Hitting Set Problem

Given a set X and a collection $\mathcal{C}$ of subsets of X , find a minimum-cardinality subset Y of X such that $Y \cap C \neq \emptyset \forall C \in \mathcal{C}$.

Let M_D be a $t_D \times n_D$ matrix obtained from M by keeping only the set T_D of the positive pools and the set N_D of n_D clones not appearing in any negative pool. Then, the decoding is associate

Outcome Vector : Can be recognized as a set, V . ⑦

Denote by $t_o^V(C) = |C \setminus V|$ where C is a column.

e.g.

$$\begin{array}{c}
 \begin{array}{cccccc}
 & 1 & \textcircled{2} & \textcircled{3} & 4 & 5 & 6 & V \\
 \begin{array}{l} t_1 \\ t_2 \\ t_3 \\ t_4 \end{array} & \left[\begin{array}{cccccc}
 1 & 1 & 1 & 0 & 0 & 0 \\
 1 & 0 & 0 & 1 & 1 & 0 \\
 0 & 1 & 0 & 1 & 0 & 1 \\
 0 & 0 & 1 & 0 & 1 & 1
 \end{array} \right] & \left[\begin{array}{c} 1 \\ 0 \\ 1 \\ 1 \end{array} \right]
 \end{array}
 \end{array}$$

$\uparrow \uparrow \uparrow \uparrow \uparrow \uparrow$
 $C_1 \ C_2 \ C_3 \ C_4 \ C_5 \ C_6$

$$V = C_2 \cup C_3$$

Note that $t_o^V(C) = |C \setminus V|$, $t_i^V = |C \cap V|$.

$$t_o^V(C_1) = 1$$

$$t_i^V(C_1) = 1$$

$$t_o^V(C_2) = 0$$

$$t_i^V(C_2) = 2$$

$$t_o^V(C_3) = 0$$

$$t_i^V(C_3) = 2$$

$$t_o^V(C_4) = 1$$

$$t_i^V(C_4) = 1$$

$$t_o^V(C_5) = 1$$

$$t_i^V(C_5) = 1$$

$$t_o^V(C_6) = 0$$

$$t_i^V(C_6) = 2$$

$\uparrow$
 Negatives
 C_1, C_4, C_5

可解是 $\begin{array}{c} \textcircled{2} \textcircled{3} \\ \textcircled{2} \textcircled{6} \\ \textcircled{3} \textcircled{6} \end{array}$

如果是 2-disjunct 就不会留下 C_2, C_3, C_6 等待判断。

(*) We can use a d -disjunct matrix to identify up-to- d positive clones. ⑦'''

Theorem If a set of n clones which contains at most d positives then a d -disjunct matrix can be applied to identify the positives. Moreover, the decoding algorithm is linear complexity of

Proof. The proof follows by evaluating $t_0^V(C)$ for each clone C . (V is the outcome vector.) We remark here that
" $t_0^V(C) = 0$ if C is positive and $t_0^V(C) \geq 1$ if C is negative."
($t_0^V(C) = |C \setminus V|$.) 高下-个 2-disjunct matrix

(说明) $V = \bigcup_{C \text{ is positive}} C$, 如果有 C' 是 negative, 而 $|C' \setminus V| = 0$, 则 $C' \subseteq V$, 这 $\Rightarrow d$ -disjunct 的假设不符。

Algorithm (Clone Selection (N, V, D, e))

1. for each clone $C \in N$, ($D \leftarrow \emptyset$)
2. compute $t_0^V(C)$
3. if $t_0^V(C) \leq e$
4. then $D \leftarrow D \cup \{C\}$
5. return

(*) 上面定理由 Clone Selection ($N, V, D, 0$)。

(8)

Definition $(d; z)$ -disjunct matrix.

A matrix is said to be $(d; z)$ -disjunct if for any $d+1$ columns $C_0, C_1, \dots, C_d$,

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

Note $(d; 1)$ -disjunct $\approx$ d -disjunct.

(*) There exist at least z rows in each of which C_0 has a "1"-entry and every $C_i, 1 \leq i \leq d$, has a "0"-entry.

$$k \geq z \left\{ \begin{array}{c} C_0 \quad C_1 \quad C_2 \quad \dots \quad C_d \\ \begin{array}{c} | \\ 1 \\ | \\ \vdots \\ | \end{array} \quad \boxed{\begin{array}{c} 0's \end{array}} \end{array} \right\}_{k \times d}$$

Theorem A $(d; z)$ -disjunct matrix can be applied to identify up-to- d positives with at most e errors.

Proof We remark that the errors occur on the conclusion of tests, i.e., the outcome vector (as a codeword) has at most e errors. (Due the fact that errors are possible, we shall have a set of possible outcomes, instead of just one outcome vector V). Now, assume the number of errors is e . For each negative clone $\bar{C}$, by definition, there exist at least $2e+1$ rows intersecting $\bar{C}$ but none of D (the set of positives). This implies that for each outcome V ...

(9)

On the other hand, for each positive clone C^+ , $t_0^V(C^+) \leq e$.
 ($t_0^V(C) = 0$ if no errors occurred.) So, the proof follows by checking $n(=|N|)$ clones by using the function $t_0^V(C)$ where V is the outcome vector (in $\mathbb{V}$). Note, there is exactly one outcome vector after the experiment is done. $\blacksquare$

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 clone in the pool.

(註) Inhibitor 出現時, 該 pool 是 陰性反應。

Lemma

⑨

Theorem (Error-tolerant inhibitor 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., $t_1^V(C) \leq e$, cannot be positive due to the $(d+h; 2e+1)$ -disjunctness property. (If C is positive, then $t_1^V(C) \geq 2e+1$ if there are no errors.) For the inhibitors, each inhibitor can occur in at most e positive pools (with e errors), i.e., $t_1^V(I) \leq e$.

Now, let O be the set of columns C such that $t_1^V(C) \leq e$. This implies that O contains all inhibitors but no positives. Hence, D , the set of positive clones is a subset of $N \setminus O$.

⑨'

Lemma (e errors and h inhibitors)

If M is an $(m; z)$ -disjunct matrix with $m \geq s$, then

- (a) $t_1^V(p) \geq m - h + z - e$ if p is positive;
- (b) $t_0^V(Q) \geq m - d + z - e$ if Q is negative; and
- (c) $t_1^V(I) \leq e$ if I is an inhibitor.

Here, V is the outcome vector.

proof. Since $m \geq h$, M is also an $(s; z + m - h)$ -disjunct matrix. Hence $t_1^V(p) \geq m - h + z$ if there are no errors. Now, if e errors occur, then $t_1^V(p) \geq m - h + z - e$. This proves (a).

(Note that $t_1^V(S) = |S \cap V|$.) Moreover, ⁱⁿ an $(m; z)$ -disjunct matrix, the outcome vector must have at least z "1"s.

The proof of (b) is similar. Let's consider (c). All 1's are from errors.



Let $C^- \in N \setminus O$. By disjointness property, we have $t_0^V(C^-) \geq (2e+1) - e = e+1$. (?) On the other hand, if C^+ is a positive clone in $N \setminus O$, then $t_0^V(C^+) \leq e$. (?) Therefore, by calculating $t_0^V(C)$ where $C \in N \setminus O$, we conclude the proof. ■

说明：在这个定理中我们只负责把 positives 找出来，基本上不知道那几个 clones 是 inhibitors。

说明'：(d+h, z)-disjointness 是用来把 Inhibitors 分出来；如果没 inhibitors (d, z)-disjointness 就够用了。



Group Testing on Complexes

In the complex model, we consider a set N of n clones and the positives are in an unknown family $P = \{P_i\}$ of subsets of N .
(known)

(有 positive outcome 要由一些 clones 所组合而成，不再是单一 clone 而已。)

(•) Each member of P is called a ^{positive} complex; (each subset of N is a candidate of a member of P .)

(•) An experiment is applied to an arbitrary subset $S \subseteq N$ with two possible outcomes; a negative outcome implies S does not contain any $P_i \in P$, and a positive outcome implies otherwise.

Complex Model

- 我们必须假设当 X, X' 皆为 complexes 时, 彼此没有互相包含的关系, 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 formidable complexes of eukaryotic DNA transcription and RNA translation could involve hundred of molecules.
- Application: protein-protein interaction (Many) (Lappe and Holm, 2004).
- (*) (Creating a protein-protein interaction network for

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

(Generalized cover-free

Definition $(d, r; z]$ -disjunct) Stinson and Wei, 1999 family.

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

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

$$\begin{array}{c} \underbrace{C_1 \dots C_r}_{\geq z} \quad \underbrace{C_{r+1}, \dots, C_{r+d}} \\ \left[\begin{array}{cc} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \end{array} \right] \end{array}$$

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.

For a complex X ,

Proof. $t_o^V(X) =_{\text{def}} |\cap X \setminus V|$ and $t_i^V(X) =_{\text{def}} |(\cap X) \cap V|$.

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

P3

Threshold Group Testing [Damaschke, 2005]

Let l and u be nonnegative integers $l < u$, called the lower and upper threshold. A group test (threshold model) for a subset S of clones is positive if S contains at least u positives, and negative if at most l positives are present. Furthermore, if the number of positives in S is between l and u , the outcome can give an arbitrary answer, either positive or negative.

$g =_{\text{def}} u - l - 1$ is called the gap.

Note: Use $(d, n; z)$ -disjunct matrices to identify the positives with at most e errors. ($z \geq 2e + 1$)

We have stop somewhere! Thank you for your patience. ^_~