GT 9 Nov. 15, 17

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

Denote by $t_0^V(C) = |C \setminus V|$ where C is a column.

e.g.

$$\begin{array}{c|cccccc|c} & 1 & 2 & 3 & 4 & 5 & 6 & V \\ \hline t_1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 \\ t_2 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ t_3 & 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ t_4 & 0 & 0 & 1 & 0 & 1 & 1 & 1 \\ \hline & C_1 & C_2 & C_3 & C_4 & C_5 & C_6 & \end{array}$$

positives: 2, 3

$\{1,2\}$
$\{1,3\}$
$\{1,4\}$
$\{2,3\}$
$\{2,4\}$
$\{3,4\}$

$t_1^V(C) = |C \cap V|$

(*) $V = C_2 \cup C_3$

Note that $t_0^V(C) = |C \setminus V|$, $t_1^V = |C \cap V|$.

$|C \setminus V| > 0 \Rightarrow C$ is a negative item.

$t_0^V(C_1) = 1$	$t_1^V(C_1) = 1$
$t_0^V(C_2) = 0$	$t_1^V(C_2) = 2$
$t_0^V(C_3) = 0$	$t_1^V(C_3) = 2$
$t_0^V(C_4) = 1$	$t_1^V(C_4) = 1$
$t_0^V(C_5) = 1$	$t_1^V(C_5) = 1$
$t_0^V(C_6) = 0$	$t_1^V(C_6) = 2$

$\uparrow$

Negatives
C_1, C_4, C_5

Two positives: $\left\{\begin{array}{l} 2\ 3 \text{ or} \\ 2\ 6 \text{ or} \\ 3\ 6. \end{array}\right.$

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

(2)

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

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

<u>Proof.</u> 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.

($t_0^V(C) = |C\setminus V|$.)

Consider C_0 (negative) $\overbrace{C_1,\cdots,C_d}^{\text{positive}}$, ‖ 多一个 positive C_0, 则 $|C_0\setminus V|=0$, $C_0\setminus V=\emptyset$ $\Rightarrow C_0\subseteq V$.

(证明) $V = \bigcup\limits_{C \text{ is positive}} C$, 如果有一 C' 是 negative, 而 $|C'\setminus V|=0$

则 $C'\subseteq V$, 这与 d-disjunct 的假设不符。

Algorithm (Clone Selection (N, V, D, e)) → possible # of errors

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

$$\boxed{V=\bigcup_{C \text{ positive}} C}$$

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

(◦) <u>Key Point</u> If C is positive, then $t_0^V(C)=0$; and

(V is the outcome vector.)

C is negative, then $t_0^V(C)\geq 1$.

<u>Definition</u> ($(d; z)$-disjunct matrix).

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

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

<u>Note</u> $(d; 1)$-disjunct $\approx$ d-disjunct.

(0) 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.

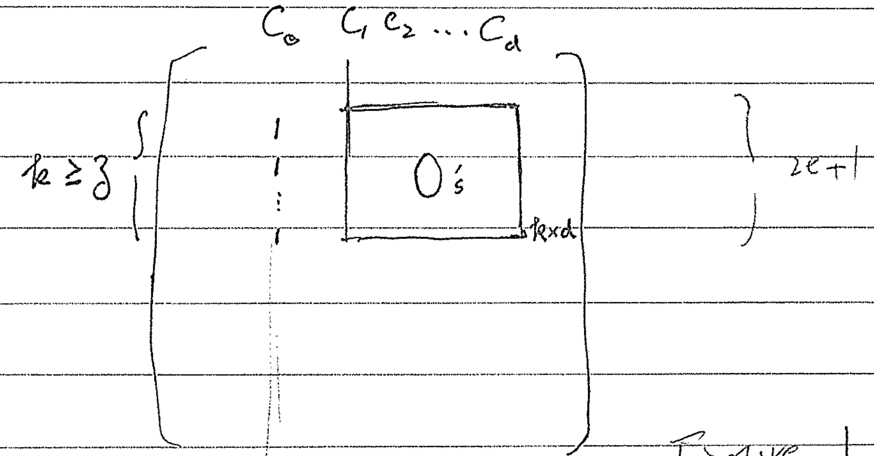


Figure 1

<u>Theorem</u> A $(d; \overset{2e+1}{\overset{\|}{z}})$-disjunct matrix can be applied to identify <u>up-to-d positives</u> with at most e errors.

<u>Proof</u> We remark that the errors occur on the <u>conclusion</u> 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, <u>instead of just one outcome vector</u>.) Now, assume the number of errors is e. For each negative clone C^-, by definition, there exist at least $2e+1$ rows intersecting C^- but none of D (the set of positives). This implies that for each clone (C^- or C^+),

$t_0^V(C^-) \geq (2e+1)-e$, and $t_0^V(C^+) \leq e$.

④

说明：

由于 Outcome vector 是所有 Positive clones 的联集；如果用 $(d;2e+1)$-disjunct 矩阵来测试的话，对于所有 positive clone C^+，$\boxed{t_1^V(C^+) \geq 2e+1}$ (没有任何 Errors 产生的情况)。这可以从上页 Figure 1 看出来：把 C_0 当成 C (positive)，$|V| \geq 2e+1$。

如果把 positives 都放入 $\{C_1, C_2, \cdots, C_d\}$ 而 $\underset{C^-}{\underset{\parallel}{C_0}}$ 为 negative，则 $t_0^V(C^-) \geq 2e+1$。

于是在有 e 个错误的情况下 $t_0^V(C^-) \geq e+1$；

另外 $t_0^V(C^+) = 0$ (没有错误)，在有 e 个错误

$$\underset{|C^+ \setminus V| = |\phi|}{\underset{\parallel}{t_0^V}}(C^+) = 0$$

的情况下，$t_0^V(C^+) \leq e$。

上述的说明也提供了 Linear Time Decoding Algorithm. (on $|N|$)

$\Uparrow$

每个 clone 都测试一次。

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

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

- 我们必需假設当 X, X' 皆為 complexes 時，彼此沒有互相包含的関係，亦即 $X \nsubseteq X'$ 且 $X' \subseteq X$。

- Or the model must be monotone on positives, i.e., if X is positive and $X \subsetneq X'$, 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 can not 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-to protein interaction (Many) (Lappe and Holm, 2004).

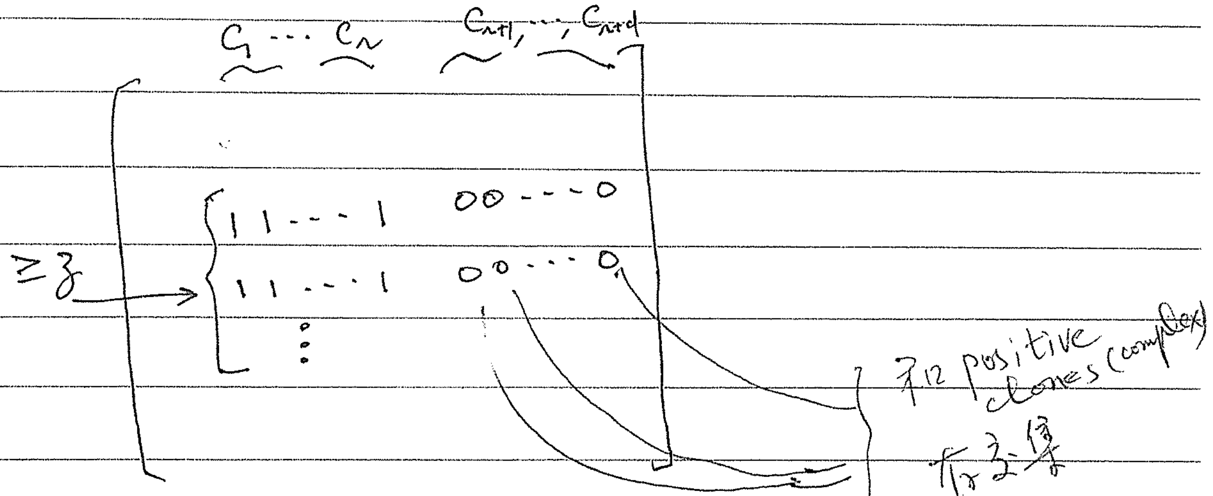
(※) (Creating a protein-protein interaction network.)

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

Definition ($(d,r;z]$-disjunct) Stinson and Wei, 2009 (~~Generalized cover-free family~~)

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

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



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.

For a complex X,

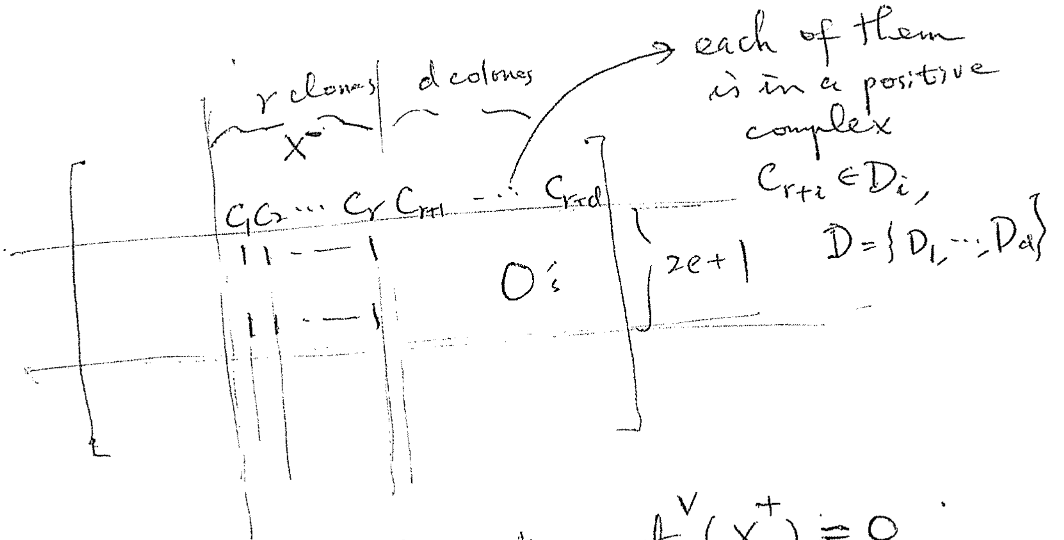
Proof. $t_0^V(X) =_{def} |\cap X \setminus V|$ and $t_1^V(X) =_{def} |(\cap X) \cap V|$.

- 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.



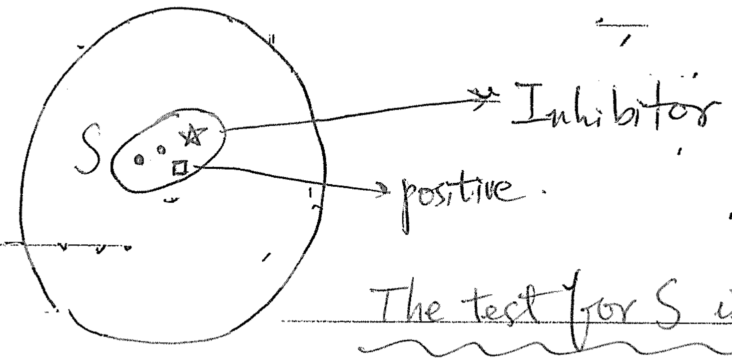
By the definition of the matrix $t_0^\vee(X^+) = 0$ (with no errors). So, if there are at most e errors, $t_0^\vee(X^+) \le e$.

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

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

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

Let $\bar{C} \in N \setminus O$. By disjunctness property, we have $t_0^{\vee}(\bar{C}) \geq e+1$. On the other hand, if $C^+ \in N \setminus O$, then $t_0^{\vee}(C^+) \leq e$. 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，我們並不需要去找到那些 Clones 是 Inhibitors。

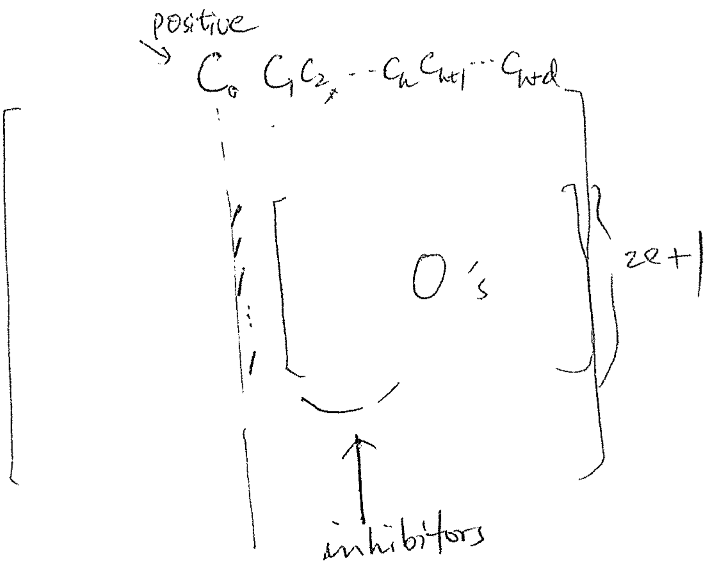
概念：先把不可能是 positive clones 的 clones 找出來，形成集合 O，這個集合中可能是有一些 negative clones.

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 個錯，$t_1^V(c) \geq e+1$。