

# Efficiency of the Two-Stage Group Testing Algorithm for DNA Library Screening

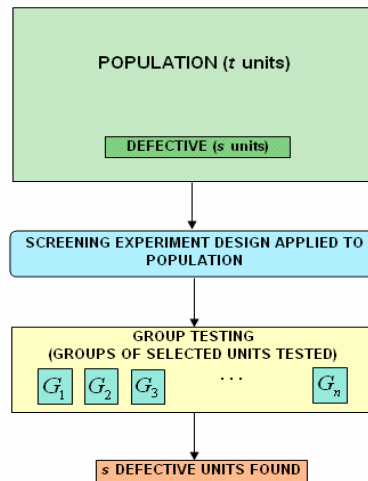
Vyacheslav Rykov, Vladimir Ufimtsev

## OVERVIEW

- Introduction to Group Testing.
- DNA Library Screening.
- Superimposed Coding Theory.
- Two-Stage Group Testing.
- Bound on parameters of DNA library screening experiments.

# GROUP TESTING & SCREENING EXPERIMENTS

- Mathematical technique commonly employed in the design of screening experiments.
- Population of  $t$  units,
- Known that *at most*  $s$  units are defective (positive).
- Defective units have a characteristic that is not present in the non-defective (negative) units.
- For a sample of any size, the presence (positive result) or absence (negative result) of the defective characteristic can be established by exactly one test.
- The defective units of the population need to be found in the least possible amount of tests.



# GROUP TESTING APPLICATIONS

Types of problems in which Group Testing is used:

- Pollution tests.
- Leakage tests.
- Flow tests.
- Identifying active users.
- Pattern recognition.
- Screening experimental factors.

## GROUP TESTING ORIGIN

- Idea originated in the spring of 1942. Dorfman and Rosenblatt.
- Blood samples of millions of draftees subjected to identical analyses for detection of syphilis.
- Instead of analyzing *each* blood sample, the samples should be analyzed in pools (groups).



## GROUP TESTING APPLICATIONS

Problems that can employ Group Testing:

- Pooling of DNA libraries to determine which strands in the DNA library contain a probe (DNA Library Screening).

# DNA LIBRARY SCREENING

- A *DNA library* is a large collection of carefully constructed single stranded DNA sequences, that can be used in computations to encode solutions to mathematical problems.
- Decoding the solutions is problematic.
- One method to decode solutions (determine the positive clones) is to employ group testing (or pooling).

# SCREENING EXPERIMENT EXAMPLE

- Suppose there are **7** single DNA strands in a library and after some computation we obtain a solution set containing **1** of the **7** DNA strands (1 positive clone).
- The goal is to identify *in the least amount of tests*, which DNA strand is present in the solution set.
- Can be done by testing for each strand in the library (7 tests).

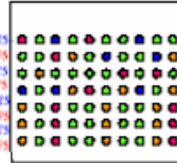
```
1  AAAAAAAAAACCCTTTCCTTAACCCGATTAATAAACACACGMB
2  AAAAAAAAAACCCTTTCCTTAACCCGATTAATAAACACACGMB
3  AAAAAAAAAACCCTTTCCTTAACCCGATTAATAAACACACGMB
4  AAAAAAAAAACCCTTTCCTTAACCCGATTAATAAACACACGMB
5  AAAAAAAAAACCCTTTCCTTAACCCGATTAATAAACACACGMB
6  AAAAAAAAAACCCTTTCCTTAACCCGATTAATAAACACACGMB
7  AAAAAAAAAACCCTTTCCTTAACCCGATTAATAAACACACGMB
```

# SCREENING EXPERIMENT EXAMPLE

- Can use group testing by augmenting every strand in the DNA library with synthetic 'tag' strands constructed from the screening experiment design.

## DNA Library Augmented

1. AAAAAAAAAACC-TTCTTAAACCATAAAACAC-T4-T5  
 2. AAAAAAAAAACC-TTCTTAAACCATAAAACAC-T4-T5  
 3. AAAAAAAAAACC-TTCTTAAACCATAAAACAC-T4-T5  
 4. AAAAAAAAAACC-TTCTTAAACCATAAAACAC-T4-T5  
 5. AAAAAAAAAACC-TTCTTAAACC-ATCTTTTCAA-T4-T5  
 6. AAAAAAAAAACC-TTCTTAAACC-ATCTTTTCAA-T4-T5  
 7. AAAAAAAAAACC-TTCTTAAACC-ATCTTTTCAA-T4-T5  
 8. AAAAAAAAAACC-TTCTTAAACC-ATCTTTTCAA-T4-T5



- To find exactly which strands constitute the solution set, probes need to be developed that will read the tags simultaneously.

# SCREENING EXPERIMENT - 1 POSITIVE CLONE

	$P_1$	$P_2$	$P_3$	$P_4$	$P_5$	$P_6$	$P_7$	
$A$	$A$	$A$	$A$	0	0	$A$	0	$A$
0	$B$	$B$	0	0	$B$	0	$B$	0
0	0	$C$	$C$	$C$	$C$	$C$	0	$C$

## SCREENING EXPERIMENT - 2 POSITIVE CLONES

$P_1$	$P_2$	$P_3$	$P_4$	$P_5$	$P_6$	$P_7$	
$A$	$A$	$A$	$0$	$0$	$A$	$0$	$A$
$0$	$B$	$B$	$0$	$B$	$0$	$B$	$B$
$0$	$0$	$C$	$C$	$C$	$C$	$0$	$C$

## SUPERIMPOSED CODING THEORY

- Consider the following matrix:

$$X = \begin{bmatrix} x_1(1) & x_1(2) & \cdot & \cdot & \cdot & x_1(t) \\ x_2(1) & x_2(2) & \cdot & \cdot & \cdot & x_2(t) \\ \cdot & \cdot & & x_i(j) & & \cdot \\ \cdot & \cdot & & \cdot & & \cdot \\ x_N(1) & x_N(2) & \cdot & \cdot & \cdot & x_N(t) \end{bmatrix} \quad x_i(j) \in \{0,1\}$$

$$\mathbf{x}(j) = \begin{bmatrix} x_1(j) \\ x_2(j) \\ \cdot \\ \cdot \\ \cdot \\ x_n(j) \end{bmatrix}$$

### Definition (Code).

The above  $N \times t$  matrix is referred to as a *code*. The columns of  $X$  are the *code packets*. Let  $\mathbf{x}(j)$  denote the  $i$ -th code packet. Code  $X$  is of size  $t$  and length  $N$ .

# SUPERIMPOSED CODING THEORY

## Definition: Cover

$\mathbf{x}(i)$  covers  $\mathbf{x}(j)$  if and only if:

$$\mathbf{x}(i) \vee \mathbf{x}(j) = \mathbf{x}(i)$$

- Example (cover):

$$\begin{bmatrix} 0 & 1 & 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 \end{bmatrix}$$

# SUPERIMPOSED CODING THEORY

## Definition: Superimposed Code

A code  $\mathbf{X}$  has strength  $s$  if and only if the Boolean sum of any  $s$  codewords does not cover any other codeword in  $\mathbf{X}$ .

A code of length  $N$ , size  $t$ , and strength  $s$  is an  $(N,s,t)$  superimposed code. The matrix  $\mathbf{X}$  is also called  $s$ -disjunct.

# CODES USED IN DNA LIBRARY SCREENING

- The codes which we will implement were introduced by Kautz-Singleton in 1964, and are built from the following code:

**Definition (Reed-Solomon Code)** Let  $k, n$  be integers that satisfy  $1 \leq k < n \leq q + 1$ , and let  $a_1, a_2, \dots, a_{q-1}$  denote the non-zero elements of  $GF(q)$ . The matrix:

$$\begin{bmatrix} 1 & 1 & \dots & 1 & 1 & 0 \\ a_1 & a_2 & \dots & a_{n-2} & 0 & 0 \\ a_1^2 & a_2^2 & \dots & a_{n-2}^2 & 0 & 0 \\ \vdots & \vdots & \dots & \vdots & \vdots & \vdots \\ a_1^{n-k-1} & a_2^{n-k-1} & \dots & a_{n-2}^{n-k-1} & 0 & 1 \end{bmatrix}$$

is the parity-check matrix of an MDS  $[n, k, n - k + 1]$  code which is known as the *Reed-Solomon code*.

Using  $[n, k, n - k + 1]$  Reed-Solomon codes where  $n$  is maximal i.e.  $n = q + 1$ , the family of superimposed codes implemented is obtained.

# SUPERIMPOSED CODES FROM REED-SOLOMON CODES

Let  $q$  be a prime or prime power and  $k \geq 2$ . Take  $C$  to be a  $q$ -nary  $[q + 1, k, q - k + 2]$  Reed-Solomon code, of size  $t = q^k$ . Code  $C$  can be represented by the following matrix, whose columns are the codewords of  $C$ :

$$C = \begin{bmatrix} y_1(1) & y_1(2) & \dots & y_1(t) \\ y_2(1) & y_2(2) & \dots & y_2(t) \\ y_3(1) & y_3(2) & \dots & y_3(t) \\ \vdots & \vdots & \dots & \vdots \\ y_{q+1}(1) & y_{q+1}(2) & \dots & y_{q+1}(t) \end{bmatrix}$$

$$y_i(j) \in GF(q), i = 1, 2, \dots, q + 1, j = 1, 2, \dots, q^k$$

$C$  can be transformed into a binary superimposed code by applying the following transformation. Each symbol of  $GF(q)$  is associated with a binary column vector of length  $q$  and weight 1 i.e.

$$[0 \ 1 \ 2 \ 3 \ \dots \ q-1] = \begin{bmatrix} 1 & 0 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 1 \end{bmatrix}$$

## SUPERIMPOSED CODES FROM REED-SOLOMON CODES

$$C = \begin{bmatrix} y_1(1) & y_1(2) & \dots & y_1(t) \\ y_2(1) & y_2(2) & \dots & y_2(t) \\ y_3(1) & y_3(2) & \dots & y_3(t) \\ \vdots & \vdots & \dots & \vdots \\ y_{q+1}(1) & y_{q+1}(2) & \dots & y_{q+1}(t) \end{bmatrix}$$

$$y_i(j) \in GF(q), i = 1, 2, \dots, q + 1, j = 1, 2, \dots, q^k$$

Each symbol in  $C$  is then replaced by the binary column associated with it. This transformation produces a binary  $q(q + 1) \times q^k$  matrix  $X$ , which is a superimposed code of size  $t = 2^K$  with the following parameters (strength is calculated by (5) using  $w = q + 1, \lambda = k - 1$ ) :

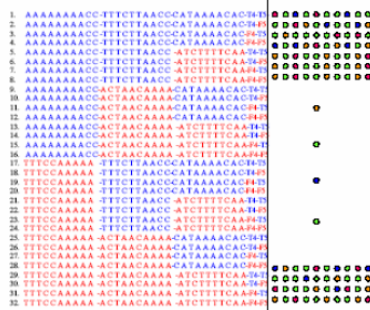
$$K = \lfloor k \log_2 q \rfloor, \quad N = q(q + 1), \quad s = \left\lfloor \frac{q}{k - 1} \right\rfloor$$

This binary superimposed code will be labelled as an  $(N, s, t)$  code.

## DNA LIBRARY SCREENING

- If the solution set contains up to  $s$  strands then:
- Augment every strand in the DNA library with synthetic 'tag' strands constructed from the screening experiment design set forth by a *superimposed code* of strength  $s$ .

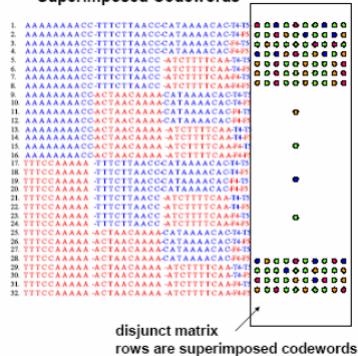
DNA Library Augmented with  
Superimposed Codewords



disjunct matrix  
rows are superimposed codewords

# DNA LIBRARY SCREENING

DNA Library Augmented with Superimposed Codewords



- Once the probes output the resulting vector (which corresponds to the Boolean sum of the superimposed tags present), the solution strands are then recovered.

# DNA LIBRARY SCREENING

- If the number of strands in the solution set (number of positive clones) exceeds the strength of the superimposed code used then *other tags (of negative clones) will be covered by the result of the experiment* (Boolean sum of positive clone tags).
- Given a value  $p$  we would like to obtain the function for *average number of extra tags (clones) covered by the Boolean sum of an arbitrary  $p$ -set of tags from the superimposed code used.*
- Studying the asymptotic behavior of such a function, allows us to obtain better bounds on the maximum number of tests required to find  $p$  positive clones among  $t$  clones and it also allows to calculate the maximum number of clones we can find using  $N$  tests among  $t$  clones.

## DNA LIBRARY SCREENING TWO-STAGE GROUP TESTING

- Stage 1: Carry out experiment to determine which clones are covered by the result (Boolean sum).
- Stage 2: Extra negative clones could be covered, thus we must establish which covered clones are indeed positive.

This is done in polynomial time since checking a possible solution to an NP problem is done in P time.

## AVERAGE NUMBER OF EXTRA CLONES COVERED

Consider an arbitrary MDS code  $C$  with parameters  $q, k, n$  of volume  $t = q^k$ ,  $k \leq n - 1 \leq q$  and codewords  $\mathbf{x}(i) = \{x_1(i), x_2(i), \dots, x_n(i)\}$ ,  $i = \overline{1, t}$ . Where  $q$  is the number of symbols in the alphabet  $GF(q)$ ,  $n$  is the length of the code,  $d = n - k + 1$  is the minimal Hamming distance and the minimal weight of the codewords. Denote by  $S_w(n)$  the number of codewords in  $C$  of weight  $w$ . Then we have:

$$S_w(n) = \binom{n}{w} (q-1) \sum_{j=0}^{w-d} (-1)^j \binom{w-1}{j} q^{w-d-j}, w = \overline{d, n}$$

the total number of  $p$ -sets of  $C$  that do not cover  $\mathbf{0}$

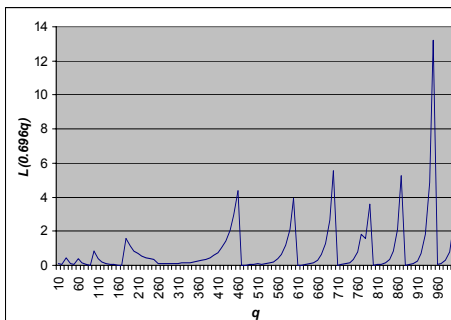
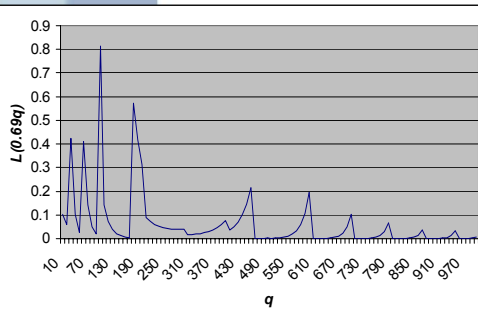
$$C_0(p, n) = \sum_{i=1}^n (-1)^{i+1} \binom{n}{i} \binom{S_i(i)}{p}$$

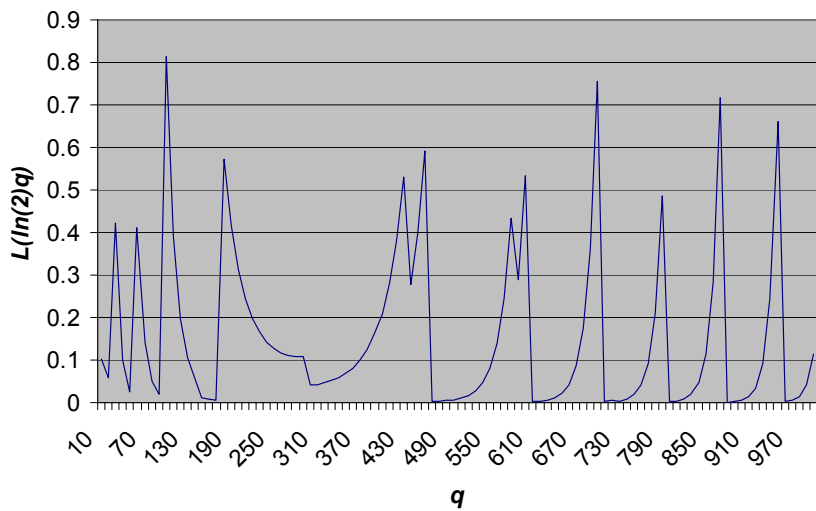
Then the average number of codewords that *do not belong to but are covered* by an arbitrary  $p$ -set of  $C$  is:

$$L(p) = \frac{\left( \binom{q^k}{p} - C_0(p, n) \right) q^k}{\binom{q^k}{p}}$$

**Theorem** Let  $p = \alpha q$ ,  $k = \lfloor \frac{q}{\log_2 q} \rfloor$ . Then as  $q \rightarrow \infty$ :

$$\lim_{q \rightarrow \infty} L(\alpha q) \rightarrow \begin{cases} \infty & \text{if } \alpha > \ln 2 \\ 0 & \text{if } \alpha < \ln 2 \end{cases}$$





## BOUNDS

- Let:  $q$  be a prime power and let  $k = \left\lfloor \frac{q}{\log_2 q} \right\rfloor$

- If the number of positive clones  $p = \ln 2 \log_2 t$   
then for sufficiently large size of library  $t$ :

$$N \leq (\log_2 t)^2$$

- If the number of tests  $N = (\log_2 t)^2$   
then for sufficiently large size of library  $t$ :

$$p \leq \ln 2 \log_2 t$$