Department of Social Systems and Management

Discussion Paper Series

No.1315

Pooling Experiments for Consecutive Positives

by

Song LUO, and Mingchao ZHANG

January 2014

UNIVERSITY OF TSUKUBA
Tsukuba, Ibaraki 305-8573
JAPAN

Pooling Experiments for Consecutive Positives

Song Luo and Mingchao Zhang

Abstract—In this paper, we study the pooling experiments for screening clone maps. Clones are overlapped and placed in a linear order, and hence the clones containing a particular DNA sequence of our interest form a consecutive set (or several consecutive subsets) under the linear order. Those clones are called consecutive positives. Pooling experiments are used to identify as many positives as possible by screening as few pools as possible. Stochastic models are introduced to explain the usefulness of the overlap structure for more efficient pooling experiments even when experimental errors exist. We describe an efficient positive detecting algorithm, called modified Markov chain pool result decoder (MMCPD), for consecutive positives. We also introduce an efficient algorithm, called random pooling designer (RPD), for estimating the minimal number of randomly generated pools required for achieving the complete identifiability of pooling experiments when consecutive positives and experimental errors exist. To our best knowledge, MMCPD is the first probabilistic group testing algorithm for detecting consecutive positives when experimental errors exist. Interestingly, RPD shows some happy coincidences between our theoretical computation results and previously known simulation results. Some simulation results are also reported in hoping of demonstrating that MMCPD and RPD may be promising for real settings.

Index Terms—pooling experiment, group testing, DNA library screening, clone map, consecutive positive, random k-set design, information-theoretic lower bound.

1 Introduction

Group testing was proposed by Robert Dorfman [15] during World War II in order to efficiently test a large number of blood samples for a rare disease. Since then, various applications of group testing have been found in many fields such as multiple communication, coding theory, information security, sparse signal recovery and others. Particularly, biology-motivated group testing has been developed into one of the most important tools in the study of gene functions. For example, in Human Genome Project, well-designed group testing schemes have been proved to be useful in both saving materials and accelerating the process of reconstructing high-quality DNA libraries. These high-quality libraries have been frequently and repeatedly used for extensive studies.

A DNA library is a collection of cloned DNA segments taken from a specific organism. Those DNA segments are called clones. Determining whether a clone contains a particular DNA sequence of interest can be accomplished by screening it with a probe. The clone is called a positive for the probe if it contains the particular DNA sequence, and a negative otherwise. Due to the large size of a DNA library and the cost of materials, instead of screening each clone individually, combinations of the clones are screened to identify and isolate the positives. The combinations of the clones are called pools. Each pool is screened with the probe to learn whether any of the clones in the pool contain the DNA sequence. Screening pools in this way is called a pooling experiment. Pooling experiments are used to identify as many positives as possible by screening as few pools as possible. The efficiency of pooling experiments has been studied by Barillot et al. [2], Berger et al. [3], Bruno et al. [5] and Sham et al. [38]. Here we refer to the books [12] and [13] by Du and Hwang for an overview.

- *S. Luo is with the Graduate school of Systems and Information Engineering, University of Tsukuba, Ibaraki, Japan, 305-8577. E-mail: luo@sk.tsukuba.ac.jp*
- *M. Zhang is with the Graduate school of Systems and Information Engineering, University of Tsukuba, Ibaraki, Japan, 305-8577. E-mail: mzhang@sk.tsukuba.ac.jp*

1.1 Classification of Group Testing

Any group testing consists of a pooling procedure and a positive detecting procedure. A pooling procedure is a procedure of constructing a collection of pools called pooling design, determining which clones are put into which pools. A positive detecting procedure is a procedure of determining which clones are positives from the outcomes of group tests.

Group testing can be classified as either adaptive or nonadaptive, based on how a pooling design is constructed in the pooling procedure. In adaptive group testing, a pooling design is constructed in multiple-stage. In each stage, the outcomes of the group tests from previous stages are learned to construct the pools in the next stage. In nonadaptive group testing, a pooling design is constructed in one-stage, before any outcome of the group test is known. In practice, it is preferable to screen pools in parallel or with few stages for minimizing time consumption.

Group testing can also be classified as either combinatorial or probabilistic.To implement a combinatorial group testing, we have to construct a pooling design with desirable combinatorial properties such that all positives can be distinguished from negatives under

the assumption on the maximum number of positives and that of experimental errors. Related studies can be found in Du and Hwang [12], [13], [14], Dyachkov et al. [16] , Macula [29], and Ngo and Du [33]. To implement a probabilistic group testing, we need not only stochastic models for positives and for pooling results but also an efficient positive detecting algorithm to infer positives from erroneous pooling results based on a stochastic inference model. Probabilistic group testing is developed by Bruno et al. [5], Knill et al. [27], Mezard and Toninelli [30] and Uehara and Jimbo [37]. Bruno et al. [5] and Knill et al. [27] proposed a positive detecting algorithm called MCPD by using Markov chain Monte Carlo simulation method. Uehara and Jimbo [37] proposed another efficient algorithm called BNPD by using Bayesian network inference. As far as we know, MCPD and BNPD are the only known efficient positive detecting algorithms when experimental errors exist.

Probabilistic group testing has three noticeable merits in practice. First, from the perspective of detecting procedure, probabilistic group testing is expected to perform stably even when a relatively larger number of positives or/and experimental errors occur than expected. Second, probabilistic group testing reduces information loss by interpreting a measurement of a test into a multilevel state. Third, from the perspective of pooling procedure, implementing probabilistic group testing requires few restriction on the combinatorial structure of pooling designs. Particularly, random pooling designs are allowed. In fact, random pooling designs are often preferred in a real setting. This is because (i) a well-designed random pooling design can have a satisfying efficiency with desirable error-tolerant ability, (ii) it is unrealistic to expect to be able to find an appropriate combinatorial pooling design for every new pooling experiment, and (iii) random pooling designs facilitate robot automation and hence are very easy to construct.

1.2 Linear DNA library and Consecutive Positives

Motivated from applications to DNA library screening, Balding and Torney [4] considered the problem of pooling experiments for screening unique-sequence on a 1530-clone map of Aspergillus nidulans. The clone map has the properties that the clones are, with possibly a few exceptions, linearly ordered and no more than two of them cover any point on the genome. The goal of screening clone maps is to identify where a particular DNA segment occurs on a clone map.

In this problem, the clones are overlapped, and hence it may happen that one segment of interest occurs in a relatively large number of clones, but typically the number is predictable as it is related to the clone coverage. Colbourn [9] introduced the d-consecutive property to capture the overlap structure with efficient combinatorial group testing algorithms. Following his work, related studies can be found, for example, in Müller and Jimbo [31] and[32], Juan and Chang [24] and Ge et al. [18]. Bruno et al. [6] used heterogeneous priors and discussed optimization issues of nonadaptive random pooling designs on the assumption that an effective detecting procedure exists and experimental errors do not exist.

However, to our best knowledge, when experimental errors exist, probabilistic group testing for consecutive positives has not yet been discussed. In this paper, we study this problem to fill the gap.

2 Stochastic Models

This section introduces a prior probability distribution for consecutive positives with overlap structure, and extends Knill et al [27]'s method for consecutive positives.

2.1 Prior Knowledge of Consecutive Positives

The following notation will be consistently used throughout. Let $\mathcal{C} = \{c_1, c_2, ..., c_n\}$ be a set of clones. $\mathcal{C}$ is called a DNA library. Each clone c_i has an associated state $\sigma_i \in \{0, 1\}$. c_i is a positive if $\sigma_i = 1$, otherwise a negative. Let $P = \{c_i : \sigma_i = 1\}$ and $X_P = (\sigma_1, \ldots, \sigma_n)$ be the set of positives and its vector form corresponding to P, respectively. For convenience, P and X_P are used interchangeably, referred to the positive set. Denote by x_i the random variable of σ_i such that

$$x_i = \begin{cases} 0, & \text{if } \sigma_i = 0, \\ 1, & \text{if } \sigma_i = 1, \end{cases}$$

and by $X = (x_1, \ldots, x_n)$ the random vector of the associated states of $\mathcal{C}$.

When a DNA library is constructed by uniformly and randomly cloning of DNA segments, the expected number of positives $\mathbb{E}\big[\sum_{i=1}^{n} x_i\big] = d$ is used as the prior knowledge, for some positive number d. This prior knowledge appeared in Knill [26], Knill et al. [27] and Uehara and Jimbo [37]. The overlap information can analogously be collected and expressed in a similar form. The basic idea is to extend the d-consecutive positives property proposed by Colbourn [9] into a probabilistic version with broader sense.

$\mathcal{C}$ is said to be linear if it is associated with an linear order $c_i \prec c_{i+1}$, for $1 \leq i < n$. The positive set of linear $\mathcal{C}$ is said to have the multi-d-consecutive property if any subset of P that forms a consecutive set (under the ordering $\prec$) contains at most d positives. A subset of positives P' is said to be a maximum consecutive subset, if P' is a consecutive set but $P' \cup \{c\}$ is not a consecutive set for any $c \in P \setminus P'$. Notice that P corresponds to a unique partition with maximum

consecutive subsets. The positive set with the multi-d-consecutive property is allowed to have more than one maximum consecutive subsets, each of which contains at most d consecutive positives.

Without loss of generality and for the sake of simplicity, we analyze the overlap structure of the clone map of Aspergillus. When the clone map is screened, if the positive set is believed to approximately have the multi-2-consecutive positive property, then the following prior knowledge of the positives can be collected:

Info 1: the portion of positives among all clones is relatively small;
Info 2: although the positives are sparse, some of the positives tend to be consecutive;
Info 3: although some of the positives tend to be consecutive, maximum consecutive subsets tend to not get close to each other;
Info 4: no more than two of the positives tend to form a maximum consecutive subset.

Therefore, based on Info 1 through Info 4, we can express the prior knowledge of consecutive positives as follows:

$$\begin{cases} \mathbb{E}\big[\sum_{i=1}^{n} x_i\big] = d_1 \\ \mathbb{E}\big[\sum_{i=1}^{n-1} x_i x_{i+1}\big] = d_2 \\ \mathbb{E}\big[\sum_{i=1}^{n-2} x_i x_{i+2}\big] = d_3 \\ \mathbb{E}\big[\sum_{i=1}^{n-2} x_i x_{i+1} x_{i+2}\big] = d_4 \end{cases} \tag{I}$$

for some positive numbers d_1, d_2, d_3 and d_4. However, how to assign proper values to d_1 through d_4 is not within our present concern. Formally, we introduce the following assumption.

Assumption 1. *Appropriate values of d_1 through d_4 have been given as parameters.*

Generally, we denote by $\mathcal{D}$ a family of polynomials of $\sigma_1, \ldots, \sigma_n$, where

$$\mathcal{D} = \Big\{ \sum_{i=1}^{t} \prod_{c_j \in \mathcal{C}_i} \sigma_j : \exists t \in \mathbb{N} \text{ and } \exists\ t \text{ distinctive sets } \mathcal{C}_1, \ldots, \mathcal{C}_t \in 2^{\mathcal{C}} \setminus \{\emptyset\} \text{ such that } \mathcal{C} \subseteq \bigcup_{i=1}^{t} \mathcal{C}_i \Big\}.$$

Each polynomial of $\mathcal{D}$ is called an overlap polynomial of $\mathcal{C}$. We can collect and explicitly express prior knowledge of consecutive positives by using overlap polynomials in accordance with overlap information. If s overlap polynomials D'_j have been chosen from $\mathcal{D}$, P is said to have positive pattern $\boldsymbol{d}$ if $(\boldsymbol{d})_j = D'_j(X_P)$, for $j = 1, \ldots, s$. In addition, we denote by $|\boldsymbol{d}|$ the number of positives in positive pattern $\boldsymbol{d}$.

2.2 Prior Probability Distribution of Consecutive Positives

This part seeks a prior probability distribution that, in some sense of optimality, approximately incorporates the prior knowledge of consecutive positives (I).

2.2.1 The Principle of Maximum Entropy

Since Shannon's theorem [39] established the uniqueness of entropy as an information measure of uncertainty, the principle of maximum entropy has been wildly used to derive prior probability distributions. Intuitively speaking, more information means less uncertainty. Any probability distribution satisfying the constraints that has less uncertainty will contain more information, and thus implies something stronger than what the prior knowledge means. The principle of maximum entropy, as a method of statistical inference, is due to Jaynes [21], [22] and [23].

Based on the principle of maximum entropy, to derive the prior probability distribution $\mathbb{P}(X)$ which incorporates (I) but is free from any other knowledge, is to solve the following problem,

$$\underset{\mathbb{P}(X)}{\text{maximize}} \quad -\sum_{X \in \{0,1\}^n} \mathbb{P}(X) \log \mathbb{P}(X)$$
$$\text{subject to} \quad \begin{cases} \mathbb{E}_{\mathbb{P}}\big[D_1(X)\big] = d_1 \\ \mathbb{E}_{\mathbb{P}}\big[D_2(X)\big] = d_2 \\ \mathbb{E}_{\mathbb{P}}\big[D_3(X)\big] = d_3 \\ \mathbb{E}_{\mathbb{P}}\big[D_4(X)\big] = d_4 \\ \sum_{X \in \{0,1\}^n} \mathbb{P}(X) = 1. \end{cases} \tag{II}$$

As is well-known, Lagrange multiplier method leads to the solution

$$\mathbb{P}_\theta(X) = \frac{1}{Z(\theta)} \exp\Big\{ -\sum_{j=1}^{4} \theta_j D_j(X) \Big\}, \tag{III}$$

for some constant vector $\theta = (\theta_1, \theta_2, \theta_3, \theta_4)$ and constant $Z(\theta)$. In literature, (III) is called a Gibbs measure. The normalization constant

$$Z(\theta) = \sum_{X \in \{0,1\}^n} \exp\Big\{ -\sum_{j=1}^{4} \theta_j D_j(X) \Big\}$$

is called the partition function. It connects θ with the constants d_i by simultaneous equations

$$\frac{\partial - \log Z(\theta)}{\partial \theta_i} = d_i, \tag{IV}$$

for $i = 1, \ldots, 4$. Before discussing how to derive $\boldsymbol{\theta}$ from (IV), we first point out three useful properties of (III).

Property 1 (Consistency Property). *If $\mathbb{E}\big[\sum_{i=1}^{n} x_i\big] = d$ $(d > 0)$ is the only prior knowledge of positives, then the*

prior probability distribution determined by the principle of maximum entropy is

$$\mathbb{P}(X) = \left(\frac{d}{n}\right)^{\sum_{i=1}^{n} x_i} \left(1 - \frac{d}{n}\right)^{n - \sum_{i=1}^{n} x_i}.$$

The proof can be found in Jaynes [23]. From this property, we see that, provided proper prior knowledge of positives, the principle of maximum entropy can be used to obtain the prior probability used in Bruno et al [5], Knill [27], Knill et al [26] and Uehara [37], where the DNA library is considered to have been constructed by uniformly and randomly cloning of DNA segments. Hence, maximum-entropy distributions may be reasonable extensions for describing the overlap structure of clone maps.

Denote by N_j the set of neighbors of j. Define $N_1 \triangleq \{2,3\}$, $N_2 \triangleq \{1,3,4\}$, $N_{n-1} \triangleq \{n-3, n-2, n\}$, $N_n \triangleq \{n-2, n-1\}$ and $N_i \triangleq \{i-2, i-1, i+1, i+2\}$, for $3 \leq i \leq n-2$. For any configuration of x_i by assigning $x_i = \sigma_i$ for each $i \in [n]$ according to $\mathbb{P}_\theta(X)$, (III) has the following Markov-type property.

Property 2 (Local Markov Property)**.** *For any $j \in [n]$,*

$$\mathbb{P}_\theta(x_j | \sigma_i, i \in [n] \setminus \{j\}) = \mathbb{P}_\theta(x_j | \sigma_i, i \in N_j).$$

We show a sketch of the proof. A more general one can be found in Kindermann and Snell [25] and Pearl [34]. To see this, we rewrite the Gibbs measure into the product form:

$$\begin{aligned}\mathbb{P}_\theta(X) =& \frac{1}{Z(\theta)} \prod_{i=1}^{n} \exp\{-\theta_1 x_i\} \prod_{i=1}^{n-1} \exp\{-\theta_2 x_i x_{i+1}\} \\ & \prod_{i=1}^{n-2} \exp\{-\theta_3 x_i x_{i+2}\} \prod_{i=1}^{n-2} \exp\{-\theta_4 x_i x_{i+1} x_{i+2}\}.\end{aligned}$$

For $\sigma_j = 0, 1$, the condition probability on the left side is,

$$\begin{aligned}&\mathbb{P}_\theta(x_j = \sigma_j | \sigma_i, i \in [n] \setminus \{j\}) \\ =& \frac{\mathbb{P}_\theta(\sigma_1, \ldots, \sigma_{j-1}, x_j = \sigma_j, \sigma_{j+1}, \ldots, \sigma_n)}{\sum_{\sigma_j=0}^{1} \mathbb{P}_\theta(\sigma_1, \ldots, \sigma_{j-1}, x_j = \sigma_j, \sigma_{j+1}, \ldots, \sigma_n)}.\end{aligned}$$

Thus, after canceling out the normalization constant, terms of $\mathbb{P}_\theta(\sigma_1, \ldots, \sigma_n)$ that do not contain σ_j cancel from both the numerator and denominator of the condition probability and therefore this probability depends only on the random value of x_j and those of its neighbors. This property serves a probabilistic interpretation of multi-2-consecutive property, saying that the state of a clone is only influenced by the states of its previous two and also next two neighbors, if the clone has such neighbors.

Property 3 (Heterogenous Property)**.** *Given $X_1, X_2 \in \{0,1\}^n$, if they have the same positive pattern, then $\mathbb{P}_\theta(X_1) = \mathbb{P}_\theta(X_2)$.*

The proof is obvious. Notice that, for any fixed positive pattern $\boldsymbol{d}$, $\mathbb{P}_\theta$ is the heterogenous prior with the greatest uncertainty.

2.2.2 Estimation of Lagrange Multiplier

The constants θ_j are called Lagrange multipliers. As the same values of d_j's may come from different information sources, different estimation methods may be needed to derive θ from (IV). Here we discuss two cases. In the first case, the values of d_j's represent the degrees of belief. The belief is needless to be relevant with the outcomes of any random experiment. While in the second case, the values are obtained from the observation of the data $Y = \{Y_1, \ldots, Y_N\}$.

In the former case, we define

$$f(\theta) = -\sum_{j=1}^{4} d_j \theta_j - \log Z(\theta).$$

From (IV), we see that θ can be any stationary point of $f(\theta)$. Therefore, to obtain θ is to solve a gradient-square-minimization problem.

$$\underset{\theta \in \mathbb{R}^4}{\text{minimize}} \quad \|\nabla f(\theta)\|^2. \tag{V}$$

However, in the latter case, we may fit a model chosen from $\{\mathbb{P}_\theta : \theta \in \mathbb{R}^4\}$ to the given data. Assuming that $Y_1, \ldots, Y_N$ are i.i.d (independent and identically distributed) random samples drawn from $\mathbb{P}_\theta$ with unknown parameter θ, we estimate θ by applying maximum likelihood estimation method. The log-likelihood is defined by

$$\log \prod_{i=1}^{N} \mathbb{P}(Y_i|\theta) = \sum_{i=1}^{N} \Big(-\sum_{j=1}^{4} N_j(Y_i)\theta_j - \log Z(\theta) \Big). \tag{$*$}$$

Let $l_Y(\theta) = \frac{1}{N} \log \mathbb{P}(\{Y_1, \ldots, Y_N\}|\theta)$ and $d_j = \frac{1}{N} \sum_{i=1}^{N} D_j(Y_i)$, for $j = 1, \ldots, 4$. To obtain a maximum likelihood estimate of θ is to maximize $(*)$. Equivalently, we solve

$$\underset{\theta \in \mathbb{R}^4}{\text{maximize}} \quad l_Y(\theta) = -\sum_{j=1}^{4} d_j \theta_j - \log Z(\theta). \tag{VI}$$

Both problems are very difficult to solve, because the partition function $Z(\theta)$ involves exponentially many computations and usually cannot be known thus. Without calculating the partition function, Geyer and Thompson [20] developed a method to numerically solve (VI) by using importance sampling and Monte Carlo simulation. Descombes et al [11] further demonstrated an efficient conjugate gradient algorithm. Motived by their work, we employ their methods to solve (V). Here we sketch that (V) is also solvable (see [20] and [11] for detailed discussions).

The key idea is to estimate, for any fixed θ, $\|\nabla f(\theta)\|^2$ and its gradient by using importance sampling. By employing

$$\mathbb{P}_\psi(X) = \frac{1}{Z(\psi)} \exp\Big\{ -\sum_{j=1}^{4} \psi_j D_j(X) \Big\},$$

we reformulate $f(\theta)$ as follows:

$$f(\theta) = -\sum_{j=1}^{4} d_j\theta_j - \log\frac{Z(\theta)}{Z(\psi)} - \log Z(\psi).$$

It follows that

$$\begin{aligned} Z(\theta) &= \sum_X \exp\Big\{ -\sum_{j=1}^{4} (\theta_i - \psi_i) D_j(X)) \Big\} \\ &\quad \cdot \exp\Big\{ -\sum_{i=j}^{4} \psi_j D_j(X) \Big\} \\ &= \mathbb{E}_\psi\Big[\exp\Big\{ -\sum_{j=1}^{4} (\theta_i - \psi_i) D_j(X) \Big\} Z(\psi) \Big], \end{aligned}$$

where $\mathbb{E}_\psi$ refers to the expectation with respect to $\mathbb{P}_\psi$. Then, we obtain

$$\frac{Z(\theta)}{Z(\psi)} = \mathbb{E}_\psi\Big[\exp\Big\{ -\sum_{j=1}^{4} (\theta_j - \psi_j) D_j(X) \Big\} \Big]. \quad (2.1)$$

The significance of (2.1) lies in that although $Z(\theta)$ is unknown, $\frac{Z(\theta)}{Z(\psi)}$ can be estimated by a sampling of the known probability distribution $\mathbb{P}_\psi$. Furthermore, we can rewrite $\frac{1}{Z(\psi)}\frac{\partial Z(\theta)}{\partial \theta_j}$ into

$$\mathbb{E}_\psi\Big[-D_j(X) \exp\Big\{ -\sum_{j=1}^{4} (\theta_i - \psi_i) D_j(X) \Big\} \Big], \quad (2.2)$$

and similarly $\frac{1}{Z(\psi)}\frac{\partial^2 Z(\theta)}{\partial \theta_j \partial \theta_k}$ into

$$\mathbb{E}_\psi\Big[D_j(X) D_k(X) \exp\Big\{ -\sum_{j=1}^{4} (\theta_i - \psi_i) D_j(X) \Big\} \Big]. \quad (2.3)$$

Next, with (2.1), (2.2) and (2.3) we can obtain,

$$\begin{aligned} &\frac{\partial}{\partial \theta_k} \|\nabla f(\theta)\|^2 \\ =&2\sum_{j=1}^{4} \left(d_j + \frac{Z(\psi)}{Z(\theta)} \frac{1}{Z(\psi)} \frac{\partial Z(\theta)}{\partial \theta_j} \right) \left(-\left(\frac{Z(\psi)}{Z(\theta)} \right)^2 \left(\frac{1}{Z(\psi)} \right. \right. \\ &\left. \left. \frac{\partial Z(\theta)}{\partial \theta_k} \right) \left(\frac{1}{Z(\psi)} \frac{\partial Z(\theta)}{\partial \theta_j} \right) + \frac{Z(\psi)}{Z(\theta)} \left(\frac{1}{Z(\psi)} \frac{\partial^2 Z(\theta)}{\partial \theta_j \partial \theta_k} \right) \right). \end{aligned} \quad (2.4)$$

Using the Gibbs sampler (Geman and Geman [19]), for any fixed θ, by $\mathbb{P}_\psi$ we can theoretically estimate (2.1), (2.2), (2.3) and even higher-order partial derivatives of $Z(\theta)$ up to constant $\frac{1}{Z(\psi)}$. This allows us to learn local variation of $\|\nabla f(\theta)\|^2$ at any fixed θ. Those estimations will be helpful to numerically sovle (V). Particularly, the local markov property of $\mathbb{P}_\psi$ makes the sampling procedure efficient. We refer to Descombes et al [11] for other algorithmic details.

In fact, (VI) is stronger than (V) in the sense that any optimal of (VI) is also an optimal of (V), but not the reverse. Therefore, we may also formulate the first case in the stronger sense,

$$\underset{\theta \in \mathbb{R}^4}{\text{maximize}} \quad f(\theta) = -\sum_{j=1}^{4} d_j\theta_j - \log Z(\theta). \quad \text{(VII)}$$

From this formulation, we can see that both cases may be numerically solvable, but the first case is more general than the second one. By estimating a numerical solution of (V) within a given tolerance, the model can be approximately determined, and it can be served as the desired prior probability distribution that approximately incorporates our prior knowledge derived from the available information.

Assumption 2. *Given the values of d_1 through d_4 under Assumption 1, $\widehat{\theta}$ is an approximate optimal of* (VII). *$\mathbb{P}_{\widehat{\theta}}(X)$ is the desired prior probability distribution.*

2.3 Stochastic Model of Pooling Results

The following notation concerning pooling experiments will be consistently used. Let $\mathcal{A} = \{p_1, \ldots, p_m\}$ be a pooling design. Each pool p_i is a subset of $\mathcal{C}$ corresponding to the clones in the pool. The pooling design $\mathcal{A}$ constructed by determining each of n clones is put into which of m pools, can be represented by an $m \times n$ binary matrix $A = (a_{ij})$. Each entry a_{ij} is defined as follows:

$$a_{ij} = \begin{cases} 1, & \text{if } c_j \in p_i, \\ 0, & \text{if } c_j \notin p_i. \end{cases}$$

Then, A is called the incidence matrix of pooling design $\mathcal{A}$ and m the size of $\mathcal{A}$. For convenience, $\mathcal{A}$ and A are interchangeably referred to a pooling design.

The pooling result of screening pool p_i is denoted by $r(p_i)$. An actual observation of $r(p_i)$ is often given in multilevel measurement, taking any value from a set of test outcomes V, such as "negative", "weak positive", "medium positive" or "strong positive". Unfortunately, the pooling results are typically corrupted due to the inclusion of additional clones in a pool or to the failure of screening test. Therefore, we need infer P from erroneous pooling results.

To begin with, we introduce the stochastic model of pooling results given in Knill et al. [27]. The following assumptions will be used.

Assumption 3 (Knill et al. [27]). *The distribution of $r(p_i)$ depends only on the number of positives, $|p_i \cap P|$, in p_i.*

Assumption 4 (Knill et al. [27]). *Since the PCR was used for screening, we assume that $\Pr(r(p_i)||p_i \cap P|)$ depends only on $|p_i \cap P| = 0$ or $|p_i \cap P| \geq 1$.*

For any integer b, we define

$$\bar{b} = \begin{cases} 0 & \text{if } b = 0, \\ 1 & \text{if } b \geq 1. \end{cases}$$

Notice that, for each i, the number of positives in pool p_i can be represented by the ith coordinate of AX_P, that is, $|p_i \cap P| = (AX_P)_i$. Let R_i be a random variable of pooling result $r(p_i)$,

$$R_i = \begin{cases} 0 & \text{if } r(p_i) \text{ is negative,} \\ 1 & \text{if } r(p_i) \text{ is weak positive,} \\ 2 & \text{if } r(p_i) \text{ is medium positive,} \\ 3 & \text{if } r(p_i) \text{ is strong positive.} \end{cases}$$

Rewrite $\Pr(R_i = r_i|(AX_P)_i)$ into $f(r_i, (AX_P)_i)$ for short, and denote by $\Pr_A(R = r|X_P)$ the likelihood for the pooling results $r(\in \{0,1,2,3\}^m)$ obtained from pooling design A with m pools. By using Assumptions 3 and 4 and the chain rule, it can be expressed as a product:

$$\Pr_A(R = r|X_P) = \prod_{i=1}^{m} f(r_i, (\overline{AX_P})_i). \tag{VIII}$$

Let $f = \{f(j,i) : j = 0, \ldots, 3, \text{ and } i = 0, 1\}$. In real applications, f can be appropriately estimated (see Knill. et al [27]). In this application, we assume that f have been estimated and provided as parameters.

Assumption 5. *The values of $f(j,i)$ are known a priori, but $f(j,i) \neq 0$, for $j = 0, \ldots, 3$, and $i = 0, 1$.*

Particularly, notice that $f(1,0) + f(2,0) + f(3,0)$ is the likelihood of a false positive, whereas $f(0,1)$ is that of a false negative. Moreover, due to Assumption 3 and Assumption 4, the model does not require any restriction on the structure of positives, and thus is compatible with the overlap structure of consecutive positives.

2.4 Bayes Inference Model

To infer the consecutive positives, Bayes inference model is used to decode the erroneous pooling results. Denoting by $\Pr_A(X_P|r)$ the posterior probability that X_P is the positive set given the pooling results r under the pooling design A, by Bayes' rule we have

$$\Pr_A(X_P|r) \propto \Pr_A(X_P)\Pr_A(r|X_P),$$

where $\Pr_A(X_P)$ denotes the probability that X_P is the positive set given pooling design A is chosen to use. Prior knowledge of positives will affect the construction of pooling design A. However, the reverse does not reasonably hold, because knowing the construction of pooling design A will not add any information to the prior knowledge of positives. Therefore, we introduce the following assumption.

Assumption 6. *Given any $A \in \mathcal{A}$, $\Pr_A(X) = \mathbb{P}_{\widehat{\theta}}(X)$, for all $X \in \{0,1\}^n$.*

From Assumption 2, Assumption 5, Assumption 6 and (VIII) , it follows that the posterior probability is computable up to a multiplicative constant, and can be written as,

$$\Pr_A(X_P|r) \propto \mathbb{P}_{\widehat{\theta}}(X_P) \prod_{i=1}^{m} f(r_i, (\overline{AX_P})_i).$$

If we are interest in $\Pr_A(c \in P|r)$ for each clone c, this probability can be written as the marginal

$$\Pr_A(x_j = 1|r) \propto \sum_{\substack{P \subseteq \mathcal{C}: \\ c_j \in P}} \mathbb{P}_{\widehat{\theta}}(X_P) \prod_{i=1}^{m} f(r_i, (\overline{AX_P})_i). \tag{IX}$$

Notice that, if we substitute Binomial distribution for $\mathbb{P}_{\widehat{\theta}}$, the inference model turns out to be the one used in Knill et al. [27].

3 Detecting Algorithm for Consecutive Positives

This section presents a positive detecting algorithm for consecutive positives, called modified Markov chain Monte Carlo pool result decoder (MMCPD).

Our primary goal is to compute $\Pr_A(x_j = 1|r)$, for each j, and to choose the clones with highest posterior probabilities of being a positive for confirmatory individual tests. Those clones are called candidate positives. However, as it stands in (IX), every exact computation will involve exponential work in n, making our primary goal impractical as n is very large. Instead, we estimate the posterior probabilities by drawing samples approximately according to $\Pr_A(X|r)$. This idea works as long as a sampling method exists and is able to efficiently produce sufficiently many samples according to $\Pr_A(X|r)$. To sample from $\Pr_A(X|r)$, we can use the Gibbs sampler, since the full conditional distributions $\Pr_A(x_i|\sigma_1, \ldots, \sigma_{i-1}, \sigma_{i+1}, \ldots, \sigma_n, r)$ are easy to obtain for each i, due to the local markov property of $\mathbb{P}_{\widehat{\theta}}(X)$ and to the product expression of $\Pr(r|X)$.

In this way, Knill et al. [27]'s method can be extended for detecting consecutive positives. Mimicking their approach, we construct a Markov chain $X_0, X_1, \ldots$ on the family of all configurations of $\mathcal{C}$ with $\Pr_A(X|r)$ as its stationary distribution. Denote by X_S the vector form of the positive set S. Let X_S be the configuration of $\mathcal{C}$ at the end of step $t-1$, that is,

state X_{t-1}. Step t of the chain updates X_S by making a random decision for each of the n clones on whether to add it to or remove it from S. At the end of step t, that is, after all clones have been processed, state X_t is the final X_S. In the Gibbs sampler, for clone c_k, the probability of changing from X_S to $X_{S\triangle\{c_k\}}$ equals to

$$\frac{1}{1+\dfrac{\Pr_A(X_S|r)}{\Pr_A(X_{S\triangle\{c_k\}}|r)}}, \tag{X}$$

where $\triangle$ is the symmetric difference, that is, $A\triangle B \triangleq (A\setminus B)\cup(B\setminus A)$. Notice that with this probability of changing from X_S to $X_{S\triangle\{c_k\}}$, X_t and hence X_S tend in distribution to $\Pr_A(X|r)$. From Bayes' Theorem, $\mathbb{P}_{\widehat{\theta}}(X)$ and (VIII) we obtain,

$$\begin{aligned}&\frac{\Pr_A(X_S|r)}{\Pr_A(X_{S\triangle\{c_k\}}|r)}\\ =&\frac{\mathbb{P}_{\widehat{\theta}}(X_S)}{\mathbb{P}_{\widehat{\theta}}(X_{S\triangle\{c_k\}})}\frac{\prod\limits_{\substack{i\in[m]:\\ c_k\in p_i}} f(r_i,\overline{(AX_S)}_i)}{\prod\limits_{\substack{i\in[m]:\\ c_k\in p_i}} f(r_i,\overline{(AX_{S\triangle\{c_k\}})}_i)}.\end{aligned}$$

Notice that, for i such that $c_k\in p_i$, we have

$$(AX_{S\triangle\{c_k\}})_i=\begin{cases}(AX_S)_i+1, & \text{if } c_k\notin S,\\ (AX_S)_i-1, & \text{if } c_k\in S.\end{cases}$$

The prior ratio $\frac{\mathbb{P}_{\widehat{\theta}}(X_S)}{\mathbb{P}_{\widehat{\theta}}(X_{S\triangle\{c_k\}})}$ can also be efficiently updated and we will see that it only depends on the associated states of c_k's neighbors. For convenience, let $X_S=(\sigma_1^S,\ldots,\sigma_k^S,\ldots,\sigma_n^S)$ and $\sigma_{-1}^S=\sigma_0^S=\sigma_{n+1}^S=\sigma_{n+2}^S=0$, for any S. If $c_k\notin S$, then

$$\begin{aligned}&\frac{\mathbb{P}_{\widehat{\theta}}(X_S)}{\mathbb{P}_{\widehat{\theta}}(X_{S\triangle\{c_k\}})}=\exp\big\{\widehat{\theta}_1^S+\widehat{\theta}_2\sigma_{k-1}^S+\widehat{\theta}_2\sigma_{k+1}^S+\widehat{\theta}_3\sigma_{k-2}^S\\ &+\widehat{\theta}_3\sigma_{k+2}^S+\widehat{\theta}_4\sigma_{k-2}^S\sigma_{k-1}^S+\widehat{\theta}_4\sigma_{k-1}^S\sigma_{k+1}^S+\widehat{\theta}_4\sigma_{k+1}^S\sigma_{k+2}^S\big\}.\end{aligned}$$

If $c_k\in S$, then

$$\begin{aligned}&\frac{\mathbb{P}_{\widehat{\theta}}(X_S)}{\mathbb{P}_{\widehat{\theta}}(X_{S\triangle\{c_k\}})}=\exp\big\{-\widehat{\theta}_1-\widehat{\theta}_2\sigma_{k-1}^S-\widehat{\theta}_2\sigma_{k+1}^S-\widehat{\theta}_3\sigma_{k-2}^S\\ &-\widehat{\theta}_3\sigma_{k+2}^S-\widehat{\theta}_4\sigma_{k-2}\sigma_{k-1}^S-\widehat{\theta}_4\sigma_{k-1}^S\sigma_{k+1}^S-\widehat{\theta}_4\sigma_{k+1}^S\sigma_{k+2}^S\big\}.\end{aligned}$$

From Assumption 4 and that $\mathbb{P}_{\widehat{\theta}}$ is strictly positive (that is, $\mathbb{P}_{\widehat{\theta}}(X)>0$ for any X), we can see that the Markov chain is aperiodic, since it has a positive probability of remaining in the same state, and that the Markov chain is also irreducible, since it is possible to go from any state to any other state. This assures uniqueness of and convergence to the stationary distribution $\Pr_A(X|r)$ as well as the ergodic property. For discussions of the Gibbs sampler, we refer to Roberts and Smith [35] and Tierney [36]. Hence, starting with any state and running Gibbs sampler algorithm after a suitable warmup period, samples obtained from a realization of the Markov chain can approximately be regarded as the desired ones drawn according to $\Pr_A(X|r)$. By taking sufficiently many samples from the Markov chain, the proportion of the samples with $\sigma_i=1$ can be thus a Bayesian estimate of $\Pr_A(x_i=1|r)$.

As an extension of MCPD, MMCPD has two attractive merits. First, only local neighborhood relations and state values are needed to update the prior ratio and likelihood ratio, making the computation efficient. Second, its implementation requires no explicit restriction on the structure of pooling designs. This provides us a considerable freedom to choose and optimize the pooling designs we want to use, rather than the ones we have to use. Particularly, MMCPD is able to decode the pooling results obtained by screening random pooling designs.

4 Random k-set Design

This section discusses random pooling designs in the presence of consecutive positives and experimental errors.

4.1 Motivation and Related work

Among all, we are particularly interested in random k-set designs. A is said to be a random k-set design if each column of A is a k-subset of $[m]$ that is generated independently and uniformly at random. k is called the replication number, deciding how many pools each clone will be put into. Since there may exist other positive detecting algorithms for consecutive positives, a practical lower bound of the optimal value of m, independent of any positive detecting algorithm, will be attractive. This motivated us to seek a nontrivial information-theoretic lower bound of m. Information-theoretic bounds have been extensively studied in many fields. For example, related studies can be found in Atia and Saligrama [1], Chan et al [7] and [8], Wang et al. [40] and Wainwright [41]. Our computation method shares some commonplaces either in purpose or in formulation with prior work such as Bruno et al. [6], Knill et al. [26], Wang et al. [40] and Wainwright [41]. However, the algorithmic consideration of consecutive positives and experimental errors with random k-set designs makes our method different from any of them.

4.2 Problem Formulation

Notice that the Gibbs sampler (Geman and Geman [19]) allows us to estimate the distribution of the positive patterns of $\mathbb{P}_{\widehat{\theta}}$. Hence, to choose proper values of m and k for $\mathbb{P}_{\widehat{\theta}}$ and f, we begin by fixing a positive pattern $\boldsymbol{d}$ and formulate the subproblem in terms of $\boldsymbol{d}$ and f.

Let $\Omega(\boldsymbol{d})\subseteq\big\{(\sigma_1,\ldots,\sigma_n)\in\{0,1\}^n:\sum\limits_{i=1}^n\sigma_i=|\boldsymbol{d}|\big\}$ be the nonempty subset consisting of all the vectors with positive pattern $\boldsymbol{d}$. If X_P belongs to $\Omega(\boldsymbol{d})$, due to

the heterogenous property of $\mathbb{P}_{\widehat{\theta}}$, we think of X_P randomly and uniformly distributed over $\Omega(\boldsymbol{d})$. Denote by $\mathcal{A}_{m,k}$ the set of all k-set designs with size m. Similar to Assumption 6, only knowing the k-set design does not change the distribution of X_P over $\Omega(\boldsymbol{d})$. We formally states this by introducing Assumption 7.

Without causing confusion, when $\boldsymbol{d}$ is fixed and known, we denote by $\Pr(X)$ the probability that X is chosen from $\Omega(\boldsymbol{d})$ as the positive set and by $\Pr_A(X)$ the probability that X is the positive set given k-set design A.

Assumption 7. *When $\boldsymbol{d}$ is given and the values of m and k have been decided, given any k-set design $A \in \mathcal{A}_{m,k}$, $\Pr_A(X) = \Pr(X)$, for all $X \in \Omega(\boldsymbol{d})$.*

Given any instance of random k-set design A, a positive detecting algorithm ϕ with respect to A is a mapping from the m-vector observation r to an estimated vector of positives, say of the form $X_{\widehat{P}} = \phi_A(r)$. Accordingly, based on the k-set design A, the average probability of decoding error of any positive detecting algorithm is defined as

$$p_{err}(A) = \frac{1}{|\Omega(\boldsymbol{d})|} \sum_{X \in \Omega(\boldsymbol{d})} \Pr[\phi_A(r) \neq X|X].$$

We apply Fano's lemma [10] to lower bound the average probability of decoding error. Notice that $X \xrightarrow{A} \overline{AX} \xrightarrow{f} r$ forms a Markov chain, we can arrive at the form used in Wang et al. [40]

$$p_{err}(A) \geq 1 - \frac{H_A(r) - H_A(r|X) + 1}{\log |\Omega(\boldsymbol{d})|}.$$

The average probability of decoding error over $\mathcal{A}_{m,k}$ can be lower bounded as follows:

$$\mathbb{E}_A[p_{err}(A)] \geq 1 - \frac{\mathbb{E}_A\left[\sum_{i=1}^{m} H_A(r_i)\right] - \mathbb{E}_A[H_A(r|X)] + 1}{\log |\Omega(\boldsymbol{d})|}. \tag{XI}$$

4.3 A lower bound of $\mathbf{E}_A[p_{err}(A)]$

Provided n, $\boldsymbol{d}$ and f are fixed and known, we begin by stating a set of lemmas and a necessary condition on m and k for $\mathbb{E}_A[p_{err}(A)] = 0$, on the conditions that the positive set X with positive pattern $\boldsymbol{d}$ is randomly and uniformly distributed over $\Omega(\boldsymbol{d})$ and A is random k-set design, independently and uniformly from $\mathcal{A}_{m,k}$. Their proofs will be given in appendices.

Lemma 1. *For any $h \in [m]$,*

$$\mathbb{E}_A[H_A(r_h)] \leq -\sum_{j=0}^{3} f_{\boldsymbol{d},m,k}(j) \log f_{\boldsymbol{d},m,k}(j),$$

where

$$f_{\boldsymbol{d},m,k}(j) = \sum_{i=0}^{1} \lambda_{\boldsymbol{d},m,k}^{1-i} (1 - \lambda_{\boldsymbol{d},m,k})^i f(j,i),$$

and

$$\lambda_{\boldsymbol{d},m,k} = (1 - \frac{k}{m})^{|\boldsymbol{d}|}.$$

We introduce some notations for an exact computation of $\mathbb{E}_A[H_A(r_h|X)]$. Let $|\overline{AX}| = |\{h \in [m] : (\overline{AX})_h = 1\}|$. Given an X with positive pattern $\boldsymbol{d}$, denote by $P_{X,m,k}^{|\boldsymbol{d}|}(t)$ the probability that a k-set design A, uniformly chosen at random from $\mathcal{A}_{m,k}$, satisfies $|\overline{AX}| = t$. Recall that the columns of random k-set design are independent and uniformly chosen at random. This suggests that $P_{X,m,k}^{|\boldsymbol{d}|}(t) = P_{X',m,k}^{|\boldsymbol{d}|}(t)$ for any X and X' with $\boldsymbol{d}$. Thus, we can write $P_{m,k}^{|\boldsymbol{d}|}(t)$ for short. For any integers a and b, we define an indicate function of a and b as follows:

$$\delta(a,b) = \begin{cases} 0 & \text{if } a < b, \\ 1 & \text{if } a \geq b. \end{cases}$$

Lemma 2. *$P_{m,k}^{|\boldsymbol{d}|}(t)$ can be iteratively computed, for $t = k, \ldots, |\boldsymbol{d}|k$.*

$$P_{m,k}^{|\boldsymbol{d}|}(t) = \sum_{w=0}^{k} \mu_{m,k,t}(w) P_{m,k}^{|\boldsymbol{d}|-1}(t - k + w),$$

where

$$\mu_{m,k,t}(w) = \delta(t-k+w,k) \frac{\binom{t-k+w}{w}\binom{m-(t-k+w)}{k-w}}{\binom{m}{k}},$$

and in particular

$$P_{m,k}^{1}(k) = 1.$$

Using Lemma 2, we can obtain a closed expression of $\mathbb{E}_A[H_A(r|X)]$.

Lemma 3.

$$\mathbb{E}_A[H_A(r|X)] = \sum_{t=k}^{|\boldsymbol{d}|k} P_{m,k}^{|\boldsymbol{d}|}(t)\big[tH_f(1) + (m-t)H_f(0)\big],$$

where

$$H_f(i) = -\sum_{j=0}^{3} f(j,i) \log f(j,i),$$

for $i = 0, 1$.

Respectively substituting the results of Lemma 1 and Lemma 3 for $\mathbb{E}_A[H_A(r)]$ and $\mathbb{E}_A[H_A(r|X)]$ in (XI), we can easily compute an information-theoretic lower bound of the average probability of decoding error, that is,

$$\mathbb{E}_A[p_{err}(A)] \geq 1 - \frac{I(\boldsymbol{d},f,m,k) + 1}{\log |\Omega(\boldsymbol{d})|}, \tag{XII}$$

where

$$I(\boldsymbol{d},f,m,k) = -m \sum_{j=0}^{3} f_{\boldsymbol{d},m,k}(j) \log f_{\boldsymbol{d},m,k}(j) - \sum_{t=k}^{|\boldsymbol{d}|k} P_{m,k}^{|\boldsymbol{d}|}(t)\big[tH_f(1) + (m-t)H_f(0)\big].$$

For convenience, we denote by $ILB(|\Omega(\boldsymbol{d})|, f, m, k)$ the information-theoretic lower bound of (XII). By letting $\mathbb{E}_A[p_{err}(A)] = 0$, (XII) derives a necessary condition on m and k.

Theorem 4 (Necessity Theorem). $\mathbf{E}_A[p_{err}(A)]$ *vanishes to* 0 *only if* m *and* k *satisfy the condition*

$$\log |\Omega(\boldsymbol{d})| \leq I(\boldsymbol{d}, f, m, k) + 1.$$

In each subproblem where $\boldsymbol{d}$ is fixed and known, the Necessity Theorem implies that roughly $m = O(\log |\Omega(\boldsymbol{d})|)$ pools are least required for complete identifiability. This casts a light on the usefulness of knowing the overlap structure of clone maps. Prior knowledge of consecutive positives shrinks the size of the sample space of positives, which hence reduces the size of pooling design. This observation is consistent with the original motivation of Balding and Torney [4] and Colbourn [9].

Additionally, $ILB(|\Omega(\boldsymbol{d})|, f, m, k)$ can be used to predict the influence of f and k on the positive detectability of random k-set designs. To show this, we re-examine one of the screening problems studied by Knill et al. [27] and Uehara and Jimbo [37]. In the problem, the DNA library consists of 1298 clones without linear orders, among which there are four positives, and thus $|\Omega(\boldsymbol{d})| = \binom{1298}{4}$. Assigning two set of values to f, we fix f_1 and f_2 as shown in TABLE 1 and TABLE 2, respectively.

TABLE 1: f_1: Knill et al. [27]

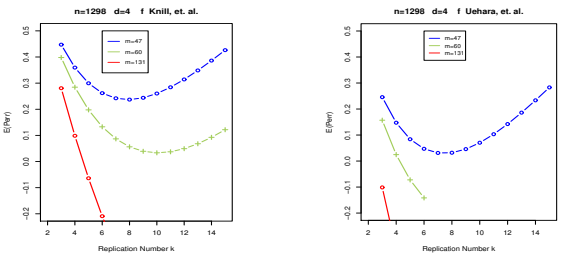
$f(0,0) = 0.871$	$f(0,1) = 0.05$
$f(1,0) = 0.016$	$f(1,1) = 0.11$
$f(2,0) = 0.035$	$f(2,1) = 0.27$
$f(3,0) = 0.078$	$f(3,1) = 0.57$

TABLE 2: f_2: Uehara and Jimbo [37]

$f(0,0) = 0.856$	$f(0,1) = 0.02$
$f(1,0) = 0.126$	$f(1,1) = 0.155$
$f(2,0) = 0.016$	$f(2,1) = 0.288$
$f(3,0) = 0.002$	$f(3,1) = 0.537$

By fixing the value of m ($m = 47, 60, 131$), Fig. 1 shows the variation of $ILB(|\Omega(\boldsymbol{d})|, f, m, k)$ as the values of f and k vary.

Similar to the LDPC (Low-Density Parity-Check) codes (see Gallager [17] and Mackay and Neal [28]), random k-set designs demonstrates a degree of error-tolerant ability, which is closely related with the value of k. The U-shaped curve of k is in accordance with our expectation, since either too small or too large a value of k will weaken the positive detectability of random k-set designs. As Uehara and Jimbo [37] mentioned, we also observe that there is a big gap



(a) f1 (b) f2

Fig. 1: Comparison of $ILB(|\Omega(\boldsymbol{d})|, f, m, k)$ subject to f_1 and f_2

between the values of $ILB(|\Omega(\boldsymbol{d})|, f, m, k)$ when $k = 3$ and $k = 4$.

Comparing Fig. 1 (a) with Fig. 1 (b), the influence of f is also noticeable. Uehara and Jimbo [37] proposed f_2 for repairing the unnatural monotonicity of f_1 in which $f(1,0) < f(2,0) < f(3,0)$. They also simulated the positive detectability of BNPD and MCPD subject to f_2 and showed that the seemingly slight modification remarkably improves the performance of MCPD and BNPD. Interestingly, as is shown in Fig. 1, the variation of $ILB(|\Omega(\boldsymbol{d})|, f, m, k)$ implies that f_1 is more difficult to cope with, which also coincides with their simulation results.

4.4 Random Pooling Designer

Given a positive pattern $\boldsymbol{d}$ it is often over complicated to obtain the exact $|\Omega(\boldsymbol{d})|$. Instead, we estimate a lower bound of $|\Omega(\boldsymbol{d})|$. We consider the positive pattern of the form $\boldsymbol{d} = (d_1, d_2, 0, 0)$ for some positive integer d_1 and nonnegative integer d_2. We claim that $|\underline{\Omega}(\boldsymbol{d})| = \binom{d_1-d_2}{d_2} \prod_{i=1}^{d_1-d_2} (n - d_1 + d_2 - 3i + 4)$ is a lower bound of $|\Omega(\boldsymbol{d})|$. This can be seen as follows. To begin with, it is obvious that there are $d_1 - 2d_2$ maximum consecutive subset with a single positive and d_2 maximum consecutive subset with 2 consecutive positives, and therefore we can treat each single positive as a red ball and each pair of consecutive positives as a blue ball. To calculate $|\Omega(\boldsymbol{d})|$, it is equivalent to counting how many ways these $d_1 - d_2$ balls with distinguishable colors can be put into $n-(d_1-d_2)+1$ urns such that any two ball have at least an urn between. This constraint implies that each ball will occupy at most 3 urns. Thus, to place the ith ball, at least $(n-(d_1-d_2)+1-3(i-1))$ urns are left to choose from, for $i = 1, 2, \ldots, d_1 - d_2$. Noticing that $\binom{d_1-d_2}{d_2}$ is the number of ways to color these balls without repetition, the desired lower bound is obtained.

With $|\underline{\Omega}(\boldsymbol{d})|$ and f, we give the algorithm called RPD (random pooling designer) for exhaustive search of the least value of m and some value of k that satisfy the Necessity Theorem. $m_{\boldsymbol{d},f}$ and $k_{\boldsymbol{d},f}$ denote by the least value of m and by the value of k with respect

to $|\underline{\Omega}(\boldsymbol{d})|$ and f, respectively. With the values of $m_{\boldsymbol{d},f}$ and $k_{\boldsymbol{d},f}$, various strategies can be used to choose the proper values of m and k for $\hat{\theta}$ and f.

Random Pooling Designer

Input: $|\underline{\Omega}(\boldsymbol{d})|$ and f
Output: $m_{\boldsymbol{d},f}$ and $k_{\boldsymbol{d},f}$
  $m_{\boldsymbol{d},f} \leftarrow 1$
  $k_{\boldsymbol{d},f} \leftarrow 1$
  $temp \leftarrow ILB(|\underline{\Omega}(\boldsymbol{d})|, f, m_{\boldsymbol{d},f}, k_{\boldsymbol{d},f})$
  while $temp > 0$ **do**
    $m_{\boldsymbol{d},f} \leftarrow m_{\boldsymbol{d},f} + 1$
    $k_{\boldsymbol{d},f} \leftarrow 1$
    while $temp > ILB(|\underline{\Omega}(\boldsymbol{d})|, f, m_{\boldsymbol{d},f}, k_{\boldsymbol{d},f})$ and $k \leq m$ **do**
      $temp \leftarrow ILB(|\underline{\Omega}(\boldsymbol{d})|, f, m_{\boldsymbol{d},f}, k_{\boldsymbol{d},f})$
      $k_{\boldsymbol{d},f} \leftarrow k_{\boldsymbol{d},f} + 1$
    end while
  end while

5 Simulation

This section shows the performance of MMCPD with random k-set designs chosen by RPD.

5.1 Simulation Method

The simulations are performed as follows.

1) Letting $n = 1298$ be the size of $\mathcal{C}$, the prior knowledge of consecutive positives is set to $d_1 = 6$, $d_2 = 2.4$, $d_3 = 0.01$ and $d_4 = 0.001$.
2) The likelihoods of experimental error f are set to f_1 given in TABLE 1.
3) Find an approximate prior probability distribution $\mathbb{P}_\theta$ that incorporates the prior knowledge by solving (V).
4) Estimate the distribution of positive patterns of $\mathbb{P}_\theta$.
5) Choose proper positive patterns, and compute $|\underline{\Omega}(\boldsymbol{d})|$, $m_{\boldsymbol{d},f}$ and $k_{\boldsymbol{d},f}$, for each chosen $\boldsymbol{d}$.
6) Choose $m_{\theta,f}$ and $k_{\theta,f}$ such that $m_{\theta,f}$ is the least value with $ILB(|\underline{\Omega}(\boldsymbol{d})|, f_1, m_{\theta,f}, k_{\theta,f}) < 0$, for all the chosen positive patterns.
7) The positive set P is given in two ways.
   - Fix $\{240, 241, 890, 891, 1001, 1002\}$ as the positive set.
   - Randomly choose a positive set approximately according to $\mathbb{P}_\theta$.
8) Generate a random k-set design A with some proper values of k and m.
9) Compute the number of positives in each pool.
10) Based on 11), determine the pooling results r randomly according to f_1.
11) Implement MMCPD to decode the corrupted pooling results r. The posterior probability of being a positive is estimated for each clone.
12) Clones are sorted in a decreasing order according to their posterior probabilities.

5.2 Preprocess of Pooling Procdedure

Solving (VII) with Descombes et al. [11]'s method, we find that $\hat{\theta} = (6.96, -7.65, 4.955, 2.01)$ is a suitable approximation for the the desired prior distribution. However, the estimation based on MCMC simulation is costly and hence a good initial θ will accelerate the convergence of Descombes et al [11]'s method. In implementation, to find a proper initial θ involves some guesswork and it can be done in a bisection way by using trial and failure method.

Next, we take 10000 samples for estimating the distribution of positive patterns of $\mathbb{P}_{\hat{\theta}}$ and list the positive patterns with sample mean above 0.0025 in TABLE 3. It also lists the corresponding $m_{\boldsymbol{d},f}$s and $k_{\boldsymbol{d},f}$s with respect to the positive patterns of our consideration and f_1. Particularly, $\boldsymbol{d} = (0, 0, 0, 0)$ is trivial and we remove it from our consideration. The total positive patterns of our consideration roughly account for 95% of the positive patterns of the samples.

TABLE 3: Estimated Distribution of $\boldsymbol{d}$ with respect to $\mathbb{P}_{\hat{\theta}}$ and Values of $m_{\boldsymbol{d},f}$ and $k_{\boldsymbol{d},f}$ with respect to f_1

$\boldsymbol{d}$	Sample Mean	$m_{\boldsymbol{d},f}$	$k_{\boldsymbol{d},f}$
(0, 0, 0, 0)	0.0286	-	-
(1, 0, 0, 0)	0.0307	17	7
(2, 0, 0, 0)	0.0192	33	9
(2, 1, 0, 0)	0.065	17	4
(3, 0, 0, 0)	0.0075	48	10
(3, 1, 0, 0)	0.0798	33	6
(4, 1, 0, 0)	0.0470	48	8
(4, 2, 0, 0)	0.0791	31	5
(5, 1, 0, 0)	0.0191	63	8
(5, 2, 0, 0)	0.0925	46	6
(6, 1, 0, 0)	0.006	96	9
(6, 2, 0, 0)	0.0617	61	7
(6, 3, 0, 0)	**0.0618**	**44**	**5**
(7, 2, 0, 0)	0.0232	75	7
(7, 3, 0, 0)	0.0717	58	6
(8, 2, 0, 0)	0.0071	88	8
(8, 3, 0, 0)	0.0459	72	6
(8, 4, 0, 0)	0.0385	55	4
(9, 3, 0, 0)	0.0189	85	7
(9, 4, 0, 0)	0.0446	69	5
(10, 3, 0, 0)	0.0048	98	7
(10, 4, 0, 0)	0.0272	82	6
(10, 5, 0, 0)	0.0181	65	4
(11, 4, 0, 0)	0.0118	95	6
(11, 5, 0, 0)	0.0217	78	5
(12, 4, 0, 0)	**0.0041**	**108**	**6**
(12, 5, 0, 0)	0.0128	91	5
(12, 6, 0, 0)	0.0059	74	4
(13, 5, 0, 0)	0.0040	104	6
(13, 6, 0, 0)	0.0087	86	5
(14, 6, 0, 0)	0.0048	99	5
(14, 7, 0, 0)	0.0028	82	4
(15, 7, 0, 0)	0.0038	95	4
Others	0.0216	-	-

5.3 Simulation 1: Fixed Positive Set

Of all the chosen positive patterns, $\boldsymbol{d} = (6, 3, 0, 0)$ attracts our attention because it has relatively high frequency, contains relatively large number of positives, but requires few pools with small value of k. By fixing the positive set $P = \{240, 241, 890, 891, 1001, 1002\}$, 500 simulations are implemented to show MMCPD's positive detectability with random 5-set designs of size 44.

In each simulation, beginning with the state drawn from i.i.d Bernoulli trials with parameter $q = \frac{6}{1298}$, the warmup period includes 5000 steps. After the warmup period, another 15000 steps are run, and the states obtained in every 3 steps are used as samples to estimate the posterior probability of being positive for each clone, that is, approximately the proportion of the obtained states including the given clone. We subsample the Markov chain in hope of weakening potential autocorrelations among the samples. Denote by CP the set of candidate positives. MMCPD outputs CP which consists of the six clones with the highest mean posterior probabilities. TABLE 4 shows the number of times among 500 simulations that $|P \cap CP|$ positives can be identified by using MMCPD.

TABLE 4: Positive Detectability of MMCPD for Fixed Positive Set

$\lvert P \cap CP \rvert$	Times
6	98
5	39
4	149
3	37
2	112
1	23
0	42

During the simulations, we observed that MMCPD detects the underlying true positives in a pairwise way, which can also be seen from the TABLE 4. Successfully decoding one underlying true positive will remarkably improve the performance of detecting the consecutive one. Besides, we can also see that the lower bound $ILB(|\underline{\Omega}(\boldsymbol{d})|, f_1, m_{\boldsymbol{d},f}, k_{\boldsymbol{d},f})$, though underestimates the underlying true minimal number of pools required, is nontrivial and useful.

5.4 Simulation 2: Random Positive Set

By using the TABLE 3, we can verify that $m_{\hat{\theta},f} = 108$ is the least value with $ILB(|\underline{\Omega}(\boldsymbol{d})|, f_1, m_{\hat{\theta},f}, k_{\hat{\theta},f}) < 0$, for all the chosen positive patterns, where $k_{\hat{\theta},f} = 6$.

1000 simulations are implemented to show MMCPD's positive detectability. In each simulation, a nonempty positive set is first randomly generated approximately according to $\mathbb{P}_{\hat{\theta}}$ and then a random 6-set design of size 108 is independently constructed. The decoding procedure of random positive set is the same with that of the fixed positive set except for the choice of the set of candidate positives. To evaluate the detectability of MMCPD for randomly generated positive set P, we introduce the success ratio $Suc(P) \triangleq \frac{|P \cap CP|}{|P|}$, where CP is the candidate positive set of $|P|$ clones which have the highest mean posterior probabilities. Obviously, the complete identifiability occurs when $Suc(P) = 1$. This measurement is nevertheless strict. Fig. 2 shows the simulation results.

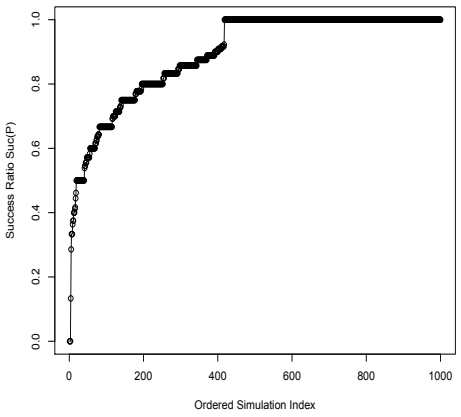


Fig. 2: Detectability of MMCPD for Random Positive set

In the simulations, MMCPD showed a stable and reliable performance with random designs. Its novel detectability is due to the prior knowledge of consecutive positives and also due to the proper choice of values of m and k suggested by RPD. There are 593 instances where MMCPD completely identified all the randomly generated positives. However, if too many positives or/and errors occur, MMCPD tends to fail. Denote by E_{FN} and E_{FP} the number of false negative errors and that of false positive errors, respectively. TABLE 5 lists all the instances with the success ratio less than 0.4.

TABLE 5: Instances Where MMCPD Performed Badly

Index	$\boldsymbol{d}$	E_{FN}	E_{FP}	$Suc(P)$
1	(1, 0, 0, 0)	0	21	0
2	(2, 0, 0, 0)	4	17	0
3	(1, 0, 0, 0)	0	14	0
4	(15, 6, 0, 0)	4	7	0.133333
5	(21, 10, 0, 0)	7	6	0.285714
6	(3, 0, 0, 0)	1	13	0.333333
7	(3, 0, 0, 0)	2	18	0.333333
8	(12, 5, 0, 0)	1	13	0.333333
9	(11, 4, 0, 0)	6	7	0.363636
10	(16, 7, 0, 0)	4	4	0.375
11	(8, 3, 0, 0)	3	17	0.375

Typically, these instances are hard. If too many positives or/and experimental errors occur, or if the positive pattern of consecutive positives deviates far from our prior knowledge, or a concurrence of some of these difficulties, will weaken the positive detectability of MMCPD. To overcome this, we may either enlarge the set of candidate positives or use a pooling design with more pools, or both.

6 Conclusion

In this paper, we study the problem of screening clone maps in the presence of experimental errors. In conclusion, as was foreseen in Balding and Torney [4] the overlap structure of clone maps facilitates efficient pooling experiments. Our method also allows non unique-screening as long as the prior knowledge can appropriately be interpreted.

Appendix A
Proof of Lemma 1

$\mathbb{E}_A\big[H_A(r_h)\big]$ can be written into

$$\sum_{A\in\mathcal{A}_{m,k}} \Pr(A)\bigg(-\sum_{j=0}^{3}\Pr_A(r_h=j)\log\Pr_A(r_h=j)\bigg).$$

Notice that $-x\log x$ defined on $\mathbb{R}^+$ is a concave function. Hence, by using Jensen's inequality we can obtain an upper bound of $\mathbb{E}_A\big[H_A(r_h)\big]$, that is,

$$\begin{aligned}
&\mathbb{E}_A\big[H_A(r_h)\big]\\
=&\sum_{j=0}^{3}\sum_{A\in\mathcal{A}_{m,k}} -\Pr(A)\Pr_A(r_h=j)\log\Pr_A(r_h=j)\\
\le&\sum_{j=0}^{3}-\bigg(\sum_{A\in\mathcal{A}_{m,k}}\Pr(A)\Pr_A(r_h=j)\bigg)\\
&\qquad\cdot\log\bigg(\sum_{A\in\mathcal{A}_{m,k}}\Pr(A)\Pr_A(r_h=j)\bigg)
\end{aligned}$$

Hence, to end our proof it suffices to calculate $\sum_{A\in\mathcal{A}_{m,k}}\Pr(A)\Pr_A(r_h=j)$. Since the sample space $\Omega(\boldsymbol{d})$ is fixed and known, due to the fact that $X\xrightarrow{A}\overline{AX}\xrightarrow{f}r$ forms a Markov chain we have

$$\Pr_A(r_h=j)=\sum_{X\in\Omega(\boldsymbol{d})}\Pr_A(X)\Pr_A(r_h=j|X).$$

Due to Assumptions 3 and 4, we have

$$\begin{aligned}
&\Pr_A(r_h=j|X)\\
=&\sum_{i=0}^{1}\Pr_A(r_h=j,(\overline{AX})_h=i|X)\\
=&\sum_{i=0}^{1}\Pr(r_h=j|(\overline{AX})_h=i)\Pr((\overline{AX})_h=i|X)\\
=&\sum_{i=0}^{1}f(j,i)\Pr((\overline{AX})_h=i|X).
\end{aligned}\tag{A. 1}$$

Since $\Pr_A(X)=\Pr(X)$ is assumed in Assumption 7, we further have

$$\begin{aligned}
&\sum_{A\in\mathcal{A}_{m,k}}\Pr(A)\Pr_A(r_h=j)\\
=&\sum_{A\in\mathcal{A}_{m,k}}\Pr(A)\sum_{X\in\Omega(\boldsymbol{d})}\Pr_A(X)\Pr_A(r_h=j|X)\\
=&\sum_{X\in\Omega(\boldsymbol{d})}\Pr(X)\sum_{A\in\mathcal{A}_{m,k}}\Pr(A)\Pr_A(r_h=j|X).
\end{aligned}\tag{A. 2}$$

Since for any fixed X with $|\boldsymbol{d}|$ positives, $\Pr((\overline{AX})_h=i|X)=1$ if pooling design A contains at least one positive in the hth pool, otherwise 0. When the hth pool is considered, if X is fixed and A is random k-set design constructed by independent column generation of uniform random k-sets and hence A is chosen independently and uniformly from $\mathcal{A}_{m,k}$, then $\Pr((\overline{AX})_h=0|X)=1$ with probability $(1-\frac{k}{m})^{|\boldsymbol{d}|}$. Therefore, we have

$$\begin{aligned}
&\sum_{A\in\mathcal{A}_{m,k}}\Pr(A)\Pr_A(r_h=j|X)\\
=&\sum_{i=0}^{1}f(j,i)\sum_{A\in\mathcal{A}_{m,k}}\Pr(A)\Pr((\overline{AX})_h=i|X)\\
=&\sum_{i=0}^{1}f(j,i)\bigg(1-\frac{k}{m}\bigg)^{|\boldsymbol{d}|(1-i)}\bigg(1-\bigg(1-\frac{k}{m}\bigg)^{|\boldsymbol{d}|}\bigg)^{i}.
\end{aligned}\tag{A. 3}$$

Consequently, we finish the proof by combining the results of (A. 1), (A. 2) and (A. 3).

Appendix B
Proof of Lemma 2

We assume $|\boldsymbol{d}|k\le m$. For a fixed X_P with $|\boldsymbol{d}|$ positives and a fixed t, let $\mathcal{A}^{|\boldsymbol{d}|}_{X_P,m,k}(t)$ be the subset of all the k-set designs of size m satisfying $|\overline{AX_P}|=t$. Obviously, the value of t can range from k to $|\boldsymbol{d}|k$, and it is determined by the columns corresponding to the positive set P. Since the columns of random k-set designs are randomly and independently generated, it follows that $P^{|\boldsymbol{d}|}_{X_P,m,k}(t)$ only depends on $|\boldsymbol{d}|$, m and k. Therefore, any X_P with $|\boldsymbol{d}|$ positives can be used to derive $P^{|\boldsymbol{d}|}_{m,k}(t)$. Let $u^{|\boldsymbol{d}|}_{X_P,m,k}(t)$ be the number of ways of choosing columns corresponding to X_P such that any k-set design with those columns belongs to $\mathcal{A}^{|\boldsymbol{d}|}_{X_P,m,k}(t)$. Due to construction method of random k-set designs, we have

$$P^{|\boldsymbol{d}|}_{m,k}(t)=\frac{|\mathcal{A}^{|\boldsymbol{d}|}_{X_P,m,k}(t)|}{\binom{m}{k}^n}=\frac{u^{|\boldsymbol{d}|}_{X_P,m,k}(t)}{\binom{m}{k}^{\boldsymbol{d}}}.$$

Hence, to end our proof it suffices to give an iterative expression of $u^{|\boldsymbol{d}|}_{X_P,m,k}(t)$. Without loss of generality and for the sake of simplicity, we express

$u^{|\boldsymbol{d}|}_{X_P,m,k}(t)$ by $u^{|\boldsymbol{d}-1|}_{X_{P'},m,k}(t')$s, for some $P' \subset P$ with $|\boldsymbol{d}|-1$ positives.

Let $v_1,\ldots,v_{|\boldsymbol{d}|-1}$ be the columns of a k-set design A corresponding the positive set P' and $v_{|\boldsymbol{d}|}$ be the column corresponding to $P \setminus P'$. Notice that, for $w = 0,1,\ldots,k$, if there are $t-k+w$ positive coordinates in the vector $\sum_{j=1}^{|\boldsymbol{d}|-1} v_j$, then we have $\binom{t-k+w}{w}\binom{m-(t-k+w)}{k-w}$ ways to choose $v_{|\boldsymbol{d}|}$ such that $\sum_{j=1}^{|\boldsymbol{d}|-1} v_j + v_{|\boldsymbol{d}|}$ has t positive coordinates. This implies that

$$u^{|\boldsymbol{d}|}_{X_P,m,k}(t) = \sum_{w=0}^{k} \delta(t-k+w,k)\binom{t-k+w}{w} \cdot \binom{m-(t-k+w)}{k-w} u^{|\boldsymbol{d}|-1}_{X_{P'},m,k}(t-k+w).$$

Dividing by $\binom{m}{k}^{|\boldsymbol{d}|}$ on both sides, the proof ends.

APPENDIX C
PROOF OF LEMMA 3

By the definition of conditional entropy and Assumption 7, we have

$$\begin{aligned}
&\mathbb{E}_A\big[H_A(r|X)\big] \\
&= \sum_{A\in\mathcal{A}_{m,k}} \Pr(A) \sum_{X\in\Omega(\boldsymbol{d})} \Pr_A(X) H_A(r|X) \\
&= \sum_{X\in\Omega(\boldsymbol{d})} \Pr(X) \sum_{A\in\mathcal{A}_{m,k}} \Pr(A)\Bigg(\sum_{r\in\{0,1,2,3\}^m} -\Pr_A(r|X) \\
&\quad \cdot \log \Pr_A(r|X)\Bigg).
\end{aligned} \tag{C. 1}$$

Given any $X \in \Omega(\boldsymbol{d})$, we have the partition

$$\mathcal{A}_{m,k} = \bigcup_{t=k}^{|\boldsymbol{d}|k} \mathcal{A}^{|\boldsymbol{d}|}_{X,m,k}(t).$$

Moreover, recall that $\Pr_A(r|X) = \prod_{h=1}^{m} f(r_h,(\overline{AX})_h)$ due to Assumptions 3 and 4. Therefore, this implies when $X \in \Omega(\boldsymbol{d})$ is fixed, for any given $A \in \mathcal{A}^{|\boldsymbol{d}|}_{X,m,k}(t)$,

$$\begin{aligned}
H_A(r|X) &= \sum_{h=1}^{m} H(r_h|(\overline{AX})_h) \\
&= t\sum_{j=0}^{3} -f(j,1)\log f(j,1) \\
&\quad + (m-t)\sum_{j=0}^{3} -f(j,0)\log f(j,0) \\
&= tH_f(1) + (m-t)H_f(0).
\end{aligned} \tag{C. 2}$$

Using the result of (C. 2) and Lemma 3, it follows that

$$\begin{aligned}
(C.1) &= \sum_{X\in\Omega(\boldsymbol{d})} \Pr(X) \sum_{t=k}^{|\boldsymbol{d}|k} \sum_{A\in\mathcal{A}^{|\boldsymbol{d}|}_{X,m,k}(t)} \Pr(A) \\
&\qquad \cdot\Big(tH_f(1) + (m-t)H_f(0)\Big) \\
&= \sum_{X\in\Omega(\boldsymbol{d})} \Pr(X) \sum_{t=k}^{|\boldsymbol{d}|k} P^{|\boldsymbol{d}|}_{m,k}(t)\Big(tH_f(1) \\
&\qquad + (m-t)H_f(0)\Big) \\
&= \sum_{t=k}^{|\boldsymbol{d}|k} P^{|\boldsymbol{d}|}_{m,k}(t)\Big(tH_f(1) + (m-t)H_f(0)\Big).
\end{aligned}$$

ACKNOWLEDGMENTS

The authors deeply thank Professor Maiko Shigeno for enlightening discussions and helpful comments.

REFERENCES

[1] G. Atia and V. Saligrama, "Boolean compressed sensing and noisy group testing," CoRR, vol. abs/0907.1061, 2009.

[2] E. Barillot, B. Lacroix, and D. Cohen, "Theoretical Analysis of Library Screening Using an N-Dimensional Pooling Strategy," Nucleic Acids Research, vol. 19, no. 22, pp. 6241-6247, 1991.

[3] T. Berger, J.W. Mandell, and P. Subrahmanya, "Maximally Efficient Two-Stage Screening," Biometrics, vol. 56, no. 3, pp. 833-840, 2000.

[4] D.J. Balding and D.C.Torney, "The design of pooling experiments for screening a clone map," Fungal Genet. Biol., 21, pp.302-307, 1997.

[5] W.J. Bruno, E. Knill, D.J. Balding, D.C. Bruce, N.A. Doggett, W.W. Sawhill, R.L. Stallings, C.C. Whittaker, and D.C. Torney, "Efficient Pooling Designs for Library Screening," Genomics, vol. 26, no. 1, pp. 21-30, 1995.

[6] W.J. Bruno, F. Sun, D.C. Torney, "Optimizing nonadaptive group tests for objects with heterogenous priors," SIAM J. Appl. Math., 58, pp. 1043-1059, 1998.

[7] C. L. Chan, P. H. Che, S. Jaggi, and V. Saligrama, "Non-adaptive probabilistic group testing with noisy measurements: Near-optimal bounds with efficient algorithms," in Communication, Control, and Computing (Allerton), 2011 49th Annual Allerton Conference on, Oct. 2010.

[8] C. L. Chan, S. Jaggi, V. Saligrama, and S. Agnihotri, "Non-adaptive group testing: Explicit bounds and novel algorithms," CoRR, vol. abs/1202.0206, 2012.

[9] C.J. Colbourn, "Group testing for consecutive positives," Ann. Combinatorics, 3, pp. 37-41, 1999.

[10] T. Cover and J. Thomas, *Elements of Information Theory*. John Wiley and Sons, 1991.

[11] X. Descombes, R. Morris, J. Zerubia, and M. Berthod, "Maximum likelihood estimation of Markov random field parameters using markov chain Monte Carlo algorithms," Proc. Int. Workshop EMMCVPR'97: Energy Minimization Methods Computer Vision Pattern Recognition, vol.1223, pp.133-148, 1997.

[12] D.Z. Du and F.K. Hwang, *Combinatorial Group Testing and Its Application*. World Scientific, 2000.

[13] D.Z. Du and F.K. Hwang, *Pooling Designs and Nonadaptive Group Testing: Important Tools for DNA Sequencing*. World Scientific, 2006.

[14] Y.X. Cheng and D.Z. Du, "New Constructions of One- and Two-Stage Pooling Designs," J. Computational Biology, vol. 15, no. 2, pp. 195-205, 2008.

[15] R. Dorfman, "The Detection of Defective Members of Large Populations," Ann. Math. Statistics, vol.14, No.4, pp. 436-440, 1943.

[16] A.G. D'yachkov, F.K. Hwang, A.J. Macula, P.A. Vilenkin, and C.Weng, "A Construction of Pooling Designs with Some Happy Surprises," J. Computational Biology, vol. 12, no. 8, pp. 1129-1136, 2005.
[17] R.G. Gallager, "Low-Density Parity-Check Codes," IRE Trans. Information Theory, vol. 8, pp. 21-28, 1962.
[18] G. Ge, Y. Miao, and X. Zhang, "On Block Sequences of Steiner Quadruple Systems with Error Correcting Consecutive Unions," SIAM J. Discrete Math., pp. 940-958, 2009.
[19] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," IEEE Trans. Pattern Anal. Machine Intell., vol. PAMl-6, pp. 721-741, 1984.
[20] C.J. Geyer and E. A. Thompson, "Constrained Monte Carlo maximum likelihood for dependent data (with discussion)," J. Roy. Statist. Soc., vol.54, pp. 657-699, 1992.
[21] E.T. Jaynes, "Information theory and statistical mechanics," Phys. Rev. 106, pp. 620-630, 1957.
[22] E.T. Jaynes, "Information theory and statistical mechanics, II," Phys. Rev. 108, pp. 171-190, 1957.
[23] E.T. Jaynes, "Prior Probabilities," IEEE Trans on Systems Science and Cybernetics, vol 4, no. 3, pp. 227-241, 1963.
[24] J.S. Juan, G.J. Chang, "Adaptive group testing for consecutive positives," Discrete Math., vol. 308, 7, pp. 1124-1129, 2008.
[25] R. Kindermann and J.L Snell, *Markov Random Fields and Their Applications*. American Mathematical Society, 1980.
[26] E. Knill, "Lower bounds for identifying subset members with subset queries," Proc. 6th ACM-SIAM sump. Discrete Algorithms, pp 369-377, 1995.
[27] E. Knill, A. Schliep, and D.C. Torney, "Interpretation of Pooling Experiments Using the Markov Chain Monte Carlo Method," J. Computational Biology, vol. 3, no. 3, pp. 395-406, 1996.
[28] D.J.C. MacKay and R.M. Neal, "Near Shannon Limit Performance of Low Density Parity Check Codes," IEE Electronics Letters, vol. 32, no. 18, pp. 1645-1655, 1996.
[29] A.J. Macula, "Probabilistic Nonadaptive Group Testing in the Presence of Errors and DNA Library Screening," Ann. Combinatorics, vol. 3, no. 1, pp. 61-69, 1999.
[30] M. Mézard and C. Toninelli, "Group Testing with Random Pools: Optimal Two-Stage Algorithms," IEEE Trans. Info. Th. 57, 1736-1745, 2011.
[31] M. Müller and M. Jimbo, "Consecutive positive detectable matrices and group testing for consecutive positives," Discrete Math., vol. 279, pp. 369-381, 2004.
[32] M. Müller and M. Jimbo, "Cyclic sequences of k-subset with distinct consecutive unions," Discrete Math., vol. 308, pp. 457-464, 2008.
[33] H.Q. Ngo and D.Z. Du, "A Survey on Combinatorial Group Testing Algorithms with Applications to DNA Library Screening," Discrete Math. Problems with Medical Applications, DIMACS Ser. Discrete Math. and Theoretical Computer Science, vol. 55, pp. 171-182, 2000.
[34] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1997.
[35] G.O. Roberts and A.F.M. Smith, "Simple conditions for the convergence of the Gibbs sampler and Metropolis-Hastings algorithms," Stochastic Process. Appl. vol. 49, pp. 207-216, 1994.
[36] L. Tierney, "Markov Chains for exploring posterior distributions, " Ann. Statistics, vol. 22, No. 4, pp. 1701-1728, 1994.
[37] H. Uehara and M. Jimbo, "A positive detecting code and its decoding algorithm for DNA library screening," IEEE/ACM Trans Computational Biology and Bioinformatics, vol. 6, no. 4, pp. 652-666, 2009.
[38] P. Sham, J.S. Bader, I. Craig, M. O'Donovan, and M. Owen, "DNA Pooling: A Tool for Large-Scale Association Studies," Nature Rev. Genetics, vol. 3, no. 11, pp. 862-871, Nov. 2002.
[39] C.E. Shannon, "A Mathematical Theory of Communication," Bell System Technical Journal, 27, pp. 379-423 & 623−656, July & October, 1948.
[40] W. Wang, M.J. Wainwright, K. Ramchandran, "Information-Theoretic Limits on Sparse Signal Recovery: Dense Versus Sparse Measurement Matrices," arXiv:0806.0604, 2008.
[41] M.J. Wainwright, "Information-theoretic bounds on sparsity recovery in the high-dimensional and noisy setting," IEEE Trans. Inf. Theory, vol. 55, pp.5728-5741, 2009.