Learning Immune-Defectives Graph through Group Tests

Abhinav Ganesan, Sidharth Jaggi, and Venkatesh Saligrama, *Senior Member, IEEE*

Abstract—This paper deals with an abstraction of a unified problem of drug discovery and pathogen identification. Here, the "lead compounds" are abstracted as inhibitors, pathogenic proteins as defectives, and the mixture of "ineffective" chemical compounds and non-pathogenic proteins as normal items. A defective could be immune to the presence of an inhibitor in a test. So, a test containing a defective is positive iff it does not contain its "associated" inhibitor. The goal of this paper is to identify the defectives, inhibitors, and their "associations" with high probability, or in other words, learn the Immune Defectives Graph (IDG). We propose a probabilistic non-adaptive pooling design, a probabilistic two-stage adaptive pooling design and decoding algorithms for learning the IDG. For the two-stage adaptive-pooling design, we show that the sample complexity of the number of tests required to guarantee recovery of the inhibitors, defectives and their associations with high probability, i.e., the upper bound, exceeds the proposed lower bound by a logarithmic multiplicative factor in the number of items. For the non-adaptive pooling design, in the large inhibitor regime, we show that the upper bound exceeds the proposed lower bound by a logarithmic multiplicative factor in the number of inhibitors.

I. Introduction

Drug discovery is known to be a tedious process, where early stages of drug discovery involve finding 'blocker' or 'lead' compounds amidst billions of chemical compounds [1], [2]. These lead compounds bind to a biomolecular target, which is a pathogenic protein, and thus inhibit the function of the protein. Such compounds are later used to produce new drugs. Finding a relatively small number of lead compounds amidst a large collection of chemical compounds is a challenging task. A complementary problem involves identifying pathogenic proteins amidst non-pathogenic ones, both of which are structurally identical in some respects. For instance, out of five known species of ebolavirus, only four of them are pathogenic to humans (see p. 5 in [2]) and a similar example can be found in arenavirus [3]. Some of these pathogenic proteins might share a common inhibitory mechanism against a lead compound which serves to distinguish them from the non-pathogenic ones [3]. So, finding potential pathogenic proteins amidst a large collection of biomolecules by testing them against known inhibitory compounds is a problem complementary to the problem of lead compound discovery. The lead compounds can be abstracted as inhibitor items, the pathogenic proteins as defective items, and the others as normal items. Now, the above problems can be

A. Ganesan and S. Jaggi are with the Chinese University of Hong Kong, Hong Kong (e-mail: abhinav@inc.cuhk.edu.hk; jaggi@ie.cuhk.edu.hk). V. Saligrama is with Boston University, Boston, USA (e-mail: srv@bu.edu).

combined to be viewed as an inhibitor-defective classification problem on the mixture of pathogenic and non-pathogenic proteins, and billions of chemical compounds. This unifies the process of finding both the pathogenic proteins and the lead compounds. An efficient means of solving this problem could potentially be applied in computer-assisted drug and pathogen identification. A natural consideration is that, while some pathogenic proteins might be inhibited by some lead compounds, other pathogenic proteins might be immune to some of these lead compounds present in the mixture of items. In other words, each defective item is possibly immune to the presence of some inhibitor items so that its expression cannot be prevented by the presence of those inhibitors when tested together. By definition, an inhibitor inhibits at least one defective. Learning this inhibitor-defective interaction as well as classifying the inhibitors and defectives efficiently through group testing is presented this work.

A representation of this model, which we refer to as the Immune-Defectives Graph (IDG) model, is given in Fig. 1. The presence of a directed edge between a pair of vertices $\left(w_{i_{k_1}}, w_{j_{k_2}}\right)$ represents the inhibition of the defective $w_{j_{k_2}}$ by the inhibitor $w_{i_{k_1}}$ and the absence of a directed edge between a pair of vertices $\left(w_{i_{k_1}}, w_{j_{k_2}}\right)$ indicates that the inhibitor $w_{i_{k_1}}$ does not affect the expression of the defective $w_{j_{k_2}}$ when tested together.
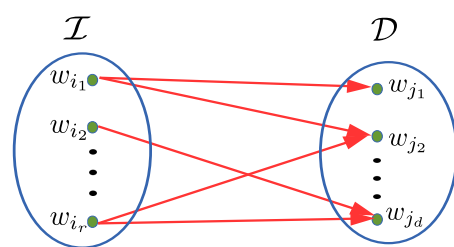

Fig. 1. A representation of the IDG Model, where $\mathcal{I}$ represents the set of inhibitors and $\mathcal{D}$ represents the set of defectives.

Example 1: An instance of the IDG model is given in Fig. 2. In this example, the outcome of a test is positive iff a defective $w_{j_{k_2}}$, for some k_2, is present in the test and its associated inhibitor $w_{i_{k_2}}$ does not appear in the test. Observe that if the item-pair $\left(w_{i_{k_1}}, w_{j_{k_2}}\right)$, for $k_1 \neq k_2$, appears in a test and $w_{i_{k_2}}$ does not appear in the test, then the outcome is positive. Also, if the item-pair $\left(w_{i_{k_2}}, w_{j_{k_2}}\right)$ appears in a test and if $w_{j_{k_2'}}$ also appears in the test but not $w_{i_{k_2'}}$, then the test outcome is positive. But if the appearance of every defective $w_{j_{k_2'}}$ in a test is compensated by the appearance of

2

its associated inhibitor $w_{i_{k'_2}}$ in the test, then the test outcome is negative. The outcome of a test is also negative when none of the defectives appear in a test.
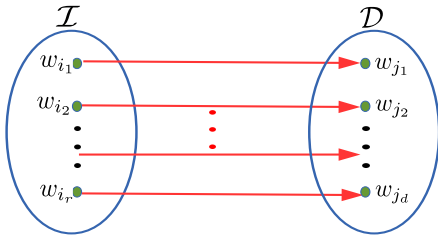

Fig. 2. An example for the IDG Model where each defective is associated with a single distinct inhibitor so that $r = d$.

The IDG model can also be viewed as a generalization of the 1-inhibitor model introduced by Farach et al. in [4]. This model was motivated by errors in blood testing where blocker compounds (i.e., inhibitors) block the expression of defectives in a test [5]. This is also motivated by drug discovery applications where the inhibitors are actually desirable items that inhibit the pathogens [6]. In the 1-inhibitor model, a test outcome is positive iff there is at least one defective and no inhibitors in the test. So, the presence of a single inhibitor is sufficient to ensure that the test outcome is negative.

Efficient testing involves pooling different items together in a every test, in other words, called group testing [7] so that the number of tests can be minimized. The pooling methodology can be of two kinds, namely non-adaptive and adaptive pooling designs. In non-adaptive pooling designs, any pool constructed for testing is independent of the previous test outcomes, while in adaptive pooling designs, some constructed pools might depend on the previous test outcomes. A k-stage adaptive pooling design is comprised of pool construction and testing in k-stages, where the pools constructed for (non-adaptive) testing in the kth stage depend on the outcomes in the previous stages. While adaptive group testing generally requires lesser number of tests than non-adaptive group testing, the latter inherently supports parallel testing of multiple pools. Thus, non-adaptive group testing is more economical (because it allows for automation) as well as saves time (because the pools can be prepared all at once) which are of concern in library screening applications [8]. The 1-inhibitor model has been extensively studied, and several adaptive and non-adaptive pooling designs for classification of the inhibitors and the defectives are known (refer, [9]–[12]). A detailed survey of known non-adaptive and adaptive pooling designs for the 1-inhibitor model is given in [13]. The best (in terms of number of tests) known non-adaptive pooling design that guarantees high probability classification of the inhibitors and defectives is proposed in [13]. The non-adaptive pooling design proposed in [13] requires $O(d \log n)$ tests in the $r = O(d)$ regime and $O\left(\frac{r^2}{d} \log n\right)$ tests in the $d = o(r)$ regime to guarantee classification of both the inhibitors and defectives with high probability[1]. In the small inhibitor, i.e., $r = O(d)$ regime, the

upper bound on the number of tests matches with the lower bound while in the large inhibitor, i.e., $d = o(r)$ regime, the upper bound exceeds the lower bound of $O\left(\frac{r^2}{d \log \frac{r}{d}} \log n\right)$ by a $\log \frac{r}{d}$ multiplicative factor. Nonetheless, the 1-inhibitor model constrains that every inhibitor must inhibit every defective, which is likely to be a tight requirement in practice. So, the IDG model is a more practical variant of the 1-inhibitor model.

A formal presentation of the IDG model and the goals of this paper are given in the next section.

Notations: The Bernoulli distribution with parameter p is denoted by $\mathcal{B}(p)$, where p denotes the probability of the Bernoulli random variable taking a value of one. The set of binary numbers is denoted by $\mathbb{B}$. Matrices are indicated by boldface uppercase letters and vectors by boldface lowercase letters. The row-i, column-j entry of a matrix $\mathbf{M}$ is denoted by $\mathbf{M}(i, j)$, and the coordinate-i of a vector $\mathbf{y}$ is denoted by $\mathbf{y}(i)$. All the logarithms in this paper are taken to the base two.

II. THE IDG MODEL

Consider a set of items $\mathcal{W}$ indexed as $w_1, \cdots, w_n$ comprised of r inhibitors, d defectives, and $n - r - d$ normal items. It is assumed throughout the paper that $r, d = o(n)$.

Definition 1: An item pair (w_i, w_j), for $i \neq j$, is said to be *associated* when the inhibitor w_i inhibits the expression of the defective w_j. An item pair (w_i, w_j), for $i \neq j$, is said to be *non-associated* if either the inhibitor w_i does not inhibit the expression of the defective w_j or if w_i is not an inhibitor or if w_j is not a defective.

In general, the mention of an item pair (w_i, w_j) need not mean that w_i is an inhibitor and w_j is a defective. This is understood from the context.

Definition 2: An *association graph* is a left to right directed bipartite graph $\mathcal{B} = (\mathcal{I}, \mathcal{D}, \mathcal{E})$, where the set of vertices (on the left hand side) $\mathcal{I} = \{w_{i_1}, w_{i_2}, \cdots, w_{i_r}\} \subset \mathcal{W}$ denotes the set of inhibitors, the set of vertices (on the right hand side) $\mathcal{D} = \{w_{j_1}, w_{j_2}, \cdots, w_{j_d}\} \subset \mathcal{W}$ denotes the set of defectives, and $\mathcal{E}$ is a collection of directed edges from $\mathcal{I}$ to $\mathcal{D}$. A directed edge $e = (w_i, w_k) \in \mathcal{E}$, for $i \in \{i_1, \cdots, i_r\}, j \in \{j_1, \cdots, j_d\}$, denotes that the inhibitor w_i inhibits the expression of the defective w_k.

We refer to $\mathcal{E}(\mathcal{I}, \mathcal{D})$ *conditioned on the sets* $(\mathcal{I}, \mathcal{D})$ to be the *association pattern* on $(\mathcal{I}, \mathcal{D})$.

A pooling design is denoted by a test matrix $\mathbf{M} \in \mathbb{B}^{T \times n}$, where the jth item appears in the ith test iff $\mathbf{M}(i, j) = 1$. A test outcome is positive iff the test contains at least one defective without any of its associated inhibitors. A positive outcome is denoted by one and a negative outcome by zero.

It is assumed throughout the paper that the defectives are not mutually obscuring, i.e., a defective does not function as an inhibitor for some other defective. In other words, the set of inhibitors $\mathcal{I}$ and the set of defectives $\mathcal{D}$ are disjoint.

The goal of this paper is to identify the association graph, or in informal terms, learn the IDG. Thus, the objectives are two-fold as represented by Fig. 3.

1) Identify all the defectives.

---

[1] The number of inhibitors, defectives and normal items are denoted by r, d, and $n - d - r$ respectively.

2) Identify all the inhibitors and also their association pattern with the defectives.
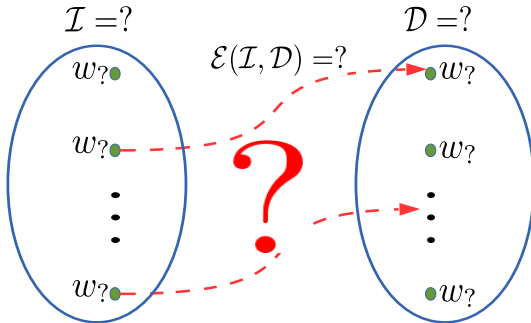

Fig. 3. Here, the presence of a directed arrow represents an association between an inhibitor and a defective. The problem statement is to identify the set of inhibitors $\mathcal{I}$, defectives $\mathcal{D}$ and the association pattern $\mathcal{E}(\mathcal{I}, \mathcal{D})$.

This problem is further mathematically formulated as follows. Denote the actual set of inhibitors, normal items, and defectives by $\mathcal{I}$, $\mathcal{N}$, and $\mathcal{D}$ respectively so that $\mathcal{I} \cup \mathcal{N} \cup \mathcal{D} = \mathcal{W}$. The actual association pattern between the actual inhibitor and defective sets is represented by $\mathcal{E}(\mathcal{I}, \mathcal{D})$. Let $\hat{\mathcal{I}}$, $\hat{\mathcal{N}}$, $\hat{\mathcal{D}}$, and $\hat{\mathcal{E}}(\hat{\mathcal{I}}, \hat{\mathcal{D}})$ denote the declared set of inhibitors, normal items, defectives, and declared association pattern between $(\hat{\mathcal{I}}, \hat{\mathcal{D}})$ respectively. The target is to meet the following error metric.

$$\max_{\mathcal{I}, \mathcal{D}, \mathcal{E}(\mathcal{I}, \mathcal{D})} \Pr \left\{ \left(\hat{\mathcal{I}}, \hat{\mathcal{D}}, \hat{\mathcal{E}} \left(\hat{\mathcal{I}}, \hat{\mathcal{D}} \right) \right) \neq (\mathcal{I}, \mathcal{D}, \mathcal{E}(\mathcal{I}, \mathcal{D})) \right\} \leq cn^{-\delta},$$

(1)

for some $c > 0$ and fixed $\delta > 0$. We propose pooling designs and decoding algorithms, and lower bounds on the number of tests required to satisfy the above error metric. It is assumed that the defective and the inhibitor sets are distributed uniformly across the items, i.e., the probability that any given subset of $r + d$ items constitutes all the defectives and inhibitors is given by $\frac{1}{\binom{n}{d}\binom{n-d}{r}}$. It is also assumed that the association pattern $\mathcal{E}(\mathcal{I}, \mathcal{D})$ is uniformly distributed over all possible association patterns on $(\mathcal{I}, \mathcal{D})$.

We consider two variants of the IDG model. The first being the case where the maximum number of inhibitors that can inhibit any defective, given by I_{max}, is known. We refer to this model as the IDG with side information (IDG-WSI) model. For example, Fig. 2 represents a case where $I_{max} = 1$. While it is known that $I_{max} = 1$, it is unknown which among the items $w_1, \cdots, w_n$ represent which inhibitors and defectives. For a given value of (r, d), not all positive integer values of $I_{max} \leq r$ might be feasible. For instance, if $(r, d) = (3, 2)$, then $I_{max} = 1$ is not feasible because, by definition, each inhibitor is associated with at least one defective. So, in the IDG-WSI model, we assume that the given value of I_{max} is feasible for the (r, d) tuple. In particular, if $(c-1)d < r \leq cd$ for some integer $c \geq 1$, then $I_{max} \geq c$. This immediately follows from the fact that each inhibitor must be associated with at least one defective.

The other variant of the IDG model we consider in this paper is the case where there is no side information about the inhibitor-defective associations, which means that each defective can be inhibited by as many as r inhibitors. We refer to this model as the IDG-No Side Information (IDG-NSI) model. For both the models, the goals (as stated in the beginning of this section) are the same.

The contributions of this paper for the IDG models are summarized below.

- The sample complexity of the number of tests sufficient to recover the association graph while satisfying the error metric (1) using the proposed
  - non-adaptive pooling design is given by $T_{NA} = O\left((r+d)^2 \log n\right)$ and $T_{NA} = O\left((I_{\max} + d)^2 \log n\right)$ tests for the IDG-NSI and IDG-WSI models respectively (Theorem 1, Section III).
  - two-stage adaptive pooling design is given by $T_A = O\left(rd \log n\right)$ and $T_A = O\left(I_{\max} d \log n\right)$ tests for the IDG-NSI and IDG-WSI models respectively (Theorem 2, Section III).

- In Section IV (Theorem 4 and Theorem 5), lower bounds of

$$\max \left\{ \Omega\left((r+d)\log n + rd\right), \Omega\left(\frac{r^2}{\log r} \log n\right) \right\} \text{ and}$$

$$\max \left\{ \Omega\left((r+d)\log n + I_{max}d\right), \Omega\left(\frac{I_{max}^2}{\log I_{max}} \log n\right) \right\}$$

are obtained for non-adaptive pooling designs for the IDG-NSI and IDG-WSI models respectively. The first lower bound in both the models is valid for adaptive pooling designs also. Thus, the upper bounds on the number of tests for the proposed two-stage adaptive pooling design exceed the lower bounds by $\log n$ multiplicative factors in both the models. For the proposed non-adaptive pooling design, the upper bounds exceed the lower bounds by $\log r$ and $\log I_{max}$ multiplicative factors in the $d = O(I_{max})$ and $d = O(r)$ regimes in the IDG-NSI and IDG-WSI models respectively.

Extension of the results on the upper and lower bounds on the number of tests to the case where only upper bounds on the number of inhibitors (given by R) and defectives (given by D) are known instead of their exact numbers is straightforward. The target error metric in (1) is re-formulated as maximum error probability criterion over all combinations of number of inhibitors and defectives. The results for this case follow by replacing r by R and d by D in the upper and lower bounds on the number of tests.

There are various generalizations of the 1-inhibitor model considered in the literature. These models are summarized in the following sub-section to show that the model considered in this paper, to the best of our knowledge, has not been studied in the literature.

A. Prior Works

The 1-inhibitor model can be generalized in various directions, mostly influenced by generalizations of the classical

4

group testing model. The various generalizations are listed below and briefly described. Though none of these generalizations include the model studied in this paper, it is worthwhile to understand the differences between these models and the IDG model.

A generalization of the 1-inhibitor model, namely k-inhibitor model was introduced in [14]. In the k-inhibitor model, an outcome is positive iff a test contains at least one defective and no more than $k-1$ inhibitors. So, the number of inhibitors must be no less than a certain threshold k to cancel the effect of any defective. This model is different from the model introduced in this paper because, in the IDG model, a single associated inhibitor is enough to cancel the effect of a defective. Further, none of the inhibitors might be able to cancel the effect of a defective because the defective might not be associated with any inhibitor. A model loosely related with the 1-inhibitor model, namely mutually obscuring defectives model was introduced in [15]. Here, it was assumed that multiple defectives could cancel the effect of each other, and hence the outcome of a test containing multiple defectives could be negative. Thus, a defective can also function as a inhibitor. However, in this paper, the sets of defectives and inhibitors are assumed to be disjoint. The threshold (classical) group testing model is where a test outcome is positive if the test contains at least u defectives, negative if it contains no more than l defectives and arbitrarily positive or negative otherwise [16]. This model was combined with the k-inhibitor model and non-adaptive pooling designs for the resulting model was proposed in [17].

A non-adaptive pooling design for the general inhibitor model was proposed in [18]. Here, the goal was to identify all the defectives with no prior assumption on the cancellation effect of the inhibitors on the defectives, i.e, the underlying unknown inhibitor model could be a 1-inhibitor, k-inhibitor model, or even the ID model introduced in this paper. However, the difference from our work is that, we aim to identify the association graph or, in other words, the cancellation effect of the inhibitors also apart from identification of the defectives. But this cancellation effect does not include the k-inhibitor model cancellation effect as noted earlier. Group testing on complex model was introduced in [19]. In the complex model, a test outcome is positive iff the test contains at least one of the defective sets. So, here the notion of defectives items is generalized to sets of defective items called defective sets. This complex model was combined with the general inhibitor model and non-adaptive pooling designs for identification of defectives was proposed in [20]. Our work is different from [20] for the same reasons as stated for [18]. Group testing on bipartite graphs was proposed in [21] as a special case of the complex model. Here, the left hand side of the bipartite graph represents the bait proteins and the right hand side represents the prey proteins. It is known a priori which items are baits and which ones are preys. The edges in the bipartite graph represent associations between the baits and preys. A test outcome is positive iff the test contains associated items and the goal was to identify these associations. Clearly, this model is different from the IDG because, in the IDG model, there are three types of items involved and the interactions between

the three types of items are different from that in [21].

In the next section, we propose a probabilistic non-adaptive and a probabilistic two-stage adaptive pooling design and decoding algorithms for both the variants of the IDG model.

III. POOLING DESIGNS AND DECODING ALGORITHM

In this section, we propose a non-adaptive pooling design and decoding algorithm as well as a two-stage adaptive pooling design and decoding algorithm for the IDG-WSI Model. The pooling designs and decoding algorithms for the IDG-NSI model follows from those for the the IDG-WSI Model by replacing I_{max} by r.

Non-adaptive pooling design: The pools are generated from the matrix $\mathbf{M}_{NA} \in \mathbb{B}^{T_{NA} \times n}$. The entries of $\mathbf{M}_{NA}$ are i.i.d. as $\mathcal{B}(p_1)$. Test the pools denoted by the rows of $\mathbf{M}_{NA}$. Let the outcome vector be given by $\mathbf{y} \in \mathbb{B}^{T_{NA} \times 1}$. The exact value of T_{NA} is specified in (11) and (12) (where, $T_{NA} = \beta_{NA} \log n$), and its scaling is given in Theorem 1. The exact value of p_1 is also given in Theorem 1.

Adaptive pooling design: A set of pools are generated from the matrix $\mathbf{M}_1 \in \mathbb{B}^{T_1 \times n}$ whose entries are i.i.d. as $\mathcal{B}(p_1)$. The pools denoted by the rows of $\mathbf{M}_1$ are tested first and all the defectives are classified from the outcome vector $\mathbf{y}_1 \in \mathbb{B}^{T_1 \times 1}$. Denote the number of items declared defectives by $\hat{d}$ and the set of declared defectives by $\{\hat{u}_1, \hat{u}_2, \cdots, \hat{u}_{\hat{d}}\}$. If $\hat{d} \neq d$, an error is declared. We keep these declared defectives aside and generate another pooling matrix $\mathbf{M}_2 \in \mathbb{B}^{T_2 \times (n-d)}$, whose entries are i.i.d. as $\mathcal{B}(p_2)$, for the rest of the items. Now, test the pools denoted by the rows of the matrix $\mathbf{M}_2$ along with each of the items declared defectives and the outcomes are denoted by $\mathbf{y}_{\hat{u}_1}, \mathbf{y}_{\hat{u}_2}, \cdots, \mathbf{y}_{\hat{u}_d} \in \mathbb{B}^{T_2 \times 1}$. The two stages of testing are done non-adaptively, and hence the pooling scheme is a two-stage adaptive pooling design. The exact values of p_1 and p_2 are given in Theorem 2. The scaling of T_1 and T_2 are also given in Theorem 2 and their exact values are given in (11) and (13) (where, $T_i = \beta_i \log n$). The total number of tests is given by $T_1 + dT_2$.

The defectives are expected to participate in a higher fraction of positive outcome tests than the normal items or the inhibitors. And, once the defectives are identified, tests of each one of them with rest of the items can be used to determine their associations. We show that this can be done non-adaptively as well. The decoding algorithm proceeds in two stages for both the non-adaptive as well as the adaptive pooling design. The first stage will identify the defectives from the outcome vectors $\mathbf{y}$ and $\mathbf{y}_1$ in the non-adaptive and adaptive pooling designs respectively, according to the fraction of positive outcome tests in which an item participates. The second stage will identify the inhibitors and their associations with the declared defectives using subsets of the outcome vector $\mathbf{y}$ in the non-adaptive pooling design and the outcome vectors $\mathbf{y}_{\hat{u}_1}, \mathbf{y}_{\hat{u}_2}, \cdots, \mathbf{y}_{\hat{u}_d}$ in the adaptive pooling design.

Let us define the following notations[2] with respect to the pools represented by $\mathbf{M}_{NA}$ and $\mathbf{M}_1$ which are eventually useful in characterizing the statistics of the different types of

---

[2]From hereon, we reserve the notation u to represent a defective, v to represent a normal item and s to represent an inhibitor.

items that are used in the decoding algorithm.

Notations:

- $\mathcal{I}(u)$ denotes the set of inhibitors that the defective u is associated with.
- $\mathscr{F}_{u_k}$ denotes the event that none of the inhibitors associated with a defective u_k appears in a test, given that the defective u_k appears in the test.
- $\mathcal{D}_i^{(j)} \subseteq \mathcal{P}(\{u_1, \cdots, u_d\})$ denotes the jth-set in the (arbitrarily) ordered set of all i-tuple subsets of the defective set denoted by $\mathcal{D}_i$, for $j = 1, \cdots, \binom{d}{i}$, where u_i denotes a defective and $\mathcal{P}\{(u_1, \cdots, u_d)\}$ denotes the power set of the set of defectives.
- $\mathcal{D}(s)$ denotes the defectives associated with the inhibitor s and its complement is given by $\overline{\mathcal{D}(s)} = \mathcal{D} - \mathcal{D}(s)$.
- $\overline{\mathcal{D}(s)}_i$ denotes the (arbitrarily) ordered set of all i-tuple subsets of the defective set $\overline{\mathcal{D}(s)}$ and the jth-set in $\overline{\mathcal{D}(s)}_i$ is denoted by $\overline{\mathcal{D}(s)}_i^{(j)}$.

Example 2: Realizations of the above notation for the association graph in Fig. 2 considered in Example 1 are given below. The inhibitor set is given by $\mathcal{I} = \{s_1, \cdots, s_r\}$ and the defective set is given by $\mathcal{D} = \{u_1, \cdots, u_d\}$ with $r = d$.

- $\mathcal{I}(u)$ for $u = u_1$ is given by $\mathcal{I}(u_1) = \{s_1\}$.
- $\mathscr{F}_{u_1}$ represents the event that the inhibitor s_1 associated with the defective u_1 does not appear in a test, given that the defective u_1 appears in the test.
- Realizations of $\mathcal{D}_i$ for $i = 1, 2$ are given by

$$\mathcal{D}_1 = \{\{u_1\}, \{u_2\}, \cdots, \{u_d\}\},$$
$$\mathcal{D}_2 = \{\{u_1, u_2\}, \{u_1, u_3\}, \cdots, \{u_1, u_d\},$$
$$\{u_2, u_3\}, \cdots, \{u_2, u_d\}, \cdots, \{u_{d-1}, u_d\}\}.$$

  Realizations of $\mathcal{D}_i^{(j)}$ for $(i, j) = (1, 2)$ and $(i, j) = (2, 3)$ are given by

$$\mathcal{D}_1^{(2)} = \{u_2\}, \mathcal{D}_2^{(3)} = \{u_1, u_4\}.$$

- $\mathcal{D}(s)$ for $s = s_1$ is given by $\mathcal{D}(s_1) = u_1$ and its complement is given by $\overline{\mathcal{D}(s_1)} = \{u_2, \cdots, u_d\}$.
- Realizations of $\overline{\mathcal{D}(s)}_i$ for $s = s_1$ and $i = 1, 2$

$$\overline{\mathcal{D}(s_1)}_1 = \{\{u_2\}, \cdots, \{u_d\}\},$$
$$\overline{\mathcal{D}(s_1)}_2 = \{\{u_2, u_3\}, \{u_2, u_4\}, \cdots, \{u_2, u_d\},$$
$$\{u_3, u_4\}, \cdots, \{u_3, u_d\}, \cdots, \{u_{d-1}, u_d\}\}.$$

  Realizations of $\overline{\mathcal{D}(s)}_i^{(j)}$ with $s = s_1$, for $(i, j) = (1, 2)$ and $(i, j) = (2, 3)$ are given by

$$\overline{\mathcal{D}(s)}_1^{(2)} = \{u_3\}, \overline{\mathcal{D}(s)}_2^{(3)} = \{u_2, u_5\}.$$

We now define the following statistics corresponding to the different types of items. The following statistics also hold good when $\mathbf{y}_1$ is replaced by $\mathbf{y}$, as entries of both $\mathbf{M}_{NA}$ and $\mathbf{M}_1$ have the same statistics.

$$q_1^{(u)} \triangleq \Pr\left\{\mathbf{y}_1(l) = 1 | \text{defective } u \text{ is present in the } l^{\text{th}}\text{-test}\right\}$$
$$\geq (1 - p_1)^{|\mathcal{I}(u)|} \geq (1 - p_1)^{I_{max}}, \qquad (2)$$

$$q_2^{(v)} \triangleq \Pr\left\{\mathbf{y}_1(l) = 1 | \text{normal item } v \text{ is present in the } l^{\text{th}}\text{-test}\right\}$$

$$= \sum_{i=1}^{d} p_1^i (1 - p_1)^{d-i} \sum_{j=1}^{\binom{d}{i}} \Pr\left\{\bigcup_{u_k \in \mathcal{D}_i^{(j)}} \mathscr{F}_{u_k}\right\} \triangleq q_2 \qquad (3)$$

$$\leq \sum_{i=1}^{d} p_1^i (1 - p_1)^{d-i} \binom{d}{i} \triangleq q_2^{UB}, \qquad (4)$$

$$q_3^{(s)} \triangleq \Pr\left\{\mathbf{y}_1(l) = 1 | \text{Inhibitor } s \text{ is present in the } l^{\text{th}}\text{-test}\right\}$$

$$= \sum_{i=1}^{|\overline{\mathcal{D}(s)}|} p_1^i (1 - p_1)^{|\overline{\mathcal{D}(s)}|-i} \sum_{j=1}^{|\overline{\mathcal{D}(s)}_i|} \Pr\left\{\bigcup_{u_k \in \overline{\mathcal{D}(s)}_i^{(j)}} \mathscr{F}_{u_k}\right\},$$

$$\qquad (5)$$

$$\text{if } \left|\overline{\mathcal{D}(s)}\right| \geq 1,$$
$$= 0, \text{ otherwise.}$$

Since the outer and inner summations in (5) is over a subset of those in (3), $\max_s q_3^{(s)} \leq q_2^{(v)} = q_2$. The upper bound in (4) follows from the upper bound of one on the probability terms of (3). In hindsight, the lower and the upper bounds can be easily obtained as follows. The lower bound on the positive outcome statistics for a defective item in (2) follows from the worst case statistics when all the inhibitors inhibit the expression of every defective. The upper bound on the statistics for a normal item in (4) follows from using the best case positive outcome statistics, in the absence of inhibitors, where the appearance of any defective gives a positive outcome. It is easy to see that positive outcome for an inhibitor in a test is less probable than that for a normal item. Denote the worst case negative outcome statistic for a defective by

$$b_{max} = 1 - (1 - p_1)^{I_{max}}. \qquad (6)$$

Denote the set of tests corresponding to outcome vector $\mathbf{y}$ in which an item w_j participates by $\mathcal{T}_{w_j}(\mathbf{y})$ and the set of positive outcome tests in which the item w_j participates by $\mathcal{S}_{w_j}(\mathbf{y})$, for $j = 1, 2, \cdots, n$. The decoding algorithm is given as follows.

1) *Stage 1 (Identifying the defectives for both non-adaptive and adaptive pooling designs):*
   For the non-adaptive pooling design, if $|\mathcal{S}_{w_j}(\mathbf{y})| > |\mathcal{T}_{w_j}(\mathbf{y})|[1 - b_{max}(1 + \tau))]$ with b_{max} as defined in (6), declare the item w_j to be a defective. For the adaptive pooling design, we use the same criterion, replacing $\mathbf{y}$ by $\mathbf{y}_1$. Denote the number of items declared as defectives by $\hat{d}$ and the set of declared defectives by $\{\hat{u}_1, \hat{u}_2, \cdots, \hat{u}_{\hat{d}}\}$. If $\hat{d} \neq d$, declare an error. Denote the remaining unclassified items in the population by $\{w'_1, \cdots, w'_{n-d}\} \triangleq \{w_1, \cdots, w_n\} - \{\hat{u}_1, \cdots, \hat{u}_d\}$.
2) *Stage 2 (Identifying the inhibitors and their associations for non-adaptive pooling design):*
   Let $\overline{\mathcal{P}_k}$ denote the sets of pools in $\mathbf{M}_{NA}$ that contain

6

only the declared defective $\hat{u}_k$ and none of the other declared defectives, for $k = 1, \cdots, d$. Also, let the outcomes corresponding to these pools be positive. This means that the pools in $\mathcal{P}_k$ do not contain any inhibitor from the set $\mathcal{I}(\hat{u}_k)$, which denotes the set of inhibitors associated with the item $\hat{u}_k$ if $\hat{u}_k$ is indeed a defective. Now, consider only the outcomes corresponding to these pools denoted by $\mathbf{y}_{\mathcal{P}_1} \subset \mathbf{y}, \cdots, \mathbf{y}_{\mathcal{P}_d} \subset \mathbf{y}$. The associations of the declared defectives are identified as follows.

- For each $k = 1$ to d, declare $(w'_j, \hat{u}_k)$ to be a non-associated inhibitor-defective pair if w'_j participates in at least one of the tests corresponding to the outcome vector $\mathbf{y}_{\mathcal{P}_k}$ and declare the rest of the items to be associated with $\hat{u}_k$.

The items declared as non-associated for all k are declared to be be normal items. If $\mathcal{P}_k = \{\emptyset\}$ for some k, declare an error.

3) *Stage 2 (Identifying the inhibitors and their associations for adaptive pooling design):*
Let $\mathcal{S}(\mathbf{y}_{\hat{u}_k})$ denote the set of positive outcome tests corresponding to $\mathbf{y}_{\hat{u}_k}$, i.e., these pools do not contain any inhibitor from the set $\mathcal{I}(\hat{u}_k)$ if $\hat{u}_k$ is a defective.

- For each $k = 1$ to d, declare $(w'_j, \hat{u}_k)$ to be a non-associated inhibitor-defective pair if w'_j participates in at least one of the tests in the set $\mathcal{S}(\mathbf{y}_{\hat{u}_k})$ and declare the rest of the items to be associated with $\hat{u}_k$.

The items declared as non-associated for all k are declared to be be normal items. If $\mathcal{S}(\mathbf{y}_{\hat{u}_k}) = \{\emptyset\}$ for some k, declare an error.

Remark 1: (Stage 1) The first stage in the decoding algorithm, which is the same for both the non-adaptive and adaptive pooling design, is similar to the defective classification algorithm used in [13] for the 1-inhibitor model. The underlying common principle used is that there exists statistical difference between the defective items and the rest of the items. Hence, with sufficient number of tests, the defectives can be classified by "matching" the tests in which an item participates and the positive outcome tests. The items involved in a large fraction of positive outcome tests are declared to be defectives. A similar decoding algorithm was used in the classical group testing framework with noisy tests [22]. Here, the inhibitors of a defective item, if any, behave like a noise due to probabilistic presence in a test. The (worst case) expected number of positive outcome tests in which a defective participates is at least $|\mathcal{T}_{w_j}(\mathbf{y})|[1 - b_{max}]$. Like in [13], the Chernoff-Hoeffding concentration inequality [23] is used to bound the error probability and obtained the exact number of tests required to achieve a target (vanishing) error probability. *It is important to note that, a priori, it is not clear if a fixed threshold technique can sieve the defectives under worst case positive outcome statistics and the rest of the items under best case positive outcome statistics, with vanishing error probability..*

Remark 2: (Stage 2) In the IDG model, the inhibitors for each defective might be distinct. Hence, an inhibitor for one defective behaves as a normal item from the perspective of

another defective. This defective-specific interaction is absent in the 1-inhibitor model. So, any inhibitor can be identified using any defective, i.e, an inhibitor's behaviour is defective-invariant in the 1-inhibitor model, which was exploited in identifying the inhibitors in [13]. Since each inhibitor's behaviour can be defective-specific in the IDG model, we need to identify the defectives first and then identify its associated inhibitors by observing the interaction of the other items with each of these defectives.

The following theorems state the values of the parameters p_1, p_2, and τ, and the scaling of the number of tests required for the proposed non-adaptive and adaptive pooling designs to determine the association graph with high probability. Similar results can be stated for the IDG-NSI model, by replacing I_{max} by r in the following theorems.

Theorem 1 (Non-adaptive pooling design): Choose the pooling design matrix $\mathbf{M}_{NA}$ of size $T_{NA} \times n$ with its entries distributed as i.i.d. $\mathcal{B}(p_1)$ with $p_1 = \frac{1}{2(I_{max}+d+1)}$ for the IDG-WSI model. Test the pools denoted by the rows of the matrix $\mathbf{M}_{NA}$ non-adaptively. The scaling of the number of tests sufficient to guarantee vanishing error probability (1) using the proposed decoding algorithm with $\tau = \frac{1-b_{max}-q_2^{UB}}{2b_{max}}$ is given by $T_{NA} = O\left((I_{max} + d)^2 \log n\right)$, where q_2^{UB} and b_{max} are defined in (4) and (6) respectively.

Theorem 2 (Adaptive pooling design): Choose the pooling design matrices $\mathbf{M}_1$ and $\mathbf{M}_2$ of sizes $T_1 \times n$ and $T_2 \times n$ with their entries distributed as i.i.d. $\mathcal{B}(p_1)$ and $\mathcal{B}(p_2)$ respectively, with $p_1 = \frac{1}{2(I_{max}+d+1)}$ and $p_2 = \frac{1}{2I_{max}}$ for the IDG-WSI model. Test the pools denoted by the rows of the matrices $\mathbf{M}_1$ non-adaptively and classify the defectives. Now, test each of the pools from $\mathbf{M}_2$ along with the d classified defectives individually. The scaling of the number of tests sufficient to guarantee vanishing error probability (1) using the proposed decoding algorithm with $\tau = \frac{1-b_{max}-q_2^{UB}}{2b_{max}}$ is given by $T_A = T_1 + dT_2 = O\left(I_{max}d \log n\right)$, where q_2^{UB} and b_{max} are defined in (4) and (6) respectively.

The following sub-section constitutes the proof of the above theorems. The exact number of tests required to guarantee vanishing error probability for recovery of the association graph are also obtained. The proof is exactly the same for the IDG-NSI model, but replacing I_{max} by r.

A. Error Analysis of the Proposed Algorithm

As mentioned in Section II, we require that

$$\max_{\mathcal{I},\mathcal{D},\mathcal{E}(\mathcal{I},\mathcal{D})} \Pr\left\{(\mathcal{I},\mathcal{D},\mathcal{E}(\mathcal{I},\mathcal{D})) \neq \left(\hat{\mathcal{I}},\hat{\mathcal{D}},\hat{\mathcal{E}}(\hat{\mathcal{I}},\hat{\mathcal{D}})\right)\right\} \leq cn^{-\delta},$$

for some $c > 0$ and fixed $\delta > 0$. We find the number of tests T_1 required to upper bound the first stage error probability by $c_1 n^{-\delta_1}$ and also the value of T_2 required to upper bound the second stage error probability by $c_2 n^{-\delta_2}$, for some constants c_1 and c_2. Finally, the values of δ_1 and δ_2 are chosen so that the total error probability is upper bounded by $cn^{-\delta}$, for some constant c and given $\delta > 0$.

1) Error Analysis of the First Stage: Since the first stage of the decoding algorithm is the same for both the non-adaptive and adaptive pooling design, the bounds on the number of

tests obtained below for adaptive pooling design applies for the non-adaptive pooling design also. The three possible error events in the first stage of the decoding algorithm for both non-adaptive and adaptive pooling design are given by

1) A defective is not declared as one.
2) A normal item is declared as a defective.
3) An inhibitor is declared as a defective.

Clearly, the defective that has the largest probability of a negative outcome, given by $b_{1_{max}} = \max_u \left(1 - q_1^{(u)}\right)$, has the largest probability of not being declared as a defective. So, with $T_1 = \beta_1 \log n$, the probability of the first error event for all the defectives can be upper bounded as

$$d \sum_{t=0}^{T_1} \binom{T_1}{t} p_1^t (1-p_1)^{T_1-t} \sum_{v=tb_{max}(1+\tau)}^{t} \binom{t}{v} b_{1_{max}}^v (1-b_{1_{max}})^{t-v}$$

$$= d \sum_{t=0}^{T_1} \binom{T_1}{t} p_1^t (1-p_1)^{T_1-t} \times$$

$$\sum_{v=tb_{1_{max}}+t(b_{max}-b_{1_{max}}+b_{max}\tau)}^{t} \binom{t}{v} b_{1_{max}}^v (1-b_{1_{max}})^{t-v}$$

$$\overset{(a)}{\leq} d \sum_{t=0}^{T_1} \binom{T_1}{t} p_1^t e^{-2t(b_{max}-b_{1_{max}}+b_{max}\tau)^2} (1-p_1)^{T_1-t}$$

$$\overset{(b)}{=} d \left[1 - p_1 + p_1 e^{-2(b_{max}-b_{1_{max}}+b_{max}\tau)^2}\right]^{\beta_1 \log n}$$

$$\overset{(c)}{\leq} d \exp\left\{-\beta_1 p_1 \log n \left(1 - e^{-2(b_{max}-b_{1_{max}}+b_{max}\tau)^2}\right)\right\} \leq n^{-\delta_1}$$

$$\overset{(d)}{\leq} d \exp\left\{-\beta_1 p_1 \log n \left(1 - e^{-2}\right) (b_{max}-b_{1_{max}}+b_{max}\tau)^2\right\}$$

$$\leq n^{-\delta_1}$$

$$\Rightarrow \beta_1 \geq \frac{\left(\frac{\ln d}{\ln n} + \delta_1\right) \ln 2}{p_1 (1-e^{-2})(b_{max}-b_{1_{max}}+b_{max}\tau)^2},$$

where (a) follows from Chernoff-Hoeffding bound [23][3], (b) follows from binomial expansion, (c) follows from the fact that $1-c \leq e^{-c}$, and (d) follows from the fact that $\left(1 - e^{-2x^2}\right) \geq \left(1 - e^{-2}\right) x^2$, for $0 < x < 1$. Using the fact that $b_{1_{max}} \leq b_{max}$, where b_{max} is defined in (6), the following bound on β_1 suffices.

$$\beta_1 \geq \frac{\left(\frac{\ln d}{\ln n} + \delta_1\right) \ln 2}{p_1 (1-e^{-2})(b_{max}\tau)^2}. \quad (7)$$

Similarly, to guarantee vanishing probability for the second error event (union-bounded over all normal items) and the third error event (union-bounded over all inhibitors), it suffices that

$$\beta_1 \geq \frac{\left(\frac{\ln(n-d-r)}{\ln n} + \delta_1\right) \ln 2}{p_1 (1-e^{-2}) \left(1 - b_{max}(1+\tau) - q_2\right)^2},$$

$$\beta_1 \geq \frac{\left(\frac{\ln r}{\ln n} + \delta_1\right) \ln 2}{p_1 (1-e^{-2}) \left(1 - b_{max}(1+\tau) - q_3^{(s)}\right)^2}, \quad (8)$$

Since $\max_s q_3^{(s)} \leq q_2$ and $r = o(n)$, the bound in (8) is asymptotically redundant for all values of τ. So, substituting

---

[3]If the term $(b_{max} - b_{1_{max}} + b_{max}\tau) > 1$, then the probability of the error event under consideration is equal to zero.

the upper bound on q_2 defined in (4) by q_2^{UB}, it suffices that

$$\beta_1 \geq \frac{\left(\frac{\ln(n-d-r)}{\ln n} + \delta_1\right) \ln 2}{p_1 (1-e^{-2}) \left(1 - b_{max}(1+\tau) - q_2^{UB}\right)^2} \quad (9)$$

Now, the value of τ chosen to optimize the denominators of (7) and (9) is given by $\tau = \frac{1 - b_{max} - q_2^{(UB)}}{2b_{max}}$. Therefore, we have

$$\beta_1 \geq \max\left\{ \frac{4\left(\frac{\ln d}{\ln n} + \delta_1\right) \ln 2}{p_1 (1-e^{-2}) \left(1 - b_{max} - q_2^{UB}\right)^2}, \quad (10) \right.$$

$$\left. \frac{4\left(\frac{\ln(n-d-r)}{\ln n} + \delta_1\right) \ln 2}{p_1 (1-e^{-2}) \left(1 - b_{max} - q_2^{UB}\right)^2} \right\}.$$

The term $1 - b_{max} - q_2$ can be lower bounded as follows.

$$1 - b_{max} - q_2 \geq (1-p_1)^{I_{max}} - \sum_{i=1}^{d} p_1^i (1-p_1)^{d-i} \binom{d}{i}$$

$$= (1-p_1)^{I_{max}} - \sum_{i=0}^{d-1} p_1^{i+1} (1-p_1)^{d-i-1} \binom{d}{i+1}$$

$$= (1-p_1)^{I_{max}} \sum_{i=0}^{d} \binom{d}{i} p_1^i (1-p_1)^{d-i}$$

$$- \sum_{i=0}^{d-1} p_1^{i+1} (1-p_1)^{d-i-1} \binom{d}{i+1}$$

$$\geq \sum_{i=0}^{d-1} \frac{d!}{i!(d-i-1)!} p_1^i (1-p_1)^{d-i-1}$$

$$\times \left(\frac{(1-p_1)^{I_{max}+1}}{d-i} - \frac{p_1}{i+1}\right)$$

The above term can be non-trivially lower bounded if the last multiplicative term is positive for all i. So, it suffices if it is positive for $i = 0$. Using the lower bound $(1-p_1)^{I_{max}+1} \geq 1 - (I_{max} + 1)\, p_1$, the above term at $i = 0$ is lower bounded as

$$(1-p_1)^{d-1} (1 - (I_{max} + d + 1)\, p_1)$$

$$\geq (1 - (d-1)p_1)(1 - (I_{max} + d + 1)\, p_1)$$

$$\geq \frac{1}{4},$$

where the last lower bound follows by substituting $p_1 = \frac{1}{2(I_{max}+d+1)}$. Hence, from (10) and using $r, d = o(n)$, it suffices that

$$\beta_{NA}, \beta_1 \geq \frac{128 \left(I_{max} + d + 1\right) \left(\frac{\ln(n-d-r)}{\ln n} + \delta_1\right) \ln 2}{(1-e^{-2})}, \quad (11)$$

where $T_{NA} = \beta_{NA} \log n$.

2) Error Analysis of the Second Stage: In the error analysis of the second stage, we assume that all the defectives have been correctly declared. Errors due to error propagation from the first stage shall be analyzed later.

Non-adaptive pooling design:
The only error event for the non-adaptive pooling design in the second stage is that there does not exist a set of pools $\mathcal{P}_k$ such that they contain only the defective u_k and none of

its associated inhibitors $\mathcal{I}(u_k)$, and all its non-associated items appear in at least one of such pools. Denote this error event by $\mathscr{U}(u_k)$. Clearly, none of the inhibitors associated with u_k will be declared as non-associated with u_k. This follows from the definition of the set of pools $\mathcal{P}_k$ and the decoding algorithm.

The probability that a non-associated item appears along with a defective u_k, but none of its associated inhibitors and none of the other defectives appear in a pool from $\mathbf{M}_{NA}$ is given by

$$b^{(u_k)} \triangleq p_1^2 (1 - p_1)^{|\mathcal{I}(u_k)|}(1 - p_1)^{d-1}.$$

Now, probability of the error event is upper bounded by

$$\Pr\{\mathscr{U}(u_k)\} \leq (n - d - |\mathcal{I}(u_k)|)\left(1 - b^{(u_k)}\right)^{T_{NA}}$$
$$\leq (n - d - |\mathcal{I}(u_k)|)e^{-T_{NA}b^{(u_k)}} \leq n^{-\delta_2}, \text{ if}$$
$$\beta_{NA} \geq \frac{\left(\frac{\ln(n-d-|\mathcal{I}(u_k)|)}{\ln n} + \delta_2\right)\ln 2}{b^{(u_k)}}.$$

Since $(1 - p_1)^{|\mathcal{I}(u_k)|} \geq (1 - p_1)^{I_{max}} \geq (1 - I_{max}p_1)$ and $(1 - p_1)^{d-1} \geq (1 - dp_1)$, substituting for p_1, it suffices that

$$\beta_{NA} \geq 16(I_{max} + d + 1)^2\left(\frac{\ln(n-d)}{\ln n} + \delta_2\right)\ln 2. \quad (12)$$

Adaptive pooling design:
Like in non-adaptive pooling design the only error event, denoted by $\mathscr{E}(u_k)$, is that items w_j not associated with u_k are declared as associated inhibitors, i.e., the item w_k does not appear in any of the positive outcome tests $\mathcal{S}(\mathbf{y}_{u_k})$. Clearly, none of the inhibitors associated with u_k will be declared as non-associated with u_k.

Let $T_2 = \beta_2 \log n$. The number of tests required to guarantee vanishing error probability for the error event $\mathscr{E}(u_k)$ is evaluated as follows. Let $w_j \notin \mathcal{I}(u_k)$. Define

$$a_{w_j}^{(u_k)} \triangleq \Pr\left\{\mathbf{y}_{u_k}(l) = 1 \mid w_j \text{ is present in } l^{\text{th}}\text{-test}\right\}$$
$$\geq (1 - p_2)^{|\mathcal{I}(u_k)|} \triangleq a^{(u_k)}.$$

Now, we have

$$\Pr\{\mathscr{E}(u_k)\} \leq (n - d - |\mathcal{I}(u_k)|)\left(1 - a^{(u_k)}p_2\right)^{T_2} \leq n^{-\delta_2}$$
$$\Leftarrow \beta_2 \geq \frac{\left(\frac{\ln(n-d-|\mathcal{I}(u_k)|)}{\ln n} + \delta_2\right)\ln 2}{p_2\left(1 - p_2\right)^{|\mathcal{I}(u_k)|}}.$$

Using the fact that $(1 - p_2)^{|\mathcal{I}(u_k)|} \geq 1 - |\mathcal{I}(u_k)|p_2$, and substituting $p_2 = \frac{1}{2I_{max}}$, we have the following bound.

$$\beta_2 \geq 4I_{max}\left(\frac{\ln(n - d - |\mathcal{I}(u_k)|)}{\ln n} + \delta_2\right)\ln 2$$
$$\Leftarrow \beta_2 \geq 4I_{max}\left(\frac{\ln(n-d)}{\ln n} + \delta_2\right)\ln 2. \quad (13)$$

3) Analysis of Total Error Probability: Assuming that the target total error probability is $O(n^{-\delta})$, the values of δ_1 and δ_2 need to be determined. Towards that end, define the following events.

$\mathscr{E}_{ij} \triangleq$ Event of declaring $(w_i, w_j), i \neq j$, to be an associated pair,

$\mathscr{W} \triangleq$ Event that at least one actual defective has not been declared as a defective.

Let $\mathcal{E}$ denote the correct association pattern for some realization $\{\mathcal{I}, \mathcal{D}\}$. Now, the total probability of error is given by

$$\Pr\left\{\bigcup_{(w_i, w_j) \notin \mathcal{E}} \mathscr{E}_{ij} \bigcup \mathscr{W}\right\} \leq \sum_{(w_i, w_j) \notin \mathcal{E}} \Pr\{\mathscr{E}_{ij}\} + \Pr\{\mathscr{W}\}$$
$$\leq \sum_{w_i \neq w_j} \sum_{w_j \in \mathcal{N} \cup \mathcal{I}} \Pr\{\mathscr{E}_{ij}\} + \sum_{w_i \notin \mathcal{I}(w_j)} \sum_{w_j \in \mathcal{D}} \Pr\{\mathscr{E}_{ij}\} \quad (14)$$
$$+ \Pr\{\mathscr{W}\}$$
$$< n \times 2n^{-\delta_1} + dn^{-\delta_2} + n^{-\delta_1}. \quad (15)$$

There are two possible ways in which the event $\mathscr{E}_{ij}$, for $(w_i, w_j) \notin \mathcal{E}$, can occur. One possibility is that the item w_j has been erroneously declared as a defective in the first stage of the algorithm, and hence any item w_i declared to be associated with w_j is an erroneous association. The first term in (14) represents this possibility. The other possibility is that w_j has been correctly identified as a defective, but the item w_i is erroneously declared to be associated with w_j. The second term in (14) represents this possibility. The last term accounts for the fact that a defective might be missed out in the first stage of the algorithm. Note that the other two terms do not capture this error event. Finally, (15) follows from the error analysis of the first and second stages of the decoding algorithm. Therefore, if the target error probability is $O(n^{-\delta})$, then choose $\delta_1, \delta_2 = \delta + 1$.

Recall that the number of tests required for non-adaptive and adaptive pooling designs are given by $T_{NA} = \beta_{NA}\log n$ and $T_A = T_1 + dT_2 = (\beta_1 + d\beta_2)\log n$ respectively. Therefore, from (11), (12), and (13) we have that $T_{NA} = O\left((I_{max} + d)^2\log n\right)$ and $T_A = O(I_{max}d\log n)$.

B. Adaptation for the IDG-NSI Model

The only modification required in the pooling design and decoding algorithm proposed for the IDG-WSI model to adapt it to the IDG-NSI model is that I_{max} is replaced by r. For the sake of clarity, we list the only changes below.

1) The pooling design parameters are chosen as $p_1 = \frac{1}{2(r+d+1)}$, $p_2 = \frac{1}{2r}$.
2) In Stage-1 of the decoding algorithm the threshold for identifying the defectives is chosen as $|\mathcal{S}_{w_j}(\mathbf{y}_1)| > |\mathcal{T}_{w_j}(\mathbf{y}_1)|[1 - b_{max}(1 + \tau)]$, where $b_{max} = 1 - (1 - p_1)^r$. Intuitively, this worst-case threshold corresponds to a scenario where every inhibitor inhibits every defective, i.e., the 1-inhibitor model.

3) The values of β_{NA}, β_1 and β_2 are chosen as

$$\beta_{NA} \geq \max \left\{ \frac{128 \, (r+d+1) \left(\frac{\ln(n-d-r)}{\ln n} + \delta_1 \right) \ln 2}{(1 - e^{-2})}, \right.$$
$$\left. 16(r+d+1)^2 \left(\frac{\ln(n-d)}{\ln n} + \delta_2 \right) \ln 2 \right\},$$

$$\beta_1 \geq \frac{128 \, (r+d+1) \left(\frac{\ln(n-d-r)}{\ln n} + \delta_1 \right) \ln 2}{(1 - e^{-2})},$$

$$\beta_2 \geq 4r \left(\frac{\ln(n-d)}{\ln n} + \delta_2 \right) \ln 2.$$

Hence, the total number of tests required for the IDG-NSI model scales as $T_{NA} = O\left((r+d)^2 \log n\right)$ for the non-adaptive pooling design and $T_A = O(rd \log n)$ for the two-stage adaptive pooling design.

In the next section, lower bounds on the number of tests for non-adaptive and adaptive pooling designs are obtained.

IV. LOWER BOUNDS FOR NON-ADAPTIVE AND ADAPTIVE POOLING DESIGN

In this section, two lower bounds on the number of tests required for non-adaptive pooling designs for solving the IDG-NSI and IDG-WSI problems with vanishing error probability are obtained. One of the lower bounds is simply obtained by counting the entropy in the system and this lower bound also holds good for adaptive pooling designs. The other lower bound is obtained using a lower bound result for the 1-inhibitor model which is stated below.

Theorem 3 (Th. 1, [13]): An asymptotic lower bound on the number of tests required for non-adaptive pooling designs in order to classify r inhibitors amidst d defectives and $n-d$ normal items in the 1-inhibitor model is given by $\Omega\left(\frac{r^2}{d \log \frac{r}{d}} \log n\right)$, in the $d = o(r), r = o(n)$ regime[4].

Theorem 4: An asymptotic lower bound on the number of tests required for non-adaptive pooling designs for solving the IDG-NSI problem with vanishing error probability for $r, d = o(n)$ is given by

$$\max \left\{ \Omega\left((r+d)\log n + rd\right), \Omega\left(\frac{r^2}{\log r} \log n\right) \right\}.$$

Proof: The proof for the first lower bound on the number of tests follows by lower bounding the number of possible combinations of the sets of defectives, inhibitors and associa-

tion patterns[5].

$$T \geq H\left(\mathcal{I}, \mathcal{D}, \mathcal{E}(\mathcal{I}, \mathcal{D})\right)$$
$$= H(\mathcal{D}) + H\left(\mathcal{I}|\mathcal{D}\right) + H(\mathcal{E}(\mathcal{I}, \mathcal{D})|(\mathcal{I}, \mathcal{D})) \quad (16)$$
$$= \log \left(\binom{n}{d} \binom{n-d}{r} \sum_{i_1=1}^{d} \cdots \sum_{i_r=1}^{d} \prod_{j=1}^{r} \binom{d}{i_j} \right)$$
$$\geq \log \left(\binom{n}{d} \binom{n-d}{r} \prod_{k=1}^{r} \binom{d}{\frac{d}{2}} \right)$$
$$= \Omega\left((r+d)\log n + rd\right), \quad (17)$$

where i_j denotes the number of defectives that the jth inhibitor can be associated with, and the last step follows by using Stirling's lower bound $\binom{d}{\frac{d}{2}} \geq 2^{\frac{d}{2}}$. This lower bound is also valid for *adaptive pooling designs*.

The proof for the second lower bound for non-adaptive pooling designs is as follows. Assume that it's required to identify the inhibitors alone. Clearly, this requires lesser number of tests than the problem of identifying the association graph. Since the objective is to satisfy the error metric in (1), the error probability criterion

$$\Pr\left\{ \left(\hat{\mathcal{I}}, \hat{\mathcal{D}}, \hat{\mathcal{E}}(\hat{\mathcal{I}}, \hat{\mathcal{D}})\right) \neq (\mathcal{I}, \mathcal{D}, \mathcal{E}(\mathcal{I}, \mathcal{D})) \right\} \leq cn^{-\delta}$$

needs to be satisfied for all possible association patterns on all possible realizations of $(\mathcal{I}, \mathcal{D})$. We choose one possible association pattern on some realization of $(\mathcal{I}, \mathcal{D})$ given in Fig. 4 to evaluate the lower bound. Now, assume that a genie
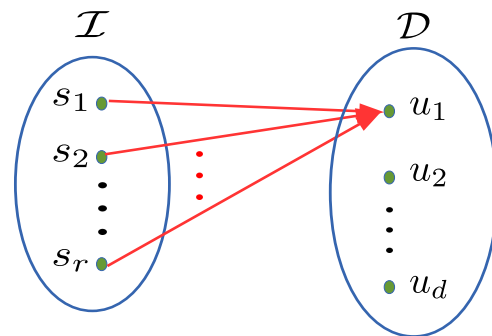

Fig. 4. A possible association pattern for a possible realization of $(\mathcal{I}, \mathcal{D})$ where a single defective is associated with all the inhibitors, but none of the other defectives are associated with any inhibitor.

reveals the set of defectives $\mathcal{D}' = \{u_2, \cdots, u_d\}$ which are not associated with any of the inhibitors. A lower bound for this problem with side information from the genie is clearly a lower bound for the original problem. Note that the presence of any defective from the set $\mathcal{D}'$ in a pool always gives a positive outcome and hence, provides zero information for distinguishing the inhibitors from the rest of the items as the entropy of such outcomes is zero. Hence, the inhibitor identification problem for items with the association pattern as

---

[4]Though Theorem 1 in [13] is stated for the classification of both the defectives and inhibitors in the 1-inhibitor model, it is also valid for classification of inhibitors alone.

[5]The proof for the fact that the entropy of the system is a lower bound on the number of tests for both non-adaptive and adaptive pooling designs can be found in the appendix of [22].

given in Fig. 4 is now reduced to the problem of identifying r inhibitors amidst $n - d$ normal items and one defective item in the 1-inhibitor model, where $d = o(n)$. For this problem, from Theorem 3, we have that the lower bound is given by $T = \Omega\left(\frac{r^2}{\log r}\log n\right)$ tests. Hence, this is also a lower bound on the number of tests required to identify the association graph with vanishing worst case error probability. ∎

The lower bounds for the IDG-WSI model are obtained in the following theorem. Since we are interested in asymptotic lower bounds, we assume that $\lim_{n\to\infty}\frac{I_{max}}{r}$ and $\lim_{n\to\infty}\frac{r}{d}$ exist, and $I_{max}\xrightarrow[n\to\infty]{}\infty$.

Theorem 5: An asymptotic lower bound on the number of tests required for non-adaptive pooling designs for solving the IDG-WSI problem with vanishing error probability for $r, d = o(n)$ is given by

$$\max\left\{\Omega\left((r+d)\log n + I_{max}d\right), \Omega\left(\frac{I_{max}^2}{\log I_{max}}\log n\right)\right\}.$$

Proof: The first lower bound is obtained by lower bounding $H\left(\mathcal{E}(\mathcal{I},\mathcal{D})|(\mathcal{I},\mathcal{D})\right)$ in (16) as follows. Two constraints need to be satisfied while counting the entropy of association pattern.

- *First constraint:* Minimum degree of a vertex in $\mathcal{I}$ is one.
- *Second constraint:* Maximum degree of a vertex in $\mathcal{D}$ is no more than I_{max}.

We now consider the three possible cases below and show that in each of the cases the lower bound on the number of association patterns scales exponentially in $I_{max}d$. Let $(c-1)d < r \le cd$, for some positive integer c, and so $I_{max} \ge c$. Define $\alpha_1 = \lim_{n\to\infty}\frac{c}{I_{max}}$ and $\alpha_2 = \lim_{n\to\infty}\frac{I_{max}}{r}$.

Case 1: $\alpha_1 < 1$ and $\alpha_2 < 1$. There exist positive constants $\beta_1 < 1$ and $\beta_2 < 1$ so that $c \le \beta_1 I_{max}$ and $I_{max} \le \beta_2 r, \forall n \ge n_0$. Define an association pattern, where each defective starting from u_1 is assigned a disjoint set of c inhibitors until every inhibitor is covered. Therefore we have, $\max_{u_i\in\mathcal{D}}|\mathcal{I}(u_i)| \le c$. Since the first constraint is satisfied, each defective is now free to choose an association pattern so that $\max_{u_i\in\mathcal{D}}|\mathcal{I}(u_i)| \le I_{max}$. The number of such possible association patterns can be lower bounded by

$$\binom{r}{I_{max}-c}^d \ge \left(\frac{r}{I_{max}-c}\right)^{(I_{max}-c)d}.$$

Thus, the entropy of association pattern in this case scales (asymptotically) as the logarithm of the above quantity, which is given by $\Omega(I_{max}d)$.

Case 2: $\alpha_1 < 1$ and $\alpha_2 = 1$. There exist positive constants $\beta_1 < 1$ and $\beta_2 \le 1$ with $\beta_2 > \beta_1$ so that $c \le \beta_1 I_{max}$ and $I_{max} \ge \beta_2 r, \forall n \ge n_0$. So, we have $I_{max} - c \ge (\beta_2 - \beta_1)r$, $\forall n \ge n_0$. Using similar arguments as in Case 1, where after satisfying the first constraint, $\beta_2 r - c$ inhibitors are chosen to associate with each defective, we now have that the entropy of association pattern in this case scales asymptotically as $\Omega(rd) = \Omega(I_{max}d)$.

Case 3: $\alpha_1 = 1$. Note that this case constitutes a large inhibitor regime with respect to the number of defectives (because $I_{max} \to \infty$). There exists a positive constant $\beta_1 \le 1$

so that $c \ge \beta_1 I_{max}, \forall n \ge n_0$. The number of ways of assigning each defective to a disjoint set of $(c-1)$ inhibitors is given by

$$\binom{r}{c-1}\binom{r-(c-1)}{(c-1)}\binom{r-2(c-1)}{(c-1)}\cdots\binom{r-(d-1)(c-1)}{(c-1)}$$

$$= \frac{r!}{((c-1)!)^d(r-d(c-1))!}$$

$$\underset{(a)}{\ge} \frac{\sqrt{2\pi}r^{r+\frac{1}{2}}e^{-r}}{e^2(c-1)^{d(c-1+\frac{1}{2})}e^{-d(c-1)}(r-d(c-1))^{r-d(c-1)+\frac{1}{2}}e^{-(r-d(c-1))}}$$

$$\underset{(b)}{\ge} \frac{\sqrt{2\pi}(d(c-1))^{d(c-1)+\frac{1}{2}}e^{-dc}}{e^2(c-1)^{d(c-1+\frac{1}{2})}e^{-d(c-1)}(r-d(c-1))^{r-d(c-1)+\frac{1}{2}}e^{-(r-d(c-1))}}$$

$$\underset{(c)}{\ge} \frac{\sqrt{2\pi}(d(c-1))^{d(c-1)+\frac{1}{2}}e^{-dc}}{e^2(c-1)^{d(c-1+\frac{1}{2})}e^{-d(c-1)}d^{d+\frac{1}{2}}}$$

$$= \frac{\sqrt{2\pi}(c-1)^{\frac{1}{2}}d^{d(c-2)}}{e^2(c-1)^{\frac{d}{2}}e^d} = \frac{\sqrt{2\pi}(c-1)^{\frac{1}{2}}d^{d(c-2)}}{e^2d^{\frac{d\log_d(c-1)}{2}}d^{d\log_d e}}$$

$$= \frac{\sqrt{2\pi}}{e^2}(c-1)^{\frac{1}{2}}d^{d\left(c-\frac{\log_d(c-1)}{2}-\log_d e-2\right)},$$

where (a) follows from Stirling's lower and upper bounds for factorial functions, (b) and (c) follow from the fact that $d(c-1) < r \le cd$. Observe that the remaining $r - d(c-1)$ inhibitors can be assigned one each to one defective without violating the second constraint. Thus, the entropy of association pattern in this case scales asymptotically as $\Omega(cd) = \Omega(\beta_1 I_{max}d) = \Omega(I_{max}d)$.
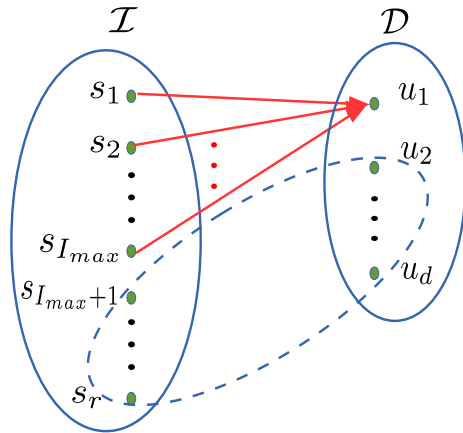

Fig. 5. A possible association pattern where, without loss of generality, u_1 is assumed to be a defective for which $|\mathcal{I}(u_1)| = I_{max}$. The set of inhibitors and defectives that are associated only among themselves are inside the dotted ellipse.

The second lower bound is obtained as shown below. There can exist at least one defective $u_1 \in \mathcal{D}$ so that $|\mathcal{I}(u_1)| = I_{max}$. Consider an association pattern where $\mathcal{I}(u_1) \cap \mathcal{I}(u_k) = \{\emptyset\}$, for $u_k \in \mathcal{D}, k \ne 1$, as depicted in Fig. 5. Now, we use a similar argument as in the proof of the second lower bound in Theorem 4. Let a genie reveal the inhibitor subset $\mathcal{I} - \mathcal{I}(u_1)$, the defective subset $\mathcal{D} - u_1$ and their associations. Now, none of the items from the sets $\mathcal{I} - \mathcal{I}(u_1)$ and $\mathcal{D} - u_1$ is useful in distinguishing the inhibitors in the set $\mathcal{I}(u_1)$ from the unknown defective and the normal items. This is because the entropy of

an outcome is zero if the test contains some defective from $\mathcal{D} - u_1$ but none of its associated inhibitors (which are only from the set $\mathcal{I} - \mathcal{I}(u_1)$) as such a test outcome is always positive. The entropy of an outcome does not change if any of the inhibitors $\mathcal{I} - \mathcal{I}(u_1)$ with or without its associated defectives (which are only from the set $\mathcal{D} - u_1$) is present in the test. Thus, the problem is now reduced to the 1-inhibitor problem of finding I_{max} inhibitors amidst $n - (r - I_{max}) - (d - 1)$ normal items and one (unknown) defective. A lower bound on the number of non-adaptive tests for this problem is clearly a lower bound on the number of tests for the original problem of determining the association graph for the IDG-WSI model. Since $r, d = o(n)$ and $I_{max} \xrightarrow[n \to \infty]{} \infty$, using Theorem 3, we get the lower bound $\Omega\left(\frac{I_{max}^2}{\log I_{max}} \log n\right)$. ∎

Thus, in the $d = O(I_{max})$ and $d = O(r)$ regimes, the upper bound on the number of tests for the proposed non-adaptive pooling design is away from the proposed lower bound for the IDG-WSI and IDG-NSI models by $\log I_{max}$ and $\log r$ multiplicative factors respectively. But in the $I_{max} = o(d)$ and $r = o(d)$ regimes, the upper bounds scale as $O(d^2 \log n)$ which is far from the proposed lower bounds by atmost $\frac{d}{I_{max}} \log n$ and $\frac{d}{r} \log n$ multiplicative factors for the IDG-WSI and IDG-NSI models respectively. For the proposed two-stage adaptive pooling design, the upper bound on the number of tests is away from the proposed lower bound by $\log n$ multiplicative factors for both the IDG-WSI and IDG-NSI models in all regimes of the number of defectives and inhibitors.

V. CONCLUSION

A new generalization of the 1-inhibitor model, termed IDG model was introduced. In the proposed model, an inhibitor can inhibit a non-empty subset of the defective set of items. Probabilistic non-adaptive pooling design and a two-stage adaptive pooling design were proposed and lower bounds on the number of tests were identified. In the large inhibitor regime, the upper bound on the number of tests for the proposed non-adaptive pooling design is shown to be close to the lower bound, with a difference of a logarithmic multiplicative factor. However, in the small inhibitor regime, the upper bound is far away from the proposed lower bound. The non-adaptive pooling design made use of the principle of testing an isolated defective with the rest of the items to determine its associated inhibitors. This is the primary reason for the upper bound to be far away from the lower bound. It is not clear if this is a necessary condition for identifying all the associations. Tightening the lower bound, based on necessity of this condition, is an interesting problem.

For the proposed two-stage adaptive pooling design, the upper bound on the number of tests is close to the lower bound in all regimes of the number of inhibitors and defectives, the difference being logarithmic multiplicative factors. An interesting research direction is to include the k-inhibitor model, for unknown k, as a part of the association pattern in the IDG model, and derive lower and upper bounds on the number of tests for the resulting model. The proposed IDG model with noisy tests is also worth studying.

REFERENCES

[1] M. Xie, K. Tatsuoka, J. Sacks, and S. Young, "Group Testing with Blockers and Synergism," *Journal of the American Statistical Association*, vol. 96, no. 453, pp. 92–102, 2001.

[2] T. Clinton et al., "Design and Characterization of Ebolavirus GP Prehairpin Intermediate Mimics as Drug Targets," *Wiley Online Library on Protein Science*, Oct. 2014.

[3] J. Xing, H. Ly, and Y. Liang, "The Z proteins of pathogenic but not non-pathogenic arenaviruses inhibit the RIG-i-like receptor (RLR)-dependent interferon production," *Journal of Virology*, pp. JVI–03349, 2014.

[4] M. Farach, S. Kannan, E. Knill and S. Muthukrishnan, "Group Testing Problems with Sequences in Experimental Molecular Biology," in *Proc. IEEE Compression and Complexity of Sequences*, pp. 357–367, 1997.

[5] R. Phatarfod and A. Sudbury, "The Use of a Square Array Scheme in Blood Testing," *Wiley Online Library on Statistics in Medicine*, vol. 13, no. 22, pp. 2337–2343, 1994.

[6] S. Langfeldt, J. Oliver, S. Ghosh, and S. Young, "Optimal Group Testing in the Presence of Blockers," *NCSU Institute of Statistics Mimeograph Series*, vol. 2297, no. 5, 1997.

[7] R. Dorfman, "The Detection of Defective Members of Large Populations," *The Annals of Mathematical Statistics*, vol. 14, no. 4, pp. 436–440, 1943.

[8] D. Balding, W. Bruno, D. Torney, and E. Knill, "A Comparative Survey of Non-Adaptive Pooling Designs," in *Genetic mapping and DNA sequencing*, vol. 81, pp. 133–154, Springer, 1996.

[9] A. De Bonis and U. Vaccaro, "Improved Algorithms for Group Testing with Inhibitors," *Elsevier Information Processing Letters*, vol. 67, no. 2, pp. 57–64, 1998.

[10] A. Dyachkov, A. Macula, D. Torney, and P. Vilenkin, "Two Models of Nonadaptive Group Testing for Designing Screening Experiments," in *Proc. 6-th International Workshop on Model Oriented Design and Analysis*, pp. 63–75, 2001.

[11] A. De Bonis, "New Combinatorial Structures with Applications to Efficient Group Testing with Inhibitors," *Springer Journal of Combinatorial Optimization*, vol. 15, no. 1, pp. 77–94, 2008.

[12] H. Chang, H. Chen, and H. Fu, "Identification and Classification Problems on Pooling Designs for Inhibitor Models," *Mary Ann Liebert Journal of Computational Biology*, vol. 17, no. 7, pp. 927–941, 2010.

[13] A. Ganesan, J. Ebrahimi, S. Jaggi, V. Saligrama, "Non-Adaptive Group Testing with Inhibitors," *arXiv:1410.8440 [cs.IT]*.

[14] A. De Bonis and U. Vaccaro, "Constructions of Generalized Superimposed Codes with Applications to Group Testing and Conflict Resolution in Multiple Access Channels," *Elsevier Journal on Theoretical Computer Science*, vol. 306, no. 1, pp. 223–243, 2003.

[15] P. Damaschke, "Randomized Group Testing for Mutually Obscuring Defectives," *Elsevier Information Processing Letters*, vol. 67, no. 3, pp. 131–135, 1998.

[16] P. Damaschke, "Threshold Group Testing," in *General theory of information transfer and combinatorics*, pp. 707–718, Springer, 2006.

[17] Y. He, H. Tian, X. Zhang, Z. Wang, and S. Gao, "Nonadaptive Algorithms for Threshold Group Testing with Inhibitors and Error-Tolerance," *Mary Ann Liebert Journal of Computational Biology*, vol. 19, no. 7, pp. 903–910, 2012.

[18] F. Hwang and F. Chang, "The Identification of Positive Clones in a General Inhibitor Model," *Elsevier Journal of Computer and System Sciences*, vol. 73, no. 7, pp. 1090–1094, 2007.

[19] D. Torney, "Sets Pooling Designs," *Springer-Verlag Annals of Combinatorics*, vol. 3, no. 1, pp. 95–101, 1999.

[20] F. Chang, H. Chang, and F. Hwang, "Pooling Designs for Clone Library Screening in the Inhibitor Complex Model," *Springer Journal of combinatorial optimization*, vol. 22, no. 2, pp. 145–152, 2011.

[21] Y. Li, T. Thai, Z. Liu, W. Wu, "Protein-protein interaction and group testing in bipartite graphs," *Inderscience International Journal of Bioinformatics Research and Applications*, vol. 1, no. 4, pp. 414–419, 2005.

[22] C. Chan, S. Jaggi, V. Saligrama, and S. Agnihotri, "Non-Adaptive Group Testing: Explicit Bounds and Novel Algorithms," *IEEE Transactions on Information Theory*, vol. 60, pp. 3019–3035, May 2014.

[23] W. Hoeffding, "Probability Inequalities for Sums of Bounded Random Variables," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 13–30, 1963.