Single-user tracing and disjointly superimposed codes

Miklós Csűrös and Miklós Ruszinkó

***Abstract*— The zero-error capacity region of r-out of T user multiple access OR channel is investigated. A family $\mathcal{F}$ of subsets of $[n] = \{1, \dots, n\}$ is an *r-single-user-tracing superimposed code* (r-SUT) if there exists such a *single-user-tracing function* $\phi\colon 2^{[n]} \mapsto \mathcal{F}$ that for all $\mathcal{F}' \subseteq \mathcal{F}$ with $1 \le |\mathcal{F}'| \le r$, $\phi(\cup_{A\in\mathcal{F}'} A) \in \mathcal{F}'$. In this paper we introduce the concept of these codes and give bounds on their rate. We also consider disjointly r-superimposed codes.**

***Index Terms*— codes, superimposed codes, group testing, physical mapping**

I. Introduction

SUPPOSE that T users share a common channel. A binary vector of length n is associated to each user. The i^{th} user transmits its vector $\mathbf{x}_i = (x_i^1, x_i^2, \dots, x_i^n)$ $(i = 1, 2, \dots, T)$ if it is active, otherwise not. It is assumed that the transmission is bit and block synchronized. The destination of the messages is a single receiver that observes the bitwise OR vector of the vectors

$$\mathbf{y} = \bigvee_{\forall i \ active} \mathbf{x}_i$$

associated to the active users. Moreover, suppose that at most r users are active simultaneously. In the classical framework of superimposed coding, the receiver has to be able to identify the set of all active users from the output vector $\mathbf{y}$ of the channel. That is, the code must satisfy the property that for all choices of $\mathbf{x}_1, \dots, \mathbf{x}_k$ and $\mathbf{z}_1, \dots, \mathbf{z}_\ell$ of codewords with $1 \le k, \ell \le r$ and $\{\mathbf{x}_1, \dots, \mathbf{x}_k\} \ne \{\mathbf{z}_1, \dots, \mathbf{z}_\ell\}$, we have

$$\bigvee_{i=1}^{k} \mathbf{x}_i \ne \bigvee_{j=1}^{\ell} \mathbf{z}_j.$$

Although the rate of these codes have been studied extensively in e.g., [1]–[6], it remains to be determined: the gap between the known upper and lower bounds is still substantially large.

Research supported by NSERC Grant 250391-02, and by OTKA Grants T038198 and T046234. This work relates to Department of the Navy Grant N00014-04-1-4034 issued by the Office of Naval Research International Field Office. The United States Government has a royalty-free license throughout the world in all copyrightable material contained herein.

AMS classification number: 94B65, 94B25, 94A40

Miklós Csűrös is with the Department of Computer Science and Operations Research, Université de Montréal, C.P. 6128, succ. Centre-Ville, Montréal, Qué. H3C 3J7, Canada. E-mail: csuros@iro.umontreal.ca.

Miklós Ruszinkó is with the Computer and Automation Research Institute of the Hungarian Academy of Sciences, Budapest 1518, Hungary. E-mail: ruszinko@lutra.sztaki.hu.

An abstract of this paper appeared at the 2004 IEEE Symposium on Information Theory, which was held June 27–July 2 in Chicago, Ill.

Here we investigate the case when the receiver has to be able to identify *just one* user out of at most r active ones. Clearly, if a code is superimposed in the classical sense then it satisfies this requirement: being able to identify all active users, the receiver can always name just one. A practical motivation for studying r-SUT families rises from applications of combinatorial designs in genomics, reviewed in Section II. Section III discusses our results on the rate of single-user tracing superimposed codes. Section IV introduces the class of disjointly superimposed codes, and analyzes their extremal properties. Section V concludes the paper with some open questions.

II. Superimposed codes for the physical mapping of genomic clones

A recently emerging application of superimposed codes, and group testing methods in general, is for the analysis of genomic data. Examples include the quality-control of DNA chips [7], and diverse applications related to genome sequencing: closing the remaining gaps at the end of a sequencing project [8], and clone library screening [9], which we consider here in more detail. The sequencing of large genomes (such as human) rely on *genomic clones*. We describe here briefly the relevant procedures, somewhat simplifying the problem. A recent overview of large genome sequencing techniques is given by E. Green [10]. The genome of an organism can be described by a sequence over a four-letter alphabet, corresponding to the four nucleotides used in DNA. Mammalian genome sizes are in the order of billions. For our purposes, a genomic clone is a random contiguous fragment of the genome. (Fragments are inserted into a host cell, which multiplies and thus creates many identical copies of the original cell containing the same piece of inserted foreign DNA fragment, hence the term "clone.") Typical clone fragment sizes are 100–200 thousand nucleotides. A *clone library* is a collection of genomic clones, produced using a large number of random fragments from many genome copies. The fragments correspond essentially to a uniform sampling of the whole genome. The information on which part of the genome the fragments originate from is lost in the course of random sampling, and needs to be determined using additional techniques. In a preliminary step to complete genome sequencing, called *physical mapping*, this information is established, by exploring overlaps between clone fragments. Using the physical map, a smaller set of minimally overlapping clones is selected in order to sequence the clones one-by-one. For instance, while sequencing the human genome, more than 300 thousand genomic clones were analyzed and about 30 thousand were selected for complete sequencing [11].

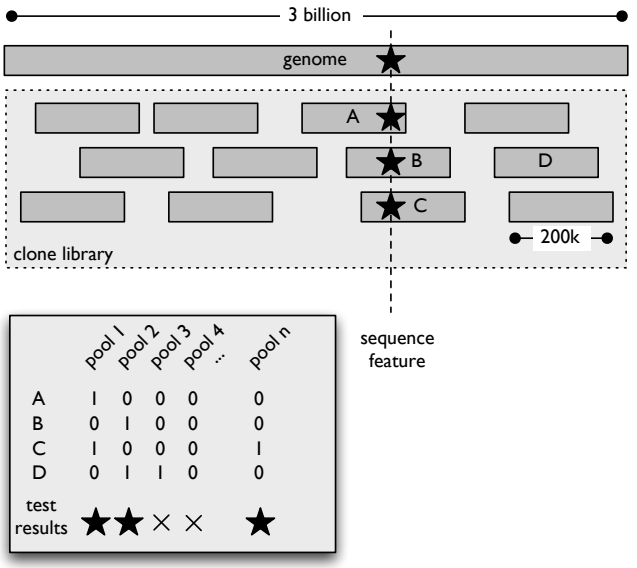

Fig. 1. Clone library screening. A clone library is a collection of random fragments from a genome. Clones in a library are tested for the presence of a sequence feature, such as homology to a given region in a related genome. The tests are carried out by pooling the clones: if the clone subset comprising the pool contains the feature, the test is positive, otherwise it is not.

The main issue in constructing a physical map is the discovery of overlaps. The key technique is to test sequence features, which are necessarily shared by overlapping clones. Often, a group-testing approach is employed by *pooling* the DNA from different clones: a pool is defined by a subset of clones that are screened together in a single experimental step. Figure 1 illustrates the concept of clone pooling. In the terminology of superimposed codes, clones correspond to users and pools correspond to the coordinates of the user vectors: the i-th clone is included in pool j if the j-th bit of user vector $\mathbf{x}_i$ equals one. Active users correspond to clones containing a particular feature. When testing a feature, pools are tested individually, and exactly those pools that contain a clone with the given feature test positive. The set (or at least one) of the clones containing the feature has to be determined from the set of positive pools, in the same way as the set of active users needs to be determined from the bitwise OR of their vectors.

Historically, the most widely used features are short (up to the order of hundreds of letters) contiguous sequences that occur once in the genome, called *Sequence Tagged Sites* (STS). All DNA in a pool can be tested for the presence of a given STS, by hybridization for example. Pooling designs for the purposes of STS screening have been studied extensively [9], [12], [13], and this particular application inspired many recent theoretical results on superimposed codes and non-adaptive group testing procedures [14]–[17].

A more recent application uses shotgun sequences [18], [19] for testing sequence features in pooled clones. Pooled Genomic Indexing [19] maps genomic clones to a reference genome sequence. Thus, the type of sequence feature that is tested by PGI is similarity to a region in the reference genome. In contrast to STS screening, the features are not defined before the experiment but are found in the analysis of the outcome. In a current application, (unsequenced) rhesus macaque clones are being mapped to the human genome. The raw outcome of the experiment is a list of mappings between sets of pools and regions in the reference sequence. Each mapping is indicative of the fact that some clones are similar to the same region in the reference sequence. The set of pools containing those clones is observed by the experimenter, along with the reference region.

The results of STS screening or a PGI experiment can be used to select clones for complete sequencing. If the purpose of the experiment is to identify clones that are particularly interesting and to sequence them completely, a single-user tracing code is more adequate for the pooling design than a "fully" superimposed code. In PGI, for instance, a number of overlapping macaque clones may include the same region that is homologous to a particular human gene: the experimenter will want to identify at least one of those clones for complete sequencing, but there is no need to identify all of them as they convey the same information about the genome.

The bound r on the number of "active users", i.e., the number of clones exhibiting a given feature is determined by the number of clones T. The size of a clone library is characterized by the *coverage*, which equals $c = TL/G$ where L is the average length of a clone, and G is the total genome length. Various aspects of clone overlaps can be studied by modeling the clone positions as arrival times in a Poisson-process. For example, the number of clones that include a given position in the genome is a Poisson random variable with expected value c [20]. Clone library coverage values are typically below ten, and are rarely above twenty. If unique sequence features are used, then every feature is shared by, say, at most $r = \lceil 2c \rceil$ clones with high probability.

III. Single-user tracing superimposed codes

As the question is rather of a combinatorial nature, we introduce a set terminology. Accordingly, codewords are characteristic vectors of subsets of a set $[n] = \{1, \dots, n\}$ where $n > 0$, i.e., the subset A corresponds to the binary vector $\mathbf{x} = (x^1, \dots, x^n)$ with $x^i = 1$ if and only if $i \in A$, and vice versa.

Throughout the paper, we use the de Finetti notation for indicator functions, i.e., $\{\cdots\}$ denotes an event, or its indicator function, depending on the context. We write $f(m) = o(g(m))$ if the sequence $f(m)/g(m) \to 0$ as $m \to \infty$. When the base of the logarithm matters, we use $\lg$ to denote binary logarithm.

Definition 3.1: A family $\mathcal{F} \subseteq 2^{[n]}$ is r-superimposed if

$$\bigcup_{i=1}^{k} A_i \neq \bigcup_{j=1}^{\ell} B_j$$

for any

$$\{A_1, A_2, \dots, A_k\} \neq \{B_1, B_2, \dots, B_\ell\},$$

$1 \leq k, \ell \leq r$; $A_1, A_2, \dots, A_k, B_1, B_2, \dots, B_\ell \in \mathcal{F}$.

We are interested in *r-single-user-tracing* (r-SUT) families, defined as follows.

Definition 3.2: A family $\mathcal{F}$ is r-SUT if for all choices of $\mathcal{F}_1, \ldots, \mathcal{F}_k \subseteq \mathcal{F}$ with $1 \le |\mathcal{F}_i| \le r$,

$$\bigcup_{A\in\mathcal{F}_1} A = \bigcup_{A\in\mathcal{F}_2} A = \cdots = \bigcup_{A\in\mathcal{F}_k} A$$

implies $\cap_{i=1}^{k}\mathcal{F}_i \neq \emptyset$. Equivalently, there exists such a *single-user-tracing function* ϕ: $2^{[n]} \mapsto \mathcal{F}$ that for all $\mathcal{F}' \subseteq \mathcal{F}$ with $1 \le |\mathcal{F}'| \le r$, $\phi(\cup_{A\in\mathcal{F}'}A) \in \mathcal{F}'$.

The following (folklore) lemma shows that it is enough to consider $k \le r+1$ in Definition 3.2.

Lemma 3.3: Let $k \ge r+1$. Let $S_1, \ldots, S_k$ be a collection of sets, each containing at most r elements. If for all choices of $1 \le i_1 < \cdots < i_{r+1} \le k$, $\cap_{j=1}^{r+1} S_{i_j} \neq \emptyset$, then $\cap_{i=1}^{k} S_i \neq \emptyset$.

Proof: For the sake of contradiction, suppose that $\cap_{i=1}^{k} S_i = \emptyset$. For all $a \in S_1$, select $i(a)$ such that $a \notin S_{i(a)}$. Then the intersection of the at most $(r+1)$ sets S_1 and $S_{i(a)}$ is empty. ■

For every base set size n and r, let $f(n,r)$ denote the maximum size of an r-superimposed family, and $g(n,r)$ denote the maximum size of an r-SUT family. In what follows, we give bounds on the *rate* of r-SUT families, which is

$$R_g(r) = \limsup_{n\to\infty} \frac{\lg g(n,r)}{n}.$$

Theorem 3.4: There exist constants $c_1, c_2 > 0$ such that

$$\frac{c_1}{r^2} \le R_g(r) \le \frac{c_2}{r} \tag{1}$$

Proof of the lower bound: Clearly, if $\mathcal{F}$ is r-superimposed then it is r-SUT. Therefore

$$g(n,r) \ge f(n,r) \ge 2^{c_1 n/r^2},$$

where the latter inequality can be found, say, in [3]. This gives the lower bound in (1). □

In order to prove the upper bound, we relate r-single-user-tracing to another property investigated in [21], [22].

Definition 3.5: (Alon, Fachini, Körner, [21]) A family $\mathcal{F}$ is r-locally thin if for all subsets $\mathcal{F}' \subseteq \mathcal{F}$ with $|\mathcal{F}'| = r$, there exists $x \in [n]$ such that

$$\sum_{A\in\mathcal{F}'} \{x \in A\} = 1,$$

i.e., there exists an element x that appears in exactly one member of $\mathcal{F}'$.

We need the following strengthening of this definition.

Definition 3.6: A family $\mathcal{F}$ is $\le r$-locally thin if for all subsets $\mathcal{F}' \subseteq \mathcal{F}$ with $1 \le |\mathcal{F}'| \le r$, there exists such $x \in [n]$ that

$$\sum_{A\in\mathcal{F}'} \{x \in A\} = 1.$$

Lemma 3.7: If $\mathcal{F}$ is r-SUT then it is $\le (r+1)$-locally thin.

Proof: Contrary to the lemma, assume that there is a subset $\mathcal{F}' = \{A_1, \ldots, A_k\}$, $1 \le k \le r+1$ for which $\sum_{i=1}^{k}\{x \in A_i\} \neq 1$ holds for all $x \in [n]$. For $i = 1, \ldots, k$, let $\mathcal{F}_i = \mathcal{F}' - \{A_i\}$. Since every element is covered at least twice by the members of $\mathcal{F}'$,

$$\bigcup_{A\in\mathcal{F}_1} A = \bigcup_{A\in\mathcal{F}_2} A = \cdots = \bigcup_{A\in\mathcal{F}_k} A, \quad \text{while} \quad \bigcap_{j=1}^{k} \mathcal{F}_j = \emptyset.$$

The existence of $\mathcal{F}_1, \cdots, \mathcal{F}_k$ contradicts the r-SUT property. ■

Let $h'(n,r)$, $h^*(n,r)$ be the maximum size of *r-locally thin*, *$\le r$-locally thin* families, respectively.

Corollary 3.8:

$$g(n,r) \le h^*(n,r+1) \le h'(n,r+1). \tag{2}$$

Proof: Here the first inequality comes from Lemma 3.7, while the second one follows directly from the definitions. ■

Alon, Fachini and Körner [21] proved the following theorem.

Theorem 3.9:

$$\begin{aligned} R_{h'}(r) &< \frac{2}{r} && \text{for } r \text{ even;} \\ R_{h'}(r) &< \frac{c\log r}{r} && \text{for } r \text{ odd, } c \text{ is constant.} \end{aligned} \tag{3}$$

Proof of the upper bound in Theorem 3.4: If r is odd, then $(r+1)$ is even. Hence, by (2) and (3),

$$R_g(r) \le R_{h^*}(r+1) \le R_{h'}(r+1) < \frac{2}{r+1}.$$

If r is even, then by the monotonicity of $h^*(n,r)$, (2), and (3),

$$R_g(r) \le R_{h^*}(r+1) \le R_{h^*}(r) \le R_{h'}(r) < \frac{2}{r}.$$

In either case, the upper bound holds in Eq. (1) with $c_2 = 2$. ■

Lemma 3.10 below allows for an alternative, self-contained proof of our upper bound on R_g, without using the (strong) bounds of Theorem 3.9. It gives a sufficient upper bound for $h^*(n,r)$ when r is even, which can then be employed with the monotonicity argument.

Lemma 3.10: Let r be even. If $\mathcal{F}$ is $\le r$-locally thin, then the modulo two sums of $(r/2)$-sets of characteristic vectors associated with members of $\mathcal{F}$ are all different.

Proof: For the sake of contradiction assume that there are two collections $\mathcal{F}_1$ and $\mathcal{F}_2$ with the same modulo two sums. Consider the symmetric difference $\mathcal{F}' = \mathcal{F}_1 \triangle \mathcal{F}_2$. Clearly, it contains at most r sets, and every element in $\cup_{A\in\mathcal{F}'}A$ is covered at least twice (in fact, even times) by members of $\mathcal{F}'$. ■

Corollary 3.11: If r is even, then $R_{h^*}(r) \le \frac{2}{r}$.

Proof: By Lemma 3.10, $\binom{h^*(n,r)}{r/2} \le 2^n$. ■

IV. Disjointly r-superimposed codes

Another important case implicated in the multiple access model of Section 1 is when the receiver must distinguish only between *disjoint sets of active users*. The following definition captures this notion.

Definition 4.1: A family $\mathcal{F} \subseteq 2^{[n]}$ is disjointly r-superimposed if

$$\bigcup_{i=1}^{k} A_i \neq \bigcup_{j=1}^{\ell} B_j \tag{4}$$

is implied by

$$\{A_1, A_2, \ldots, A_k\} \cap \{B_1, B_2, \ldots, B_\ell\} = \emptyset$$

for all $1 \le k, \ell \le r$; $A_1, A_2, \ldots, A_k, B_1, B_2, \ldots, B_\ell \in \mathcal{F}$.

Despite the seemingly slight difference between Definitions 4.1 and 3.1, the extremal properties of disjointly r-superimposed families and r-superimposed ones are completely different.

Let $h(n,r)$ be the maximum size of disjointly r-superimposed families.

Lemma 4.2: If $\mathcal{F}$ is r-superimposed then it is r-SUT. If $\mathcal{F}$ is r-SUT, then it is disjointly r-superimposed. Hence,

$$f(n,r) \le g(n,r) \le h(n,r).$$

Proof: The first part is already proved. The second part follows from the fact that if $\mathcal{F}$ is not disjointly r-superimposed, then there exist $\mathcal{A} = \{A_1, \ldots, A_k\} \subseteq \mathcal{F}$ and $\mathcal{B} = \{B_1, \ldots, B_\ell\} \subseteq \mathcal{F}$ such that $\cup_{i=1}^{k} A_i = \cup_{j=1}^{\ell} B_j$ while $\mathcal{A} \cap \mathcal{B} = \emptyset$. ∎

While we do not know if there is an exponential gap between r-superimposed and r-SUT families, the following theorem shows that there is such a gap between r-superimposed and disjointly r-superimposed ones.

Theorem 4.3: The rate of disjointly r-superimposed codes is bounded as

$$\frac{1}{2r} \le R_h(r) \le \Big(\frac{1}{2} + o(1)\Big)\frac{\lg r}{r}. \tag{5}$$

The key to the upper bound is the following observation.

Lemma 4.4: If $\mathcal{F}$ is disjointly r-superimposed then the vector sums of r-size sets of characteristic vectors associated with members of $\mathcal{F}$ are all different.

Proof: For the sake of contradiction assume that there are two collections $\mathcal{F}_1, \mathcal{F}_2 \in \binom{\mathcal{F}}{r}$, with the same vector sums. Consider $\mathcal{F}_1' = \mathcal{F}_1 \setminus \mathcal{F}_2$ and $\mathcal{F}_2' = \mathcal{F}_2 \setminus \mathcal{F}_1$. Clearly, $|\mathcal{F}_1'|, |\mathcal{F}_2'| \le r$, and the vector sums of members of $\mathcal{F}_1'$ and $\mathcal{F}_2'$ are the same. But then $\cup_{A \in \mathcal{F}_1'} A = \cup_{B \in \mathcal{F}_2'} B$, while $\mathcal{F}_1'$ and $\mathcal{F}_2'$ are disjoint, which is a contradiction. ∎

Now, in a vector sum $\mathbf{y} = (y^1, \ldots, y^n)$ of r binary vectors, $0 \le y^i \le r$ holds in every coordinate i. The number of possible vector sums is thus $(r+1)^n$, and therefore

$$\binom{h(n,r)}{r} \le (r+1)^n$$

must hold. This gives an upper bound with a constant factor of 1 in (5). In order to obtain the factor of $\frac{1}{2}$, we use a second moment method combined with a volume argument: we show that coordinates of almost all vectors in $\mathcal{F}$ deviate within $\sqrt{r}$ around the mean (instead of $r/2$, as above). In fact, we show that if a family $\mathcal{F}$ of subsets of $[n]$ has the property that for every choice of r sets, the sum of the corresponding characteristic vectors gives a different value, then the upper bound in 5 already holds. We prove Theorem 4.3 after Lemma 4.5 below.

For a set $\mathcal{A} \subseteq \{0,1\}^n$ of binary vectors of length n, $\mathbf{s}(\mathcal{A})$ stands for the sum of its elements:

$$\mathbf{s}(\mathcal{A}) = \sum_{\mathbf{x} \in \mathcal{A}} \mathbf{x}.$$

Lemma 4.5: Let $\mathcal{F}$ be a set of binary vectors of length n, and let $T = |\mathcal{F}|$. Let $\mathbf{c} = T^{-1} \sum_{\mathbf{v} \in \mathcal{F}} \mathbf{v}$ be the average vector of the set. For every integer $1 \le r \le T$, the inequality

$$\sum_{\mathcal{A} \in \binom{\mathcal{F}}{r}} \big\|\mathbf{s}(\mathcal{A}) - r\mathbf{c}\big\|^2 \le nr\binom{T}{r}$$

holds, where $\|\cdot\|$ is the Euclidean norm.

Proof: By definition of the norm,

$$\sum_{\mathcal{A} \in \binom{\mathcal{F}}{r}} \big\|\mathbf{s}(\mathcal{A}) - r\mathbf{c}\big\|^2 = \sum_{\mathcal{A} \in \binom{\mathcal{F}}{r}} \|\mathbf{s}(\mathcal{A})\|^2 + \sum_{\binom{\mathcal{F}}{r}} r^2\|\mathbf{c}\|^2 - \sum_{\mathcal{A} \in \binom{\mathcal{F}}{r}} 2r\mathbf{c}\mathbf{s}(\mathcal{A}) \tag{6}$$

Clearly, the second term in (6) gives $\binom{T}{r} r^2 \|\mathbf{c}\|^2$. The third term is

$$\sum_{\mathcal{A} \in \binom{\mathcal{F}}{r}} 2r\mathbf{c}\mathbf{s}(\mathcal{A}) = 2r\mathbf{c}\binom{T-1}{r-1}\sum_{\mathbf{v} \in \mathcal{F}} \mathbf{v} = 2\binom{T}{r} r^2 \|\mathbf{c}\|^2,$$

since in the sum $\sum_{\mathcal{A} \in \binom{\mathcal{F}}{r}} \mathbf{s}(\mathcal{A})$ every vector $\mathbf{v} \in \mathcal{F}$ appears with multiplicity $\binom{T-1}{r-1}$, which is the number of distinct r-sets in which a given vector $\mathbf{v}$ is contained. The first term of (6) can be bounded as follows.

$$\begin{aligned}
\sum_{\mathcal{A} \in \binom{\mathcal{F}}{r}} \|\mathbf{s}(\mathcal{A})\|^2 &= \sum_{\mathcal{A} \in \binom{\mathcal{F}}{r}} \Big\|\sum_{\mathbf{v} \in \mathcal{A}} \mathbf{v}\Big\|^2 \\
&\le \sum_{\mathcal{A} \in \binom{\mathcal{F}}{r}} \Big(nr + 2 \sum_{\substack{1 \le i < j \le r \\ \mathbf{v}_i, \mathbf{v}_j \in \mathcal{A}}} \mathbf{v}_i \mathbf{v}_j\Big) \\
&= nr\binom{T}{r} + 2\binom{T-2}{r-2} \sum_{\substack{1 \le i < j \le T \\ \mathbf{v}_i, \mathbf{v}_j \in \mathcal{F}}} \mathbf{v}_i \mathbf{v}_j \\
&= nr\binom{T}{r} + 2\binom{T-2}{r-2} \sum_{\substack{1 \le i < j \le T \\ \mathbf{v}_i, \mathbf{v}_j \in \mathcal{F}}} \mathbf{v}_i \mathbf{v}_j \\
&\quad + \binom{T-2}{r-2} \sum_{\mathbf{v} \in \mathcal{F}} \|\mathbf{v}\|^2 - \binom{T-2}{r-2} \sum_{\mathbf{v} \in \mathcal{F}} \|\mathbf{v}\|^2 \\
&= nr\binom{T}{r} + \binom{T-2}{r-2} \Big\|\sum_{\mathbf{v} \in \mathcal{F}} \mathbf{v}\Big\|^2 - \binom{T-2}{r-2} \sum_{\mathbf{v} \in \mathcal{F}} \|\mathbf{v}\|^2 \\
&= nr\binom{T}{r} + \binom{T-2}{r-2} T^2 \|\mathbf{c}\|^2 - \binom{T-2}{r-2} \sum_{\mathbf{v} \in \mathcal{F}} \|\mathbf{v}\|^2.
\end{aligned}$$

For the inequality, we used that the norm square of every vector is at most n, as every vector is binary. Subsequently, we used that every pair of vectors appears together in exactly $\binom{T-2}{r-2}$ sets of size r, and thus every product $\mathbf{v}_i \mathbf{v}_j$ occurs that many times.

Returning to Eq. (6), by the above computation we get

$$\begin{aligned}\sum_{\mathcal{A}\in\binom{\mathcal{F}}{r}} &\big\|\mathbf{s}(\mathcal{A})-r\mathbf{c}\big\|^2 \\ &\le nr\binom{T}{r}+\binom{T-2}{r-2}T^2\|\mathbf{c}\|^2-\binom{T-2}{r-2}\sum_{\mathbf{v}\in\mathcal{F}}\|\mathbf{v}\|^2 \\ &\quad-\binom{T}{r}r^2\|\mathbf{c}\|^2 \\ &= nr\binom{T}{r}+\binom{T-2}{r-2}T^2\|\mathbf{c}\|^2-\binom{T-2}{r-2}\sum_{\mathbf{v}\in\mathcal{F}}\|\mathbf{v}\|^2 \\ &\quad-\binom{T-2}{r-2}r^2\frac{T(T-1)}{r(r-1)}\|\mathbf{c}\|^2.\end{aligned}$$

From $r\le T$ follows that $-r^2\dfrac{T(T-1)}{r(r-1)}\le -T^2$. Therefore,

$$\begin{aligned}\sum_{\mathcal{A}\in\binom{\mathcal{F}}{r}}\big\|\mathbf{s}(\mathcal{A})-r\mathbf{c}\big\|^2 &\le nr\binom{T}{r}-\binom{T-2}{r-2}\sum_{\mathbf{v}\in\mathcal{F}}\|\mathbf{v}\|^2 \\ &\quad+\binom{T-2}{r-2}T^2\|\mathbf{c}\|^2-\binom{T-2}{r-2}T^2\|\mathbf{c}\|^2,\end{aligned}$$

which implies to the desired result. ∎

Proof of Theorem 4.3: First we prove the upper bound. Take an arbitrary set $\mathcal{F}\subseteq\{0,1\}^n$ of binary vectors of length n, such that the vector sums are different for all choices of r vectors. (By Lemma 4.4, the set of characteristic vectors for a disjointly r-superimposed family fulfills this condition.) Let $T=|\mathcal{F}|$. As in Lemma 4.5, define the average vector $\mathbf{c}=T^{-1}\sum_{\mathbf{v}\in\mathcal{F}}\mathbf{v}$. Let $\mathcal{A}\subseteq\mathcal{F}$ be a random subset of size r, chosen with uniform probability. Consider the random variable $\xi=\|\mathbf{s}(\mathcal{A})-r\mathbf{c}\|$, the distance of $\mathbf{s}(\mathcal{A})$ from its mean. By Lemma 4.5 and Jensen's inequality [23], the expected distance $\mathbb{E}\xi\le\sqrt{nr}$. By Markov's inequality [23],

$$\mathbb{P}\{\xi\ge\lambda^{-1}\sqrt{nr})\}\le\lambda$$

for all $0<\lambda<1$. This means that for any constant $0<\lambda<1$, at least the $(1-\lambda)$ fraction of all sums for r-size subsets of $\mathcal{F}$ lie within the n-dimensional ball B of radius $\lambda^{-1}\sqrt{nr}$ centered at the point $r\mathbf{c}$. Therefore, the number of integer lattice points in B is an upper bound for $(1-\lambda)\binom{T}{r}$. Consider a larger ball B' with radius $(\sqrt{nr}/\lambda+\sqrt{n}/2)$ centered at $r\mathbf{c}$. Its volume bounds the number of lattice points in B from above. To see this, draw an n-dimensional unit cube centered at each lattice point in B. All the cubes are within B', and to each integer lattice point a unit volume is associated. Using the well-known formula for the volume of an n-dimensional ball (e.g., [24]),

$$(1-\lambda)\binom{T}{r}\le\frac{\pi^{n/2}\Big(\lambda^{-1}\sqrt{nr}+\frac{1}{2}\sqrt{n}\Big)^n}{\Gamma(1+n/2)},$$

where $\Gamma(x)$ is the complete gamma function. An application of Stirling's approximation [23] to bound $\Gamma(1+n/2)$ leads to

$$\frac{\lg T}{n}\le\frac{\lg r}{2r}+\Theta\Big(\frac{1}{r}\Big)+\frac{o(n)}{n},$$

which is tantamount to the upper bound of (5).

We prove the lower bound in (5) with a probabilistic argument. (This proof was also observed by László Györfi.)

Let $\mathcal{F}$ be a randomly constructed family of size T, where T will be specified later. Every set $A_i\in\mathcal{F}$ is constructed randomly so that $x\in A_i$ with probability $(1-2^{-1/r})$ for all x independently. We prove that $\mathcal{F}$ is disjointly r-superimposed with non-zero probability for some $T=2^{\Theta(n/r)}$. Let $\mathcal{A},\mathcal{B}\subseteq\mathcal{F}$ be two disjoint sets: $\mathcal{A}=\{A_1,\dots,A_k\}$ and $\mathcal{B}=\{B_1,\dots,B_\ell\}$, where $\mathcal{A}\cap\mathcal{B}=\emptyset$ and $1\le k,\ell\le r$. Define $A=\cup_{i=1}^k A_i$ and $B=\cup_{j=1}^\ell B_j$. Equation (4) is violated if for all $x\in[n]$, either $x\in A\cap B$ or $x\notin A\cup B$. Since all A_i and B_j are independent,

$$\begin{aligned}p(k,\ell)=\mathbb{P}\bigcap_{x\in[n]}\Big(\{x\notin A;x\notin B\}\cup\{x\in A;x\in B\}\Big) \\ =\Big(p^{k+\ell}+(1-p^k)(1-p^\ell)\Big)^n,\quad(7)\end{aligned}$$

where $p=2^{-1/r}$. The expected number of disjoint set pairs that violate Eq. (4) is thus

$$N=\sum_{k=1}^{r}\sum_{\ell=0}^{k-1}\binom{T}{k}\binom{T-k}{\ell}p(k,\ell)+\sum_{k=1}^{r}\binom{\binom{T}{k}}{2}p(k,k).\quad(8)$$

By the choice of p and the fact that $k,\ell\le r$, $(1-p^k)\le p^k$ and $(1-p^\ell)\le p^\ell$. Consequently, the right-hand side of Eq. (7) is bounded from above as $p(k,\ell)\le\min\{p^{nk},p^{n\ell}\}$. Hence the right-hand side of Eq. (8) is bounded from above as

$$N\le\sum_{k=1}^{r}\sum_{\ell=0}^{k-1}\binom{T}{k}\binom{T-k}{\ell}p^{nk}+\sum_{k=1}^{r}\binom{\binom{T}{k}}{2}p^{nk}\quad(9)$$

Now, for $T\ge 2r^2+r-1$,

$$\begin{aligned}\sum_{\ell=0}^{k-1}\binom{T-k}{\ell}\le\sum_{\ell=0}^{k-1}\binom{T}{\ell}\le k\binom{T}{k-1} \\ =\frac{k^2}{T-k+1}\binom{T}{k}\le\frac{1}{2}\binom{T}{k},\end{aligned}$$

and $\binom{\binom{T}{k}}{2}<\Big(\binom{T}{k}\Big)^2/2$. Subsequently, Eq. (9) is bounded by

$$N\le\sum_{k=1}^{r}\Big(\binom{T}{k}\Big)^2p^{nk}.\quad(10)$$

In Eq. (10), the largest term is the one for $k=1$ if $T\le(k+1)p^{-n/2}+k$ for all k, i.e., if

$$T\le 1+2p^{-n/2}=1+2\cdot 2^{\frac{n}{2r}}\quad(*)$$

Then by Eq. (10),

$$N\le rT^2p^n$$

and thus $N<1$ if

$$T<\frac{p^{-n/2}}{\sqrt{r}}=\frac{2^{\frac{n}{2r}}}{\sqrt{r}}.\quad(**)$$

Between (*) and (**), (**) is more restrictive for all n and r. As a consequence, there exists a disjointly r-superimposed family of size $T=\left\lceil r^{-1/2}2^{\frac{n}{2r}}\right\rceil-1$, which implies the lower bound of (5). ∎

V. Open problems

We conclude by posing the following open problems.

Problem 5.1: It is known that

$$\frac{c_1}{r^2} \leq R_f(r) \leq \frac{c_2 \lg r}{r^2}.$$

Try to diminish the gap between the two bounds.

Problem 5.2: We show in this paper that

$$\frac{c_1}{r^2} \leq R_g(r) \leq \frac{c_2}{r}.$$

Try to diminish the gap between the two bounds.

Problem 5.3: We show in this paper that

$$\frac{1}{2r} \leq R_h(r) \leq \Big(\frac{1}{2} + o(1)\Big)\frac{\lg r}{r}.$$

Try to diminish the gap between the two bounds.

Problem 5.4: Do r-SUT and r-superimposed families differ significantly, i.e., do the functions $R_g(r)$ and $R_f(r)$ differ in magnitude?

Remark. In the course of submitting this paper we learnt that Noga Alon and Vera Asodi showed that $R_g(r) = \Omega(1/r)$ which answers Problems 5.2 and 5.4.

Acknowledgments

We are grateful for Sándor Győri for commenting on an earlier version of the manuscript. We would also like to thank Noga Alon and László Györfi for fruitful discussions.

References

[1] A. G. D'yachkov and V. V. Rykov, "Bounds on the length of disjunctive codes," *Problemi Peredachi Informatsii*, vol. 18, no. 3, pp. 7–13, 1982.

[2] P. Erdős, P. Frankl, and Z. Füredi, "Families of finite sets in which no set is covered by the union of r others," *Israel J. Math.*, vol. 51, pp. 79–89, 1985.

[3] F. K. Hwang and V. T. Sós, "Non-adaptive hypergeometric group testing," *Stud. Sci. Math. Hungar.*, vol. 22, pp. 257–263, 1987.

[4] Z. Füredi, "A note on r-cover-free families," *J. Combin. Theory Ser. A*, vol. 73, pp. 172–173, 1996.

[5] W. H. Kautz and R. C. Singleton, "Nonrandom binary superimposed codes," *IEEE Trans. Inform. Theory*, vol. IT-10, pp. 363–377, 1964.

[6] M. Ruszinkó, "On the upper bound of the size of the r-cover-free families," *J. Combin. Theory Ser. A*, vol. 66, pp. 302–310, 1994.

[7] C. C. Colbourn, A. C. H. Ling, and M. Tompa, "Construction of optimal quality control for oligo arrays," *Bioinformatics*, vol. 18, no. 4, pp. 529–535, 2002.

[8] R. Beigel, N. Alon, S. Kasif, M. S. Apaydin, and L. Fortnow, "An optimal procedure for gap closing in whole genome shotgun sequencing," in *Proc. Fifth Annual International Conference on Computational Biology*. ACM Press, 2001, pp. 22–30.

[9] D. J. Balding, W. J. Bruno, E. Knill, and D. C. Torney, "A comparative survey of non-adaptive pooling designs," in *Genetic Mapping and DNA Sequencing*, ser. IMA volumes in mathematics and its applications, T. Speed and M. S. Waterman, Eds. New York: Springer, 1996, vol. 81, pp. 133–154.

[10] E. D. Green, "Strategies for the systematic sequencing of complex genomes," *Nat. Rev. Genet.*, vol. 2, pp. 573–583, 2001.

[11] International Human Genome Sequencing Consortium, "Initial sequencing and analysis of the human genome," *Nature*, vol. 609, no. 6822, pp. 860–921, 2001.

[12] E. Barillot, B. Lacroix, and D. Cohen, "Theoretical analysis of library screening using an n-dimensional strategy," *Nucleic Acids Res.*, vol. 19, pp. 6241–6247, 1991.

[13] 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, pp. 21–30, 1995.

[14] M. A. Chateauneuf, C. J. Colbourn, D. L. Kreher, E. R. Lamken, and D. C. Torney, "Pooling, lattice square, and Union Jack designs," *Ann. Combin.*, vol. 3, pp. 27–35, 1999.

[15] A. G. D'yachkov, A. J. Macula, Jr., and V. V. Rykov, "New constructions of superimposed codes," *IEEE Trans. Inform. Theory*, vol. IT-46, pp. 284–290, 2000.

[16] A. J. Macula, "Probabilistic nonadaptive group testing in the presence of errors and DNA library screening," *Ann. Combin.*, vol. 3, pp. 61–69, 1999.

[17] H. Q. Ngo and D.-Z. Du, "New constructions of non-adaptive and error-tolerance pooling designs," *Discrete Math.*, vol. 243, pp. 161–170, 2002.

[18] W.-W. Cai, R. Chen, R. A. Gibbs, and A. Bradley, "A clone-array pooled strategy for sequencing large genomes," *Genome Res.*, vol. 11, pp. 1619–1623, 2001.

[19] M. Csűrös and A. Milosavljevic, "Pooled genomic indexing (PGI): mathematical analysis and experiment design." *J. Comput. Biol.*, 2004, in press.

[20] E. S. Lander and M. S. Waterman, "Genomic mapping by fingerprinting random clones: a mathematical analysis," *Genomics*, vol. 2, pp. 231–239, 1988.

[21] N. Alon, E. Fachini, and J. Körner, "Locally thin set families," *Combinatorics, Probability and Computing*, vol. 9, pp. 481–488, 2000.

[22] Z. Füredi, A. Gyárfás, and M. Ruszinkó, "On the maximum size of (p, Q)-free families," *Discrete Math.*, vol. 257, pp. 385–403, 2002.

[23] W. Feller, *An Introduction to Probability Theory and Its Applications*. New York: John Wiley & Sons, 1966.

[24] J. H. Conway and N. J. A. Sloane, *Sphere Packings, Lattices, and Groups*, 2nd ed. New York: Springer-Verlag, 1993.

Miklós Csűrös is an Assistant Professor at the Department of Computer Science and Operations Research, Université de Montréal. He received the Diploma degree in electrical engineering from Technical University of Budapest, Hungary in 1994, and the Ph. D. degree in computer science from Yale University in 2000. His current areas of interest include computational biology and genomics.

Miklós Ruszinkó is a senior research fellow at the Computer and Automation Research Institute of the Hungarian Academy of Sciences. He received his Ph. D. in mathematics from the Hungarian Academy of Sciences in 1995. He was a Postdoctoral Fellow at University of Cambridge in 1995–1996. He was a Visiting Assistant Professor and then a Visiting Associate Professor at Carnegie Mellon Unversity in 1998–2000 and in 2002–2003, respectively. His main research interests lie in combinatorics and codes.