Discrete Applied Mathematics 169 (2014) 68–72

Contents lists available at ScienceDirect

Discrete Applied Mathematics

journal homepage: www.elsevier.com/locate/dam

Threshold group testing with consecutive positives

Huilan Chang *, Yi-Lin Tsai

Department of Applied Mathematics, National University of Kaohsiung, Kaohsiung 811, Taiwan, ROC

ARTICLE INFO

Article history:
Received 3 July 2013
Received in revised form 26 October 2013
Accepted 16 December 2013
Available online 7 January 2014

Keywords:
Group testing
Sequential algorithm
Threshold
Consecutive positives

ABSTRACT

Threshold group testing introduced by Damaschke (2006) is a generalization of classical group testing where a group test yields a positive (negative) outcome if it contains at least u (at most l) positive items, and an arbitrary outcome for otherwise. Motivated by applications to DNA sequencing, group testing with consecutive positives has been proposed by Balding and Torney (1997) and Colbourn (1999) where n items are linearly ordered and up to d positive items are consecutive in the order. In this paper, we introduce threshold-constrained group tests to group testing with consecutive positives. We prove that all positive items can be identified in $\lceil \log_2(\lceil n/u \rceil - 1) \rceil + 2\lceil \log_2(u+2) \rceil + \lceil \log_2(d-u+1) \rceil - 2$ tests for the gap-free case ($u = l+1$) while the information-theoretic lower bound is $\lceil \log_2 n(d-u+1) \rceil - 1$ when $n \geq d+u-2$ and for $u = 1$ the best adaptive algorithm provided by Juan and Chang (2008) takes at most $\lceil \log_2 n \rceil + \lceil \log_2 d \rceil$ tests. We further show that the case with a gap ($u > l+1$) can be dealt with by the subroutines used to conquer the gap-free case.

1. Introduction

The idea of group testing, that has been used to distinguish a particular set of elements from a large population, is to group samples and then test the content of the groups. Group testing was originally proposed to detect soldiers with syphilis during World War II [10] and combinatorial group testing was first studied by Li [15]. We refer readers to the books by Du and Hwang [11,12] for a review of applications and variations of group testing. Specifically, in group testing, we have a set $\mathcal{N}$ of n items, each of which is either positive or negative; the positive ones usually stand for those of interest and their number is assumed to be at most d which is much smaller than n. A *group test*, also called a *pool*, is a subset of items that yields a positive outcome if and only if it contains at least one positive item. The main task of group testing is to identify the positive items by group tests as few as possible. Two types of algorithms are often investigated. When the outcome of a pool can be a reference to the selection of the next test, the group testing algorithm is *adaptive* or *sequential*, while in a *nonadaptive algorithm*, all tests are specified beforehand and are conducted simultaneously.

Motivated by applications in DNA sequencing, Balding and Torney [1] and Colbourn [8] studied the consecutive group testing model, where $\mathcal{N} = \{v_1, v_2, \ldots, v_n\}$ is the population set equipped with the linear order $v_1 \prec v_2 \prec \cdots \prec v_n$ and the positive items in $\mathcal{N}$ are consecutive under the order $\prec$. Colbourn [8] provided a sequential algorithm which takes at most $\log_2 d + \log_2 n + c$ tests in the worst case for some constant c; they also proposed a nonadaptive algorithm with $O(d + \log_2 n)$ tests. Subsequently, Muller and Jimbo [16] improved the nonadaptive approach by studying consecutive positive detectable matrices. Improving the sequential approaches, Juan and Chang [13] proved that sequential group testing needs at least $\lceil \log_2 dn \rceil - 1$ tests and can be accomplished by $\lceil \log_2 d \rceil + \lceil \log_2 n \rceil \leq \lceil \log_2 dn \rceil + 1$ tests if $n \geq d - 1$.

A group test is called an *interval test* if all its items are consecutive in $\mathcal{N}$ where the items in $\mathcal{N}$ are linearly ordered. To determine exon–intron boundaries within a gene [17,18], Cicalese et al. [6,7] studied *interval group testing* where all items in the population are linearly ordered and the group tests are interval tests. Motivated by applications in network monitoring

* Corresponding author. Tel.: +886 75916572.
E-mail addresses: huilan0102@gmail.com (H. Chang), d23455432b@gmail.com (Y.-L. Tsai).

0166-218X/$ – see front matter
http://dx.doi.org/10.1016/j.dam.2013.12.013

and infection propagation, Cheraghchi et al. [5] and Karbasi and Zadimoghaddam [14] considered *graph-constrained group testing* where a test should obey the restrictions imposed by a graph. More precisely, a test is admissible if it induces a connected subgraph or a path of a designated constraint graph. Therefore, if only interval tests are allowed, a group testing with consecutive positives corresponds to a graph-constrained group testing where the constraint graph is a path and an admissible group test induces a sub-path. Notice that all tests adopted by Juan and Chang [13] are interval tests.

In biology and chemistry experiments, the outcome of a test is determined by various factors such as the concentration of the positive items in the test. Damaschke [9] introduced the *threshold group testing* where a group test yields a positive (resp. negative) outcome if it contains at least u (resp. at most l) positive items and an arbitrary outcome otherwise. Recently, many improvements and variations of threshold group testing have been proposed [2–4]. Group testing with consecutive positives arises in the applications of genetic mapping and sequencing; therefore, introduction of threshold-constrained group tests to this group testing model is natural and could make it more applicable. The purpose of this paper is to study the *threshold group testing with consecutive positives*, namely, we will study the identification of the consecutive positives from $\mathcal{N} = \{v_1 \prec v_2 \prec \cdots \prec v_n\}$ by group tests with threshold constraint.

The difference $g = u-l-1$ between the thresholds is called the *gap*. Let D denote the set of positive items. Damaschke [9] showed that in threshold group testing the identification of D is possible only when $|D| \geq u$. Otherwise, any test could always yield a negative testing outcome. Moreover, in general, the set D can only be approximately identified within up to g false positives and g false negative. Such a set is especially called *g-approximate*, that is, S satisfies $|S \setminus D| \leq g$ and $|D \setminus S| \leq g$. For threshold group testing with consecutive positives, we provide algorithms composed of interval tests to identify D when the gap $g = 0$ and a g-approximate set when $g > 0$.

The rest of the paper is organized as follows. In Section 2, we introduce preliminary notions and a lower bound. For threshold group testing with consecutive positive items, we first deal with the gap-free case (Section 3) and then extend our results to the case $g > 0$ (Section 4).

2. Preliminaries

Recall that our group testing scenario consists of a set $\mathcal{N} = \{v_1, v_2, \ldots, v_n\}$ where $v_1 \prec v_2 \prec \cdots \prec v_n$ and an unknown set $D \subset \mathcal{N}$ of positive items that are consecutive under $\prec$. It is assumed that $u \leq |D| \leq d$. A test on a subset $S \subset \mathcal{N}$ is positive if $|S \cap D| \geq u$, negative if $|S \cap D| \leq l$ and arbitrary for otherwise.

We have the following information-theoretic lower bound.

Proposition 2.1. *If $n \geq d + u - 2$, then the number of group tests required to identify all positive items from $\mathcal{N}$ is at least $\lceil \log_2 n(d - u + 1) \rceil - 1$.*

Proof. The sample space consists of sets of i consecutive items for $i = u, \ldots, d$ and therefore is of size

$$\sum_{i=u}^{d} (n + 1 - i) = n(d - u + 1) - \frac{(d + u - 2)(d - u + 1)}{2} \geq \frac{n(d - u + 1)}{2}$$

when $n \geq d + u - 2$. The results follows. □

For a set S, let $\max(S)$ (resp. $\min(S)$) denote the element in S of the greatest (resp. smallest) index. We call an item in $\mathcal{N}$ *g-starter* if its index is in $\{i - g, \ldots, i\}$ and *g-terminal* if its index is in $\{j, \ldots, j + g\}$ when $D = \{v_i, v_{i+1}, \ldots, v_j\}$. We say a set $P \subset \mathcal{N}$ is *g-consecutive-approximate* if P consists of consecutive items, $\min(P)$ is a g-starter, and $\max(P)$ is a g-terminal. For a g-consecutive-approximate set P, an interval test containing at least u positives contains at least u items from P as well; an interval test containing at most l positives contains at most $g + l = u - 1$ items from P. Therefore, it would be impossible to distinguish D from a g-consecutive-approximate set by any series of interval tests. Notice that a g-consecutive-approximate set already contains all positive items and indeed removing g items of largest (or smallest) indices from a g-consecutive-approximate set yields a g-approximate one; however, some positive items could be excluded. Therefore, in the case with $g > 0$, we focus on identifying a g-consecutive-approximate set.

3. Threshold without gap

First, we consider that $g = 0$, that is, a pool is positive if and only if it contains at least u positive items. In this case, the identification of all positive items is achievable. We first partition $\mathcal{N}$ into $\lceil n/u \rceil$ parts of u consecutive items where we add some dummy negative items to the last part: $X_i = \{v_{(i-1)u+1}, \ldots, v_{iu}\}$ for $i = 1, 2, \ldots, \lceil n/u \rceil$. Let $\mathcal{X} = \{X_1, X_2, \ldots, X_{\lceil n/u \rceil}\}$. A *group test on $\mathcal{S} \subseteq \mathcal{X}$* is just a group test on $\cup_{X \in \mathcal{S}} X$.

For a set S of indexed elements, let $\mathrm{Half}_1(S)$ denote the set of $\lceil |S|/2 \rceil$ elements in S of the smallest indices and $\mathrm{Half}_2(S) = S \setminus \mathrm{Half}_1(S)$; further, define $\mathrm{Half}_2^*(S) = \{\max(\mathrm{Half}_1(S))\} \cup \mathrm{Half}_2(S)$. We use the following algorithm to find out two parts X_i and X_{i+1} such that $\min(D) \in X_i \cup \{\min(X_{i+1})\}$.

Lemma 3.1. *FIND-TWO-CANDIDATES returns $\{X_i, X_{i+1}\}$ from $\mathcal{X}$ such that $\min(D) \in X_i \cup \{\min(X_{i+1})\}$ in $\lceil \log_2(|\mathcal{X}| - 1) \rceil$ interval tests.*