Discrete Mathematics 340 (2017) 1488–1496

Contents lists available at ScienceDirect

Discrete Mathematics

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

Interval group testing for consecutive positives

Huilan Chang, Wei-Cheng Lan

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

ARTICLE INFO

Article history:
Received 18 June 2015
Received in revised form 4 February 2017
Accepted 12 February 2017
Available online 24 March 2017

Keywords:
Group testing
Pooling design
Interval group testing
Consecutive positives

ABSTRACT

To identify splice sites in a genome, Cicalese et al. (2005) studied interval group testing where all items in the search space are linearly ordered and each of them is either positive or negative. The goal is to identify all positive ones by interval group tests, each asking a query of the type "does a set of consecutive items contain any positive one?" We complete the study of error-tolerant one- and two-stage approaches by dealing with some cases which have not been well studied. Motivated by applications to DNA sequencing, group testing for consecutive positives has been proposed by Balding and Torney (1997) and Colbourn (1999) where n items are linearly ordered and up to p positive items are consecutive in the order. Juan and Chang (2008) provided an optimal sequential algorithm that consists of interval group tests. In this paper, we study error-tolerant one- and two-stage interval group testing for consecutive positives. Based on some results in the literature, we derive optimal nonadaptive algorithms. For two-stage approach, the number of interval group tests used is asymptotically upper bounded by $e\sqrt{n}$ which differs by a factor of $\sqrt{e}$ from the lower bound, where e is the number of errors allowed.

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 use the information obtained from each group. Specifically, in classical group testing, we are given a search space $\mathcal{N}$ of n items, where each item is either positive or negative. The positive items usually stand for those of interest and are needed to be identified. It is usually assumed that the number of positive items is at most p which is much smaller than n. The tool that we use to identify all positive items is a *group test* which performs a test $Q \subset \mathcal{N}$ (or, equivalently, asks a query Q) and reveals whether Q contains any positive element. The query Q is answered *yes* if it contains any positive; otherwise, it is answered *no*. Group testing was originally proposed to weed out men who were called up for induction but with syphilis during the World War [8]. Group testing has been well known for its various applications such as in network security [16], image compression [11], and molecular biology. Group testing also has strong links with several fields such as coding theory, information theory, and computational learning theory. We refer the readers to the books (Du and Hwang [9,10]) for a comprehensive review of the development and the major concepts in this area.

There are two general types of group testing algorithms: sequential and nonadaptive. When one can use the outcomes of the previous tests to design the next one, the group testing algorithm is called *sequential*. On the contrary, in a *nonadaptive algorithm*, all tests are specified beforehand and are conducted simultaneously. A compromise between these two types of algorithms is to consider a *multi-stage* algorithm that consists of several stages where all tests in the same stage are conducted simultaneously and the stages are considered sequential.

Two group testing models have been applied in DNA library screening: *interval group testing* and *group testing for consecutive positives*. The search space of these two models consists of linearly ordered items and is usually assumed to be

E-mail addresses: huilan0102@gmail.com (H. Chang), ddss12313@gmail.com (W. Lan).

http://dx.doi.org/10.1016/j.disc.2017.02.010
0012-365X/© 2017 Elsevier B.V. All rights reserved.

Table 1
Upper and lower bounds on $N^C(n, p, s, e)$.

Parameters	Lower bound	Upper bound	
$p = 1, s = 1, e \geq 0$	$\lceil \frac{(2e+1)(n+1)}{2} \rceil$	$\lceil \frac{(2e+1)(n+1)}{2} \rceil$	(Theorem 4.1)
$p \geq 2, s = 1, e \geq 0$	$(2e + 1)n$	$(2e + 1)n$	(Theorem 4.2)
$p = 1, s = 2, e \geq 1$	$\sqrt{(6e - 1)(n + 1)}$	$\sqrt{2(e + 1)(2e + 1)n} + \frac{5}{2}(e + 1)$	(Theorem 4.5)
$p = 2, s = 2, e = 1$	$\sqrt{10n}$	$2\sqrt{3n} + 5$	(Theorem 4.6)
$p \geq 2, s = 2, e = 1$	$\sqrt{10n}$	$2\sqrt{12n}$	(Theorem 4.8)
$p \geq 2, s = 2, e = 2$	$\sqrt{22n}$	$2\sqrt{30n}$	(Theorem 4.8)
$p \geq 2, s = 2, e \geq 3$	$\sqrt{2(6e - 1)n}$	$2\sqrt{3(e + 1)(2e - 1)n}$	(Theorem 4.8)

the set $\mathcal{N} = \{1, 2, \dots, n\}$. In order to determine exon–intron boundaries within a gene [15,17], Cicalese et al. [4–6] studied *interval group testing* where each group test is an interval, that is, all items in a test are consecutive in $\mathcal{N}$. Cheraghchi et al. [2] and Karbasi and Zadimoghaddam [13] studied *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. Interval group testing corresponds to graph-constrained group testing when the constraint graph is a path. On the other hand, motivated by application in DNA sequencing, Balding and Torney [1] and Colbourn [7] studied group testing for consecutive positives where the positive items are consecutive in $\mathcal{N}$. We briefly call such a model "consecutive-positive". Colbourn [7] provided a sequential algorithm which takes at most $\log_2 p + \log_2 n + c$ tests in the worst case for some constant c and also proposed a nonadaptive algorithm which takes $O(p + \log_2 n)$ tests. To improve the nonadaptive approach, Müller and Jimbo [14] studied consecutive positive detectable matrices. Juan and Chang [12] proved that sequential group testing needs at least $\lceil \log_2 pn \rceil - 1$ tests and can be accomplished by $\lceil \log_2 p \rceil + \lceil \log_2 n \rceil \leq \lceil \log_2 pn \rceil + 1$ tests if $n \geq p - 1$. It is worth pointing out that the optimal sequential algorithm provided by Juan and Chang [12] consists of interval group tests.

The graph-constrained group testing is motivated by applications in network monitoring and infection propagation: In *network tomography*, one of the key problems is to identify congested links from end-to-end path measurements. In *infection propagation*, we have a large population where a small number of people are infected by a certain epidemic disease and the task is to identify all infected individuals by sending agents to investigate people. When dealing with network monitoring problem, the links in the network could congest consecutively and when considering infection propagation problem, the certain disease could spread among people who are neighbouring. Then one may assume that the target set consists of neighbouring individuals. This motivates us to study interval group testing problem with the assumption that all positive items are consecutive. Furthermore, from the literature, we can see that constraining group tests to be intervals makes it harder to complete the work of identifying all positives. Let $N(n, p, s, e)$ be the worst-case number of interval queries that are necessary to successfully identify all positives in a search space of cardinality n and containing at most p positives under the assumption that s-stage algorithms are used and at most e erroneous answers are allowed. Generally, $N(n, p, s, e)$ is linear in n when $s = 1$ and linear in the square root of n when $s = 2$ (see [3–6]) while the worst-case performance of the classical group testing can be linear in $\log n$ for any types of algorithms. It is interesting to answer the following question: whether the assumption that positives are consecutive in the search space can reduce the degree of difficulty. We analogously use $N^C(n, p, s, e)$ to denote the worst-case number of interval queries that are necessary when assuming that all positive items appear consecutively in the search space.

An upper bound on $N(n, 2, 2, 1)$ has been studied (Theorem 7.9, [3]); however, there is a flaw in the argument used to obtain this upper bound. We amend the proof to obtain that $N(n, 2, 2, 1) \leq 2\sqrt{5n}+5$ (Theorem 4.6 in Section 3). In Section 4, we study one- and two-stage interval group testing for consecutive positives. We provide a lower bound and an upper bound on $N^C(n, p, s, e)$ for all $p, e \geq 0$ and $s = 1, 2$ (summarized in Table 1). Optimal nonadaptive algorithms are derived and for two-stage approach, the number of interval group tests used is asymptotically upper bounded by $e\sqrt{n}$ which differs by a factor of $\sqrt{e}$ from the lower bound.

In the literature, $N(n, 1, 2, e)$ was not discussed explicitly. When considering $p = 1$, it is clear that $N(n, p, s, e) = N^C(n, p, s, e)$ for all $s \geq 1$ and $e \geq 0$. We also have that

$$\sqrt{(6e - 1)(n + 1)} \leq N(n, 1, 2, e) \leq \sqrt{2(e + 1)(2e + 1)n} + \frac{5}{2}(e + 1)$$

for $e \geq 1$ (see Theorems 4.4 and 4.5).

2. Preliminaries

Some terminology and ideas for the study of interval group testing were introduced in [4,6]. We adopt most of the notations and describe their intuitive meanings before giving formal definitions. The search space is the set $\mathcal{N} = \{1, 2, \dots, n\}$ and the set of positives is denoted by P. For $i \leq j$, we use $[i, j]$ to denote the interval $\pi = \{i, i+1, \dots, j\}$ and its size is denoted by $|\pi|$. An interval query $Q = [i, j]$ has two boundaries: the left boundary is $(i - 1, i)$, and the right boundary is $(j, j + 1)$. For the sake of definiteness, we assume that, for any $k \in \mathcal{N}$, the query $[1, k]$ has left boundary $(0, 1)$, and the query $[k, n]$ has