
Journal of Combinatorial Optimization, 7, 327–337, 2003

The Incremental Group Testing Model for Gap Closing in Sequencing Long Molecules

FRANK K. HWANG fhwang@math.nctu.edu.tw
Department of Applied Mathematics, National Chiao Tung University, HsinChu 30050, Taiwan, Republic of China

WEN-DAR LIN wdlin@iis.sinica.edu.tw
Institute of Information Science, Academia Sinica, 128, Academia Road, Sec. 2, Taipei, Taiwan 115

Received February 5, 2003; Accepted July 29, 2003

Abstract. In this paper, we propose the incremental group testing model for the gap closing problem, which assumes that we can tell the difference between the outcome of testing a subset S, and the outcome of testing $S \cup \{x\}$. We also give improvements over currently best results in literature for some other models.

Keywords: contig sequencing, gap closing, group testing, multiplex PCR, detecting matrix, separating matrix, affine plane method

1. Introduction

A long molecule, hereafter referred to as the target sequence, is typically broken (several times) into fragments through shotgunning or restriction enzyme cutting for storage. Short fragments with readable lengths can be merged by using the overlapping at the end of one fragment and the head of another. Longer unreadable fragments can be sequenced by reference to a physical mapping of the target sequence, if one exists. However, due to insufficient design of coverage, contaminated vectors, errors and various other reasons, the above sequencing effort may not result in a whole piece of the long molecule, but many long pieces called contigs, with gaps in between. To sequence the target sequence, it becomes necessary to know the ordering of the contigs, and the problem is called the gap closing problem. The sequencing of the target sequence relies on some biological technology.

One such technology is the multiplex PCR, first reported by Burgart et al. (1992). From each contig, two short subsequences at its two ends are collected as primers. A multiplex PCR can test up to k primers and a PCR product will be produced if the primers in the test contain a pair of adjacent primers at the two ends of a gap (the two primers of the same contig are not an adjacent pair). In fact, the length of the PCR product reflects the length of the gap between the pair of adjacent primers. Therefore if there are several gaps with different lengths, then the test outcome will show different PCR products. Note that the multiplex PCR yields two possible mathematical models:

1. Quantitative model: The test outcome reveals the number of distinct PCR products (translated to number of pairs of adjacent primers).

2. Classical model: The test outcome reveals whether there is a PCR product in the test.

Although the quantitative model is much more powerful and requires fewer tests, it is also more liable to error as Grebinski and Kucherov (1998) warned: "However, this information (exact number of pairs of adjacent primers) has a limited value, as in practice only a restricted small number of products can be distinguished and, in addition, distinct products of similar length can be visible as a single one."

In this paper we propose the incremental model which lies between the classical and the quantitative model. The incremental model assumes that we can tell the difference between the outcome of testing a subset S, and the outcome of testing $S \cup \{x\}$, where x is a primer, provided x contributes to a PCR product. It seems that the incremental model suits the multiplex PCR experiment naturally. On one hand, it recognizes the information provided by different PCR products. On the other hand, it allows error in counting the products, as long as the error occurs in both outcomes of testing S and testing $S \cup \{x\}$. We give a nonadaptive algorithm under the incremental model (while none under the classical model is known). We also improve some currently best algorithms under both the quantitative and the classical model.

2. A nonadaptive algorithm under the incremental model

Traditionally, a group testing algorithm (see Du and Hwang (2000) for a general reference) is used to efficiently identify all positive objects among a set N of positive and negative objects. A group test is applicable to any subset S of N with two possible outcomes. A positive outcome indicates that S contains a positive object and a negative outcome indicates otherwise.

A nonadaptive group testing algorithm can be represented as a (binary) incidence matrix M where the columns are the objects and the rows are the tests, i.e., row i is the test consisting of all objects where corresponding bits in row i are 1's. Suppose M has t rows. Then the outcome of tests in M is a t-vector which is simply the boolean sum of all positive columns. Note that the t tests can be performed parallelly.

In the gap closing problem, the objects are the primers. However, there is a twist to the traditional group testing in the sense that we are not looking for positive primers, but pairs of adjacent primers. Treat the primers as nodes and adjacent primers as edges, then the problem is to identify the unknown edges instead of unknown nodes. This problem has also been studied in the group testing literature (see Chapter 12 of Du and Hwang (2000)) as "group testing on graphs." A test on a subset S of nodes reveals whether S contains an edge. It is known (p. 238 of Du and Hwang (2000)) that if the hidden graph contains a single edge, then a traditional group test and a graph group test are equivalent.

A 1-separable matrix is a binary matrix where all columns are distinct. Clearly, if only one object is positive, then a 1-separable matrix can identify it since the outcome vector will be identical to the column. Let B_n denote a $t \times n$ matrix, $2^{t-1} < n \leq 2^t$, where column i is the t-vector representing the binary number i. Then B_n is 1-separable (Du and Hwang, 2000). Hereinafter, we assume that there are n primers in the gap closing problem.

Set $M = B_{n-1}$. We use M to identify the unique neighbor of a given primer where the columns of M represent the other $n-1$ primers. Do the following steps:

1. Let $M^j(M^{(j)})$ be obtained from M by adding a column of 1's(0's) and identifying the column with node j.
2. By comparing the outcomes of test $i, i = 1, \ldots, t$, in M^j and $M^{(j)}$, we find out whether test i in M contains the neighbor of node j.
3. Since M is 1-separable, the outcomes of the t tests in M suffice to identify the neighbor of node j.

We can do the tests in M^j and $M^{(j)}$ for each $j = 1, \ldots, n$. But we can do better by recognizing duplicated tests in the sets $\{M^j\}$ and $\{M^{(j)}\}$.

Consider the n spaces separated by the $n-1$ columns of M. Let M^j be constructed by adding a column of 1's at space j. A space is called a 1(0)-space if it neighbors a 1(0). Note that a space can be both a 1-space and a 0-space. If a test contains a run of k 1's, then adding 1 at any one of the $k+1$ spaces separated by the k consecutive 1's induces the same test. Suppose test i of M has x 1's and r runs of 1's. Then it has $x+r$ 1-spaces, which induce r distinct tests. Thus the total number of distinct tests in $\{M^j\}$ corresponding to test i is $n-(x+r)+r = n-x$. On the other hand, using a similar argument, the total number of distinct tests in $\{M^{(j)}\}$ corresponding to test i is $n-(n-1-x) = x+1$ where $n-1-x$ is the number of 0's in test i. Summing up, a test in M induces $n+1$ distinct tests. So $\{M^j\}$ and $\{M^{(j)}\}$ together induces $(n+1)t = (n+1)\lceil \log(n-1) \rceil$ distinct tests.

3. Best results for the quantitative model

In this section, we will present the currently best results for the quantitative model from the literature, and then give our improvement to prepare for a comparison with the incremental model in the conclusion section.

The best sequential algorithm under the quantitative model was given by Grebinski and Kucherov (1998). They first gave a $13n$-test algorithm and then (Grebinski and Kucherov, 2000) improved it to $7n$.

Call the two primers from the same contig a couple. Then Grebinski and Kucherov's algorithm treated the couples, instead of the primers, as nodes. Thus there are $n/2$ nodes to start with, and each node has two neighbors. They gave a n-test algorithm which partitions the $n/2$ nodes into three parts X, Y, Z such that nodes in the same part are not adjacent. Therefore all edges lie in the three bipartite graphs G_{XY}, G_{XZ} and G_{YZ}. Then they gave an algorithm to identify the edges in each bipartite graph, knowing that each node is of degree at most 2.

We will mimic their method except using the n primers as nodes. Note that each primer has only one neighbor, while each couple has two. This difference brings three crucial advantages: (i) Only one bipartite graph is needed. (ii) More efficient subroutine exists to identify a single neighbor than two. (iii) Once an edge is identified, its two primers can be removed without affecting the other edges.

The partition of the n nodes into parts is to make sure that no edge exists between nodes in the same part. Since each couple has two neighbors, three parts are needed to meet the requirement. But a primer has only one neighbor, and hence two parts suffice. It takes n tests as before to partition the n nodes into two parts X and Y each with $n/2$ nodes.

A binary matrix M is called separating if $Ms \neq Ms'$ for all vectors $s \neq s'$ with nonnegative integral components; it is d-separating if s and s' are binary with at most d 1-entries. M is called d-detecting if each component takes value in the set $\{0, 1, \ldots, d-1\}$. Note that 1-separable and 1-seperating are the same; hence B_n is a 1-separating matrix. Lindstrom (1969) constructed $t \times n$ d-detecting matrices with $t = 2n/\log_d n$. Grebinski and Kucherov (2000) gave a probabilistic argument of the existence of a d-separating matrix with $4d \log_d n$ tests.

Construct a 2-detecting matrix $M_{n/2}$ on nodes of X and a 1-separating matrix $B_{n/2}$ on nodes of Y. Let $M_{n/2}(i)$ denote the ith row of $M_{n/2}$ and $B_{n/2}(j)$ the jth row of $B_{n/2}$. Define $M_{n/2} \otimes B_{n/2}$ to be the binary matrix where each row is a concatenation of $M_{n/2}(i)$ with $B_{n/2}(j)$ (denoted by $M_{n/2}(i) \oplus B_{n/2}(j)$), $1 \leq i \leq 2(n/2)/\log(n/2)$, $1 \leq j \leq \log(n/2)$. Then $M_{n/2} \otimes B_{n/2}$ has n rows.

The outcome vector of the set of rows $\{M_{n/2}(i) \oplus B_{n/2}(j)\colon 1 \leq i \leq 2(n/2)/\log(n/2)\}$ gives the degree (0 or 1) of every vertex of $M_{n/2}$ with respect to the set $B_{n/2}(j)$ since $M_{n/2}$ is a 2-detecting matrix. On the other hand, knowing the degree of a vertex v of $M_{n/2}$ in $B_{n/2}(j)$ for every $1 \leq j \leq \log(n/2)$ identifies the neighbor of v in $B_{n/2}$ since $B_{n/2}$ is 1-separating.

The n tests in $M_{n/2} \otimes B_{n/2}$ plus the n tests in partition gives a total of $2n$ tests. Further, the partition takes n rounds and $M_{n/2} \otimes B_{n/2}$ takes one round, resulting in a $(n+1)$-round algorithm.

The best nonadaptive algorithm under the quantitative model was given by Grebinski and Kucherov (2000) requiring $48(n/2)$ tests (using couples as vertices). This result is actually a special case of the general result of identifying a graph with $n/2$ nodes and maximum degree d in $24d(n/2)$ tests. The algorithm is similar to the sequential version except the partition stage is skipped. $M'_{n/2} \otimes M''_{n/2}$ is constructed, with

$$\frac{2(n/2)\lceil \log(d+1) \rceil}{\log n/2} \cdot \frac{4d \log(n/2)}{\log d} \approx 8d(n/2) \text{ tests},$$

where $M'_{n/2}$ is a $(d+1)$-detecting matrix, $M''_{n/2}$ is a d-separating matrix, and $M'_{n/2}$, $M''_{n/2}$ are defined on the same $n/2$ nodes. But the tests in $M'_{n/2} \otimes M''_{n/2}$ are not legitimate since there are edges within the two parts. A scheme is devised to translate these tests into legitimate ones in three times of tests, resulting in $24d(n/2)$ tests. For $d = 2$, $24n$ tests are required.

However, for $d = 2$, Lindstrom (1969) gave a $2\log(n/2)$-test construction of $M''_{n/2}$ (the base 2 is omitted in $\log_2$). Therefore, the number of tests becomes

$$\frac{4(n/2)}{\log(n/2)} \cdot 2\log(n/2) = 8(n/2),$$

which, when multiplied by 3, yields $24(n/2)$ tests, cutting the $48(n/2)$ into half. This observation was somehow missed in Grebinski and Kucherov (2000).

Again, by using primers instead of couples as nodes, $M'_{n/2} \otimes M''_{n/2}$ can be replaced by $M_n \otimes B_n$ with

$$\frac{2n}{\log n} \cdot \log n = 2n \text{ tests},$$

which, multiplied by 3, yields $6n$ tests.

4. Best results for the classical model

For the classical model, Grebinski and Kucherov (1998) proved that $O(n \log n)$ tests are necessary for any algorithm. Beigel et al. (2001) tightened this asymptotic lower bound to $0.5\, n \log n$ and gave a sequential algorithm achieving the bound. They also gave a 7-round algorithm with $0.75n \log n$ expected number of tests. We now give an algorithm which achieves the lower bound, and our simulation results show that its average number of rounds is bounded by $-\log_{0.22}(n/2)(\log n + 2) + 1$ for $n \leq 10000$. In fact, we prove that more than 97.5% of edges in average are identified in $3 \log n + 7$ rounds.

We use the affine plane method first proposed in Grebinski and Kucherov (1998) and Tettelin et al. (1996), but in a different way. An affine plane of order p (see Hell (1996) for general reference) is a balanced incomplete block design with p^2 elements, $p(p+1)$ blocks of size p such that each pair of elements appear together in exactly one block. To describe our method, we also need the following two results. Damaschke (1994) gave an algorithm which identifies a unique edge in a graph with n vertices in $\lceil \log \binom{n}{2} \rceil + 1$ tests. Johann (2002) observed that the algorithm works even if the graph has more than one edge to be detected.

We now state our algorithm:

Step 1. Find the smallest prime p such that $p^2 \geq n$. Randomly permute the n vertices. Add $p^2 - n$ dummy vertices labeled by $n+1, n+2, \ldots, p^2$ to obtain an affine plane. Remove the dummy vertices and test every block of the affine plane. Call a block positive if its test outcome is positive.

Step 2. Use Damaschke's algorithm to find one edge in each positive block. Remove the edge and test the remaining block. If positive, go to Step 2.

Step 3. Stop.

Note that in Step 1, we add dummy vertices and then remove them, thus the sizes of blocks vary. For example, for vertices $1, 2, \ldots, 6$, we add 7^*, 8^*, and 9^*, then apply the affine plane method of order $p = 3$ to obtain $p^2 + p = 12$ blocks: $\{1, 2, 3\}$ $\{4, 5, 6\}$ $\{7^*, 8^*, 9^*\}$ $\{1, 5, 9^*\}$ $\{1, 8^*, 6\}$ $\{1, 4, 7^*\}$ $\{4, 8^*, 3\}$ $\{4, 2, 9^*\}$ $\{2, 5, 8^*\}$ $\{7^*, 2, 6\}$ $\{7^*, 5, 3\}$ $\{3, 6, 9^*\}$. After removing the dummy vertices, we have $\{1, 5\}$ $\{1, 8\}$ $\{1, 4\}$ $\{4, 3\}$ $\{4, 2\}$ $\{2, 5\}$ $\{2, 6\}$ $\{5, 3\}$ $\{3, 6\}$ $\{1, 2, 3\}$ $\{4, 5, 6\}$, where there are 9 of them of size 2, and 2 of them of size 3 ($\{7^*, 8^*, 9^*\}$ no longer exists).

For general n and the smallst prime p such that $p^2 \geq n$, let $z \equiv n (\mathrm{mod}\, p)$, we will have

(1) $(p - z)p$ blocks of size $\lfloor n/p \rfloor$,
(2) zp blocks of size $\lceil n/p \rceil$,

(3) $\lfloor n/p \rfloor$ blocks of size p, and
(4) one block of size z (if z is nonzero)

at Step 1. For convenience, we assume that p divides n in the analysis, which means we deal with p^2 blocks of size n/p and n/p blocks of size p in our analysis. Note that we always assume n even in the gap closing problem.

It can be shown (see Appendix A) that for all n there exists a prime p such that $\lceil \log n \rceil \leq \lceil \log p^2 \rceil \leq \lceil \log n \rceil + 1$. Thus in Step 2, Damaschke's algorithm on B_i takes at most $\lceil \log \binom{p}{2} \rceil + 1 \leq \lceil \log n \rceil + 1$ rounds to identify an edge.

First we count the number of tests. The first round consumes $p(p+1)$ tests, where $p^2 \leq 2n$ for $n \geq 62$ (see Appendix A) and $2n^{1/2} \geq p \geq n^{1/2}$ (by Chebyshev's theorem, see p. 19 of Tenebaum (2000)). Afterwards, each edge requires at most $\log n + 1$ tests to be identified and an additional test to check for any other edge in the remaining block. So it takes a total of $p(p+1) + 0.5n(\log n + 2) \leq 2n + 2n^{1/2} + 0.5n \log n + n \approx 0.5n \log n$ tests for n large.

We now show that in average more than 97.5% of edges are identified in the first 3 loopings of Step 2. The probability that a random block of k vertices contains exactly i edges is

$$P(i, n, k) = \frac{\binom{n/2}{i}\binom{n/2-i}{k-2i}2^{k-2i}}{\binom{n}{k}}$$

since the number of ways of getting i edges is that the k vertices contain the i pairs of adjacent vertices (which can be chosen in $\binom{n/2}{i}$ ways), and one vertex in $k - 2i$ of the remaining $n/2 - i$ pairs. Note that

$$\frac{P(i, n, k)}{P(i-1, n, k)} = \frac{1}{i} \cdot \frac{(k-2i+1)(k-2i+2)}{4(n/2-k+i)}.$$

For the p^2 blocks of size n/p, by Appendix B,

$$\frac{P(i, n, n/p)}{P(i-1, n, n/p)} \leq \frac{1}{2i} \quad \text{for } i \geq 2.$$

Let P_i denote $P(i, n, n/p)$ and define $\alpha_i = 2iP_i/P_{i-1}$, then $0 \leq \alpha_i \leq 1$ for $i \geq 2$ because $P_i/P_{i-1} \leq 1/2i$. Further,

$$\begin{aligned}
&\frac{\text{expected number of edges internal to the p}^2 \text{ subsets of size n/p identified in the first 3 loopings}}{\text{expected number of edges internal to the p}^2 \text{ subsets of size n/p}} \\
&= \frac{1P_1 + 2P_2 + 3P_3 + 3P_4 + 3P_5 + \cdots}{1P_1 + 2P_2 + 3P_3 + 4P_4 + 5P_5 + \cdots} \\
&\geq \frac{1P_1 + 2P_2 + 3P_3 + 3P_4 + 3P_5}{1P_1 + 2P_2 + 3P_3 + 4P_4 + 5P_5 + \cdots} \\
&= \frac{1P_1 + 2P_1 \cdot \frac{1}{4}\alpha_2 + 3P_1 \cdot \frac{1}{24}\alpha_2\alpha_3 + 3P_1 \cdot \frac{1}{192}\alpha_2\alpha_3\alpha_4 + 3P_1 \cdot \frac{1}{1920}\alpha_2\alpha_3\alpha_4\alpha_5}{1P_1 + 2P_1 \cdot \frac{1}{4}\alpha_2 + 3P_1 \cdot \frac{1}{24}\alpha_2\alpha_3 + 4P_1 \cdot \frac{1}{192}\alpha_2\alpha_3\alpha_4 + 5P_1 \cdot \frac{1}{1920}\alpha_2\alpha_3\alpha_4\alpha_5 + \cdots}
\end{aligned}$$

$$\geq \frac{1+2\cdot\frac{1}{4}+3\cdot\frac{1}{24}+3\cdot\frac{1}{192}+3\cdot\frac{1}{1920}}{1+2\cdot\frac{1}{4}+3\cdot\frac{1}{24}+4\cdot\frac{1}{192}+5\cdot\frac{1}{1920}+6\cdot\frac{1}{23040}+\cdots}$$

$$\geq \frac{1+\frac{1}{2}+\frac{1}{8}+\frac{1}{64}+\frac{1}{640}}{1+\frac{1}{2}+\frac{1}{8}+\frac{1}{48}+\frac{1}{384}+\frac{1}{3840}+\frac{1}{3840}\cdot\frac{1}{12^1}+\frac{1}{3840}\cdot\frac{1}{12^2}+\cdots}$$

$$\geq \frac{6306}{6331.091} \approx 0.996037$$

For the n/p blocks of size p, by Appendix B,

$$\frac{P(i,n,p)}{P(i-1,n,p)} \leq \frac{1}{i} \quad \text{for} \begin{cases} i=2, n \neq 50, 52 \\ i \geq 3 \end{cases}.$$

By a similar method, we conclude that averagely more than 97.6% of edges internal to the n/p blocks of size p are identified in the first 3 loopings (and more than 97.5% for $n = 50$ and 52). Thus we conclude that averagely more than 97.5% of edges are identified in the first 3 loopings of Step 2. Actually, in our simulation results (for $n \leq 10000$), more than 99.7% of edges are identified in the first 3 loopings (see figure 1), which means at most $3\log n + 7$ rounds.

Another question we may consider is the maximum number of loopings of Step 2. We give a rough estimation as an upper bound from two observations: (1) By a similar method as above, we can prove that averagely more than 78% of undetected edges are detected in each looping for the edges internal to the p^2 blocks of size n/p (see Appendix 5). (2) The expected number of edges internal to the n/p blocks of size p is no more than $p/2$ (Beigel et al., 2001), where $(p/2)/(n/2) = p/n \to 0$ as $n \to \infty$. Thus we just omit the edges internal to the n/p blocks of size p and take logarithm with base $1/(1-78\%)$, that is, $-\log_{0.22}(n/2)$ to count the number of loopings in the p^2 blocks. Our simulation results show that this estimation works well as an upper bound for $n \leq 10000$ (see figure 2).

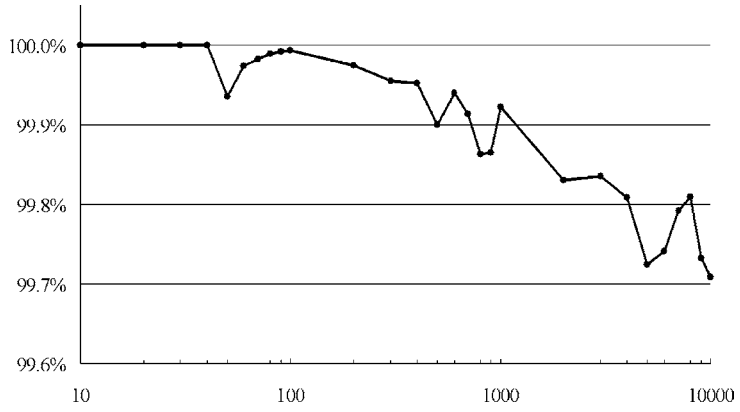

Figure 1. Simulation results for the percentage of edges identified in the first 3 loopings of Step 2, every point of data is computed from 10000 simulations.

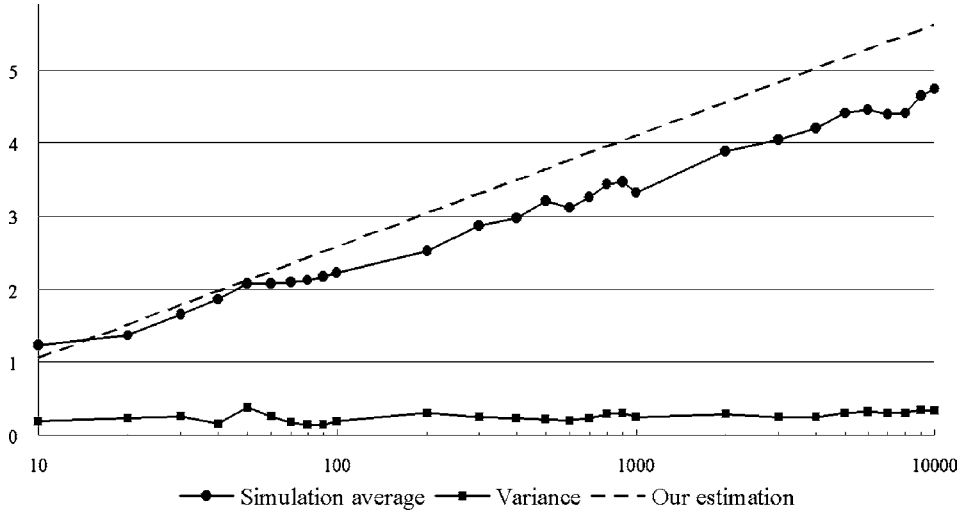

Figure 2. Simulation results for the average number of loopings of Step 2, every point of data is computed from 30000 simulations.

5. Conclusions

Three criteria have been used in the literature (Beigel et al., 2001; Grebinski and Kucherov, 1998; Lindstrom, 1969) to evaluate an algorithm:

1. Number of tests. This represents the total cost.
2. Number of rounds. All tests in a round can be tested simultaneously. So the number of rounds represents the total time required.
3. Number of pipetting operation. This is relevant as long as hand pipetting, a significant source of errors, is used (we will not be concerned with this criterion in this paper since robotic pipetting is expected to prevail).

We propose the incremental model which seems to be quite practical for the gap closing problem. We give a $n \log n$-test nonadaptive algorithm which matches the best sequential algorithm under the classical model (no nonadaptive algorithm is known within a factor of 2). We also reduce the number of tests by 3.5 times for the best sequential algorithm and 4 times for the best nonadaptive algorithm for the quantitative model.

The $n \log n$ tests of the incremental model compares favorably with the currently best $24n$-test nonadaptive algorithm under the quantitative model. For example

$$n \log n \leq 24n \quad \text{for } n \leq 2^{24} \approx 16000000.$$

However, it does not compare well with the $6n$-test nonadaptive algorithm proposed here unless $n \leq 2^6$.

Actually, the number of contigs is expected to be small if the cutting stage is designed to cover the target sequence well. For instance the example of "bacillus subtilis" quoted in Grebinski and Kucherov (1998) has 64 primers, and the example used throughout in Tettelin et al. (1996) has 48 primers. So our incremental model can be competitive sometimes with respect to the number of tests. But the more important thing is that its result is more reliable since it assumes much less than the quantitative model.

Another contribution of this paper is giving an algorithm under the classical model whose test number achieves the lower bound. While this bound is also achieved in Beigel et al.(2001), our method achieves in much fewer expected number of rounds. Moreover, our simulation results and the estimation suggest that the affine plane method is a good tool to divide a large gap closing problem to many smaller ones. A smaller problem has fewer edges, thus less likely to have PCR products with similar lengths. Consequently, it is less risky to use the more powerful quantitative or incremental model.

Appendix A

By Nagura (1952), it is known that, for two consecutive prime number p and q ($p^2 \geq n > q^2$), $p \leq 1.2q < 1.2n^{1/2}$ if $q \geq 29$ (the 10th prime number). Thus we know that for $n \geq 29^2 = 841$, there always exists a prime number p such that $n \leq p^2 \leq 1.44n < 2n$. With some exhaustive search for $n \leq 840$, the following claim is true:

Claim. For n even, let p be the smallest prime such that $p^2 \geq n$. Then $p^2 \leq 2n$ for $n \geq 62$. Moreover, $\lceil \log p^2 \rceil \leq \lceil \log n \rceil + 1$ for all n positive.

Appendix B

To prove

$$\frac{P(i, n, n/p)}{P(i-1, n, n/p)} \leq \frac{1}{2i} \quad \text{for } i \geq 2,$$

it is sufficient to prove

$$\frac{(k-2i+1)(k-2i+2)}{2(n/2-k+i)} \leq 1 \quad \text{for } k = n/p.$$

The inequality is equal to $(k^2 - n) + (-4i + 5)k + (4i^2 - 8i + 2) \leq 0$, where $k^2 - n = (n/p^2)n - n \leq 0$. For $i = 2$, $(k^2 - n) + (-4i + 5)k + (4i^2 - 8i + 2) \leq -3k + 2 \leq 0$. For $i \geq 3$, $(k^2 - n) + (-4i + 5)k + (4i^2 - 8i + 2) \leq (-2ik + 4i^2) + (5 - 2i)k + (-8i + 2) \leq 0$ since $k \geq 2i$ (the meaning of i is the number of edges in a block of size k). Thus we conclude that

$$\frac{P(i, n, n/p)}{P(i-1, n, n/p)} \leq \frac{1}{2i} \quad \text{for } i \geq 2.$$

By a similar method and the fact that $p^2 \le 2n$ for n even but not in {2, 10, 12, 50, 52, 54, 56, 58, 60} (by the exhaustive search in Appendix A), it is easy to prove that

$$\frac{P(i,n,p)}{P(i-1,n,p)} \le \frac{1}{i} \quad \text{for } i \ge 2 \text{ (where } k = p).$$

For n in $\{2, 10, 12, 50, 52, 54, 56, 58, 60\}$, just solve the inequality $(k^2-2n)+(-4i+7)k+(4i^2-10i+2) \le 0$ for integer i. For example, for $n = 50, k = 11$, by solving the inequality $(121-100)+(-4i+7)11+(4i^2-10i+2) \le 0$, we have $2.25 \le i \le 11.25 \Rightarrow i = 3, 4, 5$ (since $k \ge 2i$, we don't include 6, 7, ..., 11). Thus we conclude that

$$\frac{P(i,n,p)}{P(i-1,n,p)} \le \frac{1}{i} \quad \text{for} \begin{cases} i = 2, n \ne 50, 52 \\ i \ge 3 \end{cases}$$

Appendix C

In the x-th looping of Step 2, the average percentage of edges identified from remaining edges after $(x-1)$st looping is

$$\begin{aligned} &\frac{\text{Expected number of edges identified in } x\text{th looping of Step 2}}{\text{Expected number of unidentified edges after } (x-1)\text{st looping of Step 2}} \\ &= \frac{P_x + P_{x+1} + P_{x+2} + P_{x+3} + P_{x+4} + \cdots}{1P_x + 2P_{x+1} + 3P_{x+2} + 4P_{x+3} + 5P_{x+4} + \cdots} \\ &\ge \frac{\frac{1}{2^{x-1}x!} + \frac{1}{2^x(x+1)!} + \frac{1}{2^{x+1}(x+2)!} + \cdots}{\frac{1}{2^{x-1}x!} + \frac{2}{2^x(x+1)!} + \frac{3}{2^{x+1}(x+2)!} + \cdots} \end{aligned}$$

Let

$$Q(k,x) = \frac{\frac{1}{2^{x-1}x!} + \frac{1}{2^x(x+1)!} + \cdots + \frac{1}{2^{x+k-1}(x+k)!}}{\frac{1}{2^{x-1}x!} + \frac{2}{2^x(x+1)!} + \cdots + \frac{k+1}{2^{x+k-1}(x+k)!}},$$

if we treat $Q(k,x)$ as taking average from

$$\frac{1}{2^{x-1}x!} \text{ of } \frac{1}{1}, \frac{2}{2^x(x+1)!} \text{ of } \frac{1}{2}, \ldots,$$

and

$$\frac{k+1}{2^{x+k-1}(x+k)!} \text{ of } \frac{1}{k+1}$$

it is easy to observe that

$$\begin{aligned} Q(k,x+1) &= \frac{\frac{1}{2(x+1)} \times \frac{1}{2^{x-1}x!} + \frac{1}{2(x+2)} \times \frac{1}{2^x(x+1)!} + \cdots + \frac{1}{2(x+k+1)} \times \frac{1}{2^{x+k-1}(x+k)!}}{\frac{1}{2(x+1)} \times \frac{1}{2^{x-1}x!} + \frac{1}{2(x+2)} \times \frac{2}{2^x(x+1)!} + \cdots + \frac{1}{2(x+k+1)} \times \frac{k+1}{2^{x+k-1}(x+k)!}} \\ &\ge Q(k,x) \end{aligned}$$

since the larger ratios get more weights.

With the fact that $Q(k, 1) \geq 0.78$ for $k = 1, 2, \ldots$, we conclude that averagely more than 78% of unidentified edges are identified in each looping of Step 2 for edges internal to the p^2 blocks of size n/p.

References

1. R. Beigel, N. Alon, S. Kasif, M.S. Apaydin, and L. Fortnow, "An optimal procedure for gap closing in whole genome shotgun sequencing," in *Proceedings of the Fifth Annual International Conference on Computational Biology*, Montreal, Quebec, Canada, 2001, pp. 22–30.
2. L.J. Burgart, R.A. Robinson, M.J. Heller, W.W. Wilke, O.K. Iakoubova, and J.C. Cheville, "Multiplex polymerase chain reaction," *Mod. Pathol.*, vol. 5, pp. 320–323, 1992.
3. P. Damaschke, "A tight upper bound for group testing in graphs," *Discrete Applied Math.*, vol. 48, pp. 101–109, 1994.
4. D.-Z. Du and F.K. Hwang, *Combinatorial Group Testing and its Applications, Series on Applied Mathematics*, 2nd ed. World Scientific: Singapore, 2000.
5. V. Grebinski and G. Kucherov, "Reconstructing a Hamiltonian cycle by querying the graph: Application to DNA physical mapping," *Discrete Applied Math.*, vol. 88, pp. 147–165, 1998.
6. V. Grebinski and G. Kucherov, "Optimal reconstruction of graphs under the additive model," *Algorithmica*, vol. 28, pp. 104–124, 2000.
7. M. Hell, *Combinatorial Theory*, 2nd ed. Wiler Interscience: New York, 1996.
8. P. Johann, "A group testing problem for graphs with several defective edges," *Discrete Applied Math.*, vol. 117, pp. 99–108, 2002.
9. B. Lindstrom, "On a combinatorial problem in number theory," *Canad. Math. Bull.*, vol. 8, pp. 261–265, 1965.
10. B. Lindstrom, "Determination of two vectors from the sum," *J. Combin. Thy.*, vol. A6, pp. 402–407, 1969.
11. J. Nagura, "On the interval containing at least one prime number," *Proc. Japan Acad.*, vol. 28, pp. 177–181, 1952.
12. A. Sorokin, A. Lapidus, V. Capuano, N. Galleron, P. Pujic, and S.D. Ehrlich, "A new approach using multiplex long accurate PCR and yeast artifical chromosomes for bacterial chromosome mapping and sequencing," *Genome Res.*, vol. 6, pp. 448–453, 1996.
13. G. Tenenbaum, *The Prime Numbers and Their Distribution*. AMS: USA, 2000.
14. H. Tettelin, D. Radune, S. Kasif, H. Khouri, and S.L. Salzberg, "Optimized multiplex PCR: Efficiently closing a whole-genome shotgun sequencing project," *Genomics.*, vol. 62, pp. 500–507, 1996.