

Approximation Algorithms for Minimum PCR Primer Set Selection with Amplification Length and Uniqueness Constraints*

K. Konwar I. Măndoiu A. Russell A. Shvartsman

University of Connecticut
Department of Computer Science & Engineering
371 Fairfield Rd., Unit 2155, Storrs, CT 06269-2155, USA
E-mail: {kishori,ion,acr,aas}@cse.uconn.edu

June 2, 2006

Abstract

A critical problem in the emerging high-throughput genotyping protocols is to minimize the number of polymerase chain reaction (PCR) primers required to amplify the single nucleotide polymorphism loci of interest. In this paper we study PCR primer set selection with amplification length and uniqueness constraints from both theoretical and practical perspectives. We give a greedy algorithm that achieves a logarithmic approximation factor for the problem of minimizing the number of primers subject to a given upperbound on the length of PCR amplification products. We also give, using randomized rounding, the first non-trivial approximation algorithm for a version of the problem that requires unique amplification of each amplification target. Empirical results on randomly generated testcases as well as testcases extracted from the from the National Center for Biotechnology Information's genomic databases show that our algorithms are highly scalable and produce better results compared to previous heuristics.

1 Introduction

Availability of full genome data combined with rapid advances in high-throughput genomic technologies promises to revolutionize medical science by enabling large scale genomic analyses such as association studies between Single Nucleotide Polymorphisms (SNPs) and susceptibility to common diseases. Although recent work [3] suggests that there are only a few hundred thousand “blocks” of SNPs that recombine to provide most of the genetic variability seen in human

*Research supported in part by a Large Grant from the University of Connecticut's Research Foundation.

populations, meaningful SNP association studies will still require genotyping many thousands of SNPs in large populations.¹ This poses a daunting challenge to current SNP genotyping protocols (see [5] for a survey). A critical step in these protocols is the cost-effective amplification of DNA sequences containing the SNP loci of interest via biochemical reactions such as the *Polymerase Chain Reaction* (PCR).

PCR cleverly exploits the DNA replication machinery to create up to millions of copies of specific DNA fragments (amplification targets). In its basic form, PCR requires a pair of oligonucleotides (short single-stranded DNA sequences called *primers*) for each amplification target. More precisely, the two primers must be (perfect or near perfect) reversed Watson-Crick complements of the 3' ends of the forward and reverse strands in the double-stranded amplification target (see Figure 1).

Typically there is significant freedom in selecting the exact ends of an amplification target, i.e., in selecting PCR primers. Consequently, primer selection can be optimized with respect to various criteria affecting reaction efficiency, such as primer length, specificity, melting temperature, secondary structure, etc. Since the efficiency of PCR amplification falls off exponentially as the length of the amplification product increases, an important practical constraint is that the binding sites for the two primers must be within a certain maximum distance of each other (typically around 1000 bases).

Much of the previous work on PCR primer selection has focused on single primer pair optimization with respect to the above biochemical criteria. This line of work has resulted in the release of several robust software tools for primer pair selection, the best known of which is the Primer 3 package [9]. Another optimization objective studied in the literature is the minimization of the number of PCR primers required to carry out a given set of independent amplifications. Pearson et al. [8] were the first to consider this objective in their optimal primer cover problem formulation: given a set of DNA sequences and an integer k , find the minimum number of k -mers that cover all sequences. They showed that the primer cover problem is as hard to approximate as set cover, and hence unlikely to be approximable within a factor better than $(1 - o(1))O(\log n)$, where n is the number of DNA sequences. Pearson et al. also proposed an exact branch-and-bound algorithm for the primer cover problem and showed that the classical greedy set cover algorithm guarantees a theoretically optimum $O(\log n)$ approximation factor.

Multiplex PCR (MP-PCR) is a variation of PCR in which multiple DNA fragments are amplified simultaneously. Like the basic PCR, MP-PCR makes use of two oligonucleotide primers to define the boundaries of each amplification target. Note, however, that MP-PCR amplified targets are available only as a mixture and it may not be possible or cost-effective to separate them to the purity required, e.g., in microarray spotting. Fortunately, this is not limiting the applicability of MP-PCR to SNP genotyping, since most of the existing allelic discrimination methods are highly-parallel and thus can be applied directly to mixtures of amplified SNP loci [5]. Furthermore, effectiveness of allelic dis-

¹For example, fully powered haplotype association studies are estimated to require as much as 300,000 to 1,000,000 “haplotype-tag” SNPs [3].

crimination methods is largely unaffected by the presence of a small number of undesired amplification products, which may occur in MP-PCR.

A promising approach to further increasing MP-PCR efficiency is the use of *degenerate PCR primers* [6].² A degenerate primer is essentially a mixture consisting of multiple non-degenerate primers sharing a common pattern and can thus be used to simultaneously amplify many different SNP loci. For example, letting N to denote a position in the primer sequence where all 4 nucleotides can appear in equal proportions, the degenerate primer $aNgNc$ represents a mixture of 16 different non-degenerate primers ($aagac, aagcc, aaggc, aagtc, \dots, atgtc$). Remarkably, degenerate primer cost is nearly identical to that of non-degenerate primers, since the synthesis requires the same number of steps (the only difference is that one must add multiple nucleotides in some of the synthesis steps). However, since not all non-degenerate primers present in the degenerate primer mixture are useful, it is important to use only degenerate primers with bounded degeneracy. Linhart and Shamir [7] proved the NP-hardness of several formulations for the degenerate primer design problem, including a formulation which asks for a degenerate primer with minimum degeneracy that covers a given set of input strings. Souvenir et al. [11] proposed an iterative beam-search heuristic for the related *multiple degenerate primer design* problem, which seeks a minimum number of degenerate primers, each with bounded degeneracy, covering a given set of DNA sequences.³

A common feature of the string covering formulations in [8, 7, 11] is that they decouple the selection of forward and reverse primers, and, in particular, cannot explicitly enforce bounds on PCR amplification length. Such bounds can be enforced only by conservatively defining the allowable primer binding regions (i.e., the DNA segments to be covered). For example, in order to guarantee a distance of L between the forward and reverse primer binding sites around a SNP, one may confine the search to primers binding within $L/2$ nucleotides of the SNP locus. However, since this constraint reduces the number of candidate primer pairs by a factor of about 2,⁴ adopting this approach can lead to significant sub-optimality in the number of primers required to amplify all SNP loci.

Motivated by the requirement of unique PCR amplification in synthesis of spotted microarrays, Fernandes and Skiena [2] introduced an elegant *minimum multi-colored subgraph* formulation for the primer selection problem. In this formulation, each candidate primer is represented as a graph node and every two primers that uniquely amplify a desired target (e.g., gene) are connected by an edge labeled (or “colored”) by the target. The goal is to find a minimum subset of the nodes inducing edges of all possible colors. Fernandes and Skiena gave practical greedy and densest-subgraph based heuristics for the minimum

²Another approach is to use PCR primers that complement interspersed repetitive sequences, such as the human *Alu* sequence. Since the position of the interspersed repetitive sequences highly constrains the set of SNP loci that can be amplified, this approach is generally not applicable when a specific set of SNPs is targeted.

³The iterative beam-search heuristic of [11] is also applicable when a threshold is given for the *total-degeneracy* of the set of primers rather than individual degeneracies.

⁴E.g., assuming that all DNA k -mers can be used as primers, out of the $(L - k + 1)(L - k + 2)/2$ pairs of forward and reverse k -mers that can feasibly amplify a SNP locus, only $(L - k + 1)^2/4$ have both k -mers within $L/2$ bases of this locus.

multi-colored subgraph and showed that the problem cannot be approximated within a factor better than $(1 - o(1)) \ln n - o(1)$, where n is the number of amplification targets. While finding a minimum primer set that amplifies a given set of SNPs subject to amplification length constraints can be reduced to the minimum multi-colored subgraph problem, no non-trivial approximation factor is known for the latter problem once unique amplification is no longer required. With unique amplification constraints, the trivial algorithm of selecting two arbitrary primers for each of the n amplification target gives an approximation factor of \sqrt{n} .

In this paper we study (degenerate and non-degenerate) PCR primer selection problems with amplification length and uniqueness constraints from both theoretical and practical perspectives. Our contributions are as follows:

- We give a new string-pair covering formulation for the minimum primer set selection with amplification length constraints problem, and show that a clever modification of the classical greedy algorithm for set cover achieves a near-optimal approximation factor of $\ln(nL)$, where n is the number of amplification targets and L is the upperbound on PCR amplification length. This result is complemented by a $O(\ln n)$ inapproximability result, which implies that the approximation factor of the greedy algorithm is optimal up to an additive term of $O(\ln L)$
- We give a randomized rounding algorithm with an approximation factor of $O(\sqrt{\mathfrak{m}} \log m)$ for the minimum multi-colored subgraph problem of [2], where \mathfrak{m} is the maximum size of a color class (i.e., the maximum number of edges sharing the same color) and m is the number of colors. For the minimum primer set selection with uniqueness constraints $\mathfrak{m} = O(L^2)$ and $m = n$. Hence, our result implies an approximation factor of $O(L \log n)$, which asymptotically improves over the trivial approximation bound of \sqrt{n} . Furthermore, our algorithm has the same approximation guarantees for the minimum multi-colored subgraph problem without uniqueness requirements.
- We give the results of a comprehensive experimental study comparing our greedy approximation algorithm with previously published primer selection algorithms on randomly generated testcases as well as testcases extracted from the National Center for Biotechnology Information’s genomic databases [1].

The rest of the paper is organized as follows. In next section we introduce notations and give formal problem definitions. In Section 3 we describe and analyze the greedy algorithm for the minimum primer set selection with amplification length constraints problem. In Section 4 we give the randomized rounding algorithm for the minimum multi-colored subgraph problem. Finally, we present experimental results in Section 5 and conclude with some open problems in Section 6.

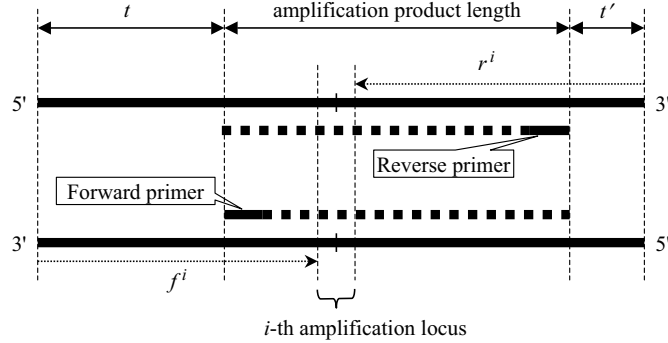


Figure 1: Strings f^i and r^i consist of the L DNA bases immediately preceding in $3' - 5'$ order the i -th amplification locus along the forward (respectively reverse) genomic sequence. If forward and reverse PCR primers cover f^i and r^i at positions t , respectively t' , then the PCR amplification product length is $(2L+x) - (t+t')$, where x is the length of the amplification locus ($x = 1$ for SNP genotyping). Thus, amplification product length is at most $L + x$ iff $t + t' \geq L$.

2 Notations and Problem Formulations

Let $\Sigma = \{a, c, g, t\}$ be the DNA alphabet. We denote by Σ^* the set of strings over Σ , and by λ the empty string. Overloading notations, we use $|\cdot|$ to denote both the length of strings over Σ and the size of sets. For a string s and an integer $t < |s|$, we denote by $s[1..t]$ the prefix of length t of s .

Following [11], we define a *non-degenerate primer of length k* as a string from Σ^k . A *degenerate nucleotide* is a non-empty subset of Σ . A *degenerate primer of length k* , or simply a *primer of length k* , is a string $d_1d_2 \dots d_k$ of degenerate nucleotides, and can equivalently be viewed as the set of non-degenerate primers $x_1x_2 \dots x_k$, $x_i \in d_i$. The *degeneracy* of a degenerate primer $d_1d_2 \dots d_k$ is the number of non-degenerate primers it represents, i.e., $\prod_{i=1}^k |d_i|$.

We denote by L the given threshold on the PCR amplification length, and by f^i (respectively r^i) the string consisting of the L DNA bases immediately preceding in $3' - 5'$ order the i -th amplification locus along the forward (respectively reverse) DNA genomic sequence (see Figure 1).

We say that degenerate primer $p = d_1d_2 \dots d_k$ *covers* (or *hybridizes at*) position i of string $s = s_1s_2 \dots s_m$ iff i is the largest index such that $s_i s_{i+1} \dots s_{i+k-1}$ is the reversed Watson-Crick complement of one of the non-degenerate primers represented by p , i.e., iff s_{i+j} is the Watson-Crick complement of one of the nucleotides in d_{k-j} for every $0 \leq j \leq k - 1$.⁵

A set of degenerate primers P is an *L -restricted primer cover* for the pairs of sequences $(f^i, r^i) \in \Sigma^L \times \Sigma^L$, $i = 1, \dots, n$, iff for every $i = 1, \dots, n$, there exist primers $p, p' \in P$, not necessarily distinct, and integers $t, t' \in \{1, \dots, L\}$,

⁵In practice, stable primer hybridization and subsequent PCR amplification occur even with a small number of mismatches if none of them is too close to the $3'$ end of the primer. Our algorithms apply unmodified to hybridization models allowing mismatches.

such that

1. p hybridizes at position t of f^i ;
2. p' hybridizes at position t' of r^i ; and
3. $t + t' \geq L$

The last constraint ensures that the PCR amplification product length is no more than $L + x$, where x is the length of the desired amplification target ($x = 1$ for SNP genotyping). We say that a primer cover has the *unique amplification property* if, for each pair (f^i, r^i) , there exists exactly one set of primers $\{p, p'\} \in P$ satisfying conditions 1-3 above.

The *minimum primer set selection problem with amplification length constraints* (MPSS-L) is defined as follows: Given primer length k , degeneracy upperbound δ , amplification length upperbound L , and n pairs of sequences (f^i, r^i) , $i = 1, \dots, n$, find a minimum size L -restricted primer cover consisting of degenerate primers of length k , each with degeneracy at most δ . The *minimum primer set selection problem with amplification length and uniqueness constraints* (MPSS-LU) is defined in the same way except that in this case we seek a minimum size L -restricted primer cover which has the unique amplification property.

3 The Greedy Algorithm for MPSS-L

MPSS-L can be viewed as a generalization of the partial set cover problem [10]. In the partial set cover problem one must cover with the minimum number of sets a given fraction of the total number of elements. In MPSS-L we can take the elements to be covered to be the non-empty prefixes of the $2n$ forward and reverse sequences; there are $2nL$ such elements. A primer p covers prefix $f^i[1..j]$ ($r^i[1..j]$) if it hybridizes to f^i (respectively r^i) at position $t \geq j$. The objective is to cover at least L (i.e., half) of the elements of $\{f^i[1..j], r^i[1..j] \mid 1 \leq j \leq L\}$ for every $i \in \{1, \dots, n\}$.

For a set of primers P , let \bar{f}^i and \bar{r}^i denote the longest prefix of f^i , respectively r^i , covered by a primer in P . Note that $|\bar{f}^i| + |\bar{r}^i|$ gives the number of elements of $\{f^i[1..j], r^i[1..j] \mid 1 \leq j \leq L\}$ that are covered by P . Let $\Phi(P) := \min\{L, |\bar{f}^i| + |\bar{r}^i|\}$. Note that $\Phi(\emptyset) = 0$, $\Phi(P) = nL$ for every feasible MPSS-L solution, and that $\Phi(P) \leq \Phi(P')$ whenever $P \subseteq P'$. Hence, $\Phi(P)$ can be used as a measure of the progress made towards feasibility by a set P of primers.

The greedy algorithm (see Figure 2) starts with an empty set of primers and iteratively selects primers which give the largest increase in Φ until reaching feasibility.

Theorem 1 *The greedy algorithm returns an L -restricted primer cover of size at most $\ln(nL)$ times larger than the optimum.*

Input: Primer length k , degeneracy upperbound δ , amplification length upperbound L , and pairs of sequences $(f^i, r^i) \in \Sigma^L \times \Sigma^L$, $i = 1, \dots, n$
Output: L -restricted primer cover P consisting of degenerate primers of length k , each with degeneracy at most δ

Function $\Delta(p, i)$:

$\Delta \leftarrow 0$

If $|\bar{f}^i| + |\bar{r}^i| \geq L$ return 0

If p covers f^i at position $t > |\bar{f}^i|$, $\Delta \leftarrow \Delta + (t - |\bar{f}^i|)$

If p covers r^i at position $t > |\bar{r}^i|$, $\Delta \leftarrow \Delta + (t - |\bar{r}^i|)$

Return $\min\{\Delta, L - (|\bar{f}^i| + |\bar{r}^i|)\}$

$P \leftarrow \emptyset$; for every $i = 1, \dots, n$, $\bar{f}^i \leftarrow \bar{r}^i \leftarrow \lambda$

While $\Phi(P) := \sum_{i=1}^n \min\{L, |\bar{f}^i| + |\bar{r}^i|\} < mL$ do

 Find the degenerate primer p maximizing $\Delta\Phi = \sum_{i=1}^n \Delta(p, i)$

 For every $i = 1, \dots, n$,

 If p covers f^i at position $t > |\bar{f}^i|$ then $\bar{f}^i \leftarrow f^i[1..t]$

 If p covers r^i at position $t > |\bar{r}^i|$ then $\bar{r}^i \leftarrow r^i[1..t]$

$P \leftarrow P \cup \{p\}$

Return P

Figure 2: The greedy algorithm for MPSS-L

Proof. Let OPT denote a minimum size L -restricted primer cover, and let p_1, \dots, p_g be the primers selected by the greedy algorithm. It can be verified that, for every A and B , $\Phi(A \cup B) \leq \Phi(A) + \sum_{p \in B} [\Phi(A \cup \{p\}) - \Phi(A)]$. By using this claim with $A = \{p_1, \dots, p_{i-1}\}$ and $B = \text{OPT}$, it follows that in the step when the greedy algorithm selects p_i , there is a primer in $\text{OPT} \setminus \{p_1, \dots, p_{i-1}\}$ whose selection increases Φ by at least $(nL - \Phi(P)) / |\text{OPT}|$. Hence, the selection of p_i must increase Φ by at least the same amount, i.e., reduce the difference between $\Phi(\text{OPT})$ and $\Phi(P)$ by a factor of at least $(1 - 1/|\text{OPT}|)$. By induction we get that

$$nL - \Phi(\{p_1, \dots, p_i\}) \leq nL \left(1 - \frac{1}{|\text{OPT}|}\right)^i \quad (1)$$

which implies that the number of primers selected by the greedy algorithm is at most $\ln(nL)$. ■

Remark. In [8] it is proved that the following primer cover problem is as hard to approximate as set cover: Given integer k and strings s_1, \dots, s_n , find a minimum set of k -length primers covering all s_i 's. A simple approximation preserving reduction of the primer cover problem to MPSS-L shows that the MPSS-L problem cannot be approximated within a factor better than $(1 - o(1)) \ln n$ unless $\text{NP} \subseteq \text{TIME}(n^{O(\log \log n)})$. Hence, the approximation factor in Theorem 1 is tight up to an additive term of $O(\ln L)$.

4 Rounding Algorithm for the Minimum Multi-Colored Subgraph Problem

In this section we consider a graph-theoretical generalization of the MPSS-LU problem. The *minimum multi-colored subgraph* problem [2] is defined as follows. Let $G = (V, E)$ be an undirected graph and $\chi_1, \dots, \chi_k \subset E$ a family of nonempty “color classes” of edges with the property that $\bigcup_i \chi_i = E$. Assigning $X = (\chi_1, \dots, \chi_k)$, let $\mathcal{I}(G, X)$ denote the minimum size of a set of vertices I for which the subgraph induced by these vertices contains at least one edge of each color. Note that $2 \leq \mathcal{I}(G, X) \leq 2|X|$ and, as an edge may belong to several distinct color classes, both of these extreme values are in fact possible.

The problem of computing $\mathcal{I}(G, X)$ is NP-hard, via, e.g., a natural reduction from SET-COVER. We show below that it can be approximated to within $O(\sqrt{\max_{\chi \in X} |\chi|} \log |X|)$ in polynomial time.

Theorem 2 $\mathcal{I}(G, X)$ can be approximated to within $O(\sqrt{m} \log |X|)$ in polynomial time, where $m = \max_{\chi \in X} |\chi|$.

Proof. We begin with the following integer program formulation of this optimization problem

$$\begin{aligned} \min \sum_v x_v, \text{ subject to} \\ \forall \chi \in X, \sum_{e \in \chi} y_e \geq 1, \\ \forall v \in V, \forall \chi \in X, \sum_{v \in e \in \chi} y_e \leq x_v, \\ \forall e \in E, y_e \geq 0, \forall v \in V, x_v \geq 0. \end{aligned}$$

Relaxing this formulation by allowing the variables x_v and y_e to take values in $[0, 1]$ results in a linear program, the optimum value for which we denote $\mathcal{I}_\ell(G, X)$. We begin by scaling the linear program to obtain the following new

linear program:

$$\begin{aligned} \min \sum_v x_v, \text{ subject to} \\ \forall \chi \in X, \sum_{e \in \chi} y_e \geq \sqrt{\mathbf{m}} \ , \\ \forall v \in V, \forall \chi \in X, \sum_{v \in e \in \chi} y_e \leq x_v \ , \\ \forall e \in E, y_e \geq 0, \forall v \in V, x_v \geq 0 \ . \end{aligned}$$

Let $\mathcal{I}_\ell^s(G, X)$ denote the optimum value for this scaled version, and note that $\mathcal{I}_\ell^s(G, X) \leq \sqrt{\mathbf{m}} \cdot \mathcal{I}_\ell(G, X)$ by scaling any solution that achieves the value $\mathcal{I}_\ell(G, X)$ by the factor $\sqrt{\mathbf{m}}$; let $x^* \in \mathbb{R}^V$ and $y^* \in \mathbb{R}^E$ denote a feasible solution to the program above, achieving the optimum value $\mathcal{I}_\ell^s(G, X)$.

Based on the solution (x^*, y^*) above, define a family of (artificial) independent $\{0, 1\}$ -valued random variables

$$\{Z_{v,e} \mid v \in e, v \in V, e \in E\}$$

where $\Pr[Z_{v,e} = 1] = p_e \triangleq \min(y_e^*, 1)$ for each $v \in e$. In terms of these variables, define, for each $v \in V$ and each $(u, v) = e \in E$, the variables

$$X_v = \bigvee_{v \in e \in E} Z_{v,e} \quad \text{and} \quad Y_e = Z_{u,e} Z_{v,e} \ .$$

Finally, we let the variables X_u determine a random set of vertices $S = \{v \mid X_v = 1\}$. Our goal is to show that, for each color class χ , the set S is likely to induce an edge in χ .

Comment. Observe that indicator variable for the event that the set S induces the edge $e = (u, v)$ is $X_u X_v$ which dominates the variable $Y_{(u,v)}$. We focus on this second, less natural, set of variables because, unlike the variables $X_u X_v$, the $Y_{(u,v)}$ are independent.

With this in mind, note that $\Pr[Y_e = 1] = (p_e)^2$ and that for each v

$$\begin{aligned} \Pr[v \in S] &= \Pr[X_v = 1] = \left(1 - \prod_{v \in e} \Pr[Z_{v,e} = 0]\right) = \left(1 - \prod_{v \in e} (1 - p_e)\right) \\ &\leq \left(1 - \left(1 - \sum_{v \in e} p_e\right)\right) \leq \sum_{v \in e} y_e^* \leq x_v^* \ . \end{aligned}$$

Hence, by linearity of expectation

$$\text{Exp}[|S|] = \text{Exp}\left[\sum_v X_v\right] \leq \mathcal{I}_\ell^s(G, X) \leq \sqrt{\mathbf{m}} \cdot \mathcal{I}_\ell(G, X) \leq \sqrt{\mathbf{m}} \cdot \mathcal{I}(G, X) \ .$$

We wish to upper bound, for each color class χ , the quantity

$$\Pr[\forall e \in \chi, Y_e = 0] = \Pr[S \text{ induces no edge from } \chi]$$

with the intention of showing that this selection S of vertices is likely to induce many color classes. So, consider now an arbitrary color class χ ; then

$$\text{Exp} \left[\sum_{e \in \chi} X_u X_v \right] \geq \text{Exp} \left[\sum_{e \in \chi} Y_e \right] = \sum_{e \in \chi} p_e^2 \geq |\chi| \cdot \left(\frac{\sqrt{\mathfrak{m}}}{|\chi|} \right)^2 \geq 1 ,$$

as $\sum_{e \in \chi} p_e \geq \sqrt{\mathfrak{m}}$ and the function $x \mapsto x^2$ is convex. Considering that the Y_e are independent, we compute

$$\begin{aligned} \Pr[\chi \text{ not induced by } S] &= \Pr[\forall (u, v) \in \chi, X_u X_v = 0] \leq \Pr[\forall e \in \chi, Y_e = 0] \\ &= \prod_{e \in \chi} (1 - p_e^2) \leq \prod_{e \in \chi} e^{-p_e^2} = e^{-\sum_{e \in \chi} p_e^2} \geq e^{-1} . \end{aligned}$$

Evidently, selection of S as above “covers” any individual class χ with constant probability. So, finally, consider the set of vertices obtained by (i.) repeating the above procedure $t = (\log |X| + 2)$ times, resulting in the vertex sets S_1, \dots, S_t followed by (ii.) forming the union $S = \bigcup_i S_i$. Then

$$\text{Exp}[|S|] \leq \sqrt{\mathfrak{m}}(\log |X| + 2) \cdot \mathcal{I}(G, X)$$

so that by Markov’s inequality, the probability that $|S|$ exceeds this value by a factor 3 is no more than $1/3$. In addition, the probability that S fails to induce an edge in all of the color classes is

$$\Pr[\exists \chi \in X, \text{ no edge of } \chi \text{ induced by } S] \leq |X| \cdot (e^{-1})^{\log |X| + 2} = e^{-2} \leq 1/3 .$$

Hence with constant probability this procedure results in a collection of vertices that induces at least one edge of each color class and has cardinality no more than $O(\sqrt{\mathfrak{m}} \log |X| \mathcal{I}(G, X))$, as desired. \blacksquare

We show below that the integrality gap of the LP defining $\mathcal{I}_\ell(G, X)$ is $\Omega(\sqrt{\mathfrak{m}})$ in general. This suggests that this particular LP formulation may have limited value in achieving approximation results beyond the $\sqrt{\mathfrak{m}}$ threshold.

Theorem 3 *For every $s \geq 0$ there is a pair (G, X) for which $\mathfrak{m} = s$ and $\mathcal{I}(G, X) \geq \Omega(\sqrt{\mathfrak{m}}) \mathcal{I}_\ell(G, X)$.*

Proof. Consider the graph on $n \gg s$ vertices obtained by selecting, independently and uniformly at random, n matchings χ_1, \dots, χ_n each of size s and assigning $E = \bigcup \chi_i$. Observe that the feasible solution obtained by setting $x_v = y_e = 1/s$ for all e and v implies that $\mathcal{I}_\ell(G, X) \leq n/s$.

On the other hand, we show that with high probability, this random selection of matchings results in a graph for which the smallest integer solution has objective value at least $\ell \triangleq (n-1)/\sqrt{2s}$. Specifically, let $L \subset V$ be a fixed collection of ℓ vertices and note that the probability that any given edge induced by L is included in, e.g., χ_1 is $s/\binom{n}{2}$; hence the probability that L induces an edge of each color is no more than

$$\left(\frac{s}{\binom{n}{2}} \binom{\ell}{2} \right)^m \leq \left(\frac{s\ell^2}{(n-1)^2} \right)^m \leq \left(\frac{1}{2} \right)^m .$$

Hence the probability that some set of ℓ vertices induces an edge of each color is no more than $\binom{n}{\ell}2^{-m} < 1$ for $m \geq n$. Evidently, there exists a family of color classes $X = (\chi_1, \dots, \chi_m)$ for which $\mathcal{I}(G, X) \geq \Theta(\sqrt{m})\mathcal{I}_\ell(G, X)$, as desired. ■

5 Experimental Results

We performed experiments on both randomly generated MPSS-L instances and instances extracted from the human genome databases. Random DNA sequences were generated from the uniform distribution induced by assigning equal probabilities for each nucleotide. The DNA sequences consisted of regions surrounding 100 known SNPs collected from National Center for Biotechnology Information’s genomic databases [1].

For all experiments we used a bound $L = 1000$ on the PCR amplification length. In all experiments we considered only non-degenerate primers ($\delta = 1$) with length k between 8 and 12. These values model the restricted degenerate primer format suggested and experimentally validated by Jordan et al. [4]. In this format, 8-12 nucleotides at the 3’ end of each primer are fully specified, followed by a middle sequence of up to 6 fully degenerate nucleotides, followed by a fixed GC-rich sequence (CTCGAG in [4]) at the 5’ end.

We compared the following four algorithms:

- The greedy primer cover algorithm of [8] (G-FIX). In this algorithm the candidate primers are collected from the reverse and forward sequences within a distance of $L/2$ around the SNP. This ensures that our final solution is a set of primers that meets the product length constraints. The algorithm repeatedly selects the candidate primer that covers the maximum number of not yet covered forward and reverse sequences.
- A naïve modification of G-FIX, which we call G-VAR, in which the candidate primers are initially collected from the reverse and forward sequences within a distance of L around the SNP. The algorithm proceeds by greedily selecting primers like G-FIX, except that after a first primer p covers one of the forward or reverse sequences corresponding to a SNP at position t , we truncate the opposite sequence to a length of $L - t$, thus ensuring that the final primer cover is L -restricted.
- The greedy approximation algorithm from Figure 2, called G-POT since it makes greedy choices based on the “potential function” Φ .
- The iterative beam-search heuristic of Souvenir et al. [11]. We used the primer-threshold version of this heuristic, MIPS-PT, with degeneracy bound set to 1 and the default beam size of 100.

Table 1 gives the number of primers selected and the running time (in CPU seconds) for the three greedy algorithms and for the iterative beam-search MIPS-PT heuristic of [11] on instances extracted from the NCBI repository. G-POT has the best performance on all testcases, reducing the number of primers by up to 24% compared to G-FIX and up to 30% compared to G-VAR. G-VAR

# SNPs	k	G-FIX		G-VAR		MIPS-PT		G-POT	
		#Primers	CPU sec.	#Primers	CPU sec.	#Primers	CPU sec.	#Primers	CPU sec.
50	8	13	0.13	15	0.30	21	48	10	0.32
50	10	23	0.22	24	0.36	30	150	18	0.33
50	12	31	0.14	32	0.30	41	246	29	0.28
100	8	17	0.49	20	0.89	32	226	14	0.58
100	10	37	0.37	37	0.72	50	844	31	0.75
100	12	53	0.59	48	0.84	75	2601	42	0.61

Table 1: Results on instances extracted from NCBI repository ($L = 1000$).

performance is neither dominated nor dominating that of G-FIX. On the other hand, the much slower MIPS-PT heuristic has the poorest performance, possibly because is fine-tuned to perform well with higher degeneracy primers.

To further characterize the performance of compared algorithms, in Figure 3(a-c) we plot the average solution quality of the three greedy algorithms versus the number of target SNPs (on a log scale) for randomly generated testcases. MIPS was not included in this comparison due to its prohibitive running time. In order to facilitate comparisons across instance sizes, the size of the primer cover is normalized by the double of the number of SNPs, which is the size of the trivial cover obtained by using two distinct primers to amplify each SNP. Although the improvement is highly dependent on primer length and number of SNPs, G-POT is still consistently outperforming the G-FIX algorithm of [8], and, with few exceptions, its G-VAR modification.

Figure 3(d) gives the log-log plot of the average CPU running time (in seconds) versus the number of pairs of sequences for primers of size 10 and randomly generated pairs of sequences. All experiments were run on a PowerEdge 2600 Linux server with 4 Gb of RAM and dual 2.8 GHz Intel Xeon CPUs – only one of which is used by our sequential algorithms – using the same compiler optimization options. The runtime of all three greedy algorithms grows linearly with the number of SNPs, with G-VAR and G-POT incurring only a small factor penalty in runtime compared to G-FIX. This suggests that a robust practical heuristic is to run all three algorithms and return the best of the three solutions found.

6 Open Problems

While the logarithmic approximation factor achieved by our greedy algorithm for PCR primer set selection with an amplification length constraint of L is optimal within an additive factor of $O(\ln L)$, the gap between the $O(\ln n)$ inapproximability bound established in [2] and the approximation factor of $O(L \ln n)$ that we obtain for PCR primer set selection with uniqueness constraints is less satisfactory. Closing this gap, either directly or via improved approximations for the minimum multi-colored subgraph problem, is an interesting open problem.

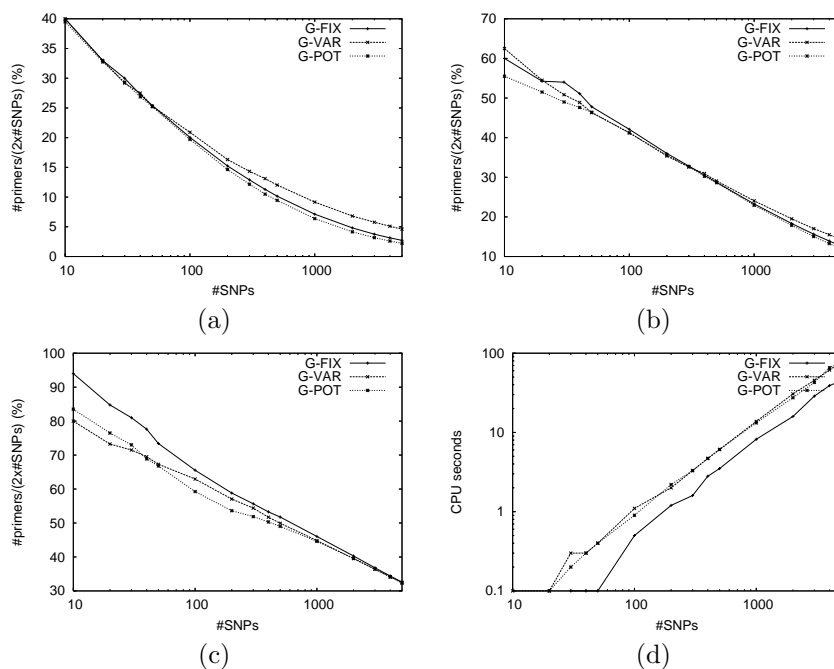


Figure 3: (a)–(c) Performance of the compared algorithms, measured by relative improvement over the trivial solution of using two primers per SNP for $k = 8, 10, 12$, $L = 1000$, and up to 5000 SNPs. (d) Runtime of the compared algorithms for $l = 10$, $L = 1000$, and up to 5000 SNPs. Each number represents the average over 10 testcases of the respective size.

References

- [1] International Human Genome Sequencing Consortium. *Homo sapiens chromosome 12 genomic contig*. National Center for Biotechnology Information, <http://www.ncbi.nlm.nih.gov>, 2004.
- [2] R.J. Fernandes and S.S. Skiena. Microarray synthesis through multiple-use PCR primer design. *Bioinformatics*, 18:S128–S135, 2002.
- [3] S.B. Gabriel, S.F. Schaffner, H. Nguyen, J.M. Moore, J. Roy, B. Blumenstiel, J. Higgins, M. DeFelice, A. Lochner, M. Faggart, S.N. Liu-Cordero, C. Rotimi, A. Adeyemo, R. Cooper, R. Ward, E.S. Lander, M.J. Daly, and D. Altshuler. The structure of haplotype blocks in the human genome. *Science*, 296:2225–2229, 2002.
- [4] B. Jordan, A. Charest, J.F. Dowd and .P. Blumenstiel, R. f. Yeh, A. Osman, D.E. Housman, and . E. Landers. Genome complexity reduction for SNP genotyping analysis. *Proc. Natl. Acad. Sci. USA.*, 99:2942–2947, 2002.
- [5] P.Y. Kwok. Methods for genotyping single nucleotide polymorphisms. *Annual Review of Genomics and Human Genetics*, 2:235–258, 2001.

- [6] S. Kwok, S.Y. Chang, J.J. Sninsky, and A. Wong. A guide to the design and use of mismatched and degenerate primers. *PCR Methods and Appl.*, 3:S539–S547, 1994.
- [7] C. Linhart and R. Shamir. The degenerate primer design problem. *Bioinformatics*, 18:S172–S181, 2002.
- [8] W.R. Pearson, G. Robins, D.E. Wrege, and T. Zhang. On the primer selection problem for polymerase chain reaction experiments. *Discrete and Applied Mathematics*, 71:231–246, 1996.
- [9] S. Rozen and H.J. Skaletsky. Primer3 on the WWW for general users and for biologist programmers. In S. Krawetz and S. Misener, editors, *Bioinformatics Methods and Protocols: Methods in Molecular Biology*, pages 365–386. Humana Press, Totowa, NJ, 2000. Code available at http://www-genome.wi.mit.edu/genome_software/other/primer3.html.
- [10] P. Slavik. Improved performance of the greedy algorithm for partial cover. *Information Processing Letters*, 64:251–254, 1997.
- [11] R. Souvenir, J. Buhler, G. Stormo, and W. Zhang. Selecting degenerate multiplex PCR primers. In *Proc. 3rd Intl. Workshop on Algorithms in Bioinformatics (WABI)*, pages 512–526, 2003.