IMPROVED ALGORITHMS FOR MULTIPLEX PCR PRIMER SET SELECTION WITH AMPLIFICATION LENGTH CONSTRAINTS*

K.M. KONWAR, I.I. MĂNDOIU, A.C. RUSSELL, AND A.A. SHVARTSMAN

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

Numerous high-throughput genomics assays require the amplification of a large number of genomic loci of interest. Amplification is cost-effectively achieved using several short single-stranded DNA sequences called primers and polymerase enzyme in a reaction called multiplex polymerase chain reaction (MP-PCR). Amplification of each locus requires that two of the primers bind to the forward and reverse DNA strands flanking the locus. Since the efficiency of PCR amplification falls off exponentially as the length of the amplification product increases, an important practical requirement is that the distance between the binding sites of the two primers should not exceed a certain threshold. In this paper we study MP-PCR primer set selection with amplification length constraints from both theoretical and practical perspectives. Our contributions include an improved analysis of a simple yet effective greedy algorithm for the problem, and a comprehensive experimental study comparing our greedy algorithm with other published heuristics on both synthetic and genomic database test cases.

1. Introduction

Numerous high-throughput genomics assays require rapid and cost-effective amplification of a large number of genomic loci. Most significantly, Single Nucleotide Polymorphism (SNP) genotyping protocols require the amplification of up to thousands of SNP loci of interest.[12] Effective amplification can be achieved using the polymerase chain reaction [16] (PCR), which cleverly exploits the DNA replication machinery in a cyclic reaction that creates an exponential number of copies of specific DNA fragments.

In its basic form, PCR requires a pair of 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 of 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, 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 requirement is that the distance between the binding sites of the two primers should not exceed a certain threshold.

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

Multiplex PCR (MP-PCR) is a variant of PCR in which multiple DNA fragments are amplified simultaneously. While MP-PCR is still making use of two oligonucleotide primers to define the boundaries of each amplification fragment, a primer may now participate in the amplification of multiple targets. A primer set is feasible as long as it contains a pair of primers that amplify each target. Note that MP-PCR amplified targets may include unintended amplification products and are available only as a mixture. However, this is not limiting the use of MP-PCR in applications such as SNP genotyping, since allelic discrimination methods (typically hybridization based) are not significantly affected by the presence of a small number of undesired amplification products, and can be applied directly to mixtures of amplified SNP loci.[12]

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 Primer3 package.[19] In the context of multiplex PCR, an important optimization objective is to minimize the total number of primers,[4,17] since reducing the number of primers reduces assay cost, increases amplification efficiency by enabling higher effective concentration of the primers, and minimizes unintended amplification. Pearson et al.[18] were the first to consider minimizing the number of primers in their optimal primer cover problem formulation: given a set of n DNA sequences and an integer ℓ, find the minimum number of ℓ-mers that cover all sequences. They proved that the primer cover problem is as hard to approximate as set cover (i.e., not approximable within a factor better than $(1 - o(1))O(\log n)$ unless NP $\subseteq$ TIME$(n^{O(\log \log n)})^5$), and that the classical greedy set cover algorithm achieves an approximation factor of $O(\log n)$.

The problem formulation in Pearson et al.[18] decouples 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. For example, in order to guarantee a distance of L between the forward and reverse primer binding sites around a SNP, one should confine the search to primers binding within $L/2$ nucleotides on each side of the SNP locus. Since this approach reduces the number of feasible candidate primer pairs by a factor of almost 2,[a] it may lead to significant sub-optimality in the total number of primers needed to amplify all given SNP loci.

Motivated by the requirement of unique PCR amplification in synthesis of spotted microarrays, Fernandes and Skiena[6] introduced an elegant *minimum multi-colored subgraph* formulation for the primer selection problem, in which each candidate primer is represented as a graph node and each two primers that feasibly amplify a desired locus define an edge "colored" by the locus number. Minimizing the number of PCR primers reduces to finding a minimum subset of the nodes inducing edges of all possible colors. Unfortunately, approximating the minimum multi-colored subgraph appears to be difficult.[7] The best approximation factor derived via this reduction is currently $O(L \log n)$, where n is the number of amplification loci and L is the upperbound on the PCR amplification length.[11]

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

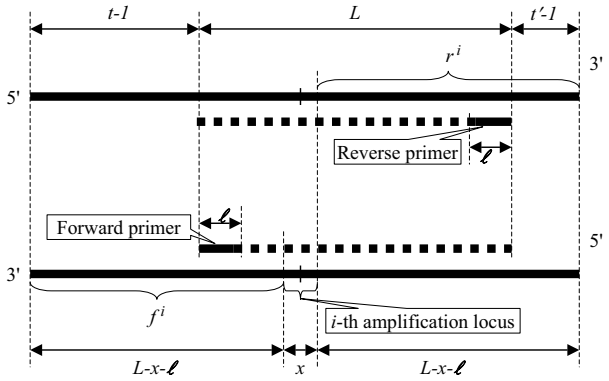


Figure 1. Strings f^i and r^i consist of the $L - x - \ell$ DNA bases immediately preceding in $3' - 5'$ order the i-th amplification locus along the forward (respectively reverse) DNA genomic sequence, where L is the given threshold on PCR amplification length, ℓ is the primer length, and x is the length of an amplification locus ($x = 1$ for SNP genotyping). If forward and reverse PCR primers cover f^i and r^i at positions t and t' respectively, then PCR amplification product length is equal to $[2(L - x - \ell) + x] - [(t - 1) + (t' - 1)]$. This is no larger than L if an only $t + t' \geq L' + 1$, where $L' = (L - x - \ell) - (\ell - 1)$.

Recently,[11] we have introduced a new string-pair covering formulation for MP-PCR primer set selection with amplification length constraints problem, proving that a modification of the classical greedy algorithm for set cover achieves an approximation factor of $1 + \ln(nL)$. In this paper we make two important contributions:

- Theoretically, we give an improved analysis of the greedy algorithm and show that it guarantees an approximation factor of $1 + \ln(\Delta)$, where Δ is the maximum "coverage gain" of a primer. The value of Δ is never more than nL, and in practice it is up to orders of magnitude smaller. The improved approximation is achieved using a novel framework for formulating and analyzing greedy algorithms based on monotonic potential functions. Our potential function technique generalizes several results for the classical set cover problem and its variants,[1,2,9,15,20] and is of interest in its own right.
- On the practical side, we give the results of a comprehensive experimental study comparing our greedy algorithm with other heuristics proposed in the literature. Experiments on both synthetic and public genomic database test cases show that our greedy algorithm obtains significant reductions in the number of primers with highly scalable running time.

The rest of the paper is organized as follows. In next section we introduce notations and give a formal problem definition. In Section 3 we describe the greedy algorithm, give its performance analysis, and discuss practical implementation issues. Finally, we present experimental results in Section 4 and conclude in Section 5.

2. Notations and Problem Formulation

Let $\Sigma = \{A, C, G, T\}$ be the four nucleotide DNA alphabet. We denote by Σ^* the set of strings over Σ, and by $|s|$ the length of string $s \in \Sigma^*$. For a string s and an integer

(1) $P \leftarrow \emptyset$
(2) While $\Phi(P) < n(L'+1)$ do
  (a) Find a primer $p \notin P$ maximizing $\Phi(P \cup \{p\}) - \Phi(P)$
  (b) $P \leftarrow P \cup \{p\}$
(3) Return P

Figure 2. The generic greedy algorithm.

$1 \leq t \leq |s|$, we denote by $s[1..t]$ the prefix of length t of s. We use ℓ to denote the required primer length, L to denote the given threshold on PCR amplification length, and n to denote the number of amplification loci. We say that primer $p = p_1 p_2 \dots p_\ell$ *hybridizes* (or *covers*) string $s = s_1 s_2 \dots s_m$ at position $t \leq m - \ell + 1$ if $s_t s_{t+1} \dots s_{t+\ell-1}$ is the reversed Watson-Crick complement of p, i.e., if s_{t+j} is the Watson-Crick complement of $p_{\ell-j}$ for every $0 \leq j \leq \ell - 1$.

For each $i \in \{1, \dots, n\}$, we denote by f^i (respectively r^i) the string preceding the amplification locus in $3' - 5'$ order in the forward (respectively reverse) DNA genomic sequence where potentially useful primer binding may occur. More precisely, if the length of the amplification locus is denoted by x ($x = 1$ for SNP genotyping), then f^i and r^i consist of the $L - x - \ell$ DNA bases immediately preceding in $3' - 5'$ order the i-th amplification locus along the forward (respectively reverse) DNA genomic sequence. Note that a primer can hybridize f^i (respectively r^i) only at positions t between 1 and L', where $L' = (L - x - \ell) - (\ell - 1)$. Simple arithmetic shows that two primers that hybridize to f^i and r^i at positions t and t' lead to an amplification product of length at most L if and only if $t + t' \geq L' + 1$ (see Figure 1, and note that f^i and r^i, and hence hybridization positions, are indexed in the respective $3' - 5'$ orders, i.e., they increase when moving towards the amplification locus).

A set of primers P is said to be an *L-restricted primer cover* for the pairs of sequences (f^i, r^i), $i = 1, \dots, n$, if, for every $i = 1, \dots, k$, there exist primers $p, p' \in P$ (not necessarily distinct) and integers $t, t' \in \{1, \dots, L - \ell + 1\}$, such that the following conditions are simultaneously satisfied

(1) p hybridizes at position t of f^i
(2) p' hybridizes at position t' of r^i
(3) $t + t' \geq L' + 1$

The *minimum primer set selection problem with amplification length constraints* (MPSS-L) is defined as follows: Given primer length ℓ, 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 primers of length ℓ.

3. The Greedy Algorithm

It is useful to view MPSS-L as a generalization of the partial set cover problem, [20] in which one must cover a certain fraction of the total number of elements of a ground set using the the minimum number of given subsets. In MPSS-L the elements to be covered are the $2nL'$ non-empty prefixes in $\{f^i[1..j], r^i[1..j] \mid 1 \leq i \leq n,\ 1 \leq j \leq L'\}$. Each primer p corresponds to the set of all prefixes $f^i[1..j]$ ($r^i[1..j]$) for which p hybridizes to

f^i (respectively r^i) at a position $t \geq j$. The objective is to choose the minimum number of primers that cover at least $L' + 1$ of the $2L'$ elements of each set $\{f^i[1..j], r^i[1..j] \mid 1 \leq j \leq L'\}$.

For a set of primers P, let $\Phi_i(P)$ denote the minimum between $L' + 1$ and the number of prefixes of $\{f^i[1..j], r^i[1..j] \mid 1 \leq j \leq L'\}$ covered by at least one primer in P. Also, let $\Phi(P) = \sum_{i=1}^{n} \Phi_i(P)$. The following properties of the integer valued set function Φ are immediate:

(A1) $\Phi(\emptyset) = 0$
(A2) $\Phi(P) = n(L' + 1)$ if and only if P is a feasible MPSS-L solution
(A3) Φ is a non-decreasing set function, i.e., $\Phi(P) \geq \Phi(P')$ whenever $P \supseteq P'$, and, furthermore, for every P such that $\Phi(P) < n(L' + 1)$, there exists $p \notin P$ such that $\Phi(P \cup \{p\}) > \Phi(P)$

Properties (A1)–(A3) suggest using $\Phi(\cdot)$ as a measure of the progress towards feasibility, and employing the generic greedy algorithm in Figure 2 to solve MPSS-L. The greedy algorithm starts with an empty set of primers and then iteratively adds the primer that gives the largest increase in Φ until reaching feasibility. By (A1)-(A3) this algorithm will end in a finite number of steps and will return a feasible MPSS-L solution.

Let us denote by $\Delta(p, P)$ the increase in Φ (also referred to as the "gain") obtained by adding primer p to set P, i.e., $\Delta(p, P) = \Phi(P \cup \{p\}) - \Phi(P)$. By (A3), it follows that the gain function Δ is non-negative. It is easy to verify that Δ is also monotonically non-increasing in the second argument, i.e.,

(A4) $\Delta(p, P) \geq \Delta(p, P')$ for every primer p and primer sets $P \subseteq P'$

Theorem 3.1. *Let $\Delta = \max_{p,P} \Delta(p, P)$. The greedy algorithm in Figure 2 returns an L-restricted primer cover of size at most $1 + \ln \Delta$ times larger than the optimum.*

Proof. We begin with some additional notations. Let $P^* = \{p_1^*, p_2^*, \ldots, p_k^*\}$ be an optimum MPSS-L solution and let $P = \{p_1, p_2, \ldots, p_g\}$ be the solution returned by the greedy algorithm, the latter one with primers indexed in the order in which they are selected by the algorithm. Let $\Phi_i^j = \Phi(\{p_1^*, \ldots, p_i^*\} \cup \{p_1, \ldots, p_j\})$, $\Delta_i^j = \Phi_i^j - \Phi_i^{j-1}$, and $\delta_i^j = \Phi_i^j - \Phi_{i-1}^j$. Note that, by (A4) and (A2), $\Delta_0^j \geq \Delta_1^j \geq \ldots \geq \Delta_k^j = 0$ for every $0 \leq j \leq g$, and $\delta_i^0 \geq \delta_i^1 \geq \ldots \geq \delta_i^g = 0$ for every $0 \leq i \leq k$. Furthermore, note that $\Delta_0^j \geq \delta_i^{j-1}$ for every $1 \leq i \leq k$ and $1 \leq j \leq g$. Indeed, Δ_0^j is the gain achieved by the greedy algorithm when selecting primer p_j. This gain must be at least $\Delta(p_i^*, \{p_1, ..., p_{j-1}\})$ since the greedy algorithm selects the primer with maximum gain in each iteration. Finally, by (A4), $\Delta(p_i^*, \{p_1, ..., p_{j-1}\}) \geq \Delta(p_i^*, \{p_1, ..., p_{j-1}\} \cup \{p_1^*, \ldots, p_{i-1}^*\}) = \delta_i^{j-1}$.

To analyze the size of the solution produced by the greedy algorithm, we use a charging scheme in which a certain cost is assigned to each primer in the optimal solution for every greedy primer. More precisely, the cost charged to p_i^* by the greedy primer p_j is

$$c_i^j = \begin{cases} \ln(\delta_i^{j-1}) - \ln(\delta_i^j), & \text{if } \delta_i^{j-1} \geq \delta_i^j > 0 \\ \ln(\delta_i^{j-1}) + 1, & \text{if } \delta_i^{j-1} > \delta_i^j = 0 \\ 0, & \text{if } \delta_i^{j-1} = \delta_i^j = 0 \end{cases}$$

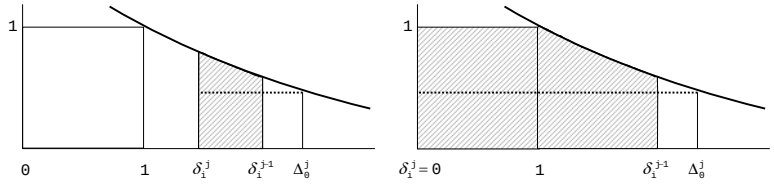


Figure 3. A graphical illustration of the cost lower-bound used in the proof of Theorem 3.1 for $\delta_i^{j-1} \geq \delta_i^j > 0$ (a), and for $\delta_i^{j-1} > \delta_i^j = 0$ (b). In each case, c_i^j is equal to the area shaded under the curve $\min\{1, 1/x\}$. Since $\Delta_0^j \geq \delta_i^{j-1}$, the shaded area is larger than the area of a rectangle with width $\delta_i^{j-1} - \delta_i^j$ and height $1/\Delta_0^j$.

Notice that the total cost charged to optimal primer p_i^*, $\sum_{j=1}^{g} c_i^j$, is a telescopic sum equal to $1 + \ln(\delta_i^0) \leq 1 + \ln \Delta$. Hence, the overall cost is at most $k(1 + \ln \Delta)$. To prove the approximation factor of $1 + \ln \Delta$ it suffices to prove that we charge at least one unit of cost for each greedy primer. Indeed, consider a fixed $j \in \{1, \ldots, g\}$. Since $\Delta_0^j \geq \delta_i^{j-1}$, it follows that

$$c_i^j \geq \frac{\delta_i^{j-1} - \delta_i^j}{\Delta_0^j}$$

for every $1 \leq i \leq k$ (see Figure 3). Using that $\delta_i^{j-1} - \delta_i^j = \Delta_j^{i-1} - \Delta_j^i$ and $\Delta_j^k = 0$ gives

$$\sum_{i=1}^{k} c_i^j \geq \sum_{i=1}^{k} \frac{\Delta_j^{i-1} - \Delta_j^i}{\Delta_0^j} = 1$$

which completes the proof. □

We remark that the value of Δ in Theorem 3.1 is much smaller than nL for practical MPSS-L instances, and hence the approximation factor in Theorem 3.1 is tighter than the one we have previously established[11] for the greedy algorithm.

3.1. *Implementation details*

In this section we discuss the details of an efficient implementation of the generic greedy algorithm in Figure 2. First, we note that although there are 4^ℓ DNA sequences of length ℓ, no more than $2nL$ of these sequences (all substrings of length ℓ of the input genomic sequences $\mathcal{S} = \{f^i, r^i \mid 1 \leq i \leq n\}$) can be used as primers. Our implementation starts by creating a list with all feasible primers by removing substrings that do not meet user-specified constraints on GC content and melting temperature T_m; masking of repetitive elements and more stringent candidate filtering based, e.g., on the sophisticated statistical scoring models developed by Yuryev et al.[22] can also be easily incorporated in this pre-processing step. For each remaining primer, we precompute all hybridization positions within the strings of $\mathcal{S}$. Using this, we can then compute the gain of any feasible primer p in time $O(n_p)$, where n_p is the number of hybridization positions for p. The primer with maximum gain is then found in step 2(a) of the algorithm by sequentially computing the gain of remaining primers.

In order to speed up the implementation, we use two further optimizations. A feasible primer is called *unique* if it hybridizes only one of the sequences in $\mathcal{S}$. The first optimization is to retain only the unique feasible primer closest to the amplification locus for each f^i and r^i. The number of eliminated unique candidate primers greatly depends on primer length ℓ, but is usually a significant fraction of the number of feasible candidate primers. Clearly, removing these primers does not worsen the quality of the returned solution.

The second optimization is to adopt a lazy strategy for recomputing primer gains in step 2(a). In first execution of step 2(a) we compute and store the gain for all feasible primers. In subsequent iterations, the gain of a primer is only recomputed if the saved gain is higher than the best gain seen in current iteration. Since gains are monotonically non-increasing, this optimization is not affecting the set of primers returned by the algorithm.

4. Experimental Results

We performed experiments on test cases extracted from the human genome databases as well as simulated test cases. The human genome test cases are regions surrounding 100 known SNPs collected from National Center for Biotechnology Information's genomic databases.[3] Random test cases were generated from the uniform distribution induced by assigning equal probabilities to each nucleotide.

For all experiments we used a bound $L = 1000$ on the PCR amplification length, and a bound ℓ between 8 and 12 on primer length. Although it has been suggested that such short primers may not be specific enough since they are likely to hybridize to many homologue sites,[8] we note that hybridization outside the target region will not result in significant amplification unless *two* primers hybridize sufficiently closely to each other, a much less likely event.[6] Indeed, the feasibility of using primers with only 8-12 target specific nucleotides has been experimentally validated by Jordan et al.[10]

We compared the following four algorithms:

- The greedy primer cover algorithm of Pearson et al.[18] (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 the resulting set of primers 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 when a primer p covers for the first time one of the forward or reverse sequences corresponding to a SNP, say at position t, we appropriately truncate the opposite sequence to ensure that the final primer cover is L-restricted.
- The greedy approximation algorithm in Figure 2, called G-POT since it makes greedy choices based on the potential function Φ. We implemented the algorithm as described in Section 3.1. However, to facilitate comparison with the other algorithms, in these experiments we did not impose any constraints on the GC content or melting temperature of candidate probes.
- The iterative beam-search heuristic of Souvenir et al.[21] We used the primer-threshold version of this heuristic, MIPS-PT, with degeneracy bound set to 1 and the default

Table 1. Results on NCBI test cases for $\ell = 8, 10, 12$ and $L = 1000$.

# SNPs	ℓ	G-FIX #Primers	G-FIX CPU sec.	G-VAR #Primers	G-VAR CPU sec.	MIPS-PT #Primers	MIPS-PT CPU sec.	G-POT #Primers	G-POT 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

values for the remaining parameters (in particular, beam size was set to 100).

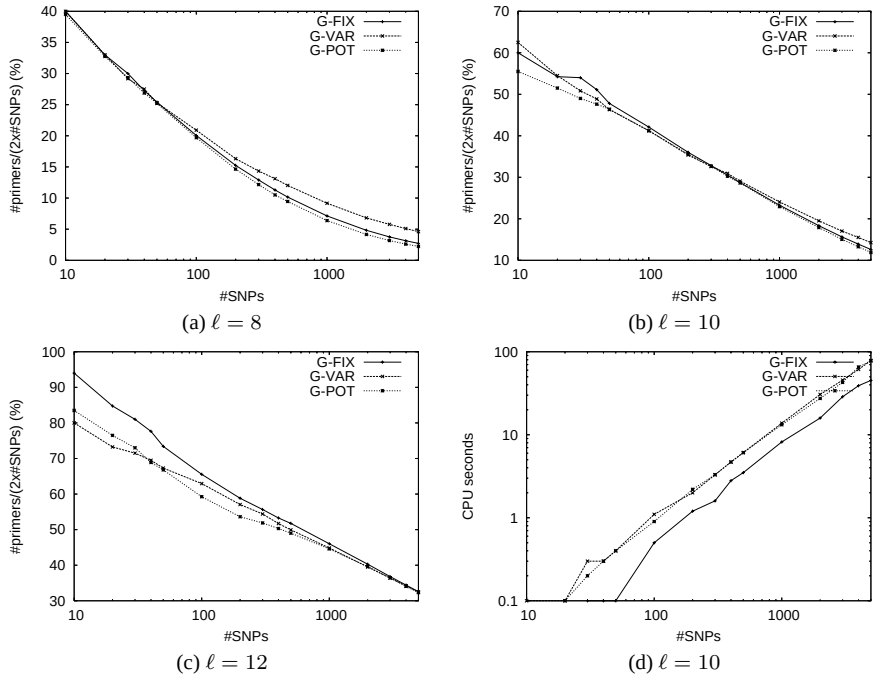
Table 1 gives the number of primers selected and the running time (in CPU seconds) for the three greedy algorithms and for the iterative beam-search MIPS-PT heuristic on instances extracted from the NCBI repository. G-POT has the best performance on all test cases, reducing the number of primers by up to 24% compared to G-FIX and up to 30% compared to G-VAR. G-VAR 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 4(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 test cases. 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 and, with few exceptions, its G-VAR modification.

Figure 4(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 meta-heuristic is to run all three algorithms and return the best of the three solutions found.

5. Conclusions

In this paper we have presented an improved analysis of a simple greedy algorithm for MP-PCR primer set selection with amplification length constraints and experimental results showing that our algorithm obtains significant reductions in the number of primers compared to previous algorithms. A promising approach to further increasing MP-PCR efficiency is the use of *degenerate PCR primers*.[13,14,21] A degenerate primer is essentially a mixture consisting of multiple non-degenerate primers sharing a common pattern. Remarkably, degenerate primer cost is nearly identical to that of a non-degenerate primer,



(a) $\ell = 8$ (b) $\ell = 10$ (c) $\ell = 12$ (d) $\ell = 10$

Figure 4. (a)–(c) Performance of the compared algorithms, measured as relative improvement over the trivial solution of using two primers per SNP, for $\ell = 8, 10, 12$, $L = 1000$, and up to 5000 SNPs. (d) Runtime of the compared algorithms for $\ell = 10$, $L = 1000$, and up to 5000 SNPs. Each number represents the average over 10 test cases of the respective size.

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). Since degenerate primers may lead to excessive unintended amplification, a bound on the degeneracy of a primer (i.e., the number of distinct non-degenerate primers in the mixture) is typically imposed. [14,21]

Our greedy algorithm extends directly to the problem of selecting, for a given set of genomic loci, a minimum size L-restricted primer cover consisting of degenerate primers with bounded degeneracy. However, even for moderate degeneracy constraints, it becomes impractical to explicitly evaluate the gain function for all candidate primers. Indeed, as remarked by Linhart and Shamir,[14] the number of candidate degenerate primers may be as large as $2nL\binom{k}{\delta}15^{\delta}$, where n is the number of loci, L is the PCR amplification length upperbound, and δ is the number of "degenerate nucleotides" allowed in a primer. To maintain a practical runtime, one may sacrifice optimality of the greedy choice in step 2(a) of the greedy algorithm, using instead approximation algorithms similar to those of Linhart and Shamir[14] for finding degenerate primers guaranteed to have near optimal gain. The analysis in Section 3 can be easily modified to prove the following approximation guarantee for this modification of the greedy algorithm.

Theorem 5.1. *Assume that the greedy algorithm in Figure 2 is modified to select in step 2(a) a primer whose gain is within a factor of α of the maximum possible gain, for some*

fixed $0 < \alpha \leq 1$. *Then, the modified algorithm returns an* L*-restricted primer cover of size at most* $(1 + \ln \Delta)\alpha$ *times larger than the optimum, where* $\Delta = \max_{p,P} \Delta(p, P)$.

References

1. P. Berman, B. DasGupta, and M.-Y. Kao. Tight approximability results for test set problems in bioinformatics. DIMACS Technical Report 2003-14, 2003.
2. V. Chvátal. A greedy heuristic for the set covering problem. *Mathematics of Operations Research*, 4:233–235, 1979.
3. International Human Genome Sequencing Consortium. *Homo sapiens chromosome 12 genomic contig.* National Center for Biotechnology Information, http://www.ncbi.nlm.nih.gov.
4. K. Doi and H. Imai. A greedy algorithm for minimizing the number of primers in multiple PCR experiments. *Genome Informatics*, 10:73–82, 1999.
5. U. Feige. A treshold of $\ln n$ for approximating set cover. *Journal of the ACM*, 45:634–652, 1998.
6. R.J. Fernandes and S.S. Skiena. Microarray synthesis through multiple-use PCR primer design. *Bioinformatics*, 18:S128–S135, 2002.
7. M.T. Hajiaghayi, K. Jain, K.M. Konwar, L.C. Lau, I.I. Mandoiu, A.C. Russell, A.A. Shvartsman, and V.V. Vazirani. The minimum k-colored subgraph problem in haplotyping and DNA primer selection. Submitted.
8. M.-H. Hsieh, W.-C. Hsu, S.-Kay, and C.M. Tzeng. An efficient algorithm for minimal primer set selection. *Bioinformatics*, 19:285–286, 2003.
9. D.S. Johnson. Approximation algorithms for combinatorial problems. *Journal of Computer and System Sciences*, 9:256–278, 1974.
10. B. Jordan, A. Charest, J.F. Dowd, J.P. Blumenstiel, R. f. Yeh, A. Osman, D.E. Housman, and J.E. Landers. Genome complexity reduction for SNP genotyping analysis. *Proc. Natl. Acad. Sci. USA*, 99:2942–2947, 2002.
11. K. Konwar, I.I. Mandoiu, A. Russell, and A. Shvartsman. Approximation algorithms for minimum PCR primer set selection with amplification length and uniqueness constraints. ACM Computing Research Repository, cs.DS/0406053, 2004.
12. P.Y. Kwok. Methods for genotyping single nucleotide polymorphisms. *Annual Review of Genomics and Human Genetics*, 2:235–258, 2001.
13. 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.
14. C. Linhart and R. Shamir. The degenerate primer design problem. *Bioinformatics*, 18:S172–S181, 2002.
15. L. Lovász. On the ratio of optimal integral and fractional covers. *Discrete Mathematics*, 13:383–390, 1975.
16. K. Mullis. Process for amplifying nucleic acid sequences. United States Patent 4,683,202, 1987.
17. P. Nicodème and J.-M. Steyaert. Selecting optimal oligonucleotide primers for multiplex PCR. In *Proc. 5th Intl. Conference on Intelligent Systems for Molecular Biology*, pages 210–213, 1997.
18. 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.
19. 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.
20. P. Slavik. Improved performance of the greedy algorithm for partial cover. *Information Processing Letters*, 64:251–254, 1997.
21. 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.
22. A. Yuryev, J. Huang, K.E. Scott, J. Kuebler, M. Donaldson, M.S. Phillipes, M. Pohl, and M.T Boyce-Jacino. Primer design and marker clustering for multiplex SNP-IT primer extension genotyping assay using statistical modeling. *Bioinformatics*, to appear.