PRIMER SELECTION METHODS FOR DETECTION OF GENOMIC INVERSIONS AND DELETIONS VIA PAMP

B. DASGUPTA*

Department of Computer Science, University of Illinois at Chicago, Chicago, IL 60607-7053
E-mail: dasgupta@cs.uic.edu

J. JUN† and I.I. MĂNDOIU†

Computer Science & Engineering Department, University of Connecticut, Storrs, CT 06269-2155
E-mail: {jinjun,ion}@engr.uconn.edu

Primer Approximation Multiplex PCR (PAMP) is a recently introduced experimental technique for detecting large-scale cancer genome lesions such as inversions and deletions from heterogeneous samples containing a mixture of cancer and normal cells. In this paper we give integer linear programming formulations for the problem of selecting sets of PAMP primers that minimize detection failure probability. We also show that PAMP primer selection for detection of anchored deletions cannot be approximated within a factor of $2 - \varepsilon$, and give a 2-approximation algorithm for a special case of the problem. Experimental results show that our ILP formulations can be used to optimally solve medium size instances of the inversion detection problem, and that heuristics based on iteratively solving ILP formulations for a one-sided version of the problem give near-optimal solutions for anchored deletion detection with highly scalable runtime.

Keywords: Genomic structural variation detection; PAMP primer selection; Integer linear programming.

1. Introduction

As described by Liu and Carson,[1] PAMP requires the selection of a large number of multiplex PCR primers from the genomic region of interest. Exploiting the fact that the efficiency of PCR amplification falls off exponentially beyond a certain product length, PAMP primers are selected such that (1) no PCR amplification results in the absence of genomic lesions, and (2) with high probability, a genomic lesion brings one or more pairs of primers in the proximity of each other, resulting in PCR amplification. Multiplex PCR amplification products are then hybridized to a microarray to identify the pair(s) of primers that yield amplification. This gives an approximate location for the breakpoints of the genomic lesion; precise breakpoint coordinates can be determined by sequencing PCR products.

As in previous multiplex PCR primer set selection formulations,[2–4] PAMP primers must satisfy standard selection criteria such as hybridizing to a unique site in the genomic

[*] Supported in part by NSF grants IIS-0346973, IIS-0612044 and DBI-0543365.
[†] Supported in part by NSF grants IIS-0546457 and DBI-0543365.

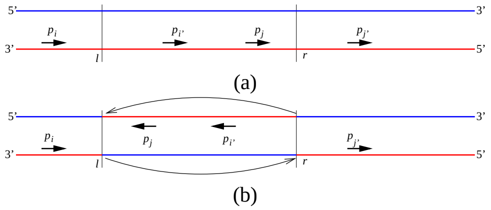
region of interest, having melting temperature in a pre-specified range, and lacking secondary structures such as hairpins. Candidate primers meeting these criteria can be found using robust software tools for primer selection, such as the Primer3 package.[5] Similar to some previous works on multiplex PCR primer set selection,[2,4] PAMP also requires subsets of non-dimerizing primers. Indeed, as observed in Bashir et al.,[6] even a single pair of dimerizing primers can lead to complete loss of amplification signal. However, unlike existing works on multiplex PCR primer set selection[2–4] which focus on minimizing the number of primers and/or multiplex PCR reactions needed to amplify a *given* set of discrete amplification targets, the objective in PAMP primer selection is to minimize the probability that an *unknown* genomic lesion fails to be detected by the assay. The only work we are aware on this novel problem is that of Bashir et al.,[6] who proposed integer linear programming (ILP) formulations and simulated annealing algorithms for PAMP primer selection when the goal is to detect genomic deletions known to include a given anchor locus.

In this paper we show that the optimization objective used in the ILP formulation of Bashir et al.[6] is not equivalent to minimization of failure probability, and propose new ILP formulations capturing the later objective in PAMP primer selection for detection of genomic inversions (Section 2) and anchored deletions (Section 3). We also show that PAMP primer selection for detection of anchored deletions cannot be approximated within a factor of $2 - \varepsilon$ (Lemma 3.1), and give a 2-approximation algorithm for a special case of the problem (Lemma 3.2). Experimental results presented in Section 4 show that our ILP formulations can be used to optimally solve medium size instances of the inversion detection problem, and that heuristics based on iteratively solving ILP formulations for a *one-sided* version of the anchored deletion detection problem give near-optimal solutions with highly scalable runtime.

2. Inversion Detection

Throughout the paper, PCR amplification is assumed to occur if and only if there is at least one pair of primers hybridizing to opposite strands at two sites that are at most L bases apart and such that the primers' $3'$ ends face each other. This model assumes that PCR amplification success probability is a simple 1-0 step function of product length, with the transition from fully efficient amplification to no amplification taking place between product lengths L and $L + 1$. Our methods can be easily modified to handle arbitrary amplification success probability functions.

Let $\mathcal{G}$ be a genomic region indexed along the forward strand in $5' - 3'$ orientation. We seek a set of non-dimerizing multiplex PCR primers that does not yield PCR amplification when a specified interval $[x_{min}, x_{max}]$ of $\mathcal{G}$ contains no inversion, and, subject to this condition, minimizes the probability of not getting amplification when an inversion is present in the sample. In order to formalize the optimization objective, we assume a known probability distribution for the pairs of endpoints of inversions within $[x_{min}, x_{max}]$, i.e., we assume that, for every pair (l, r) of endpoints with $x_{min} \leq l < r \leq x_{max}$, we are given the (conditional) probability $p_{l,r} \geq 0$ of encountering an inversion with endpoints l and r, where $\sum_{x_{min} \leq l < r \leq x_{max}} p_{l,r} = 1$. This probability distribution may be as simple as the



(a)

(b)

Fig. 1. Hybridization loci for 4 PAMP primers without (a) and with (b) an inversion with endpoints (l, r). DNA strands are color-coded blue or red according to their forward/reverse orientation in the reference genome. Multiplex PCR yields no amplicons when the sample contains no genomic inversion, but yields at least one amplicon if an inversion brings binding sites of primers p_i and p_j within L bases of each other.

uniform distribution (under which every pair of endpoints is equally likely), or can incorporate existing biological knowledge on the distribution of recombination hotspots and/or biases in inversion segment lengths.

In the pre-processing stages of the primer selection process, we collect a large number of candidate primers satisfying appropriate biochemical constraints on melting temperature, lack of hairpin secondary structures, etc. Each candidate primer must also hybridize to the reverse strand of the reference genome at a unique location within $\mathcal{G}$ (see Figure 1(a)). Clearly, multiplex PCR with any subset of the candidate primers should not yield PCR amplification when the genomic sample contains no inversion within $\mathcal{G}$.

Let $\mathcal{P} = \{p_1, p_2, \ldots, p_n\}$ denote the set of candidate primers, and let $x_1 < x_2 < \ldots < x_n$ be the positions of their $3'$ ends when hybridized to the reverse strand of $\mathcal{G}$. Furthermore, let $\mathcal{E}$ denote the set of pairs of primers in $\mathcal{P}$ that form dimers. The *PAMP primer selection problem for inversion detection (PAMP-INV)* can then be formulated as follows:

Given: set $\mathcal{P}$ of candidate primers hybridizing at unique loci of the reverse strand of $\mathcal{G}$, set $\mathcal{E}$ of dimerizing candidate primer pairs, maximum multiplexing degree N, and amplification length upper-bound L
Find: a subset $\mathcal{P}'$ of $\mathcal{P}$ such that

(1) $|\mathcal{P}'| \leq N$
(2) $\mathcal{P}'$ does not include any pair of primers in $\mathcal{E}$, and
(3) The probability that multiplex PCR using the primers of $\mathcal{P}'$ fails to yield amplification, given that $[x_{min}, x_{max}]$ contains an inversion, is minimized. In other words, $\mathcal{P}'$ minimizes

$$\sum_{x_{min} \leq l < r \leq x_{max}} f(\mathcal{P}'; l, r) p_{l,r} \tag{1}$$

where $f(\mathcal{P}'; l, r) = 1$ if $\mathcal{P}'$ fails to yield a PCR product when the inversion with endpoints (l, r) is present in the sample, and $f(\mathcal{P}'; l, r) = 0$ otherwise.

We next formulate PAMP-INV as an integer linear program (ILP). For convenience, we add to $\mathcal{P}$ "dummy" primers p_0 and p_{n+1}, assumed to uniquely hybridize to $\mathcal{G}$ at locations $x_0 = x_{min} - L$ and and $x_{n+1} = x_{max} + L$, respectively. Dummy primers are assumed not

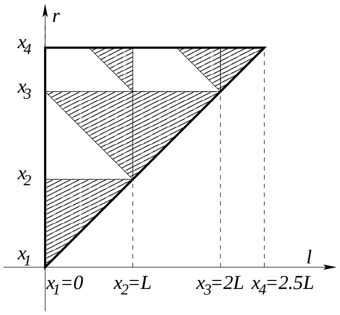


Fig. 2. Graphical representation of the space of endpoint pairs (l, r) (area within thick triangle) for a PAMP-INV instance with $x_{min} = 0$, $x_{max} = 2.5L$. If primer set $\mathcal{P}'$ consists of 4 primers hybridizing to the reference genome at positions 0, L, $2L$, and $2.5L$, respectively, inversions (l, r) corresponding to the shaded regions fail to yield PCR amplification.

to dimerize with each other or with other primers in $\mathcal{P}$, and thus they can always be included in $\mathcal{P}'$. Without loss of generality we will assume that the location of all candidate primers is between x_0 and x_{n+1}, since primers that hybridize outside the interval $[x_{min}-L, x_{max}+L]$ cannot help in detecting inversions located within $[x_{min}, x_{max}]$.

Consider an inversion with endpoints (l, r) and a set of non-dimerizing primers $\mathcal{P}' \subseteq \mathcal{P}$ with $p_0, p_{n+1} \in \mathcal{P}'$. Let $i = \max\{k : p_k \in \mathcal{P}', x_k < l\}$ and $j = \max\{k : p_k \in \mathcal{P}', x_k < r\}$. Note that if both endpoints of the inversion occur between two consecutive primers of $\mathcal{P}'$ (i.e., $i = j$), then $\mathcal{P}'$ fails to yield any amplification and the inversion remains undetected. When $i < j$, $\mathcal{P}'$ still fails to yield any amplification if $(l-1-x_i)+(r-x_j) > L$. On the other hand, when $i < j$ and $(l-1-x_i)+(r-x_j) \leq L$, the multiplex PCR reaction using the primers of $\mathcal{P}'$ yields at least one amplification product given by p_i and p_j.

For every quadruple (i, i', j, j') with $x_i < x_{i'}, x_j < x_{j'}, x_i \leq x_j$, we let $C_{i,i',j,j'} = \sum p_{l,r}$, where the sum is over all inversion endpoint pairs (l, r) such that $\max\{x_i, x_{min}\} < l \leq \min\{x_{i'}, x_{max}\}$, $\max\{x_j, x_{min}\} < r \leq \min\{x_{j'}, x_{max}\}$, $(l-1-x_i)+(r-x_j) > L$. If $(p_i, p_{i'})$ and $(p_j, p_{j'})$ are pairs of consecutive primers of $\mathcal{P}'$, then $C_{i,i',j,j'}$ gives the cumulative probability that an inversion with endpoints $l \in (x_i, x_{i'}] \cap [x_{min}, x_{max}]$ and $r \in (x_j, x_{j'}] \cap [x_{min}, x_{max}]$ fails to yield any amplification product under multiplex PCR with the primers of $\mathcal{P}'$.

To express PAMP-INV as an ILP we use three types of 0/1 variables:

- e_i, which are set to 1 if and only if $p_i \in \mathcal{P}'$,
- $e_{i,i'}$, which are set to 1 if and only if p_i and $p_{i'}$ are consecutive primers in $\mathcal{P}'$, and
- $e_{i,i',j,j'}$, which are set to 1 if and only if $(p_i, p_{i'})$ and $(p_j, p_{j'})$, are consecutive primers in $\mathcal{P}'$ and $i \leq j$.

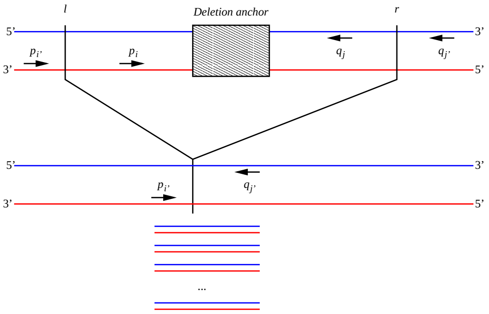
Variables of last type allow expressing the total failure probability (1) as a sum of appropriate $C_{i,i',j,j'}$'s. The complete PAMP-INV ILP is given below. Constraints (3) and (4) ensure that a variable $e_{i,i',j,j'}$ is set to 1 if and only if both $e_{i,i'}$ and $e_{j,j'}$ are set to 1. Similarly, constraints (5) ensure that a variable $e_{i,j}$ is set to 1 only if both e_i and e_j are set to 1. Variables $e_{i,j}$ which are set to 1 can be viewed as defining a path connecting p_0 to p_{n+1} via a subset of intermediate primers visited in left-to-right order, and this is captured in constraints (6) and (7). Constraint (8) can handle a limitation on the number of allowed primers (N). Finally, constraint (9) is used to ensure that no pair of dimerizing candidate primers is added to the selected set $\mathcal{P}'$.

$$
\begin{aligned}
\text{minimize} \quad & \sum_{\{(i,i',j,j')\,:\,i<i',j<j',i\leq j\}} C_{i,i',j,j'}\; e_{i,i',j,j'} && (2)\\
\text{s.t.} \quad & e_{i,i'} + e_{j,j'} \geq 2e_{i,i',j,j'},\; i<i',\, j<j', \text{ and } i \leq j && (3)\\
& e_{i,i',j,j'} \geq e_{i,i'} + e_{j,j'} - 1,\; i<i',\, j<j', \text{ and } i \leq j && (4)\\
& e_i + e_j \geq 2e_{ij},\; 1 \leq i < j \leq n && (5)\\
& \sum_{j=1}^{n+1} e_{0j} = \sum_{i=0}^{n} e_{i,n+1} = 1 && (6)\\
& \sum_{i=0}^{j-1} e_{ij} = \sum_{k=j+1}^{n+1} e_{jk},\; 1 \leq j \leq n && (7)\\
& \sum_{1\leq i\leq n} e_i \leq N && (8)\\
& e_i + e_j \leq 1,\; \text{for all } (p_i,p_j) \in \mathcal{E} && (9)\\
& e_{i,i',j,j'} \in \{0,1\},\; e_{i,j} \in \{0,1\},\; e_i \in \{0,1\}
\end{aligned}
$$

3. Anchored Deletion Detection

Bashir et al.[6] recently studied the PAMP primer selection problem for deletion detection, which we will refer to as PAMP-DEL. As in their work, we assume that the deletion spans a known genomic location, i.e., we consider detection of *anchored* deletions only. Let $\{p_1,\ldots,p_m\}$ and $\{q_1,\ldots,q_n\}$ be the two sets of forward and reverse candidate primers, indexed by increasing distance from the anchor. Given a set $\mathcal{E}$ of primer pairs that form dimers, the goal is to pick a set $\mathcal{P}'$ of at most N_f forward and at most N_r reverse primers such that no two of the selected primers dimerize, and, subject to this constraint, the probability that the selected primers fail to produce a PCR product when the sample contains a deletion is minimized. The latter probability is computed assuming given PCR amplification threshold L and probability distribution for the pairs of endpoints of the deletion.

PAMP-DEL can be formulated as an ILP using an idea similar to that in previous section. For every quadruple (i,i',j,j'), $i \leq i'$, $j \leq j'$, let $C_{i,i',j,j'}$ denote the total probability that a deletion with ends between the hybridization sites of p_i and $p_{i'}$, respectively q_j and $q_{j'}$, does not result in PCR amplification when $(p_i,p_{i'})$ and $(q_j,q_{j'})$ are consecutive sets of forward, respectively reverse primers of $\mathcal{P}'$. Using 0/1 variables f_i (r_i) to indicate when p_i (respectively q_i) is selected in $\mathcal{P}'$, $f_{i,j}$ ($r_{i,j}$) to indicate that p_i and p_j (respectively q_i and q_j) are consecutive primers in $\mathcal{P}'$, and $e_{i,i',j,j'}$ to indicate that both $(p_i,p_{i'})$ and $(q_j,q_{j'})$ are pairs of are consecutive primers in $\mathcal{P}'$, we obtain the following formulation:



(a)

Fig. 3. Deletion detection using PAMP. If a deletion with endpoints l and r brings the hybridization loci of forward primer $p_{i'}$ and reverse primer $q_{j'}$ within L bases of each other the PAMP assay results in PCR amplification.

$$\begin{aligned}
\text{minimize} \quad & \sum_{\{(i,i',j,j')\,:\,i<i',j<j'\}} C_{i,i',j,j'}\; e_{i,i',j,j'} \qquad (10)\\
\text{s.t.} \quad & f_{i,i'} + r_{j,j'} \geq 2e_{i,i',j,j'},\; i < i' \text{ and } j < j'\\
& e_{i,i',j,j'} \geq f_{i,i'} + r_{j,j'} - 1,\; i < i' \text{ and } j < j'\\
& f_i + f_j \geq 2f_{i,j},\; 1 \leq i < j \leq m\\
& r_i + r_j \geq 2r_{i,j},\; 1 \leq i < j \leq n\\
& \sum_{j=1}^{m+1} f_{0,j} = \sum_{i=0}^{m} f_{i,m+1} = \sum_{j=1}^{n+1} r_{0,j} = \sum_{i=0}^{n} r_{i,n+1} = 1\\
& \sum_{i=0}^{j-1} f_{i,j} = \sum_{k=j+1}^{m+1} f_{j,k},\; 1 \leq j \leq m\\
& \sum_{i=0}^{j-1} r_{i,j} = \sum_{k=j+1}^{n+1} r_{j,k},\; 1 \leq j \leq n\\
& \sum_{1\leq i\leq m} f_i \leq N_f,\; \sum_{1\leq i\leq n} r_i \leq N_r\\
& f_i + f_j \leq 1,\; \text{for all } (p_i, p_j) \in \mathcal{E}\\
& r_i + r_j \leq 1,\; \text{for all } (q_i, q_j) \in \mathcal{E}\\
& f_i + r_j \leq 1,\; \text{for all } (p_i, q_j) \in \mathcal{E}\\
& e_{i,i',j,j'} \in \{0,1\},\; f_{i,j}, r_{i,j} \in \{0,1\},\; f_i, r_i \in \{0,1\}
\end{aligned}$$

Bashir et al.[6] also introduced an *one-sided* version of PAMP-DEL, referred to as PAMP-1SDEL, in which one of the deletion endpoints is known in advance. For this version of the problem our ILP formulation can be simplified substantially. Let $x_1 < x_2 < \ldots < x_n$ be the hybridization positions for the reverse candidate primers $q_1, \ldots, q_n$. We introduce two dummy reverse primers that hybridize right after the location x_0 of the anchor, and at position $x_{n+1} = x_{max} + L$, respectively (as usual, dummy primers are assumed not to dimerize). Denoting by $C_{i,j}$ the probability that a deletion whose right endpoint falls between x_i and x_j does not result in PCR amplification, and using 0/1 variables r_i and $r_{i,j}$ as in the PAMP-DEL ILP, we obtain the following formulation for PAMP-1SDEL:

$$\begin{aligned}
\text{minimize } & \sum_{i<j} C_{i,j}\, r_{i,j} && (11)\\
\text{s.t. } & r_i + r_j \geq 2r_{i,j},\ 1 \leq i < j \leq n \\
& \sum_{j=1}^{n+1} r_{0,j} = \sum_{i=0}^{n} r_{i,n+1} = 1 \\
& \sum_{i=0}^{j-1} r_{i,j} = \sum_{k=j+1}^{n+1} r_{j,k},\ 1 \leq j \leq n \\
& \sum_{1\leq i \leq n} r_i \leq N, \\
& r_i + r_j \leq 1,\ \text{for all } (q_i, q_j) \in \mathcal{E} \\
& r_{i,j} \in \{0,1\},\ r_i \in \{0,1\}
\end{aligned}$$

Discussion. The PAMP-DEL formulation in Bashir et al.[6] does not actually make explicit the underlying probabilistic distribution for the endpoints of the deletion. The ILP proposed by Bashir et al. for PAMP-DEL uses an objective similar to (11) with

$$C_{i,j} = \max\{(x_j - x_i - L/2), 0\} \tag{12}$$

which is measuring the so called "uncovered area." It is not difficult to see that minimizing uncovered area as proposed by Bashir et al. may *not* result in minimizing the probability of failure, even assuming a uniform probability distribution for the deletion endpoints as suggested by (12). An example is as follows. Consider a PAMP-DEL instance in which possible deletions have left endpoint in the interval $(0, L]$ and right endpoint in the interval $(2L, 3L]$, with each endpoint position equally likely. There are non-dimerizing forward primers at *every* position between 0 and L, and three reverse primers at positions $2L$, $2.5L$, and $3L$, with the last two of these primers forming a dimer. The minimum failure probability is in this case zero, and is achieved by selecting all forward primers and the reverse primers at $2L$ and $3L$. However, the minimum uncovered area is $L/2$, since one of the primers at $2.5L$ and $3L$ cannot be selected. The ILP proposed in Bashir et al.[6] may select all forward primers and the reverse primers at $2L$ and $2.5L$, which has optimal uncovered area but fails to detect deletions with probability 1/2.

Lemma 3.1. *Assuming the UNIQUE GAMES conjecture, PAMP-1SDEL (and hence, PAMP-DEL) cannot be approximated to within a factor of* $2 - \varepsilon$ *for any constant* $\varepsilon > 0$.

Sketch of Proof. We reduce the vertex cover problem to PAMP-1SDEL. It is known[7] that, assuming that the UNIQUE GAMES conjecture holds, the vertex cover problem cannot be approximated to within a factor of $2 - \varepsilon$ for any constant $\varepsilon > 0$. Consider an instance

Table 1. Detection probability and ILP runtime for PAMP-INV instances with $x_{max} - x_{min} = 100Kb$ and $L = 20Kb$ (averages over 5 random instances).

Dimerization	n=20(ρ=3.33)			n=30 (ρ=5)		
Rate (%)	N=20	15	10	N=20	15	10
	Detection probability(%)					
0	93.91	93.83	91.17	99.25	99.20	96.79
1	93.57	93.54	91.11	98.79	98.69	96.11
2	92.68	92.68	90.55	98.69	98.60	96.06
5	89.78	89.78	88.28	97.84	97.78	95.68
10	84.41	84.41	83.57	94.99	94.98	92.95
20	71.53	71.53	71.53	81.70	81.70	81.64
	Runtime (seconds)					
0	175.01	379.87	994.76	2160.45	5238.17	86115.50
1	211.54	337.44	956.34	2461.93	4919.25	57229.18
2	259.77	260.20	913.67	2081.81	5864.61	31655.12
5	667.87	618.33	868.28	3903.71	6660.55	14266.41
10	535.20	496.97	495.14	6405.27	7081.30	18284.68
20	520.96	470.19	558.82	15506.87	14893.29	14847.14

$G = (V, E)$ of vertex cover with $V = \{1, 2, \ldots, n\}$ and $\{1, n\} \notin E$. We define an instance of PAMP-1SDEL with reverse primers $q_1, \ldots, q_n$ at positions $x_i = iL$, $i = 1, \ldots, n$, where the pairs of dimerizing primers correspond to the edges of G. Further, assume that the position of the right endpoint of the deletion is uniformly distributed in the interval $[0, nL]$.

Let V' be a vertex cover of G containing k vertices. Then, $V \setminus V'$ is an independent set of G, and the set of primers $\{q_i \,:\, i \in V \setminus V'\}$ is a feasible PAMP-1SDEL solution whose failure probability is k/n. Conversely, consider a solution $\mathcal{P}'$ of PAMP-1SDEL with failure probability k/n. Under the uniform probability distribution, it follows that $|\mathcal{P}'| = n - k$. Clearly $\{i \,:\, q_i \in \mathcal{P}'\}$ is an independent set of G, and so $\{i \,:\, q_i \notin \mathcal{P}'\}$ is a vertex cover of size k of G. ■

On the positive side we have the following result, whose proof we omit due to space the limitation.

Lemma 3.2. *There is a* 2*-approximation algorithm for the special case of PAMP-1SDEL in which candidate primers are spaced at least* L *bases apart and the deletion endpoint is distributed uniformly within a fixed interval* $(x_0, x_{max}]$.

4. Experimental Results

We used the Cplex 10.1 solver to solve ILP formulations given in Sections 2 and 3. All reported runtimes are for a Dell PowerEdge 6800 server with four 2.66GHz Intel Xeon dual-core processors (only one of which is used by Cplex).

Table 1 gives the detection probability (one minus failure probability) and runtimes for the ILP from Section 2 for randomly generated PAMP-INV instances with $x_{max} - x_{min}$=100Kb, L=20Kb (which is representative of long-range PCR), number of candidate primers n between 20 and 30 (candidate primer density $\rho = nL/(x_{max} - x_{min} + L)$ be-

Table 2. Comparison of PAMP-DEL ILP, ITERATED-1SDEL, and INCREMENTAL-1SDEL for instances with $m = n = N_f = N_r = 15$, $x_{max} - x_{min} = 5Kb$, and $L = 2Kb$ (averages over 5 random instances for each dimerization rate between 0 and 20%).

Dimerization	PAMP-DEL ILP		ITERATED-1SDEL		INCREMENTAL-1SDEL	
Rate	Detection	#Primers	Detection	#Primers	Detection	#Primers
(%)	Prob. (%)		Prob. (%)		Prob. (%)	
0	97.29	(15.0,15.0)	97.29	(15.0,15.0)	97.29	(10.4, 8.8)
1	96.81	(14.2,12.6)	96.81	(14.4,12.6)	96.81	(11.4, 9.6)
2	96.73	(13.4,11.6)	96.70	(13.6,11.4)	96.73	(11.6,10.0)
5	93.13	(10.8, 8.0)	88.91	(10.4, 7.4)	91.60	(10.0, 7.8)
10	87.58	(8.2, 6.2)	84.34	(8.4, 6.4)	83.19	(7.0, 5.8)
20	72.95	(6.0, 4.8)	56.03	(6.4, 3.8)	68.89	(5.4, 4.0)

tween 3.33 and 5), maximum multiplexing degree N between 10 and 20, and primer dimerization rate between 0 and 20%. Both the hybridization locations for candidate primers and the pairs of candidate primers that dimerize were selected uniformly at random. In this experiment all inversions longer than 10Kb were assumed to be equally likely. The PAMP-INV ILP can usually be solved to optimality within a few hours, and the runtime is relatively robust to changes in dimerization rate, candidate primer density, and constraints on multiplexing degree. The detection probability varies from 75% to over 99% depending on instance parameters.

Unfortunately the runtime for solving the PAMP-DEL ILP in Section 3 is impractical for all but very small problem instances. In contrast, the PAMP-1SDEL ILP can be solved efficiently for very large instances. Therefore, we considered a practical PAMP-DEL heuristic which relies on iteratively solving simpler PAMP-1SDEL instances, as follows. First, we solve a PAMP-1SDEL for one side – say, for reverse primers – assuming that the position of the other deletion endpoint is right next to the anchor. Then we solve a PAMP-1SDEL for selecting a set of forward primers from candidates that do not dimerize with the already selected reverse primers. The PAMP-1SDEL ILP in this second step is as in Section 3, however, coefficients $C_{i,j}$ in (11) represent the two-sided failure probability reflecting the fixed set of reverse primers. The process is repeated until there is no further decrease in failure probability.

The above iterative heuristic is referred to as ITERATED-1SDEL. One drawback of ITERATED-1SDEL is that it may result in unbalanced sets of primers for high dimerization rates. This happens since the first step will typically select the maximum possible number of reverse primers, and this may leave very few non-dimerizing forward primers. To avoid this drawback, we have also implemented a version of ITERATED-1SDEL, referred to as INCREMENTAL-1SDEL, which in the first iteration limits the number of selected reverse and forward primers to some proportional number of the given bounds N_r and N_f, for example, half of the given bounds, then increments these limits by a fixed factor in each of the subsequent iterations.

Table 2 compares the detection probability and average number of forward and reverse primers selected using the PAMP-DEL ILP, ITERATED-1SDEL, and INCREMENTAL-1SDEL on a set of small randomly generated instances for which the PAMP-DEL ILP

Table 3. Detection probability and runtime (in seconds) of INCREMENTAL-1SDEL.

Instance size	Dimer. Rate (%)	N=55	N=44	N=33	N=22
$2 \times 200Kb$ $n = 55$	0	93.24 (3902.16)	93.23 (3901.92)	93.02 (3901.68)	91.73 (3900.54)
	1	91.91 (93.80)	91.91 (93.70)	91.89 (93.60)	90.86 (93.40)
	2	90.54 (12.24)	90.54 (12.14)	90.54 (12.04)	89.90 (11.94)
	3	86.40 (5.58)	86.40 (5.50)	86.40 (5.42)	86.05 (5.34)
	4	82.68 (5.36)	82.68 (5.20)	82.68 (5.04)	82.56 (4.88)
	5	76.09 (2.46)	76.09 (2.40)	76.09 (2.34)	76.09 (2.28)
		N=105	N=84	N=63	N=42
$2 \times 400Kb$ $n = 105^*$	1	91.04 (1258.70)	91.04 (1258.22)	91.04 (1257.74)	90.13 (1257.24)
	2	78.28 (56.48)	78.28 (55.90)	78.28 (55.32)	77.30 (54.74)
	3	65.88 (29.31)	65.88 (28.03)	65.88 (26.75)	65.86 (25.45)
	4	54.12 (89.33)	54.12 (85.39)	54.12 (81.45)	54.12 (76.43)
	5	54.66 (276.93)	54.66 (272.19)	41.87 (267.45)	41.22 (257.21)

Note: * runtime for 0 dimerization rate exceeded 48 hours.

can be solved in practical runtime. The results show that both ITERATED-1SDEL and INCREMENTAL-1SDEL solutions are very close to optimal for low dimerization rates. For larger dimerization rates INCREMENTAL-1SDEL detection probability is still close to optimal, while ITERATED-1SDEL detection probability degrades substantially. As shown in Table 3, the runtimes of INCREMENTAL-1SDEL remain practical for large random instances except for the largest instance with no dimerization rate.

5. Conclusions

In this paper we propose ILP formulations for selecting sets of PAMP primers with high probability of detecting genomic inversions and anchored deletions in cancer tumors. In ongoing work we are assessing the performance of our methods on real biological datasets[6] and exploring scalable heuristics and approximation algorithms for un-anchored deletion detection via PAMP.

References

1. Y.-T. Liu and D. Carson, *PLoS ONE* **2**, p. e380 (2007).
2. K. Doi and H. Imai, *Genome Informatics* **10**, 73 (1999).
3. K. Konwar, I. Mǎndoiu, A. Russell and A. Shvartsman, Improved algorithms for multiplex PCR primer set selection with amplification length constraints, in *Proc. 3rd Asia-Pacific Bioinformatics Conference (APBC)*, 2005.
4. P. Nicodème and J.-M. Steyaert, Selecting optimal oligonucleotide primers for multiplex PCR, in *Proc. 5th Intl. Conference on Intelligent Systems for Molecular Biology*, 1997.
5. S. Rozen and H. Skaletsky, Primer3 on the WWW for general users and for biologist programmers, in *Bioinformatics Methods and Protocols: Methods in Molecular Biology*, eds. S. Krawetz and S. Misener (Humana Press, Totowa, NJ, 2000).
6. A. Bashir, Y.-T. Liu, B. Raphael, D. Carson and V. Bafna, *Bioinformatics* (Advance Access published online on August 30, 2007).
7. S. Khot and O. Regev, Vertex cover might be hard to approximate to within 2-ε, in *Proc. 18th IEEE Annual Conference on Computational Complexity*, 2003.