

# A Review of the Primer Approximation Multiplex PCR (PAMP) Technique for Detecting Large Scale Cancer Genomic Lesions

Kedsuda Apichonbancha<sup>1</sup>, Bhaskar Dasgupta<sup>1,\*</sup>, Jin Jun<sup>2</sup>, Ion Mandoiu<sup>2</sup>, and Emma Mendonca<sup>3</sup>

<sup>1</sup>Department of Computer Science, University of Illinois at Chicago, Chicago, IL 60607-7053, USA; <sup>2</sup>Computer Science & Engineering Department, University of Connecticut, Storrs, CT 06269-2155, USA; <sup>3</sup>Department of Bioengineering, University of Illinois at Chicago, Chicago, IL 60607-7053, USA

**Abstract:** 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 chapter we will first review previous solutions for the problem of selecting sets of PAMP primers that minimize detection failure probability and subsequently review our approach based on integer programming formulations for inversion and deletion detections.

**Keywords:** Primer approximation multiplex pcr, cancer genomic lesions, efficient algorithms.

## 1. INTRODUCTION

A major widely known cause of cancer is the abnormalities in the genetic material of the transformed cells. More widely accepted is the fact that these abnormalities are a result of the accumulation of genetic alterations in oncogenes and tumor suppressor genes, followed by clonal evolution [1]. Hence, the success rate of the prognosis of many cancer treatments depends for the most part on early detection and notably on the often microscopically undetectable tumor cells (MRD) [2]. Currently, two main experimental approaches are in vogue to address these issues: 1) antigen-antibody interaction and 2) amplified nucleic acids [1].

Among the two, the expectations with amplification methods, especially PCR is very high as it has the potential to amplify minute amounts of DNA more than 1 million-fold. DNA is an ideal substrate for molecular diagnosis because it readily survives the adverse conditions experienced by many clinical specimens and it can be rapidly amplified by polymerase chain reaction (PCR)-based techniques, thus diminishing the amount of starting material needed [4]. Therefore, the genetic alterations that arise during tumorigenesis can be used as targets for detection of cancer cells in clinical samples where PCR-based methods can detect low numbers of tumor cells in the presence of excess of normal cells [3, 4]. Given these arguments, a trigger has been set for novel sensitive technologies such as RT-PCR or digital PCR [5] to name a few for diagnosing cancer.

Existing well-known DNA-based approaches to detect genomic changes include Southern blotting [6], fluorescent in situ hybridization (FISH) [7], quantitative PCR [7-11], and array-CGH [12]. Application of these methods as a diagnostic tool to detect genetic alterations is based on a negative result or as noted by Liu and Carson [13] as absence of a detectable wildtype signal. On the other hand PAMP gives a positive signal based on the pairing between the primer

and the targeted genomic changes like deletion or insertion. Therefore the result is more reliable.

The second aspect to be considered in these methods is the sensitivity of the protocol. Tissue specimens may contain heterogeneous cell populations, which may further decrease the ability to detect copy number change in genes in the aberrant tumor cells because the population may contain normal cells. Furthermore, the use of tissue from clinical specimens severely limits the amount of DNA available for analysis. Given these kind of samples most of these methods give a poor performance/resolution for detection of genomic lesions. PAMP, however, has the potential to detect genomic changes of interest even in the presence of minute quantities of cancerous cells in the sample complemented with accurate mapping of the genomic breakpoint.

In comparison to the other PCR approaches the detection ability of FISH is not subjected to sample quality or quantity of tissue specimens and is able to detect varied genomic rearrangements like translocations or inversions. However, the process of preparing probes for FISH is complex due to the fact that it is necessary to tailor the probes to identify specific sequences of DNA. Also, it is difficult to count total numbers in probe-stained clusters of cells. Hence it becomes an impractical technique for high-throughput analysis. Moreover, if other approaches are considered like genome sequencing techniques like ESP (End sequencing profiling) [14] the cost of a whole genome analysis is very high. Also, it is not well-understood how to restrict these methods to lesions of interest in the genome.

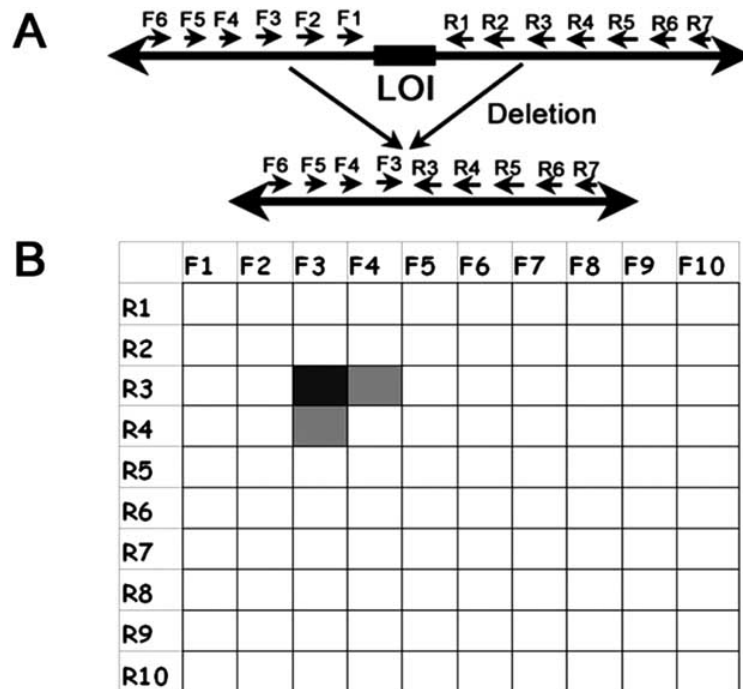
In this review paper, we hereby describe the unique and simplistic approach of Primer Approximation Multiplex PCR (PAMP) which is able to enrich small amounts of altered DNA in the presence of wild type DNA.

## 2. WHAT IS PAMP?

### 2.1. Experimental Method

PAMP is a novel multiplex primer technique designed by Liu and Carson [13] which allows for the assaying of many possible lesion boundaries at once. As observed in Fig. (1), it utilizes a set of primers which focus on amplifying the spe-

\*Address correspondence to this author at the Department of Computer Science, University of Illinois at Chicago, Chicago, IL 60607-7053, USA; Tel: 312-355-1319; Fax: 312-413-0024; E-mail: dasgupta@cs.uic.edu



**Fig. (1).** [Liu and Carson D, PLoS ONE 2007, <http://www.plosone.org/article/info:doi%2F10.1371%2Fjournal.pone.0000380>] Schematic of PAMP design. Pairs of forward and reverse primers flank the left and right breakpoints of the fusing genomic regions. The amplified product is detected by hybridization to probes on an array, the dark spots on the array correspond to amplification of primers most proximal to the breakpoint.

cific regions of genomic DNA where the precise breakpoints of alterations such as deletions or translocations occur. A multiple primer set is required as these breakpoints may vary between patients. A pair of forward and reverse primer can amplify a region of interest only if they are brought in close proximity of each other due to a genomic lesion. The resulting amplicon can then be assayed on an array leading to identification of the precise breakpoint. Given the experimental conditions, analysis of the array indicates colored spots corresponding to the expected breakpoints. This technique offers precise mappings of a genomic alteration with resolution of less than 1kb.

Primer approximation PCR screening is already in use for isolation of deletion mutants in *C.elegans* [15]. The results of the PAMP technique is completely based on the observation of a band brought about by a successful PCR reaction when a genomic change has occurred. However, since primers corresponding to multiple regions are used, it is difficult to discern the approximate location of the genomic region of interest. Hence, this raises the possibility of false positive results. To overcome this problem, coupling of PCR with Genomic Tiling array can concurrently increase the coverage area of the genomic region. Assuming that the tiling array covers one of the endpoints in the neighbor regions of the genomic lesion, the identification of the breakpoint then becomes straightforward. Moreover, the commercial availability of tiling arrays further increases the feasibility of this approach.

This PAMP technique was successfully designed and implemented for the CDKN2A locus [13]. Deletion of CDKN2A locus is major cause of chromosomal reassembly in human cancers. Precise breakpoints were mapped in De-

troit 562 cell lines using PAMP for contaminated samples. It was capable of detecting genomic DNA deletions in the presence of more than 99.9% wild type DNA.

Nevertheless, success of any PCR-based method is highly dependent on its primer design which includes problems such as self-complementarity or primer length. Combining these issues with constraints of accurately covering the cancer prone genomic breakpoint gives rise to an interesting optimization problem. The optimal primer design for PAMP should include primers which adequately cover the region of interest while avoiding dimerization between the primer pairs and satisfying the physico-chemical constraints of a PCR technique. Therefore, the focus of this review paper is to put together the computational formulation of the problem, complexity and the solution designed by Bashir *et al.* 2007 [16] in Section 2 followed by further improvements and methods implemented by Dasgupta *et al.* 2007 [17] in Section 3.

## 2.2. Formal Definitions of a Basic Version of PAMP Primer Design Problem

Consider a set of forward primers  $F_n, \dots, F_2, F_1$  and reverse primers  $R_n, \dots, R_2, R_1$  as observed in Fig. (1). Let  $d$  be the maximum distance between a pair of forward and reverse primers that allows amplification. The objective of primer design here is that if the target region is deleted then the distance between a pair of  $F_i$  and  $R_j$  should be at the most  $2d$  so that the probability of that region being amplified is high. At the same time it should obey the following constraints: 1) no cross hybridization between primers, 2) primers in the same direction should be non-overlapping and 3) they should satisfy all the physiochemical parameters.

The first mathematical formulation of this type of problem was described by Bashir *et al.* 2007 [16] when one deletion end-point is known in advance (and thus we consider forward primers only). Consider a set of primers E consisting of primers that do not dimerize with each other. Look at the genomic location of two adjacent forward primers, denoted as  $l_{j1}$  and  $l_{j2}$ . If the distance between the two locations is less than or equal to  $d$  then any deletion with breakpoint between the two should cause an amplification. Therefore, we can assign a coverage cost of  $C(f_1, f_2) = \max\{0, |l_{j1} - l_{j2}| - d\}$  to each consecutive pair of primers. Now consider designing a chain of primers  $P = \{p_1, p_2, \dots, p_n\}$  with forward primers followed by a reverse primer such that  $l_{p_i} < l_{p_{i+1}}$  for all  $i$ . Cost of this design is to minimize

$$C(P) = \sum_{(p_i, p_j) \in E} w_p + \sum_j w_c C(p_j, p_{j+1})$$

where  $w_c$  is the coverage cost and  $w_p$  is the weight to retain some hybridizing pairs as eliminating all of them would lead to small primer set. Adjusting  $w_p$  controls the number of hybridizing pairs to be included.

Further, this problem can handle two more improvements. An intuitive understanding of experimental setup suggests that all the primers need not be included in one reaction. Therefore, dimerizing primers can be segregated into different sets. So now in the definition of the problem cross-hybridizing forward or reverse primers are allowed when are in different sets  $N$ . Second interesting issue here is the total number of primers required to cover a genomic region. Hence a parameter primer density  $p$  is introduced which is the average number of primers every  $d$  base pairs. Consequently,  $p \geq 1$  holds for a complete coverage of the target region. A small change in the problem design helps us to control primer density,

$$\text{Minimize } C(P) = \sum_{(p_i, p_j) \in E} w_p + \sum_j w_c C(p_j, p_{j+1}) + w_p P$$

$w_p$  being  $\infty$  if  $p$  exceeds the desired density else  $w_p$  is set to zero.

The above PAMP design in restricted form was proven to be NP-hard via reduction from Max-2SAT [16]. This restricted form is known as the One sided PAMP design (OPAMP) problem where the breakpoint was exactly known along with a set of either forward or reverse primers targeting that region.

## 2.3. Solution for each problem in PAMP

### 2.3.1. Cross-Hybridization of Primers

To begin with we need to know the possible primers which dimerize with each other. In computational terms, it is interpreted as finding a set  $E$  with conflict edges. Previous studies [16, 18] suggest while aligning two primers if the ungapped alignment has more matches than mismatches ( $\geq 7$ ) preferable in 3' region then the probability of dimerization between these primers is high. Hence, this is the conflict criteria. To compute the conflict graph, a hash table of 3-mers is created. Primers that hash to the same table are aligned to compute  $E$ .

Also, it is not necessary that all forward and reverse primers are to be included in a single set of the multiplex

PCR reaction. Dimerizing forward and reverse primers can be partitioned into different sets to avoid cross-hybridization. In this way, more primers can be employed to increase the chances of detection of genomic lesions while allowing hybridization among forward or reverse primer sets.

However, the segregation of cross-hybridizing primers can increase the cost of the experimental setup. However the cost of this setup is directly proportional to the length of the genomic region of interest and on the accuracy of the knowledge required for the precise breakpoint. For example, if the scope of the experiment is only to confirm the presence or absence of deletion of well-studied specific genomic regions then the primer design can be standardized to be restricted to unique primers or relatively small set of PCR reactions. In case of diagnostic assay, only the knowledge of the approximate breakpoint is required and therefore the cost can be feasible. On the other hand, for an experimental setup where there is none or less information about the possible breakpoints or the length of genomic region, the number of primers required is high thereby making the experiment expensive in terms of labor or cost. However, once the information is known we will require fewer primers for the subsequent experiments. Hence, the cost of the experiment is inversely balanced by the amount of information about the possible breakpoints in the genomic region of interest.

### 2.3.2. Selection of Unique Primers

A logical approach to select unique primers is to filter the repeats. However in this case, to get more coverage area some primers are chosen from the repeat region. For filtering criteria, parameters were adopted from Wang *et al.* 2005 [10]. For each repeat in the target region, make a hash table 20-mer denoting its raw occurrence in the genome, as well as the occurrence of the 13 bp sub-string from its 3'-end. After satisfying standard primer criteria, a primer is selected if it did not have its 3'-end occur more than was expected by chance. Thereafter, recheck the resulting sequence set rigorously for uniqueness in the target region. Hence, now a list of unique and probable dimerizing primers is present from which low cost candidate pairs need to be chosen for PAMP design.

### 2.3.3. Selection of Low Cost Candidate Primers

Due to the complexity of the problem, three different approaches are used to select the optimal primers. They are greedy heuristic and simulated annealing (fast but sacrifice on optimality) and Integer Linear Programming (slow but selects optimal primers).

#### 2.3.3.1. Greedy Heuristic

In this method, a low cost primer set is extended using a greedy approach. Define  $P_u$  as the chain whose penultimate primer is  $u$  (the primer at  $l_u$  being the last) with cost  $C(P_u)$ .  $E_u$  corresponds to the set of primers which has dimerizing edges with  $u$ .

$$C(P_u) = \min_{v: l_v < l_u} \{C(P_v + \{u\})\}$$

$$P_v = \arg \min \{C(P_v + \{u\})\}$$

$$v: l_v < l_u$$

$$P_u = P_{v^*} + \{u\}$$

$P_u$  with the minimum cost is the solution. Disadvantage of this approach is that one may get unevenly distributed Primer set, with bias towards regions that were looked at first.

### 2.3.3.2. Simulated Annealing

In this approach, an initial random solution is required. Then at each iteration the present cost  $C(P')$  is compared to the initial cost  $C(P)$ . Here, the cost of each transition is  $\Delta_S = C(P') - C(P)$ . If the cost is lower it is accepted else it is accepted with probability proportional to  $e^{-\Delta_S/T}$ , where the temperature  $T$  is an adjustable parameter. Simulated annealing approach attempts to sample all putative solutions in the solution space. Two solution spaces with  $w_p = \infty$  and  $w_p < \infty$  with its neighborhood are considered. While  $w_p = \infty$  every candidate-set  $P$  induces an independent set, i. e., no pair of dimerizing primers is allowed.  $P'$  is in the neighborhood of  $P$  ( $P' \in N_p$ ) if there exists a primer  $u$  such that  $P' = P + \{u\} - \{v\}$  ( $(u,v) \in E$ ). In the case where  $w_p < \infty$ , each subset  $P$  has a size constraint on it such that  $P' \in N_p$  and  $|P - P'| \leq 1$ ;  $P'$  can be obtained from  $P$  by deleting or adding a primer. Both these cases have different convergence properties.

### 2.3.3.3. Integer Linear Programming

Here, the optimization problem is formulated as binary integer linear program (ILP). Every primer is represented as a binary variable  $x_i$  where  $i$  is the start location of the primer. Here, the objective is to minimize the uncovered region  $d_i$ .

$$\begin{aligned} & \min \sum_i d_i \\ & s.t. \\ & x_i + x_j \leq 1 \\ & x_i + x_j \leq 1 \\ & \sum_i x_i \leq p * \frac{L_{\max}}{d} \\ & q_{ij} \leq x_j \\ & \sum_{j < i} q_{ij} \geq x_i \\ & d_i \geq \max\{0, \sum_{ij < li} (l_i - l_j - d) q_{ij}\} \\ & q_{ij}, x_i \in \{0,1\} \end{aligned}$$

In the above formulation,  $x_i = 1$  holds if primer starting at location  $l_i$  is chosen and  $L$  is the length of the region of interest. For each dimerizing primer pair  $i$  and  $j$  we put the constraint  $x_i + x_j \leq 1$ . If both primers  $i$  and  $j$  are selected then  $q_{ij} = 1$  contributing to the cost therefore lower and upper bounds are set to it. In the constraint for penalty for uncovered regions when  $l_i - l_j - d < 0$ , its value is replaced by 0. Assign  $q_{ij} = 1$  to minimize the penalty when the primer  $j$  is selected before primer  $i$ . This converts  $d_i$  to exactly the same number of uncovered bases.

Comparing the above three methods, it is observed that simulated annealing outperforms the greedy approach. Therefore the next choice is between simulated annealing and Integer Linear Programming (ILP) methods. Based on the implementations of Bashir *et al.* it was observed that

though ILP gives optimal solution in theory the size of the ILP is enormous. Simulated Annealing on the other hand converges faster but giving an approximate solution. Hence, present efforts are concentrated on improving the ILP or designing other heuristics to find an optimal or near-optimal solution in reasonable running time. A progress in this direction was done by the work of Dasgupta *et al.* which is described in the next section.

## 3. A NEW ILP APPROACH FOR DETECTING DELETIONS AND INVERSIONS VIA PAMP

Unlike the one proposed by Bashir *et al.* [16] which focused on attempting to minimize the coverage cost related to uncovered region of interest, DasGupta *et al.* [17] proposed the new optimization objective and the corresponding ILP formulations of PAMP primer selection for detecting genomic rearrangement of both deletions as well as inversions. The proposed objective is to minimize the so called the *probability of failure*, namely the probability that an unknown genomic rearrangement will not be amplified by the PAMP assay. This objective is to be contrasted with the objective of Bashir *et al.* [16] which does not actually make explicit the underlying probabilistic distribution for the endpoints of the deletion but rather uses an objective that intuitively minimizes the proportion of the uncovered area. It is not difficult to see that minimizing uncovered area may not result in minimizing the probability of failure even assuming a uniform probability distribution for the deletion endpoints; see [17] for details.

### 3.1. Anchored Deletion Detection

The PAMP primer selection problem for deletion detection called PAMP-DEL and one-sided version of PAMP-DEL called PAMP-1SDEL when one of the deletion endpoints is recognized in advance were firstly introduced in [16], with the assumption that the deletion spans a known genomic location. In [17], the new ILP formulations for these two PAMP primer selection problems for detecting anchored deletions were introduced appropriately. It was proved in [17] that PAMP-DEL cannot be approximated to within a factor of  $2 - \epsilon$  for any constant  $\epsilon > 0$  by reducing the inapproximability result of the vertex cover problem proved in [19] under the assuming the UNIQUE GAMES conjecture to PAMP-1SDEL. Furthermore, it was also proved that there is a 2-approximation algorithm for a special case of PAMP-1SDEL in which candidate primers are spaced sufficiently far apart and the deletion endpoint is distributed uniformly within a fixed interval.

#### PAMP-DEL

To simulate the optimization problem of PAMP primer selection in [17], the model of 0-1 step function is assumed for the amplification of PCR product, namely that the probability of obtaining the amplification of two opposite strands (forward and reverse primers) is 1 if their distance is within  $L$  bases apart and 0 otherwise. However, the approach will work for other models, such as the exponential decay model. For the probabilistic model of lesion location, the uniform distribution, that is when a lesion with breakpoints  $l$  and  $r$  is equally likely over all  $l$  and  $r$ , is assumed, though other non-uniform distributions will work as well. For each pair of

breakpoints, its probability of occurring such pair is denoted by  $p_{l,r}$ .

PAMP-DEL can now be formulated as follows. Given the sets of forward candidate primers  $\{p_1, p_2, \dots, p_m\}$  and reverse candidate primers  $\{q_1, q_2, \dots, q_n\}$  which are indexed by increasing distance from the deletion anchor, the set  $E$  of pairs of primers that form cross-hybridization, maximum multiplexing degree  $N_f$  and  $N_r$ , and amplification length upper-bound  $L$ , our goal is to find the subset  $P'$  of at most  $N_f$  forward and at most  $N_r$  reverse primers such that  $P'$  does not include primer pairs in  $E$  and *minimizes* the probability of failure

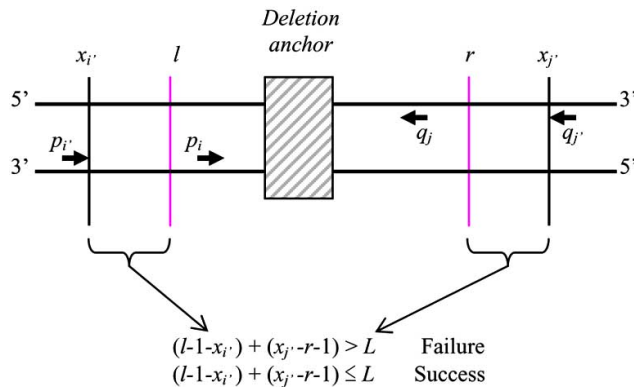
$$\sum_{x_{\min} \leq l, r \leq x_{\max}} f(P'; l, r) p_{l,r}$$

where  $f(P'; l, r) = 1$  if  $P'$  fails to get the amplification when the deletion with breakpoints  $(l, r)$  is present and  $f(P'; l, r) = 0$  otherwise; (see Fig. 2) for an illustration.

Next, to formulate PAMP-DEL as ILP, the dummy forward primers ( $p_0$  and  $p_{m+1}$ ) and reverse primers ( $q_0$  and  $q_{n+1}$ ) are introduced and assumed to uniquely hybridize at locations  $x_0 = x_{\min} - L$  and  $x_{m+1} = x_{\max} + L$  on the forward strand, respectively  $x_0 = x_{\min} - L$  and  $x_{n+1} = x_{\max} + L$  on the reverse strand. Furthermore, they are assumed not to dimerize with each other and with other candidate primers, and therefore they are definitely included in  $P'$ . Suppose the deletion with breakpoints  $(l, r)$ ,  $p_i$  and  $q_j \in P'$ , and the location  $x_i < x_j$  are given. On one hand, if  $(l-1-x_i) + (x_j-r-1) > L$  then  $P'$  fails to yield the PCR amplification. On the other hand, if  $(l-1-x_i) + (x_j-r-1) \leq L$ , at least one amplification product is obtained by  $P'$ .

For every quadruple  $(i, i', j, j')$ ,  $i \leq i'$ ,  $j \leq j'$ , let  $C_{i,i',j,j'}$  be the total probability that forward primers  $(p_i, p_{i'})$  and reverse primers  $(q_j, q_{j'})$  fail to produce the amplification when the deletion, with one end located between the sites of  $p_i$  and  $p_{i'}$ , and the other end located between the sites of  $q_j$  and  $q_{j'}$ , is present. The 0/1 variables using in ILP formulation are defined as follows:

- $f_i(r_i)$  is set to 1 if  $p_i$  (respectively  $q_i$ ) is selected in  $P'$  and to 0 otherwise,



**Fig. (2).** The PCR amplification succeeds when the deletion with endpoints  $l$  and  $r$  takes place and brings the forward primer  $p_i$  and reverse primer  $q_i$ , uniquely hybridized at location  $x_i$  and  $x_j$ , respectively, into proximity within  $L$  bases apart.

- $f_{i,j}(r_{i,j})$  is set to 1 if  $p_i$  and  $p_j$  (respectively  $q_i$  and  $q_j$ ) are consecutive forward (respectively reverse) primers in  $P'$  and to 0 otherwise,
- $e_{i,i',j,j'}$  is set to 1 if both  $(p_i, p_{i'})$  and  $(q_j, q_{j'})$  are pairs of consecutive forward and reverse primers in  $P'$  for any  $i \leq j$  and to 0 otherwise.

The ILP optimization formulation of PAMP-DEL corresponding to the proposed objective can then be obtained as below:

$$\text{Minimize} \quad \sum_{(i,i',j,j'): i \leq i', j < j', i \leq j} C_{i,i',j,j'} e_{i,i',j,j'} \quad (1)$$

$$\text{Subject to:} \quad f_{i,i'} + r_{j,j'} \geq 2e_{i,i',j,j'}, \quad i < i' \text{ and } j < j' \quad (1.1)$$

$$e_{i,i',j,j'} \geq f_{i,i'} + r_{j,j'} - 1, \quad i < i' \text{ and } j < j' \quad (1.2)$$

$$f_i + f_j \geq 2f_{i,j}, \quad 1 \leq i < j \leq m \quad (1.3)$$

$$r_i + r_j \geq 2r_{i,j}, \quad 1 \leq i < j \leq n \quad (1.4)$$

$$\sum_{j=1}^{m+1} f_{0,j} = \sum_{i=0}^m f_{i,m+1} = \sum_{j=1}^{n+1} r_{0,j} = \sum_{i=0}^n r_{i,n+1} = 1 \quad (1.5)$$

$$\sum_{i=0}^{j-1} f_{i,j} = \sum_{k=j+1}^{m+1} f_{j,k}, \quad 1 \leq j \leq m \quad (1.6)$$

$$\sum_{i=0}^{j-1} r_{i,j} = \sum_{k=j+1}^{n+1} r_{j,k}, \quad 1 \leq j \leq n \quad (1.7)$$

$$\sum_{1 \leq i \leq m} f_i \leq N_f, \quad \sum_{1 \leq i \leq n} r_i \leq N_r \quad (1.8)$$

$$f_i + f_j \leq 1, \text{ for all } (p_i, p_j) \in E \quad (1.9)$$

$$r_i + r_j \leq 1, \forall (q_i, q_j) \in E \quad (1.10)$$

$$f_i + r_j \leq 1, \text{ for all } (p_i, q_j) \in E \quad (1.11)$$

$$e_{i,i',j,j'} \in \{0,1\}, f_{i,j}, r_{i,j} \in \{0,1\}, f_i, r_i \in \{0,1\} \quad (1.12)$$

Constraints (1.1) and (1.2) are compatibility constraints attempting to guarantee that a variable  $e_{i,i',j,j'}$  is set to 1 if and only if  $f_{i,i'}$  and  $r_{j,j'}$  are both set to 1. Similarly, the constraint (1.3) and (1.4) can be explained by the same idea for both sorts of primers. Constraints (1.5) – (1.7) can be considered as the path connecting constraints for the forward and reverse primers such that primers of each type in  $P'$  are linked in left-to-right order. The limitations for both kind of primers allowed in the selected set  $P'$  (maximum multiplex degree  $N_f$  and  $N_r$ ) are treated by constraint (1.8). Lastly, constraints (1.9) – (1.11) insure that no pair of chosen primers in  $P'$  can cause a dimerization.

### PAMP-ISDEL

As mentioned earlier, PAMP-ISDEL was a special case of PAMP-DEL in which one endpoint is recognized in advance. In other words, for the sake of simplicity, it can be said that the reverse primer is already known relative to the known deletion. Hence, the ILP formulation for PAMP-ISDEL can be obtained by considerably simplifying the

PAMP-DEL ILP through focusing on the reverse candidate primers solely and the ILP formulation of PAMP-1SDEL was able to handle substantially larger-sized instances as compared to the ILP formulation of PAMP-DEL.

### 3.2. Inversion Detection: PAMP-INV

The ILP formulation of PAMP primer selection problem for detecting inversions, referred as PAMP-INV was initially proposed by [17]. The similar concept using in PAMP-DEL ILP can be applied to PAMP-INV ILP. However, only one set of candidate primers which all lie in the same orientation, is considered. When an inversion takes place, it causes the primers hybridizing at unique loci of an inversion region to lie in the opposite orientation, and consequently, bring the primer  $p_i$  and  $p_j$  into proximity. The amplification product is then generated when an inversion leads binding sites of these two primers  $p_i$  and  $p_j$  within  $L$  bases apart.

Like PAMP-DEL, the objective of PAMP-INV is to seek a set of non-dimerizing primers that produces at least one amplification product when an inversion is present in a specified region, as well as subject to the condition, minimize the probability that the selected PCR primers fail to result in the amplification when an inversion is present. Thus, the generalization of the optimization formulation and its constraints for PAMP-DEL ILP can be applied and interpreted in the same manner toward PAMP-INV ILP.

### 3.3. Summary of Experimental Results

To experimental evaluate these ILP approaches, the following experimental setup was used by DasGupta *et al.* [17]. For inversion detection, the ILP approach was tested on randomly generated instances of 100Kb long sequences with  $L=20$ Kb (which is representative of long-range PCR), number of candidate primers between 20 and 30 (candidate primer density between 3.33 and 5), maximum multiplexing degree 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. 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 and is relatively insensitive to the length of the inversion.

A similar experimental setup was used for the deletion detection as well. Unfortunately the runtime for solving the PAMP-DEL ILP was impractical for all but very small problem instances. In contrast, the PAMP-1SDEL ILP can be solved efficiently for very large instances. Therefore, they considered a practical PAMP-DEL heuristic (called ITERATED-1SDEL) which relies on iteratively solving simpler PAMP-1SDEL instances. One drawback of ITERATED-1SDEL is that it may result in unbalanced sets of primers for high dimerization rates. To avoid this drawback, they 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  $N_f$ . Simulation results showed that both ITERATED-1SDEL and IN-

CREMENTAL-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.

## 4. CONCLUSION AND FUTURE WORK

We have reviewed a PCR-based PAMP that covers a genomic region of interest with unique non-dimerizing primers. PAMP has the potential to be applied in varied setup in experimental and diagnostic area for study of genomic alterations. It is a promising technique for the age of personalized medicine as it can approximate map the genomic breakpoint allowing it to target specific biomarkers. Successful multiplex PCR with more than 1000 primer pairs have already been carried out [20]. Therefore the cost and labor is markedly reduced. The ultimate goal of PAMP is to be used as a diagnostic assay for cancer patients. Once the primers are selected for a specific genomic region, the simplistic technique of PAMP can be readily assimilated in a robotic environment. However, to achieve this, we need to have a > 95% success rate in detecting the altered genomic region. Therefore, the current efforts are concentrated on finding a computational method for solving the primer design problem. Presently the ILP approach proposed by Bashir *et al.* does not necessarily minimize the failure probability and works for one-sided deletion only; in contrast the work of Dasgupta *et al.* does minimize the failure probability for the more general version of the problem but has a practical runtime for small and medium size instances. Hence, a future direction is to look at more scalable heuristics and approximation algorithms to solve the PAMP primer design problem in practical time and high accuracy. Also, the focus area of PAMP can be shifted from the detection of inversions and anchored deletions to other more challenging problems such as unanchored deletions. This would definitely propel PAMP from the experimental laboratories to real life diagnostic assay.

## 5. ACKNOWLEDGEMENTS

Apichonbancha, DasGupta and Mendenco were supported by NSF grants IIS-0346973, IIS-0612044 and DBI-0543365. DasGupta also thankfully acknowledges support from the DIMACS computational epidemiology program. Jun and Mandoiu were supported by NSF grants IIS-0546457 and DBI-0543365.

## REFERENCES

- [1] van Houten VM, Tabor MP, van den Brekel MW, Denkers F, Wisshaupt RG, Kummer JA, Snow GB, Brakenhoff RH. Molecular assays for the diagnosis of minimal residual head-and-neck cancer: methods, reliability, pitfalls, and solutions. *Clin Cancer Res* **2000**; 6: 3803-16.
- [2] Leemans CR, Tiwari R, Nauta JJ, van der Waal I, Snow GB. Regional lymph node involvement and its significance in the development of distant metastases in head and neck carcinoma. *Cancer (Phila.)* **1993**; 71: 452-6.
- [3] Cairns P, Sidransky D. Molecular methods for the diagnosis of cancer. *Biochim Biophys Acta* **1999**; 1423: C11-C18.
- [4] Sidransky D. Nucleic acid-based methods for the detection of cancer. *Science (Washington DC)* **1997**; 278: 1054-9.
- [5] Vogelstein B, Kinzler KW. Digital PCR. *Proc Natl Acad Sci USA* **1999**; 96: 9236-41.

- [6] Petrij-Bosch A, Peelen T, van Vliet M, *et al.* BRCA1 genomic deletions are major founder mutations in Dutch breast cancer patients. *Nat Genet* **1997**; 17: 341-5.
- [7] Perry A, Nobori T, Ru N, *et al.* Detection of p16 gene deletions in gliomas: a comparison of fluorescence in situ hybridization (FISH) versus quantitative PCR. *J Neuropathol Exp Neurol* **1997**; 56: 999-1008.
- [8] Kees UR, Terry PA, Ford J, *et al.* Detection of hemizygous deletions in genomic DNA from leukaemia specimens for the diagnosis of patients. *Leuk Res* **2005**; 29: 165-71.
- [9] Carter TL, Watt PM, Kumar R, *et al.* Hemizygous p16 (INK4A) deletion in pediatric acute lymphoblastic leukemia predicts independent risk of relapse. *Blood* **2001**; 97: 572-4.
- [10] M'Soka TJ, Nishioka J, Taga A, *et al.* Detection of methylthioadenosine phosphorylase (MTAP) and p16 gene deletion in T cell acute lymphoblastic leukemia by real-time quantitative PCR assay. *Leukemia* **2000**; 14: 935-40.
- [11] Batova A, Diccianni MB, Nobori T, *et al.* Frequent deletion in the methylthioadenosine phosphorylase gene in T-cell acute lymphoblastic leukemia: strategies for enzyme-targeted therapy. *Blood* **1996**; 88: 3083-90.
- [12] Pinkel D, Albertson DG. Array comparative genomic hybridization and its applications in cancer. *Nat Genet* **2005**; 37 Suppl: S11-7.
- [13] Liu YT, Carson D. A novel approach for determining cancer genomic breakpoints in the presence of normal DNA. *PLoS ONE* **2007**; 2(4): e380.
- [14] Volik S, Zhao S, Chin K, *et al.* End-sequence profiling: sequence-based analysis of aberrant genomes. *Proc Natl Acad Sci USA* **2003**; 100: 7696-701.
- [15] Jansen G, Hazendonk E, Thijssen KL, Plasterk RH. Reverse genetics by chemical mutagenesis in *Caenorhabditis elegans*. *Nat Genet* **1997**; 17: 119-21.
- [16] Bashir A, Liu YT, Raphael B, Carson D, Bafna V. Optimization of primer design for the detection of variable genomic lesions in cancer. *Bioinformatics* **2007**; 23(21): 2807-15.
- [17] Dasgupta B, Jun J, Mandoiu I. Primer Selection Methods for Detection of Genomic Inversions and Deletions via PAMP. *Proc. 6th Asia-Pacific Bioinformatics Conference* **2008**; 353-62.
- [18] Vallone PM, Butler JM. Autodimer: a screening tool for primer-dimer and hairpin structures. *Biotechniques* **2004**; 37: 226-31.
- [19] Knot S, Regev O. Vertex cover might be hard to approximate to within  $2 - \epsilon$ . In: *Proc. 18<sup>th</sup> Annual IEEE Conference on Computational Complexity* **2003**; 379-86.
- [20] Wang HY, Luo M, Tereshchenko IV, *et al.* A genotyping system capable of simultaneously analyzing > 1000 single nucleotide polymorphisms in a haploid genome. *Genome Res* **2005**; 15: 276-83.