Integer Linear Programming in Designing Universal Arrays with Multiplexed Applications

Atif Rahman
Mahmuda Naznin
Masud Hasan

Department of Computer Science and Engineering
Bangladesh University of Engineering and Technology
Dhaka, Bangladesh

e-mail: atif.bd@gmail.com, mahmudanaznin@cse.buet.ac.bd, masudhasan@cse.buet.ac.bd

Abstract — Single Nucleotide Polymorphisms (SNPs) are variations in a single nucleotide within a generally conserved genomic sequence across the population. SNPs can be genotyped using a universal DNA tag array. But cross-hybridization involving assay-specific components disrupts this process. The problem of identifying the most economic experimental configuration of the assay specific components that avoids cross-hybridization can be formulated as an optimization problem. More specifically, the problem translates into the problem of covering the vertices of one side of a bipartite graph by a minimum number of balanced subgraphs of maximum degree one. The problem was shown to be NP-complete by Ben-Dor et al. [1] In this paper, we give integer linear programming formulation of this problem. We also provide an integer linear programming formulation of a mathematically related problem.

I. INTRODUCTION

Single Nucleotide Polymorphisms (SNPs) are variations in a single nucleotide within a generally conserved genomic sequence across the population [2]. Efficient SNP detection, genotyping, and measurement techniques are important issues in computational biology. The sequence variations due to SNPs often result in variations in phenotypic traits. This happens when the variation occurs in coding or regulatory regions [3]. SNPs in some regions determine cancer susceptibility and are often related to pathogenesis [4, 5, 6, 7]. SNPs also serve as genetic markers that can be used in mapping genomes. The process that determines the variants present in a given sample over a set of SNPs is known as genotyping.

SNP genotyping has been subject to a considerable amount of research and many different techniques have been developed [8, 9, 10, 11, 12]. In array-based hybridization assay a target specific set of oligonucleotides is synthesized or deposited on a solid surface such as silicon or glass. Then the target DNA or RNA fragments are fluorescently labeled and brought in contact with the surface and allowed to hybridize with the oligonucleotides. The content of the sample mixture can be inferred from the resulting fluorescence pattern. Although hybridization is supposed to occur only in sites on the surface that are Watson-Crick complements to some substring in the target, cross-hybridization makes determination of the pattern difficult.

Recently, an alternative approach based on universal arrays containing oligonucleotides called antitags was suggested by Brenner [13] and others [14, 15]. The Watson-Crick complement of an antitag is called a tag. The tag-antitag pairs are designed so that each tag hybridizes strongly to its complementary antitag, but not to other antitags. The system is referred to as DNA tag/antitag system and in short a DNA TAT system. In this approach, a set of reporter molecules are synthesized. Each molecule consists of a primer part that is the complement of the upstream sequence that immediately precedes the polymorphic site and a tag, an element of the universal set of tags, ligated together. Then the sample is mixed with the reporter molecules and hybridization is allowed to take place. If the specificity is perfect, the reporter molecules bind only at the sites they are designed for. Single dideoxynucleotides, fluorescently labeled with four distinct chemical dyes, are then added to the mixture. In a polymerase-driven reaction, each hybridized reporter molecule is extended by exactly one labeled dideoxynucleotide. When the molecules are brought into contact with the universal array, the tag part of each reporter molecule hybridizes to its complementary antitag on the array. The SNP variations can be determined by detecting the fluorescent dyes present at the site. A detailed illustration of the process is mentioned in [16]. This technique can also be used to determine allele frequencies and to analyze gene expression.

There is a tradeoff in designing DNA TAT systems. When thousands of SNPs are genotyped, it is desirable to have as many tags as possible, in order to maximize the number of SNPs that can be genotyped in parallel. How-
A set of SNPs can be measured in a single array operation without cross-hybridization if all corresponding members can be assigned to tags such that cross-hybridization is avoided. Each subset corresponds to one array experiment and proper partitions of $P$ into subsets correspond to the solution to the multiplexing problem.

**Definition 1.** Let $R = \{(p_1, t_1), \ldots, (p_k, t_k)\}$ be a set of reporter molecules with distinct $p_i \in P$ and distinct $t_j \in T$, $R$ is said to be non-cross-hybridizing if $A_{p_i,t_j} = 0$ for all $i \neq j$.

**Definition 2.** A set of $k$ distinct primers $P' = \{p_1, \ldots, p_k\} \subset P$ is called assignable if there exists a non-cross-hybridizing set of reporter molecules $\{(p_1, t_1), \ldots, (p_k, t_k)\}$ for $k$ distinct tags $t_1, \ldots, t_k \in T$. An assignable set of tags can be defined in the same way.

The input can be thought of as a bipartite graph $G = (P, T, A)$, whose vertices are primers ($P$) and tags ($T$) and whose edges represent potential cross-hybridization between primers and the corresponding antitags. Therefore $A = \{(p,t) : p \in P, t \in T, A_{p,t} = 1\}$. A subgraph $H = (P', T', E')$, of $G$ is called balanced if $|P'| = |T'|$. Subgraph $H$ is called an assignable subgraph if $H$ is a balanced induced subgraph of maximum degree 1.

**Definition 3.** A partition $\mathcal{E}$ of the primer set $P$ is called a primer cover if each $P' \in \mathcal{E}$ is assignable.

We know define the problem formally.

**Minimum Cover Problem (MPC) Given** a bipartite graph $G = (P, T, A)$, find a minimum primer cover of $P$.

A related problem is identifying the largest assignable subset in $P$.

**Maximum Assignable Primer-set (MAP) Given** a bipartite graph $G = (P, T, A)$, find a maximum assignable subgraph $H$ of $G$.

A mathematically related problem is optimal partitioning of bipartite graph into a set of vertex-disjoint assignable subgraphs that cover the set of primers.

**Minimum Partition into Disjoint Assignable Subgraphs (MPDAS)** Given a bipartite graph $G = (P, T, A)$, find a minimum set of vertex-disjoint assignable subgraphs that cover $P$.

## II. PRELIMINARIES

We denote the set of DNA tag sequences associated with the universal array by $T$ and the set of antitags by $\overline{T}$. The set of primers is denoted by $P$. Let $m = |P|$ and $n = |T|$. A reporter molecule is a primer-tag pair $(p, t)$, where $p \in P$ and $t \in T$. For a graph $G$ and a subset of its vertices $R$, we denote by $G_R$ the subgraph of $G$ induced by $R$. We denote by $V(G)$ and $E(G)$ the sets of vertices and edges of $G$, respectively. We assume that cross-hybridization potentials are given in the form of a binary $m \times n$ matrix $A$, such that

$$A_{p,t} = \begin{cases} 1 & \text{if } p \in P \text{ potentially hybridizes with } t \in T, \\ 0 & \text{otherwise.} \end{cases}$$

A set of SNPs can be measured in a single array operation without cross-hybridization if all corresponding members can be assigned to tags such that cross-hybridization is avoided.
They also proposed a greedy approach to MPC that mimics approximation algorithms for SET COVER [17]. In every step the largest assignable subset in $P$ is identified and removed and the algorithm proceeds recursively on the rest of the graph. So, each step requires a solution to the maximum assignable primer-set(MAP) problem. However, MAP was also shown to be NP-hard by Ben-Dor et al. [1] and they gave an integer linear programming (ILP) formulation of the problem.

Ben-Dor et al. [1] also proved the intractability of minimum partition into disjoint assignable subgraphs (MPDAS) and proposed an $O(d)$ approximation algorithm for $d$-bounded instance of the problem.

IV. INTEGER LINEAR PROGRAMMING FORMULATIONS

In this section we give integer linear programming (ILP) formulations of MPC and MPDAS.

A. Minimum Primer Cover (MPC)

We observe that a primer cover of $P$ can always be obtained by partitioning $P$ into $m$ subsets. In the ILP formulation we introduce binary variables $s_k$ for $k \in S = \{1, 2, \ldots , m\}$. $s_k = 1$, if and only if some $p \in P$ are assigned to $k$-th partition. We are interested in a partition where minimum number of $s_k$’s are 1. We also introduce binary variables $e_{p,t}^k$ for every primer-tag pair and for each partition. $e_{p,t}^k = 1$ if and only if in the solution $p \in P$ is matched with $t \in T$, and $p \in P$ is assigned to $k \in S$. The constraints subjected to as follows:

- If $s_k = 0$ for some $k \in S$ then no primer $p \in P$ can be assigned to it. Otherwise, $m$ primers can be assigned to it provided the constraints mentioned below are satisfied.

This is ensured by the constraint

$$s_k m - \sum_{p \in P, t \in T} e_{p,t}^k \geq 0 \text{ for all } k$$

- The $s_k$’s with value 1 form a partition over $P$. For this we need

  - Each $p \in P$ is assigned to at most one partition

    $$e_{p,t_1}^k + e_{p,t_2}^k \leq 1 \text{ for } k \neq j \text{ and any } t_1, t_2 \in T$$

  - Each $p \in P$ is assigned to exactly one partition

    $$\sum_{p \in P, t \in T, k \in S} e_{p,t}^k \geq m$$

- There must be 1-1 correspondence in each subset and no cross-hybridization between primers and tags in any of the subsets. This can be ensured by constraints analogous to the ILP formulation of MAP by Ben-Dor et al. [1]

  - To ensure that there is 1-1 correspondence between primers and tags

    $$e_{p_1,t}^k + e_{p_2,t}^k \leq 1 \text{ for } p_1 \neq p_2$$

    $$e_{p,t_1}^k + e_{p,t_2}^k \leq 1 \text{ for } t_1 \neq t_2$$

  - For no cross-hybridization we need

    $$e_{p,t}^k + e_{p,\tau}^k \leq 1 \text{ for } A_{p,t} = 1, p \neq \tau, \tau \neq t$$

So, the ILP formulation for MPC is

$$\min \sum_{k \in S} s_k$$

$$\text{s.t. } s_k m - \sum_{p \in P, t \in T} e_{p,t}^k \geq 0 \text{ for all } k$$

$$e_{p_1,t}^k + e_{p_2,t}^k \leq 1 \text{ for } k \neq j \text{ and any } t_1, t_2 \in T$$

$$\sum_{p \in P, t \in T, k \in S} e_{p,t}^k \geq m$$

$$e_{p,t_1}^k + e_{p,t_2}^k \leq 1 \text{ for } t_1 \neq t_2$$

$$e_{p,t}^k + e_{p,\tau}^k \leq 1 \text{ for } A_{p,t} = 1, p \neq \tau, \tau \neq t$$

$$s_k, e_{p,t}^k \in \{0, 1\}$$

B. Minimum Partition into Disjoint Assignable Subgraphs (MPDAS)

The ILP formulation of MPDAS is similar to that of MPC. We need an additional constraint

- No $t \in T$ is assigned to more than one subset

$$e_{p_1,t}^k + e_{p_2,t}^k \leq 1 \text{ for } k \neq j \text{ and any } p_1, p_2 \in P$$

The ILP formulation for MPDAS is

$$\min \sum_{k \in S} s_k$$

$$\text{s.t. } s_k m - \sum_{p \in P, t \in T} e_{p,t}^k \geq 0 \text{ for all } k$$

$$e_{p_1,t}^k + e_{p_2,t}^k \leq 1 \text{ for } k \neq j \text{ and any } t_1, t_2 \in T$$

$$\sum_{p \in P, t \in T, k \in S} e_{p,t}^k \geq m$$

$$e_{p_1,t}^k + e_{p_2,t}^k \leq 1 \text{ for } k_1 \neq k_2 \text{ and any } p_1, p_2 \in P$$

$$e_{p_1,t}^k + e_{p_2,t}^k \leq 1 \text{ for } t_1 \neq t_2$$

$$e_{p,t}^k + e_{p,\tau}^k \leq 1 \text{ for } A_{p,t} = 1, p \neq \tau, \tau \neq t$$

$$s_k, e_{p,t}^k \in \{0, 1\}$$
V. CONCLUSION

In this paper we presented ILP formulations for minimum primer cover (MPC) and minimum partition with disjoint assignable subgraphs (MPDAS). Both problems are NP-complete and no approximation algorithms are known so far. The complexity of \(d\)-bounded instances of these problems are open and no constant factor approximation algorithm has been developed as yet. An interesting variant of MPC arises if primer-primer cross-hybridization is also considered. The graph is no longer bipartite and the goal is to find a minimum primer cover so that there is no cross-hybridization.

ACKNOWLEDGMENT

This paper is an outcome of research conducted in the Department of Computer Science and Engineering of Bangladesh University of Engineering and Technology (BUET).

REFERENCES