JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 11, Numbers 2–3, 2004
© Mary Ann Liebert, Inc.
Pp. 477–493

Towards Optimally Multiplexed Applications of Universal Arrays*

AMIR BEN-DOR,[1] TZVIKA HARTMAN,[2] RICHARD M. KARP,[3]
BENNO SCHWIKOWSKI,[4] RODED SHARAN,[3] and ZOHAR YAKHINI[5]

ABSTRACT

We study a design and optimization problem that occurs, for example, when single nucleotide polymorphisms (SNPs) are to be genotyped using a universal DNA tag array. The problem of optimizing the universal array to avoid disruptive cross-hybridization between universal components of the system was addressed in previous work. Cross-hybridization can, however, also occur assay specifically, due to unwanted complementarity involving assay-specific components. Here we examine the problem of identifying the most economic experimental configuration of the assay-specific components that avoids cross-hybridization. Our formalization translates this problem into the problem of covering the vertices of one side of a bipartite graph by a minimum number of balanced subgraphs of maximum degree 1. We show that the general problem is NP-complete. However, in the real biological setting, the vertices that need to be covered have degrees bounded by d. We exploit this restriction and develop an $O(d)$-approximation algorithm for the problem. We also give an $O(d)$-approximation for a variant of the problem in which the covering subgraphs are required to be vertex disjoint. In addition, we propose a stochastic model for the input data and use it to prove a lower bound on the cover size. We complement our theoretical analysis by implementing two heuristic approaches and testing their performance on synthetic data as well as on simulated SNP data.

Key words: SNP, genotyping, multiplexing, universal array, cross-hybridization, graph algorithm.

1. INTRODUCTION

Single nucleotide polymorphisms (SNPs) are differences, across the population, in a single base, within an otherwise conserved genomic sequence (Wang *et al.*, 1998). The sequence variation repre-

*A preliminary version of this manuscript appeared in the *Proceedings of the Seventh Annual International Conference on Research in Computational Molecular Biology*, pages 48–56, 2003.

[1] Agilent Laboratories, 395 Page Mill Rd., Palo Alto, CA 94303.
[2] Department of Computer Science and Applied Mathematics, Weizmann Institute, Rehovot, Israel.
[3] International Computer Science Institute, 1947 Center Street, Suite 600, Berkeley, CA 94704.
[4] Institute for Systems Biology, 1441 N. 34th Street, Seattle, WA 98103.
[5] Agilent Laboratories and Computer Science Dept., Technion, Haifa, Israel.

sented by SNPs is often directly related to phenotypic traits. Such is the case when the variation occurs in coding or other functional (e.g., regulatory) regions (Cargill *et al.*, 1999). Somatic or native SNPs in oncogenes or in related regions can determine cancer susceptibility and are often related to pathogenesis (see, e.g., Venitt [1994, 1996], Kristensen *et al.* [2002], Watanabe *et al.* [2002]). *Genotyping* is a process that determines the variants present in a given sample, over a set of SNPs. SNPs also serve as genetic markers that can be used in linkage and association studies (see, e.g., Risch [2000]). In the latter case, a population of samples is jointly measured and the frequencies of the different variants are inferred. Efficient SNP detection, genotyping, and measurement techniques have, therefore, great clinical, scientific, and commercial value.

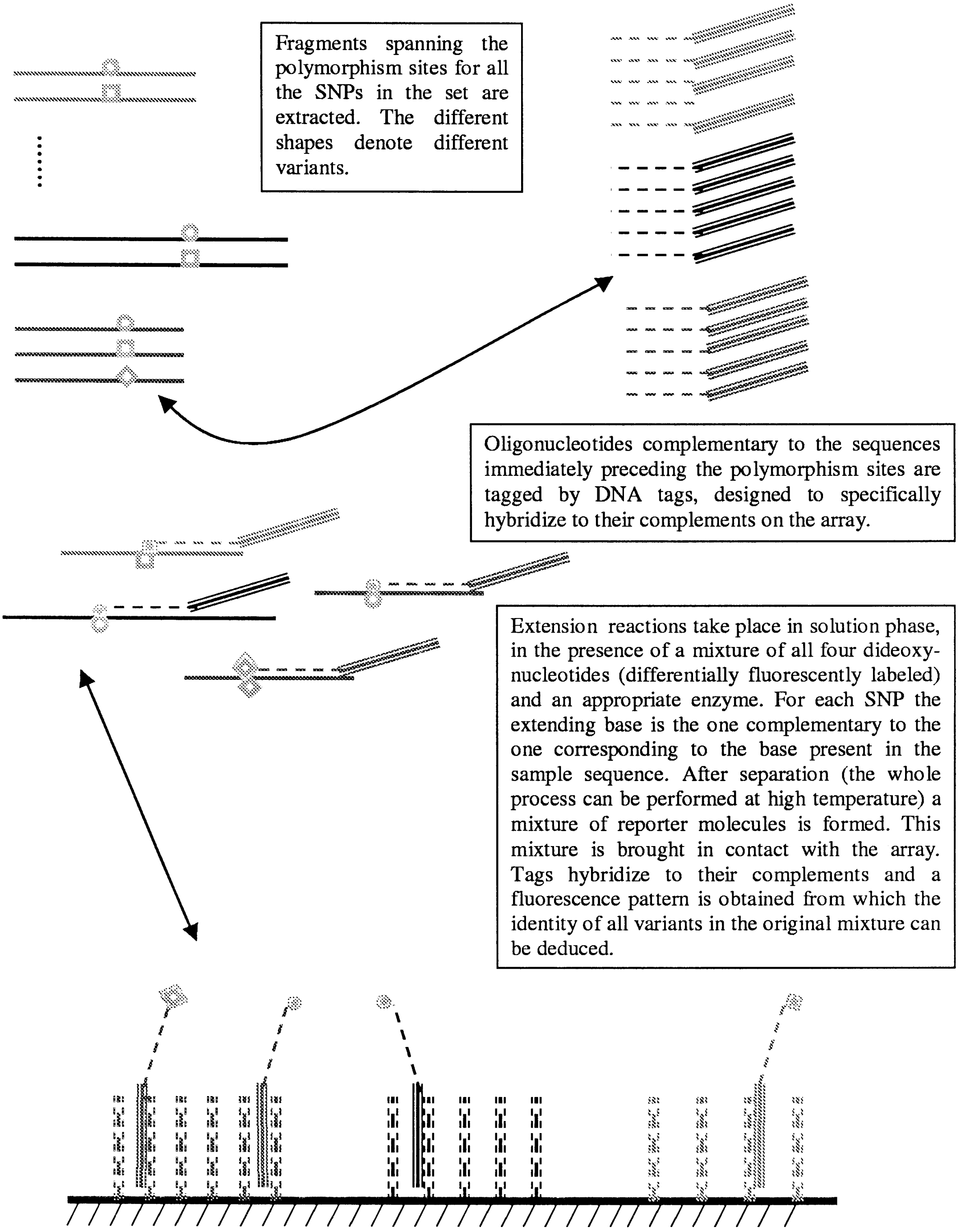
Methods for high throughput SNP genotyping are under fast development and evolution. This task enlists various molecular biology techniques, separation technologies, and detection methods. Methods based on mass spectrometry or length separation are described, e.g, in Grant and Phillips (2001) and Schouten *et al.* (2002). Other methods are based on hybridization array technology (Syvanen, 1999; Hacia, 1999; Drmanac *et al.*, 1991). In an array-based hybridization assay a target-specific set of oligonucleotides is synthesized or deposited on a solid support surface (e.g., silicon or glass). A fluorescently labeled target sample, a mixture of DNA or RNA fragments, is then brought in contact with the treated surface and allowed to hybridize with the surface oligonucleotides. Scanning the resulting fluorescence pattern reveals information about the content of the sample mixture. Theoretically, the assay conditions are such that hybridization occurs only in sites on the surface that are Watson–Crick complements to some substring in the target. In practice, cross-hybridization[1] is a main source of signal contamination in any array-based hybridization assay.

Recently, S. Brenner (1997) and others (Brenner, 1997; Davis *et al.*, 1997; Gerry *et al.*, 1999) suggested an alternative genotyping approach based on *universal arrays* containing oligonucleotides called *antitags*. The Watson–Crick complement of each antitag is called a *tag*. The tag–antitag pairs are designed so that each tag hybridizes strongly to its complementary antitag, but not to any other antitag. We shall call the entire system a *DNA tag/antitag system* and in short a *DNA TAT system*. To exemplify the approach, we describe in detail the application of universal arrays to SNP genotyping. The method is illustrated in Fig. 1 and consists of the following steps:

1. A set of reporter molecules (one for each SNP) is synthesized. Each reporter molecule consists of two parts that are ligated together. The *primer* part is the Watson–Crick complement of the upstream sequence that immediately precedes the polymorphic site of the SNP. The other part is a unique tag—an element of the universal set of tags.
2. When an individual is to be genotyped, a sample is prepared that contains the sequences flanking each of the SNP sites. Typically, these are PCR amplicons. The sample is mixed with the reporter molecules and solution-phase hybridization takes place. Assuming that specificity is perfect, the reporter molecules bind only at the sites they are designed for.
3. Single dideoxynucleotides, `ddA,ddC,ddT,ddG`, fluorescently labeled with four distinct chemical dyes, are added to the mixture. In a polymerase-driven reaction, each hybridized reporter molecule is extended by exactly one labeled dideoxynucleotide.
4. The extended reporter molecules are separated from the sample fragments and brought into contact with the universal array. Assuming that specificity is perfect, the tag part of each reporter molecule will hybridize only to its complementary antitag on the array.
5. For each site of the array, the fluorescent dyes present at that site are detected. The colors indicate which bases participated in the extension reaction at the corresponding SNP site and, thus, reveal the SNP variations possessed by the tested individual.

This method, with appropriate modifications, is also applicable in a pooled genotyping strategy, where PCR is applied to pooled DNA from several individuals and the purpose is to determine allele frequencies. In addition, the general idea of a universal array is also applicable for other measurement purposes.

[1] Hybridization between array-bound probes and sample molecules other than their intended target. See, e.g., www.arep.med.harvard.edu/labgc/adnan/projects/YeastCrossHyb/ and Ben-Dor *et al.* (2000).



Fragments spanning the polymorphism sites for all the SNPs in the set are extracted. The different shapes denote different variants.

Oligonucleotides complementary to the sequences immediately preceding the polymorphism sites are tagged by DNA tags, designed to specifically hybridize to their complements on the array.

Extension reactions take place in solution phase, in the presence of a mixture of all four dideoxy-nucleotides (differentially fluorescently labeled) and an appropriate enzyme. For each SNP the extending base is the one complementary to the one corresponding to the base present in the sample sequence. After separation (the whole process can be performed at high temperature) a mixture of reporter molecules is formed. This mixture is brought in contact with the array. Tags hybridize to their complements and a fluorescence pattern is obtained from which the identity of all variants in the original mixture can be deduced.

FIG. 1. A scheme for SNP genotyping using a DNA TAT system (taken from Ben-Dor *et al.* [2000]).

For example, if the reporter molecules are designed to be specific, each for some target mRNA, then the same protocol can be used for expression profiling. The main advantage of DNA TAT systems is the design universality they enable. Reagents for specific queries have to be designed and obtained only for the solution phase of the assay, while all array components are universal and no design changes are necessary.

Designing DNA TAT systems presents a tradeoff. A large-scale experiment may involve the genotyping of thousands of SNPs. Clearly, it is desirable to have as many tags as possible, in order to maximize the number of SNPs that can be genotyped in parallel (which is directly correlated with the cost of the experiment). On the other hand, if too many tags are used, similar tags will necessarily entail cross-hybridization events (where tags hybridize to foreign antitags), reducing accuracy. The design of DNA TAT systems is independent of any particular application scenario and is, thus, optimized to avoid cross-hybridization between tags and foreign antitags. This issue was studied by Ben-Dor *et al.* (2000).

In performing an actual genotyping assay, there are also assay-specific sources of potential cross-hybridization. One major source involves the primer parts of the reporter molecules hybridizing to array bound antitags, producing a confusing signal unless the corresponding site on the array is designated to the primed SNP site. As this problem is specific to the actual set of SNPs to be studied, it is impossible to address it in the TAT system design stage.

Figure 2 shows a typical case, taken from one of our examples on real data from Section 6. The design used for this figure forbids complementary substrings in tags and foreign antitags with a predicted melting temperature of 20°C or more. However, the primer part introduces substrings complementary to antitag substrings that violate this bound significantly: The predicted melting temperature between substrings of antitags 20 and 2022 and their complements in the primer part is 30°C, which may be too high for washing off the reporter molecules from those foreign antitags. As a consequence, the cross-hybridization signal generated by the WIAF-909 primer at the seven antitag sites prevents these tags from being useful to type any other SNP in the same experiment.

In this work, we assume that the set of primers for the reaction was designed to achieve the desired level of specificity in the target genome, i.e., that reporter molecules will not extend on unintended genomic sequences. Remaining assay-specific sources of potential confusing signal are as follows: *Primer to antitag cross-hybridization* is as described above. *Sandwich cross-hybridization*: A duplex of two reporter molecules hybridizes to a single site in the array. The duplex is formed due to high complementarity of the sequences and hybridizes to the array site through one of the tags. Sandwich cross-hybridization involves a complicated configuration and is rare. *Primer to primer misextension*: The primer parts of reporter molecules can hybridize to other primers in the extension step in a configuration that allows for polymerase extension. *Primer to tag misextension*: Similar, but with primer parts of reporter molecules hybridizing to tags in the extension step. The latter two are similar to primer–dimer formations in PCR. Note that the cross-hybridization needs to be perfect at the 3′ end of the reporter for these problems to occur.

Primer to antitag cross-hybridization is, therefore, by far the most probable source of confusing signal. This is the only problem we explicitly address in this work, although the methods we develop can be extended to handle primer to tag misextension (as discussed in Section 3.3). In addition, any multiplexing

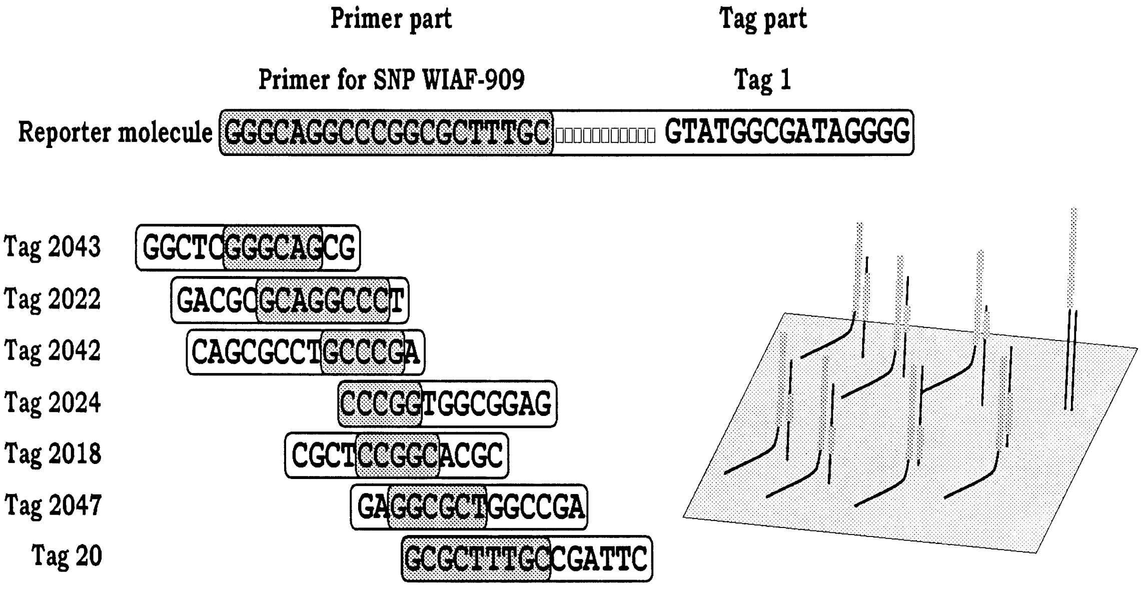


FIG. 2. Primer to antitag cross-hybridization. Sequence similarities between the primer part of reporter molecules and the tags of the universal array lead to multiple cross-hybridizations between each reporter molecule and foreign antitags on the array.

scheme for a universal array-based assay can be screened for any undesired properties prior to performing the measurement. Avoiding less common configurations should be deferred to such a screening stage rather than taken into account in the design stage.

Maximizing the multiplexing rate for a given set of SNPs (alternatively, minimizing the number of arrays to be used for a single genotyping measurement of this set), under given primer to antitag cross-hybridization constraints, is the main subject of this work. Every time we say cross-hybridization, we mean primer-to-antitag cross-hybridization. To control the multiplexing rate, we use our freedom to choose how to partition the set of SNPs into *assignable* subsets (subsets that can be measured using one array without cross-hybridization) and to assign tags to SNP sites. The assignment of a primer to a tag means that they will form a reporter molecule. A proper assignment $(p_1, t_1), \ldots, (p_k, t_k)$ of primers to tags should avoid cross-hybridization between every primer p_i and antitag $\overline{t_j}$, unless $i = j$.

To approach the multiplexing problem, we model the input data using a bipartite graph, in which the primers are on one side and the tags are on the other side. Each edge in the graph indicates potential cross-hybridization between a primer and the corresponding antitag. The multiplexing problem then translates to the problem of covering the primer vertices of the graph using a minimum number of balanced induced subgraphs of maximum degree one. We prove that the general problem is NP-complete. However, in actual applications, the primer vertices have degrees bounded by some constant d. We exploit this restriction and develop an $O(d)$-approximation algorithm for the problem, which produces a cover of cardinality $\lceil \frac{m}{\lfloor n/d \rfloor} \rceil$ for m primers and n tags. We also give an $O(d)$-approximation for a variant of the problem in which the covering subgraphs are required to be vertex disjoint. We then propose a stochastic model for the input cross-hybridization data and use it to prove a lower bound on the cover size. We complement our theoretical analysis by implementing two heuristic approaches for the problem that are based on the theoretical results. The algorithms are tested on synthetic data, with randomly generated primer sequences, and on simulated SNP data, where real primer sequences are used, and cross hybridization potential is assessed by a simplistic model of hybridization thermodynamics.

The paper is organized as follows: In Section 2, we mathematically model and formalize the optimization problem. Section 3 presents our hardness and approximation results on the covering problem and its variants. We develop a lower bound under a stochastic data model in Section 4. In Section 5, we present the two practical algorithms. The experimental results of both algorithms on simulated data are presented in Section 6.

2. FORMAL PROBLEM DEFINITION

Denote the set of DNA tag sequences associated with the universal array by T and the corresponding set of antitags by $\overline{T}$. By P we denote the set of primers (the sequences complementary to the upstream regions of the SNPs). 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.

Potential cross-hybridization between primers and antitags can be determined experimentally or predicted computationally, e.g., on the basis of the sequence (Ben-Dor *et al.*, 2000). The methods presented here are not specific to any such determination mechanism. We only assume that cross-hybridization potentials are given in the form of a binary $m \times n$ matrix A, such that

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

Solutions to our multiplexing problem correspond to proper partitions of P into subsets, where each subset corresponds to one array experiment. 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. Formally:

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* $P' = \{p_1, \ldots, p_k\} \subseteq P$ *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 is similarly defined.*

We now give a characterization of assignable primer sets directly in terms of the cross-hybridization matrix A.

Definition 3. *A* subpermutation matrix *is a square* 0–1-*matrix whose rows and columns can be permuted such that all entries outside the main diagonal are* 0.

Lemma 1. *A set of primers* $P' \subseteq P$ *is assignable if and only if* P' *corresponds to the row set of a subpermutation submatrix of* A.

Proof. Let $P' = \{p_1, \ldots, p_k\}$. If P' is assignable, then there exists a set of non-cross-hybridizing reporter molecules $\{(p_1, t_1), \ldots, (p_k, t_k)\}$ with $A_{p_i,t_j} = 0$ for all $i \neq j$. Hence, the submatrix of A induced by the rows in P' and the columns in $T' = \{t_1, \ldots, t_k\}$ is a subpermutation matrix.

Conversely, suppose there exists a subpermutation submatrix of A with row set P'. Let $R = \{(p_1, t_1), \ldots, (p_k, t_k)\}$ denote those elements of the submatrix that end up on the diagonal if its rows and columns are permuted such that all other entries are 0. Clearly, R corresponds to a non-cross-hybridizing set of reporter molecules, and its row set P' is assignable. ■

An alternative point of view models the input matrix A as a bipartite graph $G = (P, T, A)$, whose vertices are primers (P) and tags (T) and whose edges represent potential cross-hybridizations between primers and the corresponding antitags. Throughout the paper, we use graph and matrix language interchangeably. For convenience, we use A to denote both the input cross-hybridization matrix and the edge set of G, which is the set of primer–tag pairs $\{(p, t) : p \in P, t \in T, A_{p,t} = 1\}$. 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.

Observation 1. *A matrix A with a set of rows P and a set of columns T is a subpermutation matrix if and only if the bipartite graph* $G = (P, T, A)$ *is an assignable graph.*

The following proposition formalizes a necessary and sufficient condition for a set of primers to be assignable:

Proposition 1. *Let* $G = (P, T, A)$ *be a bipartite graph, with* $T = \{t_1, \ldots, t_n\}$ *and* $P = \{p_1, \ldots, p_m\}$. *For* $j = 1, \ldots, n$, *let* $Y(j) = 1$ *if* t_j *has degree zero, and* 0 *otherwise. For* $i = 1, \ldots, m$, *let* $X(i) = 1$ *if G contains a tag of degree 1, which is adjacent to* p_i, *and* 0 *otherwise. Then P is assignable if and only if*

$$\sum_{j=1}^{n} Y(j) + \sum_{i=1}^{m} X(i) \geq m.$$

Proof. According to Observation 1, P is assignable if and only if G has an assignable subgraph H that covers the m vertices of P. On the one hand, if such a subgraph exists, each of its m tags contributes a value of 1 to either $\sum_{j=1}^{n} Y(j)$ or $\sum_{i=1}^{m} X(i)$. On the other hand, if one attributes one tag to each 1 generated in the above sum, the subgraph induced by m of these tags and P is an assignable graph. ■

Definition 4. *A partition* $\mathcal{E}$ *of the primer set P is called a* primer cover *if each* $P' \in \mathcal{E}$ *is assignable.*

Our main optimization problem is stated as follows:

Problem 1 (Minimum Primer Cover [MPC]). *Given a bipartite graph* $G = (P, T, A)$, *find a minimum primer cover of* P.

Equivalently, MPC calls for finding a minimum set of assignable subgraphs that cover P, or a minimum set of subpermutations submatrices of A that cover all the rows of A.

3. COMPUTATIONAL COMPLEXITY

In this section, we address the computational complexity of MPC and variants thereof.

3.1. *Minimum primer cover*

In this section, we show that MPC is NP-complete and give an approximation algorithm for the problem when the degrees of the primer vertices in the input graph are bounded.

Theorem 1. *MPC is NP-complete.*

Proof. Membership in NP is trivial. We reduce from SET COVER, where all input subsets are required to have cardinality at least 2. This problem is known to be NP-complete (Garey and Johnson, 1979, Problem SP5). Given an instance $(P, \mathcal{S}, l)$ of SET COVER, where $\mathcal{S}$ is a collection of subsets of a finite set P and l is an integer, we construct an instance $(G = (P, T, A), l)$ of MPC as follows: For every subset $S_i = \{s_{i,1}, \ldots, s_{i,k}\} \in \mathcal{S}$, we add vertices $T_i = \{t_{i,1}, \ldots, t_{i,k}\}$ to T, such that every $t_{i,j}$ $(1 \leq j \leq k)$ is adjacent to all the vertices in $P \setminus S_i$ (and to no other vertex).

A set cover $S_{i_1}, \ldots, S_{i_l}$ induces a primer cover $\mathcal{E} = (S_{i_1}, \ldots, S_{i_l})$ with the same cardinality, since each S_{i_j} is assignable. Conversely, suppose there exists a primer cover $\mathcal{E}$ of size l. A set $S \in \mathcal{E}$ is called *homogeneous* if all its primers belong to the same subset S_i. Set S is called *crossing* if $S = \{p, p'\}$ and no $S' \in \mathcal{S}$ contains both p and p'.

Observe that every assignable primer set is either homogeneous or crossing. If all the primer sets in the cover are homogeneous, taking the corresponding subsets yields a set cover of size l. Otherwise, we can apply a series of transformations to the cover that eliminate the crossing sets in the cover and preserve its cardinality. In each step, we consider a crossing primer set $S = \{s_i, s_j\} \in \mathcal{E}$, where $s_i \in S_i, s_j \in S_j$, and $i \neq j$. If some $s_i' \in S_i$ is covered by a homogeneous set S', we can move s_i from S to S', eliminating one crossing set. Otherwise, there exists a crossing set $S'' = \{s_i', s_k\}$, which contains some $s_i' \in S_i$. By moving s_i' to S and s_j to S'', we eliminate one crossing set. Applying these modifications to the cover, we necessarily end with a cover of size l whose subsets are all homogeneous. ■

Note that the proof of Theorem 1 implies that MPC is NP-complete even if the number of tags is required to be greater than or equal to the number of primers.

In the context of DNA TAT systems, as constructed in Ben-Dor *et al.* (2000), the choice of tags implies that the degree of every $p \in P$ is bounded by some constant d. This is true since, by construction, strings that are long enough to potentiate cross-hybridization are not common to any two tags. Each primer (of bounded length) can have at most d such substrings (where d depends on the primer length and the thermodynamical model) and can, therefore, form edges with at most d tags. Practical values of d range between 5–15. We call an instance $G = (P, T, A)$ of MPC a *d-bounded instance* if the degree of every $p \in P$ is bounded by d. We shall exploit this restriction and develop an $O(d)$ approximation algorithm for MPC on d-bounded instances.

By Proposition 1, instances that can be covered by one subgraph are easily determined. Henceforth, we assume that the input MPC instance has an optimum solution of cardinality at least 2. The case $d = 1$ is polynomial for $m \leq n$:

Lemma 2. *Let $G = (P, T, A)$ be a 1-bounded instance of MPC, with $m \leq n$. Then a minimum primer cover for G can be found in polynomial time.*

Proof. W.l.o.g., $m = n$ (if $n > m$, arbitrarily remove $n - m$ vertices from T). Let G_1 be the subgraph induced by the vertices of a maximal matching M in G. Let G_2 be the subgraph induced by all other vertices. Clearly, G_1 and G_2 are balanced and together they span the entire set of primers. Since the degree of every primer is at most 1, G_1 has maximum degree 1. By maximality of M, G_2 contains no edges. The claim follows. ■

The algorithm for the general case is based on the following Lemma:

Lemma 3. *Let $G = (P, T, A)$ be a d-bounded instance of MPC with $d > 1$. Then, any set of at most $\lfloor n/d \rfloor$ primers in G is assignable.*

Proof. Let P' be a set with $s \leq \lfloor n/d \rfloor$ primers. Let T' be the set of neighbors of P' in G. Finally, let H be the subgraph induced on $P' \cup T'$, and let $C_1, \ldots, C_k$ be its connected components. To prove that P' is assignable, we shall identify a set of tags that forms a non-cross-hybridizing set with P'. For each C_i, denote its number of primers and tags by p_i and t_i, respectively. For every $1 \leq i \leq k$, we have $t_i \leq (d-1)p_i + 1$, due to the degree constraints and the connectivity of C_i. Furthermore, if C_i contains no tag of degree 1, then $t_i \leq (d-1)p_i$. Denote by r the number of connected components that contain a tag of degree 1 in H. We conclude that $|T \setminus T'| \geq s - r$. Let T'' be a subset of tags that consists of r tags of degree 1, one from each component of H that has such, and $s - r$ tags from $T \setminus T'$. Then T'' forms a non-cross-hybridizing set with P'. ■

Corollary 1. *Let $G = (P, T, A)$ be a d-bounded instance of MPC with $d > 1$. Then we can find, in $O(m)$ time, a primer cover for G of cardinality at most $\lceil \frac{m}{\lfloor n/d \rfloor} \rceil$.*

Interestingly, one can get a similar result by identifying assignable sets of tags. The following algorithm is the basis for one of the practical heuristics that are presented in Section 5.

Theorem 2. *Let $G = (P, T, A)$ be a d-bounded instance of MPC with $m \leq n$. Assume that $n > 2d(d+1)$. Then we can find, in polynomial time, a solution to MPC on G of cardinality at most $\lceil \frac{m}{n} \rceil (d+1)$.*

Proof. **Algorithm:** Let $x = \lceil \frac{m}{n} d \rceil$. First, assume for simplicity that $m \leq n$ and $\frac{m}{x}$ is an integer. If $m \leq 2d$, then P is assignable, and we are done. Henceforth, we assume that $m > 2d$. Consider the following iterative algorithm: The algorithm has x steps. At each step, it identifies an assignable subgraph of size $\frac{m}{x}$, deletes the primer vertices of this subgraph (as they are covered), and continues the next iteration on the remaining graph. The assignable subgraph at each iteration is constructed on the basis of the $\frac{m}{x}$ tags of lowest degree. Altogether, we obtain a cover with x assignable subgraphs. We now prove the correctness of the algorithm.

In the i-th iteration ($0 \leq i < x$), the current graph G_i consists of $m\left(\frac{x-i}{x}\right)$ primers and n tags. Since the degree of each primer is bounded by d, the number of edges in G_i is at most $md\left(\frac{x-i}{x}\right)$. Let T' be the set of $\frac{m}{x}$ tags with lowest degrees in G_i.

Claim 1. *The degree of each tag in T' is bounded by $x - i$.*

Proof. Suppose there exists a tag in T' with degree at least $x - i + 1 \geq 1$. Hence, all tags in $T \setminus T'$ have degree at least $x - i + 1$, so the number of edges incident on $T \setminus T'$ is at least

$$r = \left(n - \frac{m}{x}\right)(x - i + 1) = (nx - m)\left(\frac{x - i + 1}{x}\right) \geq m(d-1)\left(\frac{x-i}{x}\right) + \frac{m(d-1)}{x}$$

where the last inequality follows from the fact that $nx \geq md$. Since $m(d-1) > m > x - i$, we have $r \geq |E(G_i)|$, a contradiction. ■

Applying Lemma 3, while exchanging the roles of primers and tags, we conclude that there is a set P' of $\lfloor \frac{m(x-i)}{x(x-i)} \rfloor = \frac{m}{x}$ primers, such that the graph induced on $P' \cup T'$ is an assignable graph. This completes the correctness proof.

If $\frac{m}{x}$ is not an integer, we first find the largest number $m' < m$ such that that $\frac{m'}{x'}$ is an integer, where $x' = \lceil \frac{m'}{n} d \rceil$. Using simple calculations, one can show that such m' exists and $m' \geq m - 2d$. Applying the above algorithm to m' of the primers and noting that the other primers form an assignable set, we obtain a cover of the required size.

If $m > n$, we arbitrarily partition P into sets of size at most n and apply the above algorithm. ■

We have given a polynomial algorithm which produces a primer cover of size $O(d)$ for any d-bounded instance of MPC with $n = m$. The following theorem shows that this result is almost tight. Precisely, using the probabilistic method, we show that there are instances of MPC for which an optimum primer cover has cardinality $\Omega(d/\log d)$.

Theorem 3. *There exists a d-bounded instance of MPC with $m = n$, whose optimum primer cover has cardinality $\Omega(d/\log d)$.*

Proof. First, observe that an assignable primer set of size k implies a balanced bipartite graph with $k/2$ vertices on each side and no induced edges. Thus, it suffices to prove the existence of a bipartite graph $G = (U, V, E)$ with $|U| = |V| = n$, such that the degree of every vertex in U is bounded by d and for any subset $U' \subset U$ with $|U'| = cn \log d/d$ (for some constant $c > 0$), the set $N(U)$ of neighbors of U satisfies $|N(U')| > n - |U'|$. The latter expansion property implies that any subset of U with at least $2cn \log d/d$ vertices is not assignable and, hence, the bound on the size of an optimum primer cover. The existence of such an expander graph (for, e.g., $c = 2$) can be proven using the probabilistic method (see, e.g., Motwani and Raghavan [1995, pages 108–110]). ■

3.2. Maximum assignable primer set

In this section we study a greedy approach to MPC that mimics approximation algorithms for SET COVER (cf. Corman *et al.* [1990]). The scheme is recursive: The largest assignable subset in P is identified and removed, and the algorithm proceeds recursively on the remaining graph. This approach could guarantee an $O(\log m)$ approximation for MPC and would typically perform better. However, each of the stages is NP-hard:

Problem 2 (Maximum Assignable Primer-set [MAP]). *Given a bipartite graph G, find a maximum assignable subgraph H of G.*

Theorem 4. *MAP is NP-hard.*

Proof. By reduction from the complete balanced bipartite subgraph problem, where the input is a bipartite graph and an integer k and the objective is to find a complete balanced subgraph with k vertices on each side. This problem is known to be NP-complete (Garey and Johnson, 1979, Problem GT24) and can be trivially reduced to the empty balanced bipartite subgraph problem, where the objective is to find a balanced subgraph with no induced edges. We reduce the latter problem to MAP.

Given an instance $(G = (U, V, E), k)$ of the empty balanced bipartite subgraph problem, where $|U|, |V| < l$, we construct an instance $(G' = (U', V', E'), lk)$ of MAP. Each vertex v in G is duplicated l times $v^1, \ldots, v^l$ in G'. For every edge $(u, v) \in E$, we add the edges (u^i, v^j) to E' for all $1 \le i, j \le l$.

Clearly, an empty balanced induced subgraph of size k induces a solution to MAP of size at least lk. Conversely, suppose that $H = (X, Y, F)$ is an assignable subgraph of G' and $|X| \ge lk$. We first claim that $|F| < l$. If $|F| \ge l$, then F contains, w.l.o.g., two edges $(u^1, a), (u^2, b)$ for some $u \in U$. But then either $a = b$, implying that a has degree at least 2 in H, or both u^1 and u^2 have degree at least 2 in H, a contradiction.

Removing all vertices incident to edges in F, we obtain a solution to MAP with size (strictly) greater than $(k - 1)l$, since F is a matching. This implies an empty balanced induced subgraph of size k in G. ■

Note that the related problem of finding a maximum induced matching in a bipartite graph is also NP-hard (Cameron, 1989).

We next present a polynomial algorithm for MAP in the case that every primer has at most one adjacent tag, and an integer programming formulation for the general case.

Lemma 4. *Let $G = (P, T, A)$ be a bipartite graph in which the degree of each $p \in P$ is at most 1. Then a maximum assignable subgraph of G can be found in polynomial time.*

Proof. For a vertex v, denote by $N(v)$ its set of neighbors in G. Let $T^* = \{t \in T : |N(t)| > 1\}$, and let $n^* = |T^*|$. Let $H = (P', T', E')$ be a solution to MAP on G. We observe that for every $t \in T^*$, either $t \notin T'$ or $|N(t) \cap P'| = 1$. Moreover, in the latter case, for $\{p'\} = P' \cap N(t)$ and for every other $p \in N(t)$, replacing p' with p yields a solution as well.

The observation motivates the following algorithmic scheme for creating a solution: Enumerate the number k of tags from T^* in an optimum solution. Remove the $n^* - k$ highest degree tags in T^* from G. For every remaining tag $t \in T^*$, arbitrarily choose some $p \in N(t)$ and remove $N(t) \setminus \{p\}$ from G. The result is a graph of maximum degree 1, which contains the required number of tags and a maximum number m^* of primers. Any subset of these primers of size $\min\{m^*, |T| - n^* + k\}$ induces a maximum assignable set. ■

We now give an integer programming formulation for MAP. Such formulation allows applying algorithmic methods for solving integer programs to MAP. In this formulation there are binary variables $e_{p,t}$ for every primer–tag pair. These variables are constrained so that $e_{p,t} = 1$ if and only if in a maximum assignable set of primers $p \in P$ is matched with $t \in T$.

$$\max \sum_{p \in P, t \in T} e_{p,t} \tag{1}$$

$$s.t. \quad e_{p,t} \in \{0, 1\} \tag{2}$$

$$e_{p_1,t} + e_{p_2,t} \leq 1 \quad p_1 \neq p_2 \tag{3}$$

$$e_{p,t_1} + e_{p,t_2} \leq 1 \quad t_1 \neq t_2 \tag{4}$$

$$e_{\pi,t} + e_{p,\tau} \leq 1 \quad A_{p,t} = 1, \pi \neq p, \tau \neq t \tag{5}$$

Lemma 5. *The above integer programming solves MAP.*

Proof. Suppose that $\{e_{p,t}\}$ is a vector satisfying the constraints. We show that it induces a non-cross-hybridizing set of reporter molecules. Constraints 3 and 4 ensure that there will be a 1–1 correspondence between primers and tags. Let this correspondence be $\{(p_1, t_1), \ldots, (p_k, t_k)\}$ (i.e., $e_{p_i,t_i} = 1$), then it suffices to prove that $A_{p_i,t_j} = 0$ for $i \neq j$. Suppose to the contrary that $A_{p_i,t_j} = 1$ for $i \neq j$. But constraint 5 ensures that $e_{p_i,t_i} + e_{p_j,t_j} \leq 1$, a contradiction.

Conversely, suppose that $\{(p_1, t_1), \ldots, (p_k, t_k)\}$ is a non-cross-hybridizing set. Define $e_{p_i,t_j} = 1$ if and only if $i = j \leq k$. It is easy to check that that this assignment satisfies the constraints of the programming. ■

3.3. Minimum partition into disjoint assignable subgraphs

In our discussion so far, we did not require the covering subsets in a solution of MPC to be tag disjoint. From the assay point of view, there is no need for such requirement. In this section, we study a mathematically related question of optimally partitioning a bipartite graph into a set of vertex-disjoint assignable subgraphs that cover the set of primers. Note that it is meaningful only when the number of primers is at most the number of tags. We henceforth assume this is the case. We give an algorithm which produces a cover of size at most $2d$ for a graph with d-degree bounded primer vertices. The problem is formally stated as follows:

Problem 3 (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.*

MPDAS is NP-complete by essentially the same reduction as in the proof of Theorem 1. A d-bounded instance of MPDAS is defined in the same way as for MPC. Our covering algorithm is based on graph coloring and is given below.

Theorem 5. *Let $G = (P, T, A)$ be a d-bounded instance of MPDAS with $m \leq n$. Then we can find, in time $O(m^2 + md)$, a solution to MPDAS on G of cardinality at most $2d$.*

Proof. Assume $n = m$ (if $n > m$, remove arbitrarily $n - m$ vertices from T). We shall find a collection of at most $2d$ assignable subgraphs that span the vertices of P. Let M be a maximal matching in G. Let

H be the subgraph induced by the set of vertices that are not incident to edges of M. Clearly, H contains no induced edges and is assignable.

We now construct a directed graph $G' = (V', E')$ as follows: Every vertex $v \in V'$ corresponds to a pair of vertices $p \in P, t \in T$ that were matched by M. An edge $e \in E'$ is directed from $v_1 = (p_1, t_1)$ to $v_2 = (p_2, t_2)$ if and only if $(p_1, t_2) \in A$. By construction, every vertex in G' has out-degree at most $d - 1$. Hence, G' can be colored using at most $2d - 1$ colors using SLO (smallest-last ordering) coloring (Matula and Beck, 1983). Each color class corresponds to the vertices of an assignable subgraph, and together with H these subgraphs cover P.

The running time is dominated by the coloring, which costs $O(m^2 + md)$ time. ■

Note that if we avoid the step of finding a maximal matching in the algorithm of Theorem 5, then the complexity of the algorithm reduces to $O(m^2 + md)$ at the expense of possibly obtaining a cover with one more set.

In fact, we can produce smaller covers if the number of tags is strictly greater than the number of primers as the following theorem shows.

Theorem 6. *Let $G = (P, T, A)$ be a d-bounded instance of MPDAS with $n \geq (k + 1)m$, for some $k \geq 1$. Then we can find, in polynomial time, a solution to MPDAS on G with cardinality at most $2\lfloor \frac{d}{k} \rfloor$.*

Proof. We first remove from G the $n - m \geq mk$ tags with highest degrees. Clearly, the degree of each remaining tag is bounded by $\lfloor \frac{d}{k} \rfloor$. By changing the roles of tags and primers in the proof of Theorem 5, we obtain a solution of cardinality $2\lfloor \frac{d}{k} \rfloor$. ■

We end this section by commenting on the applicability of MPDAS: There is a protocol solution to avoiding primer-to-antitag cross-hybridization. The idea is to introduce blocking oligonucleotides, perfect Watson–Crick complements of the primers used in the assay, right after the extension reaction and prior to the array hybridization step. As these occupy the primer parts of the reporter molecule, they block any potential hybridization of these primers. The main source of confusing signal now becomes primer-to-tag misextensions. Solving MPDAS in this case can be done using the same algorithmic methods presented above. By solving MPDAS for multiplexing the solution-phase experiments, it is possible to perform the genotyping using a single array, at the cost of performing several solution-phase experiments. This protocol has not been experimentally tested, to our knowledge. The principal motivation for MPDAS, therefore, remains purely mathematical.

3.4. Approximating MPC and MPDAS

In this section, we study the approximation ratios of our algorithms for MPC and MPDAS. Since by Proposition 1 we can check, in polynomial time, if a set of primers is assignable, we assume henceforth that the optimum solution for instances of either problem has cardinality at least 2.

Claim 2. *There exists a polynomial approximation algorithm for MPC on d-bounded instances with $n > 2d(d+1)$, which guarantees an approximation ratio of $\left(\lceil \frac{m}{n} d \rceil / 2 + \frac{1}{2}\right)$ if $m \leq n$ and a ratio of $(d+1)$ if $m > n$.*

Proof. The proof directly from Theorem 2 and the fact that at least $\lceil \frac{m}{n} \rceil$ subgraphs are needed in order to cover the primer set of an instance with $m > n$. ■

Claim 3. *There exists a d-approximation algorithm for MPDAS on d-bounded instances. The approximation ratio can be improved to $\lfloor \frac{d}{k} \rfloor$ if $n \geq (k + 1)m$.*

Proof. The first part follows from Theorem 5. The second part follows from Theorem 6 . ■

Next, we show how to improve the approximation ratio of our algorithm for MPDAS in the case that $m = n$ and $d = \omega(\log^2 m)$. We include this result as it gives more insight on the problem. The result is based on an improved algorithm for the case that the optimum solution has cardinality 2, which uses an approximation algorithm for the minimum bisection problem by Feige and Krauthgamer (2000). A *minimum bisection* of a graph G with $2l$ vertices is a partition of $V(G)$ into two sets of size l, such that the number s of edges that cross the partition is minimum. The algorithm of Feige and Krauthgamer (2000) produces a bisection of size $O(s\log^2 l)$.

Theorem 7. *Let $G = (P, T, A)$ be a d-bounded instance of MPDAS with $m = n$ and an optimum solution of cardinality 2. Then we can find, in polynomial time, a solution to MPDAS on G with cardinality $O(\log m\sqrt{d})$.*

Proof. Since P can be covered using two vertex-disjoint assignable subgraphs, these subgraphs induce a balanced cut (bisection) in G of size at most m. This bisection is formed by taking one side of the cut to be the union of the set of primers of one subgraph and the set of tags of the other subgraph. Our algorithm for MPDAS in this case starts by applying to G the minimum bisection approximation algorithm of Feige and Krauthgamer (2000), producing a bisection of size $O(m\log^2 m)$. This bisection induces a partition into two balanced subgraphs (by joining the primers on one side of the cut with the tags on the other side). Clearly, in each subgraph there are $O(m\log^2 m)$ edges. Next, these two subgraphs are handled using the following lemma.

Lemma 6. *Let $H = (P', T', E')$ be a d-bounded instance of MPDAS with $|P'| = |T'| = r$ and $|E| \leq kr\log^2 r$, for some constant k. Then we can find, in polynomial time, a solution to MPDAS on H with cardinality at most $4\log r\sqrt{kd}$.*

Proof. Let $c = \log r\sqrt{kd}$. We partition the set of primers into two subsets and cover each one separately:

(1) $S_1 = \{p \in P : deg(p) > c\}$. Clearly, $|S_1| \leq \frac{kr\log^2 r}{c}$ and, thus, by Theorem 6 S_1 can be covered by $\frac{2kd\log^2 r}{c} = 2\log r\sqrt{kd}$ subgraphs.

(2) $S_2 = \{p \in P : deg(p) \leq c\}$. By Theorem 5, S_2 can be covered by $2c = 2\log r\sqrt{kd}$ subgraphs. ■

Using Lemma 6, we can cover each of the two subgraphs produced by the bisection approximation algorithm using $O(\log m\sqrt{d})$ vertex-disjoint assignable subgraphs. Overall, we obtain a solution to MPDAS on G with cardinality $O(\log m\sqrt{d})$. ■

Corollary 2. *For $d = \omega(\log^2 m)$, we can approximate MPDAS on d-bounded instances with $n = m$, in polynomial time, to within a factor of $\frac{2d}{3}$ of optimum.*

4. A STOCHASTIC MODEL

In this section, we formulate a stochastic model for the cross-hybridization matrix A. The purpose is twofold: to generate a platform on which we can test the performance of algorithmic approaches and to study the distribution of affordable multiplexing rates for random sets of SNPs.

Let A be a binary matrix. Let $n(A)$ denote the minimum t such that the rows of A can be partitioned into t assignable sets. Ideally, we would like to specify a probability distribution over A that corresponds to the actual distribution of matrices that arise from genotyping problems using universal arrays and then study the distribution of $n(A)$ for matrices drawn from this distribution. However, this distribution will depend on the particular system of tags chosen, the primers occurring in the genotyping problem, and the criterion for cross-hybridization between a primer and an antitag. Because of these complications, we shall instead consider a simple parameterized family of distributions of 0–1-matrices. The model is governed by m and n, the dimensions of A, and by p, which represents the expected fraction of the antitags that potentially hybridize to a given primer used in the assay. The value of p depends on the primer length and on the cross-hybridization thermodynamical model.

Let $D(m, n, p)$ be a probability distribution of $m \times n$ matrices, where each matrix entry independently is equal to 1 with probability p and to 0 with probability $1 - p$. We shall derive a lower bound on $n(A)$ for matrices drawn from $D(m, n, p)$. We require the following Chernoff bound (cf. Grant and Phillips, 2001):

Lemma 7 (Chernoff). *Let X be a random variable with a binomial distribution $X \sim Binom(n, p)$. For every $\epsilon > 0$,*

$$Prob[X \geq (1 + \epsilon)np] \leq \left(\frac{e^{\epsilon}}{(1 + \epsilon)^{1+\epsilon}} \right)^{np} .$$

Theorem 8. *Let matrix A be drawn from the probability distribution $D(m, n, p)$. For a positive integer t, define $h = \lceil \frac{m}{t} \rceil$ and $x = n(1 - p)^{h-1}(1 - p + hp)$. Then the following bound holds for all t such that $h > x$:*

$$Prob[n(A) \leq t] \leq \frac{t^m}{t!} \left(\frac{x e^{\frac{h-x}{h}}}{h} \right)^{ht} .$$

Proof. Consider a matrix C drawn from $D(h, n, p)$. We shall derive an upper bound on the probability that C is assignable. Call a column of C *useful* if it contains at most one 1. Clearly, C is assignable only if it contains at least h useful columns. Each column independently is useful with probability $(1 - p)^h + hp(1 - p)^{h-1}$. Hence, the probability that C is assignable is at most $Prob[X \geq h]$, where $X \sim Binom(n, (1 - p)^h + hp(1 - p)^{h-1})$. For fixed n and p, denote the Chernoff bound of Lemma 7 on this probability by $f(h)$. Simple calculus shows that $\log(f(h))$ is a concave function for $h > x$.

Now let A be drawn from $D(m, n, p)$ and consider a row-partition of A into sets of sizes $h_1, h_2, \ldots, h_t$. Then the probability that all of these submatrices are assignable is at most $\prod_{i=1}^{t} f(h_i)$. Using the concavity of $\log(f(h))$, we conclude that this probability is maximized when the h_i's are all equal and is, therefore, at most $f(\lceil \frac{m}{t} \rceil)^t$.

If $n(A) \leq t$, then A can be row partitioned into t assignable sets. The number of such partitions is at most $\frac{t^m}{t!}$. For any given partition, the probability that all its row sets are assignable is at most $f(\lceil \frac{m}{t} \rceil)^t$. Therefore, the probability that $n(A) \leq t$ is bounded by $\frac{t^m}{t!} f(\lceil \frac{m}{t} \rceil)^t$. ■

We illustrate this result with a numerical example: Let $m = 10^5, n = 10^4$, and $p = 10^{-3}$. Evaluating the lower bound for $t = 16$, we find that $Prob[n(A) \leq 16] \leq e^{-4,878}$. In general, we determine the lower bound of a given instance as the minimal t such that $\frac{t^m}{t!} \left(\frac{x e^{\frac{h-x}{h}}}{h} \right)^{ht} \geq 0.001$.

5. ALGORITHMIC APPROACHES

The theoretical analysis in Sections 3.1 and 3.2 inspires two practical approaches to MPC. The first is based on the approximation algorithm for MPC. The second is based on the set cover approximation method alluded to in Section 3.2. We assume below that the input instance is a graph $G = (P, T, A)$ with m primers and n tags.

By Proposition 1, we can check whether a set of primers P' is assignable. Symmetrically, we can check the assignability of a set of tags. Building on this simple test of assignability, Algorithm A produces a cover with size at most $\lceil \frac{m}{n} \rceil (d + 2)$, where d is the maximum primer degree in G. It is described in Fig. 3.

The general scheme of Algorithm B is described in Fig. 4. To complete its description, we need to specify the heuristic rule used in step 3 to select which primer to remove from the set P'. The purpose of removing primers is to progress towards assignability by creating useful tags, i.e., tags of degree zero and tags of degree one that are adjacent to distinct primers. We have experimented with a family of *potential-based rules* in which each tag is assigned a potential for becoming useful, based on its degree: The higher the degree, the lower the potential, since a tag cannot possibly become useful until primer deletions have reduced its degree to 0 or 1. We then define the *potential* of a primer as the sum of the potentials of

```
1. $\mathcal{E} \leftarrow \emptyset$.
2. Unmark all vertices of $T$.
3. Sort the tags in $T$ in non-decreasing order based on their degrees in $G_{P \cup T}$.
4. $T' \leftarrow \emptyset$.
5. **While** there are unmarked tags **do:**
   (a) Find an unmarked tag $t \in T \setminus T'$ with lowest degree.
   (b) Mark $t$.
   (c) **If** $T' \cup \{t\}$ is assignable **then** $T' \leftarrow T' \cup \{t\}$.
6. Find a set $P'$ of $|T'|$ primers that form a non-cross-hybridizing set with $T'$.
7. $\mathcal{E} \leftarrow \mathcal{E} \cup \{P'\}$ (add $P'$ to the cover).
8. Update $P \leftarrow P \setminus P'$.
9. **If** $P = \emptyset$ **then** halt **else** go to 2.
```

FIG. 3. Algorithm A.

```
1. $\mathcal{E} \leftarrow \emptyset$.
2. $P' \leftarrow P$.
3. **While** $P'$ is not assignable, remove a primer of maximum potential from $P'$.
4. $\mathcal{E} \leftarrow \mathcal{E} \cup \{P'\}$ (add $P'$ to the cover).
5. Update $P \leftarrow P \setminus P'$.
6. **If** $P = \emptyset$ **then** halt **else** go to 2.
```

FIG. 4. Algorithm B.

its adjacent tags. Our heuristic rule is to choose for removal a primer of maximum potential, where the potential of a tag of degree w is defined as 2^{-w}. Whenever a primer is adjacent to at least one tag of degree 1, we adjust its potential by subtracting $\frac{1}{2}$, since this tag is useful even if the primer is not deleted.

6. EXPERIMENTAL RESULTS

6.1. Performance on synthetic data

In this section, we report on the performance of algorithms A and B on synthetic data of two types. The first type of synthetic data was generated according to the stochastic model presented in Section 4. The number of tags ranged from 500 to 2,000, the number of primers ranged from 1,000 to 5,000, and p was determined so that an average of 10 antitags potentially cross-hybridize with each primer. The results of applying both algorithms to the data are summarized in Table 1. We list both the average size of the cover achieved by the algorithms in 10 runs and the lower bound of Theorem 8. Notably, Algorithm B

TABLE 1. COMPARISON BETWEEN ALGORITHMS A AND B ON SYNTHETIC DATA[a]

	Tags (p)								
	500 (0.02)			*1,000 (0.01)*			*2,000 (0.005)*		
SNPs	*A*	*B*	*L*	A	*B*	*L*	A	*B*	*L*
1,000	9	7	5	5	4	3	3	2	2
2,000	15.3	12.5	8	9	7	5	5	4	3
5,000	33.7	28	17	18.9	15	9	10	8	5

[a] The data was generated using the stochastic model with different parameter combinations. Recorded for each set of parameters are the average cover size of algorithms A (column A) and B (column B) and the stochastic lower bound of Theorem 8 (column L).

TABLE 2. COMPARISON BETWEEN ALGORITHMS A AND B ON SYNTHETIC DATA[a]

	λ					
	6		7		8	
SNPs	*A*	*B*	*A*	*B*	*A*	*B*
1,000	10	9	3	3	1	1
2,000	17.7	15.3	5	4	2	2
5,000	38	34	10	9	3	3

[a]The data was generated using the combinatorial model with different parameter combinations.

outperforms Algorithm A in all simulations and produces covers that have cardinality within a factor of $\frac{5}{3}$ of the lower bound (which is not necessarily tight).

The second type of synthetic data was generated as follows: We assume potential cross-hybridization to depend on common substrings of length λ. This is the simpler combinatorial model described by Ben-Dor *et al.* (2000). We applied a de Bruijn sequence construction (also described therein) to generate sets of tags of length 20. Here we used $\lambda = 6, 7, 8$ resulting in 273, 1170, and 5041 tags, respectively. We then randomly generated primers of length 13. The average results over 10 runs are given in Table 2. Again, Algorithm B produces smaller covers than does Algorithm A.

6.2. *Performance on simulated data*

We complemented our analyses on synthetic data by applying both algorithms to matrices derived from real genomic sequence data. Specifically, we retrieved the first 20,000 SNP entries of human chromosome 22 from the public SNP database at NCBI (*www.ncbi.nlm.nih.gov/SNP/index.html*). For each SNP, we extracted the 20 nucleotides immediately upstream of the respective SNP site and defined a corresponding primer as as reverse complement of the upstream sequence.

We then employed the full combinatorial construction scheme of Ben-Dor *et al.* (2000) to generate two tag sets T_1 and T_2. The construction of Ben-Dor *et al.* (2000) takes into account two parameters, c and h, where $c < h$. Parameter c represents the maximal allowable hybridization potential for a tag and a foreign antitag. Parameter h represents the minimal allowable hybridization potential for a tag and its corresponding (perfect match) antitag. Full details on this representation can be found in Ben-Dor *et al.* (2000). Set T_1 was generated using the parameters $c = 12$ and $h = 58$ and contains 4,966 tags. Set T_2 contains 997 tags and was generated using the parameters $c = 10$ and $h = 40$. The parameters for the sets T_1 and T_2 were chosen as representatives of employing large and medium sized universal arrays to SNP genotyping.

To derive the entries $A_{p,t}$ in the cross-hybridization matrices A_1 and A_2, we employed the 2–4-rule as in Ben-Dor *et al.* (2000). The 2–4-rule (Strachen and Read, 1996) estimates the melting temperature of a DNA sequence and its complement as twice the number of As and Ts, plus four times the number of Cs and Gs, in degrees Celsius. Whenever the result of this rule, applied to any perfectly complementary substring between p and $\bar{t}$, exceeded the threshold of 24 (for A_1) or 20 (for A_2), we considered p and $\bar{t}$ as potentially cross-hybridizing, i.e., we set $A_{p,t} = 1$. In all other cases, we set $A_{p,t} = 0$.

Densities of A_1 and A_2 were 0.0016 and 0.0085, respectively. For A_1, both algorithms produced a cover of size 8, while the stochastic lower bound is 7 (using an estimated $p = 0.0016$). For A_2, Algorithm A used 36 arrays and Algorithm B used 34 arrays, while the lower bound lies at 27 (using an estimated $p = 0.0085$).

7. CONCLUDING REMARKS

In this paper, we formulated three combinatorial problems arising in multiplexing universal array experiments. MPC, MAP, and MPDAS were shown to be NP-complete. The complexity of these problems in the case of d-bounded instances remains open.

We have given $O(d)$-approximation algorithms for MPC and MPDAS on d-bounded instances. The approximation ratios are not known to be tight. An attempt to improve the ratios can build on the results of Section 3.4.

A variant of MPC arises if we also wish to solve simultaneously the primer-to-primer misextension problem. The corresponding graph is no longer bipartite, but contains edges between potentially cross-hybridizing primers. The goal here is to find a minimum primer cover, such that for each covering set no pair of primers cross-hybridize.

ACKNOWLEDGMENTS

We thank Moni Naor, Pavol Hell, and Noga Alon for helpful discussions. We also thank an anonymous JCB referee for carefully reading the manuscript. T.H. and R.M.K. were supported in part by a grant from the US–Israel Binational Science Foundation (BSF). R.S. was supported by a Fulbright grant.

REFERENCES

Alon, N., and Spencer, J.H. 1992. *The Probabilistic Method*, John Wiley, NY.

Ben-Dor, A., Karp, R.M., Schwikowski, B., and Yakhini, Z. 2000. Universal DNA tag systems: A combinatorial design scheme. *J. Comp. Biol.*, 7(3), 503–519.

Brenner, S. 1997. Methods for sorting polynucleotides using oligonucleotide tags. US Patent 5,604,097, 1997.

Cameron, K. 1989. Induced matchings. *Dis. Appl. Math.* 24, 97–102.

Cargill, M., Altshuler, D., Ireland, J., Sklar, P., Ardlie, K., *et al.* 1999. Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genet.* 22(3), 231–238.

Cormen, T.H., Leiserson, C.E., and Rivest, R.L. 1990. *Introduction to Algorithms*, MIT Press, Cambridge, MA.

Davis, R.W., Morris, M.S., Shoemaker, D.D., and Mittmann, M.P. 1997. Methods and compositions for selecting tag nucleic acids and probe arrays. European Patent Application 97,302,313.

Drmanac, R., Lennon, G., Drmanac, S., Labat, I., Crkvenjakov, R., and Lehrach, H. 1991. Partial sequencing by oligohybridization: Concept and applications in genome analysis. *Proc. 1st Int. Conf. Electrophoresis Supercomputing and the Human Genome*, 60–75, World Scientific.

Feige, U., and Krauthgamer, R. 2000. A polylogarithmic approximation of the minimum bisection. *Proc. 41st Symp. Foundations of Computer Science (FOCS)*, 105–115.

Garey, M.R., and Johnson, D.S. 1979. *Computers and Intractability: A Guide to the Theory of NP-Completeness*, W.H. Freeman, San Francisco.

Gerry, N.P., Witowski, N.E., Day, J., Hammer, R.P., Barany, G., *et al.* 1999. Universal DNA microarray method for multiplex detection of low abundance point mutations. *J. Mol. Biol.*, 292(2), 251–262.

Grant, D.M., and Phillips, M.S. 2001. *Technologies for the Analysis of Single-Nucleotide Polymorphisms: An Overview*, Marcel Dekker, NY.

Hacia, J.G. 1999. Resequencing and mutational analysis using oligonucleotide micro arrays. *Nature Genet.* 21(1), 42–47.

Kristensen, V.N., Harada, N., Yoshimura, N., Haraldsen, E., Lonning, P.E., *et al.* 2000. Genetic variants of cyp19 (aromatase) and breast cancer risk. *Oncogene* 19(10), 1329–1333.

Matula, D., and Beck, L. 1983. Smallest-last ordering and clustering and graph coloring algorithms. *J ACM* 30, 417–427.

Motwani, R., and Raghavan, P. 1995. *Randomized Algorithms*, Cambridge University Press, London.

Risch, N.J. 2000. Searching for genetic determinants in the new millennium. *Nature* 405(6788), 847–856.

Schouten, J.P., McElgunn, C.J., Waaijer, R., Zwijnenburg, D., Diepvens, F., and Pals, G. 2002. Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. *Nucl. Acids Res*, 30(12).

Strachen, T., and Read, A.P. 1996. *Human Molecular Genetics*, Bios Scientific Publishers.

Syvanen, A.C. 1999. From gels to chips: “Minisequencing” primer extension for analysis of point mutations and single nucleotide polymorphisms. *Human Mutat.* 13(1), 1–10.

Venitt, S. 1994. Mechanisms of carcinogenesis and individual susceptibility to cancer. *Clin. Chem.* 40(7.2), 1421–1425.

Venitt, S. 1996. Mechanisms of spontaneous human cancers. *Environ. Health Perspect.* 104(3), 633–637.

Wang, D.G., Fan, J.B., Siao, C.J., Berno, A., Young, P.P., *et al.* 1998. Large-scale identification, mapping, and genotyping of single nucleotide polymorphisms in the human genome. *Science* 280(5366), 1077–1082.

Watanabe, Y., Fujiyama, A., Ichiba, Y., Hattori, M., Yada, T., Sakaki, Y., and Ikemura, T. 2002. Chromosome-wide assessment of replication timing for human chromosomes 11q and 21q: Disease-related genes in timing-switch regions. *Human Mol. Genet.* 11(1), 13–21.

Address correspondence to:
Roded Sharan
International Computer Science Institute
1947 Center Street, Suite 600
Berkeley, CA 94704

E-mail: roded@icsi.berkeley.edu