Towards Optimally Multiplexed Applications of Universal DNA Tag Systems

Amir Ben-Dor * Tzvika Hartman † Benno Schwikowski ‡ Roded Sharan §
Zohar Yakhini ¶

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 a previous work. However, cross-hybridization can 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 simulated and real SNP data.

*Agilent Laboratories (amir_ben-dor@agilent.com).
†Dept. of Computer Science and Applied Mathematics, Weizmann Institute (tzvi@cs.weizmann.ac.il).
‡Institute for Systems Biology, 1441 N. 34th St., Seattle, WA 98103 (benno@systemsbiology.org).
§Corresponding author: International Computer Science Institute, 1947 Center St., Suite 600, Berkeley CA 94704-1198 (roded@icsi.berkeley.edu).
¶Agilent Laboratories and Computer Science Dept., Technion (zohar_yakhini@agilent.com).

RECOMB'03, April 10–13, 2003, Berlin, Germany.

Categories and Subject Descriptors

G.2.2 – graph algorithms, J.3 computer applications – biology and genetics.

General Terms

Algorithms, theory, experimentation.

Keywords

SNP genotyping, universal array, cross-hybridization, minimum primer cover, stochastic model.

1. INTRODUCTION

SNPs (single nucleotide polymorphisms) are differences, across the population, in a single base, within an otherwise conserved genomic sequence [21]. The sequence variation represented 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 (see [5]). Somatic or native SNPs in oncogenes or in related regions can determine cancer susceptibility and are often related to pathogenesis (see, e.g., [19, 20, 12, 22]). 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 [15]). 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 [10, 16]. Other methods are based on hybridization array technology (cf. [18, 11, 7]). 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 only occurs in sites on the surface that are Watson-Crick complements to some substring in the target. In practice, cross-hybridization is a main source of signal contamination in any array-based hybridization assay.

Recently, S. Brenner and others [3, 13, 9] suggested an alternative 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 Figure 1 and consists of the following steps:

1. A set of reporter molecules (one for each SNP) is synthesized. Each 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 flanking sequences of the SNPs only hybridize with the appropriate reporter molecule.

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 only hybridize 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 the 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. 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.

Designing DNA TAT systems presents a tradeoff. 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. 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 addressed in [2].

When 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.

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., reporter molecules will not extend on unintended genomic sequences. Remaining assay-specific sources of potential confusing signal are: <u>Primer to antitag cross-hybridization</u> as described above. <u>Sandwich cross-hybridization</u>: 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. <u>Primer to primer mis-extension</u>: The primer parts of reporter molecules can hybridize to other primers in the extension step in a configuration that allows for polymerase extension. <u>Primer to tag mis-extension</u>: 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 this problem 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. The methods we develop can be extended to handle primer to tag mis-extension. In addition, any multiplexing 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 ac-

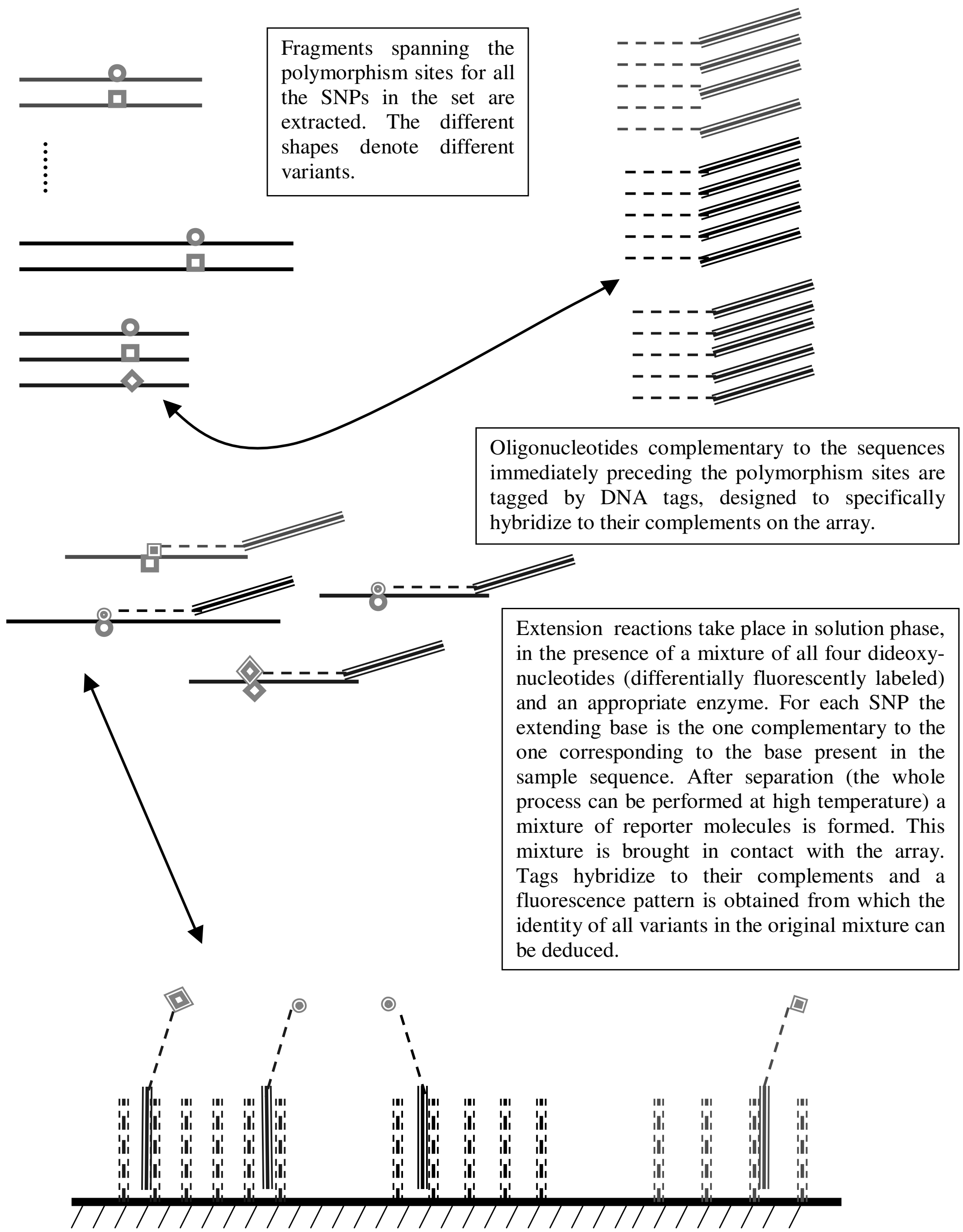

Figure 1: A scheme for SNP genotyping using a DNA TAT system.

tual 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. Specifically, we give an algorithm that produces a cover of cardinality $\lceil \frac{m}{\lfloor n/(d+1) \rfloor} \rceil$ 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. In addition, we 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 and testing their performance on simulated and real SNP data.

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 two practical heuristic algorithms. The experimental results of both algorithms on simulated and real data are presented in Section 6. For lack of space some proofs are sketched or omitted.

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 $n = |T|$. 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.

Solutions to our multiplexing problem correspond to partitions of P into subsets, where each subset corresponds to one array experiment. Potential cross-hybridization between primers and antitags can be determined experimentally or predicted computationally, e.g., on the basis of the sequence [2]. The methods presented here are not specific to any such determination mechanism. We only assume that potential cross-hybridizations are given in the form of a binary $m \times n$ matrix A, such that:

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

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

Note that in the above definition the primer part of a reporter molecule may potentially cross-hybridize with the tag part (but with no other tag). 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 2. *A set of* k *distinct primers* $\{p_1, \ldots, p_k\} \subseteq P$ *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* $t_1, \ldots, t_k \in T$.

The following definition allows us to cast these conditions directly in terms of 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.

Observation 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.

An alternative point of view models the input matrix A as a bipartite graph. Let $G = (P, T, A)$ be a bipartite graph, whose vertices are primers (P) and tags (T), and whose edges represent potential cross-hybridizations between primers and the corresponding antitags. A subgraph $H = (P', T', E')$ of G is called *balanced* if $|P'| = |T'|$. H is called an *assignable subgraph* if H is a balanced induced subgraph of maximum degree 1.

Observation 2. *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_j, *and* 0 *otherwise. Then* P *is assignable if and only if*

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

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

Observations 1 and 2 lead to a short statement of our main optimization problem:

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

Throughout the paper we use graph and matrix language interchangeably.

3. COMPUTATIONAL COMPLEXITY

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

3.1 Minimum Primer Cover

Here we show that MPC is NP-complete and give an approximation algorithm for the problem when the primer degrees 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 [8, 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. 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 modifications to the cover that eliminate the crossing sets in it 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 a crossing set. Applying these modifications to the cover we necessarily end with a homogeneous cover of size l. $\square$

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

In the context of DNA TAT systems, as constructed in [2], 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 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 1. *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 $n > m$, remove arbitrarily $n - m$ 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. $\square$

Our main result in this section is a polynomial algorithm which guarantees finding a solution of cardinality at most $\lceil \frac{m}{\lfloor n/(d+1) \rfloor} \rceil$ for a d-bounded input instance.

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

PROOF. Observe that any primer subset with size at most $x = \lfloor \frac{n}{d+1} \rfloor$ is assignable, since there are at least $n - dx \geq x$ tags that are not adjacent to these primers. Our algorithm for MPC is straightforward. We form a primer cover by partitioning the set of primers into disjoint subsets of size at most x. The size of the cover is bounded by $\lceil \frac{m}{x} \rceil$. $\square$

Note that our algorithm actually gives a solution in which every subgraph is an independent set or, equivalently, the submatrices covering A are all 0. Proposition 1 implies that the algorithm achieves an approximation ratio of $\lceil \frac{m}{\lfloor n/(d+1) \rfloor} \rceil / 2$ for $m \leq n$. When $m > n$ at least $\lceil \frac{m}{n} \rceil$ subgraphs are needed in order to cover the primer set, so the approximation ratio obtained is $d + 2$.

3.2 Maximum Assignable Primer Sets

In this section we study a greedy approach to MPC that mimics approximation algorithms for SET COVER (cf. [6]). The scheme is recursive: The largest assignable subset in P is identified and removed, and the algorithm proceeds recursively on the remaining graph. If possible, this approach could guarantee an $O(\log m)$ approximation, 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 of G.*

THEOREM 3. *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 [8, 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 build 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 $1 \leq i, j \leq l$.

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 $|X| \geq lk$. We first claim that $|F| < l$. If $|F| \geq l$, 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. $\square$

Note that the related problem of finding a maximum induced matching in a bipartite graph is also NP-hard [4]. In the case that each primer has at most one adjacent tag, MAP can be solved in polynomial time. We omit the details.

3.3 Minimum Partition into Disjoint Assignable Subgraphs

Until now 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. Our covering algorithm is based on graph coloring and is given below.

THEOREM 4. *Let $G = (P, T, A)$ be an input bipartite graph in which the degree of each $p \in P$ is bounded by d, and $m \leq n$. Then we can find, in polynomial time, 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 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 $v_2 = (p_2, t_2)$ iff $(p_1, t_2) \in A$. 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 [14]. Each coloring class corresponds to the vertices of an assignable subgraph, and together with H these subgraphs cover P. □

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 5. *Let $G = (P, T, A)$ be an input bipartite graph with the degree of every vertex in P bounded by d. Suppose that $n \geq (k+1)m$, 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 all $n - m \geq mk$ 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 4, we obtain a solution of cardinality $2\lfloor \frac{d}{k} \rfloor$. □

Observing that Proposition 1 holds for the MPDAS problem as well, we conclude that the algorithms of Theorem 4 and Theorem 5 have approximation ratios of d and $\lfloor \frac{d}{k} \rfloor$, respectively.

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 mis-extensions. 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 slightly more solution-phase experiments (since, typically, a solution for MPC would have smaller cardinality than a solution for MPDAS on the same instance). This protocol has not been experimentally tested, to our knowledge. The principal motivation for MPDAS, therefore, remains purely mathematical.

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 to 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 A can be partitioned by rows into t assignable row 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, that represents the expected fraction of the antitags that potentially hybridize to a given primer used in the assay. 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 0 with probability $1 - p$. We shall derive a lower bound on $n(A)$ for matrices drawn from $D(m, n, p)$ and use it in testing our algorithmic approaches.

THEOREM 6. *Let matrix A be drawn from the probability distribution $D(m, n, p)$. Then, for every positive integer t,*

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

where $x = n(1-p)^{h-1}(1 - p + hp)$ and $h = \lceil \frac{m}{t} \rceil$.

PROOF. Let $X \sim \text{Binom}(n, s)$. We require the following Chernoff bound (cf. [1]):

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

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 on this probability by $f(h)$. It can be shown that $\log(f(h))$ is a concave function.

Now let A be drawn from $D(m,n,p)$ and consider a row-partition of A, into sets of sizes $h_1, h_2, \cdots, h_t$. Then the probability that all of these subsets 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, bounded from above by $f(\lceil \frac{m}{t} \rceil)^t$.

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

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 $\text{Prob}[n(A) \le 16] \le e^{-4,878}$. In general, we determine the lower bound of a given instance as the minimal t such that $\frac{t^m}{t!} \left(\frac{xe^{\frac{h-x}{h}}}{h} \right)^{ht} \ge 0.001$.

5. ALGORITHMIC APPROACHES

In this section we describe two heuristic approaches to MPC. Algorithm A is based on the theoretical analysis of Section 3.1. Algorithm B is a heuristic approach based on the set cover approximation method alluded to in Section 3.2.

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 builds a cover with size at most $\lceil \frac{m}{\lfloor n/(d+1) \rfloor} \rceil$, for any set of primers with degree bounded by d. It is described in Figure 2.

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.

Figure 2: Algorithm A.

The general scheme of Algorithm B is described in Figure 3. 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 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.

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.

Figure 3: Algorithm B.

6. EXPERIMENTAL RESULTS

6.1 Performance on Simulated 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 2000, the number of primers ranged from 1000 to 5000, and p was determined so that an average of 10 tags 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 6. Notably, Algorithm B 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).

	Tags (p)								
SNPs	500 (0.02)			1,000 (0.01)			2,000 (0.005)		
	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

Table 1: Comparison between Algorithms A and B on data simulated using the stochastic model for different parameter combinations. For each set of parameters recorded are the average cover size of algorithms A (column A) and B (column B), and the stochastic lower bound of Theorem 6 (column L).

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 model described in [2]. 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 Algorithm A.

	λ					
SNPs	6		7		8	
	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

Table 2: Comparison between Algorithms A and B on data simulated using the combinatorial model for different parameter combinations.

6.2 Performance on Real Data

We complemented our analyses on simulated data by applying Algorithm B to matrices derived from real genomic sequence data. Specifically, we retrieved 3304 SNP entries of the public Human SNP database [21] that were annotated with the 20 nucleotides immediately upstream of the respective SNP site. 3304 primers were obtained as reverse complements of these sequences.

We then employed the combinatorial construction scheme in [2] to generate two tag sets T_1 and T_2. The construction of [2] takes into account two parameters c and h, where $c < h$. c represents the maximal allowable hybridization potential for a tag and a foreign antitag. h represents the minimal allowable hybridization potential for a tag and its corresponding (perfect match) antitag. The thermodynamical model used in this representation employs the 2–4-rule [17], which estimates the melting temperature of a DNA sequence and its complement as twice the number of `A`s and `T`s, plus four times the number of `C`s and `G`s, in degrees Celsius. T_1 was generated using the parameters $c = 10$ and $h = 24$ and contains 2047 tags. T_2 contains 314 tags and was generated from the parameters $c = 8$ and $h = 20$. 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 [2]. Whenever the result of this rule, applied to any perfectly complementary substring between p and $\bar{t}$, exceeded the threshold of 20 (for A_1) or 16 (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 0.0043 and 0.0299, respectively. For A_1, Algorithm B found a cover of size 5 (where the last array contains only 13 tags), while the stochastic lower bound is 4 (using an estimated $p = 0.0043$). For A_2, Algorithm B used 24 arrays, while the lower bound lies at 16 (using an estimated $p = 0.0299$).

Acknowledgements

We thank Richard M. Karp for contributing valuable ideas and for many insightful comments on the manuscript. We thank Moni Naor and Pavol Hell for helpful discussions. T.H. was supported in part by a grant from the US-Israel Binational Science Foundation (BSF). R.S. was supported by a Fulbright grant.

7. REFERENCES

[1] N. Alon and J. H. Spencer. *The Probabilistic Method*. John Wiley and Sons, Inc., 1992.

[2] A. Ben-Dor, R. M. Karp, B. Schwikowski, and Z. Yakhini. Universal DNA tag systems: A combinatorial design scheme. *Journal of Computational Biology*, 7(3):503–519, 2000.

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

[4] K. Cameron. Induced matchings. *Discrete Applied Mathematics*, 24:97–102, 1989.

[5] M. Cargill, D. Altshuler, J. Ireland, P. Sklar, K. Ardlie, et al. Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genetics*, 22(3):231–8, 1999.

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

[7] R. Drmanac, G. Lennon, S. Drmanac, I. Labat, R. Crkvenjakov, and H. Lehrach. Partial sequencing by oligohybridization: Concept and applications in genome analysis. In *Proceedings of the first international conference on electrophoresis supercomputing and the human genome*, pages 60–75. World Scientific, 1991.

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

[9] N.P. Gerry, N.E. Witowski, J. Day, R.P. Hammer, G. Barany, et al. Universal DNA microarray method for multiplex detection of low abundance point mutations. *J. Mol. Biol.*, 292(2):251–62, 1999.

[10] D.M. Grant and M.S. Phillips. *Technologies for the Analysis of Single-Nucleotide Polymorphisms: An Overview*. Marcel Dekker, Inc., New York, 2001.

[11] J.G. Hacia. Resequencing and mutational analysis using oligonucleotide micro arrays. *Nature Genetics*, 21(1):42–47, January 1999.

[12] V. N. Kristensen, N. Harada, N. Yoshimura, E. Haraldsen, P. E. Lonning, et al. Genetic variants of cyp19 (aromatase) and breast cancer risk. *Oncogene*, 19(10):1329–33, March 2000.

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

[14] D. Matula and L. Beck. Smallest-last ordering and clustering and graph coloring algorithms. *Journal of the ACM*, 30:417–427, 1983.

[15] N. J. Risch. Searching for genetic determinants in the new millennium. *Nature*, 405(6788):847–56, 2000.

[16] J. P. Schouten, C. J. McElgunn, R. Waaijer,

D. Zwijnenburg, F. Diepvens, and G. Pals. Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. *Nucleic Acids Research*, 30(12), June 2002.

[17] T. Strachen and A.P. Read. *Human Molecular Genetics*. Bios scientific publishers, 1996.

[18] A. C. Syvanen. From gels to chips: "minisequencing" primer extension for analysis of point mutations and single nucleotide polymorphisms. *Hum. Mutat.*, 13(1):1–10, 1999.

[19] S. Venitt. Mechanisms of carcinogenesis and individual susceptibility to cancer. *Clin. Chem.*, 40(7.2):1421–5, July 1994.

[20] S. Venitt. Mechanisms of spontaneous human cancers. *Environ. Health Perspect.*, 104(3):633–7, May 1996.

[21] D. G. Wang, J. B. Fan, C. J. Siao, A. Berno, P. P. Young, et al. Large-scale identification, mapping, and genotyping of single nucleotide polymorphisms in the human genome. *Science*, 280(5366):1077–82, 1998.

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