High-Throughput SNP Genotyping by SBE/SBH

Ion I. Măndoiu and Claudia Prăjescu

***Abstract*— Despite much progress over the past decade, current Single Nucleotide Polymorphism (SNP) genotyping technologies still offer an insufficient degree of multiplexing when required to handle user selected sets of SNPs. In this paper we propose a new genotyping assay architecture combining multiplexed solution-phase single-base extension (SBE) reactions with sequencing by hybridization (SBH) using universal DNA arrays such as all k-mer arrays. Simulation results on datasets both randomly generated and extracted from the NCBI dbSNP database suggest that the SBE/SBH architecture provides a flexible and cost-effective alternative to genotyping assays currently used in the industry, enabling genotyping of up to hundreds of thousands of SNPs per assay.**

***Index Terms*— Single nucleotide polymorphisms, genotyping assay, universal DNA arrays, multiplexing algorithms.**

I. Introduction

AFTER the completion of the Human Genome Project has provided a blueprint of the DNA present in each human cell [2], genomics research is now focusing on the study of DNA variations that occur between individuals, seeking to understand how these variations confer susceptibility to common diseases such as diabetes or cancer. The most common form of genomic variation are the so called *single nucleotide polymorphisms* (SNPs), i.e., the presence of different DNA nucleotides, or *alleles*, at certain chromosomal locations. Determining the identity of alleles present in a DNA sample at a given set of SNP loci is called *SNP genotyping*.

The continuous progress in high-throughput genomic technologies has resulted in numerous SNP genotyping platforms combining a variety of allele discrimination techniques (sequencing, direct hybridization, primer extension, allele-specific PCR/ligation/cleavage, etc.), detection mechanisms (fluorescence, mass spectrometry, etc.) and reaction formats (solution phase, solid support, bead arrays), see, e.g., [3], [4] for comprehensive reviews. However, current technologies still offer an insufficient degree of multiplexing for fully-powered genome wide disease association studies that require genotyping of large sets of user selected SNPs [5]. The highest throughput is achieved by high-density mapping arrays produced by Affymetrix, which currently can simultaneously genotype about 250 thousands of *manufacturer selected* SNPs per array. Genotyping a comparable number of user selected SNPs would require an expensive and time-consuming re-design of array probes as well as a difficult re-engineering of the DNA amplification protocol.

Among technologies that allow genotyping of custom sets of SNPs one of the most successful ones is the use of DNA tag arrays [6], [7], [8], [9]. DNA tag arrays consist of a set of DNA strings called *tags*, designed such that each tag hybridizes strongly to its own *antitag* (Watson-Crick complement), but to no other antitag. The flexibility of tag arrays comes from combining solid-phase hybridization with the high sensitivity of single-base extension reactions, which has also been used for SNP genotyping in combination with MALDI-TOF mass spectrometry [10]. Commercially available tag arrays have between 2,000 and 10,000 tags [11], [12]. However, the number of SNPs that can be genotyped per array is typically smaller than the number of tags since some of the tags must remain unassigned due to cross-hybridization with the primers [13], [14]. Another factor limiting the wider use of tag arrays is the relatively high cost of synthesizing the reporter probes, which have a typical length of 40 nucleotides.

In the k-mer array format [15], all 4^k DNA probes of length k are spotted or synthesized on the solid array substrate. This format was originally proposed for performing *sequencing by hybridization (SBH)*, which seeks to reconstruct an unknown DNA sequence based on its k-mer spectrum [16]. However, the sequence length for which unambiguous reconstruction is possible with high probability is surprisingly small [17], and, despite several suggestions for improvement, such as the use of gapped probes [18] and pooling of target sequences [19], sequencing by hybridization has not become practical so far.

In this paper we propose a new genotyping assay architecture combining multiplexed solution-phase single-base extension (SBE) reactions with sequencing by hybridization (SBH) using universal DNA arrays such as all k-mer arrays. SNP genotyping using SBE/SBH assays requires the following steps (see Figure 1): (1) Synthesizing primers complementing the genomic sequence immediately preceding SNPs of interest; (2) Hybridizing primers with the genomic DNA; (3) Extending each primer by a single base using polymerase enzyme and dideoxynucleotides labeled with 4 different fluorescent dyes; and finally (4) Hybridizing extended primers to a universal DNA array and determining the identity of the bases that extend each primer by hybridization pattern analysis.

To the best of our knowledge the combination of the two technologies in the context of SNP genotyping has not been explored thus far. The most closely related genotyping assay is the generic Polymerase Extension Assay (PEA) recently proposed in [20]. In PEA, short amplicons containing the SNPs of interest are hybridized to an all k-mers array of *primers* that are subsequently extended via single-base extension reactions. Hence, in PEA the SBE reactions take place on solid support,

Manuscript received June 21, 2006; revised October 6, 2006. A preliminary version of this paper was published in [1]. Work supported in part by NSF awards IIS-0546457 and DBI-0543365, and a Faculty Large Research Grant from the University of Connecticut's Research Foundation.

IIM is with the Computer Science and Engineering Department, University of Connecticut, 371 Fairfield Rd., Unit 2155, Storrs, CT 06269-2155, USA (e-mail: ion@engr.uconn.edu).

CP is with Siemens VDO Automotive, Bd. Poitiers 10, 700671 Iaşi, Romania. Work done while she was with the Computer Science and Engineering Department at the University of Connecticut.

similar to *arrayed primer extension* (APEX) assays which use SNP specific primers spotted on the array [21].

As the SBH multiplexing technique of [19], the SBE/SBH assays lead to high array probe utilization since we hybridize to the array a large number of short extended primers. However, the main power of the method lies in the fact that the sequences of the labeled oligonucleotides hybridized to the array are known a priori (up to the identity of extending nucleotides). While genotyping with SBE/SBH assays uses similar general principles as the PEA assays proposed in [20], there are also significant differences. A major advantage of SBE/SBH is the much shorter length of extended primers compared to that of PCR amplicons used in PEA. A second advantage is that *all* probes hybridizing to an extended primer are informative in SBE/SBH assays, regardless of array probe length (in contrast, only probes hybridizing with a substring containing the SNP site are informative in PEA assays). As shown by the experimental results in Section IV these advantages translate into an increase by orders of magnitude in multiplexing rate compared to the results reported in [20]. We further note that PEA's effectiveness crucially depends on the ability to amplify very short (preferably 40bp or less) genomic fragments spanning the SNP loci of interest. This limits the achievable degree of multiplexing in PCR amplification [22], making PCR amplification the main bottleneck for PEA assays. Full flexibility in picking PCR primers is preserved in SBE/SBH assays.

Under the assumption of perfect hybridization, unambiguous SBE/SBH genotyping of a set of SNPs requires selecting primers upstream of the SNPs such that each primer hybridizes to at least one array probe that hybridizes to no other primer that can be extended by a common base. Our contributions in this paper include a study of multiplexing algorithms for SBE/SBH genotyping assays and preliminary experimental results showing the achievable tradeoffs between the number of array probes and primer length on one hand and the number of SNPs that can be simultaneously genotyped on the other. We prove that the problem of selecting a maximum size subset of SNPs that can be unambiguously genotyped in a single SBE/SBH assay is NP-hard, and propose efficient heuristics with good practical performance. Our heuristics take into account the freedom of selecting primers from both strands of the genomic DNA as well as the presence of disjoint allele sets among genotyped SNPs. Furthermore, our heuristics can enforce redundancy constraints facilitating reliable genotyping in the presence of hybridization errors. Preliminary simulation results presented in Section IV suggest that the SBE/SBH architecture provides a flexible and cost-effective alternative to genotyping assays currently used in the industry, enabling genotyping of up to hundreds of thousands of user selected SNPs per assay.

The rest of the paper is organized as follows. In Section II we formalize two optimization problems that arise in genotyping large sets of SNPs using SBE/SBH assays: the problem of partitioning a set of SNPs into the minimum number of decodable subsets of SNPs, and that of finding a maximum size decodable subset of a given set of SNPs. We also establish hardness results for the latter problem. In Section III we propose several efficient heuristics. Finally, we present experimental results on both randomly generated datasets and instances extracted from the NCBI dbSNP database in Section IV and conclude in Section V.

II. Problem Formulations and Complexity

A set of SNP loci can be unambiguously genotyped by SBE/SBH if every combination of SNP genotypes yields a different hybridization pattern (defined as the vector of dye colors observed at each array probe). To formalize the requirements of unambiguous genotyping, we first consider a simplified SBE/SBH assay consisting of four parallel *single-color* SBE/SBH reactions, one for each possible SNP allele. Under this scenario, only one type of dideoxynucleotide is added to each SBE reaction, corresponding to the complement of the tested SNP allele. Therefore, a primer is extended in such a reaction if the tested allele is present at the SNP locus probed by the primer, and is left un-extended otherwise.

Let $\mathcal{P}$ be the set of primers used in a single-color SBE/SBH reaction involving dideoxynucleotide $e \in$ {A,C,G,T}. From the resulting hybridization pattern we must be able to infer for every $p \in \mathcal{P}$ whether or not p was extended by e. The extension of p by e will result in a fluorescent signal at all array probes that hybridize with pe. However, some of these probes can give a fluorescent signal even when p is not extended by e, due to hybridization to other extended primers. Since in the worst case *all* other primers are extended, it must be the case that at least one of the probes that hybridize to pe does not hybridize to any other extended primer.

Formally, let $X \subset \{A,C,G,T\}^*$ be the set of array probes. For every string $y \in \{A,C,G,T\}^*$, let the *spectrum of* y *in* X, denoted $Spec_X(y)$, be the set of probes of X that hybridize with y. Under the assumption of perfect hybridization, $Spec_X(y)$ consists of those probes of X that are Watson-Crick complements of substrings of y. Then, a set of primers $\mathcal{P}$ is said to be *decodable* with respect to extension e if and only if, for every $p \in \mathcal{P}$,

$$Spec_X(pe) \setminus \bigcup_{p' \in \mathcal{P} \setminus \{p\}} Spec_X(p'e) \neq \emptyset \qquad (1)$$

Decoding constraints (1) can be directly extended to 4-color SBE/SBH experiments, in which each type of extending base is labeled by a different fluorescent dye. As before, let $\mathcal{P}$ be the set of primers, and, for each primer $p \in \mathcal{P}$, let $E_p \subseteq \{A,C,G,T\}$ be the set of possible extensions of p, i.e., Watson-Crick complements of corresponding SNP alleles. If we assume that any combination of dyes can be detected at an array probe location, unambiguous decoding is guaranteed if, for every $p \in \mathcal{P}$ and every extending nucleotide $e \in E_p$,

$$Spec_X(pe) \setminus \bigcup_{p' \in \mathcal{P} \setminus \{p\}, e' \in E_{p'}} Spec_X(p'e') \neq \emptyset \qquad (2)$$

In the following, we refine (2) to improve practical reliability of SBE/SBH assays. More precisely, we impose additional constraints on the set of probes considered to be *informative* for each SNP allele. First, to enable reliable genotyping of genomic samples that contain SNP alleles at very different

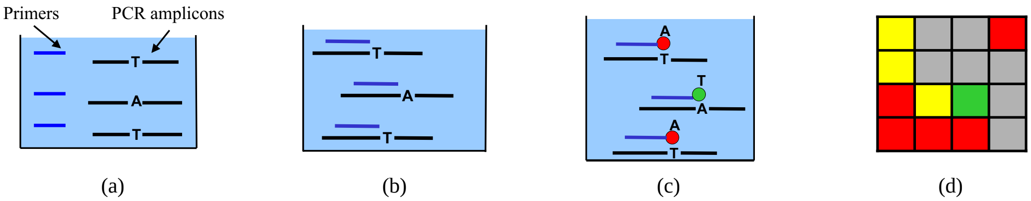


Fig. 1. *SBE/SBH assay: (a) Primers complementing genomic sequence upstream of each SNP locus are mixed in solution with the genomic DNA sample. (b) Temperature is lowered allowing primers to hybridize to the genomic DNA. (c) Polymerase enzyme and dideoxynucleotides labeled with 4 different fluorescent dyes are added to the solution, causing each primer to be extended by a nucleotide complementing the SNP allele. (d) Extended primers are hybridized to a universal DNA array and genotypes are determined by analyzing the resulting hybridization pattern.*

concentrations (as a result of uneven efficiency in the PCR amplification step or of pooling DNA samples from different individuals), we require that a probe that is informative for a certain SNP locus must not hybridize to primers corresponding to other SNP loci. Second, since recent studies by Naef et al. [23] suggest that fluorescent dyes can significantly interfere with oligonucleotide hybridization on solid support, possibly destabilizing hybridization to a complementary probe on the array, in this paper we use a conservative approach and require that each probe that is informative for a certain SNP allele must hybridize to a substring of the corresponding *un-extended* primer. On the other hand, informative probes are required not to hybridize with any other extended primer, even if such hybridizations involve the fluorescently labeled extension nucleotides. Finally, we introduce a *decoding redundancy* parameter $r \geq 1$, and require that each SNP have at least r informative probes. Such a redundancy constraint facilitates reliable genotype calling in the presence of hybridization errors. Clearly, the larger the value of r, the more hybridization errors that can be tolerated. If a simple majority voting scheme is used for making allele calls, the assay can tolerate up to $\lfloor r/2 \rfloor$ hybridization errors involving the r informative probes of each SNP. Furthermore, since the informative probes of a SNP are required to hybridize *exclusively* with the primer corresponding to the SNP, the redundancy requirement provides a powerful mechanism for gauging the extent of hybridization errors. Indeed, each unintended hybridization at an informative probe for a bi-allelic SNP has a dye complementary to one of the SNP alleles with probability of only 1/2, and the probability that k such errors pass undetected decreases exponentially in k.

The refined set of constraints is captured by the following definition, where, for every primer $p \in \{A, C, G, T\}^*$ and set of extensions $E \subseteq \{A, C, G, T\}$, we let

$$Spec_X(p, E) = \bigcup_{e \in E} Spec_X(pe)$$

Definition 1: A set of primers $\mathcal{P}$ is said to be *strongly r-decodable* with respect to extension sets E_p, $p \in \mathcal{P}$, if and only if, for every $p \in \mathcal{P}$,

$$\left| Spec_X(p) \setminus \bigcup_{p' \in \mathcal{P} \setminus \{p\}} Spec_X(p', E_{p'}) \right| \geq r \tag{3}$$

Note that testing whether or not a given set of primers is strongly r-decodable can be easily accomplished in time linear in the total length of the primers.

Genotyping a large set of SNPs will, in general, require more than one SBE/SBH assay. This rises the problem of partitioning a given set of SNPs into the smallest number of strongly r-decodable subsets. For each SNP locus there are typically two different primers that can be used for genotyping. As shown in [14] for the case of SNP genotyping using tag arrays, exploiting this degree of freedom significantly increases achievable multiplexing rates. Therefore, we next extend Definition 1 to capture this degree of freedom. Let P_i be the *pool of primers* that can be used to genotype the SNP at locus i. Similarly to Definition 1, we have:

Definition 2: A set of primer pools $\mathcal{P} = \{P_1, \ldots, P_n\}$ is said to be *strongly r-decodable* if and only if there is a primer p_i in each pool P_i such that $\{p_1, \ldots, p_n\}$ is strongly r-decodable with respect to the respective extension sets E_{p_i}, $i = 1, \ldots, n$.

Primers $p_1, p_2, \ldots, p_n$ above are called the *representative primers* of pools $P_1, P_2, \ldots, P_n$, respectively. The SNP partitioning problem can then be formulated as follows:

Minimum Pool Partitioning Problem (MPPP): *Given primer pools $\mathcal{P} = \{P_1, \ldots, P_n\}$, extension sets E_p, $p \in \cup_{i=1}^n P_i$, probe set X, and redundancy parameter r, find a partitioning of $\mathcal{P}$ into the minimum number of strongly r-decodable subsets.*

A natural strategy for solving MPPP, is to find a maximum size strongly r-decodable subset of pools, remove it from $\mathcal{P}$, and then repeat the procedure until no more pools are left in $\mathcal{P}$. This greedy strategy for solving MPPP has been shown to empirically outperform other algorithms for solving the similar partitioning problem for PEA assays [20]. In the case of SBE/SBH, the optimization involved in the main step of the greedy strategy is formalized as follows:

Maximum r-Decodable Pool Subset Problem (MDPSP): *Given primer pools $\mathcal{P} = \{P_1, \ldots, P_n\}$, extension sets E_p, $p \in \cup_{i=1}^n P_i$, probe set X, and redundancy parameter r, find a strongly r-decodable subset $\mathcal{P}' \subseteq \mathcal{P}$ of maximum size. In addition, for each pool $P_i \in \mathcal{P}'$, find its representative primer.*

Unfortunately, as shown in next theorem, MDPSP is NP-hard even for the case when the redundancy parameter is 1 and each pool has exactly one primer.

```
Input: Pools $\mathcal{P} = \{P_1, \ldots, P_n\}$, extension sets $E_p$, probe set $X$,
and redundancy parameter $r$
Output: Strongly $r$-decodable subset of pools $\mathcal{P}' \subseteq \mathcal{P}$ and set $R$
of representative primers for the pools in $\mathcal{P}$'
--------------------------------------------------------------------
$\mathcal{P}' \leftarrow \emptyset$,  $R \leftarrow \emptyset$
For each $P \in \mathcal{P}$ do
    For each $p \in P$ do
        If $R \cup \{p\}$ satisfies (3)
        Then
            $\mathcal{P}' \leftarrow \mathcal{P}' \cup P$
            $R \leftarrow R \cup \{p\}$
            Exit inner For
```

Fig. 2. The Sequential Greedy algorithm.

Theorem 1: MDPSP is NP-hard, even when restricted to instances with $r = 1$ and $|P| = 1$ for every $P \in \mathcal{P}$.

Proof: We will use a reduction from the *maximum induced matching* problem in bipartite graphs, which is defined as follows:

Maximum Induced Matching (MIM) Problem in Bipartite Graphs: *Given a bipartite graph $G = (U \cup V, E)$, find maximum size subsets $U' \subseteq U$, $V' \subseteq V$, with $|U'| = |V'|$ such that the subgraph of G induced by $U' \cup V'$ is a matching.*

The MIM problem in bipartite graphs is known to be NP-hard even for graphs with maximum degree 3 [24]. Let $G = (U \cup V, E)$ be such a bipartite graph with maximum degree 3. Without loss of generality we may assume that every vertex in G has degree at least 1. We will denote by $N(u)$ the *neighborhood* of vertex $u \in U \cup V$, i.e., the set of vertices adjacent with u in G.

We construct an instance of MDPSP as follows: Let $r = 1$ and $l = \lceil \log_2 |V| \rceil$. For every $v \in V$ we add to X a distinct probe $x_v \in \{\text{A,T}\}^l$; note that this can be done since $|\{\text{A,T}\}^l| = 2^l > |V|$ by our choice of l. For every $u \in U$, with neighborhood $N(u) = \{v_1, v_2, v_3\}$, we construct a primer $p_u = x_{v_1}\text{C}x_{v_2}\text{C}x_{v_3}$ and a pool $P_u = \{p_u\}$. We use a similar construction for vertices $u \in U$ with only 1 or 2 neighbors. Note that in each case the pool P_u consists of a single primer p_u of length at most $3l + 2$. For each constructed primer p, the set of possible extensions is defined as $E_p = \{\text{G,C}\}$. Since the probes of X contain only A's and T's, for every primer p_u, $u \in U$, $Spec_X(p_u, E_{p_u}) = Spec_X(p_u) = \{x_v \in X |\ v \in N(u)\}$.

Let $U' \subseteq U$, $V' \subseteq V$, $|U'| = |V'|$, be subsets of vertices such that $U' \cup V'$ induces a matching in G. Let $\mathcal{P}' = \{P_u |\ u \in U'\}$. For every $u \in U'$, exactly one of u's neighbors, denoted v_u, appears in V', because $U' \cup V'$ induces a matching. Furthermore, for each $u' \in U' \setminus \{u\}$, $(u', v_u) \notin E$, and therefore $x_{v_u} \notin Spec_X(p_{u'}, E_{p_{u'}})$. Thus, for every $u \in U'$,

$$x_{v_u} \in Spec_X(p_u) \setminus \bigcup_{\{p_{u'}\} \in \mathcal{P}' \setminus \{p_u\}} Spec_X(p_{u'}, E_{p_{u'}})$$

which means that $\mathcal{P}'$ is a strongly 1-decodable subset of pools of the same size as the induced matching of G.

```
Input: Pools $\mathcal{P} = \{P_1, \ldots, P_n\}$, extension sets $E_p$, probe set $X$,
and redundancy parameter $r$
Output: Strongly $r$-decodable subset of pools $\mathcal{P}' \subseteq \mathcal{P}$ and set $R$
of representative primers for the pools in $\mathcal{P}$'
--------------------------------------------------------------------
Construct hybridization graph $G$; $\mathcal{P}' \leftarrow \emptyset$; $R \leftarrow \emptyset$
While $G$ is not empty do
    Find a minimum degree primer $p$, and let $P$ be
        its pool
    $\mathcal{P}' \leftarrow \mathcal{P}' \cup \{P\}$
    $R \leftarrow R \cup \{p\}$
    For each $(p') \in P \setminus \{p\}$ do remove-primer$(p')$
    Let $|N^+(p)| = k$ and let $\{x_1, \ldots, x_k\}$ be the probes
        in $N^+(p)$, indexed in increasing degree order
    For each $x \in \{x_1, \ldots, x_r\}$ do
        For each $(p') \in N^+(x) \cup N^-(x)$ do
            remove-primer$(p')$
        Delete vertex $x$ from $G$
    For each $x \in \{x_{r+1}, \ldots, x_k\} \cup N^-(p)$ do
        remove-probe$(x)$
```

Fig. 3. MinPrimerGreedy greedy algorithm.

Conversely, let $\mathcal{P}'$ be a strongly 1-decodable subset of $\mathcal{P}$, and let $U' = \{u \in U |\ \{p_u\} \in \mathcal{P}'\}$. Since $\mathcal{P}'$ is 1-decodable, for every primer p_u with $\{p_u\} \in \mathcal{P}'$, there must exist a probe $x \in X$ such that $x \in Spec_X(p_u)$ and $x \notin Spec_X(p_{u'}, E_{p_{u'}})$ for every $\{p_{u'}\} \in \mathcal{P}' \setminus \{p_u\}$. Because $Spec_X(p_u) = \{x_v \in X |\ v \in N(u)\}$, it follows that every vertex $u \in U'$ has a neighbor $v \in V$ that is not a neighbor of any other $u' \in U' \setminus \{u\}$. Let v_u be such a neighbor (pick v_u arbitrarily if more than one vertex in V satisfies above property), and let $V' = \{v_u |\ u \in U'\}$. It is clear that $U' \cup V'$ induce a matching of size $|\mathcal{P}'|$ in G.

Thus, for every integer k, there is a one-to-one correspondence between induced matchings of size k in G and strongly 1-decodable subsets of k pools in the constructed instance of MDPSP, and NP-hardness of MDPSP follows. ■

The reduction in the proof of Theorem 1 preserves the size of the optimal solution, and therefore any hardness of approximation result for the MIM in bipartite graphs will also hold for MDPSP, even when restricted to instances with $r = 1$ and $|P| = 1$ for every $P \in \mathcal{P}$. Since Duckworth et al. [25] proved that it is NP-hard to approximate MIM in bipartite graphs with maximum degree 3 within a factor of 6600/6659, we get:

Theorem 2: It is NP-hard to approximate MDPSP within a factor of 6600/6659, even when restricted to instances with $r = 1$ and $|P| = 1$ for every $P \in \mathcal{P}$.

III. Algorithms

In this section we describe three heuristic approaches to MDPSP. The first one is a naive greedy algorithm that sequentially evaluates the primers in the given pools in an arbitrary order. The algorithm picks a primer p to be the representative of pool $P \in \mathcal{P}$ if p together with the representatives already picked satisfy condition (3). The pseudocode of this algorithm, which we refer to as Sequential Greedy, is given in Figure 2.

The next two algorithms are inspired by the Min-Greedy algorithm in [25], which approximates MIM in d-regular graphs within a factor of $d-1$. For the MIM problem, the Min-Greedy algorithm picks at each step a vertex u of minimum

```
Input: Pools $\mathcal{P} = \{P_1, \ldots, P_n\}$, extension sets $E_p$, probe set $X$,
and redundancy parameter $r$
Output: Strongly $r$-decodable subset of pools $\mathcal{P}' \subseteq \mathcal{P}$ and set $R$
of representative primers for the pools in $\mathcal{P}$'
------------------------------------------------------------------------
Construct hybridization graph $G$; $\mathcal{P}' \leftarrow \emptyset$; $R \leftarrow \emptyset$
While $G$ is not empty do
    Find a minimum degree probe $x$
    Find a minimum degree primer $p$, and let $P$ be
        its pool
    $\mathcal{P}' \leftarrow \mathcal{P}' \cup \{P\}$
    $R \leftarrow R \cup \{p\}$
    For each $p' \in P \setminus \{p\}$ do remove-primer$(p')$
    Let $|N^+(p)| = k$ and let $\{x_1, \ldots, x_k\}$ be the probes
        in $N^+(p)$, indexed in increasing degree order
    For each $x \in \{x_1, \ldots, x_r\}$ do
        For each $p' \in N^+(x) \cup N^-(x)$ do
            remove-primer$(p')$
        Delete vertex $x$ from $G$
    For each $x \in \{x_{r+1}, \ldots, x_k\} \cup N^-(p)$ do
        remove-probe$(x)$
```

Fig. 4. MinProbeGreedy greedy algorithm.

```
remove-primer $(p)$
------------------------------------------------------------------------
For all $x \in N^+(p)$ do
    $N^+(x) \leftarrow N^+(x) \setminus \{p\}$
    If $|N^+(x)| = 0$ then remove-probe $(x)$
For all $x \in N^-(p)$ do $N^-(x) \leftarrow N^-(x) \setminus \{p\}$
    Delete $p$ from $G$
```

Fig. 5. The remove-primer subroutine.

```
remove-probe $(x)$
------------------------------------------------------------------------
For all $p \in N^+(x)$ do
    $N^+(p) \leftarrow N^+(p) \setminus \{x\}$
    If $|N^+(p)| < r$ then remove-primer $(p)$
For all $p \in N^-(x)$ do $N^-(p) \leftarrow N^-(p) \setminus \{x\}$
    Delete $x$ from $G$
```

Fig. 6. The remove-probe subroutine.

degree and a vertex v, which is a minimum degree neighbor of u. All the neighbors of u and v are deleted and the edge (u, v) is added to the induced matching. The algorithm stops when the graph becomes empty.

Each instance of MDPSP can be represented as a bipartite *hybridization graph* $G = ((\bigcup_{i=1}^{n} P_i) \cup X, E)$, with the left side containing all primers in the given pools and the right side containing the array probes. There is an edge between primer p and probe $x \in X$ if and only if $x \in Spec_X(p, E_p)$. As discussed in Section II, we need to distinguish between the hybridizations that involve fluorescently labeled nucleotides and those that do not. Thus, for every primer p, we let $N^+(p) = Spec_X(p)$ and $N^-(p) = Spec_X(p, E_p) \setminus Spec_X(p)$. Similarly, for each probe $x \in X$, we let $N^+(x) = \{p | \ x \in N^+(p)\}$ and $N^-(x) = \{p | \ x \in N^-(p)\}$.

We considered two versions of the Min-Greedy algorithm when run on the bipartite hybridization graph, depending on the side from which the minimum degree vertex is picked. In the first version, referred to as MinPrimerGreedy, we pick first a minimum degree node from the primers side, while in the second version, referred to as MinProbeGreedy, we pick first a minimum degree node from the probes side. Thus, MinPrimerGreedy greedy picks at each step a minimum degree primer p and pairs it with a minimum degree probe $x \in N^+(p)$. MinProbeGreedy greedy, selects at each step a minimum degree probe x and pairs it with a minimum degree primer p in $N^+(x)$. In both algorithms, all neighbors of p and x and their incident edges are removed from G. Also, at each step, the algorithms remove all vertices u, for which $N^+(u) = \emptyset$. These deletions ensure that the primers p selected at each step satisfy condition (3). Both algorithms stop when the graph becomes empty.

As described so far, the MinPrimerGreedy and MinProbeGreedy algorithms work when each pool contains only one primer and when the redundancy is 1. We extended the two variants to handle pools of size greater than 1 by simply removing from the graph all primers $p' \in P \setminus \{p\}$ when picking primer p from pool P. If the redundancy r is greater than 1, then whenever we pick a primer p, we also pick it's r probe neighbors from $N^+(p)$ with the smallest degrees (breaking ties arbitrarily). The primer neighbors of all these r probes will then be deleted from the graph. Moreover, the algorithm maintains the invariant that $|N^+(p)| \geq r$ for every primer p and $|N^+(x)| \geq 1$ for every probe x by removing primers/probes for which the degree decreases below these bounds. Figures 3 and 4 give the pseudocode for the MinPrimerGreedy, respectively the MinProbeGreedy greedy algorithms. For the sake of clarity, they use two subroutines for removing a primer vertex, respectively a probe vertex, which are separately described in Figures 5 and 6.

Algorithms MinPrimerGreedy and MinProbeGreedy can be implemented efficiently using a Fibonacci heap for maintaining the degrees of primers, respectively of probes. Let N be the total number of primers in the n pools, m be the number of probes in X, and k be the size of the r-decodable set returned by the algorithm. Since each primer has bounded degree, the sorting of probe degrees requires $O(k)$ total time. The total number of edges in the hybridization graph is $O(N + m)$. By using a Fibonacci heap, finding a minimum degree primer (probe) can be done in $O(\log N)$ (respectively $O(\log m)$) and each primer degree update can be done in amortized $O(1)$ time. Thus, the total runtime for MinPrimerGreedy algorithm is $O(k \log N + N + m)$, and the total runtime for MinProbeGreedy algorithm is $O(k \log m + N + m)$.

IV. Experimental Results

We considered two types of data sets: randomly generated datasets containing between 1,000 to 200,000 pools of 1 or 2 primers, and 2-primer pools representing over 9 million reference SNPs extracted from the NCBI dbSNP database build 125. We simulated two different types of array probe sets. First, we simulated all 10-mer arrays, which are well studied in the context of sequencing by hybridization. Since a major drawback of all k-mer arrays is the wide range of probe melting temperatures, we also simulated probe sets consisting of all *c-tokens* for $c = 13$. Following [26], a DNA string is called a c-token if it has a weight of c or more and all its proper suffixes have weight strictly less than c, where the weight of a DNA string is the number of A and T bases plus twice the number of C and G bases. Since the weight of c-tokens is either c or $c+1$, it follows that their melting temperature computed according to the 2-4 rule of Wallace [27] varies in a range of only 2°C. There are approximately 1 million 10-mers and 645 thousand 13-tokens. Probe sets of this size can be synthesized using existing photolithographic technologies such as those employed by Affymetrix.

A. Results on Synthetic Datasets

In a first set of experiments we compared the three MDPSP algorithms on randomly generated datasets. In these experiments we used a primer length of 20, which is a typical length used in genotyping using tag arrays. For each of the two considered probe sets, we ran simulations with pools of size 1 or 2, to see how much can be gained by allowing primers to be selected from either strand of the genomic DNA. We also considered both primer extension sets of size 4 – modeling the testing of all four nucleotides at every SNP locus, as commonly done in current Affymetrix genotyping assays – or 2 – modeling the testing of possible alleles only.

The results in Table I show that using the flexibility of picking primers from either strand of the genomic sequence yields an increase in the size of the strongly r-decodable pool subset when the number of pools is large. All algorithms benefit from this degree of freedom, with maximum increases of 25%, 22%, and 28% for SequentialGreedy, MinPrimerGreedy, and MinProbeGreedy respectively. Taking into account the reduced number of possible extensions further increases the size of computed decodable pool subsets, by up to an additional 9-11%.

The MinProbeGreedy algorithm produces consistently better results compared to the MinPrimerGreedy variant. On the other hand, neither Sequential Greedy nor MinProbeGreedy dominates the other for all range of instance parameters – Sequential Greedy generally gives better results for 10-mer experiments with high redundancy values, while MinProbeGreedy generally gives better results for 10-mer experiments with low redundancy requirements and for 13-token experiments. Since the algorithms are very fast, a feasible practical meta-heuristic is to run all three algorithms and take the best solution.

In a second set of experiments we explored the degree of freedom given by the primer length. Figure 7 gives the tradeoff between primer length and the size of the strongly r-decodable pool subsets computed by the MDPSP meta-heuristic suggested above for pools with 2 primers, 2 possible extensions per primer and arrays with all 10-mers, respectively all 13-tokens. We notice that, for both probe sets, the optimal primer length increases with the redundancy parameter. For any fixed array probe set and redundancy requirement, we need a minimum primer length to be able to satisfy constraints (3). Increasing the primer length beyond this minimum primer length is at first beneficial, since it increases the number of array probes that hybridize with the primer. However, if primer length increases too much, a large number of these probes become non-specific, and the multiplexing rate starts to decline, especially for low redundancy requirements. We further notice that, for both probe sets, the optimal primer length increases with the redundancy parameter.

B. Results on dbSNP Data

To stress-test our methods, we extracted a total of over 9 million primer pools corresponding to reference SNPs in human chromosomes 1-22, X, and Y in the NCBI dbSNP database build 125. We constructed a dataset for each of the 24 chromosomes by creating a 2-primer pool for each reference SNP for which dbSNP contains at least 20 non-degenerate base pairs of flanking sequence on both sides. Since these large sets of pools must be partitioned between multiple SBE/SBH experiments, we used a simple MPPP algorithm which iteratively finds maximum r-decodable pool subsets using the sequential greedy algorithm.

Table II gives the number of arrays required to cover $10-50\%$ of the extracted reference SNPs for each chromosome when using primers of length 20. In practical association studies even lower SNP coverage (and hence fewer arrays) may suffice due to the high degree of linkage disequilibrium between SNPs [28].

V. Conclusions

Simulation results presented in this paper suggest that SBE/SBH offers a promising alternative to current SNP genotyping assays. SBE/SBH can also be used in other applications that require detecting the presence or absence of a large number of substrings in a sample of genomic DNA, such as large-scale species identification [29]. For such applications, assay specificity and sensitivity can be further enhanced by combining primer extension with ligation to a second locus-specific probe and PCR amplification of the ligation product prior to hybridization to the universal array, similar to the steps of the GoldenGate genotyping protocol used by Illumina [30].

References

[1] I. Mandoiu and C. Prajescu, “High-throughput SNP genotyping by SBE/SBH,” in *Proc. 6th International Conference on Computational Science (ICCS 2006), Part II*, ser. Lecture Notes in Computer Science, V.N. Alexandrov et al., Ed., vol. 3992. Berlin: Springer-Verlag, 2006, pp. 742–749.

[2] International Human Genome Sequencing Consortium, “Finishing the euchromatic sequence of the human genome,” *Nature*, vol. 431, pp. 931–945, 2004.

[3] S. Jenkins and N. Gibson, “High-throughput SNP genotyping,” *Comparative and Functional Genomics*, vol. 3, pp. 57–66, 2002.

[4] P. Kwok, “Methods for genotyping single nucleotide polymorphisms,” *Annual Review of Genomics and Human Genetics*, vol. 2, pp. 235–258, 2001.

[5] C. Carlson, M. Eberle, M. Rieder, Q. Yi, L. Kruglyak, and D. Nickerson, “Selecting a maximally informative set of snps for association analyses using linkage disequilibrium,” *American Journal of Human Genetics*, vol. 74, pp. 106–120, 2004.

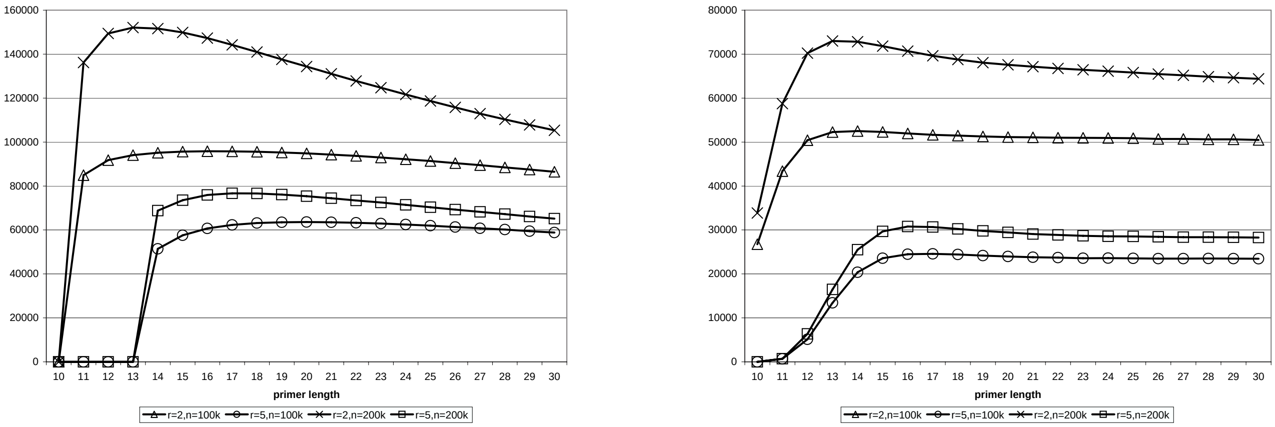
[6] S. Brenner, “Methods for sorting polynucleotides using oligonucleotide tags,” *US Patent 5,604,097*, 1997.

[7] N. Gerry, N. Witowski, J. Day, R. Hammer, G. Barany, and F. Barany, “Universal DNA microarray method for multiplex detection of low abundance point mutations,” *J. Mol. Biol.*, vol. 292, no. 2, pp. 251–262, 1999.

TABLE I

SIZE OF THE STRONGLY r-DECODABLE POOL SUBSET COMPUTED BY THE THREE MDPSP ALGORITHMS FOR PRIMERS OF LENGTH 20 WITH REDUNDANCY $r \in \{1, 2, 5\}$ (AVERAGES OVER 10 TEST CASES).

r	# pools	Algorithm	All 10-mer array			All 13-token array		
			$\lvert P_i\rvert = 1$ $\lvert E_p\rvert = 4$	$\lvert P_i\rvert = 2$ $\lvert E_p\rvert = 4$	$\lvert P_i\rvert = 2$ $\lvert E_p\rvert = 2$	$\lvert P_i\rvert = 1$ $\lvert E_p\rvert = 4$	$\lvert P_i\rvert = 2$ $\lvert E_p\rvert = 4$	$\lvert P_i\rvert = 2$ $\lvert E_p\rvert = 2$
1	10000	Sequential	10000	10000	10000	9420	9927	9953
		MinPrimer	10000	10000	10000	9472	9801	9866
		MinProbe	10000	10000	10000	9550	9980	9990
	20000	Sequential	19999	20000	20000	16656	18931	19197
		MinPrimer	19999	20000	20000	16673	18204	18573
		MinProbe	19999	20000	20000	17430	19613	19746
	100000	Sequential	93632	98630	99478	45064	56064	59498
		MinPrimer	93642	96712	98209	42824	51540	55031
		MinProbe	93837	99601	99885	51448	65877	69188
	200000	Sequential	140820	157908	166796	61351	76037	81443
		MinPrimer	139787	154028	164696	57530	70048	75470
		MinProbe	141614	160532	173910	72230	91488	97154
2	10000	Sequential	10000	10000	10000	8616	9611	9716
		MinPrimer	10000	10000	10000	8572	9214	9381
		MinProbe	10000	10000	10000	8896	9783	9851
	20000	Sequential	19992	20000	20000	14060	16839	17409
		MinPrimer	19992	20000	20000	13699	15613	16231
		MinProbe	19993	20000	20000	15152	17980	18396
	100000	Sequential	82315	90627	94420	32223	39839	42814
		MinPrimer	81056	85852	90098	30138	36595	39542
		MinProbe	82522	90935	94868	38246	48131	51125
	200000	Sequential	109450	122470	130911	41783	50811	54858
		MinPrimer	104891	114624	125287	39125	47357	51390
		MinProbe	109252	122986	134342	51198	63112	67567
5	10000	Sequential	9972	10000	10000	6284	7662	8000
		MinPrimer	9969	9998	9999	5939	6976	7324
		MinProbe	9970	10000	10000	6324	7651	7988
	20000	Sequential	19498	19967	19985	9139	11399	12143
		MinPrimer	19435	19804	19905	8504	10308	11007
		MinProbe	19468	19931	19973	9222	11530	12240
	100000	Sequential	52078	59021	63631	17580	21359	23232
		MinPrimer	47922	52711	57521	16252	19645	21421
		MinProbe	49329	55573	61043	18048	22119	23977
	200000	Sequential	62791	70334	75361	21762	25859	28234
		MinPrimer	56160	61406	67565	20226	24058	26297
		MinProbe	58565	65344	72313	22602	27186	29439



(a) (b)

Fig. 7. Size of the best strongly r-decodable pool subset computed by the MDPSP algorithms as a function of primer length, for pools with 2 primers, 2 possible extensions per primer, and array probes consisting of all 4^{10} 10-mers (a), respectively all 645,376 13-tokens (b) (averages over 10 test cases).

[8] J. Hirschhorn, P. Sklar, K. Lindblad-Toh, Y.-M. Lim, M. Ruiz-Gutierrez, S. Bolk, B. Langhorst, S. Schaffner, E. Winchester, and E. Lander, “SBE-TAGS: An array-based method for efficient single-nucleotide polymorphism genotyping,” *PNAS*, vol. 97, no. 22, pp. 12 164–12 169, 2000.

[9] M. Morris, D. Shoemaker, R. Davis, and M. Mittmann, “Selecting tag nucleic acids,” *U.S. Patent 6,458,530 B1*, 2002.

[10] Y. Aumann, E. Manisterski, and Z. Yakhini, “Designing optimally mul-

TABLE II

NUMBER OF ARRAYS NEEDED TO COVER 10-50% OF THE REFERENCE SNPS WITH UNAMBIGUOUS PRIMERS OF LENGTH 20.

Chr	#	#	# 10-mer arrays									# 13-token arrays								
ID	Ref.	Extracted	r=1			r=2			r=5			r=1			r=2			r=5		
	SNPs	Pools	10%	25%	50%	10%	25%	50%	10%	25%	50%	10%	25%	50%	10%	25%	50%	10%	25%	50%
1	786058	736850	1	1	3	1	2	4	2	3	7	1	2	4	1	3	6	2	5	12
2	758368	704415	1	1	3	1	2	4	2	3	7	1	2	3	1	3	5	2	5	11
3	647918	587531	1	1	2	1	2	3	1	3	6	1	2	3	1	2	5	2	4	9
4	690063	646534	1	1	3	1	2	3	1	3	6	1	2	3	1	2	5	2	5	9
5	590891	550794	1	1	2	1	2	3	1	3	5	1	2	3	1	2	4	2	4	9
6	791255	742894	1	2	3	1	2	4	2	4	7	1	2	4	1	3	6	2	5	12
7	666932	629089	1	1	2	1	2	3	1	3	6	1	2	3	1	3	5	2	5	10
8	488654	456856	1	1	2	1	1	3	1	2	5	1	1	3	1	2	4	2	4	7
9	465325	441627	1	1	2	1	1	3	1	2	5	1	1	3	1	2	4	2	4	8
10	512165	480614	1	1	2	1	2	3	1	2	5	1	2	3	1	2	4	2	4	8
11	505641	476379	1	1	2	1	2	3	1	2	5	1	2	3	1	2	4	2	4	8
12	474310	443988	1	1	2	1	1	3	1	2	5	1	1	3	1	2	4	2	4	8
13	371187	347921	1	1	2	1	1	2	1	2	4	1	1	2	1	2	3	1	3	6
14	292173	271130	1	1	2	1	1	2	1	2	3	1	1	2	1	1	3	1	2	5
15	277543	258094	1	1	1	1	1	2	1	2	3	1	1	2	1	1	3	1	2	5
16	306530	288652	1	1	2	1	1	2	1	2	3	1	1	2	1	2	3	1	3	6
17	269887	249563	1	1	1	1	1	2	1	2	3	1	1	2	1	2	3	1	3	6
18	268582	250594	1	1	1	1	1	2	1	2	3	1	1	2	1	1	3	1	2	5
19	212057	199221	1	1	1	1	1	2	1	1	3	1	1	2	1	2	3	1	3	6
20	292248	262567	1	1	1	1	1	2	1	2	3	1	1	2	1	2	3	1	3	6
21	148798	138825	1	1	1	1	1	1	1	1	2	1	1	1	1	1	2	1	2	3
22	175939	164632	1	1	1	1	1	1	1	1	2	1	1	2	1	1	3	1	2	5
X	380246	362778	1	1	2	1	1	2	1	2	4	1	1	2	1	2	3	1	3	6
Y	50725	49372	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2

tiplexed SNP genotyping assays," in *Proc. 3rd Workshop on Algorithms in Bioinformatics (WABI)*, 2003, pp. 320–338.

[11] Affymetrix, Inc. (2001) GeneFlex tag array technical note no. 1. [Online]. Available: http://www.affymetrix.com/support/technical/technotes/genflex_technote.pdf

[12] ——. (2005) Custom and application-specific genotyping with the Affymetrix GeneChip MegAllele system. [Online]. Available: http://www.affymetrix.com/support/technical/other/parallele_brochure.pd%f

[13] A. BenDor, T. Hartman, B. Schwikowski, R. Sharan, and Z. Yakhini, "Towards optimally multiplexed applications of universal DNA tag systems," in *Proc. 7th Annual International Conference on Research in Computational Molecular Biology*, 2003, pp. 48–56.

[14] I. Mandoiu, C. Prajescu, and D. Trinca, "Improved tag set design and multiplexing algorithms for universal arrays," *LNCS Transactions on Computational Systems Biology*, vol. II, no. LNBI 3680, pp. 124–137, 2005.

[15] R. Dramanac and R. Crkvenjakov, "DNA sequencing by hybridization," *Yugoslav patent application*, 1987.

[16] P. Pevzner, *Computational Molecular Biology: An Algorithmic Approach*. MIT Press, 2000.

[17] P. Pevzner and R. Lipshutz, "Towards DNA sequencing chips," in *Proc. 19th Int. Conf. on Mathematical Foundations of Computer Science*, 1994, pp. 143–158.

[18] S. Heath and F. Preparata, "Enhanced sequence reconstruction with DNA microarray application," in *Proc. 7th Annual International Conference on Computing and Combinatorics (COCOON)*, 2001, pp. 64–74.

[19] E. Hubbell, "Multiplex sequencing by hybridization," *Journal of Computational Biology*, vol. 8, no. 2, pp. 141–149, 2001.

[20] R. Sharan, J. Gramm, Z. Yakhini, and A. Ben-Dor, "Multiplexing schemes for generic SNP genotyping assays," *Journal of Computational Biology*, vol. 12, no. 5, pp. 514–533, 2005.

[21] N. Tonisson, A. Kurg, E. Lohmussaar, and A. Metspalu, "Arrayed primer extension on the DNA chip - method and application," in *Microarray Biochip Technology*, M. Schena, Ed. Eaton Publishing, 2000, pp. 247–263.

[22] K. Konwar, I. Mandoiu, A. Russell, and A. Shvartsman, "Improved algorithms for multiplex PCR primer set selection with amplification length constraints," in *Proc. 3rd Asia-Pacific Bioinformatics Conference (APBC)*, Y.-P. Phoebe Chen and L. Wong, Eds. London: Imperial College Press, 2005, pp. 41–50.

[23] F. Naef and M. Magnasco, "Solving the riddle of the bright mismatches: Labeling and effective binding in oligonucleotide arrays," in *Physical Review E.*, vol. 68, 2003, pp. 11 906–11 910.

[24] V. Lozin, "On maximum induced matchings in bipartite graphs," in *Infomation Processing Letters*, vol. 81, 2002, pp. 7–11.

[25] W. Duckworth, D. Manlove, and M. Zito, "On the approximability of the maximum induced matching problem," in *Journal of Discrete Algorithms*, vol. 3, 2005, pp. 79–91.

[26] A. Ben-Dor, R. Karp, B. Schwikowski, and Z. Yakhini, "Universal DNA tag systems: a combinatorial design scheme," *Journal of Computational Biology*, vol. 7, no. 3-4, pp. 503–519, 2000.

[27] R. Wallace, J. Shaffer, R. Murphy, J. Bonner, T. Hirose, and K. Itakura, "Hybridization of synthetic oligodeoxyribonucleotides to phi chi 174 DNA: the effect of single base pair mismatch," *Nucleic Acids Res.*, vol. 6, no. 11, pp. 6353–6357, 1979.

[28] N. Patil et al., "Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21," *Science*, vol. 294, pp. 1719–1723, 2001.

[29] S. Angelov, B. Harb, S. Kannan, S. Khanna, J. Kim, and L.-S. Wang, "Genome identification and classification by short oligo arrays," in *Proc. 4th International Workshop on Algorithms in Bioinformatics*, 2004, pp. 400–411.

[30] R. Shen et al., "High-throughput SNP genotyping on universal bead arrays," *Mutation Research*, vol. 573, no. 1-2, pp. 70–82, 2005.

Ion I. Măndoiu's biography appears with this issue's Guest Editorial.

Claudia Prăjescu attended "Gr.C. Moisil" Highschool of Computer Science in Iasi, Romania between 1996 and 2000. She received a B.Sc. degree in Computer Science from the "A.I. Cuza" University of Iasi in 2004, and a M.Sc. degree in Computer Science and Engineering from the University of Connecticut in 2005, with a thesis on multiplexing algorithms for high-throughput genomic based assays. She is currently a Software Developer with Siemens VDO Automotive in Iasi, Romania.