MULTIPLEXING SCHEMES FOR GENERIC SNP GENOTYPING ASSAYS

R. SHARAN
International Computer Science Institute, 1947 Center St., Berkeley CA 94704.
roded@icsi.berkeley.edu.

A. BEN-DOR
Agilent Laboratories. amir_ben-dor@agilent.com.

Z. YAKHINI
Agilent Laboratories and Computer Science Dept., Technion.
zohar_yakhini@agilent.com.

Abstract

A generic genotyping assay utilizes a fixed set of reagents, which is independent of the actual target sample, to determine all present alleles. An example is the interrogation of several amplicons spanning polymorphic sites using an all k-mer array. Due to the high cost associated with a genotyping experiment, it is desirable to design a set of experiments, which maximizes the number of SNPs that can be genotyped in parallel per assay. In this study we investigate algorithmic approaches for optimally multiplexing SNP genotyping using generic assays. We devise a graph theoretic formulation of the problem and use it to derive an approximation algorithm for the problem, and several practical heuristics. We apply our methods to simulated and real data, for evaluating the multiplexing rates afforded by generic techniques. The results on real human data show the practicality of generic approaches for genotyping, allowing, e.g., the genotyping of 5000 SNPs using four all 7-mer arrays.

1 Introduction

Single nucleotide polymorphisms (SNPs) are differences across the population, in a single base, within an otherwise conserved genomic sequence [1]. 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 [2]. Somatic or native SNPs in oncogenes or in related regions can determine cancer susceptibility and are often related to pathogenesis [3,4,5,6]. SNPs in all regions of the genome are useful in studies aimed at finding genomic regions linked to clinical or otherwise significant properties. Such studies are performed by seeking correlations between the inheritance pattern of the target properties and polymorphic genetic variations. Linkage, association and linkage disequilibrium studies are examples of specific methodologies employed for genetic studies [7,8].

Genotyping is a process that determines the variants present in a given sample, over a set of SNPs. In the case of association studies, a population of samples is jointly measured and the frequencies of the different variants need to be inferred. The development of efficient SNP detection, genotyping and measurement techniques is an active research area as they have great clinical, scientific and commercial value.

Most current SNP genotyping techniques [9,10] are problem specific in the sense that at least some of the reagents used in the assay have to be specifically tailored to the set of SNPs under interrogation. *Generic methods* are techniques that defer all problem specific components to the assay planning stage and to the data analysis and result interpretation stage. For example, Sampson et al. [11] present a method that uses natural and mass modified generic mixtures of oligonucleotides, and a target mediated enzymatic reaction, to produce a mixture, the mass-spectrum of which is indicative of the genotype of the sample over a set of sites.

SNP genotyping is time-consuming and may be an expensive procedure. This cost is directly related to the number of assays actually performed. Thus, we are interested in minimizing the number of assays that need to be performed in a given study. Under certain circumstances, genotyping of multiple SNP sites can be performed simultaneously, in a single genotyping assay; a process called *multiplexed genotyping.* Examples include utilizing primer extension and MALDI-TOF mass spectrometry, relying on the natural masses of the extended specifically designed primers [10,12]. Typically, not all SNPs in a set of interest can be genotyped together. Specifically, any given genotyping method imposes a set of constraints regarding which SNPs can be assayed together, and which cannot. Thus, in order to achieve high multiplexing rates, it is necessary to carefully plan the genotyping assays, in order to allow simultaneous genotyping of as many SNPs as possible, on the one hand, while conforming to the constraints, on the other.

In this paper we present methods for achieving high multiplexing rates for a family of generic SNP genotyping techniques. We model all the applications in a unified framework, in which each SNP is assigned a set of features and the multiplexing problem translates to partitioning the SNPs into sets, such that in each set every SNP has a unique feature (Section 2). We give a constant approximation algorithm for the problem which is based on graph coloring, and provide several practical heuristics (Section 3). Finally, we apply our algorithms to simulated data as well as to real human SNP data (Section 4).

2 Generic Genotyping Techniques

Polymerase extension is a widely used technique for interrogating DNA sequences. Typically, all methods based on this technique utilize extension of specifically designed primers, and are not generic. For example, in an Array Polymerase EXtension assay (APEX) [13,9] the target sample is annealed to array bound probes, that are complementary to subsequences upstream the polymorphic site. Four differentially fluorescently labeled terminator nucleotides are used by DNA polymerase in primer extension reaction, extending the array probes. As a result each probe represents a polymorphic site and the fluorescence observed therein indicates the measured genotype there. Note that the array needs to be specifically designed to address the input set of SNPs.

In a generic Polymerase Extension Assay (PEA), the target sample reacts with a generic set of features (e.g., primers). These are extended, or not, depending on the target. A detection step follows, wherein the extended primers are determined, based on their altered properties. Information on the target is obtained by an interpretation process. We provide several examples below. Throughout, two alleles that correspond to the same SNP are called *mates.*

All k-mer Arrays (C_k). Ben-Dor et al. [14] study aspects of a system that uses a generic array design but a specifically designed set of solution primers. A completely generic approach uses an array of all k-mers, denoted C_k, and no specific reagents, to perform the measurement as follows. First assume that a single site is to be genotyped:

1. The target region is PCR amplified.
2. The sequence is hybridized to the C_k array and a polymerase reaction is started, in the presence of single labeled dideoxynucleotides.
3. k-mers that are complementary to non-polymorphic parts of the amplicon will hybridize to the target, get extended and produce fluorescence signals.
4. The hybridization signals obtained for k-mers that span the site, depend on the alleles of this SNP in the genotyped individual.

The genotype of the sample, at the interrogated site, can therefore be determined by analyzing the hybridization signature, provided that there is at least one k-mer for each allele that does not appear in the sequence of its mate.

In a multiplexed assay several targets are jointly interrogated. The set can be jointly interrogated as long as each allele has at least one unique k-mer that does not occur in the sequence of any other allele-pair in the set.

PEA and Native/Tagged Mass-Spectrometry. This process involves the following components [11]:

1. A mixture of primers is applied to the target in the presence of polymerase and all 4 dideoxynucleotides, allowing for single base extension to occur in a specific, target mediated manner.
2. Products (extended primers) are separated, e.g by HPLC.
3. The mixture of extended primers is analyzed by mass-spectrometry.

Under complete stringency assumptions the output mass spectrum will only have peaks at masses that correspond to extended primers that are Watson-Crick complements of some target subsequence. A set of SNPs can be jointly interrogated as long as each of the respective alleles has a corresponding extended primer with a unique mass, different from that potentially arising from any other allele-pair in the set.

A similar genotyping process uses cleavable mass-tags that are attached to the original primers and then cleaved after the separation of extended products. (Here we assume that the number of available distinguishable tags exceeds the number of primers.) The tags, rather than the extended primers are analyzed by mass-spectrometry. The spectrum will have peaks at masses of tags that correspond to primers that are Watson-Crick complements of some target subsequence. Again, a set of SNPs can be jointly interrogated as long as each of the respective alleles has a corresponding extended primer with a unique tag, different from that potentially arising from any other allele-pair in the set.

2.1 Problem Formulation

In any of the embodiments, the target is typically a collection of short PCR amplicons, spanning bi-allelic SNP sites. A SNP allele in a target can be determined if and only if the extension event, for one of the k-mers spanning this site and corresponding to this allele, can be uniquely detected under the assay conditions. This requirement can be abstracted as follows: Associate with each target sequence a list of features at which it registers, e.g., all complementary k-mers, the masses corresponding to all complementary extended primers in the mixture, etc. This is the set of features potentially *activated* by the given target sequence. Furthermore, the set of activated features can be partitioned into *informative* ones, spanning the polymorphic nucleotide, and *common* ones, being all features activated by the amplicon corresponding to both alleles expected at this site. A set of allele-pairs is *assignable* if each allele in the set has an informative feature that is not potentially activated by any other allele in the set.

Assume we are given a set of target sequences, each containing a bi-allelic polymorphic site. To genotype this set of SNPs we need to partition them into assignable subsets. This partition constitutes a *multiplexing scheme*. We seek a multiplexing scheme under which the number of assignable subsets in the partition is minimum. W.l.o.g., we shall assume that for every given target sequence, each of the two alleles of the corresponding SNP has at least one informative feature that is not shared by its mate.

The objective of the multiplexing scheme can be modeled in two ways. Both formulations reflect the fact that when a specific site is genotyped, both its alleles may activate features (indeed, this will be the case if the sample is heterozygous) and there is no easy way to separate these sets of features one from the other. In the first formulation we seek a partition of the SNPs into a minimum number of assignable subsets. The basic units here are *allele-pairs* (corresponding to SNPs). In the second variant we seek a partition of the alleles into a minimum number of assignable subsets. The basic units here are *single alleles*, dropping the constraint that two mates should be put in the same subset in a partition. Solutions to the first variant have the advantage that they require a smaller number of PCR reactions, compared to the second variant. However, when studying the multiplexing problem in isolation, the latter formulation is the more general one.

3 Algorithmic Approaches

In this section we provide theoretical analysis and practical heuristics for the multiplexing problem.

3.1 An Approximation Scheme

We present an approximation algorithm for the multiplexing problem under the single-allele variant. First, we devise a graph-theoretic formulation of the problem: We view the input allele sequences and list of features as a bipartite graph $G(U, V, E, F)$, where U is the set of alleles and V is the set of features. We put an edge $(u, v) \in E$, connecting an allele $u \in U$ and a feature $v \in V$, if v is an informative feature of u. We put an edge $(u, v) \in F$, if v is a common feature of u or if v is an informative feature of u's mate. We call this graph the *Alleles-Features (AF) graph*. We call E the set of *informative* edges in G. Note that every allele with a sequence of length l has k' (at most k) informative edges incident to it, corresponding to the k-mers spanning the polymorphic site; and at most $(l - k + 1)$ non-informative edges in F, corresponding to $(l - k + 1 - k')$ k-mers that do not involve the polymorphic site and k' additional k-mers that constitute the informative features of the allele's mate.

Consider a set of alleles $X \subseteq U$ of cardinality t. The set X is called *assignable* if there exists an induced matching over X consisting of only informative edges. That is, there exists a set $R \subseteq E$ of t informative edges that form a matching between X and a set of t features. In addition, the matching R is *induced*, that is, no two edges in R have a third edge adjacent to both of them. Clearly, if a set of alleles is assignable, it can be assayed together. We define the following two decision problems:

Maximum Assignable Set (MAS): Given an AF graph G and an integer k, is there an assignable set of size at least k?

Minimum Assignable Cover (MAC): Given an AF graph G and an integer k, is there a set of k assignable subsets that together cover its allele set?

Observe that for a given set of alleles X, one can test in linear time whether X is assignable, by checking if each allele in X has an informative edge to a unique feature. By considering only the informative edges, **MAS** is equivalent to the maximum induced matching problem on an appropriate bipartite graph. For general bipartite graphs, the maximum induced matching problem is known to be NP-hard[15]. However, not all bipartite graphs can be realized as AF graphs. The complexity of **MAS** on AF graphs is currently open.

We now show a lower bound for the **MAC** problem. A *good matching* in G is a matching which consists of informative edges only. Clearly, any assignable subset of alleles corresponds to a good induced matching in G between the alleles and their unique features. If we restrict attention to informative edges only, and drop the requirement that the matching should be an induced one, our problem variant can be stated as follows:

Minimum Matching Cover (MMC): Given a bipartite graph $G'(U, V, E)$, find a minimum number of matchings that cover U.

Given an instance $G(U, V, E, F)$ of **MAC**, the cardinality of an optimum solution to **MMC** on $G'(U, V, E)$ is a lower bound on the cardinality of any solution to **MAC** on G and, in particular, a lower bound on the optimum **MAC** solution. We can compute this lower bound in polynomial time using the algorithm of Aumann et al. for **MMC**[12].

The following theorem states our approximation result for **MAC**.

Theorem 1 *Let $G(U, V, E, F)$ be an instance of* **MAC** *with allele sequences of length l. Then there is a $(2l + 1)$-approximation algorithm to* **MAC** *on G.*

Proof: We find the approximate solution in two stages. The reader is referred to the example in Figure 3.1 for further explanation and intuition about the algorithm. First, we construct the graph $G'(U, V, E)$ by removing the non-informative edges from G. We find a minimum matching cover $E_1, E_2, \ldots, E_r$ of G' using the algorithm of Aumann et al.[12]. Next, we show below that for each matching E_i, its set of alleles can be partitioned into at most $(2l+1)$ assignable subsets. Overall, the cardinality of our solution is bounded above by $opt_{MMC}(G') \cdot (2l+1) \leq opt_{MAC}(G) \cdot (2l+1)$.

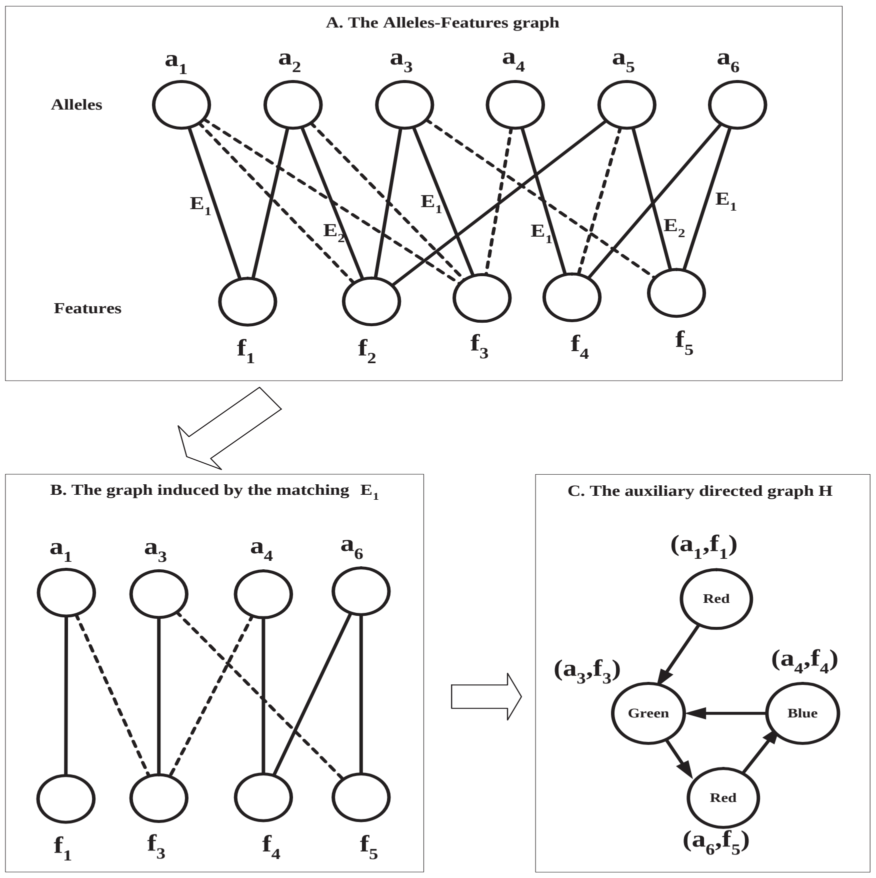

Figure 1: An example demonstrating the approximation algorithm for MAC. **A.** The Alleles-Features graph. Informative edges are solid, common edges are dashed. E_1 and E_2 represent one possible optimal matching cover. **B.** The graph induced by the matching E_1 in part A. **C.** The auxiliary directed graph H constructed from the graph in part B. The outdegree ≤ 1, and thus H can be colored with 3 colors. Each color class corresponds to a set of independent edges in E_1. For example, the red color corresponds to the edges $(a_1, f_1), (a_6, f_5)$. These edges corresponds to the assignable allele set $\{a_1, a_6\}$.

It remains to show how to partition the alleles included in a matching E_i into at most $(2l+1)$ assignable sets. To this end we use the coloring approach of Ben-Dor et al. [14]: We build an auxiliary directed graph H, whose vertices correspond to the edges in the matching E_i. We direct an edge in H from (u,v) to (u',v') if $(u,v') \in E_i$. Note that since each allele has at most $l+1$ incident edges in the original graph G, by the construction of H, each of its vertices has outdegree at most l. Therefore, H can be colored using smallest-last ordering (SLO) coloring [16] by at most $2l+1$ colors. Each color class represent an independent set of vertices, which correspond to an independent set of informative edges. Thus, each color class induces an assignable set, and together they cover the alleles of E_i. ■

3.2 Practical Heuristic Approaches

In this section we propose two greedy heuristic procedures for the multiplexing problem. Both approaches work on allele-pairs as well as on single alleles. We describe them only in their allele-pair version, but apply both variants in the sequel.

The first heuristic is called *minimal partition (MP)*. We allocate one SNP at a time, inserting it into the subset that best accommodates it: This is a subset which after adding the allele-pair remains assignable and has the smallest number of activated features. We start a new subset only when the target cannot be accommodated in an existing subset. In the following we denote by $\sigma(S)$ the number of activated features in a set of allele-pairs S. We let $q_1,\ldots,q_n$ be the input allele-pairs. The algorithm is given in Figure 2.

```
Randomly order the allele-pairs $q_1,\ldots,q_n$.
$Q_1 = \emptyset$, $k = 1$.
For $i = 1 \ldots n$ consider the pair $q_i$:
    Find an index $j_0$ s.t. $Q_{j_0} \cup \{q_i\}$ is assignable
    and $\sigma(Q_{j_0} \cup \{q_i\})$ is minimal.
    If such $j_0$ exists then $Q_{j_0} = Q_{j_0} \cup \{q_i\}$.
    Else $Q_{k+1} = \{q_i\}$ and $k = k+1$.
```

Figure 2: The minimal partition (MP) algorithm.

The second heuristic is called *maximal set (MS)*. We attempt to construct the largest assignable subset of SNPs. When this set cannot be extended anymore, we iteratively call the process on the remaining SNPs. The algorithm is given in Figure 3.

```
$Q_1 = \emptyset$, $U = \{q_1, \ldots, q_n\}$, $k = 1$.
While $U \neq \emptyset$
        Find a pair $q \in U$ s.t. $Q_k \cup \{q\}$ is assignable
        and $\sigma(Q_k \cup \{q\})$ is minimal.
        If such $q$ exists then $Q_k = Q_k \cup \{q\}$, $U = U \setminus \{q\}$.
        Else $Q_{k+1} = \emptyset$ and $k = k + 1$.
```

Figure 3: The Maximal Set (MS) algorithm.

The complexity of the MP algorithm is $O(nr)$ for a solution of cardinality r. The complexity of the MS algorithm is $O(n^2)$, which is higher than the former since, typically, $r \ll n$.

4 Results

In this section we report on the performance of the two algorithmic schemes, MP and MS, on simulated and real SNP data.

The synthetic data was generated as follows: We generated at random 41-long sequences for varying number of SNPs (between 1000 and 5000). For each sequence we chose at random two distinct nucleotides, representing two alleles, to occupy the 21-st base of the sequence. We used as features all k-mers of an array, where k ranged from 6 to 8. The results of applying both algorithms to the data, when using the allele-pair version, are summarized in Table 1. The total running time of one simulation was less than a minute on a single processor. Notably, the maximal set algorithm outperforms the minimal partition algorithm in all simulations.

Next, we applied the algorithms in their single-allele version to the synthetic data. The results are summarized in Table 2. Again the MS heuristic outperforms the MP heuristic. Overall, the results are compa-

Table 1: A comparison between Algorithms MP and MS, in their allele-pair version, on simulated data for different parameter combinations. Each entry contains the solution's cardinality, averaged over 10 runs. All respective standard deviations were smaller than 1.

	Array					
SNPs	C_6		C_7		C_8	
	MP	MS	MP	MS	MP	MS
1000	7	5	3	2.2	2	1.3
2000	11.7	7	4.5	3.1	2.1	2
5000	24	12.7	8.1	5	3.5	2.5

Table 2: Comparison between Algorithms MP and MS, in their single-allele version, on simulated data for different parameter combinations. Each entry contains the solution's cardinality, averaged over 10 runs. All respective standard deviations were smaller than 1.

SNPs	Array					
	C_6		C_7		C_8	
	MP	MS	MP	MS	MP	MS
1000	7.1	4.9	3	2.3	2	1.3
2000	10.9	7	4.7	3	2.3	2
5000	22.5	12	8	5	3.6	2.6

rable to those obtained for allele-pairs. This is a result of the assignment criterion employed by the algorithms, which is strongly biased towards assigning mates to the same set in the solution.

We complemented our analyses by designing multiplexing schemes for 5000 SNPs taken from human chromosome 21. We extracted from the public SNP database [17] 41-long sequences flanking the first 5000 reference SNPs of chromosome 21 (with the polymorphic site at the middle position). We then applied our algorithms to the data, again using as features all k-mers of an array, where k ranged from 6 to 8. The best results were obtained using the MS heuristic with pairs: 7 assays for C_8, 9 assays for C_7 and 19 assays for C_6.

When comparing to the above simulation results, we observed that the solutions on real data had higher cardinality than the corresponding solutions on simulated data. Looking more closely at the results, we observed that the real data solutions contained several large sets covering most of the SNPs and some small sets containing only few SNPs each. Specifically, if one wishes to cover at least 95% of the SNPs then the following numbers of assays are required: 2 assays for C_8, 4 assays for C_7 and 11 assays for C_6. We could further see that most of the SNPs that were included in small sets had very degenerate sequences, often consisting of repeats of a single nucleotide around the polymorphic site.

5 Concluding Remarks

In this paper we studied the problems arising in designing multiplexing schemes for SNP genotyping using generic assays. We devised a graph theoretic formulation for the multiplexing problem and used it to find a constant approximation algorithm for the problem. We also suggested two practical heuristics to approach the problem. We applied our algorithms to simulated and real SNP data. The results on real data show the practicality of generic approaches for genotyping, allowing, e.g., the genotyping of about 5000 SNPs using four C_7 arrays.

It is important to note that the procedure described in the paper assumes full stringency of the C_k array measurements, and that current assays do not support this assumption. The multiplexing methods that we presented here can be modified to handle non-perfect hybridization under weaker assumptions. For example, one could define a measure of similarity between k-mers, and solve the multiplexing problem assuming that a k-mer may hybridize to its complement or to any other sufficiently similar k-mer.

While focusing here on the application of our algorithmic approaches to SNP genotyping, multiplexing problems arise in other domains, e.g., primer design, gene expression measurements, etc. Our algorithmic approaches could be applicable to multiplexing problems in other domains that have graph theoretic models similar to the ones presented here.

Acknowledgments

The authors thank Jeff Sampson, Anya Tsalenko and Bo Curry of Agilent Technologies for discussion and early testing of the heuristics. The authors also thank an anonymous PSB'04 referee for helpful comments. RS was supported by a Fulbright grant.

References

1. 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.
2. 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.
3. S. Venitt. Mechanisms of carcinogenesis and individual susceptibility to cancer. *Clin. Chem.*, 40(7.2):1421–5, July 1994.
4. S. Venitt. Mechanisms of spontaneous human cancers. *Environ. Health Perspect.*, 104(3):633–7, May 1996.
5. 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.
6. Y. Watanabe, A. Fujiyama, Y. Ichiba, M. Hattori, et al. 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.
7. J. Ott. *Analysis of Human Genetic Linkage.* Johns Hopkins University Press, 1991.

8. N. J. Risch. Searching for genetic determinants in the new millennium. *Nature*, 405(6788):847–56, 2000.
9. 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.
10. P. Ross et al. High level multiplex genotyping by MALDI-TOF mass-spectometry. *Nature Biotechnology*, 16:1347–1351, 1998.
11. J.R. Sampson et al. Method and mixture reagents for analyzing the nucleotide sequence of nucleic acids by mass spectrometry. US patent 6,218,118., 2001.
12. Y. Aumann, E. Manisterski, and Z. Yakhini. Designing optimally multiplexed SNP genotyping assays. In *Proc. of Workshop on Algorithms in Bioinformatics (WABI)*, 2003. To appear.
13. `http://www.asperbio.com/APEX.htm`.
14. A. Ben-Dor, T. Hartman, B. Schwikowski, R. Sharan, and Z. Yakhini. Towards optimally multiplexed applications of universal DNA tag systems. In *Proc. of seventh annual conference on research in computational molecular biology (RECOMB)*, pages 48–56, 2003.
15. K. Cameron. Induced matchings. *Discrete Applied Mathematics*, 24:97–102, 1989.
16. D. Matula and L. Beck. Smallest-last ordering and clustering and graph coloring algorithms. *Journal of the ACM*, 30:417–427, 1983.
17. `http://www.ncbi.nlm.nih.gov/SNP`.