Available online at www.sciencedirect.com
ScienceDirect
Theoretical Computer Science 390 (2008) 120–125
Theoretical Computer Science
www.elsevier.com/locate/tcs

Note

On the complexity of non-unique probe selection

Yongxi Cheng[a,*], Ker-I Ko[b], Weili Wu[c]

[a] *Department of Computer Science, Tsinghua University, Beijing 100084, China*
[b] *Department of Computer Science, State University of New York at Stony Brook, Stony Brook, NY 11794-4400, USA*
[c] *Department of Computer Science, University of Texas at Dallas, Richardson, TX 75083, USA*

Received 3 June 2007; received in revised form 1 October 2007; accepted 6 October 2007

Communicated by D.-Z. Du

Abstract

We investigate the computational complexity of some basic problems regarding non-unique probe selection using separable matrices. In particular, we prove that the minimal $\bar{d}$-separable matrix problem is *DP*-complete, and the $\bar{d}$-separable submatrix with reserved rows problem, which is a generalization of the decision version of the minimum $\bar{d}$-separable submatrix problem, is Σ_2^P-complete.

Keywords: Non-unique probe selection; Separable matrices; *DP*-complete; Σ_2^P-complete

1. Introduction

Given a collection of n targets and a sample S containing at most d of these targets, and a collection of m probes each of which hybridizes to a subset of the given targets, we want to select a subset of probes such that we can identify all targets in S by observing the hybridization reactions between the selected probes and S. For each probe p, there is a hybridization reaction between p and S if S contains at least one target that hybridizes with p; otherwise there is no hybridization reaction. The above probe selection problem has been extensively studied recently [5,1,9,10,13] due to its important applications, particularly in molecular biology. For example, one application of this identification problem is in identifying viruses (targets) from a blood sample. We establish the presence or absence of the viruses by observing the hybridization reactions between the blood sample and some probes; here, each probe is a short oligonucleotide of size 8–25 that can hybridize with one or more of the viruses.

A probe is called *unique* if it hybridizes with only one target; otherwise it is called *non-unique*. Identifying targets using unique probes is straightforward. However, in situations where the targets have a high degree of similarity, for instance when identifying closely related virus subtypes, finding unique probes for all targets is difficult. In [11], Schliep, Torney and Rahmann proposed a group testing method using non-unique probes to identify targets in a given

* Corresponding author. Tel.: +86 10 51537918.
E-mail addresses: cyx@mails.tsinghua.edu.cn (Y. Cheng), keriko@cs.sunysb.edu (K.-I Ko), weiliwu@utdallas.edu (W. Wu).

0304-3975/$ - see front matter
doi:10.1016/j.tcs.2007.10.014

sample. Since each non-unique probe can hybridize with more than one target, the identification problem becomes more complicated. One important issue is how to select a subset from the given non-unique probes so that we can decode the hybridization results, i.e., determine the presence or absence of targets in the sample S. Also, the number of selected probes is exactly the number of hybridization experiments required, so we hope to select as few probes as possible to reduce the experimental cost. In [11,6], two heuristics using greedy and linear programming based techniques respectively are proposed for choosing a suitable subset of non-unique probes. In this paper, we investigate the computational complexity of some basic problems in non-unique probe selection, in the context of the theory of *NP*-completeness (see Chapter 10 in [2–4]).

2. Preliminaries

The non-unique probe selection problem can be formulated as follows. We are given a collection of n targets $t_1, t_2, \ldots, t_n$, and a collection of m non-unique probes $p_1, p_2, \ldots, p_m$. A sample S is known to contain at most d of the n targets. The probe–target hybridizations can be represented by an $m \times n$ 0–1 matrix M. $M_{i,j} = 1$ indicates that probe p_i hybridizes with target t_j, and $M_{i,j} = 0$ indicates otherwise. The subset of probes selected corresponds to a subset of rows in M, which forms a submatrix H of M with the same number of columns. The results for hybridization between the selected probes and S also can be represented as a 0–1 vector V. $V_i = 1$ indicates that there is a hybridization reaction between p_i and S, i.e., p_i hybridizes with at least one target in S, and $V_i = 0$ indicates otherwise. If there is no error in the hybridization experiments, then V is equal to the union of the columns of H that correspond to the targets in S. Here, the union of a subset of columns is simply the Boolean sum of these column vectors. In order to identify all targets in S, the submatrix H should satisfy that all unions of up to d columns in H are different; in other words H should be $\bar{d}$*-separable*. Also, as mentioned above, we hope to minimize the number of rows in H.

A matrix H is said to be $\bar{d}$-separable if all unions of up to d columns in H are different. However, the following equivalent definition is more useful in our proofs. Let H be a $t \times n$ Boolean matrix. For each $i \in \{1, 2, \ldots, t\}$, define $H_i = \{j \mid 1 \le j \le n, H_{i,j} = 1\}$. For any subset S of $\{1, 2, \ldots, n\}$ and any $i \in \{1, 2, \ldots, t\}$, we write $H_i(S) = 1$ if $H_i \cap S \ne \emptyset$, and $H_i(S) = 0$ otherwise. We say two sets $S_1, S_2 \subseteq \{1, 2, \ldots, n\}$ can be *separated* by H if there exists an integer i, $1 \le i \le t$, such that $H_i(S_1) \ne H_i(S_2)$. We say H is $\bar{d}$-separable if for any two different subsets S_1, S_2 of $\{1, 2, \ldots, n\}$, with $|S_1| \le d$ and $|S_2| \le d$, S_1 and S_2 can be separated by H.

3. Complexity of the minimal $\bar{d}$-separable matrix

In non-unique probe selection, one natural problem of interest is determining whether a submatrix H chosen is $\bar{d}$-separable and minimal. By minimal we mean that the removal of any row from H will make it no longer $\bar{d}$-separable. The problem can be formulated as follows.

> MIN-SEPARABILITY (MINIMAL SEPARABILITY): Given a $t \times n$ Boolean matrix H and an integer $d \le n$, determine whether it is true that (a) H is $\bar{d}$-separable, and (b) for any submatrix Q of H of size $(t-1) \times n$, Q is not $\bar{d}$-separable.

For a given binary matrix H and a positive integer d, the problem of determining whether H is $\bar{d}$-separable is known to be *coNP*-complete ([2], Theorem 10.2.1). In this section, we will show that MIN-SEPARABILITY is *DP*-complete. The class *DP* is the collection of sets A which are the intersection of a set $X \in NP$ and a set $Y \in coNP$. The notion of *DP*-completeness has been used to characterize the complexity of the "exact-solution" version of many *NP*-complete problems. For instance, the exact traveling salesman problem, which asks, for a given edge-weighted complete graph G and a constant K, whether the minimum weight of a traveling salesman tour of the graph G is equal to K, is *DP*-complete (see [7], Theorem 17.2). In addition, the "critical" versions of some *NP*-complete problems are also known to be *DP*-complete. For instance, the following problem is the critical version of the 3-satisfiability problem, and has been shown to be *DP*-complete by Papadimitriou and Wolfe [8]:

> MIN-3-UNSAT: Given a 3-CNF Boolean formula φ which consists of clauses $C_1, C_2, \ldots, C_m$, determine whether it is true that (a) φ is not satisfiable, and (b) for any j, $1 \le j \le m$, the formula φ_j that consists of all clauses C_ℓ, $\ell \in \{1, 2, \ldots, m\} - \{j\}$, is satisfiable.

Although most exact-solution versions of *NP*-complete problems have been shown to be *DP*-complete, many critical versions are not known to be *DP*-complete. The problem MIN-SEPARABILITY may be viewed as a critical version of the $\bar{d}$-separability problem. We will prove it to be *DP*-complete by constructing a reduction from MIN-3-UNSAT.

Theorem 1. MIN-SEPARABILITY *is DP-complete.*

Proof. Recall that $DP = \{X \cap Y \mid X \in NP, Y \in coNP\}$. A problem A is *DP-complete* if $A \in DP$ and, for all $B \in DP$, $B \le_m^P A$. For convenience, we write, for any $t \times n$ matrix H, $\widetilde{H}_j$ to denote the $(t-1) \times n$ submatrix of H with the jth row removed.

First, to see that MIN-SEPARABILITY $\in DP$, let $X = \{(H, d) \mid H$ is a $t \times n$ Boolean matrix, $1 \le d \le n$, $(\forall j, 1 \le j \le t)$ $\widetilde{H}_j$ is not $\bar{d}$-separable$\}$, and $Y = \{(H, d) \mid H$ is a $t \times n$ Boolean matrix, $1 \le d \le n$, H is $\bar{d}$-separable$\}$. It is clear that MIN-SEPARABILITY $= X \cap Y$. It is also not hard to see that $X \in NP$ and $Y \in coNP$. In particular, to see that $X \in NP$, we note that $(H, d) \in X$ if and only if there exist $2t$ subsets $S_{j,1}, S_{j,2}$ of $\{1, 2, \ldots, n\}$, for $j \in \{1, 2, \ldots, t\}$, such that, for each j, $H_k(S_{j,1}) = H_k(S_{j,2})$ for all $k \in \{1, 2, \ldots, t\} - \{j\}$.

Next, we describe a reduction from MIN-3-UNSAT to MIN-SEPARABILITY. Let φ be a 3-CNF Boolean formula which consists of m clauses $C_1, C_2, \ldots, C_m$, over n variables $x_1, x_2, \ldots, x_n$. For each $j \in \{1, 2, \ldots, m\}$, let φ_j denote the Boolean formula that consists of all clauses C_ℓ for $\ell \in \{1, 2, \ldots, m\} - \{j\}$. From φ, we will construct a $(3n + m + 1) \times (2n + 2)$ Boolean matrix H, and define $d = n + 1$. For convenience, we denote the columns of H by $X = \{x_i, \bar{x}_i \mid 1 \le i \le n\} \cup \{y, z\}$; and denote the rows of H by $T = \{x_i, \bar{x}_i, u_i \mid 1 \le i \le n\} \cup \{y\} \cup \{C_j \mid 1 \le j \le m\}$. We define H by defining each row of H:

(1) For each $1 \le i \le n$, let $H_{x_i} = \{x_i\}$, $H_{\bar{x}_i} = \{\bar{x}_i\}$, and $H_{u_i} = \{x_i, \bar{x}_i, z\}$.
(2) $H_y = \{y\}$.
(3) For each $1 \le j \le m$, let $H_{C_j} = \{x_i \mid x_i \in C_j\} \cup \{\bar{x}_i \mid \bar{x}_i \in C_j\} \cup \{y, z\}$ (so that $|H_{C_j}| = 5$).

To prove the correctness of the reduction, we first verify that, if φ is not satisfiable, then H is $\bar{d}$-separable. To see this, let S_1 and S_2 be two subsets of X, each of size $\le n + 1$.

Case 1. $S_1 - \{z\} \ne S_2 - \{z\}$. Then, there exists $v \in X - \{z\}$ such that $v \in S_1 \Delta S_2$. Then, $H_v(S_1) \ne H_v(S_2)$.

Case 2. $S_1 - \{z\} = S_2 - \{z\}$. Then, it must be true that $S_1 \Delta S_2 = \{z\}$. Without loss of generality, assume $S_2 = S_1 \cup \{z\}$. Note that $|S_2| \le n + 1$ implies $|S_1| \le n$.

Subcase 2.1. There exists an integer i such that $|S_1 \cap \{x_i, \bar{x}_i\}| \ne 1$. First, if $|S_1 \cap \{x_i, \bar{x}_i\}| = 0$ for some i, then $H_{u_i}(S_1) = 0$ and $H_{u_i}(S_2) = 1$ (because $z \in S_2$). Next, if $|S_1 \cap \{x_i, \bar{x}_i\}| = 2$ for some i, then we must have $|S_1 \cap \{x_k, \bar{x}_k\}| = 0$ for some k, because $|S_1| \le n$. Then, again $H_{u_k}(S_1) = 0 \ne 1 = H_{u_k}(S_2)$.

Subcase 2.2. $|S_1 \cap \{x_i, \bar{x}_i\}| = 1$ for all $i \in \{1, 2, \ldots, n\}$. We note that, in this case, $y \notin S_1$. Define a Boolean assignment $\tau : \{x_1, x_2, \ldots, x_n\} \to \{\text{TRUE}, \text{FALSE}\}$ by $\tau(x_i) = \text{TRUE}$ if and only if $x_i \in S_1$. Since φ is not satisfiable, there exists a clause C_j that is not satisfied by τ. This means that $C_j \cap S_1 = \emptyset$, and so $H_{C_j}(S_1) = 0$. However, $H_{C_j}(S_2) = 1$ since $z \in S_2$.

The above completes the proof that H is $\bar{d}$-separable.

Next, we show that if φ_j is satisfiable for all $j = 1, 2, \ldots, m$, then $\widetilde{H}_v$ is not $\bar{d}$-separable for all $v \in T$. First, for $v \in X - \{z\}$, let $S_1 = \{z\}$ and $S_2 = \{v, z\}$. Then, we can see that for all rows $w \in X - \{z, v\}$, $H_w(S_1) = 0 = H_w(S_2)$. Also, for all other rows $w \in T - X$, $H_w(S_1) = H_w(S_2) = 1$ since $z \in H_w$. So, S_1 and S_2 are not separable by $\widetilde{H}_v$.

Next, consider the case $v = u_i$ for some $i \in \{1, 2, \ldots, n\}$. Let $S_1 = \{x_k \mid 1 \le k \le n, k \ne i\} \cup \{y\}$ and $S_2 = S_1 \cup \{z\}$. It is clear that $|S_1| = n$ and $|S_2| = n + 1$. We claim that S_1 and S_2 are not separable by $\widetilde{H}_{u_i}$.

To prove the claim, we note that the rows $H_{x_k}, H_{\bar{x}_k}$, for $1 \le k \le n$, and row H_y cannot separate S_1 from S_2, since $S_1 - \{z\} = S_2 - \{z\}$. Also, rows $H_{u_k}(S_1) = H_{u_k}(S_2) = 1$, for all $k \in \{1, 2, \ldots, n\} - \{i\}$, because $|S_1 \cap \{x_k, \bar{x}_k\}| = 1$ if $k \ne i$. In addition, for any $j = 1, 2, \ldots, m$, we have $H_{C_j}(S_1) = 1 = H_{C_j}(S_2)$, since $y \in S_1$. It follows that $\widetilde{H}_{u_i}$ cannot separate S_1 from S_2.

Finally, consider the case $v = C_j$ for some $j \in \{1, 2, \ldots, m\}$. We note that φ_j is satisfiable. So, there is a Boolean assignment $\tau : \{x_1, x_2, \ldots, x_n\} \to \{\text{TRUE}, \text{FALSE}\}$ satisfying all clauses C_ℓ, except C_j. Define $S_1 = \{x_i \mid \tau(x_i) = \text{TRUE}\} \cup \{\bar{x}_i \mid \tau(x_i) = \text{FALSE}\}$, and $S_2 = S_1 \cup \{z\}$. Then, like with the argument for the case $v = u_i$, we can verify that $H_w(S_1) = H_w(S_2)$ for $w \in X - \{z\}$, and for $w \in \{u_i \mid 1 \le i \le n\}$. In addition, for any clause C_ℓ, with $\ell \ne j$, C_ℓ is satisfied by τ. It follows that $C_\ell \cap S_1 \ne \emptyset$ and $H_{C_\ell}(S_1) = 1 = H_{C_\ell}(S_2)$. This completes the proof that $\widetilde{H}_v$ is not $\bar{d}$-separable, for all $v \in T$.

Conversely, we show that if $\varphi \notin$ MIN-3-UNSAT, then $(H, n+1) \notin$ MIN-SEPARABILITY. First, we consider the case where φ is a satisfiable formula. Let $\tau : \{x_1, x_2, \ldots, x_n\} \to \{\text{TRUE}, \text{FALSE}\}$ be a Boolean assignment satisfying φ. Define $S_1 = \{x_i \mid \tau(x_i) = \text{TRUE}\} \cup \{\bar{x}_i \mid \tau(x_i) = \text{FALSE}\}$, and $S_2 = S_1 \cup \{z\}$. Then, like in the earlier proof, we can verify that H cannot separate S_1 from S_2. In particular, $H_{C_j}(S_1) = 1$ for all $j \in \{1, 2, \ldots, m\}$, because τ satisfies C_j and so $C_j \cap S_1 \neq \emptyset$. Thus, $(H, n+1) \notin$ MIN-SEPARABILITY.

Next, assume that there exists an integer $j \in \{1, 2, \ldots, m\}$ such that φ_j is not satisfiable. We claim that $\widetilde{H}_{C_j}$ is $\bar{d}$-separable. The proof of the claim is similar to the proof for the statement that if φ is not satisfiable then H is $\bar{d}$-separable.

Case 1. $S_1 - \{z\} \neq S_2 - \{z\}$. Then, there exists $v \in X - \{z\}$ such that $v \in S_1 \Delta S_2$. So, $H_v(S_1) \neq H_v(S_2)$.

Case 2. $S_1 - \{z\} = S_2 - \{z\}$. Then, it must be true that $S_1 \Delta S_2 = \{z\}$, and we may assume $S_2 = S_1 \cup \{z\}$. We must have $|S_2| \leq n+1$ and $|S_1| \leq n$.

Subcase 2.1. There exists an integer i such that $|S_1 \cap \{x_i, \bar{x}_i\}| \neq 1$. Like in the earlier proof, if $|S_1 \cap \{x_i, \bar{x}_i\}| = 0$ for some $i = 1, 2, \ldots, n$, then we can use H_{u_i} to separate S_1 from S_2. If $|S_1 \cap \{x_i, \bar{x}_i\}| = 2$ for some $i = 1, 2, \ldots, n$, then $|S_1 \cap \{x_k, \bar{x}_k\}| = 0$ for some k, and again H_{u_k} separates S_1 from S_2.

Subcase 2.2. $|S_1 \cap \{x_i, \bar{x}_i\}| = 1$ for all $i \in \{1, 2, \ldots, n\}$. Then, since $|S_1| \leq n$, $y \notin S_1$. Define a Boolean assignment $\tau : \{x_1, x_2, \ldots, x_n\} \to \{\text{TRUE}, \text{FALSE}\}$ by $\tau(x_i) = \text{TRUE}$ if and only if $x_i \in S_1$. Since φ_j is not satisfiable, there exists a clause C_ℓ, $\ell \neq j$, such that $\tau(C_\ell) = \text{FALSE}$. This means that $C_\ell \cap S_1 = \emptyset$, and so $H_{C_\ell}(S_1) = 0$. However, $H_{C_\ell}(S_2) = 1$ since $z \in S_2$. So, H_{C_ℓ} separates S_1 from S_2. This completes the proof that $\widetilde{H}_{C_j}$ is $\bar{d}$-separable, and hence $(H, n+1) \notin$ MIN-SEPARABILITY. □

4. Minimum $\bar{d}$-separable submatrix

A more important problem in non-unique probe selection is finding a minimum subset of probes that can identify up to d targets in a given sample. In the matrix representation, the problem can be formulated as the following: Given a binary matrix M and a positive integer d, find a minimum $\bar{d}$-separable submatrix of M with the same number of columns (problem MIN-$\bar{d}$-SS in [2], Chapter 10).

For $d = 1$, MIN-$\bar{d}$-SS has been proved to be *NP*-hard ([2], Theorem 10.3.2), by modifying a reduction used in the proof of the *NP*-completeness of the problem MINIMUM-TEST-SETS in [4]. For a fixed $d > 1$, MIN-$\bar{d}$-SS is believed to be *NP*-hard; however up to now no formal proof has been known. We consider the decision version of MIN-$\bar{d}$-SS.

> $\bar{d}$-SS ($\bar{d}$-SEPARABLE SUBMATRIX): Given a $t \times n$ Boolean matrix M and two integers $d, k > 0$, determine whether there is a $k \times n$ submatrix H of M that is $\bar{d}$-separable.

Recall that Σ_2^P is the complexity class of problems that are solvable in nondeterministic polynomial time with the help of an *NP*-complete set as an oracle. For instance, the following problem SAT_2 is Σ_2^P-complete ([3], Theorem 3.13): Given a Boolean formula φ over two disjoint sets X and Y of variables, determine whether there exists an assignment to variables in X so that the resulting formula (over variables in Y) is a tautology. It is easy to see that $\bar{d}$-SS is in Σ_2^P. We conjecture that it is actually Σ_2^P-complete. Here, we consider a similar problem that is a little more general than $\bar{d}$-SS, and prove that it is Σ_2^P-complete.

> $\bar{d}$-SSRR ($\bar{d}$-SEPARABLE SUBMATRIX WITH RESERVED ROWS): Given a $t \times n$ Boolean matrix M and three integers $d > 0$, $s, k \geq 0$, determine whether there is a $\bar{d}$-separable $(s+k) \times n$ submatrix H of M that contains the first s rows of M and k rows from the remaining $t - s$ bottom rows of M.

Let φ be a Boolean formula; an *implicant* of φ is a conjunction C of literals that implies φ. The following problem is proved to be Σ_2^P-complete by Umans [12].

> SHORTEST IMPLICANT CORE: Given a DNF formula $\varphi = T_1 + T_2 + \cdots + T_m$, and an integer p, determine whether φ has an implicant C that consists of p literals from the last term T_m.

By a reduction from SHORTEST IMPLICANT CORE, we can obtain the following result.

Theorem 2. $\bar{d}$-SSRR *is* Σ_2^P*-complete.*

Proof. The problem $\bar{d}$-SSRR can be solved by a nondeterministic machine that guesses an $(s+k)\times n$ submatrix H of M which contains the first s rows of M, and then determines whether H is $\bar{d}$-separable. We note that the problem of determining whether a given matrix H is $\bar{d}$-separable is in *coNP*. Thus, $\bar{d}$-SSRR $\in \Sigma_2^P$.

Next, we prove that $\bar{d}$-SSRR is Σ_2^P-complete by constructing a polynomial-time reduction from SHORTEST IMPLICANT CORE to it. To define the reduction, let (φ, p) be an instance of the problem SHORTEST IMPLICANT CORE, i.e., let $\varphi = T_1 + T_2 + \cdots + T_m$ be a DNF formula over n variables $x_1, x_2, \ldots, x_n$, and let p be an integer > 0. We note that each term T_j, $1 \le j \le m$, of φ is a conjunction of some literals. We also write T_j to denote the set of these literals. Assume that the last term T_m of φ has q literals $\ell_1, \ell_2, \ldots, \ell_q$. We define a $(3n+m+q)\times(2n+1)$ Boolean matrix M as follows:

(1) Let the $2n+1$ columns of M be $X = \{x_1, \bar{x}_1, x_2, \bar{x}_2, \ldots, x_n, \bar{x}_n, z\}$, and the $3n+m+q$ rows of M be $T = \{x_i, \bar{x}_i, u_i \mid 1 \le i \le n\} \cup \{t_j \mid 1 \le j \le m\} \cup \{c_j \mid 1 \le j \le q\}$.
(2) For $i = 1, 2, \ldots, n$, $M_{x_i} = \{x_i\}$, $M_{\bar{x}_i} = \{\bar{x}_i\}$, and $M_{u_i} = \{x_i, \bar{x}_i, z\}$.
(3) For $j = 1, 2, \ldots, m$, $M_{t_j} = \{x_i \mid \bar{x}_i \in T_j\} \cup \{\bar{x}_i \mid x_i \in T_j\} \cup \{z\}$. (Note that $M_{t_j} \cap T_j = \emptyset$.)
(4) The bottom q rows of M are $M_{c_j} = \{\ell_j, z\}$, for $j = 1, 2, \ldots, q$.

We let $d = n+1$, $s = 3n+m$, $k = p$, and consider the instance (M, d, s, k) for the problem $\bar{d}$-SSRR.

First assume that φ has an implicant C of size p that is a subset of T_m. Let H be the submatrix of M that consists of the first $s = 3n+m$ rows plus the $k = p$ rows M_{c_j} for which $\ell_j \in C$. We claim that H is $\bar{d}$-separable. That is, for any subsets S_1 and S_2 of $\{x_1, \bar{x_2}, \ldots, x_n, \bar{x}_n, z\}$ of size $\le d$, there exists a row in H that separates them.

Case 1. $S_1 - \{z\} \ne S_2 - \{z\}$. Then, there exists $v \in X - \{z\}$ such that $v \in S_1 \Delta S_2$. Then, $M_v(S_1) \ne M_v(S_2)$, and so H separates S_1 from S_2.

Case 2. $S_1 - \{z\} = S_2 - \{z\}$. Then, it must be true that $S_1 \Delta S_2 = \{z\}$. Without loss of generality, assume $S_2 = S_1 \cup \{z\}$. Note that $|S_2| \le n+1$ implies $|S_1| \le n$.

Subcase 2.1. There exists an integer i such that $|S_1 \cap \{x_i, \bar{x}_i\}| \ne 1$. First, if $|S_1 \cap \{x_i, \bar{x}_i\}| = 0$ for some i, then $M_{u_i}(S_1) = 0$ and $M_{u_i}(S_2) = 1$ (because $z \in S_2$). Next, if $|S_1 \cap \{x_i, \bar{x}_i\}| = 2$ for some i, then we must have $|S_1 \cap \{x_k, \bar{x}_k\}| = 0$ for some k, because $|S_1| \le n$. Then, again $M_{u_k}(S_1) = 0 \ne 1 = M_{u_k}(S_2)$. It follows that H separates S_1 from S_2.

Subcase 2.2. $|S_1 \cap \{x_i, \bar{x}_i\}| = 1$ for all $i \in \{1, 2, \ldots, n\}$. Define a Boolean assignment $\tau : \{x_1, x_2, \ldots, x_n\} \to \{\text{TRUE}, \text{FALSE}\}$ by $\tau(x_i) = \text{TRUE}$ if and only if $x_i \in S_1$. We further divide this into two subcases:

Subcase 2.2.1. τ satisfies the conjunction C. Since C is an implicant of $\varphi = T_1 + T_2 + \cdots + T_m$, τ must satisfy some T_j, $1 \le j \le m$. Thus, we have $T_j \subseteq S_1$: for any $x_i \in T_j$, $\tau(x_i) = \text{TRUE}$ and so $x_i \in S_1$; and for any $\bar{x}_i \in T_j$, $\tau(x_i) = \text{FALSE}$ and so $\bar{x}_i \in S_1$. It follows that $M_{t_j}(S_1) = 0$ since $M_{t_j} \cap T_j = \emptyset$. On the other hand, $M_{t_j}(S_2) = 1$ since $z \in M_{t_j} \cap S_2$. So, M_{t_j}, and hence H, separates S_1 from S_2.

Subcase 2.2.2. τ does not satisfy C. Then, for some literal $\ell_j \in C$, $\tau(\ell_j) = 0$. Thus, $\ell_j \notin S_1$, and $M_{c_j}(S_1) = 0$. On the other hand, $M_{c_j}(S_2) = 1$ since $z \in M_{c_j}$. Thus, M_{c_j}, which is a row in H, separates S_1 from S_2.

Conversely, assume that H is a $(3n+m+k)\times(2n+1)$ submatrix of M that contains the first $3n+m$ rows of M and is $\bar{d}$-separable. Let C be the conjunction of literals ℓ_j for which M_{c_j} is a row in H. Then, obviously, $|C| = k$. We claim that C is an implicant of φ.

Let $\tau : \{x_1, x_2, \ldots, x_n\} \to \{\text{TRUE}, \text{FALSE}\}$ be a Boolean assignment that satisfies C. We need to show that τ satisfies φ. Let $S_1 = \{x_i \mid \tau(x_i) = \text{TRUE}\} \cup \{\bar{x}_i \mid \tau(x_i) = \text{FALSE}\}$ and $S_2 = S_1 \cup \{z\}$. Then, S_1 and S_2 can be separated by some row in H. Since $S_2 = S_1 \cup \{z\}$, we know that they are not separable by a row M_{x_i} or $M_{\bar{x}_i}$, for any $i = 1, 2, \ldots, n$. In addition, since $|S_1 \cap \{x_i, \bar{x}_i\}| = 1$ for all $i = 1, 2, \ldots, n$, we know that they cannot be separated by row M_{u_i}, for any $i = 1, 2, \ldots, n$. Furthermore, we note that for any literal $\ell_j \in C$, $\tau(\ell_j) = 1$ and so $\ell_j \in S_1$ and $M_{c_j}(S_1) = M_{c_j}(S_2) = 1$. Thus, S_1 and S_2 cannot be separated by any row M_{c_j} of H.

Therefore, S_1 and S_2 must be separable by a row M_{t_j}, for some $j = 1, 2, \ldots, m$. That is, $M_{t_j}(S_1) = 0 \ne 1 = M_{t_j}(S_2)$. Since M_{t_j} contains the complements of the literals in T_j, we see that $T_j \subseteq S_1$. It follows that τ satisfies the term T_j, and hence φ. □

5. Conclusion

In the previous sections, we investigated the computational complexity of problems related to non-unique probe selection. We have shown that the problem of verifying the minimality of a $\bar{d}$-separable matrix is *DP*-complete, and hence is intractable, unless $DP = P$. For the problem of finding a minimum $\bar{d}$-separable submatrix, we conjecture that it is Σ_2^P-complete and, hence, is even more difficult than the minimal $\bar{d}$-separability problem. To support this conjecture, we showed that the problem $\bar{d}$-SSRR, which is a little more general than the minimum $\bar{d}$-separable submatrix problem, is Σ_2^P-complete. The complexity of the original problem MIN-$\bar{d}$-SS remains open.

Acknowledgements

The first author was supported in part by the National Natural Science Foundation of China under grant No. 60553001 and the National Basic Research Program of China under grant No. 2007CB807900, 2007CB807901. The work of the second author was partially supported by the National Science Foundation under grant No. 0430124.

References

[1] J. Borneman, M. Chrobak, G. Della Vedova, A. Figueroa, T. Jiang, Probe selection algorithms with applications in the analysis of microbial communities, Bioinformatics 17 (Suppl.) (2001) S39–S48.
[2] D.-Z. Du, F.K. Hwang, Pooling Designs and Nonadaptive Group Testing: Important Tools for DNA Sequencing, World Scientific, 2006.
[3] D.-Z. Du, K.-I Ko, Theory of Computational Complexity, Wiley & Sons, New York, 2000.
[4] M.R. Garey, D.S. Johnson, Computers and Intractability: A Guide to the Theory of *NP*-Completeness, Freeman, San Francisco, 1979.
[5] R. Herwig, A.O. Schmitt, M. Steinfath, J. O'Brien, H. Seidel, S. Meier-Ewert, H. Lehrach, U. Radelof, Information theoretical probe selection for hybridisation experiments, Bioinformatics 16 (2000) 890–898.
[6] G.W. Klau, S. Rahmann, A. Schliep, M. Vingron, K. Reinert, Optimal robust non-unique probe selection using integer linear programming, Bioinformatics 20 (2004) i186–i193.
[7] C.H. Papadimitriou, Computational Complexity, Addison-Wesley, New York, 1994.
[8] C.H. Papadimitriou, D. Wolfe, The complexity of facets resolved, J. Comput. Systems Sci. 37 (1988) 2–13.
[9] S. Rahmann, Rapid large-scale oligonucleotide selection for microarrays, in: Proceedings of the 1st IEEE Computer Society Conference on Bioinformatics, CSB' 02, 2002, pp. 54–63.
[10] S. Rahmann, Fast and sensitive probe selection for DNA chips using jumps in matching statistics, in: Proceedings of the 2nd IEEE Computer Society Bioinformatics Conference, CSB' 03, 2003, pp. 57–64.
[11] A. Schliep, D.C. Torney, S. Rahmann, Group testing with DNA chips: Generating designs and decoding experiments, in: Proceedings of the 2nd IEEE Computer Society Bioinformatics Conference, CSB' 03, 2003, pp. 84–93.
[12] C. Umans, The minimum equivalent DNF problem and shortest implicants, in: Proceedings of 39th IEEE Symposium on Foundation of Computer Science, 1998, pp. 556–563.
[13] X. Wang, B. Seed, Selection of oligonucleotide probes for protein coding sequences, Bioinformatics 19 (2003) 796–802.