Non-Unique Probe Selection and Group Testing

Feng Wang * Hongwei David Du † Xiaohua Jia † Ping Deng ‡
Weili Wu ‡ David MacCallum *

Abstract

A minization problem arosen from the study of the non-unique probe selection with group testing technique is as follows: Give a binary matrix, find a d-disjunct submatrix with the minimum number of rows and the same number of columns. We show that when every probe hybridizes at most two viruses, i.e., every row contains at most two 1s, this minimization is still MAX SNP-complete, but has a polynomial-time approximation with performance $1 + 2/(d + 1)$. This approximation is constructed based on an interesting result that the above minimization is polynomial-time solvable when every probe hybridizes exactly two viruses.

*Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN 55455, USA. Email: fwang@cs.umn.edu.

†Department of Computer Science, City University of Hong Kong, China. Email: {hongwei,jia}@cs.cityu.edu.hk.

‡Department of Computer Science, University of Texas at Dallas, Richardson, Tx 75083, USA. Email: {pxd010100,weiliwu}@utdallas.edu. Support in part by National Science Foundation under grant ACI-0305567.

1 Introduction

Currently, non-unique probe selection is a hot topic in computational molecular biology. A probe is a short oligonucleotide of size 8-25, used for identifying viruses in a biological sample through hybridization. When each probe hybridizes to a unique virus, identification is straightforward. However, unique probes are very hard to be obtained, especially for virus subtypes which are closely related, although temperature and salt concentration are very helpful to reduce the number of viruses bybridized by a probe. Schilep, Torney and Rahman [9] introduced a method to use non-unique probes with group testing techniques. They consider each virus as an item and for each probe, the set all viruses hybridized to it as a pool. Based on classical theory of nonadaptive group testing, when the incidence matrix between items and pool is $\bar{d}$-separable, the test-outcome can identify up to d viruses in biological sample.

For n items with t pools, the incidence matix is an $t \times n$ binary matrix with rows labeled by pools and columns labeled by items and cell (i, j) contains 1-entry if and only if the ith pool contains item j. A binary matrix is d-separable if all boolean sums of at most d columns are distinct. If each column is seen as a set of rows corresponding to 1-entries in the column, then the boolean sum can be seen as a union of columns which is a classic statement in the study of group testing.

When a probe is hybridized by some virus in a biological sample, we say that the test-outcome is positive; otherwise, the test-outcome is negative. Test-outcomes for all probes can be written as a column vector which is exactly the union of columns corresponding

viruses contained in the biological sample, where 1 denotes a positive outcome and 0 denotes a negative outcome. Therefore, the definition of d-separable matrix means that different sets of at most d viruses receive differen t test-outcomes.

Torney and Rahman [9] suggested a method consists of three steps:

Step 1. Collect a large set of non-unique probes.

Step 2. From this large set of probes, find a minimum subset of probes to identify up to d viruses.

Step 3. Decode the presence or absence of viruses in the given biological sample from test-outcome.

The minimization problem in Step 2 can be described as follows:

> MIN-$\bar{d}$-SS (Minimum $\bar{d}$-Separable Submatrix). Given a binary matrix M, find a minimum $\bar{d}$-separable submatrix with the same number of columns.

For any fixed d, MIN-$\bar{d}$-SS is NP-hard[3]. They suggested a greedy algorithm which adds probe one by one until the incidence matrix with considered viruses form a $\bar{d}$-separable matrix.

Since it is hard to decode the test-outcome from a $\bar{d}$-separable matrix[3], Thai *et al.* [10] considered to use a d-disjunct matrix instead. A binary matrix is *d-disjunct* if any union of d columns cannot contains the $(d+1)$th column. Decoding test-outcome from a d-disjunct matrix is very easy[3]. This introduces another minimization problem:

MIN-d-DS (Minimum d-Disjunct Submatrix). Given a binary matrix M, find a minimum $\bar{d}$-disjunct submatrix with the same number of columns.

For $d = 1$, MIN-d-SS is exactly the well-known minimum test cover problem [5] (also called the minimum test set problem [2, 1, 7] or the minimum test collection [6]). The minimum test cover problem has a greedy approximation with performance $1 + 2 \ln n$ where n is the number of items [1, 2].

Often, the pool size cannot be too large since selected candidate probes is usually nearly unique. This motivated the study of above minimization problems with bounded pool size. For instance, let us consider the case that every pool has size at most 2. Halldórsson *et al.* [6] and De Bontridder *et al.* [2] proved that in this case, MIN-1-SS is still APX-hard, which means that there is no polynomial-time approximation scheme for it unless NP=P. They also showed that MIN-1-SS in this case has a polynomial-time approximation with performance ratio $7/6 + \varepsilon$ for any fixed $\varepsilon > 0$.

In this paper, we will present some interesting results that while MIN-d-DS is polynomial-time solvable in the case that all pools have size two, it is MAX SNP-complete in the case that all pools have size at most two, but there is a polynomial-time approximation with performance ratio $1 + 2/(d + 1)$ for $d \geq 1$.

2 Main Results

First, let us indicate that

Theorem 1 *There exist greedy approximations with performance ratio* $1 + (d+1)\ln n$ *for* MIN-d-DS *and* $1 + 2d\ln(n+1)$ *for* MIN-$\bar{d}$-SS.

Proof. The proof is quite easy. For example, let us consider MIN-d-DS. Consider the collection $\mathcal{S}$ of all possible pairs (C, D) of one column C and a subset D of d columns. Clearly $|\mathcal{S}| < n^{d+1}$. A row is said to *cover* such a pair (C, D) if at this row, the entry of column C is 1 and all entries of columns in D are 0. Now, we choose rows one by one to maximize the total number of pairs newly covered by the row. This is a special case of the set cover problem. It is well-known that the greedy algorithm for the set cover has performance ratio $1 + \ln|\mathcal{S}| < 1 + (d+1)\ln n$. □

Our main interest in this paper is to study MIN-d-DS in the case that all pools have size at most two. For simplicity of notation, we denote by MIN-d-DS-2 the MIN-d-DS in this special case. The following lemma plays an important role.

Lemma 2 *Consider a collection C of pools of size at most 2. Let G be the graph with all items as vertices and all pools of size 2 as edges. Then C gives a d-disjunct matrix if and only if every item not in a singleton pool has degree at least $d+1$ in G.*

Proof. Suppose there exists an item a_0 not in any singleton pool of C and its degree in G is at most d. Let (a_0, a_1), (a_0, a_2), ..., $a_0, a_k)$ $(k \leq d)$ be all edges of G at a_0. Then a_0 is contained in the union of columns with label $a_1, a_2, ..., a_k$. Therefore, $\mathcal{C}$ does not form a d-disjunct matrix.

Conversely, suppose every item is either in a singleton pool or of degree at least $d+1$. Then in the former case, the singleton pool does not contain any other item, and in the latter case, for any d other items $a_1, ..., a_d$, there is a pool of size two containing a_0 but not anyone of $a_1, ..., a_d$. Hence, $\mathcal{C}$ forms a d-disjunct matrix. □

As a consequence of above lemma, we have

Theorem 3 MIN-d-DS *is polynomial-time solvable in the case that all given pools have size exactly 2*

Proof. Let H be the graph with all items as vertices and all given pools as edges. By Lemma 2, MIN-d-DS is equivalent to find a subgraph G, with minimum number of edges, such that every vertex has degree at least $d+1$ in G. It is equivalent to maximize the number of edges in $H-G$ such that every vertex v has degree at most $d_H(v)-d-1$ in $H-G$ where $d_H(v)$ denotes the degree of v in H. The latter maximization problem has been known to be polynomial-time solvable for a long-time. □

Theorem 4 *Min-d-DS-2 for $d \geq 2$ is MAX SNP-complete.*

Proof. Cnsider a well known MAX SNP-complete problem [4]:

> VC-CUBIC: Given a cubic graph G (a graph is *cubic* if every vertex has degree exactly three), find the minimum vertex-cover of G.

We show a L-reduction from VC-CUBIC to MIN-2-DS-2.

Suppose $G = (V, E)$ is an input of VC-CUBIC. For each edge (u, v) of G, we add break (u, v) into two edge (u, w_0) and (w_0, v), and add other four vertices w_1, w_2, w_3, w_4 and seven edges (w_0, w_1), (w_0, w_2), (w_1, w_3), (w_1, w_4), (w_2, w_3), (w_2, w_4), (w_3, w_4) (Fig.1). The result

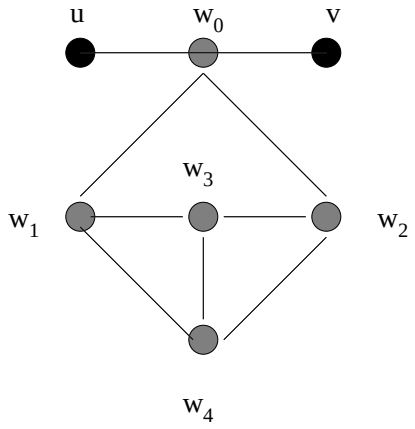

Figure 1: Construction from G to G'.

graph is denote by $G' = (V', E')$. Denote $E'' = \{(u, w_0), (w_0, v) \mid (u, v) \in E\}$.

Let $\mathcal{C} = E' \cup \{\{u\} \mid u \in V\}$ be an input of MIN-1-DS. By Lemma 2, every feasible solution of MIN-1-DS must contain all edges in $E' - E''$. Since w_0 is not contained by a singleton pool, every feasible solution must contains either (u, w_0) or (w_0, v). Consider a minimum solution $\mathcal{C}^*$ with smallest number of pools of size 2. Then $\mathcal{C}^*$ does not contain both (u, w_0) and (w_0, v). In fact, if it contains both, then we may replace (u, w_0) by $\{u\}$ to reduce the number of pools of size 2. It follows that either u or v is contained in a singleton pool of $\mathcal{C}^*$. Therefore, all items in singleton pools of $\mathcal{C}^*$ form a vertex-cover of G.

Conversely, if X is a vertex-cover of G, then $C(X) = \{\{u\} \mid u \in X\} \cup \{e \in E' \mid e \text{ is incident to a vertex in } X\}$ is a feasible solution of MIN-2-DS-2 with input $\mathcal{C}$. Therefore,

$C(X)$ is a minimum vertex-cover of G if and only if $C(X)$ is a minimum solution of MIN-2-DS.

Suppose X^* is a minimum vertex-cover of G. Note that $|X^*| \geq |E|/3$. Hence

$$|C(X^*)| = 8|E| + |X^*| \leq 25|X^*|.$$

Moreover, suppose X is a vertex-cover of G. Then $C(X)$ is a feasible solution of MIN-2-DS-2 satisfying

$$|C(X) - C(X^*)| = ||X| - |X^*||.$$

Therefore, VC-CUBIC is L-reducible to MIN-2-DS-2.

Since MIN-2-DS-2 is a special case of MIN-d-DS-2 for any $d > 2$, we can easily construct L-reduction from MIN-2-DS-2 to MIN-d-DS-2 for any $d > 2$. By Theorem 1, MIN-d-DS has polynomial-time approximation with constant performance ratio. Therefore, MIN-d-DS-2 for $d \geq 2$ is MAX-SNP-complete. □

Corollary 5 *There exists a positive number r such that* MIN-d-DS-2 *has no polynomial-time approximation with performance ratio r unless NP=P.*

Next, we present a better approximation for MIN-d-DS-2.

Lemma 6 *Let s be the number of given singleton pools. Then any feasible solution of* MIN-d-DS-2 *contains at least $s + (n - s)(d + 1)/2$ pools.*

Proof. Suppose $\mathcal{C}$ is a feasible solution of MIN-d-DS. By Lemma 2, every item is either in a singleton pool or in at least $d + 1$ pools of size 2. Suppose $\mathcal{C}$ contains s singleton pools.

Then $\mathcal{C}$ contains at least $s + (n-s)(d+1)/2$ pools. □

Now, we describe an approximation algorithm with two steps.

Step 1. Compute a minimum solution of the following polynomial-time solvable problem: Let G be the graph with all items as vertices and all given pools of size 2 as edges. Find a subgraph H, with minimum number of edges, such that every item not in a singleton pool has degree at least $d+1$.

Step 2. Suppose H is a minimum solution obtained in Step 1. Choose all singleton pools at vertices with degree less than $d+1$ in H. All edges of H and chosen singleton pools form a feasible solution of MIN-d-DS-2.

Theorem 7 *The feasible solution obtained in the above algorithm is a polynomial-time approximation with performance ratio* $1 + 2/(d+1)$.

Proof. Suppose H contains m edges and k vertices of degree at least $d+1$. Suppose an optimal solution containing s^* singletons and m^* pools of size 2. Then $m < m^*$ and $(n-k) - s^* < 2m^*/(d+1)$. Hence,

$$(n-k) + m < s^* + m^* + 2m^*/(d+1) < (s^* + m^*)(1 + 2/(d+1)).$$

□

References

[1] P. Berman, B. Dasgupta and M.-Y. Kao, Tight approximability results for test set problems in bioinformatics, *Journal of Computer and System Sciences* 71 (2005) 145-162.

[2] K.M.J. De Bontridder, B.V. Halldórsson, M.M. Halldórsson, C.A.J. Hurkens, J.K. Lenstra, R. Ravi and L. Stougie, Approximation algorithms for the test cover problem, *Mathematical Programming*, 98 (2003) 477-491.

[3] D.-Z. Du and F.K. Hwang, *Pooling Designs and Nonadaptive Group Testing*, (World Scientific, 2006).

[4] D.-Z. Du and K.-I Ko, *Theory of Computational Complexity*, (John Wiley, 2000).

[5] M.R. Garey and D.S. Johnson, *Computers and Intractability*, (W.H. Freeman, San Francisco, 1979).

[6] B.V. Halldórsson, M.M. Halldórsson and R. Ravi, On the approximability of the minimum test collection problem, *Lecture Notes in Computer Science*, 2161 (2001) 158-169.

[7] R. M. Karp, R. Stougton and K. Y. Yeung, Algorithms for choosing differential gene expression experiments, *Proceedings of the Third Annual International Conference on Computational Molecular Biology*, 1999, pp. 208-217.

[8] G. Klau, S. Rahmann, A. Schliep, M. Vingron, and K. Reinert, Optimal robust non-unique probe selection using integer linear programming, *Bioinformatics*, 20 (2004) I186-I193.

[9] A. Schliep, D. C. Torney and S. Rahmann, Group testing with DNA chips: generating designs and decoding experiments, *Proceedings of the 2nd IEEE Computer Society Bioinformatics Conference*, 2003.

[10] M. Thai, P. Deng, W. Wu and T. Znati, Approximation algorithm of nonunique probe selection for biological target identification, manuscript, 2006.