The First International Symposium on Optimization and Systems Biology (OSB'07)
Beijing, China, August 8–10, 2007
Copyright © 2007 ORSC & APORC，pp. 1–4

Non-unique Probe Selection with Group Testing

Ping Deng[1] Feng Wang[2] Ding-Zhu Du[3,*]

[1] Department of Computer Science, University of Texas at Dallas, Richardson, TX 75083, USA.
[2] Department of Computer Science, University of Minnesota, Minneapolis, MN 55455, USA.
[3] School of Science, Xi'an Jiaotong University, Xi'an, P.R. China.

Abstract Non-unique probe selection is an important research problem in computational molecular biology. One of approaches to study it is group testing. A minization problem arosen from such study 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 give a mini survey on its computational complexityshow and approximation algorithms.

Every person has an ID and different persons have different IDs. Therefore, it is easy to identify a person with ID. However, if different persons may have the same ID and a person may have more than one IDs, it would be a problem to identify a person with IDs. In computational molecular biology, there is a similar problem. A probe is a short oligonucleotide of size 8-25, which plays a role of ID for identifying a virus 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. Given a biological sample and a set of probes each of which may hybridize not only one viruses, how to select probes to identify viruses in the biological sample. This problem is called the *non-unique probe selection*, which is a hot research topic currently in computational molecular biology.

If the biological sample contains only one virus, selected probes should satisfy the condition that different viruses hybridize different sets of probes. In general, if the biological sample contains at most d viruses, selected probes should satisfy the condition that different sets of at most d viruses should hybridize different sets of probes. Schilep, Torney and Rahman [9] 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, the above condition means that incidence matrix between items and pool is $\bar{d}$-separable; in such a case, the test-outcome can identify up to d viruses in biological sample.

[*] Suppose in part by National Natural Science Foundation of China.

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.

The non-unique probe selection can be solved in the following steps [9]:

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]. Since it is hard to decode the test-outcome from a $\bar{d}$-separable matrix[3], Thai *et al.* [11] 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] or the minimum test collection [6]). Minimum test cover problem has a greedy approximation with performance $1 + 2\ln n$ where n is the number of items [2]. This suggests us to study greedy approximations for MIN-d-SS, MIN-$\bar{d}$-SS and MIN-d-DS.

Actually, it is not hard to obtain greedy approximations with performance ratio $1 + 2d\ln n$ for MIN-d-SS, $1 + (d+1)\ln n$ for MIN-d-DS and $1 + 2d\ln(n+1)$ for MIN-$\bar{d}$-SS. For example, let us consider MIN-d-DS. Consider the collection $\mathscr{S}$ of all possible pairs (C, D) of one column C and a subset D of d columns. Clearly

$|\mathscr{S}| < n^{d+1}$. 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 $1+\ln|\mathscr{S}| < 1+(d+1)\ln n$.

This greedy algorithm works well only for small d because its running time is $O(n^{d+1})$. When d is large, it is too slow. Therefore, we must look for other smart ways. Schilep, Torney and Rahman [9] proposed greedy algorithm which adds probe one by one until the incidence matrix with considered viruses form a $\bar{d}$-separable matrix. This doesn't work for large d, neither. In fact, if d is not bounded, then testing whether a binary matrix is d-separable, or $\bar{d}$-separable, or d-disjunct is co-NP-complete. There exist other methods [8] in the literature, which work well for small d. However, no efficient method has been found to produce good solutions for larger d.

When error-tolerance is considered, the design of greedy approximation becomes more interesting. Indeed, when d-disjunct is replaced by $(d;z)$-disjunct, we meet a set z-multiple cover problem, that is, each element should be covered z times.

In some applications, the pool size cannot be too big due to the sensitivity of tests. For example, UNH suggested in ADS testing, each pool should not contain more than five blood samples. When the pool size is bounded, the problem becomes easier. For instance, let us consider the case that every pool has size at most 2 so that all pools of size 2 together with items form a graph G where pools are edges and item are vertices. 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 $\varepsilon > 0$.

A surprising result was showed by Wang *et al.* [10] that a subgraph H of G represents a d-disjunct matrix if and only if every vertex in H has degree at least $d+1$ and hence finding such an H with minimum number of edges is polynomial-time solvable. What is about the case that all pools have size 3? Wang *et al.* proved that in this case MIN-d-DS is still NP-hard. However, there may exist approximations with better performance.

Actually, design of nonadaptive group testing is a special case of the non-unique probe selection. In this case, every kind of probes exists, that is, for any subset of viruses, there exists a probe to hybridize exactly only viruses in this subset. Therefore, a near-optimal group testing is a constant-bounded approximation for non-unique probe selection in this special case and any good approximation solution for the non-unique probe selection can also be used to find a good design for nonadaptive group testing. Conversely, any new discovery for good design of nonadaptive group testing may also provide a hint to motivate some idea for design of good approximations of the non-unique probe selection. For example, we may get some idea to design random algorithms for the non-unique probe selection by some observation on random pooling designs [3]. So far, the best-known design of nonadaptive group testing is within a factor of $O(\log d)$ from the lower bound and the best-known approximation for the non-unique probe selection is within a factor of $O(\log n)$ from optimal solution. We intend to integrate results in these two research directions and find a new research points through observing interactions between these two research directions.

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] Feng Wang, Hongwei Du, Xiaohua Jia and Ping Deng, Non-unique probe selection and group testing, *Theoretical Computer Science*, to appear.

[11] M. Thai, P. Deng, W. Wu and T. Znati, Efficient algorithms for non-unique probes selection using d-disjunct matrix, muscript, 2006.