
Protein-Protein Interaction and Group Testing in Bipartite Graphs

Yingshu Li^{*}, My T. Thai[†], Zhen Liu[‡]
and Weili Wu[‡]

^{*}Department of Computer Science, Georgia State University, P.O. Box 3994, Atlanta, GA 30302. [†]Department of Computer Science, University of Minnesota, Twin Cities, 200 Union Street S.E., Minneapolis, MN 55455. [‡]Department of Computer Science, University of Texas at Dallas, Richardson, TX 75083. E-mail: yili@cs.umn.edu
mythai@cs.umn.edu
zhenliu@utdallas.edu
weiliwu@utdallas.edu

Abstract: The interactions between bait proteins and prey proteins are often critical in many biological processes, such as the formation of macromolecular complexes and the transduction of signals in biological pathways. Thus, identifying all protein-protein interactions is an important task in those processes, which can be formulated as a group testing problem in bipartite graphs. In this paper, we take the advantages of the characteristics of bipartite graphs and present two nonadaptive algorithms for this problem. Furthermore, we illustrate a generalization of our solution in a more general case.

Keywords: Protein-protein Interaction, Group Testing, Bipartite Graphs, Nonadaptive Algorithms.

Reference to this paper should be made as follows: Li, Y., Thai, M. T., Liu, Z. and Wu, W. (2005) 'Protein-Protein Interaction and Group Testing in Bipartite Graphs', *Int. J. Bioinformatics Research and Applications*, Vol. *x*, No. *x*, pp.xxx-xxx.

Biographical notes: Yingshu Li received her Ph.D. degree in computer science from the University of Minnesota, Twin Cities, in 2005. She is an assistant professor in the Department of Computer Science at the Georgia State University, Atlanta. Her research interests include Wireless networks, Optimization algorithm design and Computational Biology. She is a member of the IEEE.

My T. Thai is a PhD candidate in the Department of Computer Science and Engineering at the University of Minnesota, Twin Cities. She received the B.S. degree in Computer Science and the B.S. degree in Mathematics from Iowa State University in 1999. Her current research interests include Wireless Networks, Biological Networks, Algorithms and Combinatorics.

Zhen Liu received the B.E. of Information Engineering at WTUSM in 1992, and the M.S. degree of Mapping and Remote Sensing from Chi-

nese Academy of Sciences in 1992, and M.S. on Computer Science at University of Texas at Dallas in 1999. Zhen Liu is a Ph.D. candidate in the department of computer Sciences at University of Texas at Dallas.

Weili Wu received her MS and PhD degrees in computer science from University of Minnesota, in 1998 and 2002 respectively. She is currently an assistant professor and a lab director of the database research lab at the Department of Computer Science and Engineering, the University of Texas at Dallas. Her research interest is mainly in database systems, especially in spatial database with applications in geographic information systems and bioinformatics, distributed database in internet system, and wireless database systems with connection to wireless communication. She has published more than 30 research papers in various prestigious journals and conferences such as IEEE Transaction on Multimedia, Theoretical Computer Science, Journal of Complexity, Discrete Mathematics, Discrete Applied Mathematics, ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, SIAM Conference on Data Mining, UCGIS Summer Assembly, International Conference on Computer Science and Informatics. She is an author of the textbook *Mathematical Theory of Optimization* and an editor of the research monograph *Clustering and Information Retrieval*. She is a member of the IEEE Computer Society.

1 Introduction

Recently, many applications of group testing have been found in molecular biology, such as DNA library screening [6, 7, 16], physical mapping, contig sequencing, and gene detection [3]. Motivated by those applications, some new models for group testing have been proposed. Especially, the group testing in a complex model has been studied extensively [9, 10, 14, 15]. The group testing in a bipartite graph is a special case of the group testing in a complex model. It plays an important role in the identification of protein-to-protein interactions [13]. The interactions between bait proteins and prey proteins are often critical in many biological processes, such as the formation of macromolecular complexes and the transduction of signals in biological pathways. Thus, identifying all protein-to-protein interactions is an important task in those processes.

In the formulation of protein-to-protein interaction, all bait proteins are represented by the vertices which form a set A , all prey proteins are represented by the vertices which form another set B , and the interaction between a bait protein and a prey protein is represented by an edge connecting a vertex in A that represents the bait protein and a vertex in B that represents the prey protein. An edge is *positive* if there exists an interaction between its two endpoints, otherwise, the edge is *negative*. A test is conducted on a *pool* consisting of a subset C of A and a subset D of B . The test outcome is *positive* if there is an interaction between a vertex in C and a vertex in D , *i.e.*, the edge connecting these two vertices is positive, otherwise, the test outcome is *negative*. A pool is said to be *positive* if the test outcome on the pool is positive, and *negative* otherwise. This model is called the *group testing in a complete bipartite graph $K_{A,B}$* .

Usually, group testing can save a lot of effort compared with individual testing.

For example, the Life Science Division of Los Alamos National Laboratories in 1998 [11] was facing 220,000 clones in DNA library screening. Testing those clones individually requires 220,000 tests. However, only 376 tests were conducted with group testing.

The technology of group testing was initiated from Wasserman-type blood test in World War II. Since then, many research works have been published in the literature [2, 3]. An algorithm for group testing is *nonadaptive* if all the tests are arranged in a single round, that is, no information on the test outcomes is available for determining the pool of another test. It has been very well-known that compared with sequential group testing, the nonadaptive group testing usually takes a shorter time with a little more number of tests. Therefore, for applications in molecular biology, nonadaptive group testing is promoted due to the shorter time consumption of each test.

An algorithm for nonadaptive group testing can be represented by a binary incidence matrix. Its columns are labelled with all the vertices and its rows are labelled with all the pools. The cell (i, j) contains an 1-entry if and only if the i th pool contains the j th vertex. The binary incidence matrix M of a nonadaptive algorithm for group testing in a bipartite graph H is $d(H)$ -disjunct if for any $d + 1$ edges e_0, e_1, \dots, e_d of H , there exists a row indicating that a pool contains e_0 , but not e_1, \dots, e_d . A $d(H)$ -disjunct matrix can identify all the positive edges in a sample with up to d positive edges in a very simple way, and an edge is negative if and only if it is contained in a negative pool.

Several constructions of nonadaptive algorithms for group testing in a complex model have been proposed [3]. Especially, Du, Hwang, Wu and Znati [4] designed a general construction of $d(H)$ -disjunct matrix for any hyper-graph H . However, we found that if H is a bipartite graph, we can take the advantage of the special characteristics of a bipartite graph to reduce the number of the tests. In this paper, we present two constructions of $d(H)$ -disjunct matrices for a bipartite graph H as well as a generalization of our solution for H in a more general case.

2 The First Construction

Our first construction is based on the existence of d -disjunct matrices. A $t \times n$ binary matrix is said to be d -disjunct if for any $d + 1$ columns C_0, C_1, \dots, C_d , there exists a row where C_0 has an 1-entry and all C_1, \dots, C_d have 0-entries. There exist many ways to construct d -disjunct matrices in the literature [2, 1, 8, 12].

Consider a bipartite graph $G = (A, B, E)$. Suppose M_A is a d -disjunct $t_A \times |A|$ matrix with columns labelled by the vertices in A and M_B a d -disjunct $t_B \times |B|$ matrix with column labelled by the vertices in B . Construct a binary matrix M with columns labelled by the vertices in $A \cup B$ and for each row i of M_A and each row i' of M_B , M has a row with a label $\langle i, i' \rangle$ such that the cell $(\langle i, i' \rangle, u)$ contains an 1-entry if and only if either, in the case that $u \in A$, the cell (i, u) in M_A contains a 1-entry or, in the case that $u \in B$, the cell (i', u) in M_B contains a 1-entry.

Theorem 2.1 M is a $d(G)$ -disjunct matrix.

Proof. Consider the $d + 1$ edges e_0, e_1, \dots, e_d of $G = (A, B, E)$. Denote $e_j =$

(x_j, y_j) for $j = 0, 1, \dots, d$, where $x_j \in A$ and $y_j \in B$. Since M_A is d -disjunct, there exists a row i such that at row i , the column with label x_0 has an 1-entry and the columns with labels in $\{x_1, \dots, x_d\} \setminus \{x_0\}$ all have 0-entries. Similarly, there exists a row i' of M_B such that at row i' , the column with label y_0 has an 1-entry and the columns with labels in $\{y_1, \dots, y_d\} \setminus \{y_0\}$ all have 0-entries. Now, we look at row $\langle i, i' \rangle$ of M . This row gives a pool containing edge e_0 , but not e_j for all $j = 1, \dots, d$, since for $j = 1, \dots, d$, either $x_j \neq x_0$ or $y_j \neq y_0$, which means that either x_j or y_j has a 0-entry at row $\langle i, i' \rangle$. \square

3 The Second Construction

The second construction is a modification of the construction given in [4] for general $d(H)$ -disjunct matrices in a complex model.

Consider a bipartite graph $G = (A, B, E)$. Let $GF(q)$ be a finite field of order q . For each vertex $u \in A$, we associate it with a polynomial f_u of degree $k - 1$ over $GF(q)$ and for each vertex $v \in B$, we associate it with a polynomial g_v of degree $k - 1$ over $GF(q)$. Thus, each edge $e = (u, v)$ of G would be associated with a pair of polynomials (f_u, g_v) .

We first construct a $t \times |E|$ matrix $A_G(q, k, t)$ with the rows labelled by t elements in $GF(q)$ and the columns labelled by all the edges of G such that each cell (x, e) contains a vector $(f_u(x), g_v(x))$ where $e = (u, v)$.

Theorem 3.1 *Suppose $t \geq d(k - 1) + 1$. Then $A_G(q, k, t)$ has the property that for any $d + 1$ columns C_0, C_1, \dots, C_d , there exists a row at which the entry of C_0 does not equal the entry of C_j for all $j = 1, 2, \dots, d$.*

Proof. Suppose to the contrary that such a row does not exist. Then at any row, the entry of C_0 equals the entry of C_j for some $j \in \{1, 2, \dots, d\}$. Thus, there exists a $j \in \{1, 2, \dots, d\}$ such that the entries of C_0 equal the corresponding entries of C_j at at least k rows. Let $e_0 = (u_0, v_0)$ and $e_j = (u_j, v_j)$ be the edges associated with columns C_0 and C_j , respectively. Then $f_{u_0}(x) = f_{u_j}(x)$ and $g_{v_0}(x) = g_{v_j}(x)$ for at least k distinct values of x . Therefore, $f_{u_0} = f_{u_j}$ and $g_{v_0} = g_{v_j}$. Hence, $e_0 = e_j$, a contradiction. \square

We now construct a $d(G)$ -disjunct matrix $B_G(q, k, t)$ from $A_G(q, k, t)$. $B_G(q, k, t)$ has $|A \cup B|$ columns labelled with all the vertices of G . For each row x of $A_G(q, k, t)$ and each entry (y, z) at row x , we construct a row with label $\langle x, (y, z) \rangle$ for $B_G(q, k, t)$ such that the cell $(\langle x, (y, z) \rangle, u)$ for $u \in A$ contains an 1-entry if and only if $f_u(x) = y$ and cell $(\langle x, (y, z) \rangle, v)$ for $v \in B$ contains an 1-entry if and only if $g_v(x) = z$.

Theorem 3.2 *Suppose $t \geq d(k - 1) + 1$. Then $B_G(q, k, t)$ is $d(G)$ -disjunct.*

Proof. Consider $d + 1$ edges e_0, e_1, \dots, e_d of G . By Theorem 3.1, $A_G(q, k, t)$ has a row x such that the entry (y, z) in cell (x, e_0) does not equal the entry in cell (x, e_j) for all $j = 1, 2, \dots, d$. This means that the row $\langle x, (y, z) \rangle$ of $B_G(q, k, t)$ corresponds to a pool which contains e_0 , but not e_j for all $j = 1, 2, \dots, d$. Therefore, $B_G(q, k, t)$ is $d(G)$ -disjunct. \square

4 A Generalization

Both of the constructions in Sections 2 and 3 can be generalized in the following way.

Consider a hyper-graph $G = (V, E)$. Suppose G is c -colorable, that is, G can be partitioned into c disjoint independent sets with different colors. Let $GF(q)$ be a finite field of order q . For each vertex $u \in V$, we associate it with a polynomial p_u of degree $k - 1$ over $GF(q)$ such that for two vertices u and v with the same color, p_u and p_v must be distinct and for two vertices u and v with different colors, p_u and p_v are not necessarily distinct.

We first construct a $t \times |E|$ matrix $A_G(q, k, t)$ with the rows labelled by t elements in $GF(q)$ and the columns labelled by all the edges of G such that each cell (x, e) contains a set

$$\{(p_u(x), i) \mid u \in V \text{ with color } i\}.$$

Theorem 4.1 *Suppose $t \geq d(k - 1) + 1$. Then $A_G(q, k, t)$ has the property that for any $d + 1$ columns C_0, C_1, \dots, C_d , there exists a row at which the entry of C_0 does not contain the entry of C_j for all $j = 1, 2, \dots, d$.*

Proof. Suppose to the contrary that such a row does not exist. Then at any row, the entry of C_0 contains the entry of C_j for some $j \in \{1, 2, \dots, d\}$. Thus, there exists a $j \in \{1, 2, \dots, d\}$ such that entries of C_0 contain the corresponding entries of C_j at at least k rows. Let e_0 and e_j be the edges associated with columns C_0 and C_j , respectively. Consider a vertex $u \in e_j$. Suppose u is in color i . Then e_0 must contain a vertex v in color i and $p_v(x) = p_u(x)$ for at least k values of x . Hence, $p_v = p_u$, contradicting the assignment of polynomials. \square

We now construct a $d(G)$ -disjunct matrix $B_G(q, k, t)$ from $A_G(q, k, t)$. $B_G(q, k, t)$ has $|V|$ columns labelled with all the vertices of G . For each row x of $A_G(q, k, t)$ and each entry Q at row x , we construct a row with label $\langle x, Q \rangle$ for $B_G(q, k, t)$ such that the cell $(\langle x, Q \rangle, u)$ contains a 1-entry if and only if u is in color i and $p_u(x) = y$ for $(y, i) \in Q$.

Theorem 4.2 *Suppose $t \geq d(k - 1) + 1$. Then $B_G(q, k, t)$ is $d(G)$ -disjunct.*

Proof. Consider $d + 1$ edges e_0, e_1, \dots, e_d of G . By Theorem 4.1, $A_G(q, k, t)$ has a row x such that the entry Q in cell (x, e_0) does not contain the entry in cell (x, e_j) for all $j = 1, 2, \dots, d$. This means that the row $\langle x, Q \rangle$ of $B_G(q, k, t)$ corresponds to a pool which contains e_0 , but not e_j for all $j = 1, 2, \dots, d$. Therefore, $B_G(q, k, t)$ is $d(G)$ -disjunct. \square

Note that the number of the rows in $B_G(q, k, t)$ is bounded by $t \binom{qc}{r}$ where r is the maximum number of vertices on an edge of G and c is the number of the colors. Therefore, we would prefer to have less colors. This implies an interesting application of hyper-graph coloring problem.

5 Conclusion

In this paper, we study how to identify all protein-protein interactions between bait proteins and prey proteins which is critical in many biological applications.

We formulate the problem as a group testing problem in bipartite graphs. Two nonadaptive algorithms are proposed. Another generalization of our idea in a more general case is also discussed.

References and Notes

- 1 E. Barillot, B. Lacroix and D. Cohen (1991) 'Theoretical analysis of library screening using a N -dimensional pooling designs', *Nucleic Acids Res*, Vol. 19, pp.6241–6247.
- 2 D.-Z. Du and F. K. Hwang (1999) *Combinatorial Group Testing and Its Applications (2nd ed.)*, World Scientific, Singapore.
- 3 D.-Z. Du and F. K. Hwang (2005) 'Pooling Designs: Group Testing in Biology', manuscript.
- 4 D.-Z. Du, F.K. Hwang, W. Wu and T. Znati (2004) 'A new construction of transversal designs', manuscript.
- 5 D.-Z. Du, F.K. Hwang, W. Wu, Z. Liu and T. Znati (2005) 'Construction of $d(H)$ -Disjunct Matrix for Group Testing in Complex Model', manuscript.
- 6 A. G. D'ychkov, A. J. Macula, D. C. Torney, and P. A. Vilenkin (2001) 'Two models of nonadaptive group testing for designing screening experiments', *Proc. 6th Int. Workshop on Model-Oriented Designs and Analysis*, pp.63–75.
- 7 M. Farach, S. Kannan, E. Knill and S. Muthukrishnan (1997) 'Group testing problem with sequences in experimental molecular biology', *Proc. Compression and Complexity of Sequences*, pp.357–367.
- 8 H. Q. Ngo and D.-Z. Du (2000) 'A survey on combinatorial group testing algorithms with applications to DNA library screening', *Discrete mathematical problems with medical applications* (New Brunswick, NJ, 1999), pp.171–182, DIMACS Ser. Discrete Math. Theoret. Comput. Sci., 55, Amer. Math. Soc., Providence, RI.
- 9 A. J. Macula, D.C. Torney and P.A. Villenkin (2000) 'Two-stage group testing for complexes in the presence of errors', *DIMACS Sries in Disc. Math. and Theor. Comput. Sci.*, Vol. 55, pp.145–157.
- 10 A.J. Macula, V.V. Rykov and S. Yekhanin 'Trivial two-stage group testing for complexes using almost disjunct matrices', *Disc. Appl. Math.*
- 11 M. V. Marathe, A. G. Percus, and D. C. Torney (2000) 'Combinatorial optimization in biology', manuscript.
- 12 H. Park, W. Wu, X. Wu, and H.G. Zhao (2003) 'DNA screening, nonadaptive group testing, and simplicial complex', *Journal of Combinatorial Optimization*, Vol. 7, pp.389–394.
- 13 N. Thierry-Mieg, L. Trilling, and J.-L. Roch (2004) 'Anovel pooling design for protein-protein interaction mapping', manuscript.
- 14 D.C. Torney (1999) 'Sets pooling designs', *Ann. Combin.* Vol. 3, pp.95–101.
- 15 E. Triesch (1996) 'A group testing problem for hypergraphs of bounded rank', *Disc. Appl. Math.* Vol. 66, pp.185–188.
- 16 W. Wu, C. Li, X. Huang and Y. Li (2004) 'On error-tolerant DNA screening', submitted to *Discrete Applied Mathematics*.