

Finding Bicliques in Digraphs: Application into Viral-host Protein Interactome

Malay Bhattacharyya, SRF, MIU, ISI Kolkata, India

Co-authors: Sanghamitra Bandyopadhyay & Ujjwal Maulik

Overview

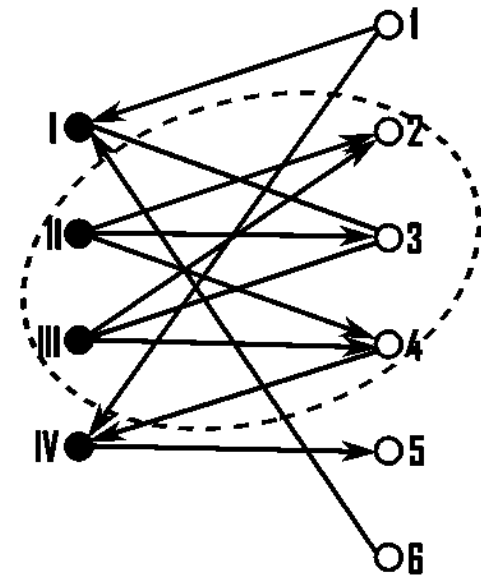
- ▶ The problem
- ▶ Related works
- ▶ Addressing the problem
- ▶ Application into host-viral protein interaction network

Directed Bipartite Graph

- ▶ If $V1, V2$ are two distinct sets of vertices and E is a subset of $V1 \times V2$ then a directed bipartite graph is definable as

$$G = (V1, V2, E)$$

where the edges (i, j) and (j, i) in E are distinct.



Finding DBCliques

Definition 1 (DBClique). *A DBClique is a fully connected subgraph $G' = (V'_1, V'_2, E') \subseteq G$ of a directed bipartite graph G such that either $i \in V'_1, j \in V'_2, \forall (i, j) \in E'$ or $i \in V'_2, j \in V'_1, \forall (i, j) \in E'$.*

Biclustering

A	A	A	A
A	A	A	A
A	A	A	A
A	A	A	A

(a)

A	A	A	A
B	B	B	B
C	C	C	C
D	D	D	D

(b)

A	A+a	A+b	A+c
A+p	A+p+a	A+p+b	A+p+c
A+q	A+q+a	A+q+b	A+q+c
A+r	A+r+a	A+r+b	A+r+c

(c)

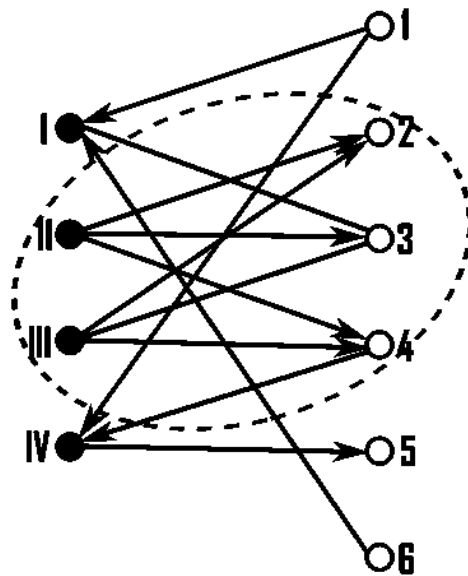
(d)

Pandey *et al.*, KDD, France, 2009

Related Works

- ▶ **Biclustering approaches**
 - ▶ Cheng and Church's algorithm (CCA)
 - ▶ SAMBA
 - ▶ Co-clustering algorithm (CA)
 - ▶ Divide-and-conquer based algorithm (DBA)
- ▶ Biclusters are classified into the types – fixed value (CCA, SAMBA, CA, DA), fixed row/column (CCA), additive coherent value, and coherent evolution
- ▶ Some are able to find overlapping biclusters (CCA, CA)
- ▶ The equivalence of biclique finding and biclustering

Correspondence of a DBClique to an Interaction Matrix



	1	2	3	4	5	6
I	-1	0	X	0	0	-1
II	0	1	1	1	0	0
III	0	1	X	1	0	0
IV	-1	0	0	-1	1	0

Formalization of an Interaction Matrix for a Directed Bipartite Graph

Definition 2 (Interaction matrix of a directed bipartite graph). *The interaction matrix of a directed bipartite graph $G = (V_1, V_2, E)$ is defined as a $|V_1| \times |V_2|$ matrix $\mathcal{I}$ such that*

$$\mathcal{I}_{ij} = \begin{cases} 0, & \text{if } (i, j) \notin E \text{ and } (j, i) \notin E \\ 1, & \text{if } (i, j) \in E \text{ and } (j, i) \notin E \\ -1, & \text{if } (i, j) \notin E \text{ and } (j, i) \in E \\ X, & \text{if } (i, j) \in E \text{ and } (j, i) \in E \end{cases},$$

An Observation

Lemma 1. *Given a directed bipartite graph $G = (V_1, V_2, E)$, a DBClique $G' = (V'_1, V'_2, E') \subseteq G$ corresponds to a bicluster in the interaction matrix of G such that all the elements in the submatrix are either '1' or '-1', with the entries of 'X' additionally allowed.*

The Approach

Algorithm 1 An Algorithm for Finding out Bicliques in Digraphs

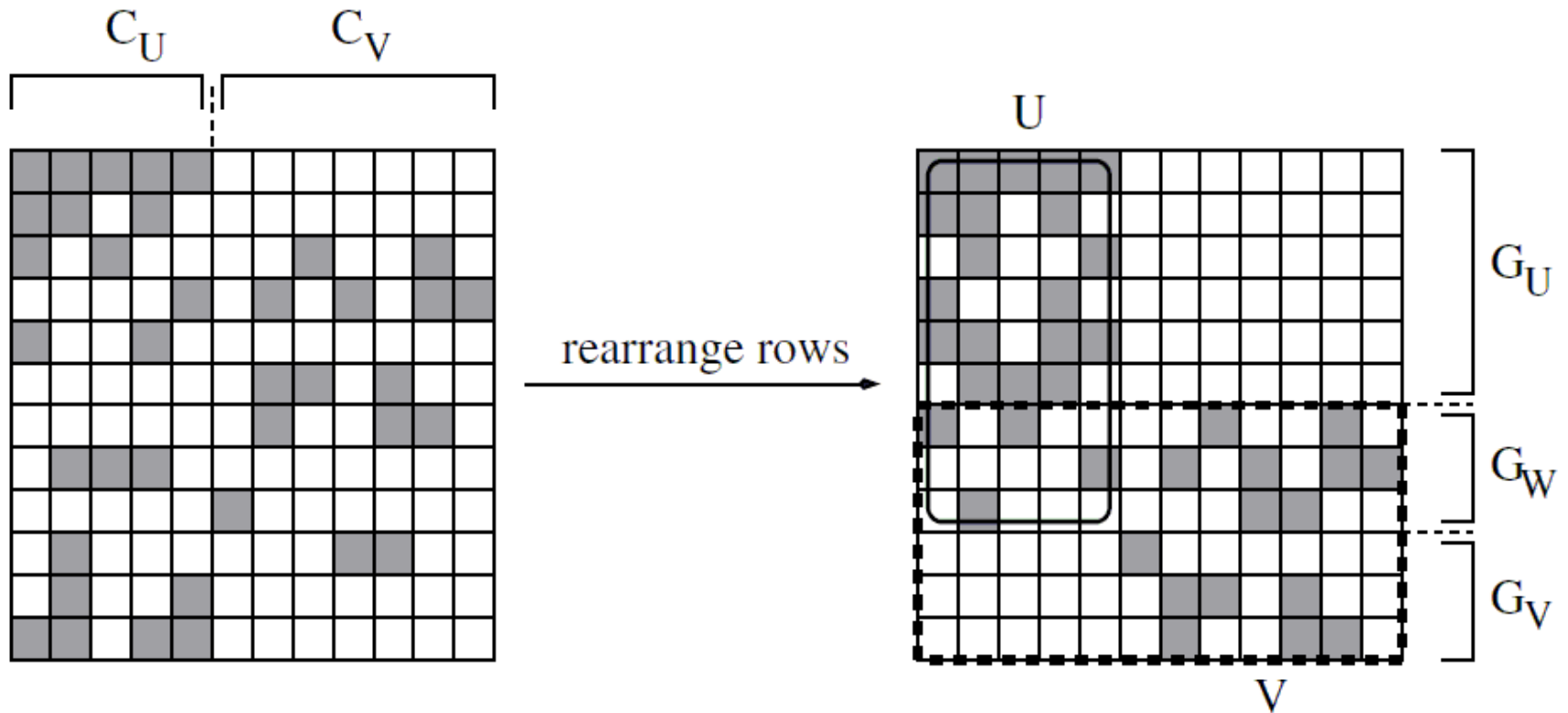
Input: A directed bipartite graph $G = (V_1, V_2, E)$.

Output: The set of maximal DBCliques.

Steps of the algorithm:

- 1: Obtain the correspondent interaction matrix $\mathcal{I}$ from G
 - 2: Replace the entries 'X' with '1' and '-1' with '0' in $\mathcal{I}$ // Finding the all '1' biclusters
 - 3: Partition $\mathcal{I} = \mathcal{I}_0 \cup \mathcal{I}_1 \cup \mathcal{I}_2$ such that the size of $\mathcal{I}_0$ maximizes and it contains only 0's.
 - 4: Go to the previous step and apply the same individually on $\mathcal{I}_1$ and $\mathcal{I}_2$ until no further partitioning is possible.
 - 5: Return the DBCliques corresponding to the biclusters
 - 6: Replace the entries 'X' with '-1' and '1' with '0' in $\mathcal{I}$ // Finding the all '-1' biclusters
 - 7: Partition $\mathcal{I} = \mathcal{I}_0 \cup \mathcal{I}_1 \cup \mathcal{I}_2$ such that the size of $\mathcal{I}_0$ maximizes and it contains only 0's.
 - 8: Go to the previous step and apply the same individually on $\mathcal{I}_1$ and $\mathcal{I}_2$ until no further partitioning is possible.
 - 9: Return the DBCliques corresponding to the biclusters
-

Division of the Interaction Matrix



Prelic et al., *Bioinformatics*, 22(9):1122-1129, 2006

Nature Precedings : doi:10.1038/npre.2012.7148.1 : Posted 4 Apr 2012



Details of the Data

- ▶ Direct physical interactions/indirect interactions – categorized into 65 more specific types
- ▶ 19 HIV-1 proteins and 1448 human proteins
- ▶ 5134 interactions (18.66% of the total possible)

DBCliques Obtained

Table 1. The DBCliques obtained from the HIV-1-human protein interaction network containing at least three HIV-1 and human proteins each. The size of a DBClique is defined based on the number of edges it contains.

Biclustertype	<i>Don't care</i> allowed	# DBCliques obtained	Maximum size (HIV-1, Human)
All '1'	Yes	113	(6, 5)
All '-1'	Yes	25	(3, 13)
All '1'	No	54	(4, 5)
All '-1'	No	7	(3, 8)

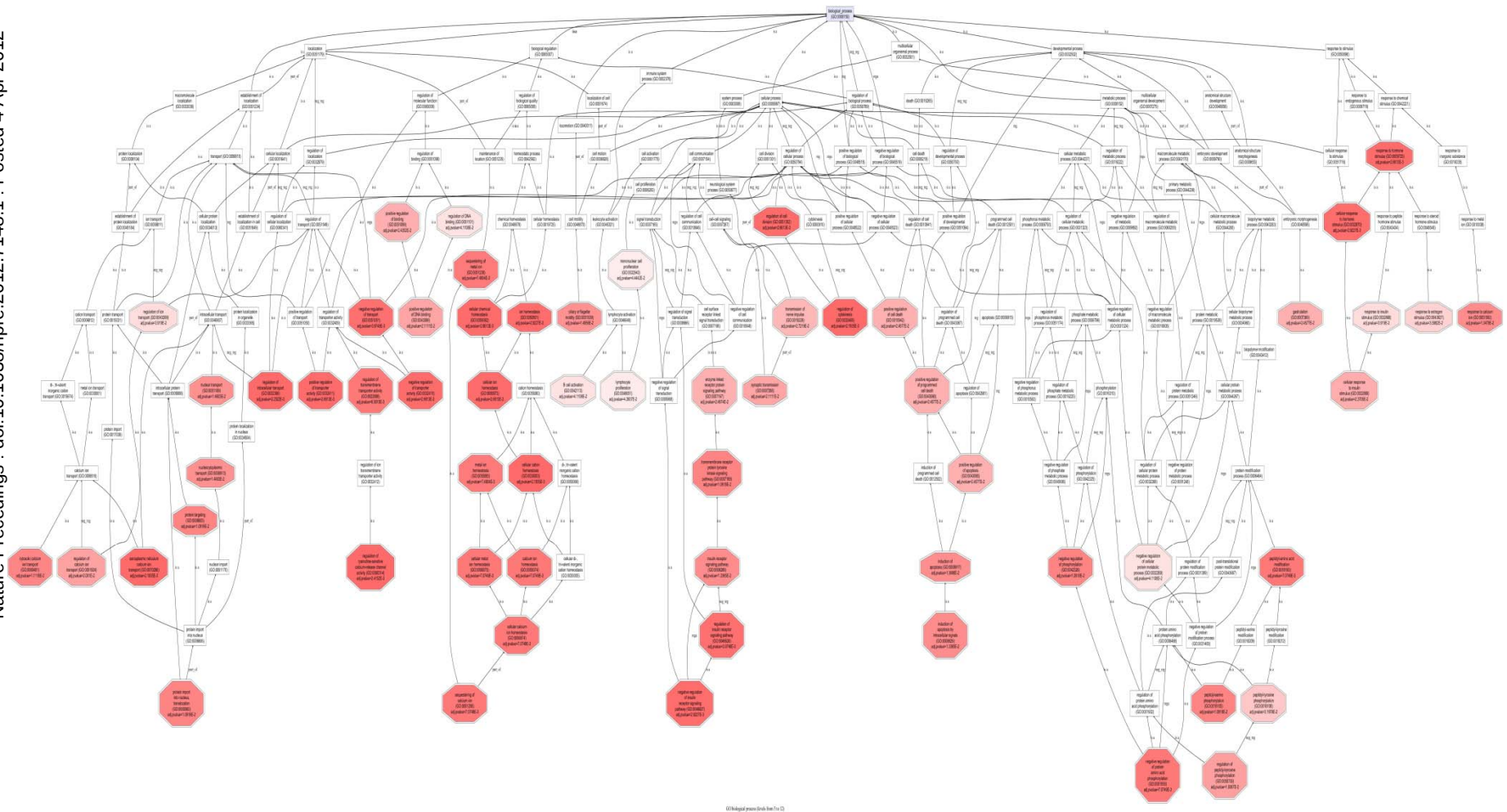
Comparative Results

Table 2. Comparison of the largest bicliques (consisting of at least three HIV-1 and human proteins) derived by various algorithms from the HIV-1-human protein interaction network. The proposed method exclude the *Don't care* conditions and returns DBCliques. Crossed cells in the third column represent insignificant p -values.

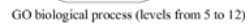
Analytical details	Bimax	CC	ISA	Proposed
# Bicliques obtained	197	60	10	61
Largest biclique found	(4, 9)	(19, 392)	(5, 76)	(3, 8)
Best p -value from GO	1.9E−6	×	×	2.3E−12
Best annotation (GO Term)	Regulation of cytokinesis (GO:0032465)	Not applicable	Not applicable	Response to protein stimulus (GO:0051789)

GO – Biological Process

Nature Precedings : doi:10.1038/npre.2012.7148.1 : Posted 4 Apr 2012



PReMI, Russia June 29, 2011



References

1. Barkow, S., Bleuler, S., Prelić A., Zimmermann, P., Zitzler, E.: BicAT: a Biclustering Analysis Toolbox. *Bioinformatics* 22(10), 1282–1283 (2006)
2. Brass, A.L., Dykxhoorn, D.M., Benita, Y., Yan, N., Engelman, A., Xavier, R.J., Lieberman, J., Elledge, S.J.: Identification of Host Proteins Required for HIV Infection Through a Functional Genomic Screen. *Science* 319(5865), 921–926 (2008)
3. Cheng, Y., Church, G.: Biclustering of Expression Data. *Proceedings of the 8th ISMB Conference, AAAI Press*, 93–103 (2000)
4. Ding, C., Zhang, Y., Li, T.: Biclustering Protein Complex Interactions with a Biclique Finding Algorithm. *Proceedings of the Sixth International Conference on Data Mining, Hong Kong*, 178–187 (2006)
5. Fu, W., Sanders-Beer, B.E., Katz, K.S., Maglott, D.R., Pruitt, K.D., Ptak, R.G.: Human immunodeficiency virus type 1, human protein interaction database at NCBI. *Nucleic Acids Research (Database Issue)* 37, D417–D422 (2009)
6. Madeira, S.C., Oliveira, A.L.: Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 1, 24–45 (2004)
7. Pandey, G., Atluri, G., Steinbach, M., Myers, C.L., Kumar, V.: An Association Analysis Approach to Biclustering. *Proceedings of the 15th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Paris, France* (2009)
8. Prelić, A., Bleuler, S., Zimmermann, P., Wille, A., Bühlmann, P., Gruissem, P., Hennig, L., Thiele, L., Zitzler, E.: A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics* 22(9), 1122–1129 (2006)
9. Tanay, A., Sharan, R., Shamir, R.: Discovering statistically significant biclusters in gene expression data. *Bioinformatics* 18, S136–S144 (2002)

Thank You