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

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

Malay Bhattacharyya, SRF, MIU, ISI Kolkata, India

Co-authors: Sanghamitra Bandyopadhyay & Ujjwal Maulik

Overview

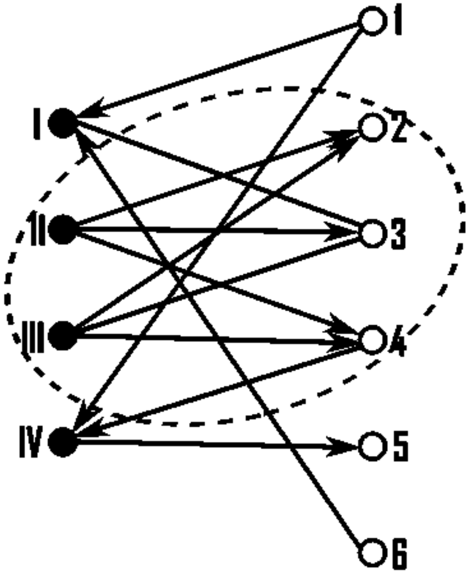
- The problem
- Related works
- Addressing the problem
- Application into host-viral protein interaction network

Directed Bipartite Graph

- If *V1*, *V2* are two distinct sets of vertices and *E* is a subset of *V1*x*V2* then a directed bipartite graph is definable as

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

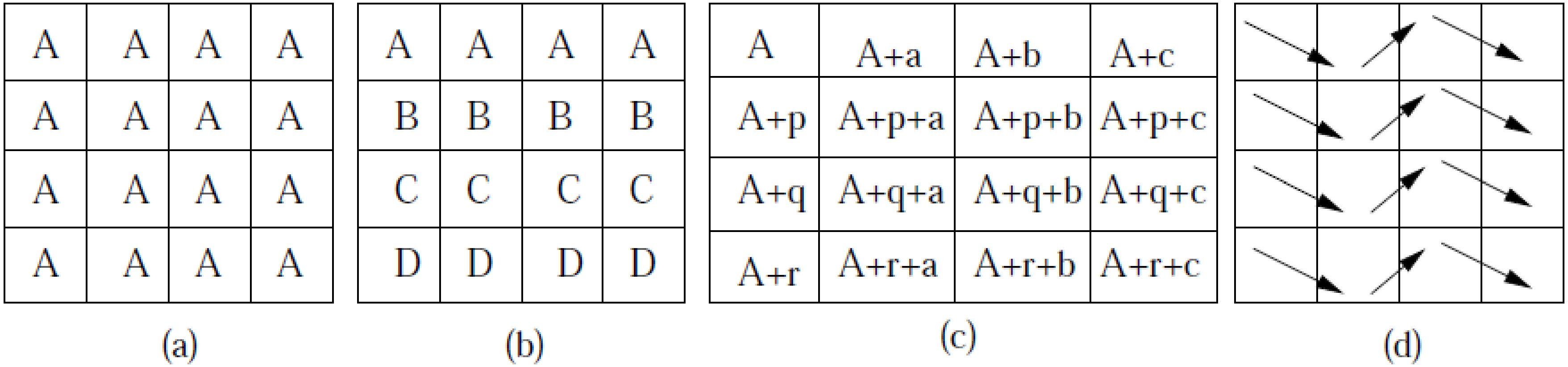
where the edges (*i*, *j*) and (*j*, *i*) in *E* are distinct.



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)

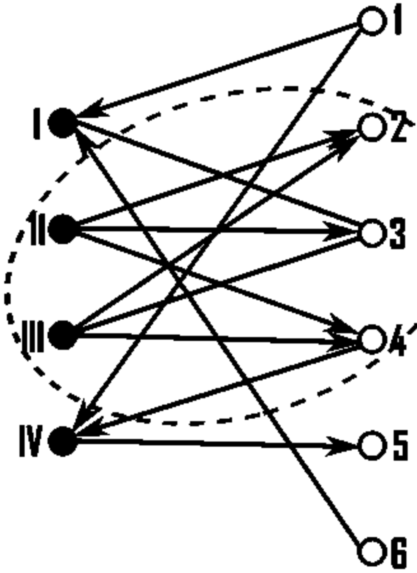
(d)

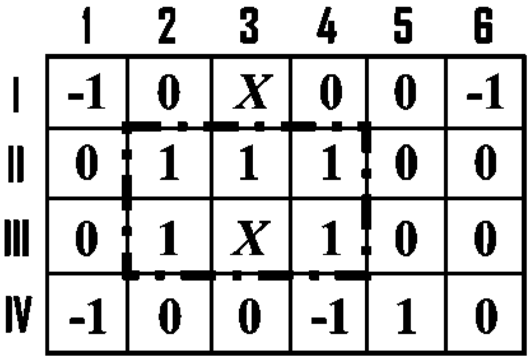
Pandey *et al.*, KDD, France, 2009

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

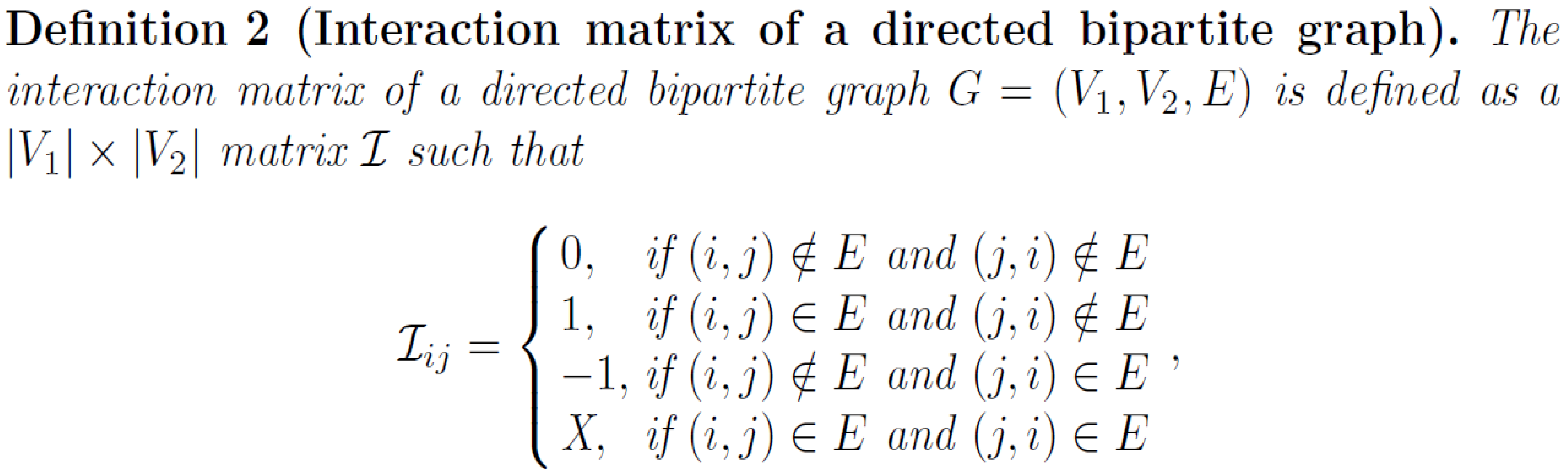
Correspondence of a DBClique to an Interaction Matrix





	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

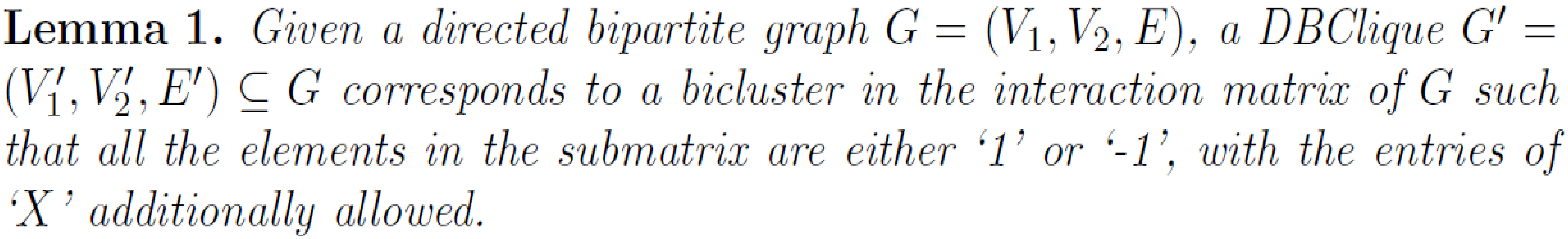
Formalization of an Interaction Matrix for a Directed Bipartite Graph



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*

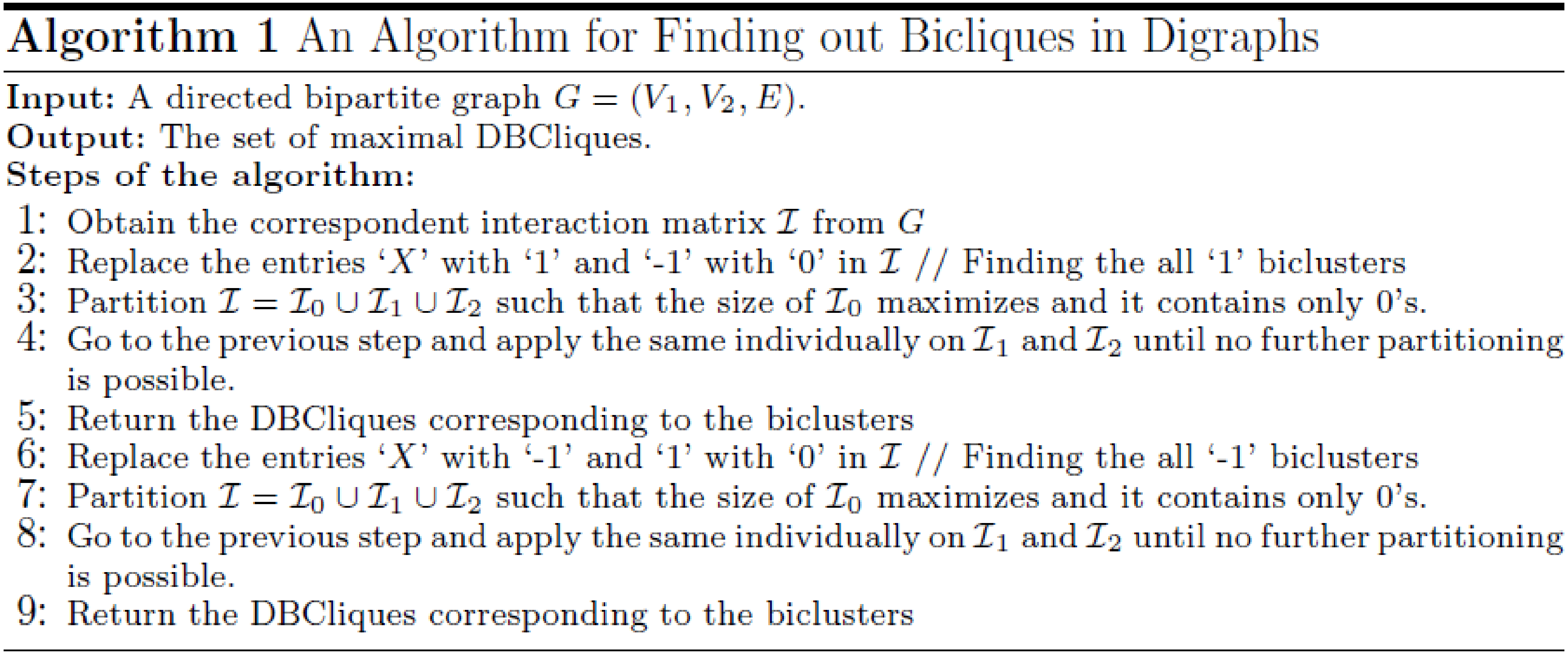
$$\mathcal{I}_{ij} = \begin{cases} 0, & \textit{if } (i,j) \notin E \textit{ and } (j,i) \notin E \\ 1, & \textit{if } (i,j) \in E \textit{ and } (j,i) \notin E \\ -1, & \textit{if } (i,j) \notin E \textit{ and } (j,i) \in E \\ X, & \textit{if } (i,j) \in E \textit{ and } (j,i) \in E \end{cases},$$

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

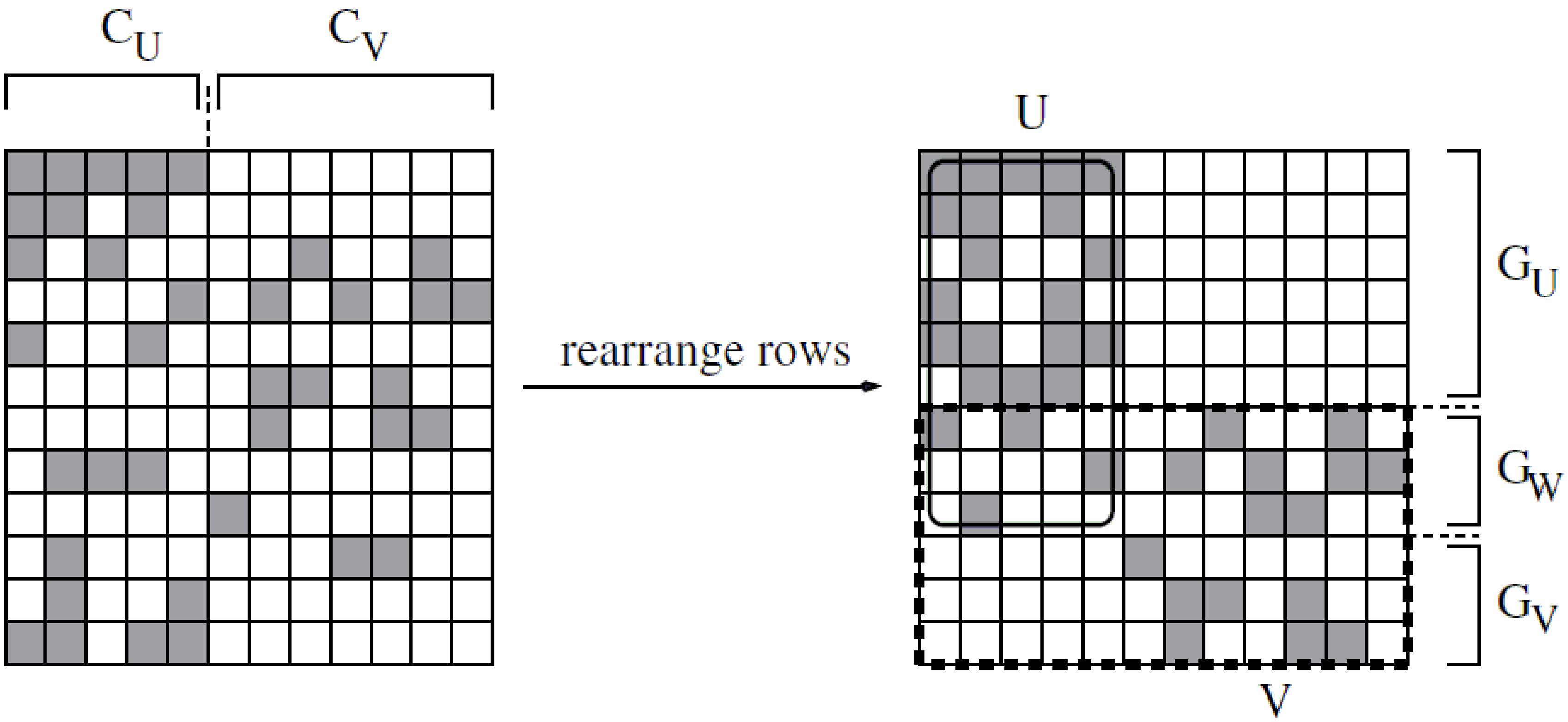


Algorithm 1 An Algorithm for Finding out Bicliques in Digraphs

Input: A directed bipartite graph $G = (V_1, V_2, E)$.
Output: The set of maximal DBCliques.
Steps of the algorithm:

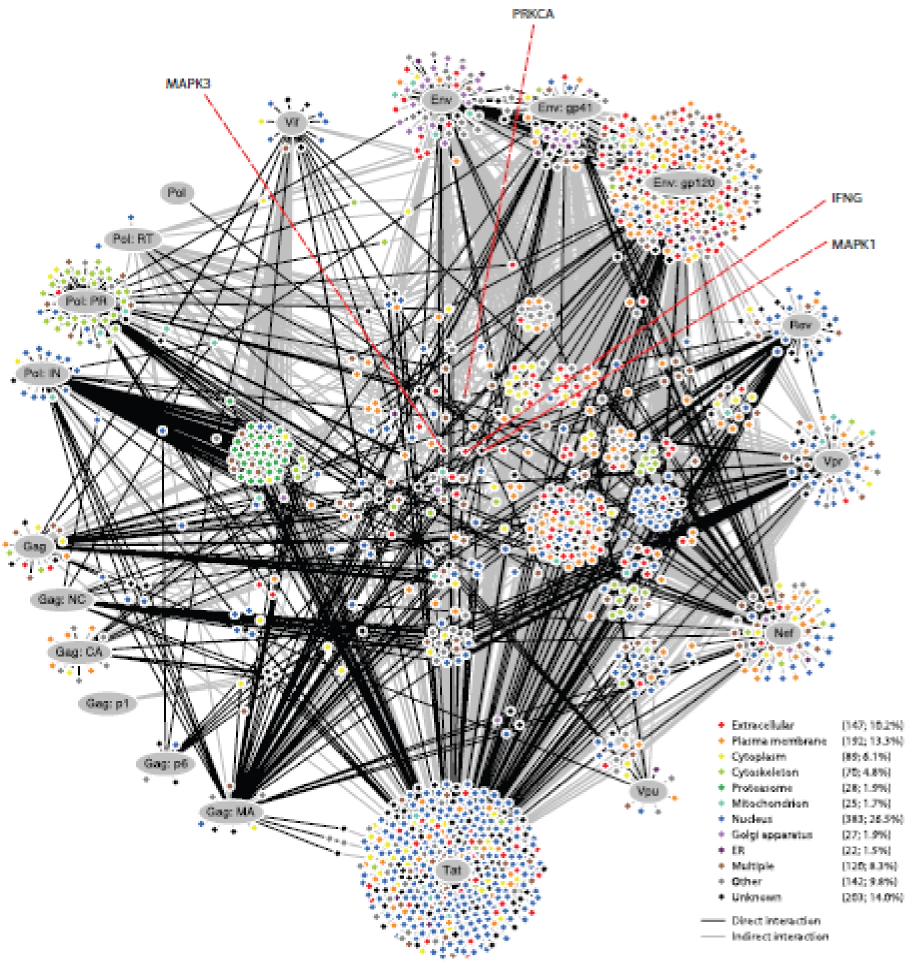
```
Obtain the correspondent interaction matrix $\mathcal{I}$ from $G$
Replace the entries ‘$X$’ with ‘1’ and ‘-1’ with ‘0’ in $\mathcal{I}$ // Finding the all ‘1’ biclusters
Partition $\mathcal{I} = \mathcal{I}_0 \cup \mathcal{I}_1 \cup \mathcal{I}_2$ such that the size of $\mathcal{I}_0$ maximizes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

Division of the Interaction Matrix



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

HIV-1–Human Protein Interaction Network



Ptak *et al.*, *AIDS Res Hum Retroviruses*, 24(12):1497-502, 2008

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

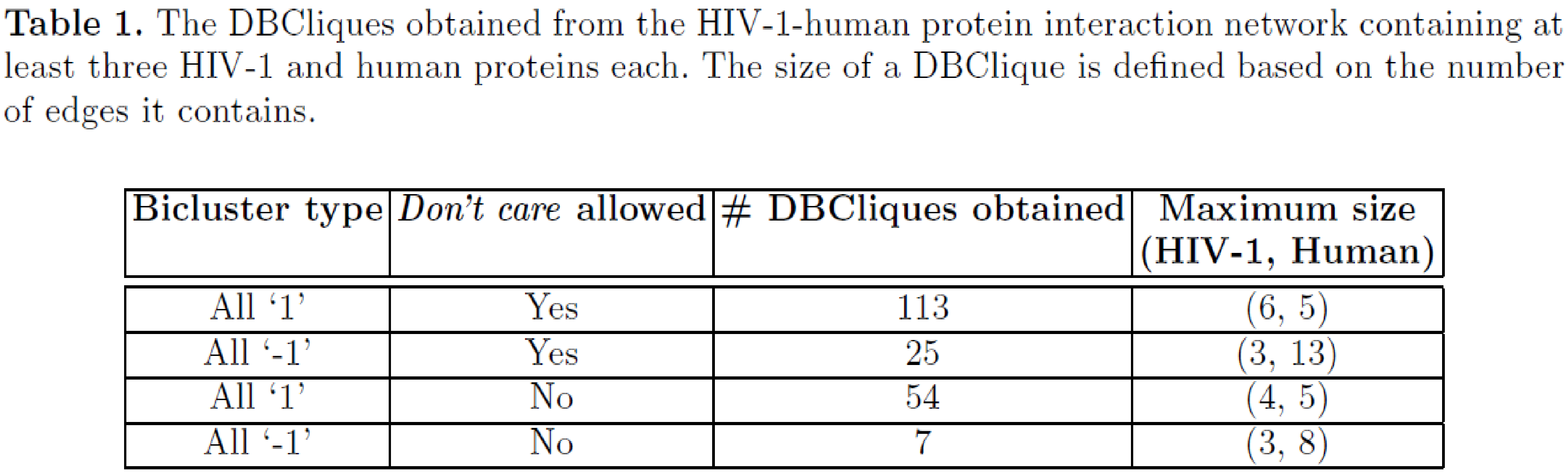


Table 1. The DBCliques obtained from the HIV-1-human protein interaction network containing at least three HIV-1 and human proteins each. The size of a DBClique is defined based on the number of edges it contains.

Bicluster type	***Don't care* 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

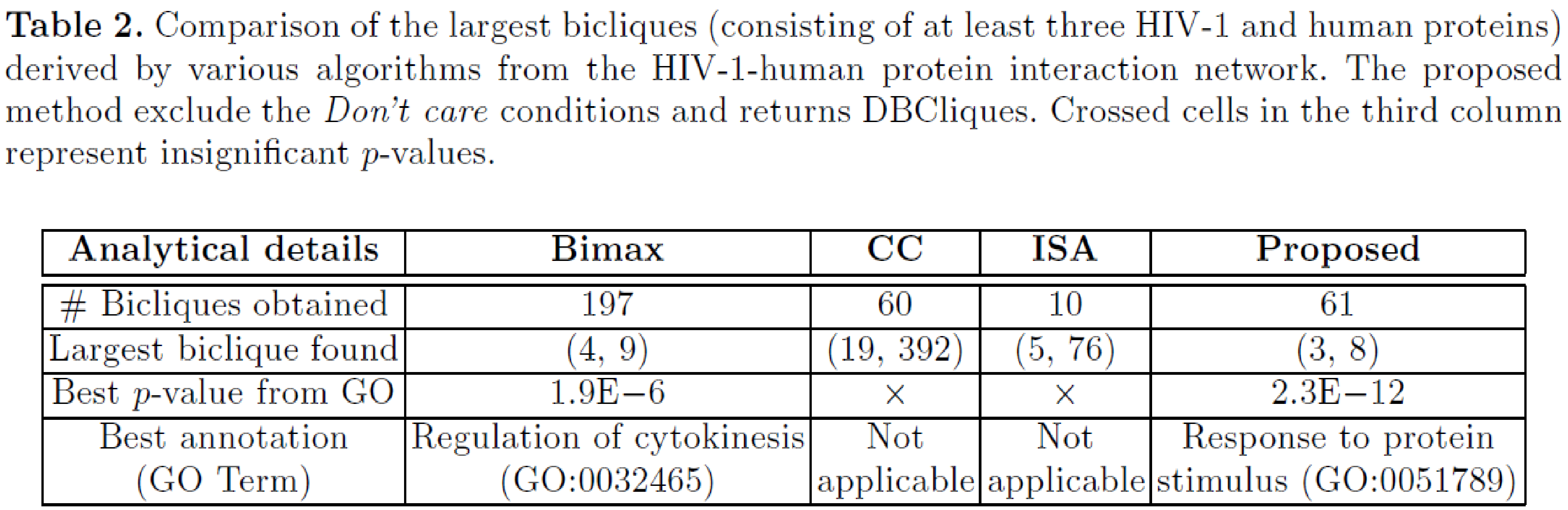
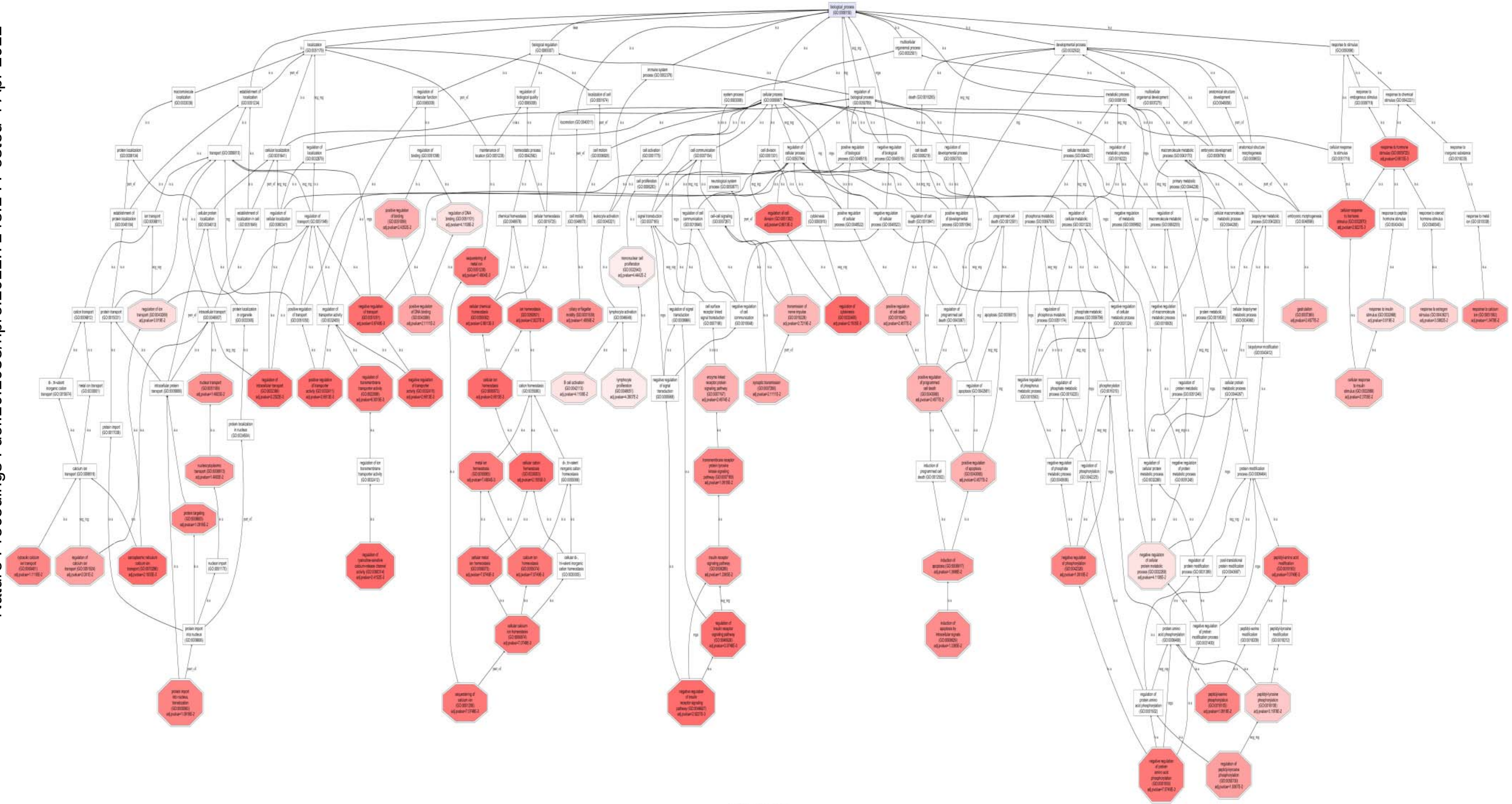


Table 2. Comparison of the largest bicliques (consisting of at least three HIV-1 and human proteins) derived by various algorithms from the HIV-1-human protein interaction network. The proposed method exclude the *Don't care* conditions and returns DBCliques. Crossed cells in the third column represent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

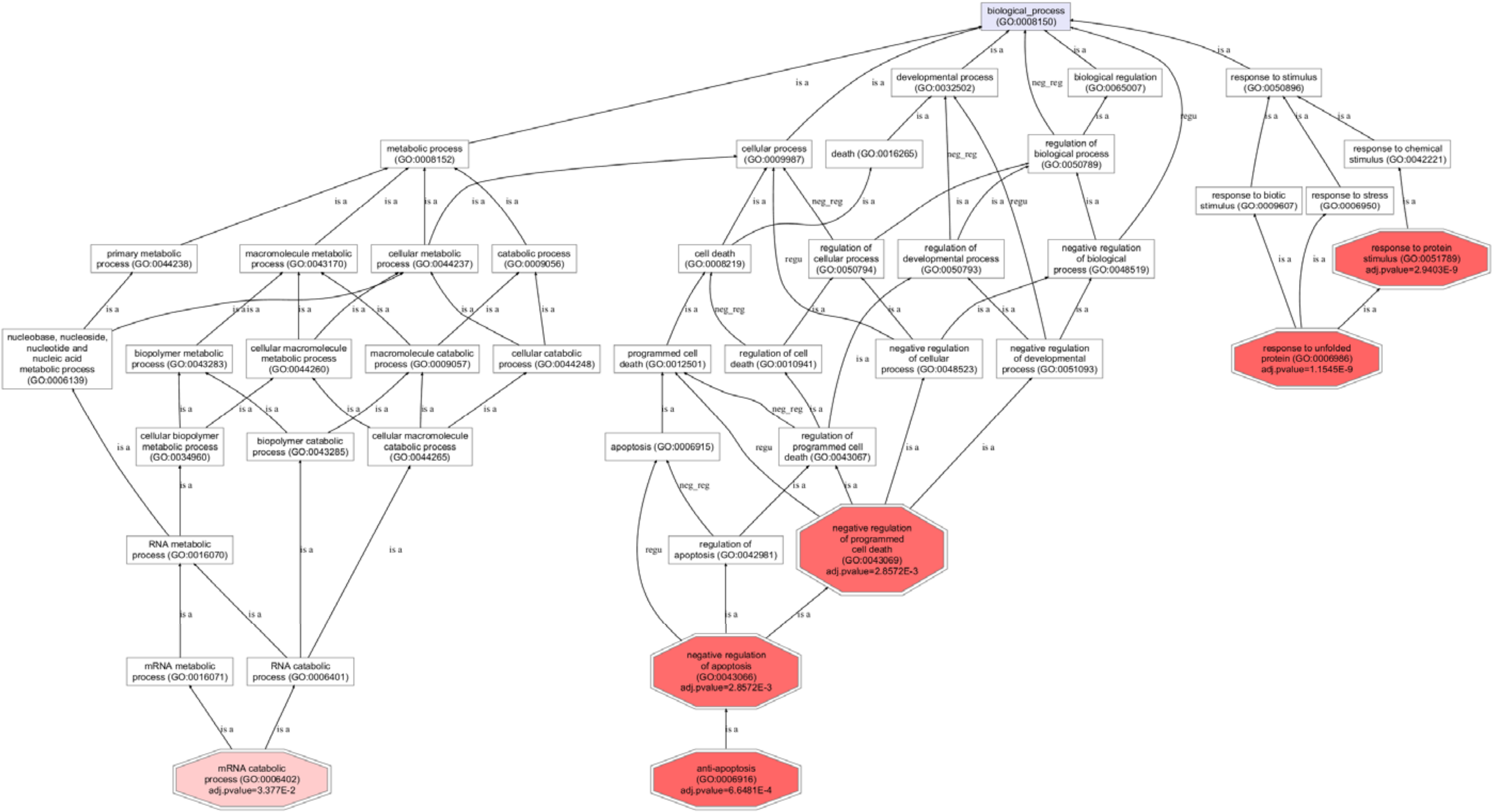
Analytical details	**Bimax**	**CC**	**ISA**	**Proposed**
# Bicliques obtained	197	60	10	61
Largest biclique found	(4, 9)	(19, 392)	(5, 76)	(3, 8)
Best *p*-value from GO	1.9E−6	×	×	2.3E−12
Best annotation (GO Term)	Regulation of cytokinesis (GO:0032465)	Not applicable	Not applicable	Response to protein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

GO – Biological Process


biological_process (GO:0008150)
multicellular organismal process (GO:0032501)

GO biological process (levels from 5 to 12)

GO – Molecular Function


biological_process (GO:0008150)
developmental process (GO:0032502)
biological regulation (GO:0065007)
response to stimulus (GO:0050896)
metabolic process (GO:0008152)
cellular process (GO:0009987)
death (GO:0016265)
regulation of biological process (GO:0050789)
response to chemical stimulus (GO:0042221)
response to biotic stimulus (GO:0009607)
response to stress (GO:0006950)
primary metabolic process (GO:0044238)
macromolecule metabolic process (GO:0043170)
cellular metabolic process (GO:0044237)
catabolic process (GO:0009056)
cell death (GO:0008219)
regulation of cellular process (GO:0050794)
regulation of developmental process (GO:0050793)
negative regulation of biological process (GO:0048519)
response to protein stimulus (GO:0051789) adj.pvalue=2.9403E-9
nucleobase, nucleoside, nucleotide and nucleic acid metabolic process (GO:0006139)
biopolymer metabolic process (GO:0043283)
cellular macromolecule metabolic process (GO:0044260)
macromolecule catabolic process (GO:0009057)
cellular catabolic process (GO:0044248)
programmed cell death (GO:0012501)
regulation of cell death (GO:0010941)
negative regulation of cellular process (GO:0048523)
negative regulation of developmental process (GO:0051093)
response to unfolded protein (GO:0006986) adj.pvalue=1.1545E-9
cellular biopolymer metabolic process (GO:0034960)
biopolymer catabolic process (GO:0043285)
cellular macromolecule catabolic process (GO:0044265)
apoptosis (GO:0006915)
regulation of programmed cell death (GO:0043067)
negative regulation of programmed cell death (GO:0043069) adj.pvalue=2.8572E-3
RNA metabolic process (GO:0016070)
regulation of apoptosis (GO:0042981)
mRNA metabolic process (GO:0016071)
RNA catabolic process (GO:0006401)
negative regulation of apoptosis (GO:0043066) adj.pvalue=2.8572E-3
mRNA catabolic process (GO:0006402) adj.pvalue=3.377E-2
anti-apoptosis (GO:0006916) adj.pvalue=6.6481E-4
is a
neg_reg
regu

GO biological process (levels from 5 to 12)

References

1. Barkow, S., Bleuler, S., Preli´c 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. Preli´c, A., Bleuler, S., Zimmermann, P., Wille, A., B¨uhlmann, 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