Nature Precedings : doi:10.1038/npre.2012.7119.1 : Posted 3 Apr 2012

Univerza *v Ljubljani*

STATISTICAL INDIAN INSTITUTE
UNITY IN DIVERSITY

Mining Co-expression Graphs: Applications to MicroRNA Regulation and Disease Analysis

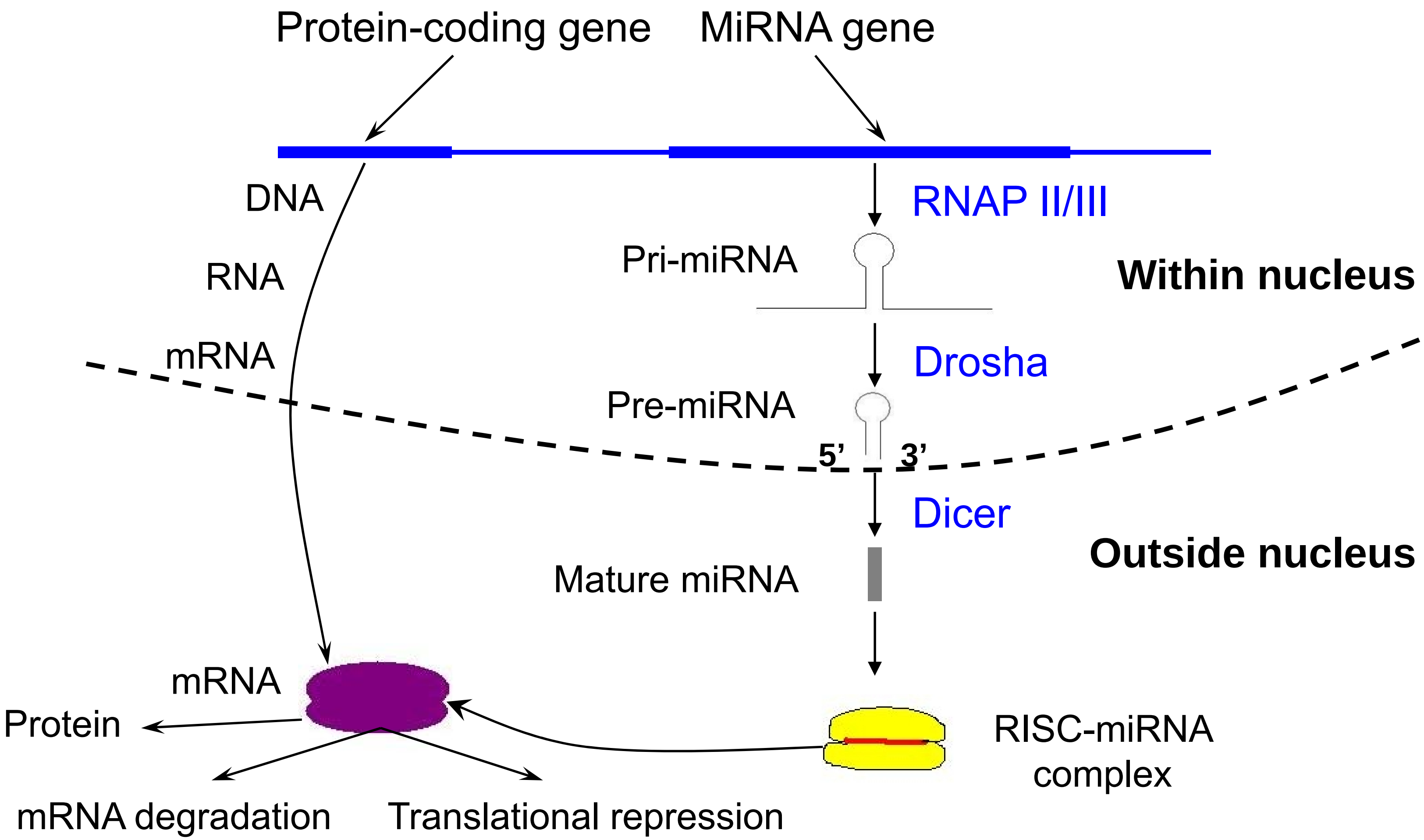
Malay Bhattacharyya

Senior Research Fellow

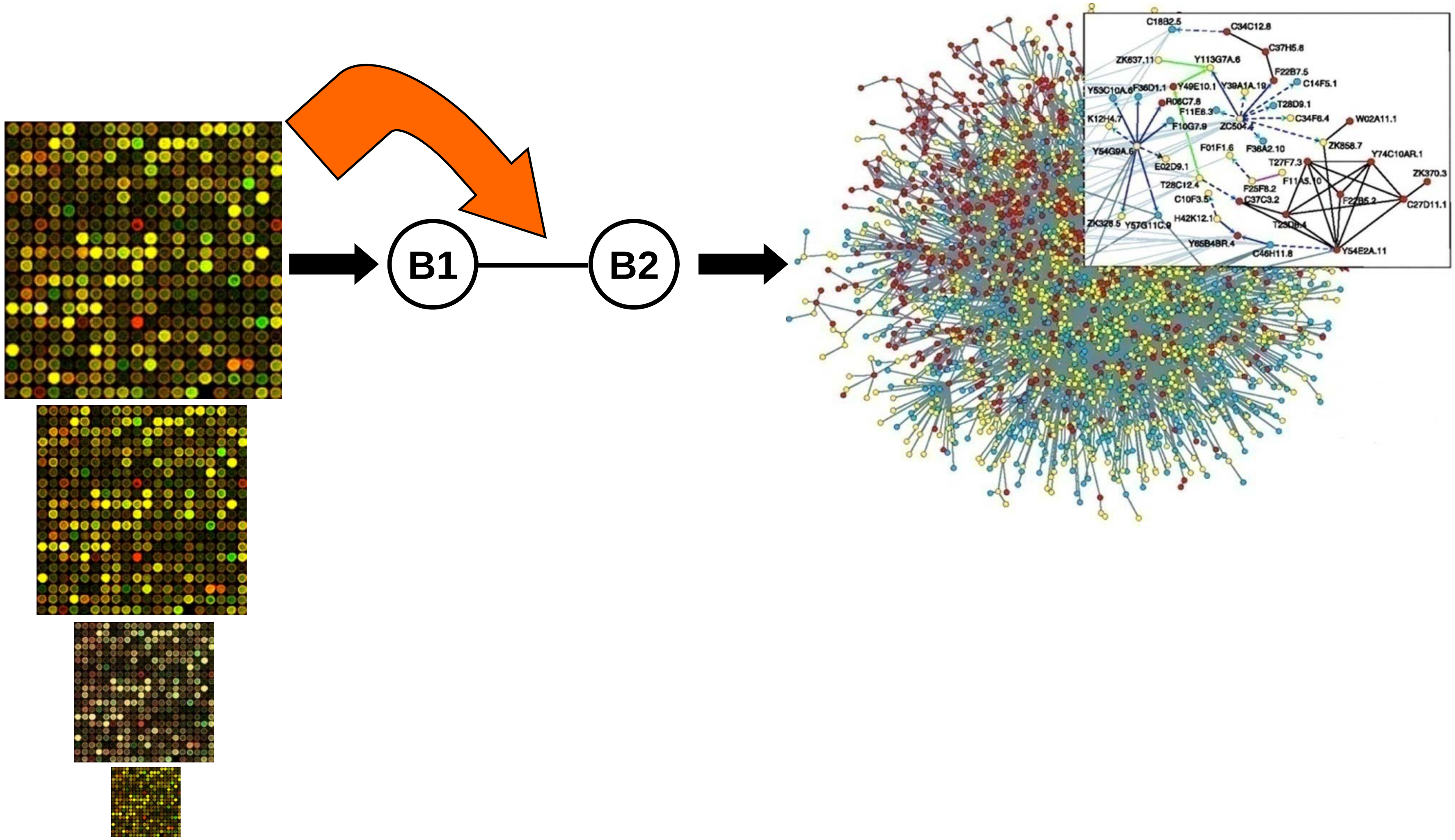
MIU, Indian Statistical Institute, Kolkata, India

Some Concepts from Biology

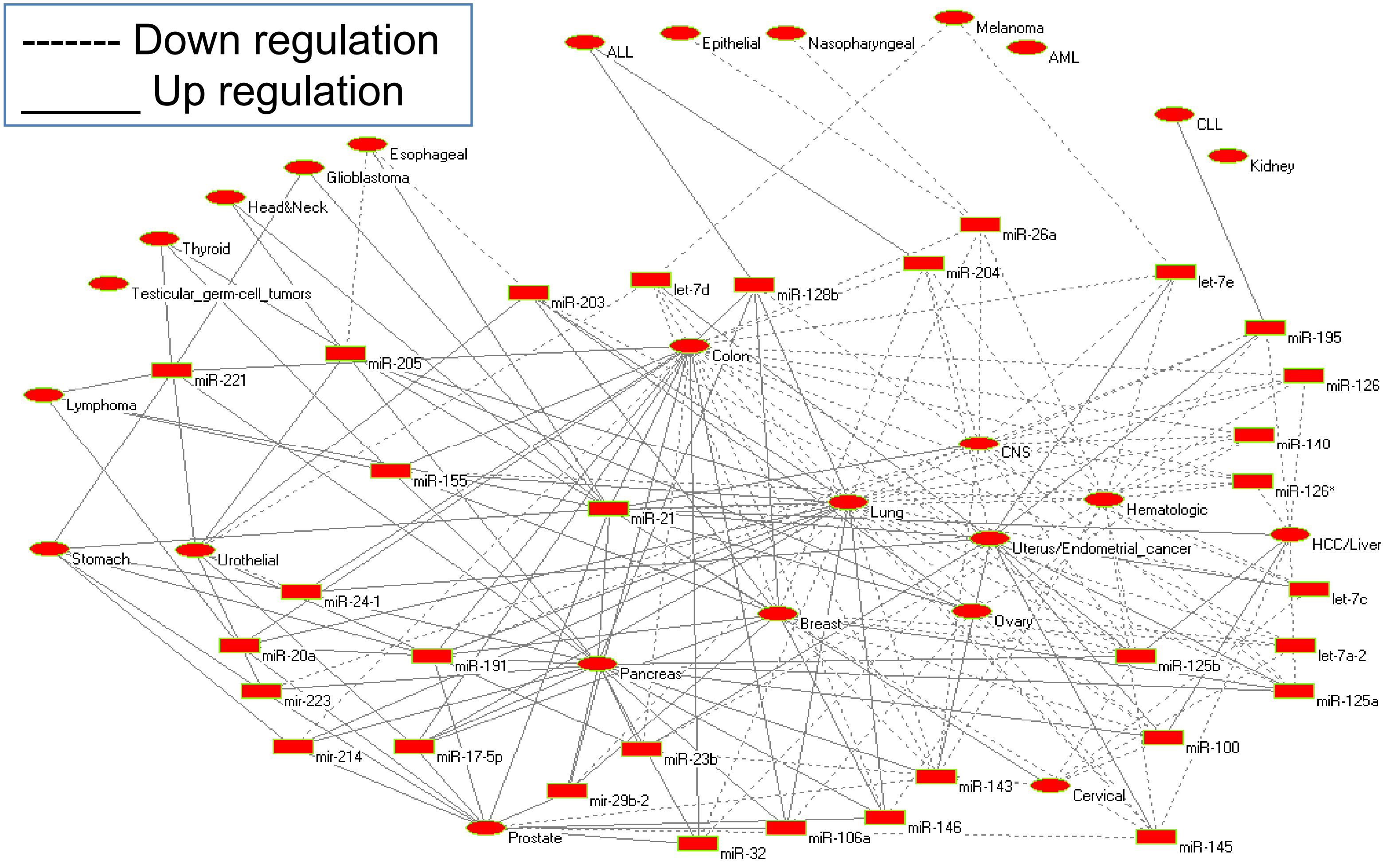
Biogenesis of Genes and MicroRNAs



Mapping from Molecules to Systems



Disease Association of MicroRNAs



------- Down regulation

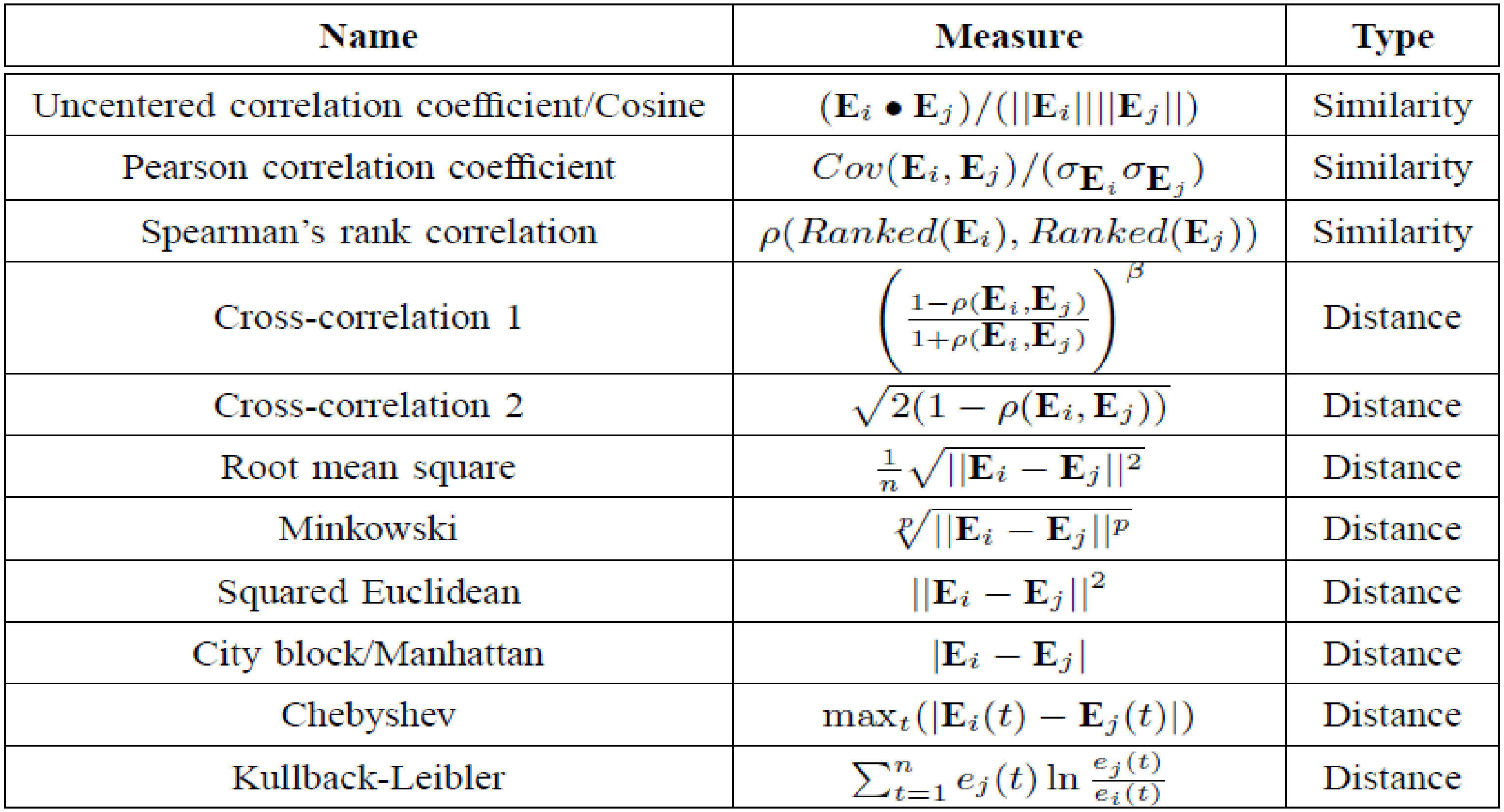
______ Up regulation

Outline of the Talk

- Expression studies to co-expression graph construction
- Co-expression graph mining
- Studying the differential co-expression graphs
- Co-expression to coregulation
- Disease analysis
- Future Goals

Expression Studies to Co-expression Graph Construction

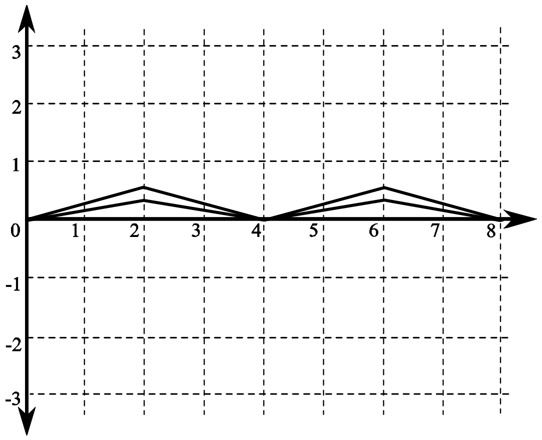
Co-expression Measures



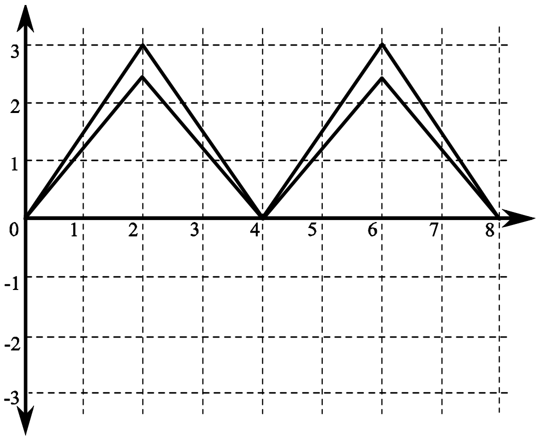
Name	Measure	Type
Uncentered correlation coefficient/Cosine	$(\mathbf{E}_i \bullet \mathbf{E}_j)/(\|\|\mathbf{E}_i\|\|\|\|\mathbf{E}_j\|\|)$	Similarity
Pearson correlation coefficient	$Cov(\mathbf{E}_i, \mathbf{E}_j)/(\sigma_{\mathbf{E}_i}\sigma_{\mathbf{E}_j})$	Similarity
Spearman's rank correlation	$\rho(Ranked(\mathbf{E}_i), Ranked(\mathbf{E}_j))$	Similarity
Cross-correlation 1	$\left(\frac{1-\rho(\mathbf{E}_i,\mathbf{E}_j)}{1+\rho(\mathbf{E}_i,\mathbf{E}_j)}\right)^{\beta}$	Distance
Cross-correlation 2	$\sqrt{2(1-\rho(\mathbf{E}_i, \mathbf{E}_j))}$	Distance
Root mean square	$\frac{1}{n}\sqrt{\|\|\mathbf{E}_i - \mathbf{E}_j\|\|^2}$	Distance
Minkowski	$\sqrt[p]{\|\|\mathbf{E}_i - \mathbf{E}_j\|\|^p}$	Distance
Squared Euclidean	$\|\|\mathbf{E}_i - \mathbf{E}_j\|\|^2$	Distance
City block/Manhattan	$\|\mathbf{E}_i - \mathbf{E}_j\|$	Distance
Chebyshev	$\max_t(\|\mathbf{E}_i(t) - \mathbf{E}_j(t)\|)$	Distance
Kullback-Leibler	$\sum_{t=1}^{n} e_j(t) \ln \frac{e_j(t)}{e_i(t)}$	Distance

- Very few measures for quantifying positive and negative dependence
- How to signify the amount of deviation?

Variation of Deviation in Co-expression

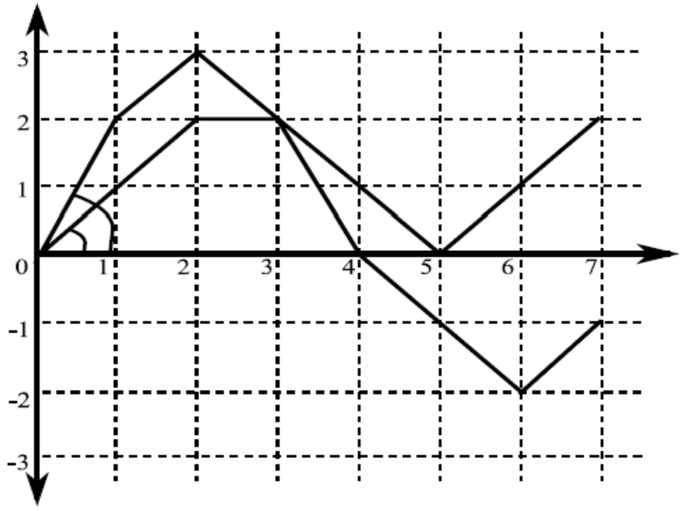


(I)



(II)

Modeling both similarity and deviation

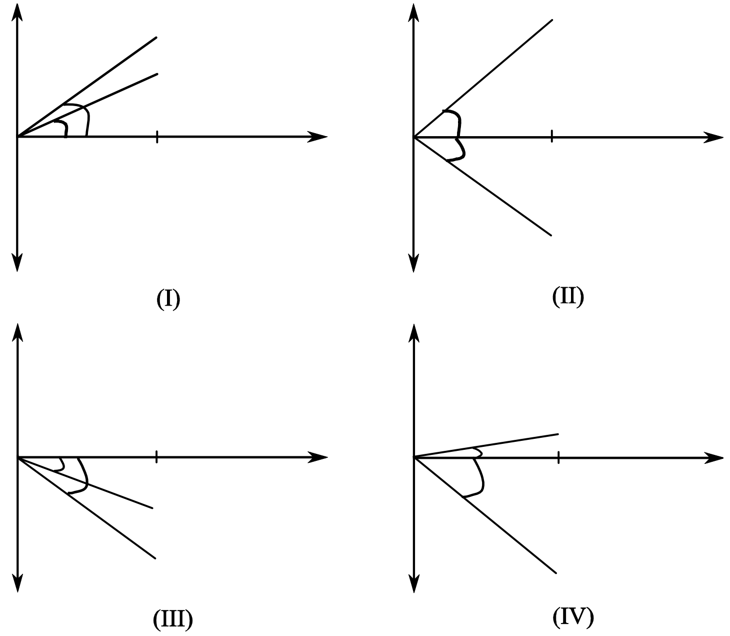


[1]

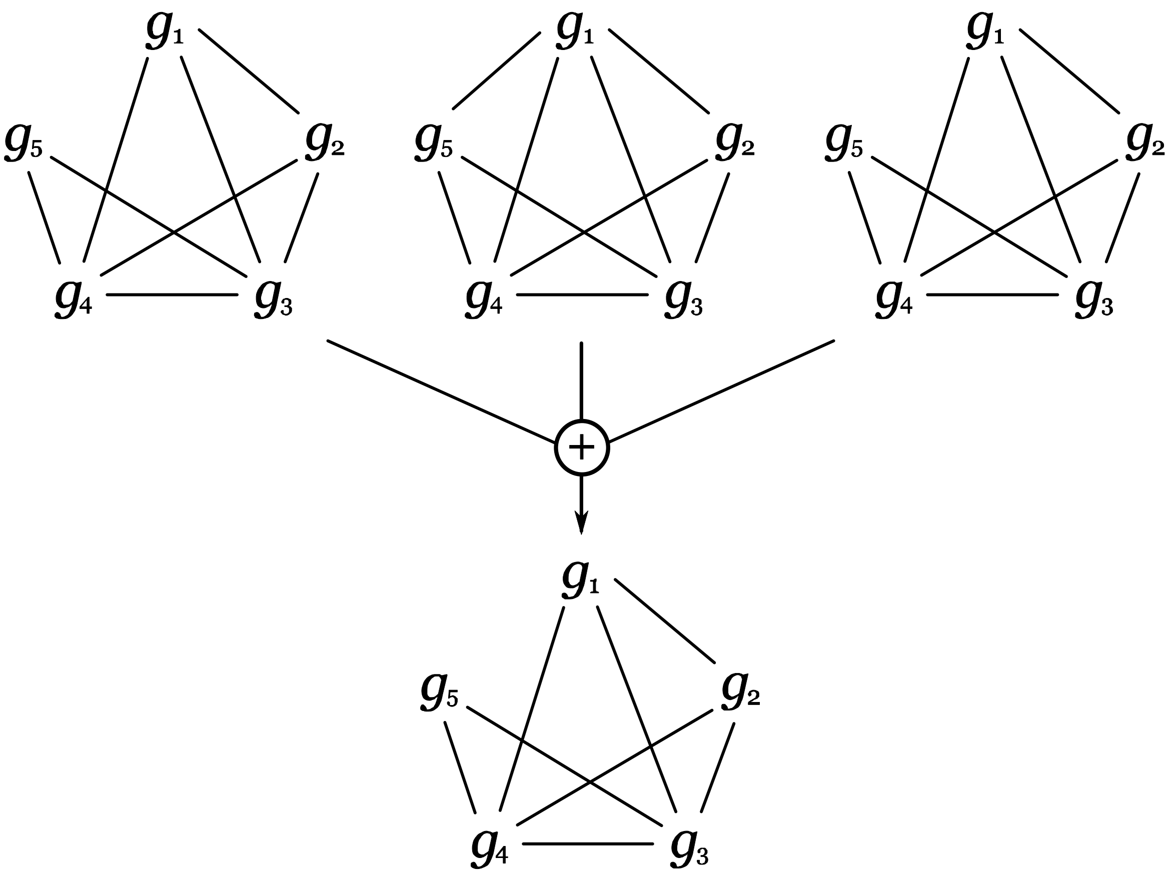
A Novel Measure of Co-expression

$$M_{norm} = \frac{\alpha_1\alpha_2}{|\alpha_1\alpha_2|} \cdot \frac{\cos(|\alpha_1| - |\alpha_2|)}{1 + \cos(\min(|\alpha_1|, |\alpha_2|)}$$

$$BioSim = \frac{1}{N-1}\sum_{i=1}^{N-1} M_{norm_i}.$$



Combining Co-expression Graphs



[2, 3]

Consensus Gene Co-expression Graph

$$N'_1 = (N, A, W_1) \qquad\qquad N'_2 = (N, A, W_2)$$

$$S(N'_1, N'_2) = \frac{1}{|A|} \sum_{\forall i \in N} \sum_{\forall j \in N, i \neq j} 1 - |W_1(i, j) - W_2(i, j)|$$

A consensus gene co-expression graph, ***N'$_c$* = (N, A, W$_c$)**, of a set of *n* graphs **{*N'$_1$* = (N, A, W$_1$), *N'$_2$* = (N, A, W$_2$), . . . , *N'$_n$* = (N, A, W$_n$)}**, is defined to be a graph for which

$$\prod_{i=1}^{n} S(N'_i, N'_c)$$ becomes maximum.

$$\mathbf{W_c(i, j) = \sqrt[\alpha]{\sum_{k=1}^{n} \xi_k(i, j) W_k(i, j)^{\alpha}}} \qquad \mathbf{\xi_k(i, j) = \frac{\#Condition_k}{\sum_{k=1}^{n} \#Condition_k}}$$

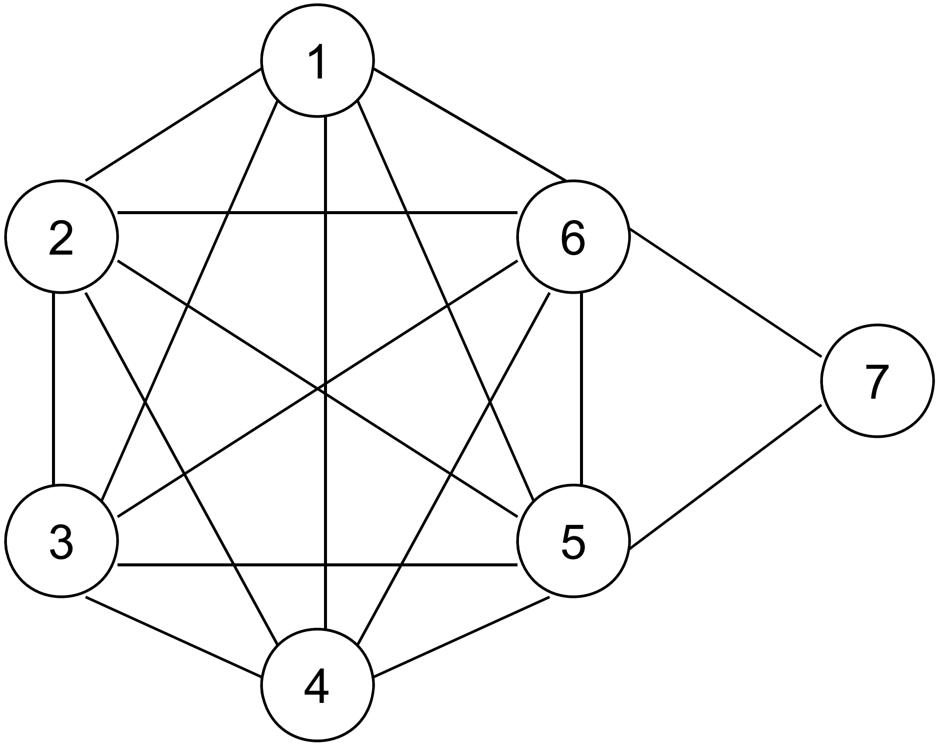
Co-expression Graph Mining

Co-expression Graph Mining

- Dense subgraphs in gene co-expression networks based on conventional definition of density
  - Requires relaxation in the definition of density
  - Significant participation of all the members should be considered
- Quasi-cliques in unweighted graphs
- Dense cores of autonomous systems in communication networks
  - What about unweighted graphs?
- CLIQUE-like problems on graphs
  - On finding all the dense groups
  - study specific to scale-free graphs

Generation of False DAV

Density of the graph = 2.(15+2)/[7.(7–1)] ~ 0.8



* Participation density of 7 is = 2/6 ~ 0.3

[4]

Statement of *MAX-DAV*

Given a weighted graph,

$$\tilde{G} = (V, \tilde{E}, \Omega)$$

and an *association density* threshold of an N-vertexlet δ, locate a dense N-vertexlet,

$$V_{let}^{N_{\max}}$$

that has the maximum cardinality, i.e.,

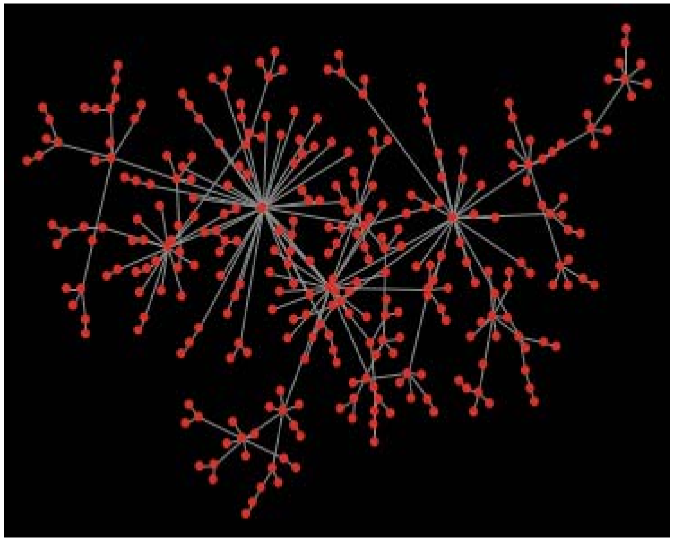
$$N_{\max} \geq N_i : \forall \mu_{V_{let}}^{N_i} \geq \delta, \forall N_i = \{1,2,\ldots,|V|\}$$

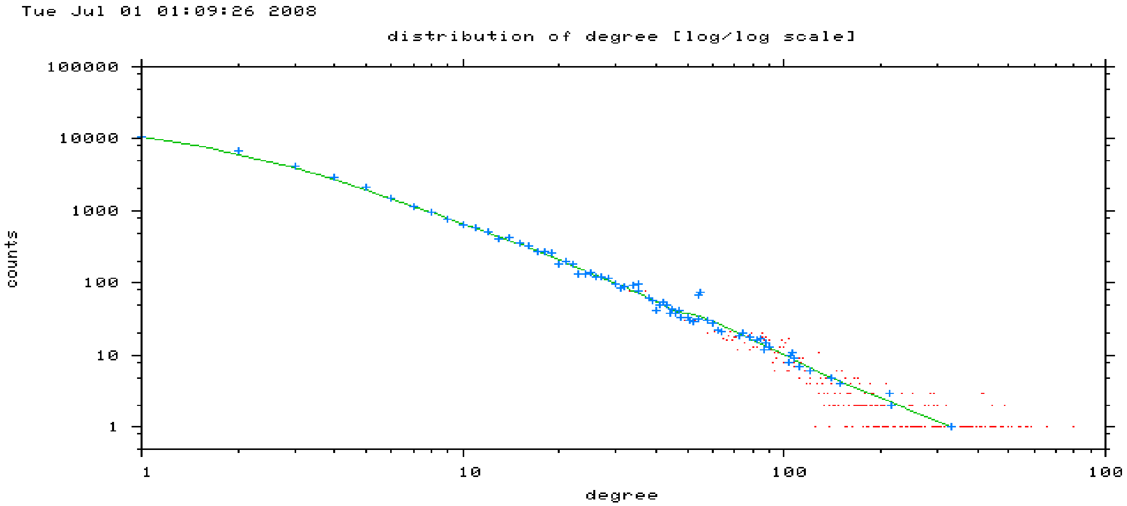
[5, 6]

Scale-free Graphs

- Degree distribution

$$P(k) \sim k^{-\gamma}; \gamma > 0$$





[7]

Equivalence with Quadratic 0-1 Programming Problem

- *Theorem* 3: The *MAX-DAV* in a weighted graph,

$$\tilde{\boldsymbol{G}} = (\boldsymbol{V}, \tilde{\boldsymbol{E}}, \boldsymbol{\Omega})$$

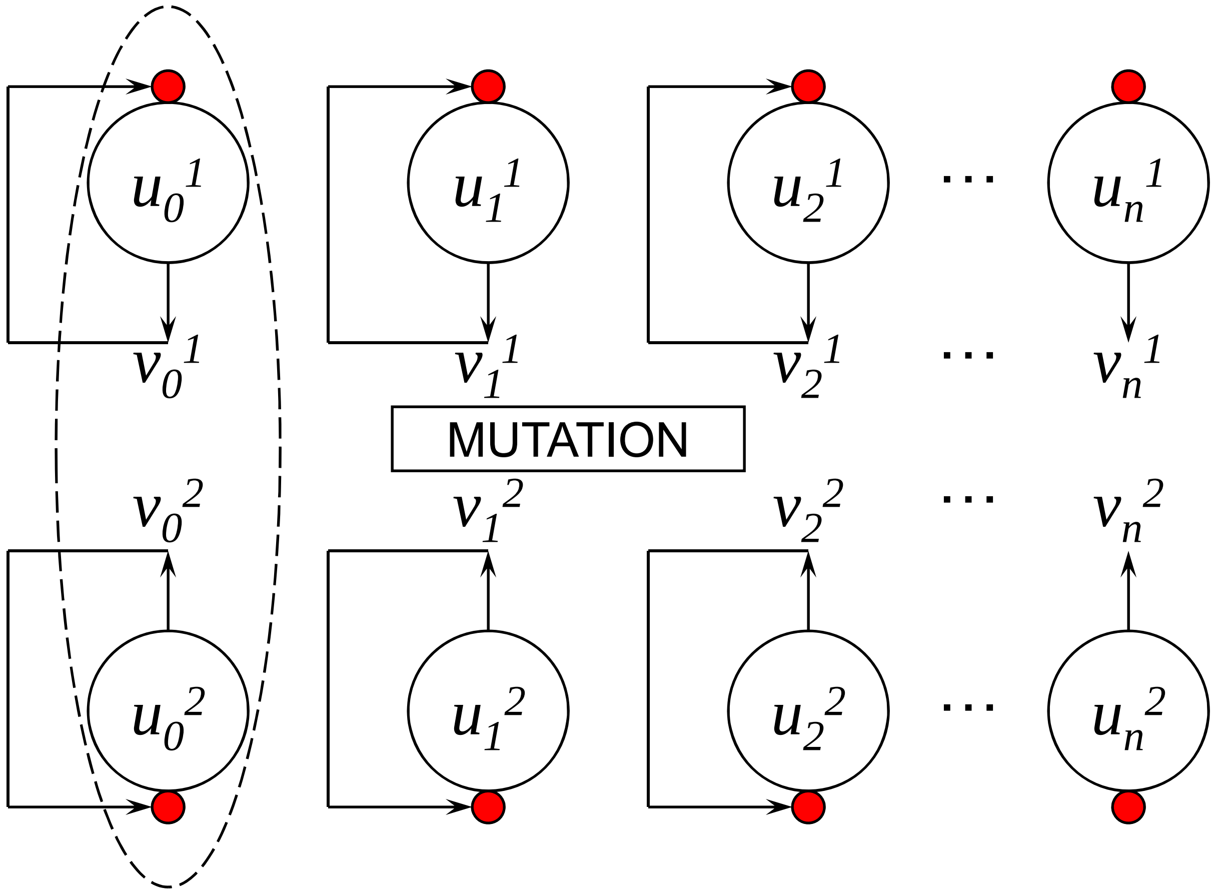
for a given association density threshold *δ*, is equivalent to the following optimization problem,

$$\min \ Z = \underbrace{-\sum_{i=1}^{|V|} v_i}_{\boxed{\text{Maximizing DAV size}}} + \underbrace{\sum_{i=1}^{|V|} [\sum_{j=1}^{|V|} \delta(\sum_{j=1}^{|V|} v_j - 1) - \sum_{j=1}^{|V|} \Omega_{ij} v_i v_j]}_{\boxed{\text{Satisfying density constraint}}}; \forall v_i, v_j \in \{0,1\}$$

where, $v_i = 1$, if x_i is in the maximum DAV, else $v_i = 0$, $\forall \boldsymbol{x_i} \in \boldsymbol{V}$.

[8, 9]

Schematic Diagram of the Maximum Neural Network Model

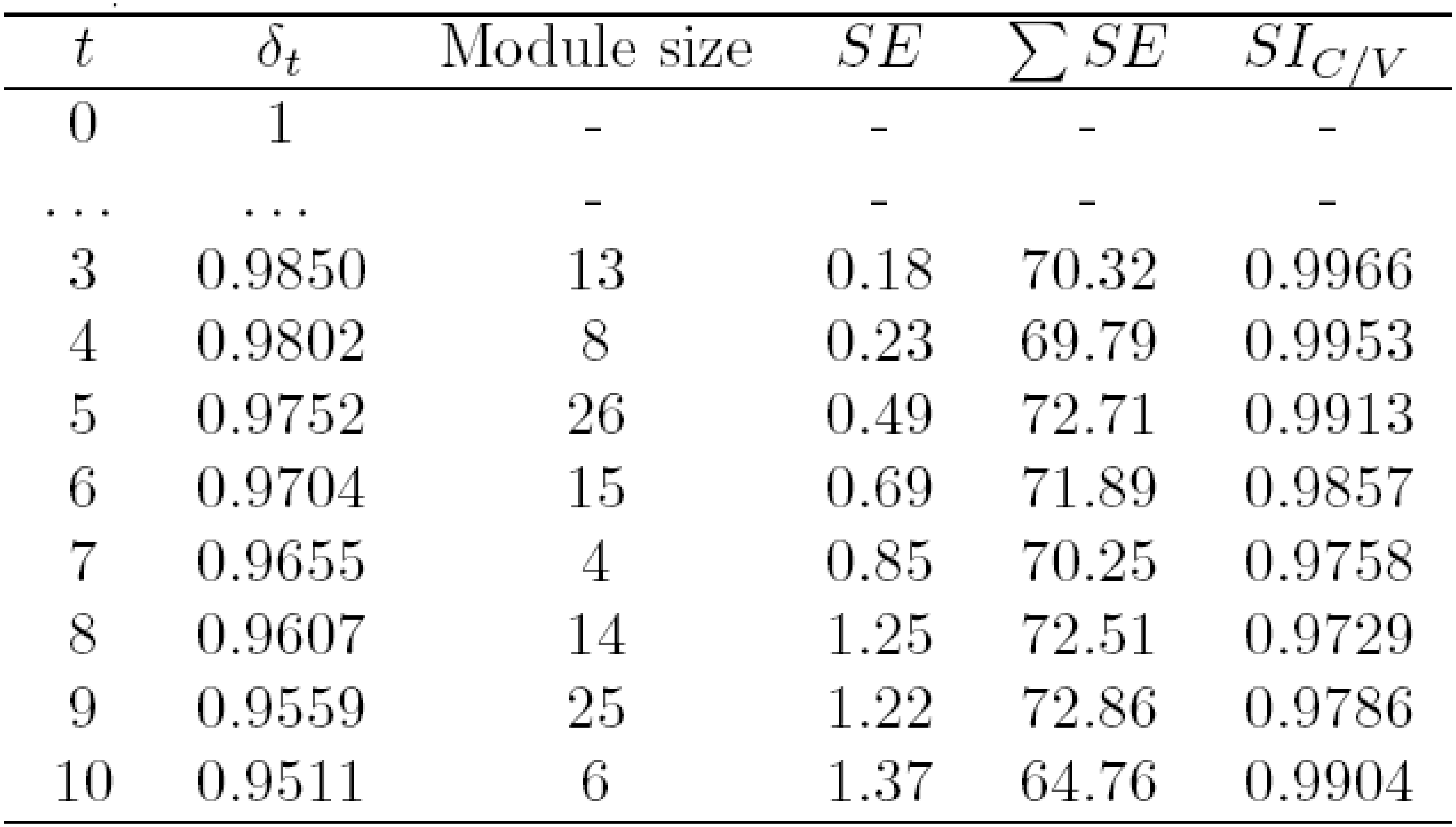


Co-expression to Coregulation

Analysis of MicroRNA Regulation

- Mostly biological analysis
  - Not for an exhaustive collection of miRNAs (<10%)
  - No significant computational analysis based on expression profiling
- Sequence-based analyses based on the construction of positional weight matrices
  - How to establish a correspondence between co-expression, co-functionality and co-regulation?

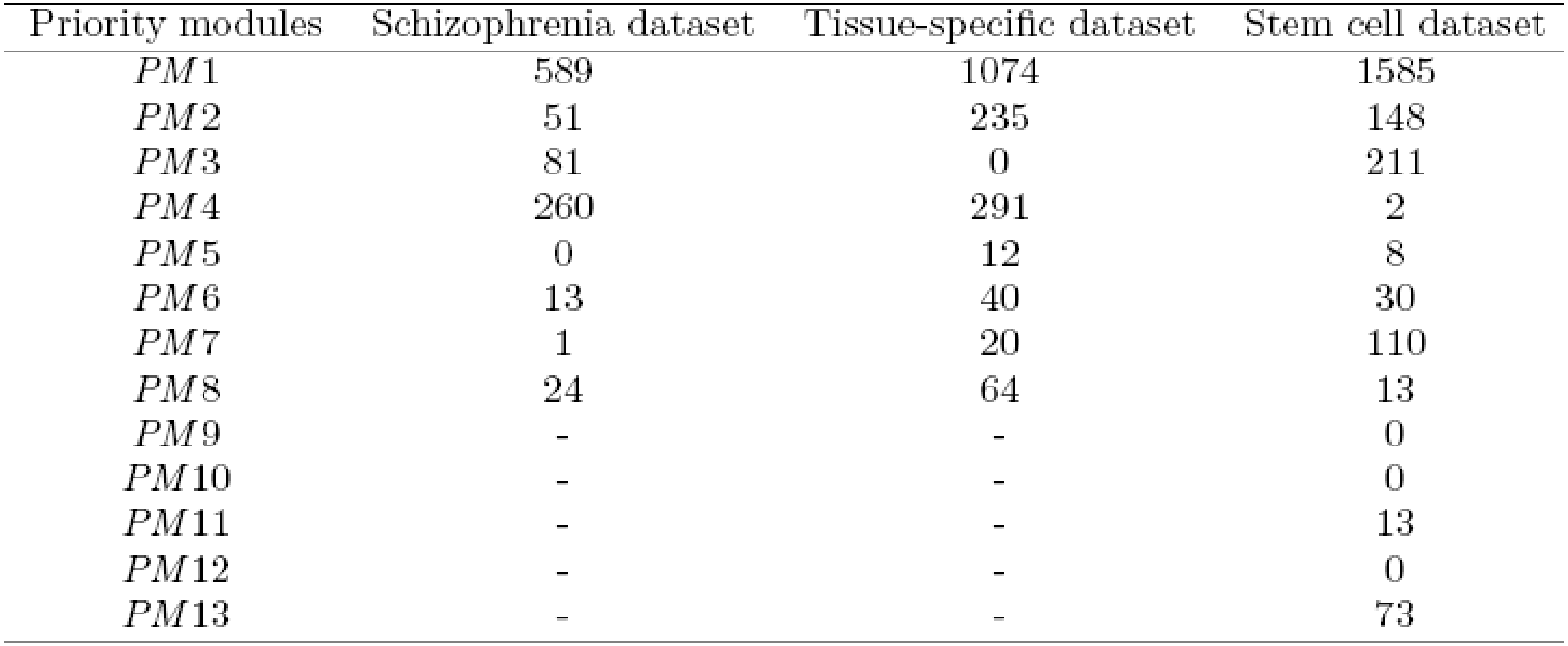
Studies on miRNAs (Schizophrenia Dataset)



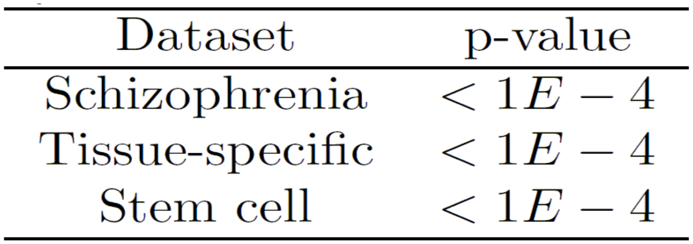
t	δ_t	Module size	SE	$\sum SE$	$SI_{C/V}$
0	1	-	-	-	-
...	...	-	-	-	-
3	0.9850	13	0.18	70.32	0.9966
4	0.9802	8	0.23	69.79	0.9953
5	0.9752	26	0.49	72.71	0.9913
6	0.9704	15	0.69	71.89	0.9857
7	0.9655	4	0.85	70.25	0.9758
8	0.9607	14	1.25	72.51	0.9729
9	0.9559	25	1.22	72.86	0.9786
10	0.9511	6	1.37	64.76	0.9904

[10]

Biological and Statistical Validation



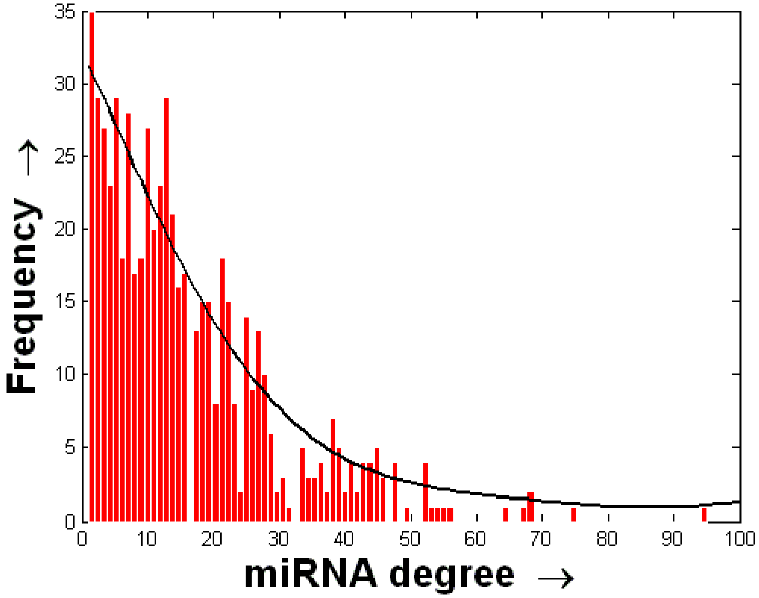
Priority modules	Schizophrenia dataset	Tissue-specific dataset	Stem cell dataset
$PM1$	589	1074	1585
$PM2$	51	235	148
$PM3$	81	0	211
$PM4$	260	291	2
$PM5$	0	12	8
$PM6$	13	40	30
$PM7$	1	20	110
$PM8$	24	64	13
$PM9$	-	-	0
$PM10$	-	-	0
$PM11$	-	-	13
$PM12$	-	-	0
$PM13$	-	-	73

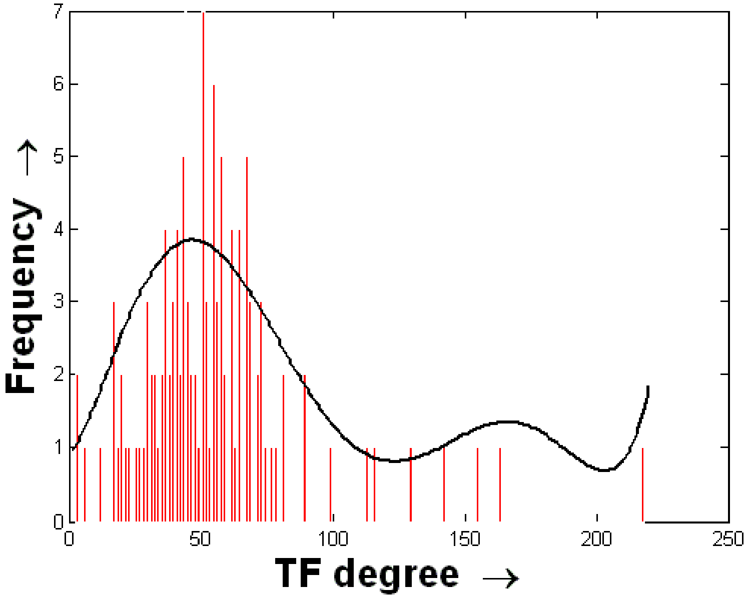


Dataset	p-value
Schizophrenia	$< 1E-4$
Tissue-specific	$< 1E-4$
Stem cell	$< 1E-4$

Degree Distribution in TF-microRNA Interaction Graph

[11]

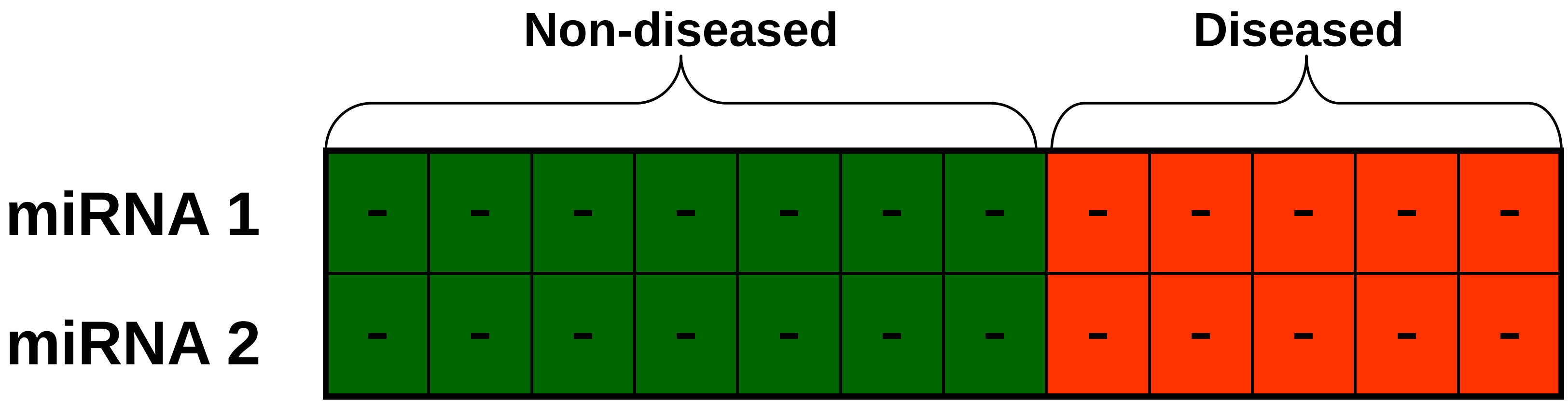




Studying the Differential Co-expression Graphs

[12, 13]

Relative Co-expression Score

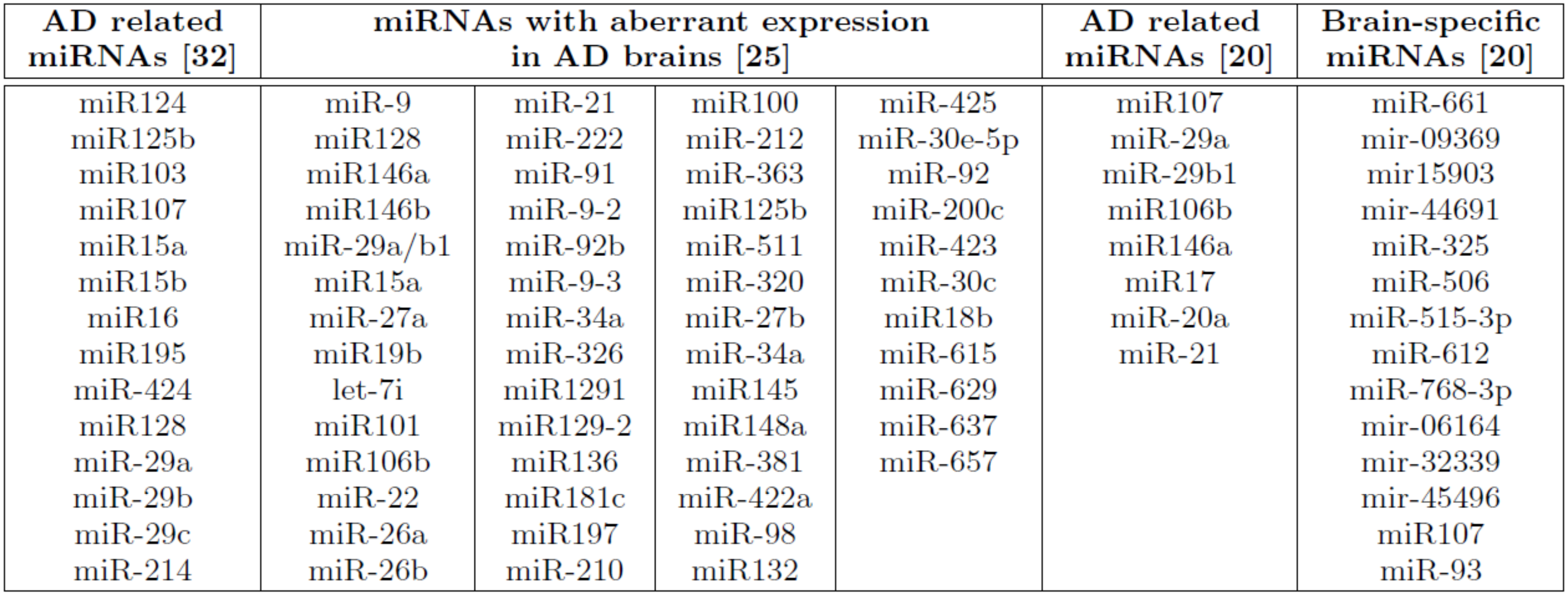


$$RCS\,(X, Y) = \frac{Cor^2(X_{nd} @ X_d, Y_{nd} @ Y_d)}{\varepsilon + Cor^2(X_d @ Y_d)}, \text{ for } X \neq Y$$

$$= 0, \text{ otherwise}$$

Here, the constant $\varepsilon > 0$ is incorporated to map the range of *RCS* from $[0, \infty]$ to $[0, 1/\varepsilon]$ and *Cor* (*X*, *Y*) denotes the Pearson correlation coefficient between *X* and *Y*.

Association of MicroRNAs with Alzheimer’s Disease



AD related miRNAs [32]	**miRNAs with aberrant expression in AD brains [25]**				**AD related miRNAs [20]**	**Brain-specific miRNAs [20]**
miR124	miR-9	miR-21	miR100	miR-425	miR107	miR-661
miR125b	miR128	miR-222	miR-212	miR-30e-5p	miR-29a	mir-09369
miR103	miR146a	miR-91	miR-363	miR-92	miR-29b1	mir15903
miR107	miR146b	miR-9-2	miR125b	miR-200c	miR106b	mir-44691
miR15a	miR-29a/b1	miR-92b	miR-511	miR-423	miR146a	miR-325
miR15b	miR15a	miR-9-3	miR-320	miR-30c	miR17	miR-506
miR16	miR-27a	miR-34a	miR-27b	miR18b	miR-20a	miR-515-3p
miR195	miR19b	miR-326	miR-34a	miR-615	miR-21	miR-612
miR-424	let-7i	miR1291	miR145	miR-629		miR-768-3p
miR128	miR101	miR129-2	miR148a	miR-637		mir-06164
miR-29a	miR106b	miR136	miR-381	miR-657		mir-32339
miR-29b	miR-22	miR181c	miR-422a			mir-45496
miR-29c	miR-26a	miR197	miR-98			miR107
miR-214	miR-26b	miR-210	miR132			miR-93

Differential Co-expression Analyses

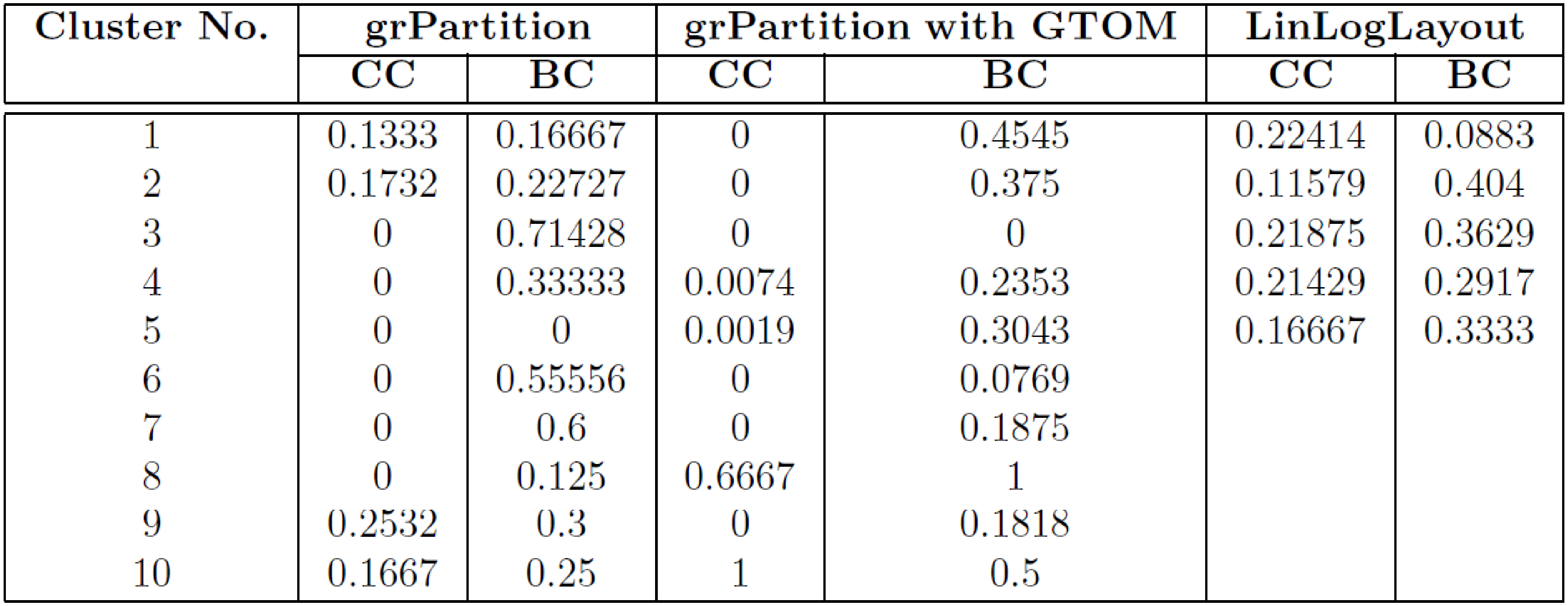
Method	**Gray Matter**		**White Matter**	
	# AD related microRNAs	*p*-value	# AD related microRNAs	*p*-value
Student's paired t-test	8 (15)	3.74E-02	4 (15)	6.96E-01
SAM	5 (15)	4.66E-01	3 (15)	8.77E-01

* Among the top 15 microRNAs selected by the respective analysis

Method	**No constraint**		**Excluding miR-423-5p**	
	# AD related microRNAs	*p*-value	# AD related microRNAs	*p*-value
Correlation-based analysis	11 (16)	7.44E-04	5 (17)	4.66E-01

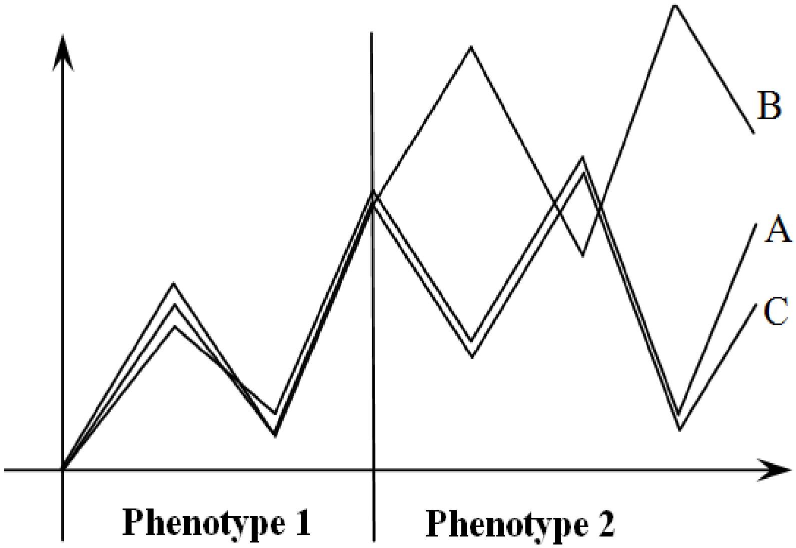
* Among the top 15 pairs of microRNAs selected by the respective analysis

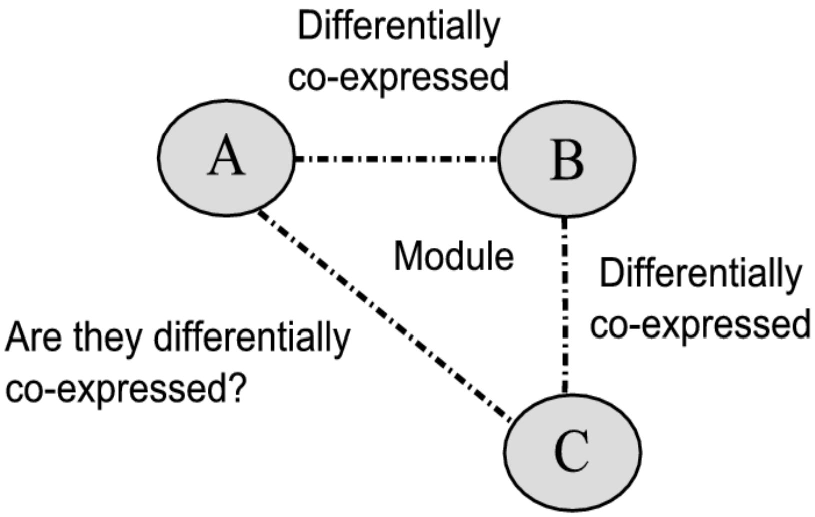
Comparative Studies on Different Graph Clustering Algorithms



Cluster No.	grPartition		grPartition with GTOM		LinLogLayout	
	CC	BC	CC	BC	CC	BC
1	0.1333	0.16667	0	0.4545	0.22414	0.0883
2	0.1732	0.22727	0	0.375	0.11579	0.404
3	0	0.71428	0	0	0.21875	0.3629
4	0	0.33333	0.0074	0.2353	0.21429	0.2917
5	0	0	0.0019	0.3043	0.16667	0.3333
6	0	0.55556	0	0.0769		
7	0	0.6	0	0.1875		
8	0	0.125	0.6667	1		
9	0.2532	0.3	0	0.1818		
10	0.1667	0.25	1	0.5		

Whether Differential Co-expression Patterns Do Exist within a Module?



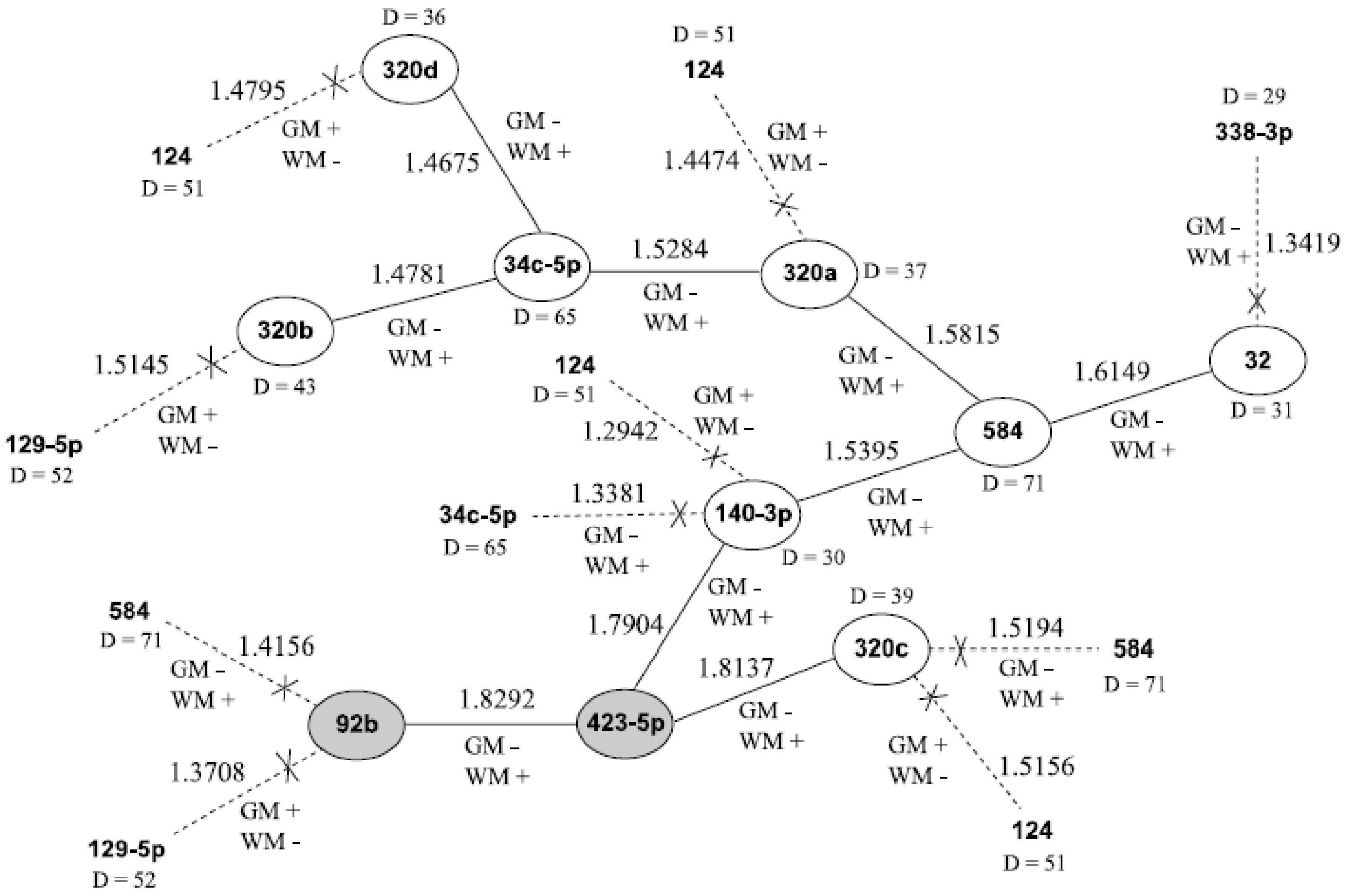

Differentially co-expressed
A
B
Module
Differentially co-expressed
Are they differentially co-expressed?
C

[14]

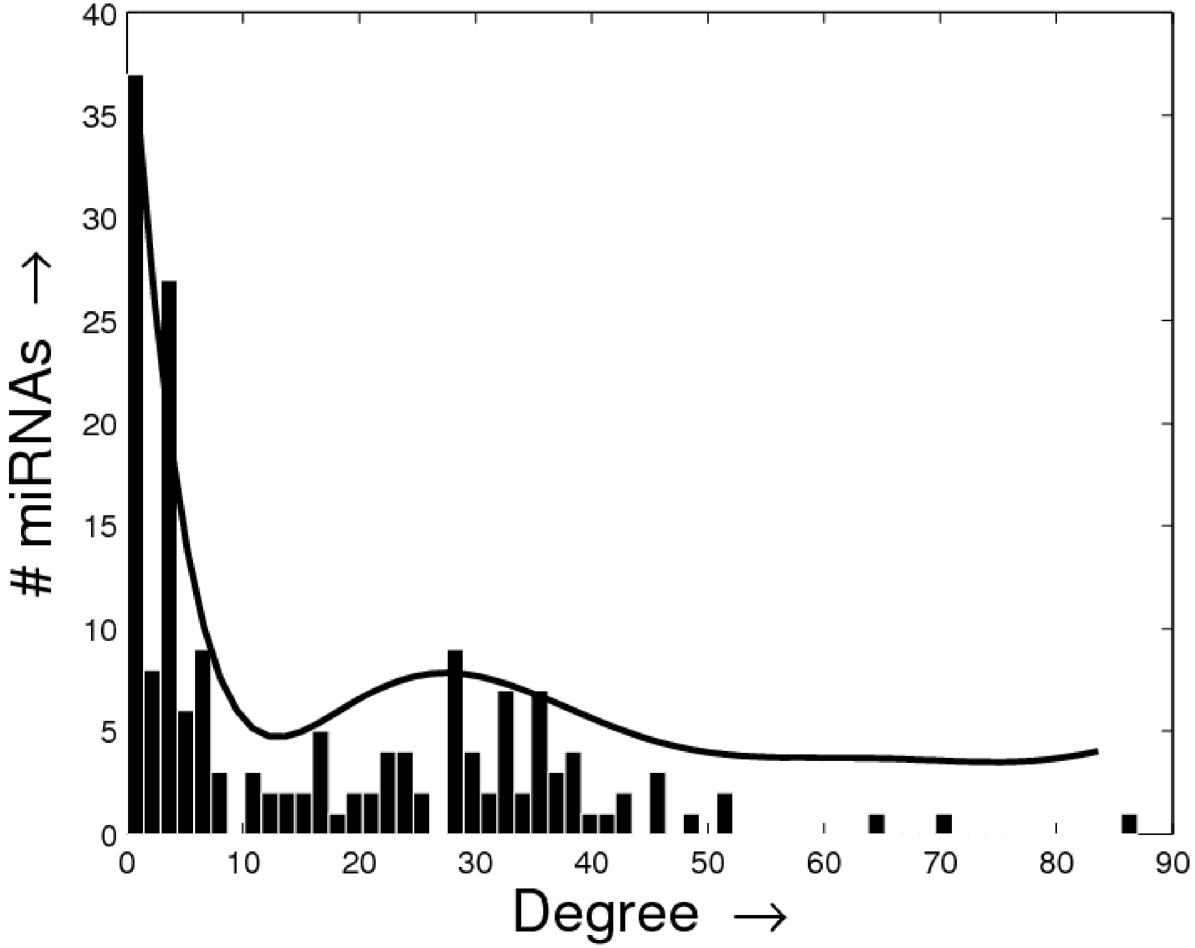
The Proposed Approach

```
Input: A differentially co-expressed graph $G = (V, E, W, S)$, a strict lower threshold of differential co-expression value $T$,
    and a degree threshold $TD$.
Output: The set of largest DCSTs.
Remove the edges having weight $W \leq T$.
repeat
    Remove the nodes from the reduced graph having no connectivity with the others.
    repeat
        Find the edge having the strongest weight (seed edge) and initialize it as a DCST.
        Find the switching pattern of the seed edge.
        Find the nodes $s_i$ forming the strongest edges with the nodes $i$ in the DCST such that none of them is
        present in the DCST. Find the degree values of the $s_i$'s.
        Select the node $s_i$ to expand the DCST further by the inclusion of the edge $(i, s_i)$ such that it possesses:
            a. comparatively higher weight than the other edges.
            b. degree value greater than $TD$.
            c. same switching pattern as that of the seed edge.
        Resolve the conflicts in 7.a by randomization.
    until The current DCST is no more expandable by the inclusion of edges
    if the DCST is empty then
        Exit.
    else
        Return the current DCST and remove its belonging nodes from the original graph $G$.
    end if
until The reduced graph is empty
```

Construction of the Differentially Co-expressed Switching Tree



Frequency Distribution of the Degree Values of MicroRNAs

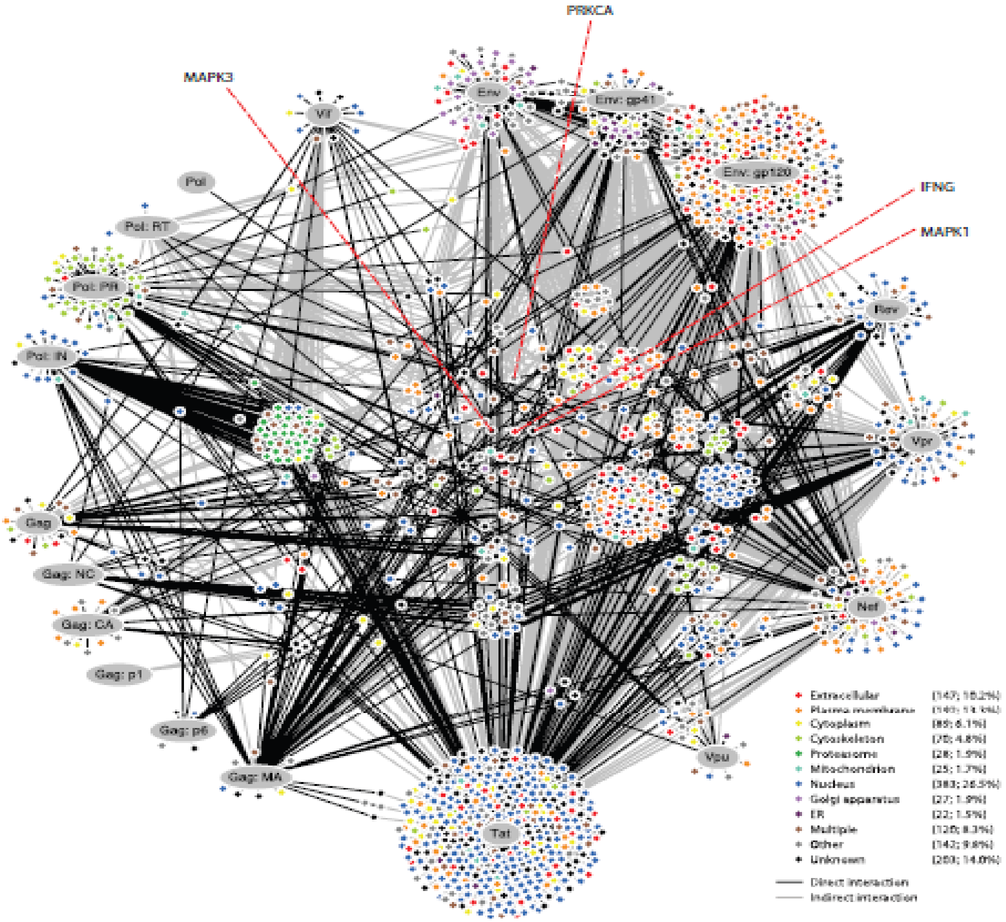


Disease Analysis

Disease Analysis

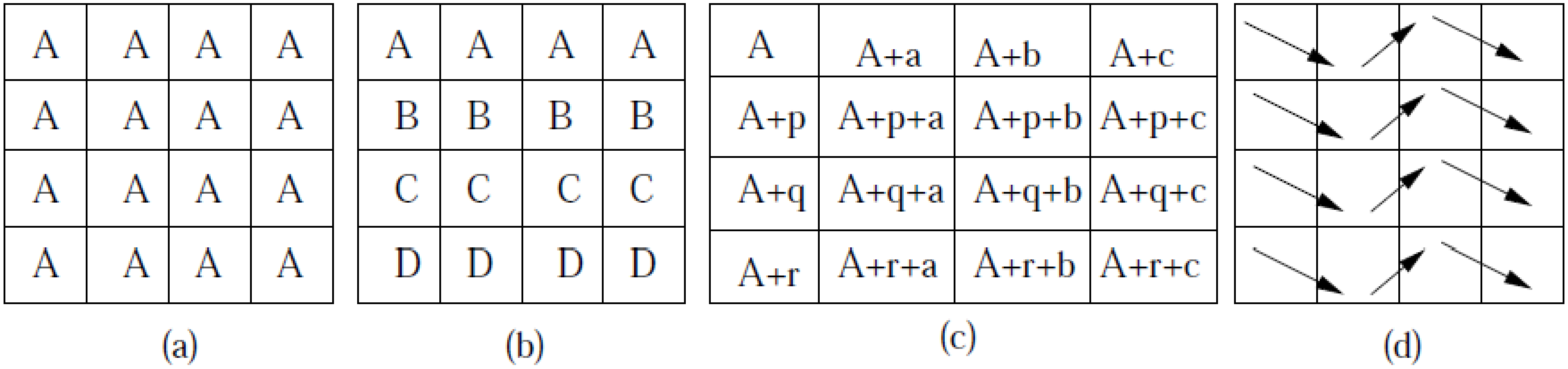
- Differential expression pattern analysis based on different phenotypes and mostly biological
  - How to define co-expression/differential co-expression/co-expression dynamics
- Network based analysis
  - System level analyses are problem-specific

HIV-1–Human Protein Interaction Network



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

Biclustering Approaches



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)

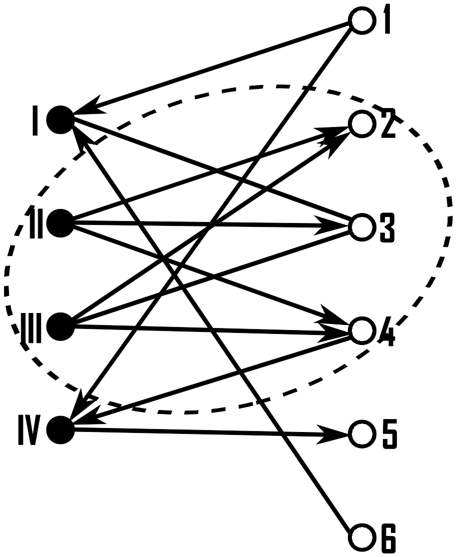
(d)

- Biclustering approaches
  - Cheng and Church's algorithm (CCA), SAMBA, Co-clustering algorithm (CA), Divide-and-conquer based algorithm (DBA)
- Bicluster 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)

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

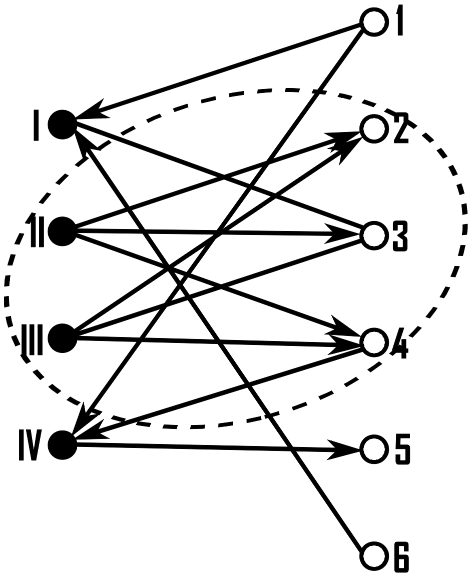
$$G = (V1, V2, E)$$



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

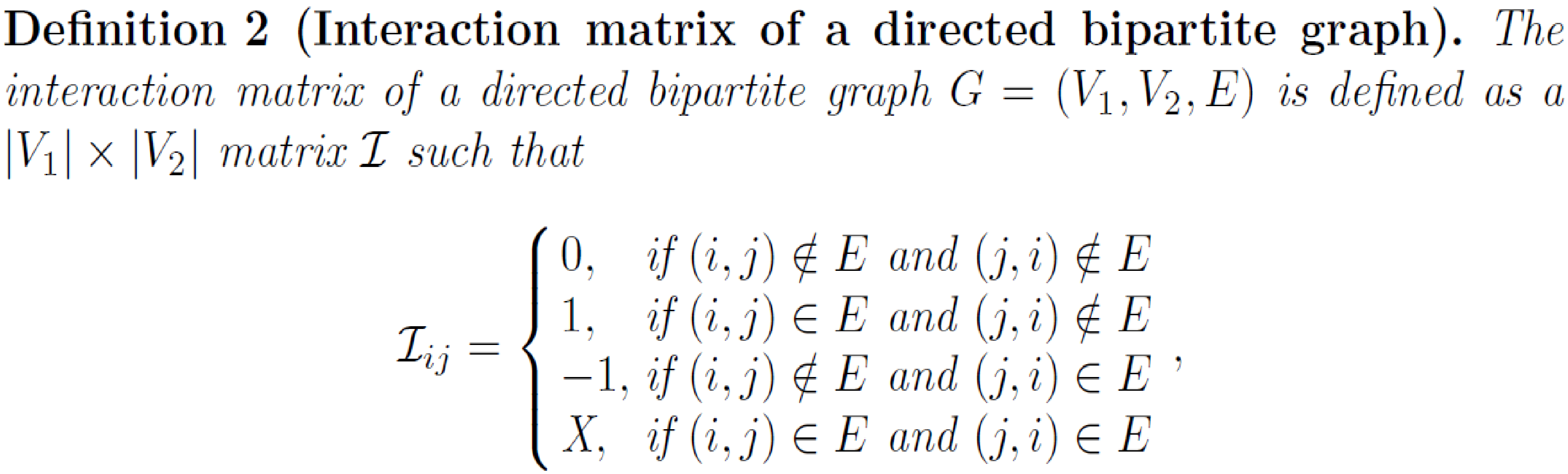
Definition 1 (DBClique). *A DBClique is a fully connected subgraph* $G' = (V_1', V_2', E') \subseteq G$ *of a directed bipartite graph* G *such that either* $i \in V_1', j \in V_2', \forall(i,j) \in E'$ *or* $i \in V_2', j \in V_1', \forall(i,j) \in E'$.

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

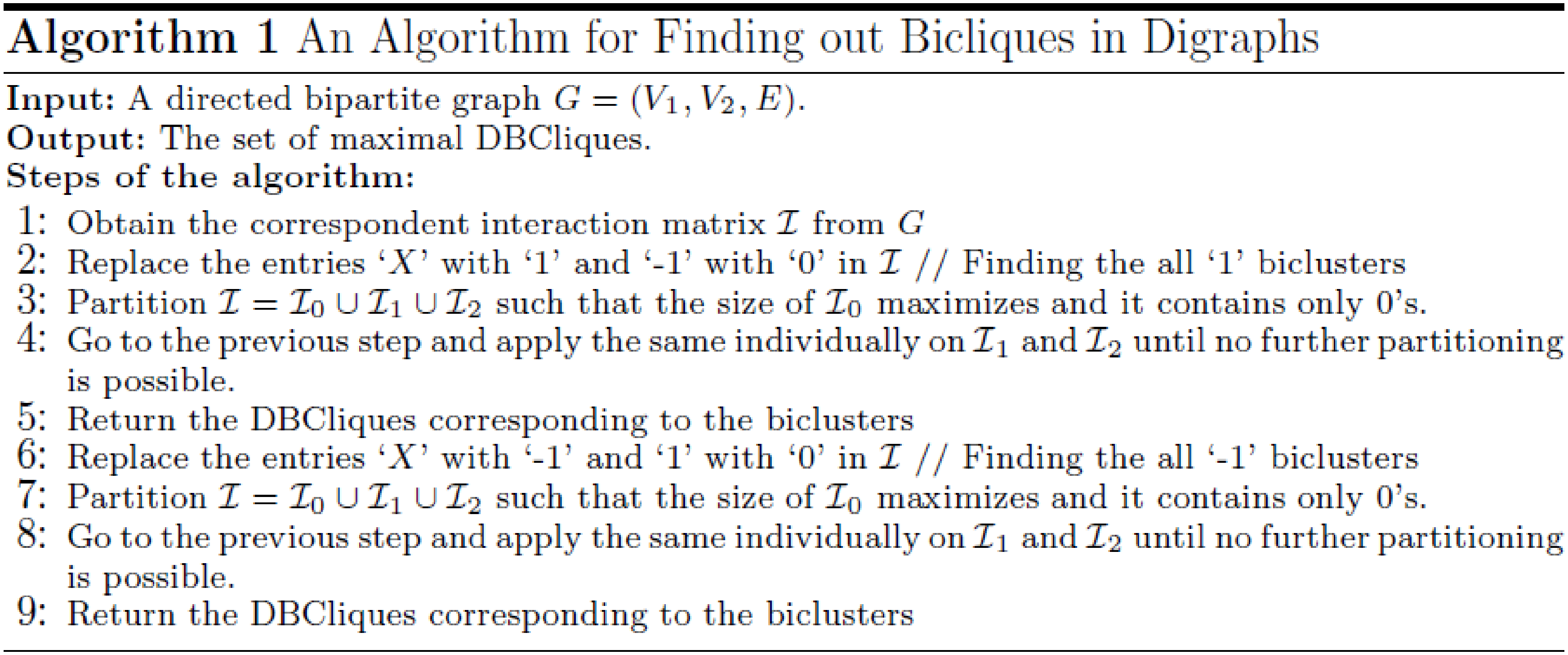


Definition 2 (Interaction matrix of a directed bipartite graph). *The interaction matrix of a directed bipartite graph* $G = (V_1, V_2, E)$ *is defined as a* $|V_1| \times |V_2|$ *matrix* $\mathcal{I}$ *such that*

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

[15]

The Approach



Algorithm 1 An Algorithm for Finding out Bicliques in Digraphs

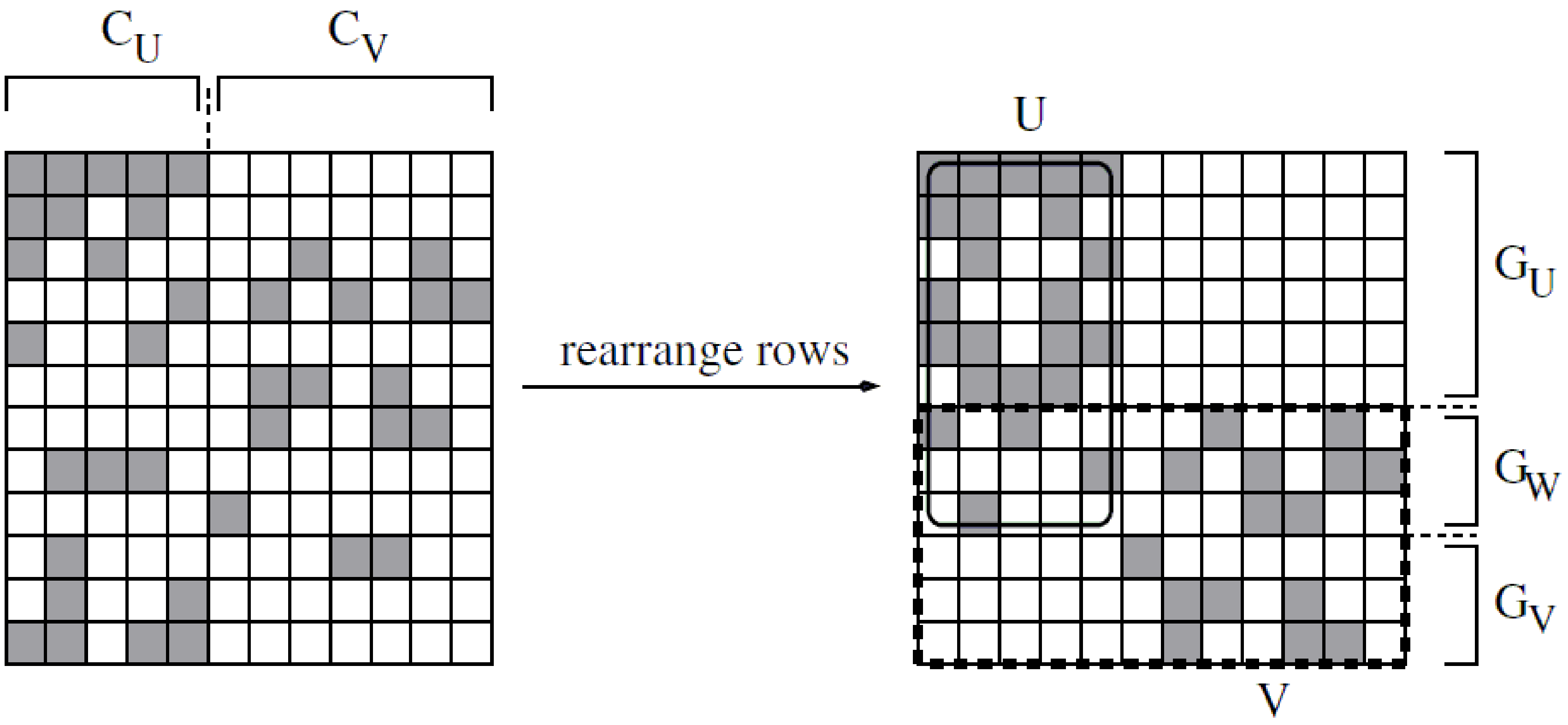
Input: A directed bipartite graph $G = (V_1, V_2, E)$.

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



* Figure taken from [21]

Details of the Data and the DBCliques Obtained

- 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)

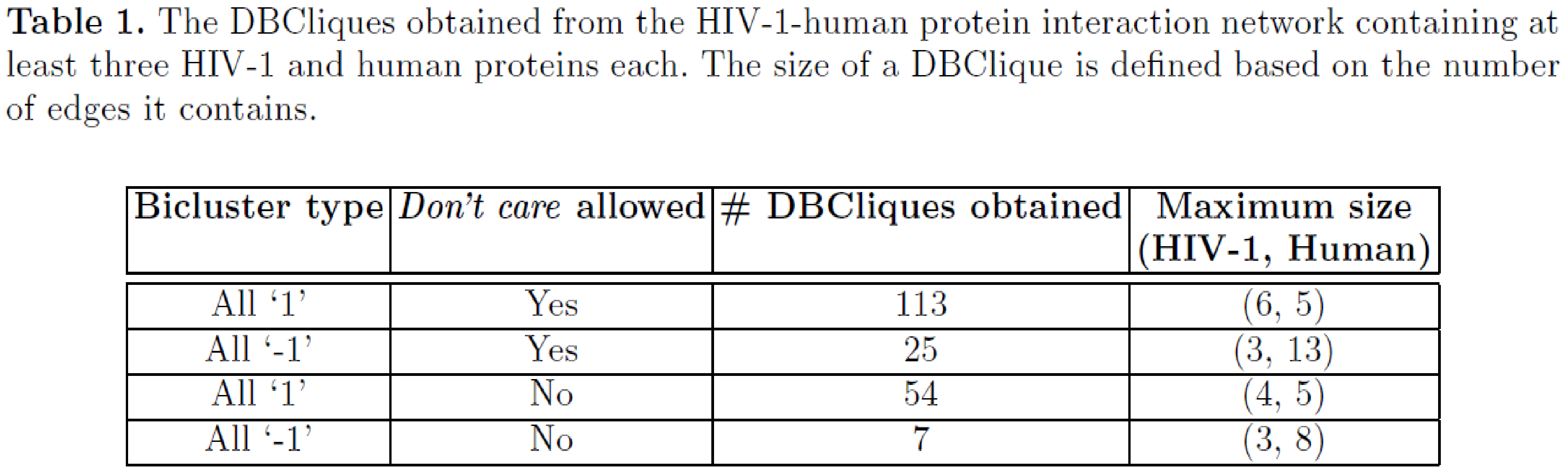


Table 1. The DBCliques obtained from the HIV-1-human protein interaction network containing at least three HIV-1 and human proteins each. The size of a DBClique is defined based on the number of edges it contains.

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

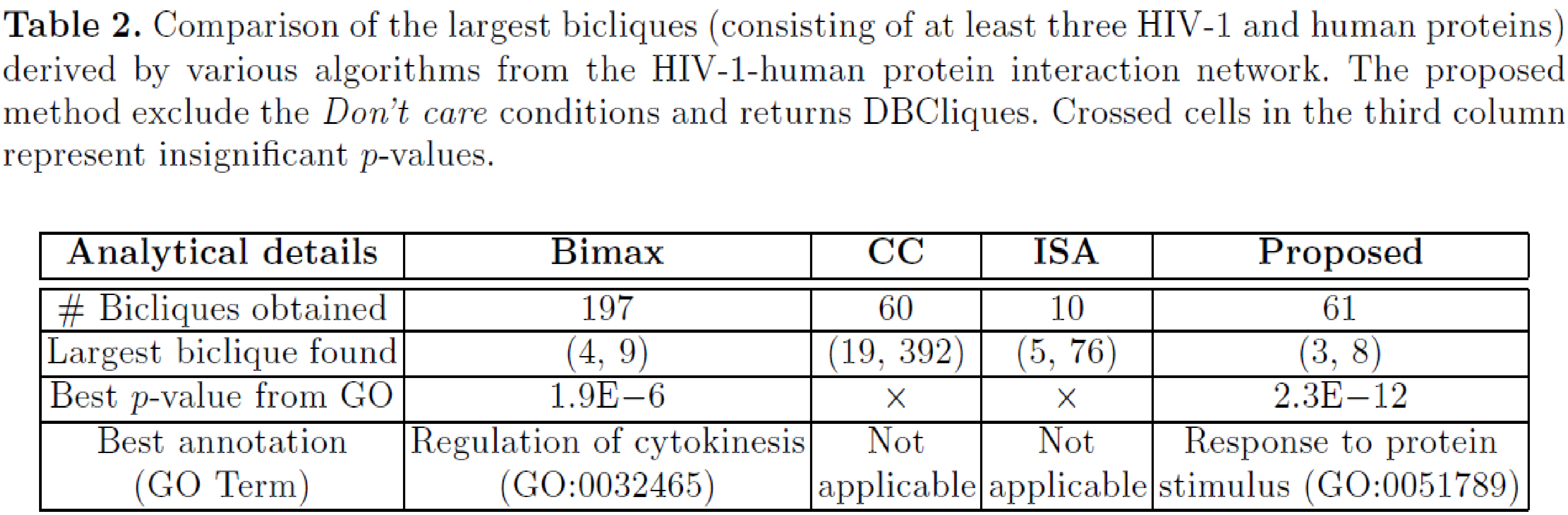


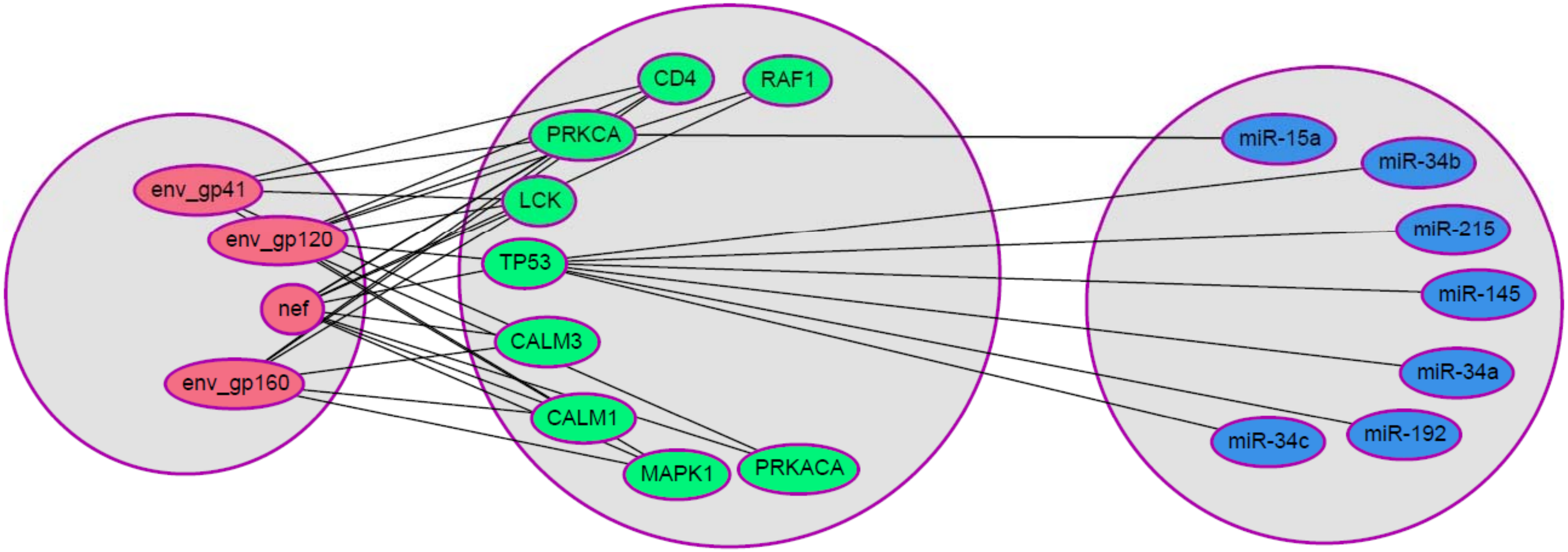
Table 2. Comparison of the largest bicliques (consisting of at least three HIV-1 and human proteins) derived by various algorithms from the HIV-1-human protein interaction network. The proposed method exclude the *Don't care* conditions and returns DBCliques. Crossed cells in the third column represent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

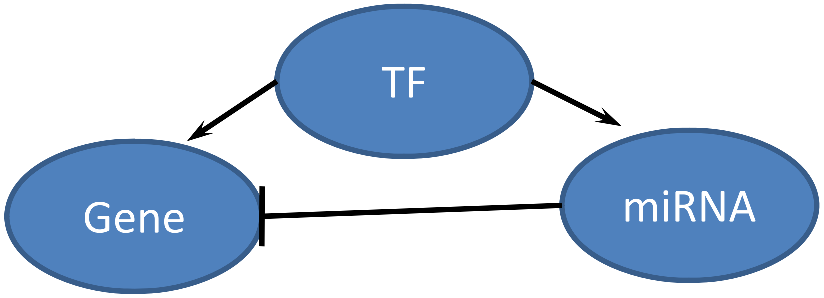
Analytical details	**Bimax**	**CC**	**ISA**	**Proposed**
# Bicliques obtained	197	60	10	61
Largest biclique found	(4, 9)	(19, 392)	(5, 76)	(3, 8)
Best p-value from GO	1.9E−6	×	×	2.3E−12
Best annotation (GO Term)	Regulation of cytokinesis (GO:0032465)	Not applicable	Not applicable	Response to protein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

Future Goals

[16]

Extended Regulation Including MicroRNAs





Major references

1. S. Bandyopadhyay and M. Bhattacharyya, A Biologically Inspired Measure for Co-expression Analysis, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8(4), pp. 929-942, 2011.
2. M. Bhattacharyya and S. Bandyopadhyay. Integration of Co-expression Networks for Gene Clustering. In *Proceedings of the 7th International Conference on Advances in Pattern Recognition*, pp. 355–358, Kolkata, India, 2009.
3. M. Bhattacharyya and S. Bandyopadhyay. Information Fusion in Bioinformatics. *Technorama*, The Institution of Engineers (India), Summer-Winter 2008-09:21–25, 2009.
4. S. Bandyopadhyay and M. Bhattacharyya. Mining the Largest Dense Vertexlet in a Weighted Scale-free Graph. *Fundamenta Informaticae*, 96(1-2):1–25, 2009.
5. M. Bhattacharyya and S. Bandyopadhyay. Analyzing Topological Properties of Protein-protein Interaction Networks: A Perspective towards Systems Biology. In *Computational Intelligence and Pattern Analysis in Biology Informatics*, U. Maulik, S. Bandyopadhyay and J. T. L. Wang (Eds.), John Wiley & Sons, Inc., pp. 349–368, 2010 (ISBN: 978-0-470-58159-9).

Major references (continued)

6. M. Bhattacharyya and S. Bandyopadhyay. Mining the Largest Quasi-clique in Human Protein Interactome. In *Proceedings of the International Conference on Adaptive and Intelligent Systems*, pp. 194–199, Klagenfurt, Austria, 2009.
7. M. Bhattacharyya and S. Bandyopadhyay. A Combinatorial Counterpart of the Maximum Quasi-clique Problem. In *Proceedings of the International Conference on Discrete Mathematics, Algebra and their Applications*, pp. 129–130, Minsk, Belarus, 2009.
8. M. Bhattacharyya and S. Bandyopadhyay, Solving Maximum Fuzzy Clique Problem with Neural Networks and its Applications, *Memetic Computing*, Thematic Issue on “Adaptive Soft Computing Techniques and Applications”, 1(4), pp. 281-290, 2009.
9. S. Bandyopadhyay and M. Bhattacharyya. A Chaotic Neuro-GA Synergism for Solving Maximum Fuzzy Clique Problem, *In Proceedings of the 15th International Conference on Neural Information Processing*, Auckland, New Zealand, November 25–28, pp. 209-210, 2008.
10. S. Bandyopadhyay and M. Bhattacharyya. Analyzing miRNA co-expression networks to explore TF-miRNA regulation. *BMC Bioinformatics*, 10:163, 2009.

Major references (continued)

11. S. Bandyopadhyay and M. Bhattacharyya. PuTmiR: A database for extracting neighboring transcription factors of human microRNAs. *BMC Bioinformatics*, 11:190, 2010.
12. M. Bhattacharyya and S. Bandyopadhyay. Computational Discovery of Different Categories of Human Oncogenic MicroRNAs. In *Proceedings of the 1st IFIP International Conference on Bioinformatics*, no. 94, Surat, India, 2010.
13. S. Bandyopadhyay and M. Bhattacharyya. A Novel Method of Studying the Disease Regulatory Activities of MicroRNAs. *Current Bioinformatics*, 4(3):234–241, 2009.
14. M. Bhattacharyya, S. Bandyopadhyay and U. Maulik, Finding Bicliques in Digraphs: Application into Viral-host Protein Interactome, In Proceedings of the 4th International Conference on Pattern Recognition and Machine Intelligence, Moscow, Russia, June 27-July 01, Springer LNCS 6744, pp. 412-417, 2011.
15. M. Bhattacharyya and S. Bandyopadhyay, Co-expression Toggling of MicroRNAs in Alzheimer's Brain, *SiNAPSA Neuroscience Conference*, Ljubljana, Slovenia, 2011.
16. U. Maulik, M. Bhattacharyya, A. Mukhopadhyay, S. Bandyopadhyay, Identifying the Immunodeficiency Gateway Proteins in Human and their Involvement in MicroRNA Regulation, *Molecular Biosystems*, 7(6), pp. 1842-1851, 2011.

Thank you