Nature Precedings : doi:10.1038/npre.2012.7126.1 : Posted 3 Apr 2012

Integration of Co-expression Networks for Gene Clustering

Malay Bhattacharyya
&
Sanghamitra Bandyopadhyay
Machine Intelligence Unit,
Indian Statistical Institute, Kolkata
Year – 2009

Outline of the presentation

- Motivation
- Related works
- Contribution
- Results

Motivation

Microarray experimentation

- Microarray profiling is prone to noise
- Multi-experimental data integration
- Avoiding missing value estimation

Related works

Probabilistic integration

- Log-likelihood score [Lee *et al*., 2004]

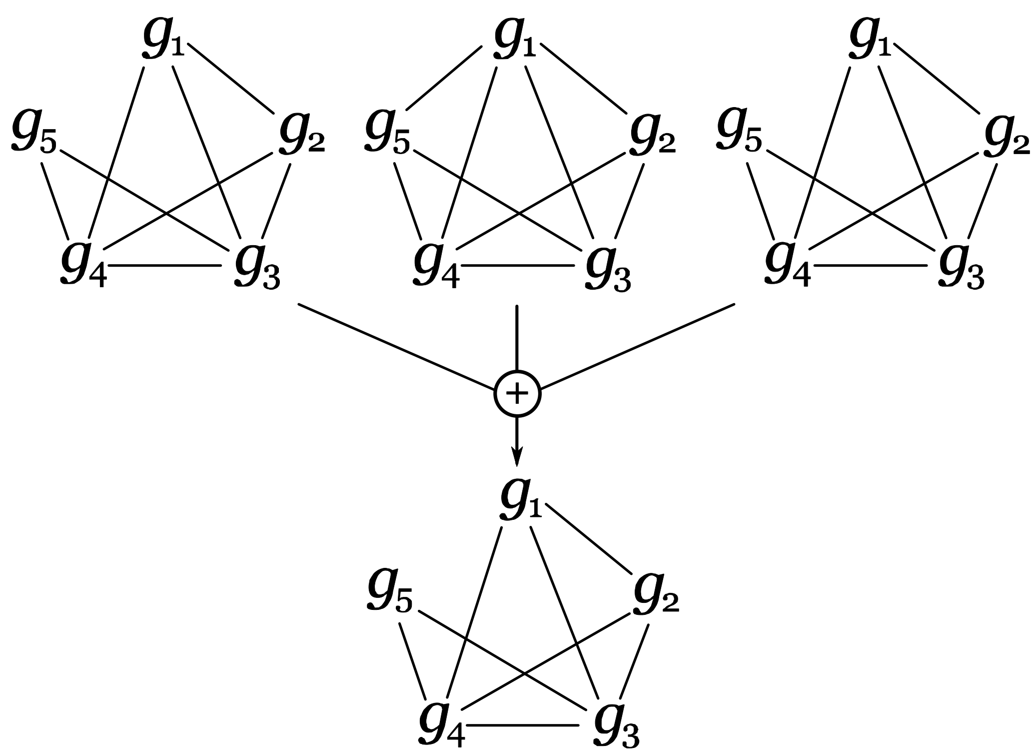
$$LLS = \ln\left(\frac{P(L\mid E)/\sim P(L\mid E)}{P(L)/\sim P(L)}\right)$$

  – Depends on the ground truth knowledge

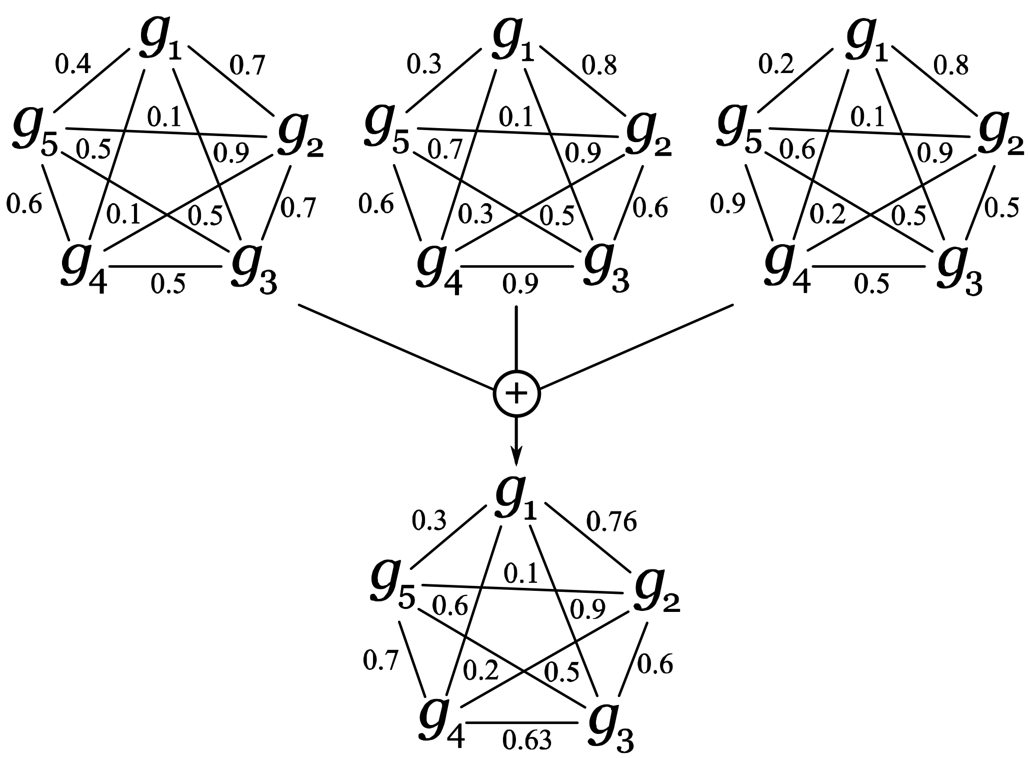
Summary graph approach

- On unweighted graphs [Hu *et al.*, 2005]
  - Frequency count of the edges
- On weighted graphs [Yan *et al.*, 2007]
  - Cutoff threshold on the aggregated weights

Consensus unweighted graph



Consensus weighted graph



Cross platform normalization

- On gene expression studies [Shablin *et al.*, 2008]
  - Merging expression studies

Contribution

Co-expression networks

$$N' = (N, A, W)$$

$$N = \{n_1, n_2, ..., n_{|N|}\}$$

$$A \subseteq N \times N - \bigcup_{i=1}^{|N|} (n_i, n_i)$$

$$W : A \to [0,1]$$

Gene co-expression network similarity

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

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

Characteristics of the similarity measure

$$S(N'_1, N'_2) \in [0,1]$$

$$S(N', N') \in 1$$

$$S(N'_1, N'_2) = S(N'_2, N'_1)$$

Consensus gene co-expression network

A consensus gene co-expression network, $\boldsymbol{N'_c = (N, A, W_c)}$*, of a set of n networks* $\boldsymbol{\{N'_1 = (N, A, W_1), N'_2 = (N, A, W_2), \ldots, N'_n = (N, A, W_n)\}}$*, is defined to be a network having the maximum similarity with the given set of n networks, i.e.,*

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

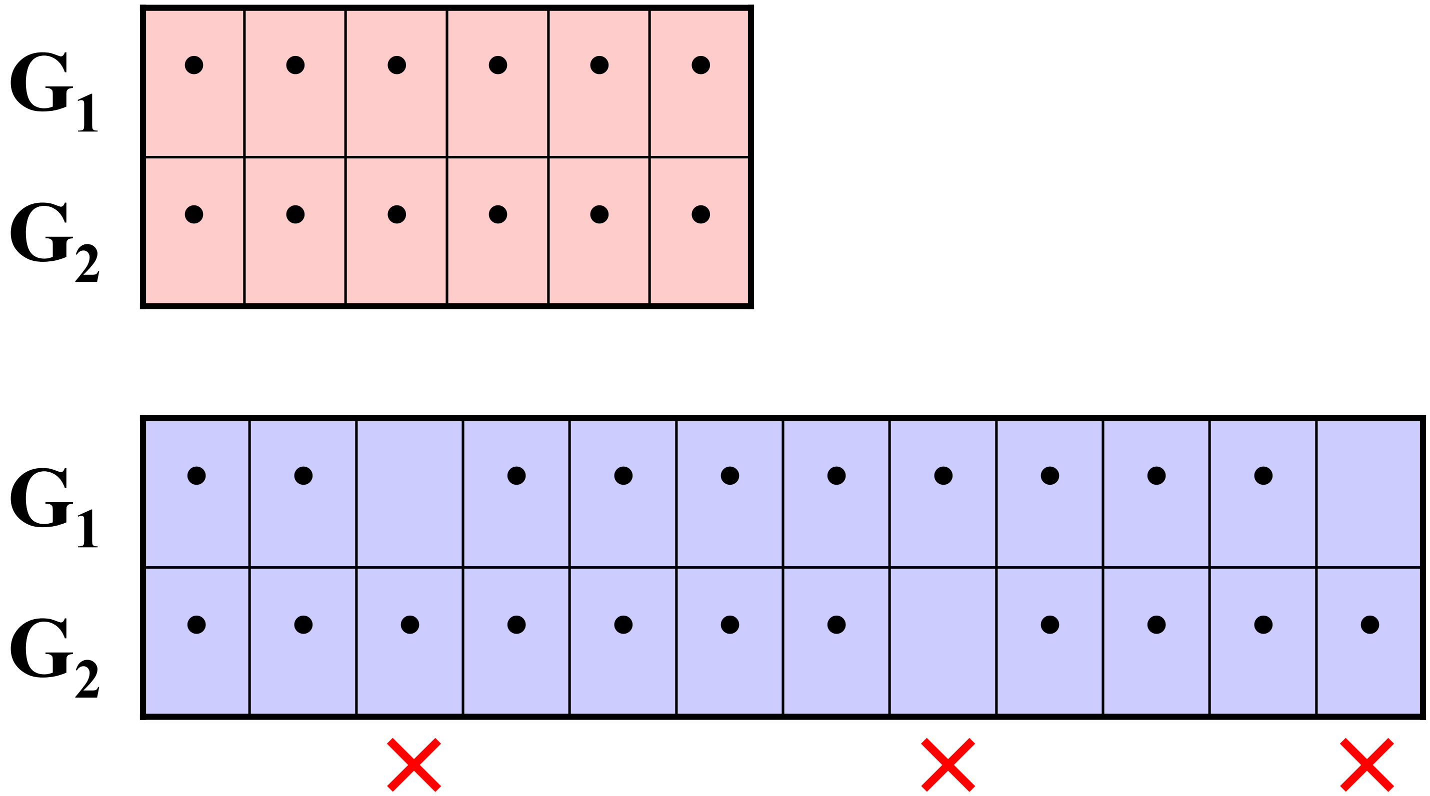
becomes maximum.

Normalized fuzzy integration (NFA)

$$\boldsymbol{W_c(i,j) = \sqrt[\alpha]{\sum_{k=1}^{n} \xi_k(i,j) W_k(i,j)^{\alpha}}}$$

$$\xi_k(i,j) = \frac{\#Condition_k}{\sum_{k=1}^{n} \#Condition_k}$$

Avoiding missing value estimation



Results

Comparative results on the Eisen dataset [Eisen *et al.*, 1998]

Integration Method	**#Clusters**	***SI***	**z-score**
NIL	10	0.035	4.56
AVERAGE	10	−0.057	8.56
NFA ($\alpha = 2$)	10	−0.023	11

Comparative results on the Gasch dataset [Gasch *et al.*, 2001]

Integration Method	**#Clusters**	***SI***	**z-score**
NIL	10	−0.154	0.626
AVERAGE	10	−0.201	1.2
NFA ($\alpha = 2$)	10	−0.19	1.62

Major references

1. I. Lee *et al.* A probabilistic functional network of yeast genes. *Science*, 306:1555–1558, 2004.
2. H. Hu *et al.* Mining coherent dense subgraphs across massive biological networks for functional discovery. *Bioinformatics*, 21:i213–i221, 2005.
3. X. Yan *et al.* A graph-based approach to systematically reconstruct human transcriptional regulatory modules. *Bioinformatics*, 23:i577–i586, 2007.
4. A. Shabalin *et al.* Merging two gene-expression studies via cross-platform normalization. *Bioinformatics*, 24(9):1154–1160, 2008.
5. F. D. Gibbons and F. P. Roth. Judging the quality of gene expression-based clustering methods using gene annotation. *Genome Research*, 12:1574–1581, 2002.
6. M. B. Eisen *et al.*.Cluster analysis and display of genome-wide expression patterns. *Proceedings of National Academy of Sciences*, 95:14863–14868, 1998.
7. A. P. Gasch *et al.* Genomic expression responses to dna-damaging agents and the regulatory role of the yeast atr homolog mec1p. *Molecular Biology of the Cell*, 12:2987–3003, 2001.

Thank you