Annotating Gene Function by Combining Expression Data with a Modular Gene Network

Motoki Shiga, Ichigaku Takigawa,
Hiroshi Mamitsuka

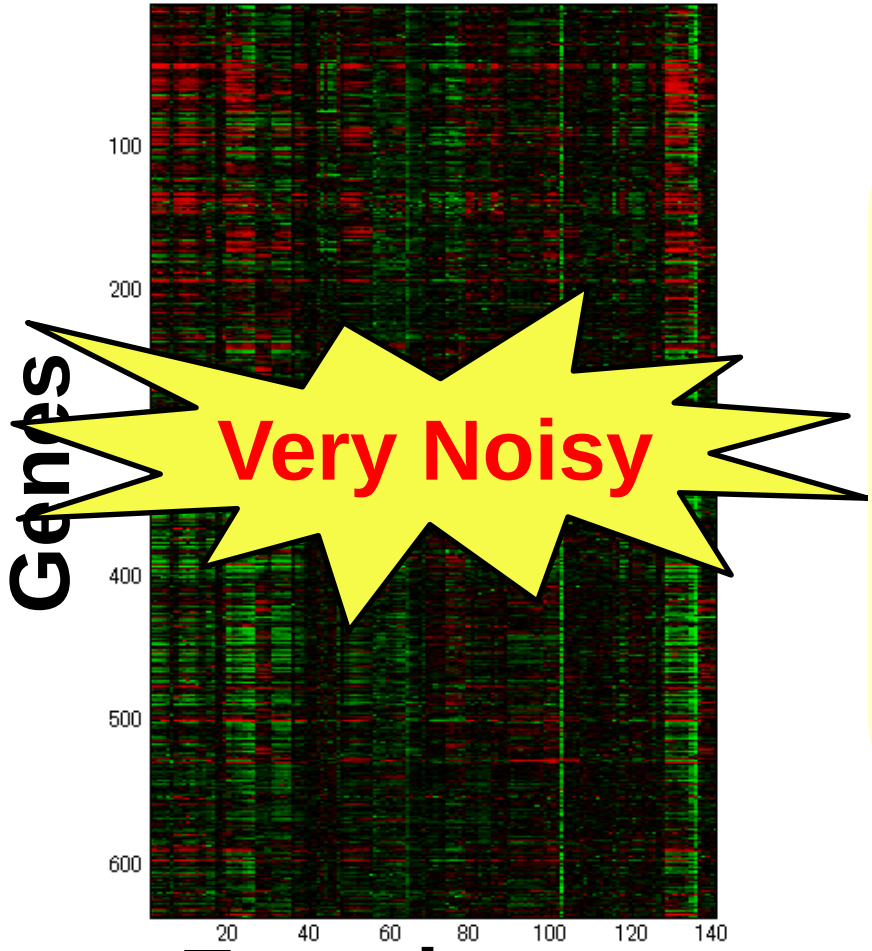
Bioinformatics Center, ICR, Kyoto University, Japan

Table of Contents

Gene Clustering using Expression Data

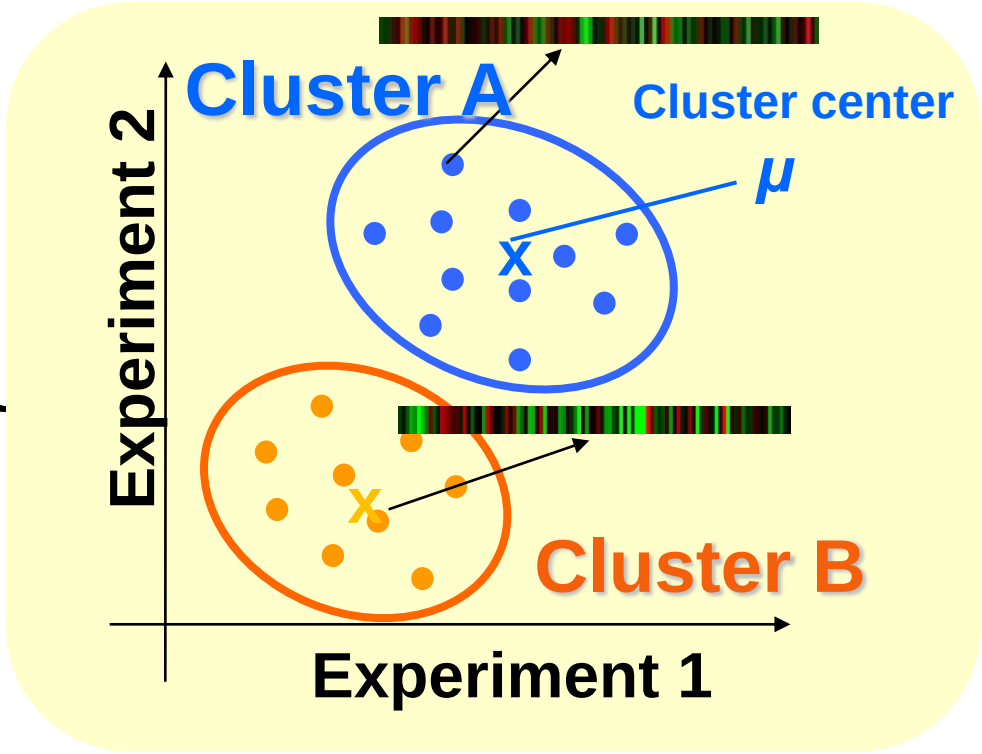
Expression Data

(Gasch, et al., Mol. Biol. Cell, 2000)



Clustering analysis

Ex. Hierarchical clustering, K-means, Self-Organized Map



To improve accuracy, use additional information

Eisen, et al., PNAS, 1998. Tamayo et al., PNAS, 1999., Tavazoie, et al., Nat. Genet.1999. Keer, et al., PNAS 2001.

Gene Expression Analysis Using Additional Information

Related work

Gene Ontology (GO) ⇨ *a* priori probability distribution
(Pan,Bioinformatics,2006)

GO ⇨ Corrected distance
(Huang and Pan,Bioinformatics,2006)

Metabolic pathway ⇨ Length of the shortest path
(Hanisch, et al.,Bioinformatics,2002)

Proposed method

Metabolic pathway

⇨ ***Hidden modular random fields (Network modularity)***

2. Proposed Method

Probabilistic model

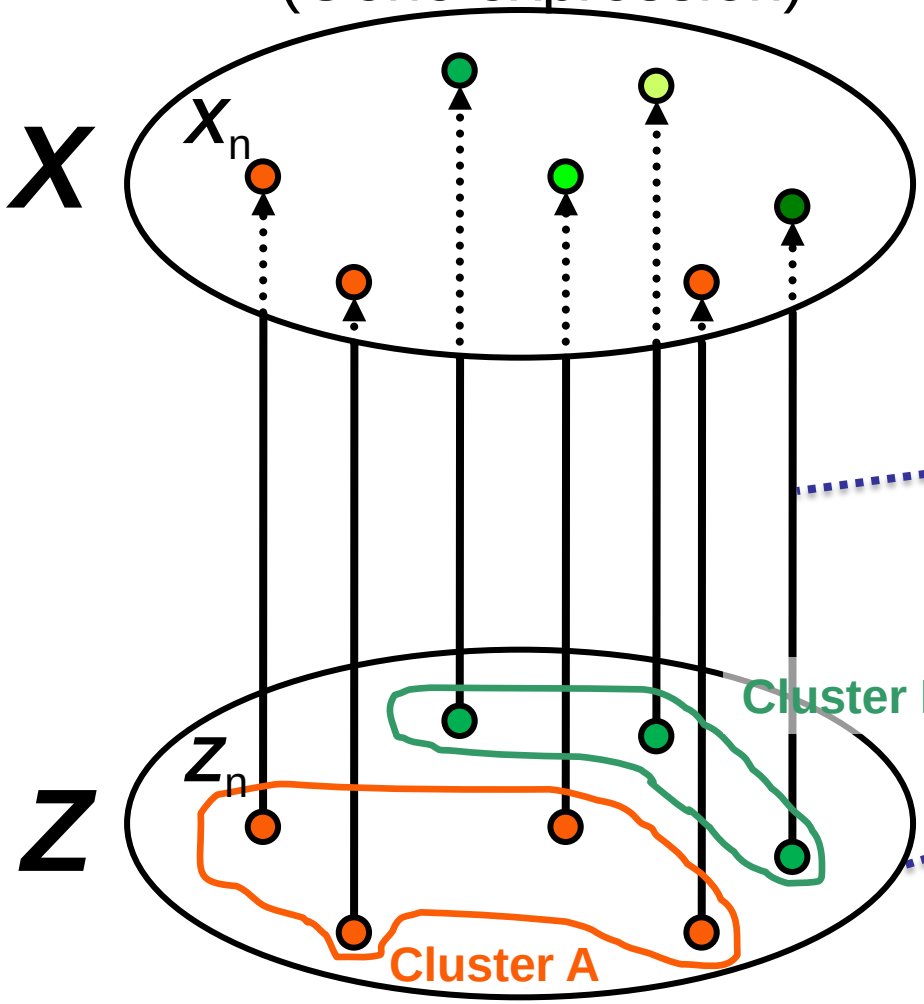
- Hidden modular random field using network modularity

Clustering algorithm

- EM algorithm + ICM

Existing Probabilistic Model (k-means)

Observable
(Gene expression)



Hidden (Cluster labels)

Probability distribution

$$p(\mathbf{X},\mathbf{Z}) = p(\mathbf{X}|\mathbf{Z})p(\mathbf{Z})$$

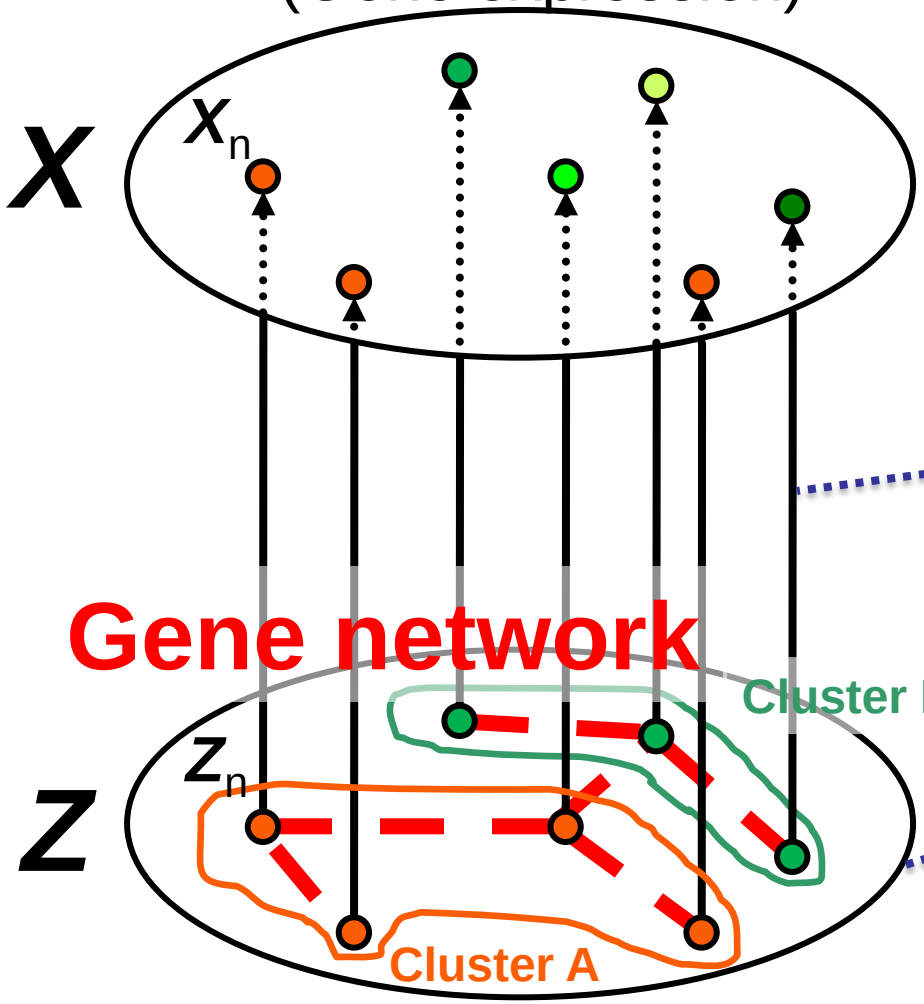
$$p(\mathbf{X}|\mathbf{Z}) = \prod_{n=1}^{N} p(\mathbf{X}_n|Z_n)$$

$$p(\mathbf{Z}) = \prod_{n=1}^{N} p(Z_n)$$

Independent

Our Probabilistic Model

Observable
(Gene expression)



Hidden (Cluster labels)

Hidden random field

Probability distribution

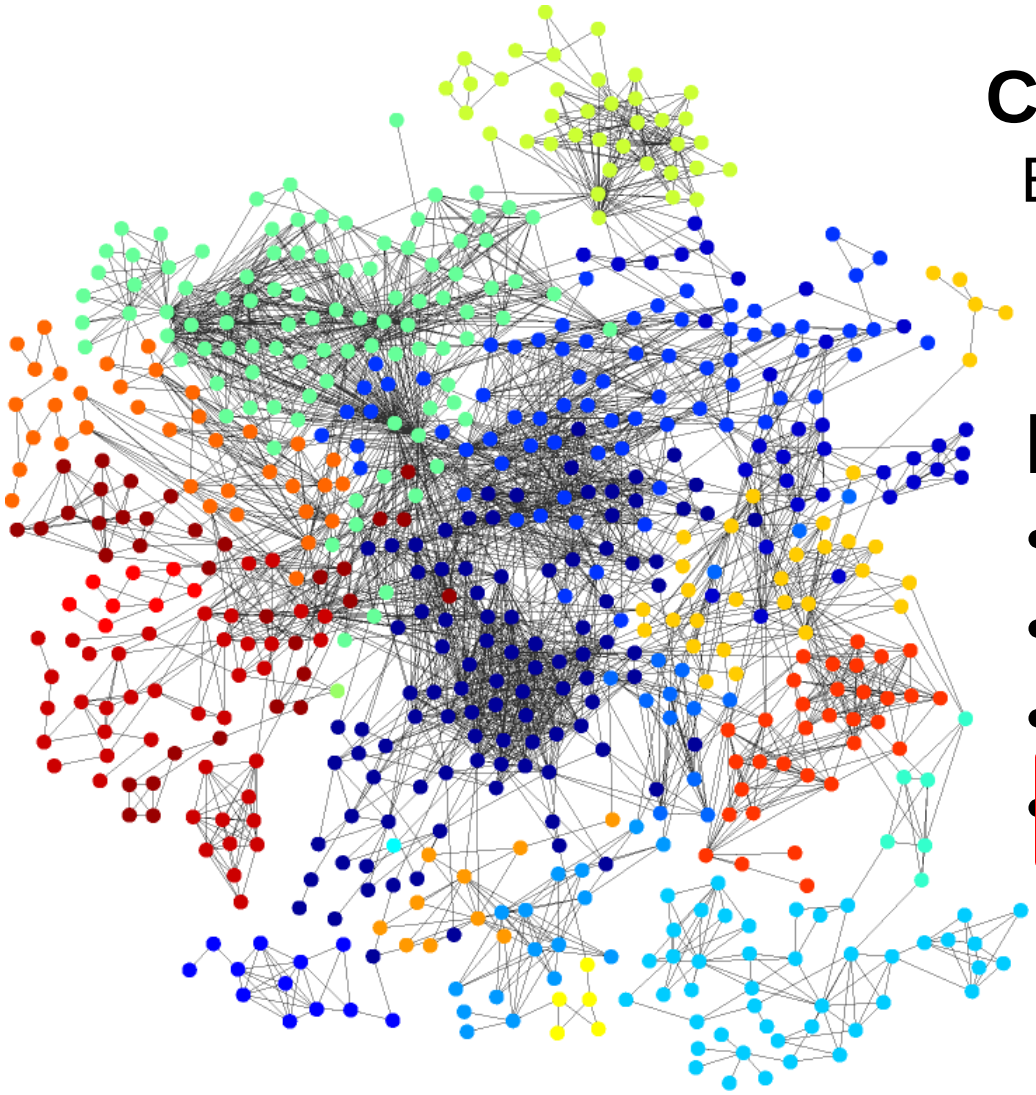
$$p(\mathbf{X},\mathbf{Z}) = p(\mathbf{X}|\mathbf{Z})p(\mathbf{Z})$$

$$p(\mathbf{X}|\mathbf{Z}) = \prod_{n=1}^{N} p(\mathbf{X}_n|Z_n)$$

$$p(\mathbf{Z}) =$$ ***Random Field***

dependent

Gene Networks



Complex networks

Ex. Gene network, WWW,
Social network, …, etc.

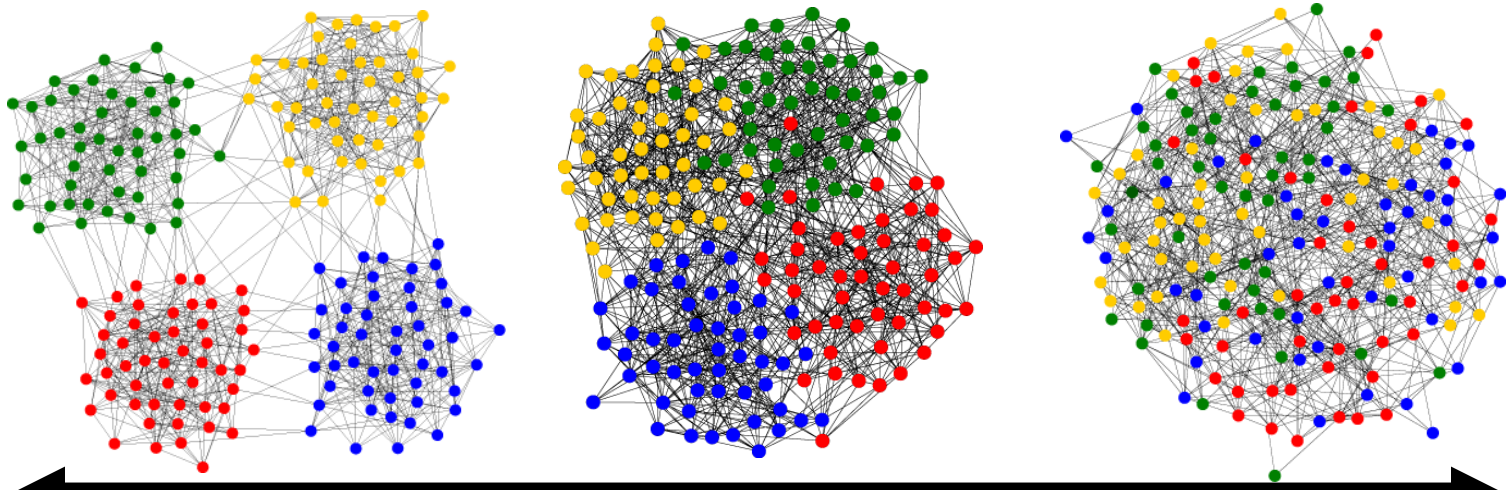
Properties

- Small world phenomena
- Scale-free
- Hierarchical structure
- Network modularity

Ravasz, et al., Science, 2002.
Guimera, et al., Nature, 2005.

Network Modularity

= density of intra-cluster edges



High ⟷ Low

$$V_G(Z) = N \sum_{k=1}^{K} \left\{ \frac{l_k(Z)}{L} - \left(\frac{d_k(Z)}{2L} \right)^2 \right\}$$

intra-edges

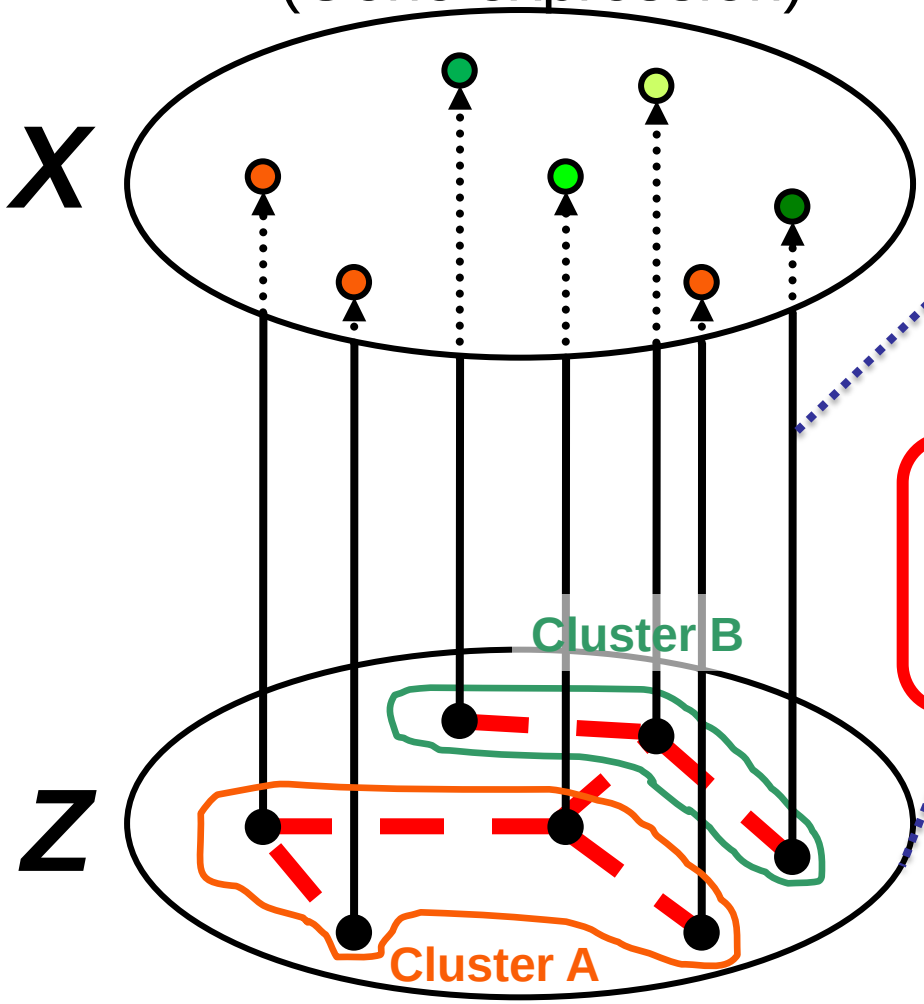
total edges

Guimera, et al., Nature, 2005., *Newman, et al., Phy. Rev. E, 2004.*

Hidden Modular Random Field

Observable

(Gene expression)



Hidden (Cluster labels)

Probability distribution

$$p(\mathbf{X},\mathbf{Z}) = p(\mathbf{X}|\mathbf{Z})p(\mathbf{Z})$$

$$p(\mathbf{X}|\mathbf{Z}) = \prod_{n=1}^{N} p(\mathbf{X}_n|Z_n)$$

Modular random field

$$p(Z) =$$ Random field

Network modularity

$$V_G(\boldsymbol{Z}) = N \sum_{k=1}^{K} \left\{ \frac{l_k(\boldsymbol{Z})}{L} - \left(\frac{d_k(\boldsymbol{Z})}{2L} \right)^2 \right\}$$

2. Proposed Method

Probabilistic model

- Hidden modular random field using network modularity

Clustering algorithm

- EM algorithm + ICM

Cost function

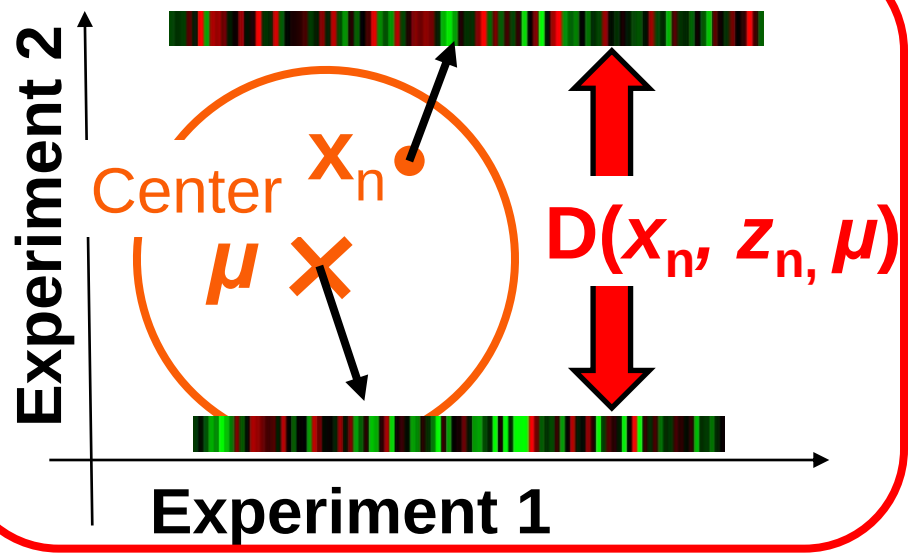
is the **negative pseudo-likelihood**

$$-\log P(\boldsymbol{X}, \boldsymbol{Z}) = -\sum_{n=1}^{N} \log p(\boldsymbol{x}_n | z_n) - \log P(\boldsymbol{Z})$$

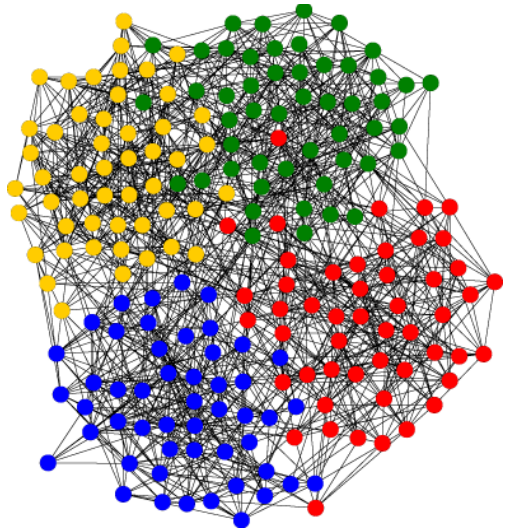
substitute **our probabilistic model** into log *P(X,Z)*

$$J = (1-\omega)\,D(X, Z, \mu) - \omega\, V_G(Z)$$

Dissimilarity



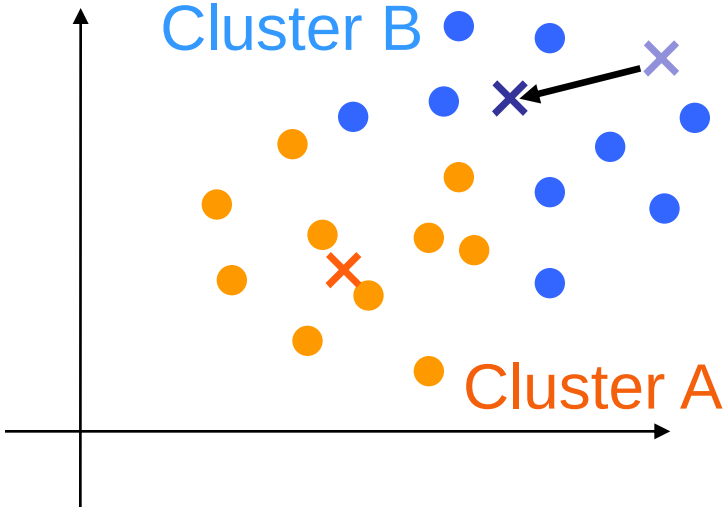
Network modularity



Clustering Algorithm

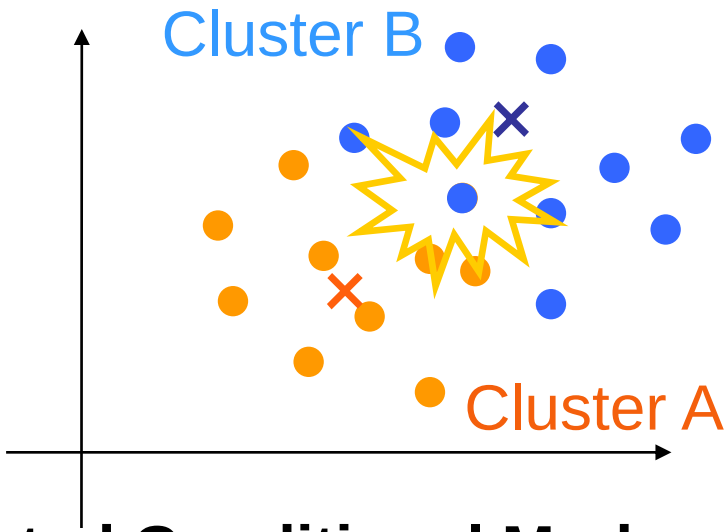
1. Set the number of cluster ***K***
2. Initialize cluster labels ***Z*** and the cluster centers ***μ***.
3. Iterate (i) and (ii) alternately until convergence.

(i)Update the cluster centers by minimizing the cost *J*



Iteration

(ii)Update cluster labels by minimizing *J*



Iterated Conditional Modes

(J. Besag, J. Roy. Stat. Soc. B, 1986)

3. Experiments

- Datasets
- Standard cluster
- Evaluation measure
- Results

Datasets

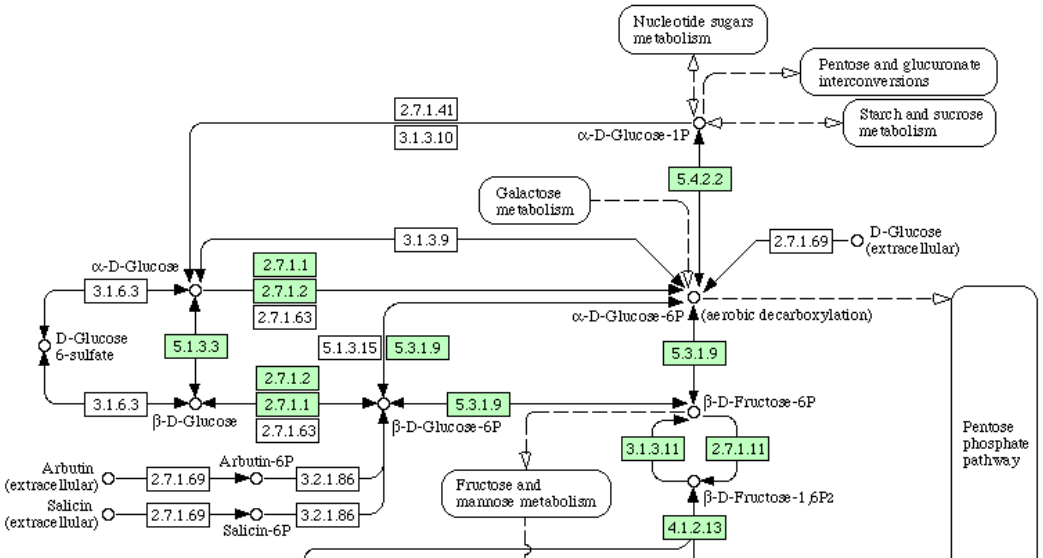
Species : *Saccharomyces cerevisiae (Yeast)*

Gene expression dataset

- Used 6 dataset (# experiments > 50)
- Results almost similar.
- Show an example : **Hughes'** dataset *(Hughes, et al., cell, 2000)*
  - **# experiments = 300**.
  - Missing 0.001% (= 241 / (636x300)).

Gene network

KEGG (Kyoto Encyclopedia of Genes and Genomes)



 # genes = 636
 # edges = 3,104

Evaluation Measure

Normalized Mutual Information (NMI)

between estimated cluster and the standard cluster

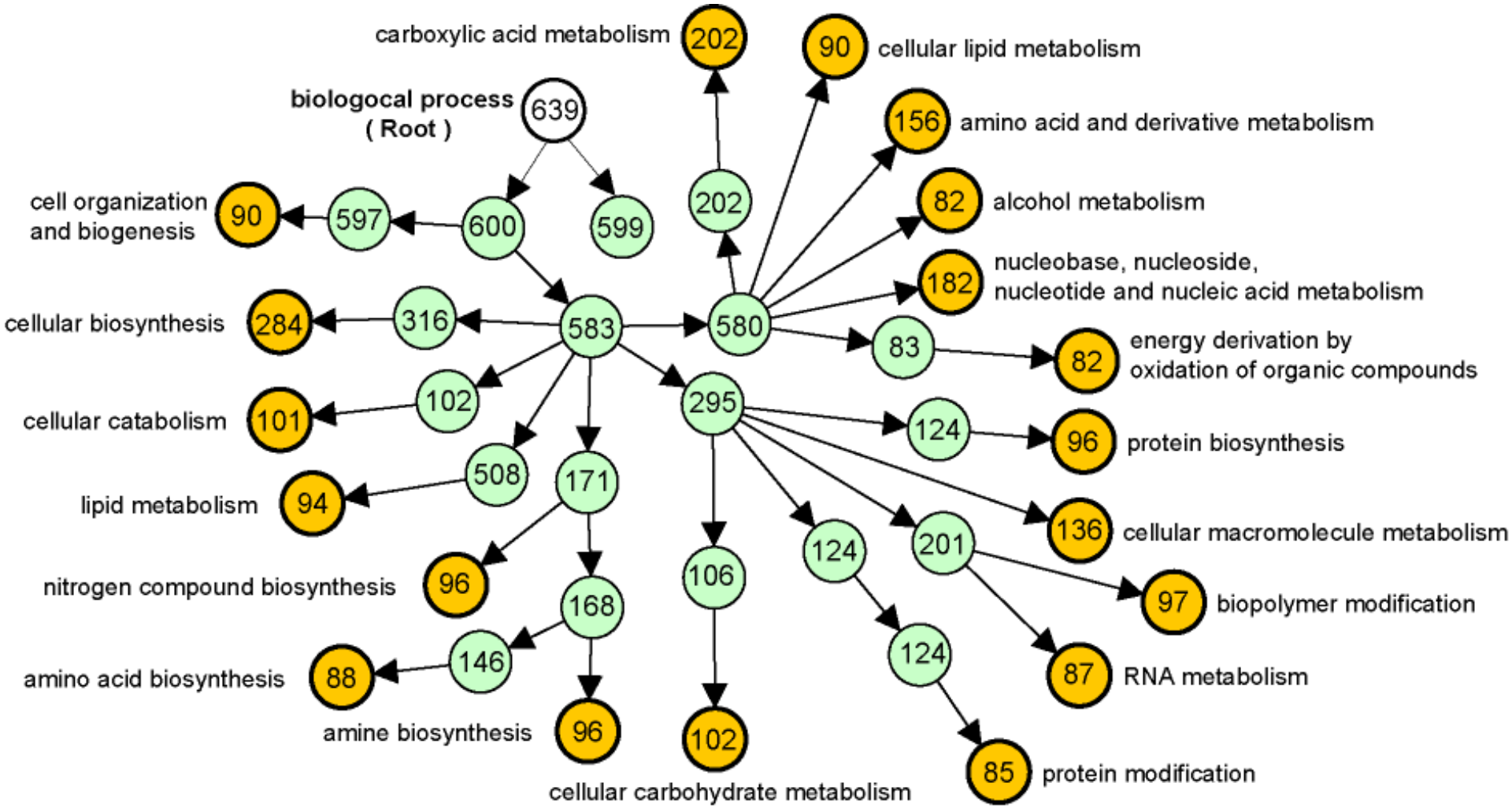
$$NMI = \frac{H(C) + H(G) - H(C,G)}{\sqrt{H(C)}\sqrt{H(G)}}$$

H(***C***) : Entropy of probability variable ***C***,
C : Estimated clusters, ***G*** : Standard clusters

The more **similar** clusters
C and ***G*** are, **the larger the NMI**.

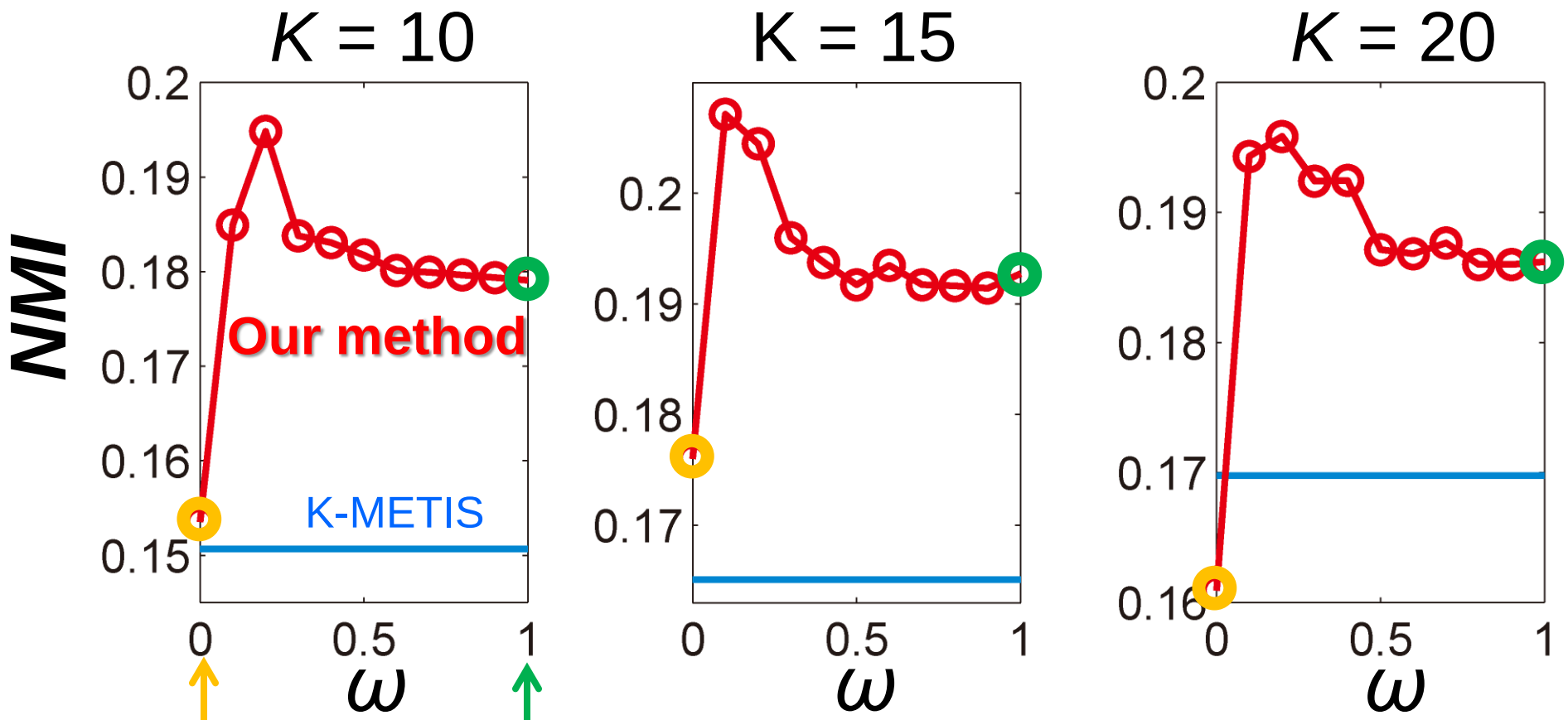
Standard cluster generated from GO

clusters = 19



(Threshold = 75)

Performance results



Gene expression only
(**k-means**)

Gene network only
(**maximum modularity**)

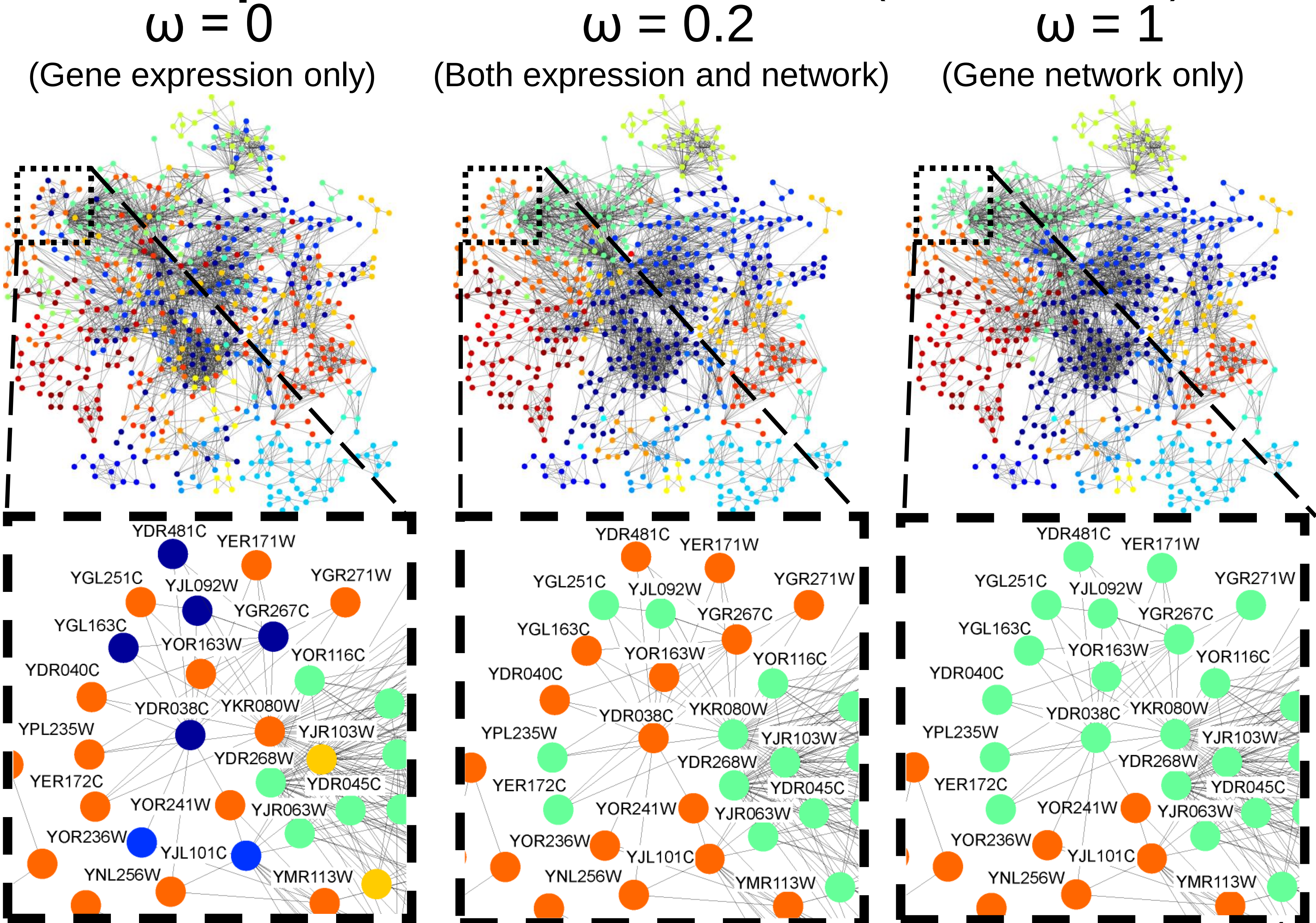
∗ Best NMI is in 0 < ω < 1.

∗ Better than three others.

(i) k-means, (ii) max. Mod., (iii) k-METIS

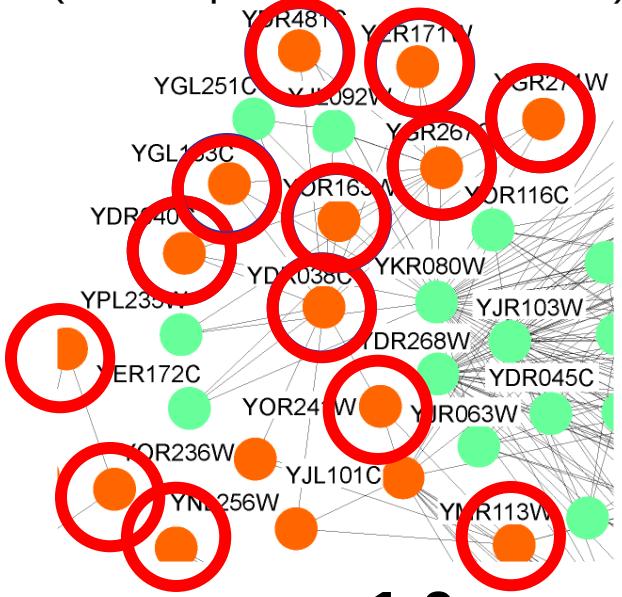
k-METIS (Karypis, G. and Kumar, V., J. Parallel Distrib. Comput., 1998)

Examples of Resultant Cluster (#cluster = 20)



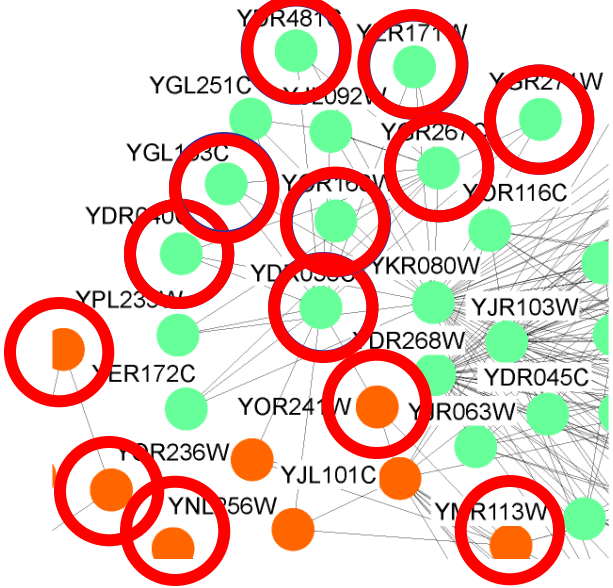
ω = 0.2

(Both expression and network)

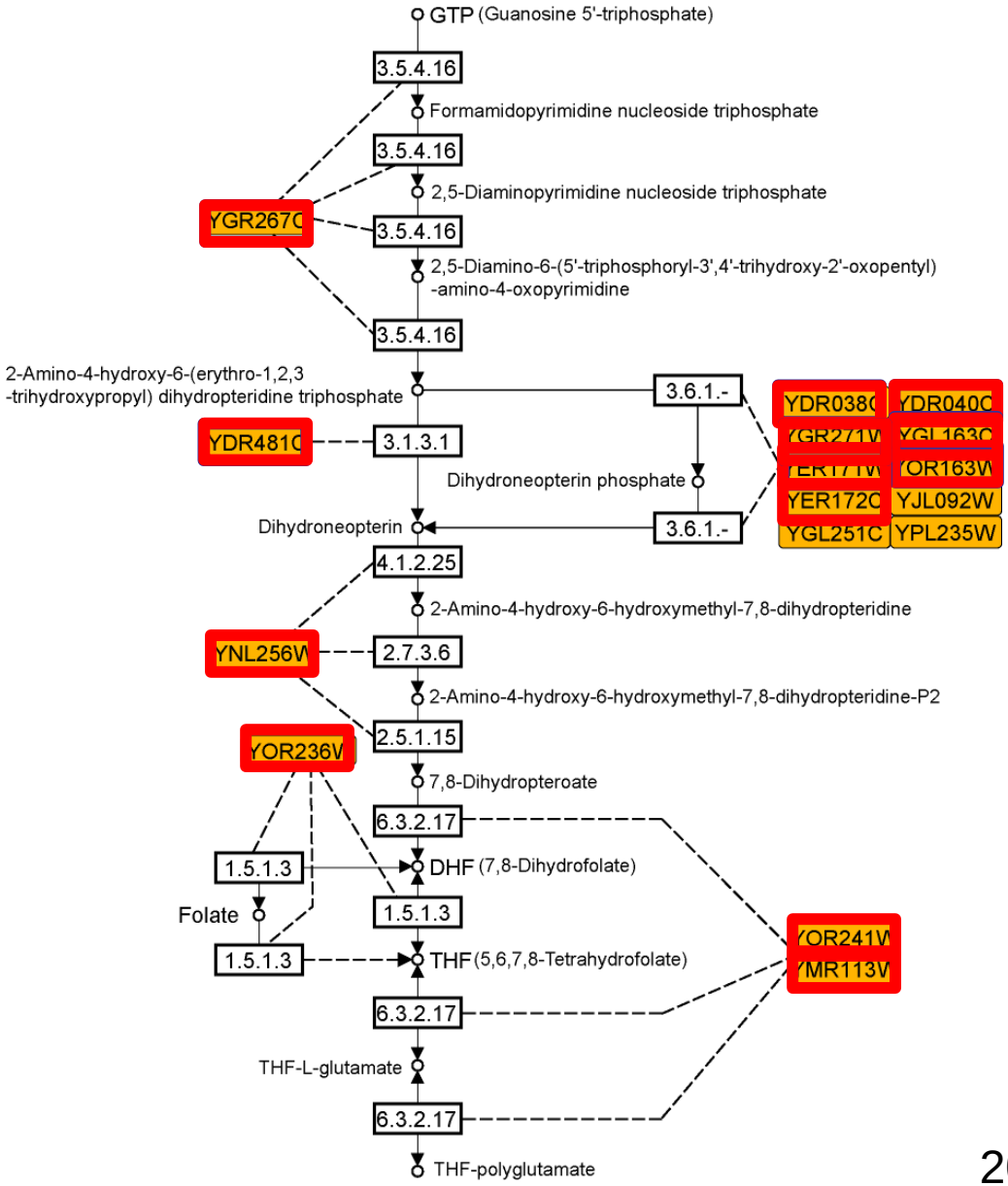


ω = 1.0

(Network only)



Folate biosynthesis pathway



Summary

1. **Proposed a new clustering method combining expression data and a network**
   - Probabilistic model
     Hidden modular random fields (gene expression + network)
   - Clustering algorithm
     EM algorithm + ICM
2. **Experimental results**
   - Best **NMI** by $0 < \omega < 1$ **(expression + network)**.
   - Better than three others.
     - $\omega=0$ (k-means)
     - $\omega=1$ (Maximizing network modularity)
     - k-METIS (Graph partitioning)

Thank you for your attention!