Functions and Phenotypes by Integrative Network Analysis

Xianghong Jasmine Zhou

Molecular and Computational Biology

University of Southern California

Biological Networks

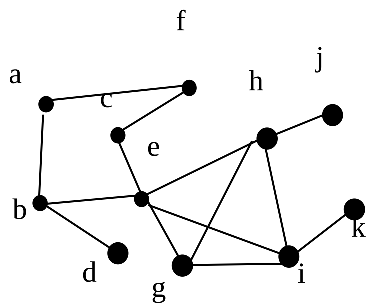
- Protein-protein interaction network
- Metabolic network
- Transcriptional regulatory network
- Co-expression network
- Genetic Interaction network
- ...

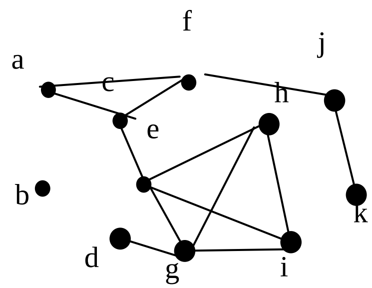
Challenges in biological network analysis

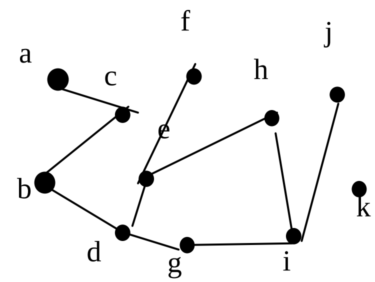
- Most current network algorithms can only be applied to a single network.

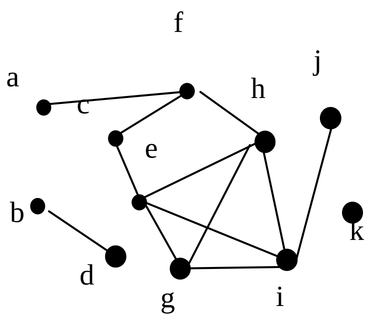
- The rapid accumulation of biological networks translates into an urgent need of methods for integrative network analysis

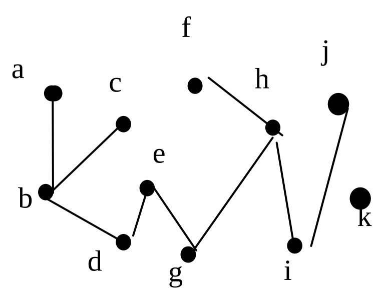
Data Mining Across Multiple Networks

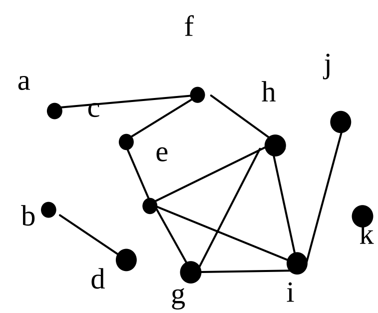




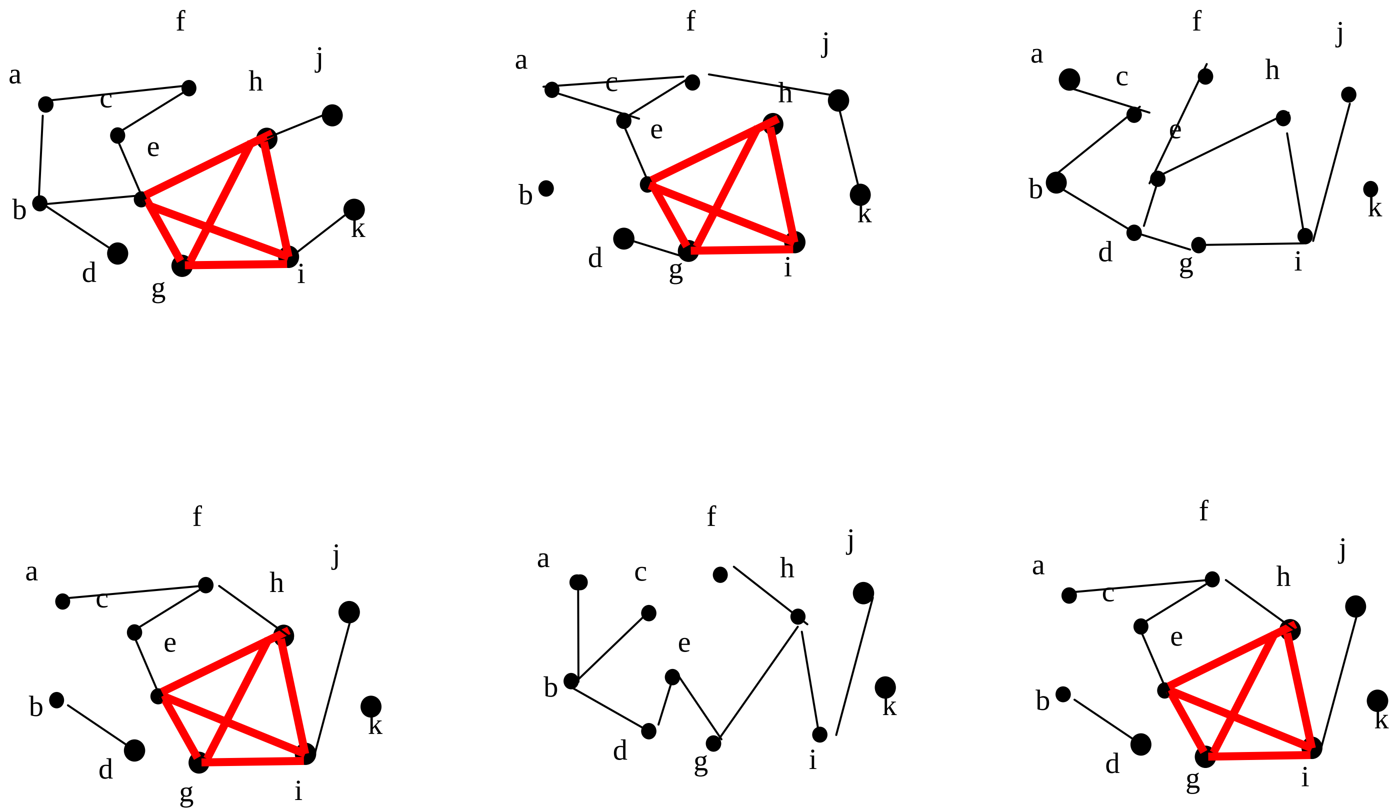








Data Mining Across Multiple Networks



Microarray technology

- Microarray technology is used to measure the expression (activities) of tens of thousand genes in cells simultaneously.
- The results can be summarized into a matrix

$$
\begin{array}{cc}
\text{genes} & \begin{bmatrix} X_{11} & X_{12} & X_{13} & \cdots \\ X_{21} & X_{22} & X_{23} & \cdots \\ X_{31} & X_{32} & X_{33} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \\
 & \text{target samples}
\end{array}
$$

Rapid accumulation of microarray data in public repositories

- NCBI Gene Expression Omnibus

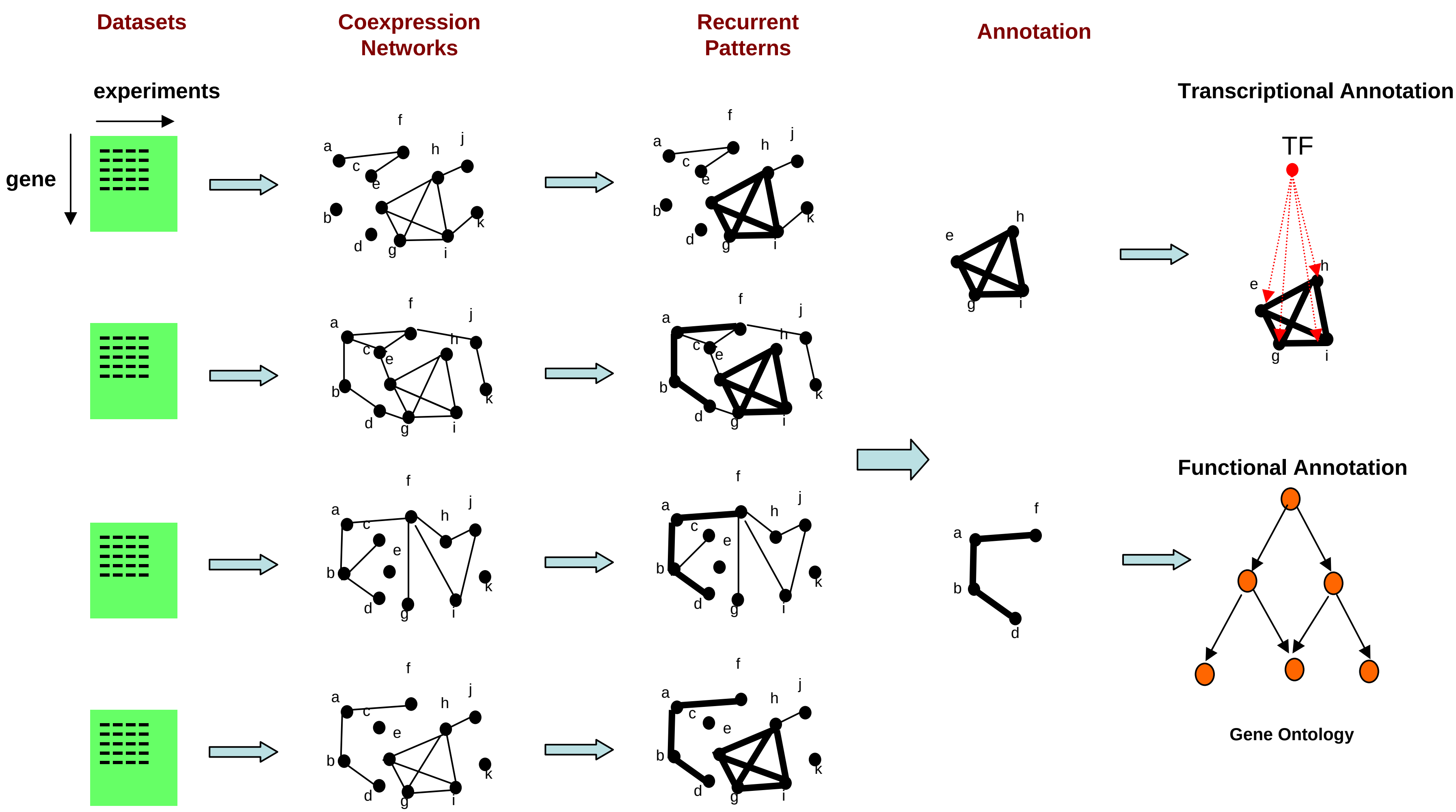
137231 experiments

- EBI Array Express

55228 experiments

The public microarray data increases by 3 folds per year

Graph-based Approach for the Integrative Microarray Analysis



Frequent Subgraph Mining Problem is hard!

Problem formulation: Given *n* graphs, identify subgraphs which occur in at least *m* graphs (m ≤ *n*)

Our graphs are massive!

The traditional pattern growth approach (expand frequent subgraph of *k* edges to *k+1* edges) would not work, since the time and memory requirements increase exponentially with increasing size of patterns and increasing number of networks.

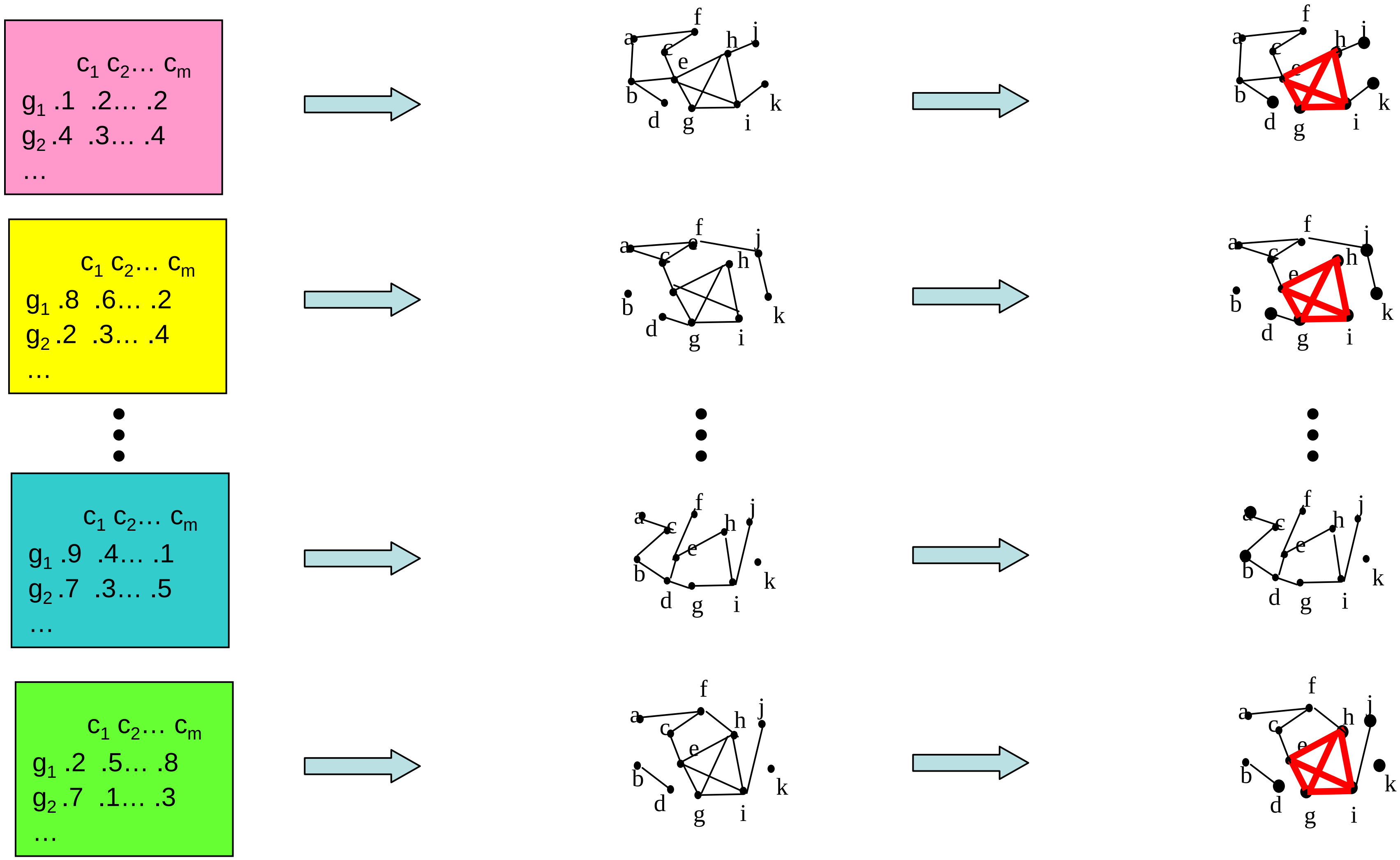
Novel Algorithms to identify diverse frequent network patterns

- CoDense (ISMB 2005)
  - identify frequent coherent dense subgraphs across many massive graphs
- Network Biclustering (ISMB 2007)
  - identify frequent subgraphs across many massive graphs
- Network Modules (ISMB 2007)
  - identify frequent dense vertex sets across many massive graphs

CODENSE: identify frequent coherent dense subgraphs across massive graphs

Hu et al, ISMB 2005

Identify frequent co-expression clusters across multiple microarray data sets



The common pattern growth approach

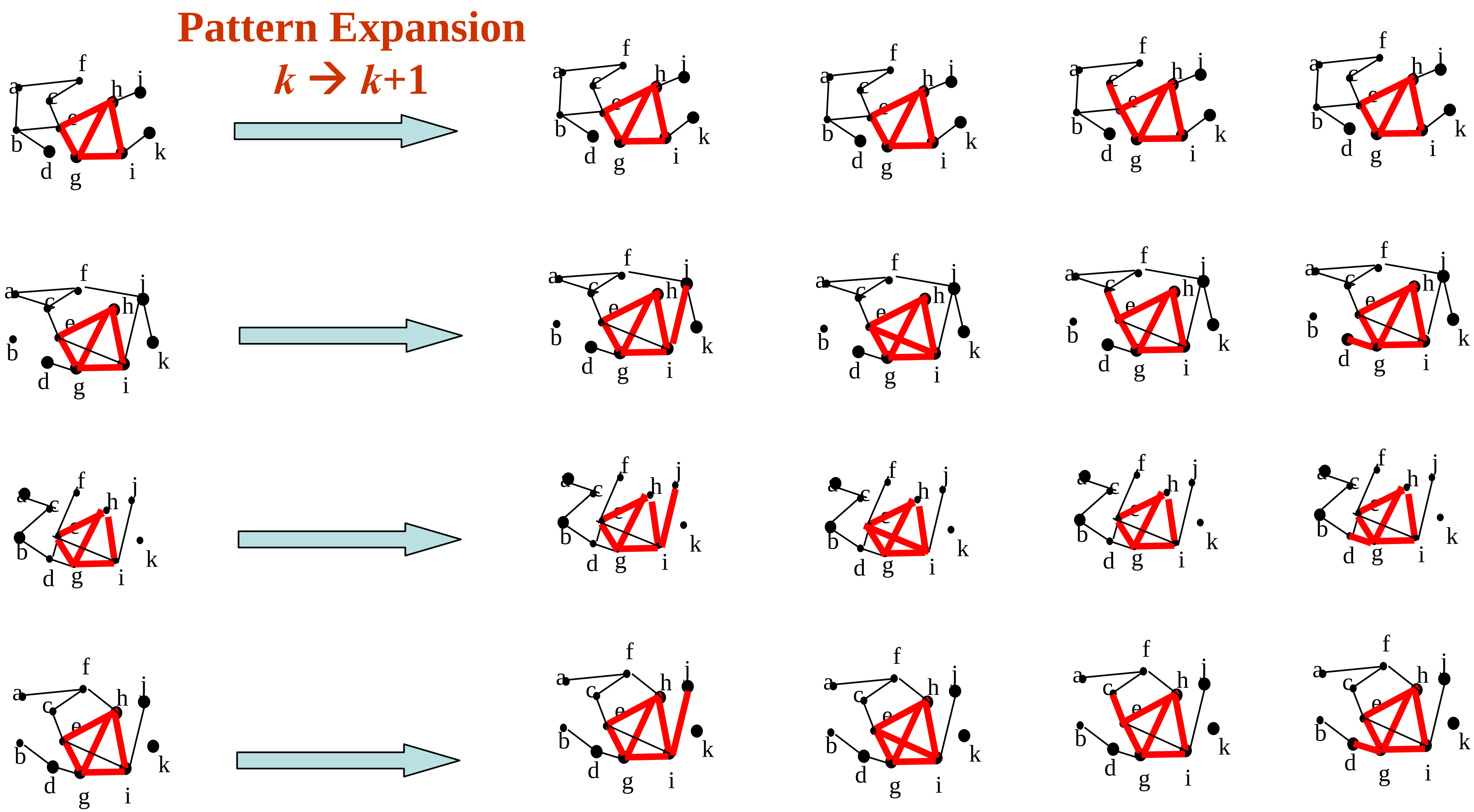
Find a frequent subgraph of *k* edges, and expand it to *k+1* edge to check occurrence frequency

- Koyuturk M., Grama A. & Szpankowski W. *An efficient algorithm for detecting frequent subgraphs in biological networks*. ISMB 2004
- Yan, Zhou, and Han. *Mining Closed Relational Graphs with Connectivity Constraints*. ICDE 2005

Problem of the Pattern-growth approach

The time and memory requirements increase exponentially with increasing size of patterns and increasing number of networks. The number of frequent dense subgraphs is explosive when there are very large frequent dense subgraphs, e.g., subgraphs with hundreds of edges.

Problem of the Pattern-growth approach



Our solution

We develop a novel algorithm, called *CODENSE*, to mine frequent ***co****herent* ***dense*** subgraphs. The target subgraphs have three characteristics:

(1) All edges occur in >= k graphs (frequency)

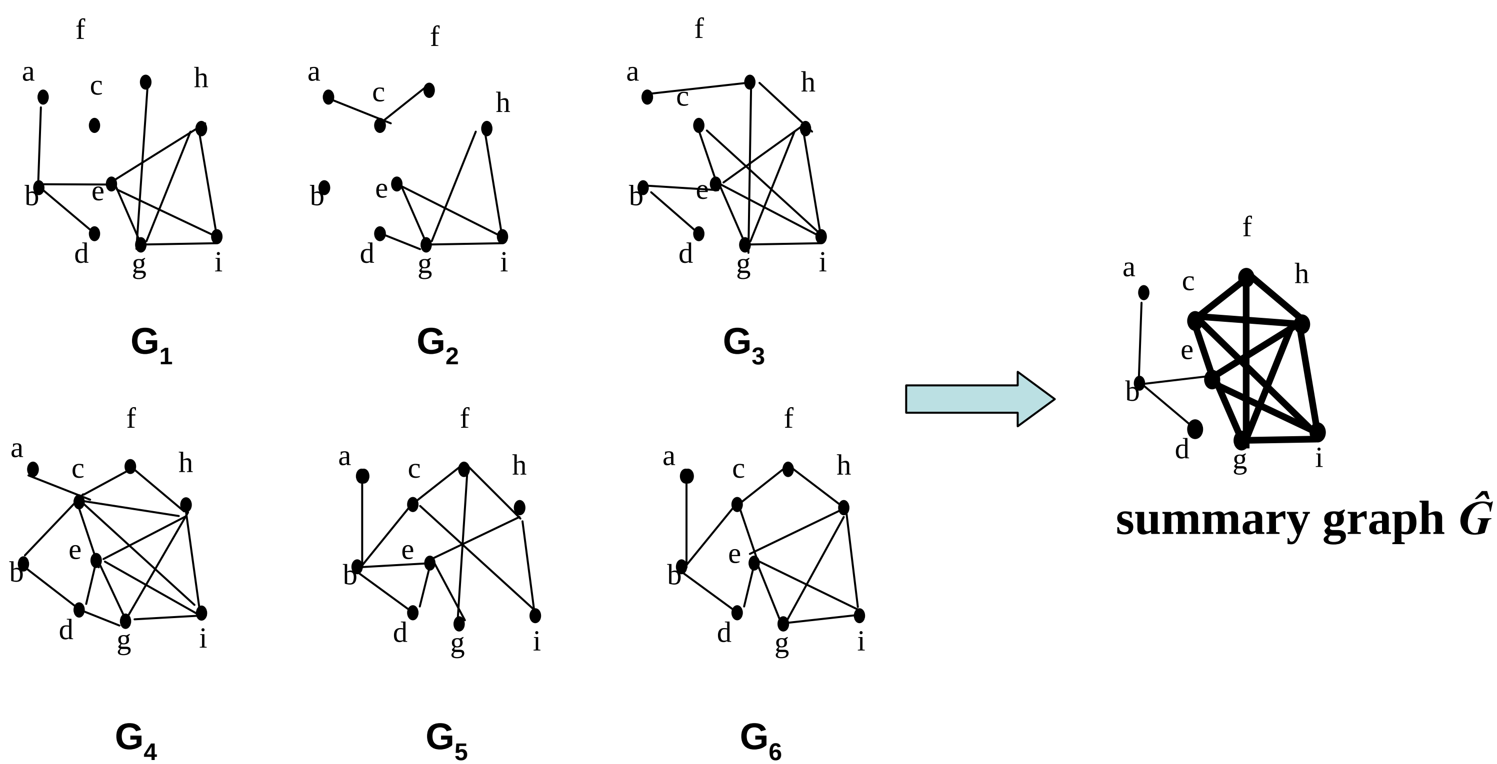
(2) All edges should exhibit correlated occurrences in the given graph set. (coherency)

(3) The subgraph is dense, where density *d* is higher than a threshold γ and *d=2m/(n(n-1))* (density)

m: #edges, n: #nodes

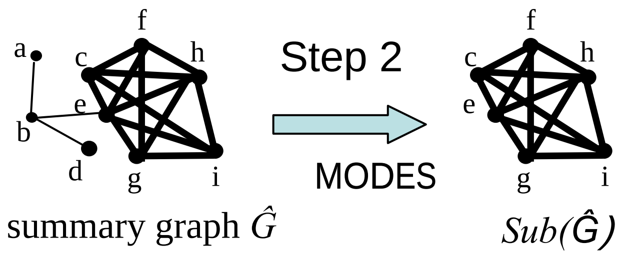
CODENSE: Mine coherent dense subgraph

(1) Builds a summary graph by eliminating infrequent edges



CODENSE: Mine coherent dense subgraph

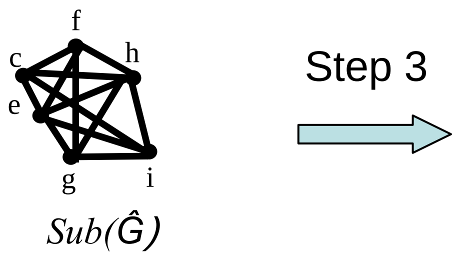
(2) Identify dense subgraphs of the summary graph



Observation: If a frequent subgraph is dense, it must be a dense subgraph in the summary graph. However, the reverse conclusion is not true.

CODENSE: Mine coherent dense subgraph

(3) Construct the edge occurrence profiles for each dense summary subgraph



E	G1	G2	G3	G4	G5	G6
c-e	0	0	1	1	0	1
c-f	0	1	0	1	1	1
c-h	0	0	0	1	1	1
c-i	0	0	1	1	1	0
e-f	0	0	0	1	1	1
…	…	…	…	…	…	…

edge occurrence profiles

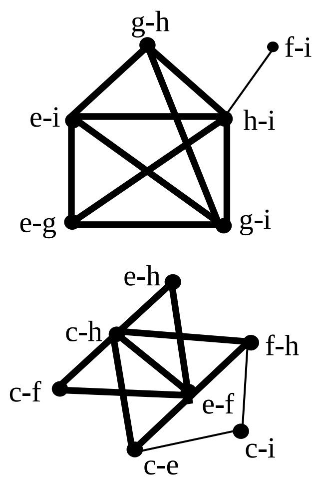
CODENSE: Mine coherent dense subgraph

(4) builds a second-order graph for each dense summary subgraph

E	G1	G2	G3	G4	G5	G6
c-e	0	0	1	1	1	1
c-f	0	1	0	1	1	1
c-h	0	0	0	1	1	1
c-i	0	0	1	1	1	0
e-f	0	0	0	1	1	1
…	…	…	…	…	…	…

edge occurrence profiles

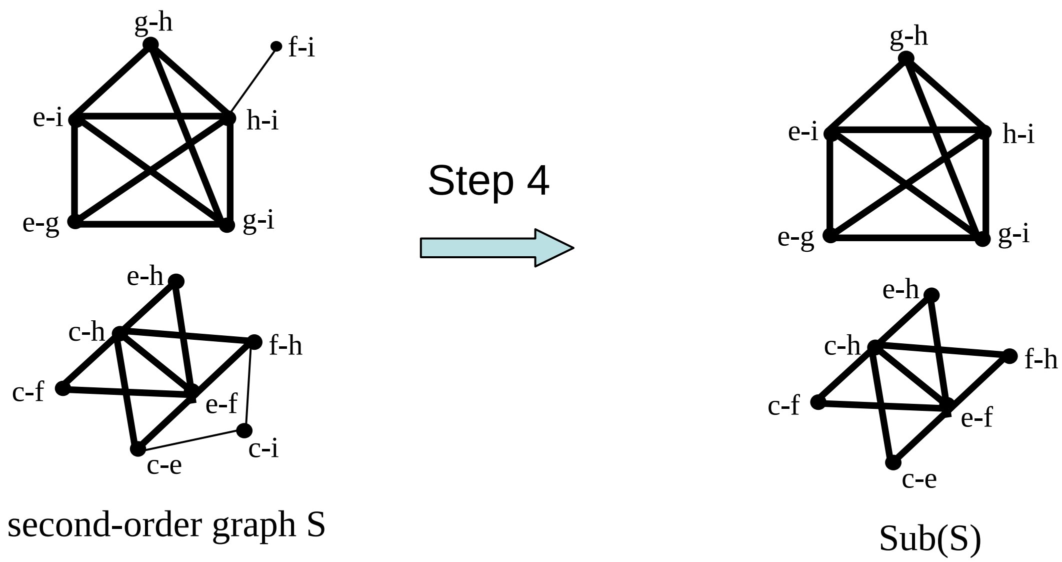
Step 4



second-order graph S

CODENSE: Mine coherent dense subgraph

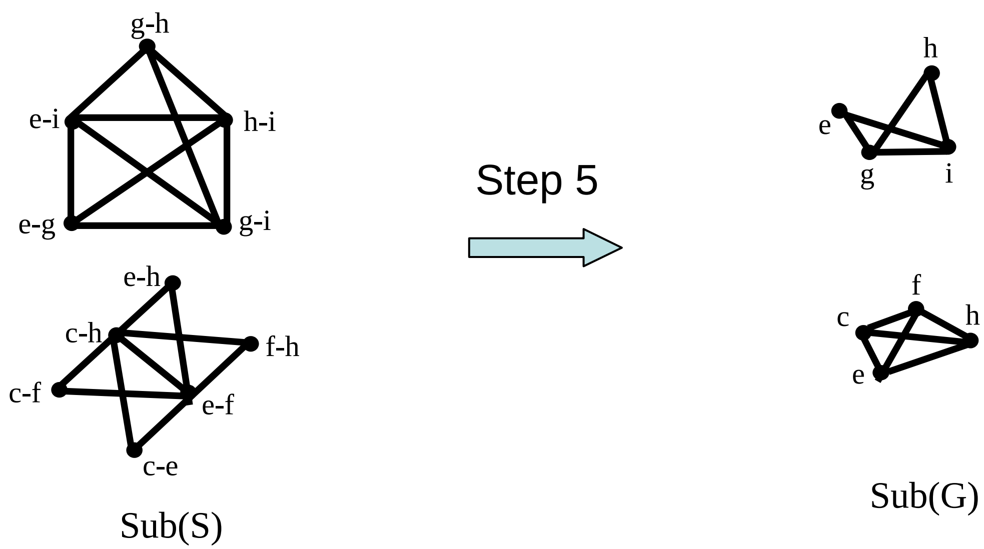
(5) Identify dense subgraphs of the second-order graph



Observation: if a subgraph is coherent (its edges show high correlation in their occurrences across a graph set), then its 2nd-order graph must be dense.

CODENSE: Mine coherent dense subgraph

(6) Identify the coherent dense subgraphs



Our solution

The identified subgraphs by definition satisfy the three criteria:

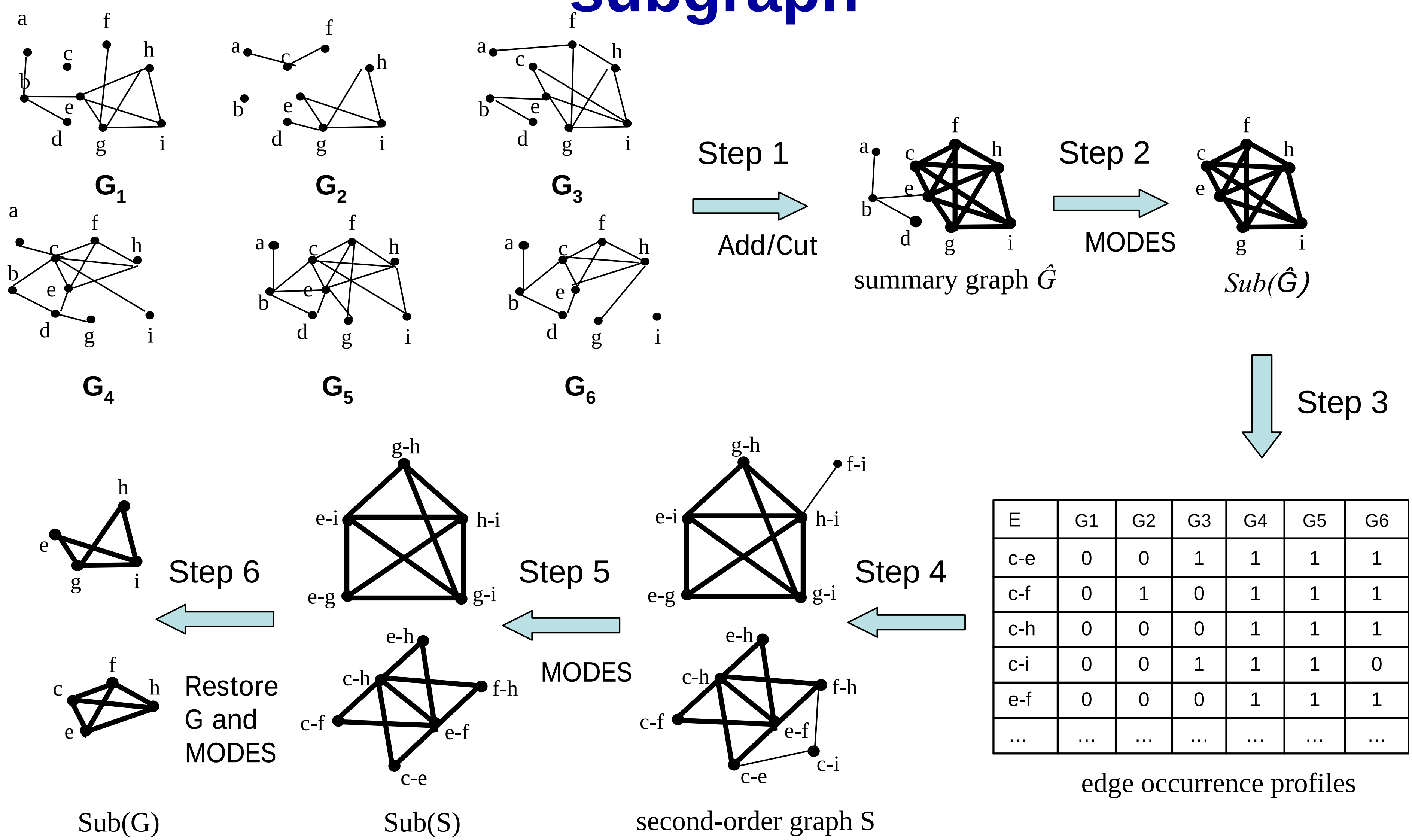
(1) All edges occur in >= k graphs (frequency)

(2) All edges should exhibit correlated occurrences in the given graph set. (coherency)

(3) The subgraph is dense, where density *d* is higher than a threshold γ and $d=2m/(n(n-1))$ (density)

m: #edges, n: #nodes

CODENSE: Mine coherent dense subgraph



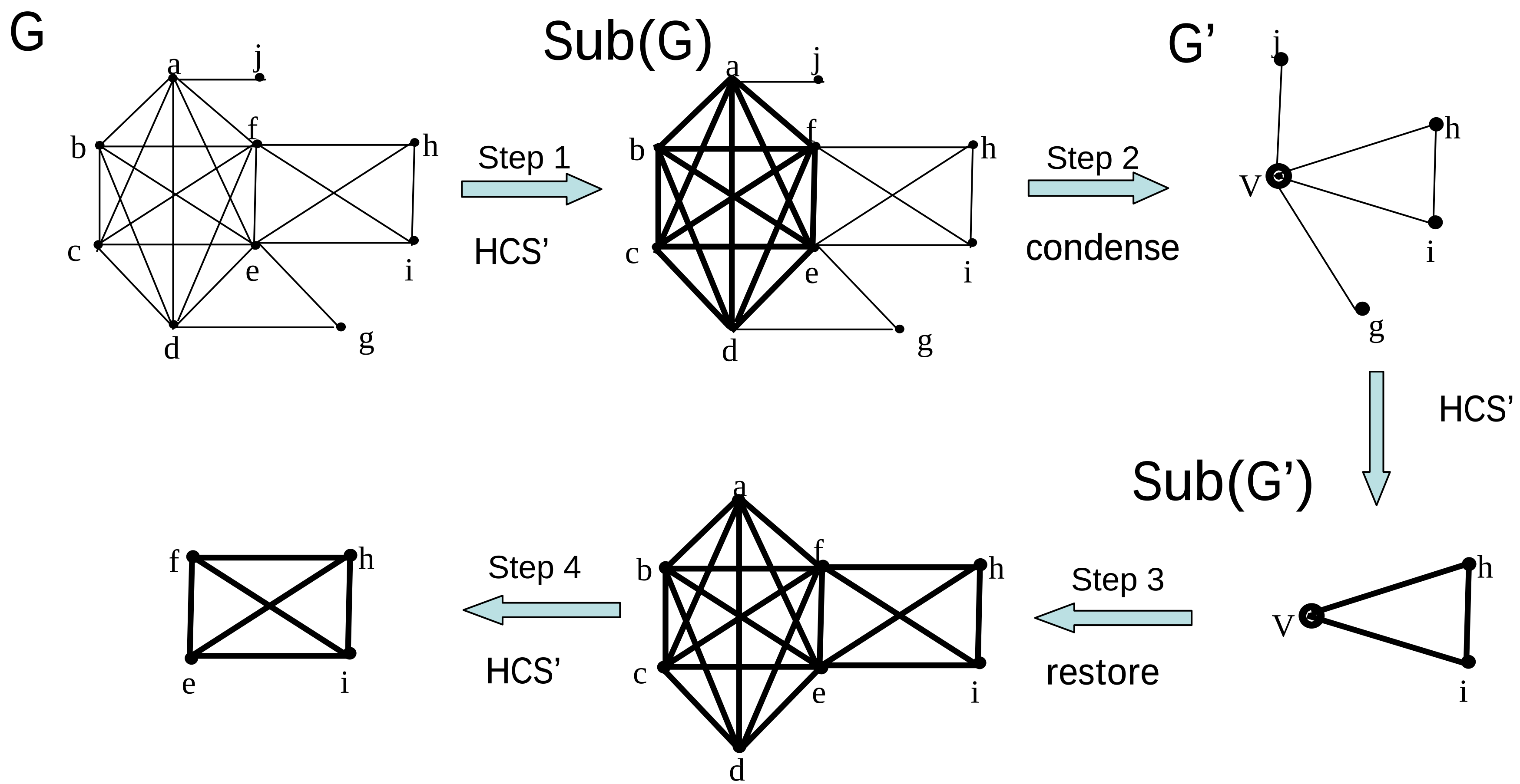
E	G1	G2	G3	G4	G5	G6
c-e	0	0	1	1	1	1
c-f	0	1	0	1	1	1
c-h	0	0	0	1	1	1
c-i	0	0	1	1	1	0
e-f	0	0	0	1	1	1
…	…	…	…	…	…	…

edge occurrence profiles

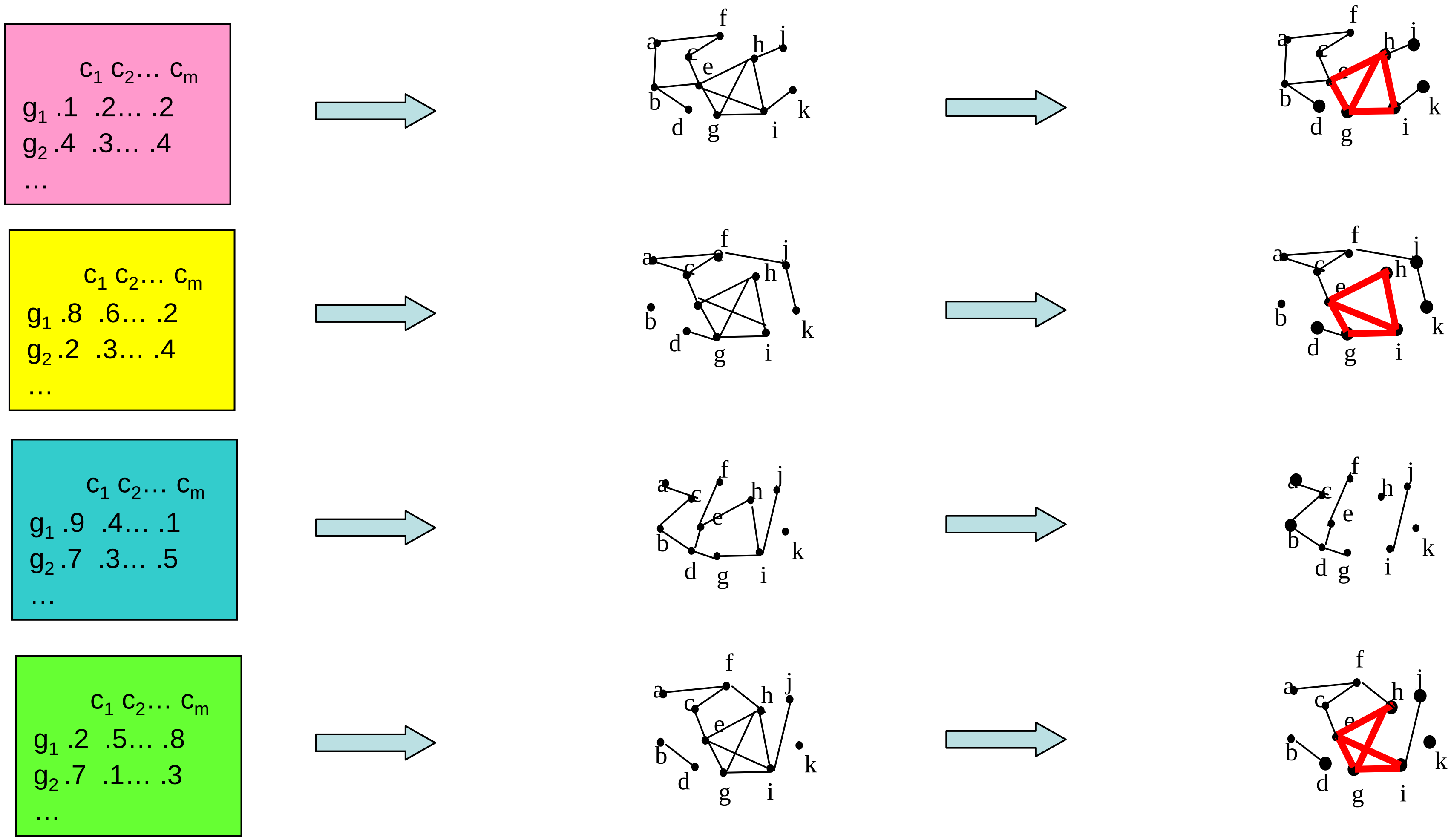
CODENSE

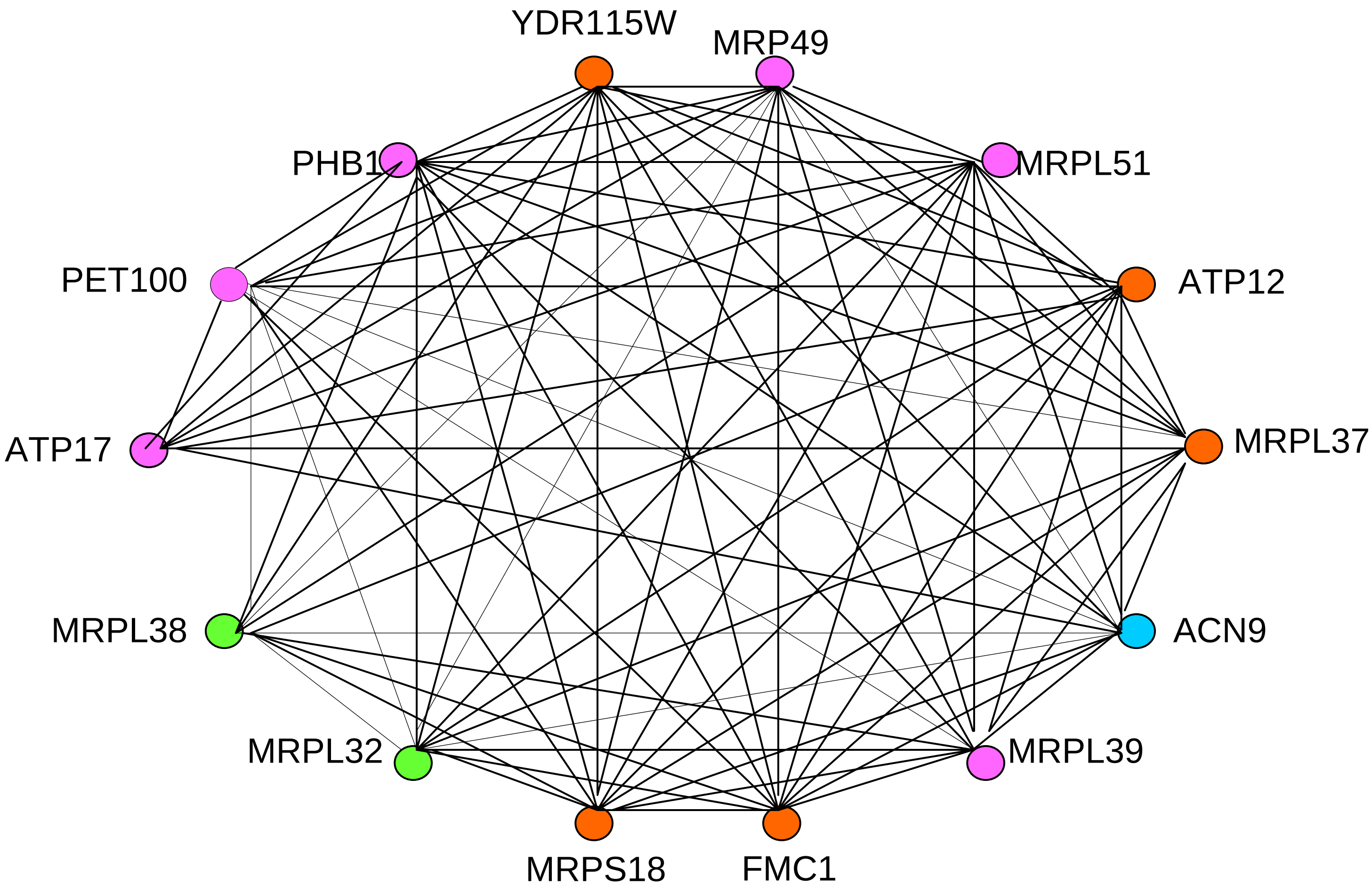
The design of CODENSE can solve the scalability issue. Instead of mining each biological network individually, CODENSE compresses the networks into two meta-graphs and performs clustering in these two graphs only. Thus, CODENSE can handle any large number of networks.

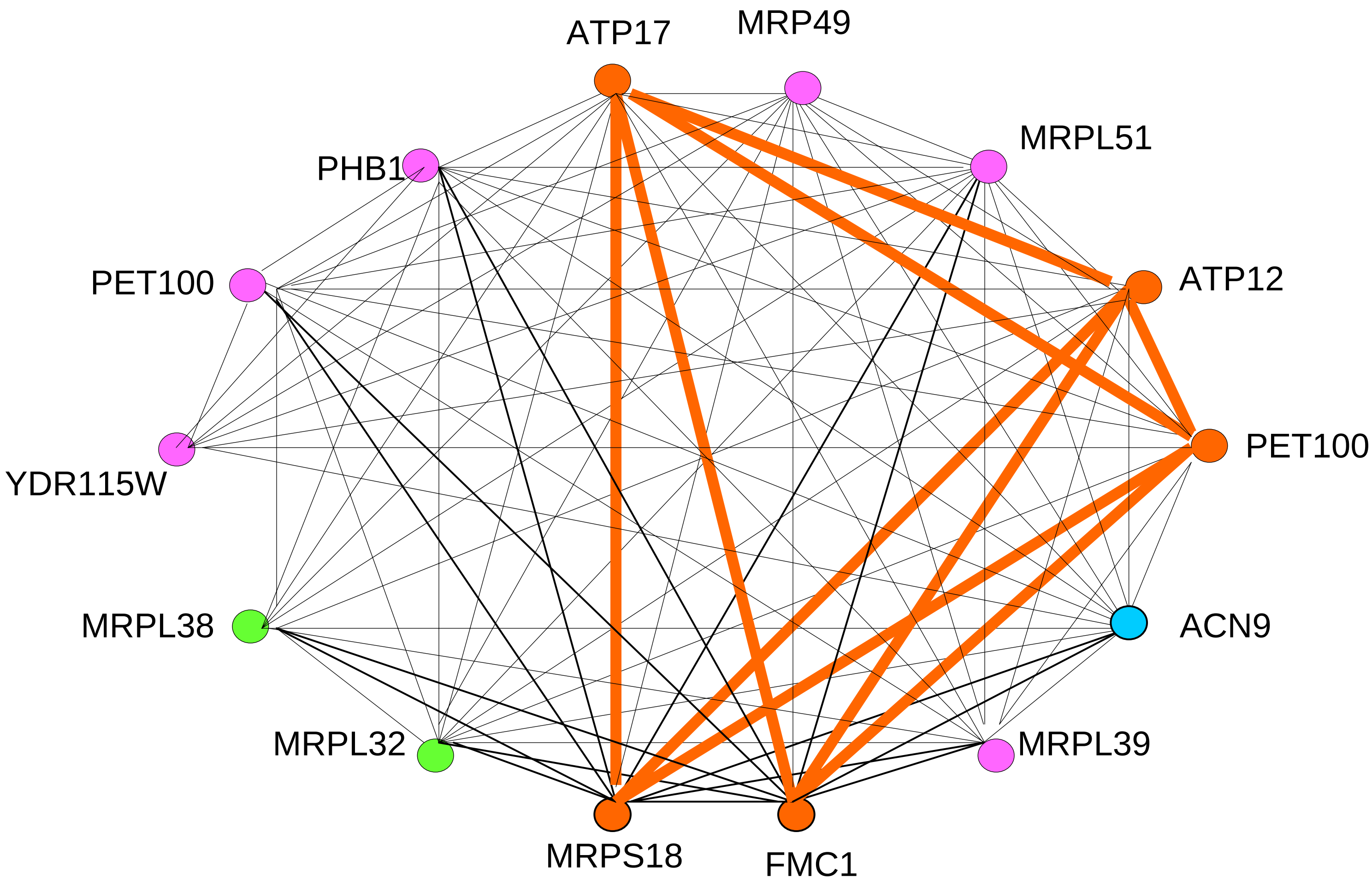
MODES: Mine overlapped dense subgraph



Applying CoDense to 39 yeast microarray data sets

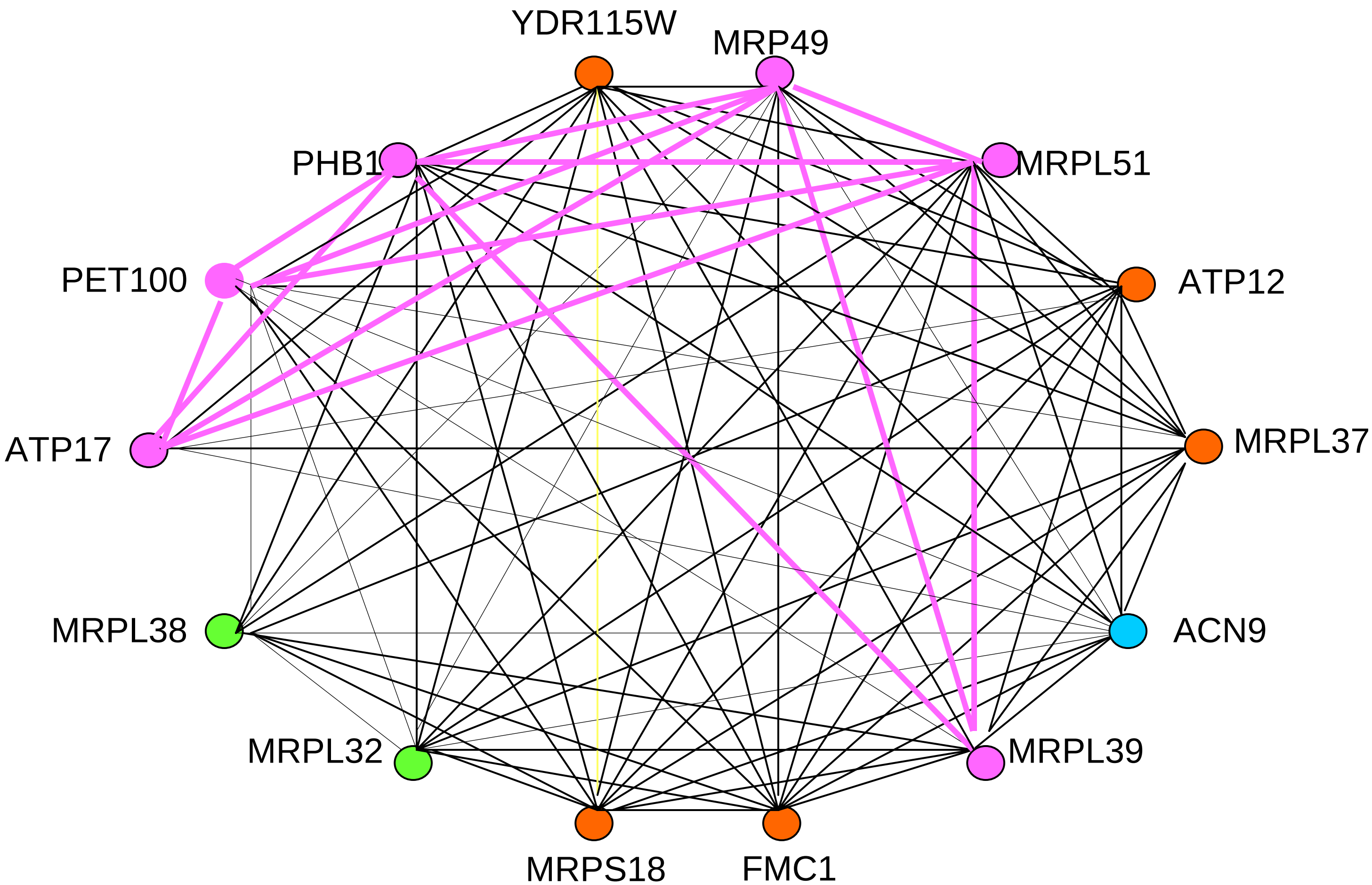



YDR115W
MRP49
PHB1
MRPL51
PET100
ATP12
ATP17
MRPL37
MRPL38
ACN9
MRPL32
MRPL39
MRPS18
FMC1



Yellow: YDR115W, FMC1, ATP12,MRPL37,MRPS18

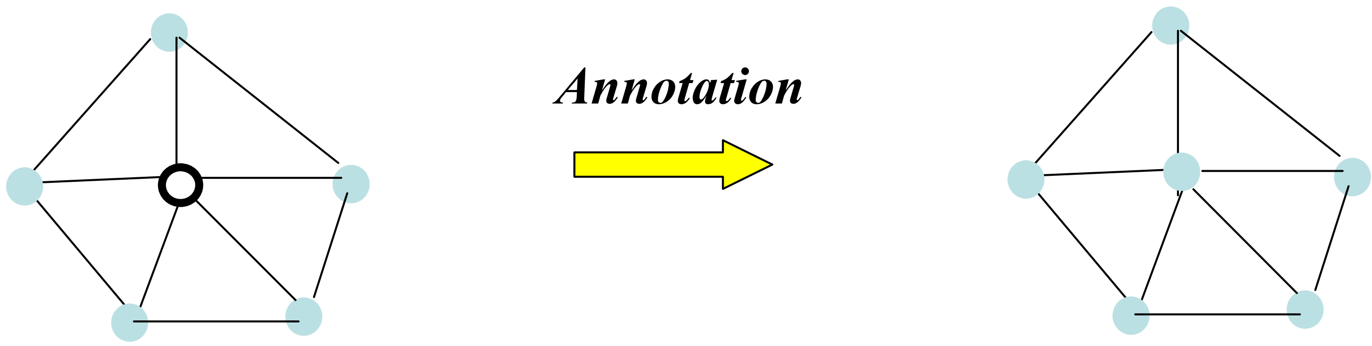
GO:0019538(protein metabolism; pvalue = 0.001122)



Red:PHB1,ATP17,MRPL51,MRPL39, MRPL49, MRPL51,PET100

GO:0006091(generation of precursor metabolites and energy; pvalue=0. 001339)

Functional annotation



Functional Annotation (Validation)

Method: leave-one-out approach - masking a known gene to be unknown, and assign its function based on the other genes in the subgraph pattern.

Functional categories: 166 functional categories at GO level at least 6

Results: 448 predictions with accuracy of 50%

Functional Annotation (Prediction)

We made functional predictions for 169 genes, covering a wide range of functional categories, e.g. amino acid biosynthesis, ATP biosynthesis, ribosome biogenesis, vitamin biosynthesis, etc. A significant number of our predictions can be supported by literature.

However…

- How about frequent non-dense graphs?
  - Many biological modules may form paths

- How about subgraphs which are coherent across only a subset of the graphs?
  - Not all modules are activated across all conditions, and genes may form modules with diff. other genes under diff. conditions

Network Biclustering:

Identify frequent subgraphs across massive graphs

Huang et al, ISMB 2007

Using 65 human co-expression network as an illustration example

- 65 co-expression networks generated from 65 microarray data sets

- each graph contains 8297 genes, and 1%-10% edges of a complete graph

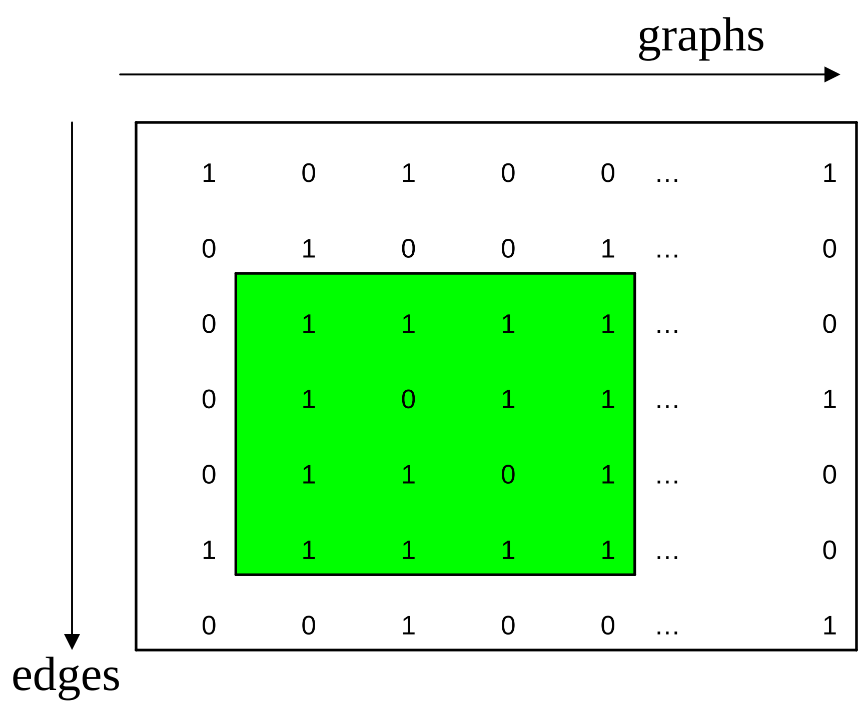
Basically, it is a biclustering problem

Network Biclustering

- Objective function

$$f = \frac{c'}{mn + \lambda c}$$

c': number of 1 in the bicluster
c: number of 1 in the whole matrix
mn: size of the bicluster
λ: regularization factor



However, the matrix is very large with millions of edges …

We will first identify robust seed to narrow down the search space

Identify Bicluster seed

The property of relation graphs: edge labels are unique.

Hence, each graph can be treated as a collection of items

Thus, Frequent subgraph Mining can be modeled as frequent item set mining

Problem: current frequent item set mining algorithms can only efficiently mine across many small item sets
In our problem, we have 65 very large item set…

We use a trick….

Identify Bicluster seed

Edge occurrence profiles:

E	G1	G2	G3	G4	G5	G6	...	G60	G61	G62	G63	G64	G65
e1	1	1	0	1	0	1	...	0	0	1	1	1	1
e2	1	1	0	1	0	1	…	0	1	0	1	1	1
e3	1	1	1	1	1	1	…	0	0	0	1	1	1
e4	0	0	1	1	1	0	...	0	0	1	1	1	0
e5	1	1	0	1	0	1	…	0	0	0	1	1	1
…	…	…	…	…	…	…	…	…	…	…	…	…	…

Frequent pattern tree

Graph set with more than 5 members
and with > 1000 common edges

{G1, G3, G5, G6, G7,…} {G2, G3, G5, G7, G8,…} …. {G8, G9, G15, G26, G29}

common edges common edges common edges

{e1, e10, e56, e100, e1000,…} {e4, e12, e33, e56, e890,…} …. {e99, e220, e1545, e2629,…}

Very time consuming! It takes more than 2 weeks on 40 Pentium IV nodes

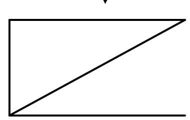
Expanding the Biclusters

E	G1	G2	G3	G4	G5	G6	G7	…	G61	G62	G63	G64	G65
e1	1	1	0	1	1	1	0	0	0	1	1	1	1
e2	1	1	1	1	1	0	1	0	1	0	1	1	1
e3	1	1	1	1	1	1	0	0	0	0	1	1	1
e4	1	1	1	1	1	1	.0	0	0	1	1	1	0
e5	1	0	0	1	0	1	1	0	0	0	1	1	1
…	…	…	…	…	…	…	…	…	…	…	…	…	…

Simulated Annealing

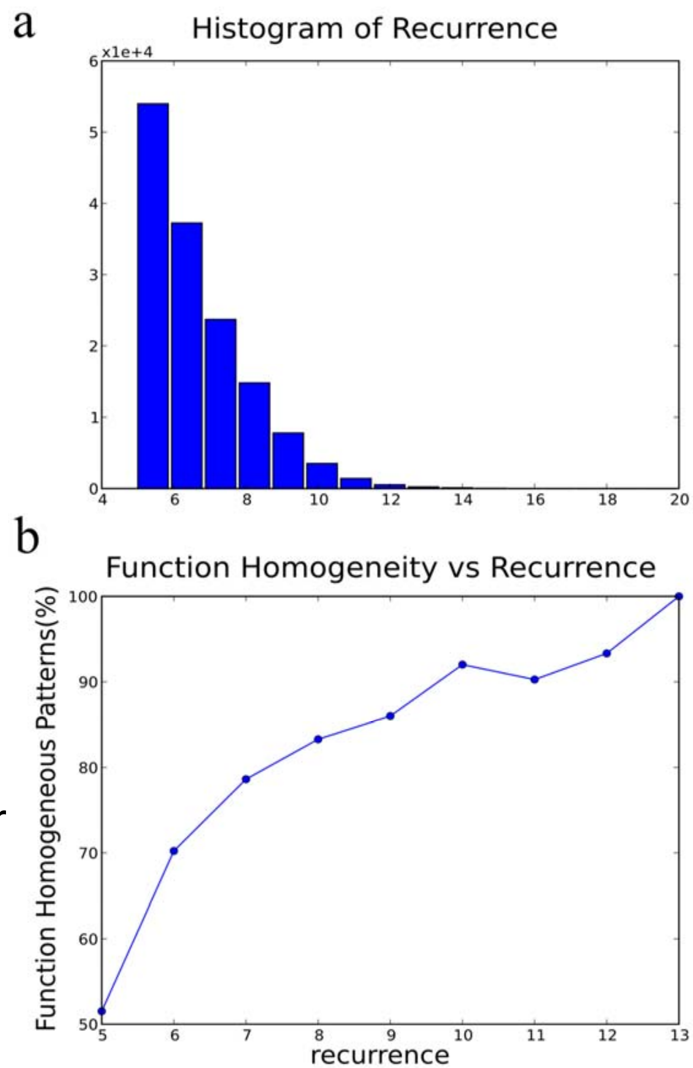
E	G1	G2	G3	G4	G5	G6	G7	…	G61	G62	G63	G64	G65
e1	1	1	0	1	1	1	0	0	0	1	1	1	1
e2	1	1	1	1	1	0	1	0	1	0	1	1	1
e3	1	1	1	1	1	1	0	0	0	0	1	1	1
e4	1	1	1	1	1	1	.0	0	0	1	1	1	0
e5	1	0	0	1	0	1	1	0	0	0	1	1	1
…	…	…	…	…	…	…	…	…	…	…	…	…	…

Identify connected components



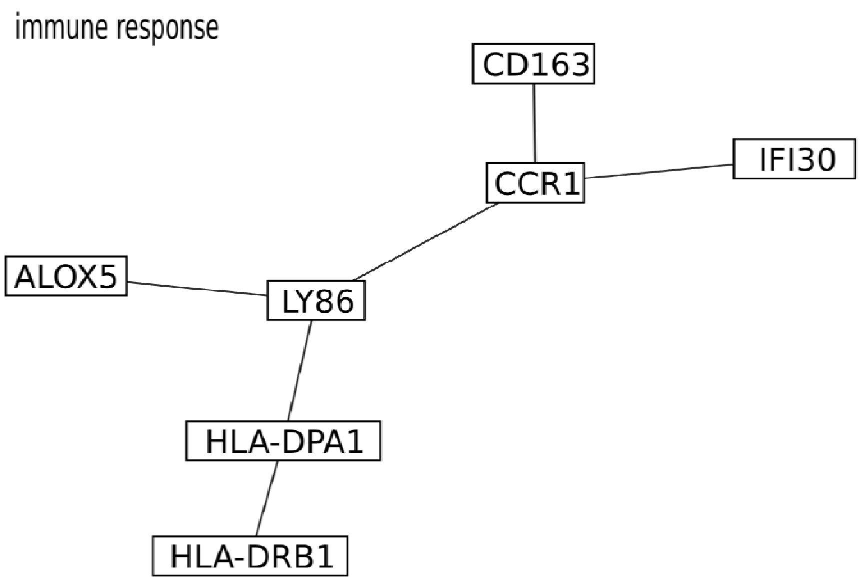
Systematic identification of functional modules in human genome

- We identified 143,400 network modules with recurrence >= 5. They vary in size from 4 to 180.
- 77.0% of the patterns are functionally homogenous (GO hyper-geometric *P*-value less than 0.01)
- Figure (a) shows the histogram of network recurrence, which resembles an exponential distribution..
- Figure (b) shows that the functional homogeneity of modules increase with their recurrences.

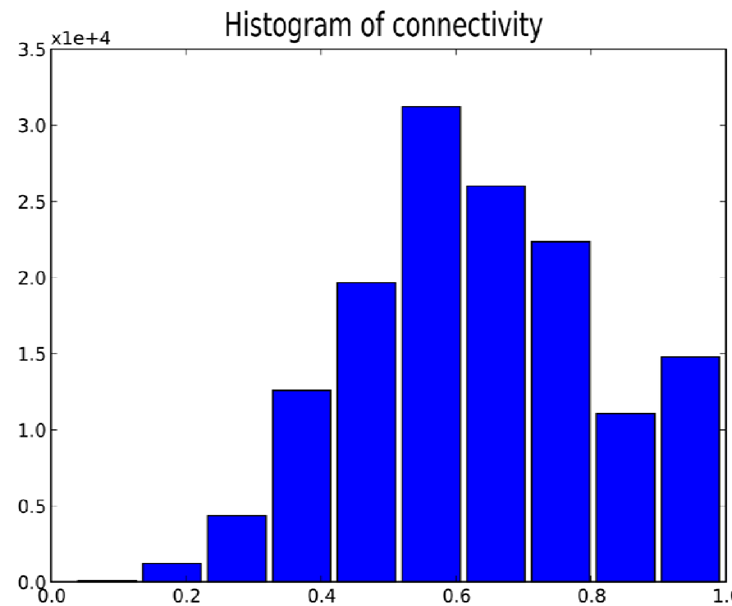


Loosely connected network patterns with high recurrence can represent functional modules

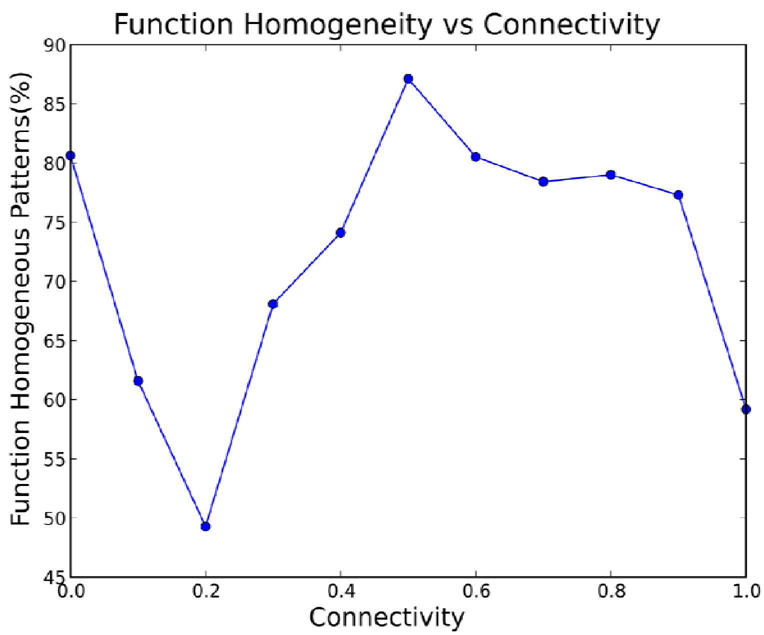
a



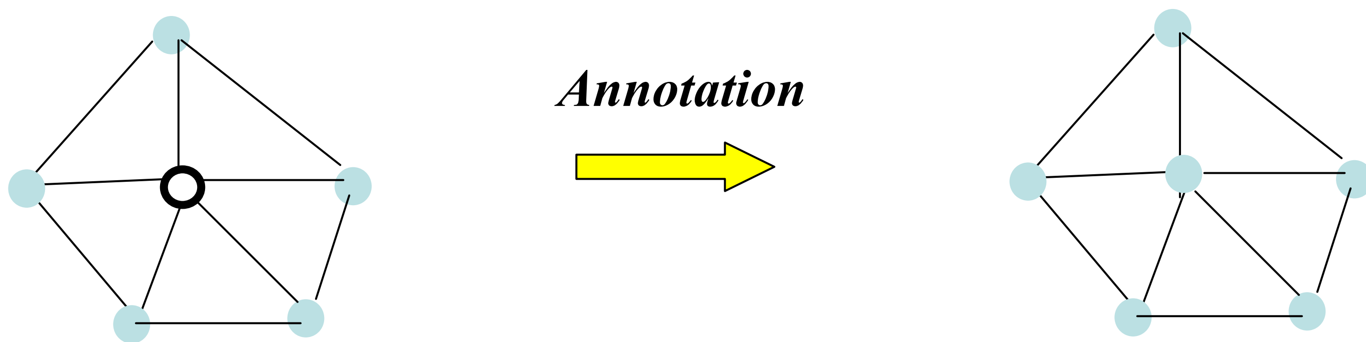
b



c



Functional annotation



We made functional predictions for 779 known and 116 unknown genes by random forest classification with 71% accuracy.

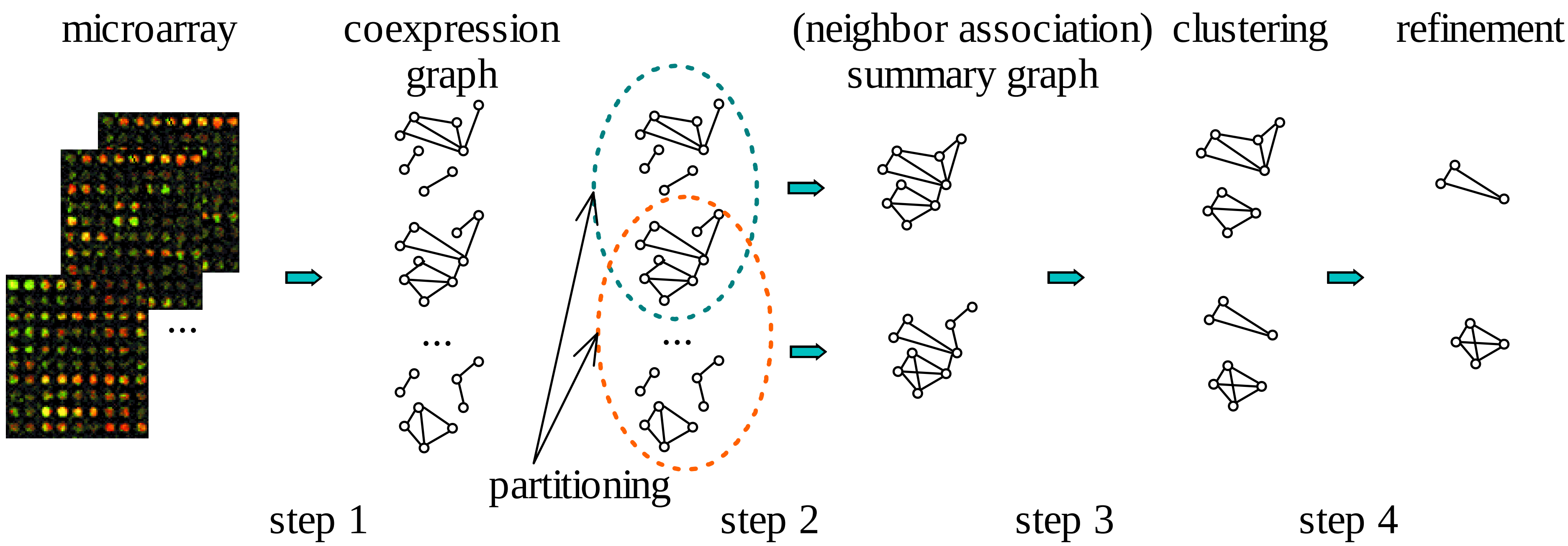
Variables for random forest classification:

functional enrichment P-value
network connectivity
average node degree
Network size

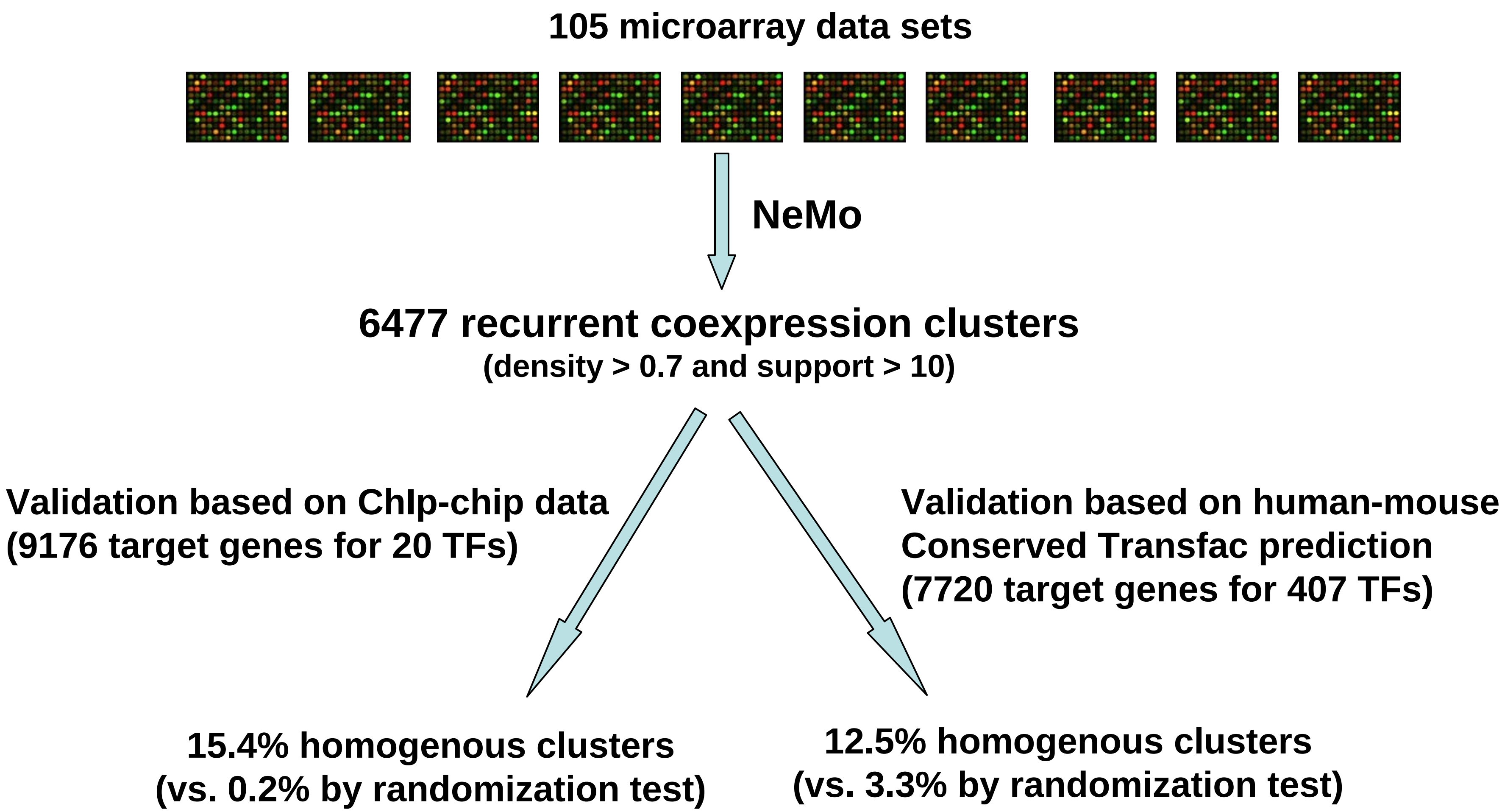
network topology score
pattern recurrence numbers
unknown gene ratio

Network Modules (NeMo)

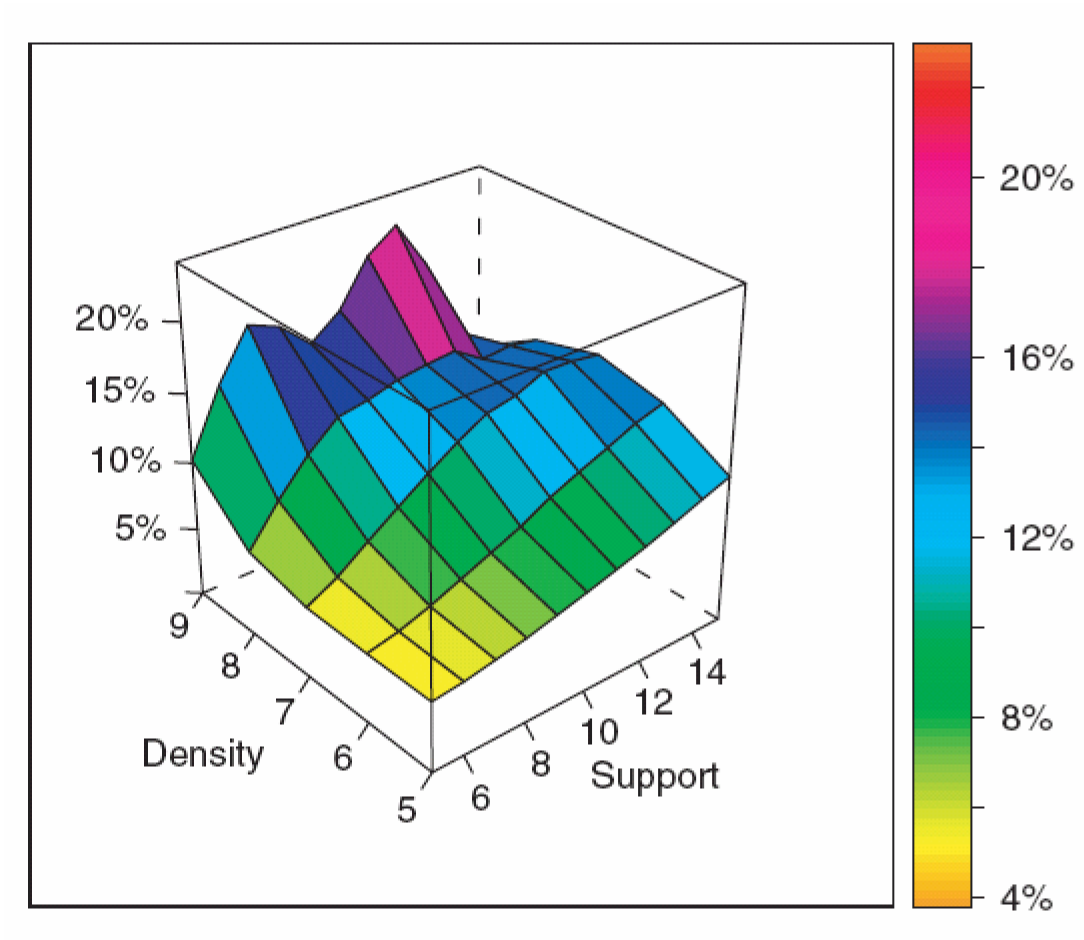
Identify <u>frequent dense vertex sets</u> across many massive graphs



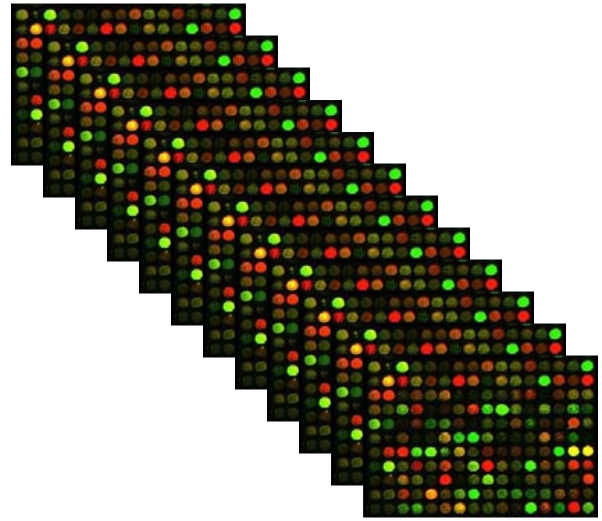
Yan et al. ISMB 2007


105 microarray data sets
NeMo
6477 recurrent coexpression clusters
(density > 0.7 and support > 10)
Validation based on ChIp-chip data
(9176 target genes for 20 TFs)
Validation based on human-mouse
Conserved Transfac prediction
(7720 target genes for 407 TFs)
15.4% homogenous clusters
(vs. 0.2% by randomization test)
12.5% homogenous clusters
(vs. 3.3% by randomization test)

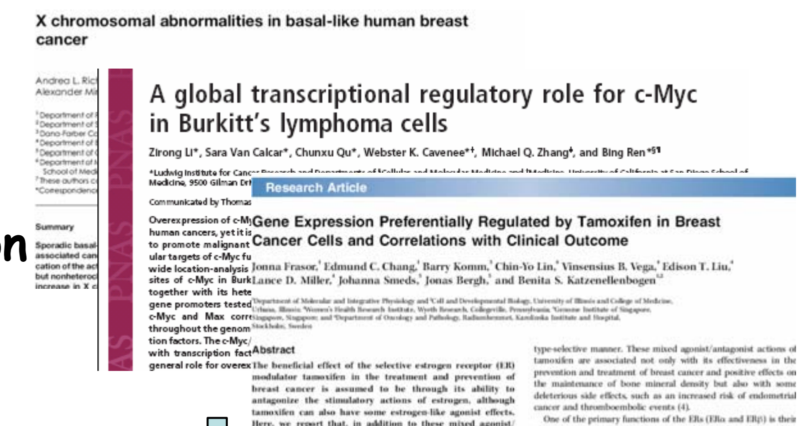
Percentage of potential transcription modules validated by ChIP-Chip data increase with cluster density and recurrence



Expression data



Microarray Data

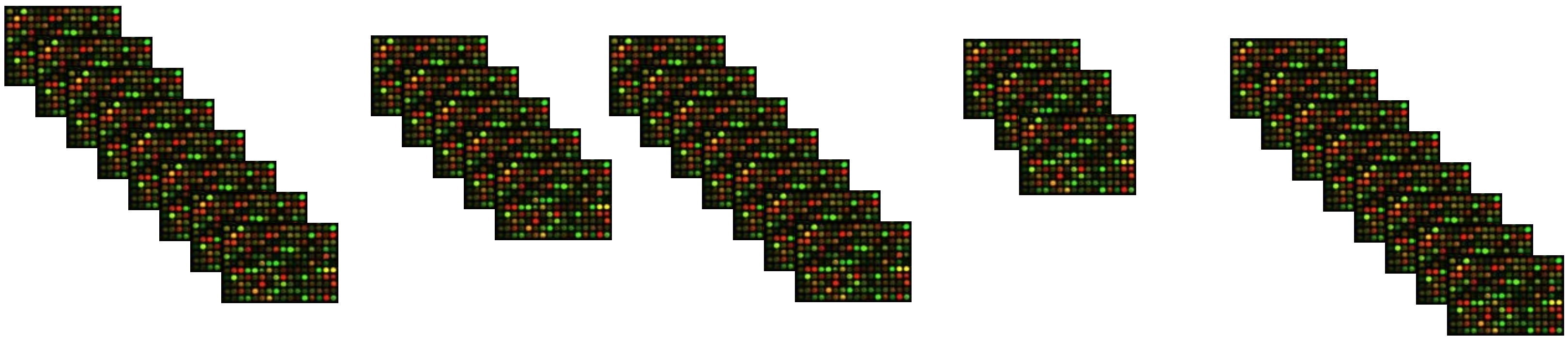


Phenotype information

Phenotype Concepts (e.g. diseases, perturbations, tissues)
in Unified Medical Language System (UMLS)

Classifying microarray data based on phenotype

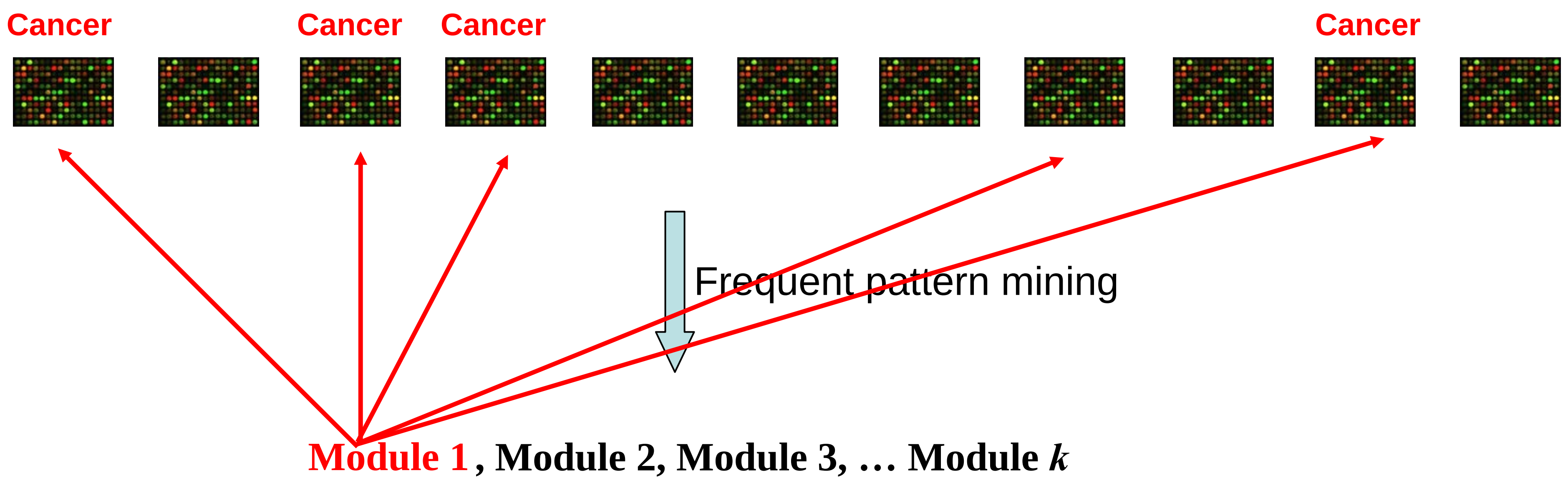
Adenocarcinoma Arthritis Asthma ··· Glaucoma HIV



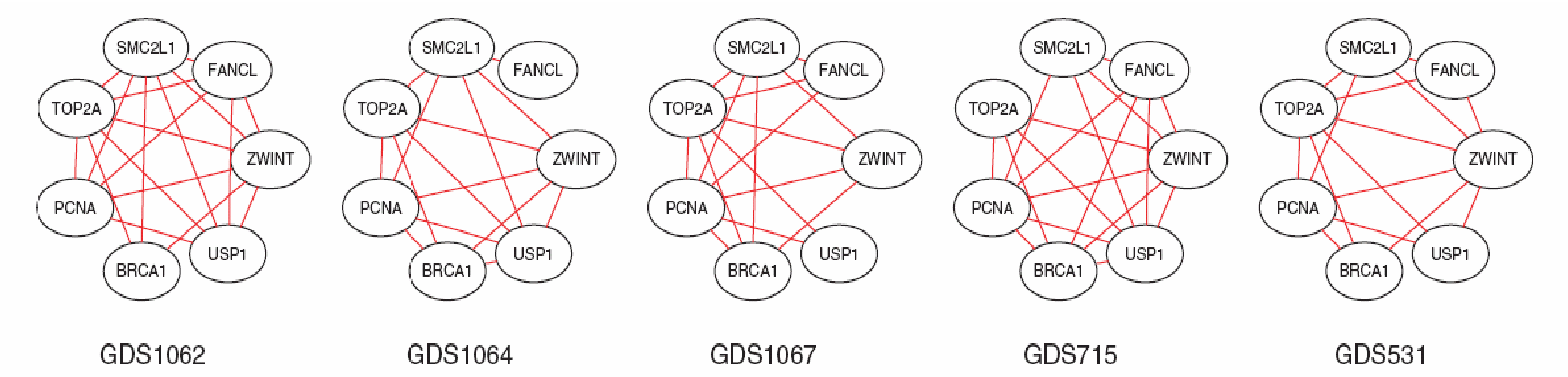
For example, the current NCBI GEO database contains >60 cancer datasets, among which 11 leukemia datasets.

Identify phenotype-specific functional or transcriptional modules

- Unsupervised approach



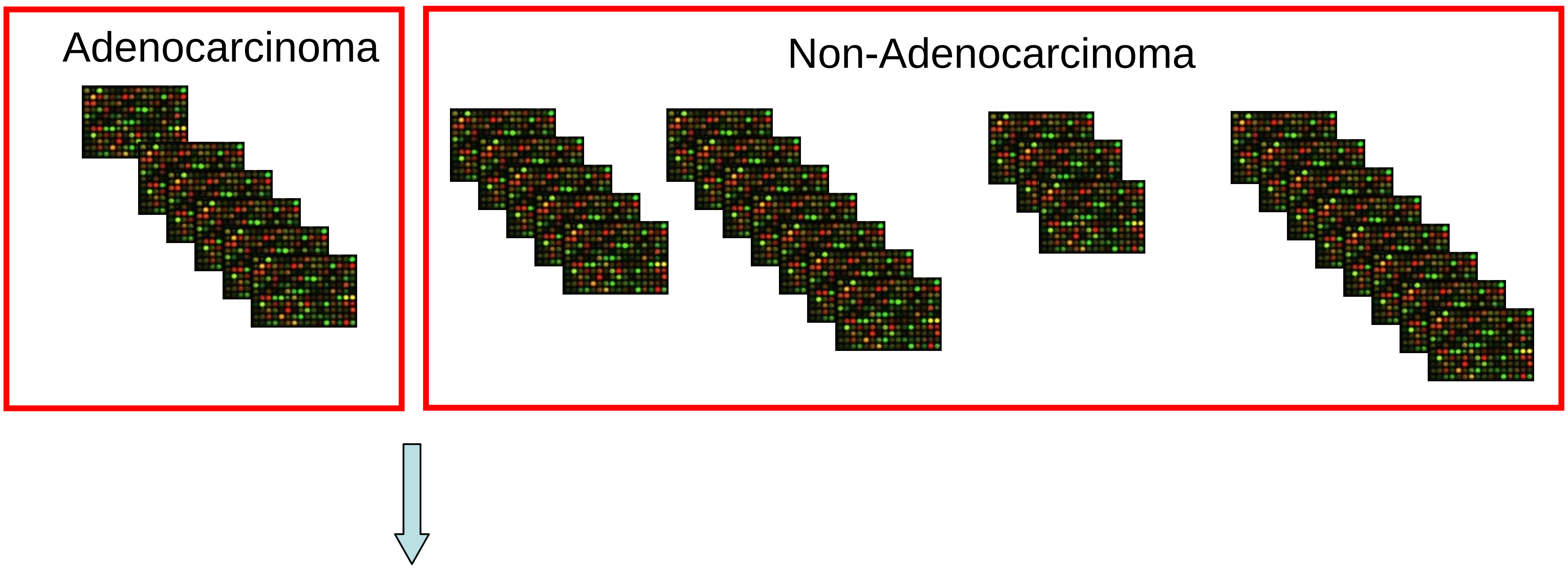
An example



5 out of the 9 support datasets are leukemia datasets (*P*-value 0.0039). It is potentially regulated by E2F4, and majority genes are involved in cell cycle and DNA repair.

Identify phenotype-specific functional or transcriptional modules

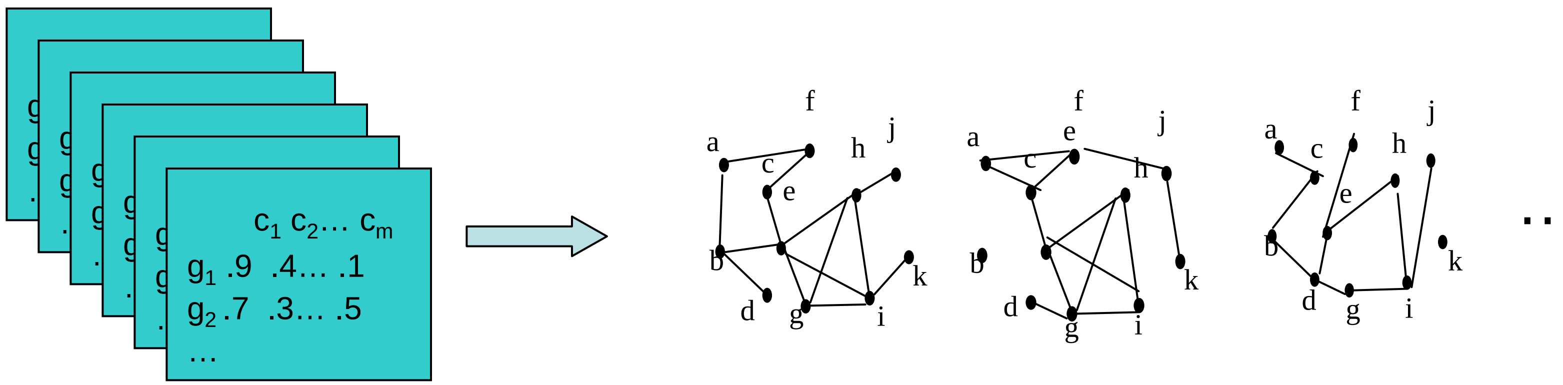
- Supervised approach



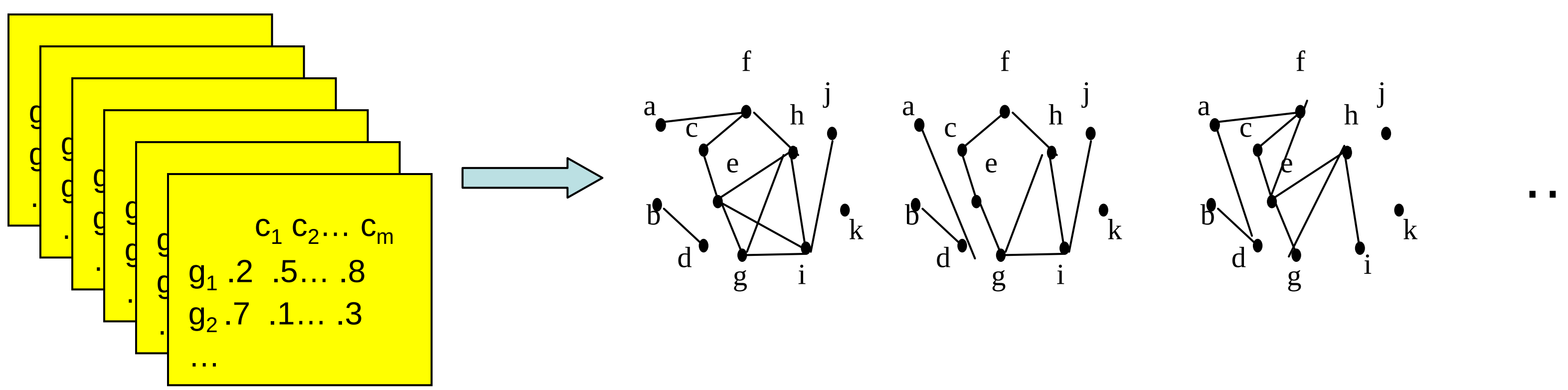
Functional and transcriptional modules
which are active ONLY in Adenocarcinoma
Related data sets

A case study: Identify Network Modules Characterizing Cancer

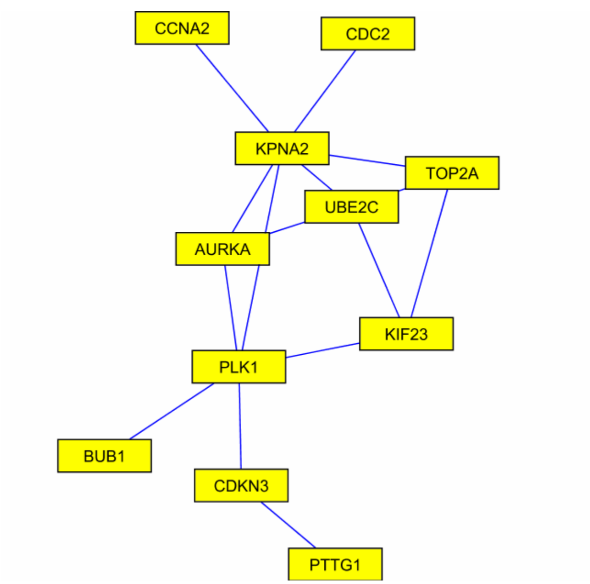
32 Cancer Datasets



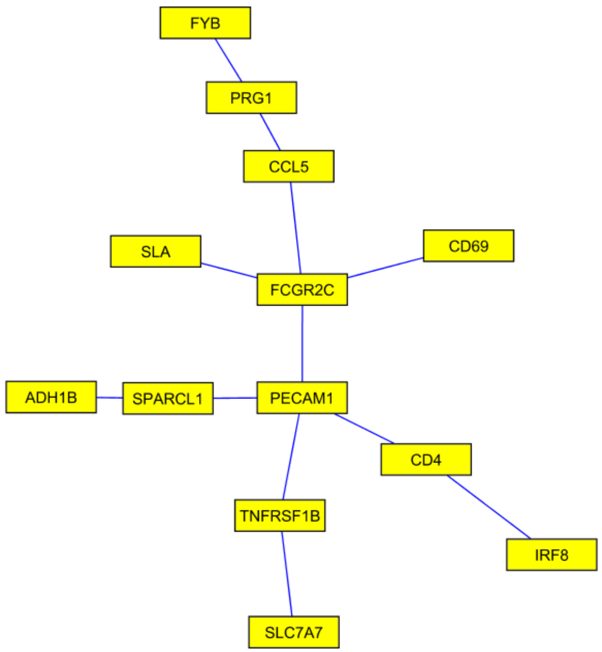
25 Non-cancer Datasets



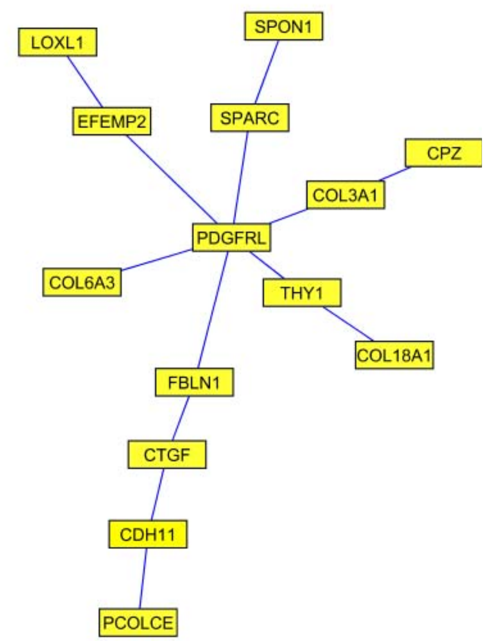
Examples of identified modules



Cell cycle Module
across all cancer datasets



Cell adhesion
across all solid tumor datasets

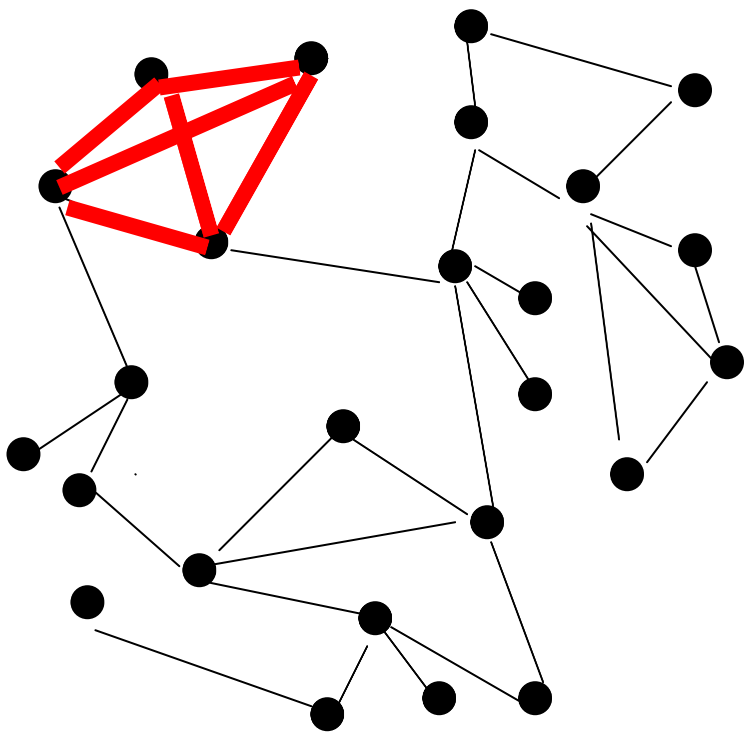


PDGF-signaling
in breast cancer datasets

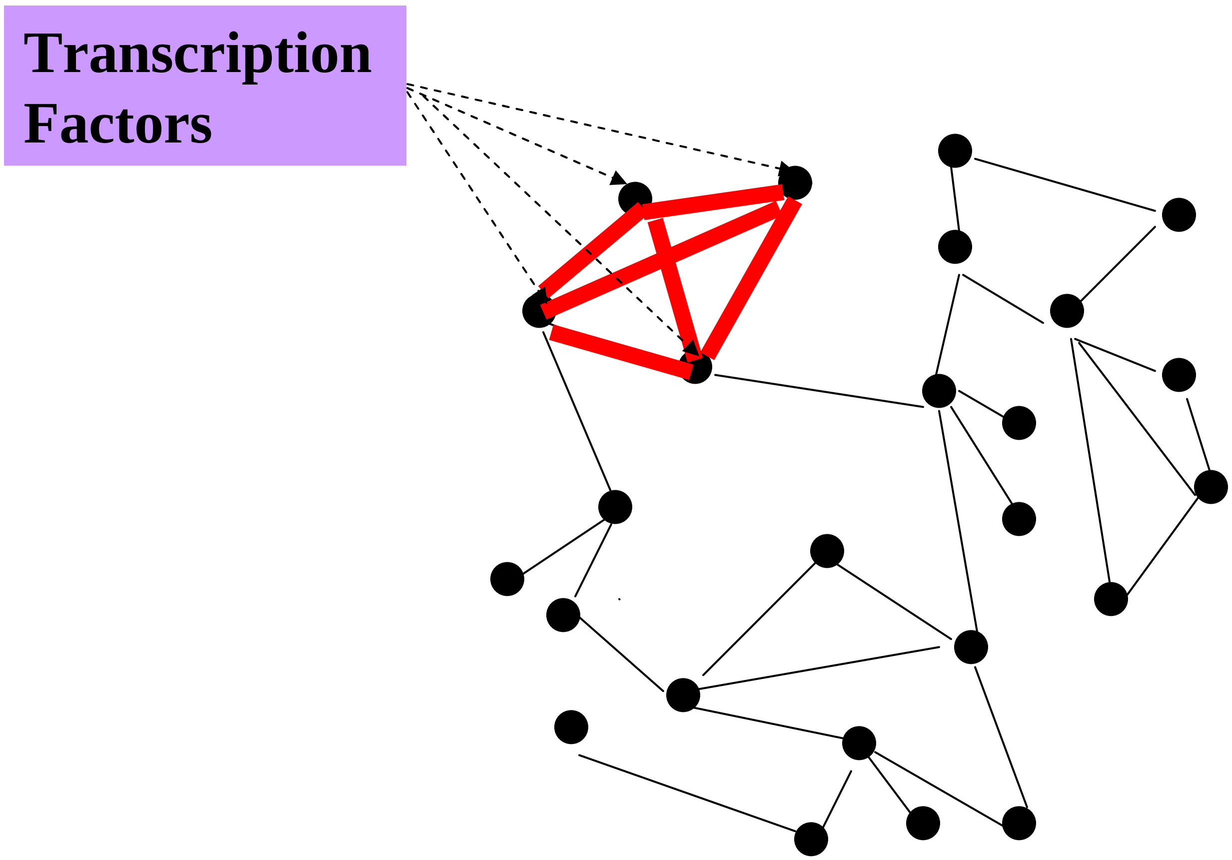
Reconstruct transcriptional cascades by second-order correlation

Zhou et al. Nature Biotech 2005

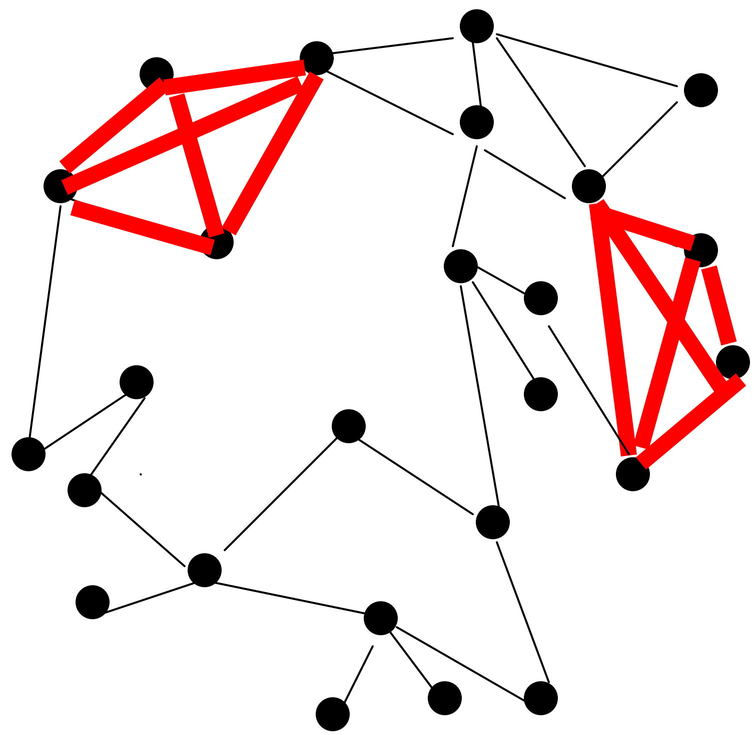
Frequently occurring tight clusters



Frequently occurring tight clusters

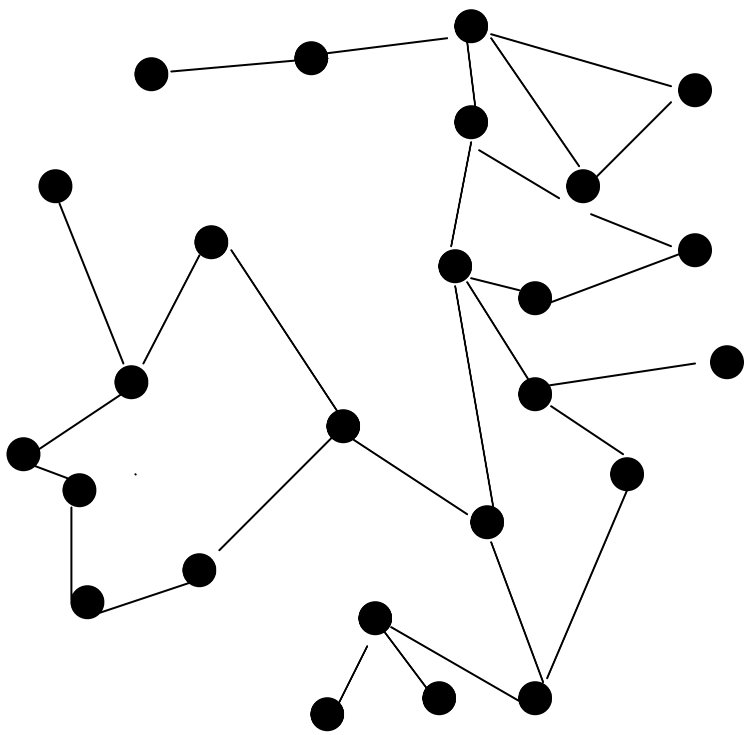


Co-occurrence of tight clusters



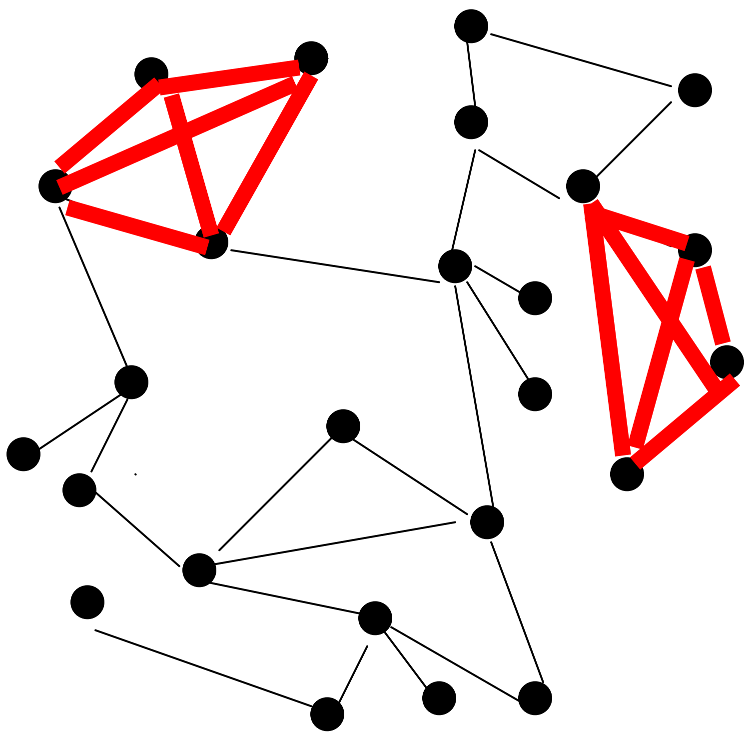
Coexpression network constructed with the dataset 1

Co-occurrence of tight clusters



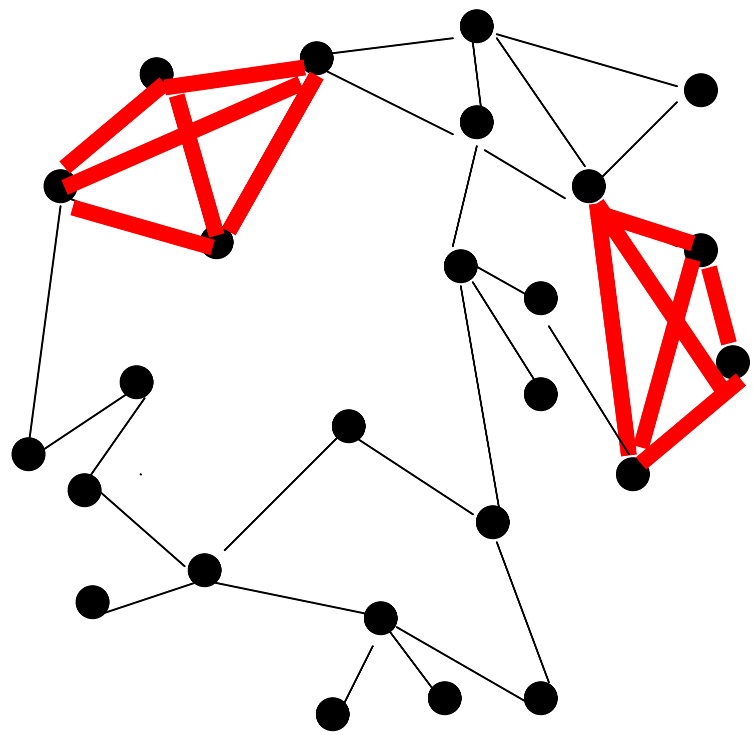
Coexpression network constructed with the dataset 2

Co-occurrence of tight clusters



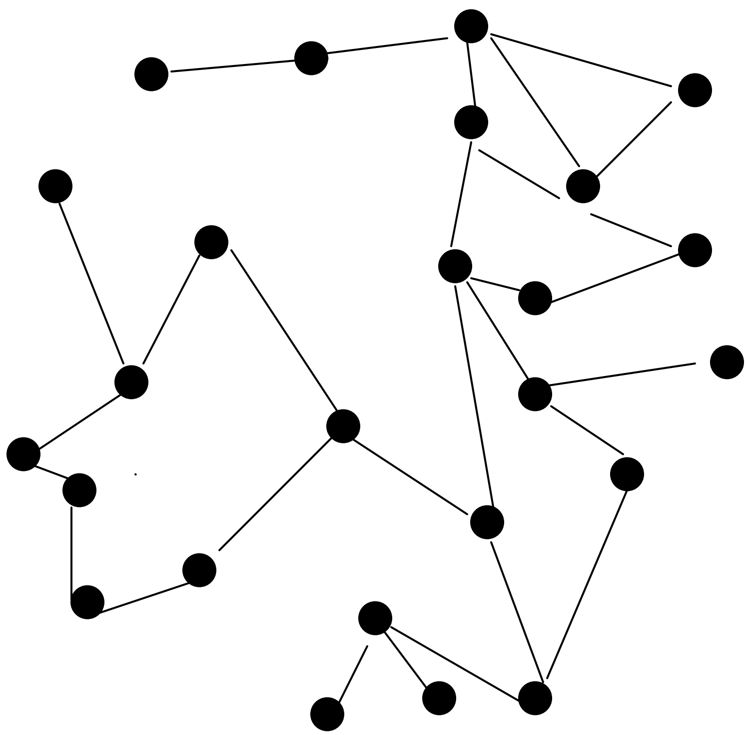
Coexpression network constructed with the dataset 3

Co-occurrence of tight clusters



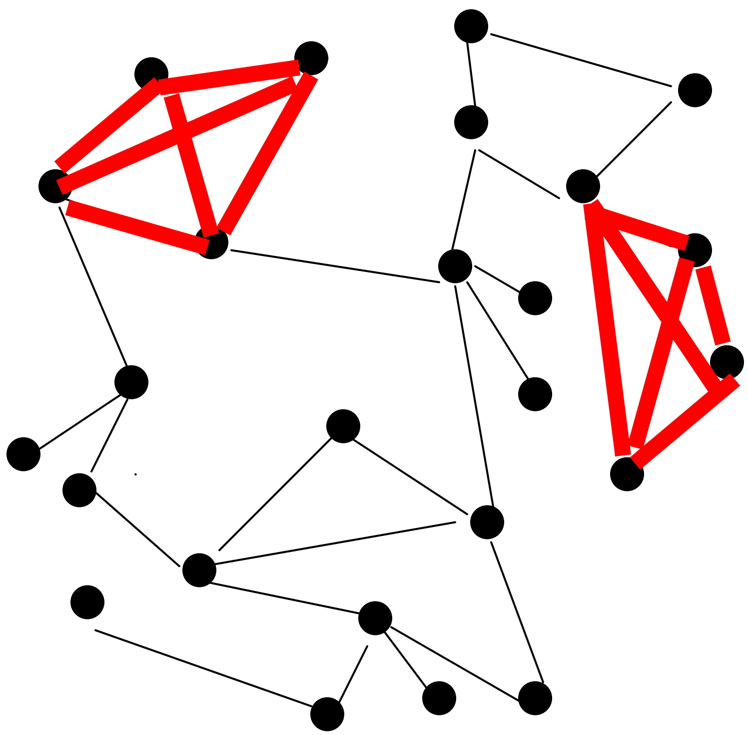
Coexpression network constructed with the dataset 1

Co-occurrence of tight clusters



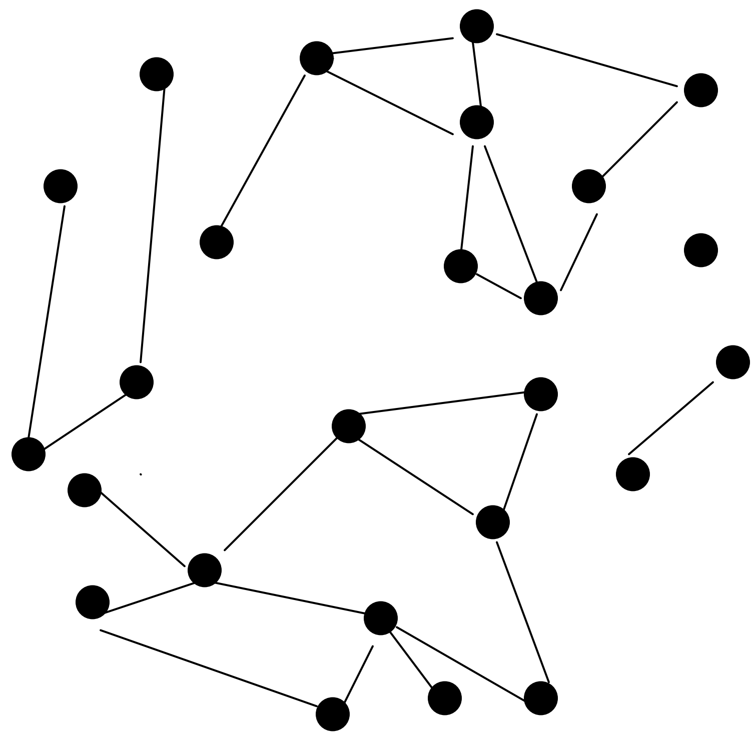
Coexpression network constructed with the dataset 2

Co-occurrence of tight clusters



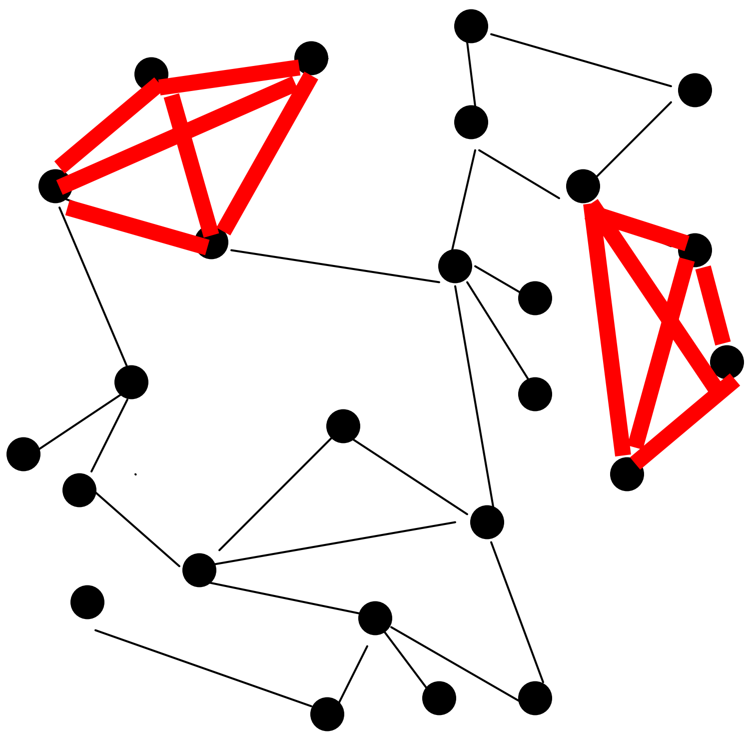
Coexpression network constructed with the dataset 3

Co-occurrence of tight clusters

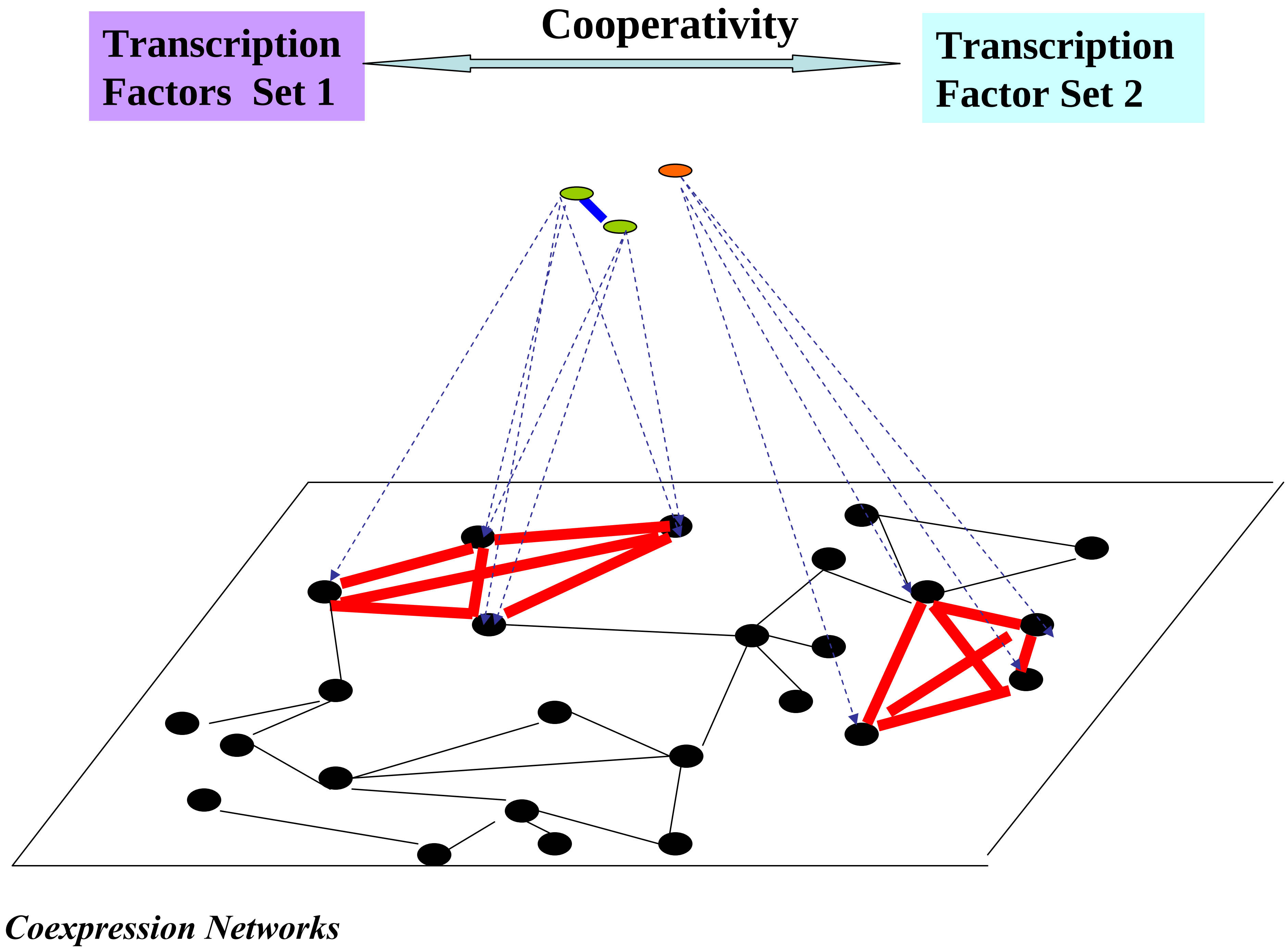


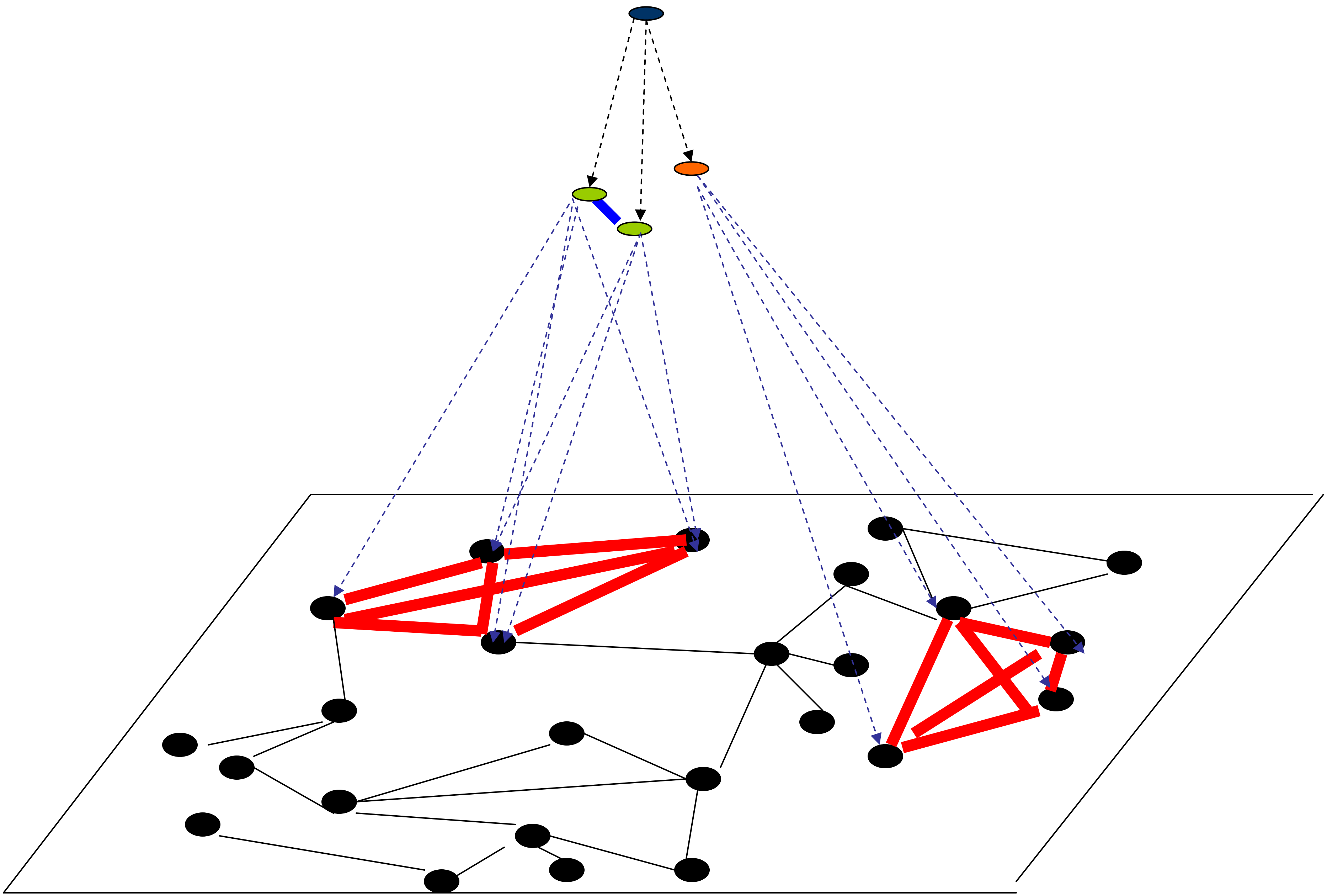
Coexpression network constructed with the dataset 4

Co-occurrence of tight clusters



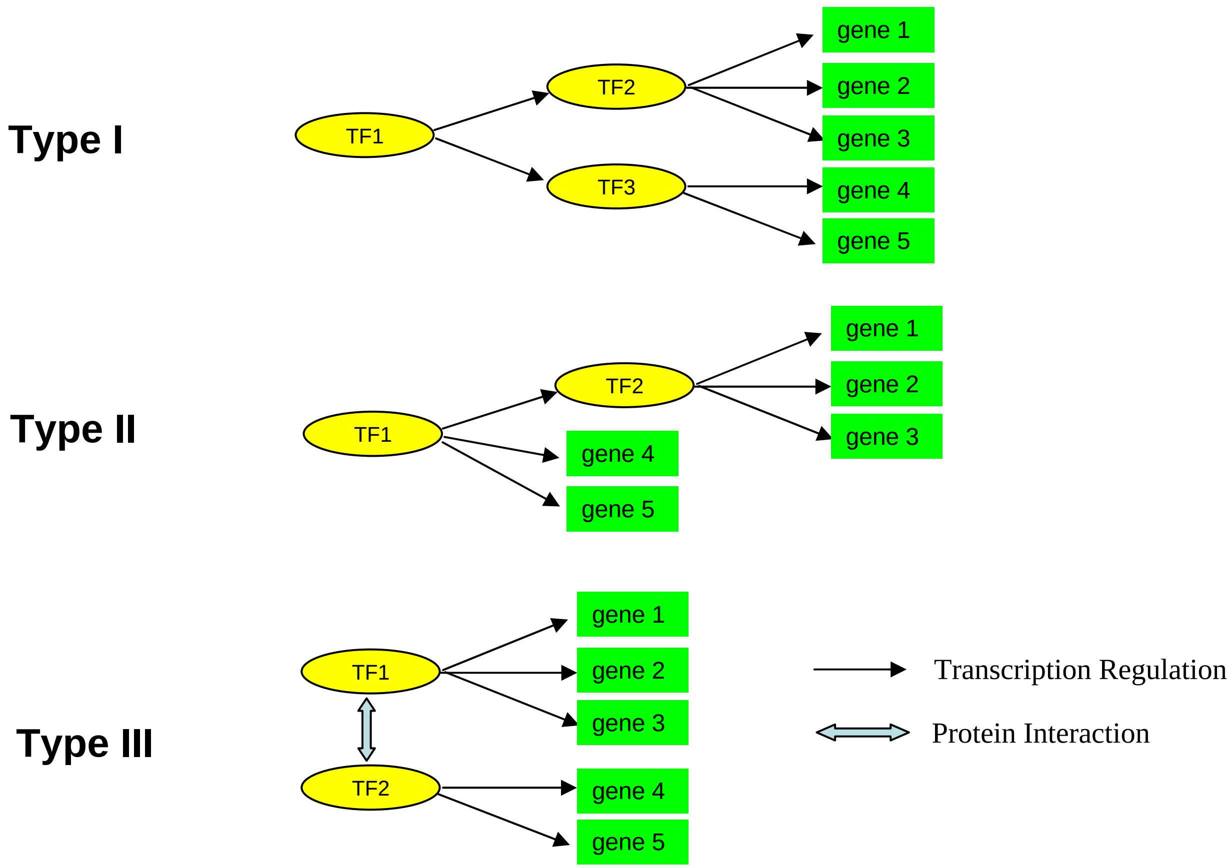
Coexpression network constructed with the dataset 5


Transcription Factors Set 1
Cooperativity
Transcription Factor Set 2
Coexpression Networks



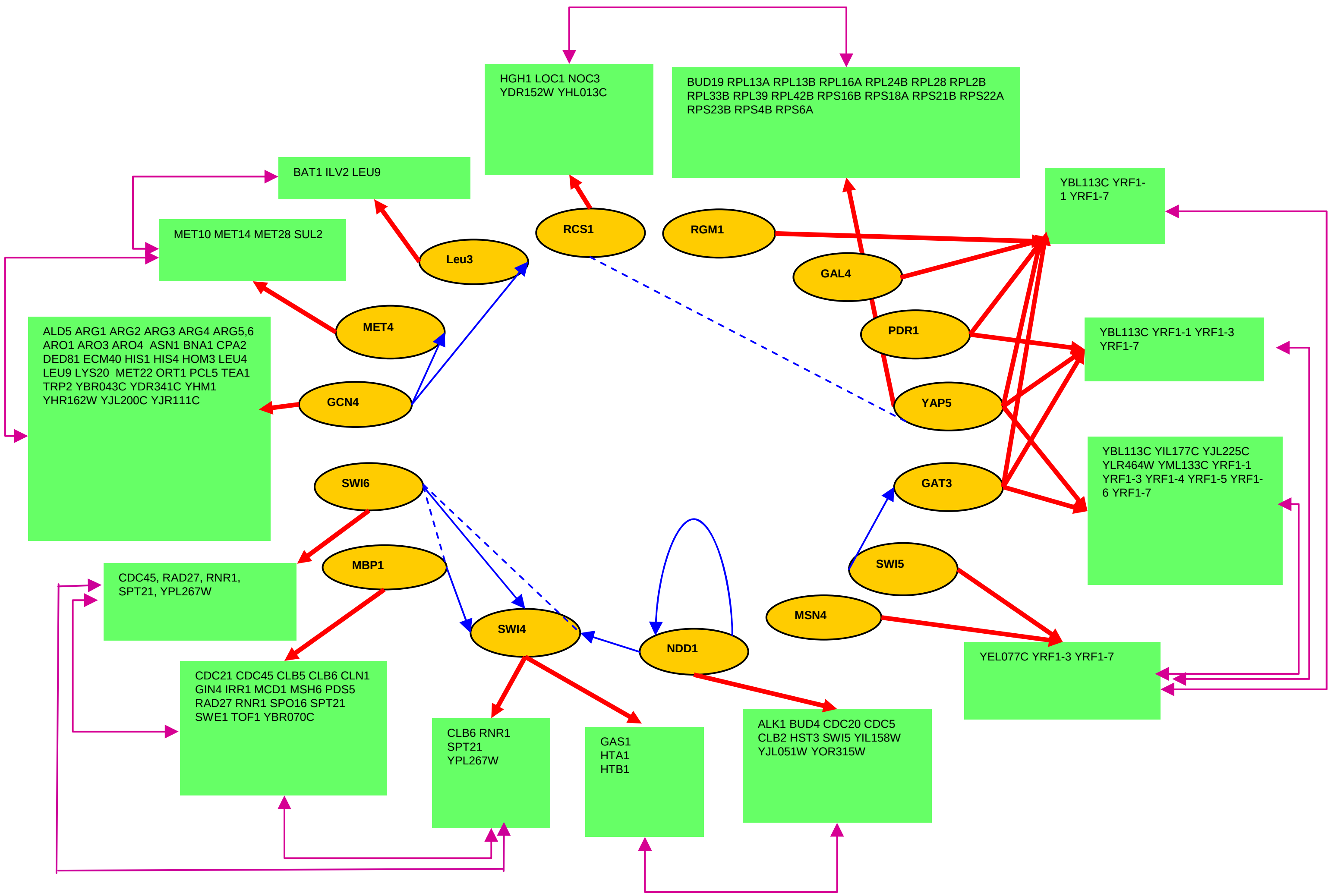
Relevance Networks

Three types of transcription cascades



Applying to 39 yeast microarray data sets

- We identified 60 transcription modules. Among them, we found 34 pairs that showed high 2nd-order correlation. A significant portion (29%, *p*-value<10-5 by Monte Carlo simulation) of those modules pairs are participants in transcription cascades: 2 pairs in Type I, 8 pairs in Type II, and 3 pairs in type III cascades. In fact, these transcription cascades inter-connect into a partial cellular regulatory network.



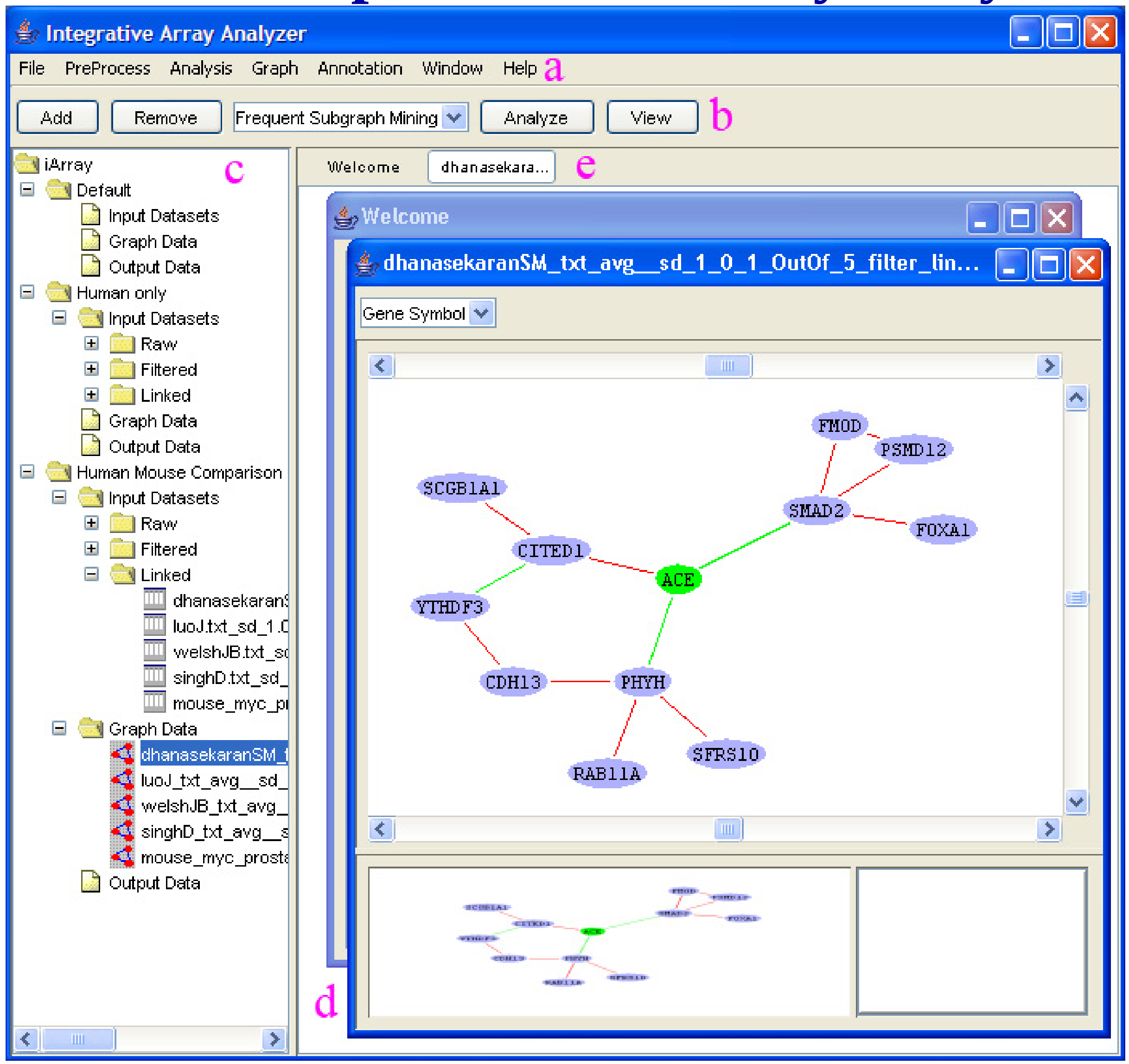
Regulation of transcription modules by transcription factors, based on ChIP-chip data and supported by recurrent expression clusters

Regulation between transcription factors, based on ChIP-chip data and supported by 2nd-order expression correlation

Protein interactions between two transcription factors, based on experimental data and supported by 2nd-order expression correlation

Two transcription modules with high 2nd-order expression correlations

Integrative Array Analyzer (*i*Array): a software package for cross-platform and cross-species microarray analysis



Bioinformatics. 22(13):1665-7

Acknowledgement

- Haiyan Hu
- Yu Huang
- Haifeng Li
- Xifeng Yan (IBM)
- Mike Mehan
- Min Xu
- Min-Chih Kao (Michigan)
- Juan Nunez-Iglesias
- Mrinal Kalakrishnan
- Michael Waterman
- Jiawei Han (UIUC)
- Haiyan Huang (UC Berkeley)
- Wing H. Wong (Stanford)

Funding agencies: NIGMS, NCI, NSF, Seaver Foundation, Sloan Foundation, Zumberg Foundation

Thank you!