# *Network Modules Identification Biclustering*

Viktorian Miok

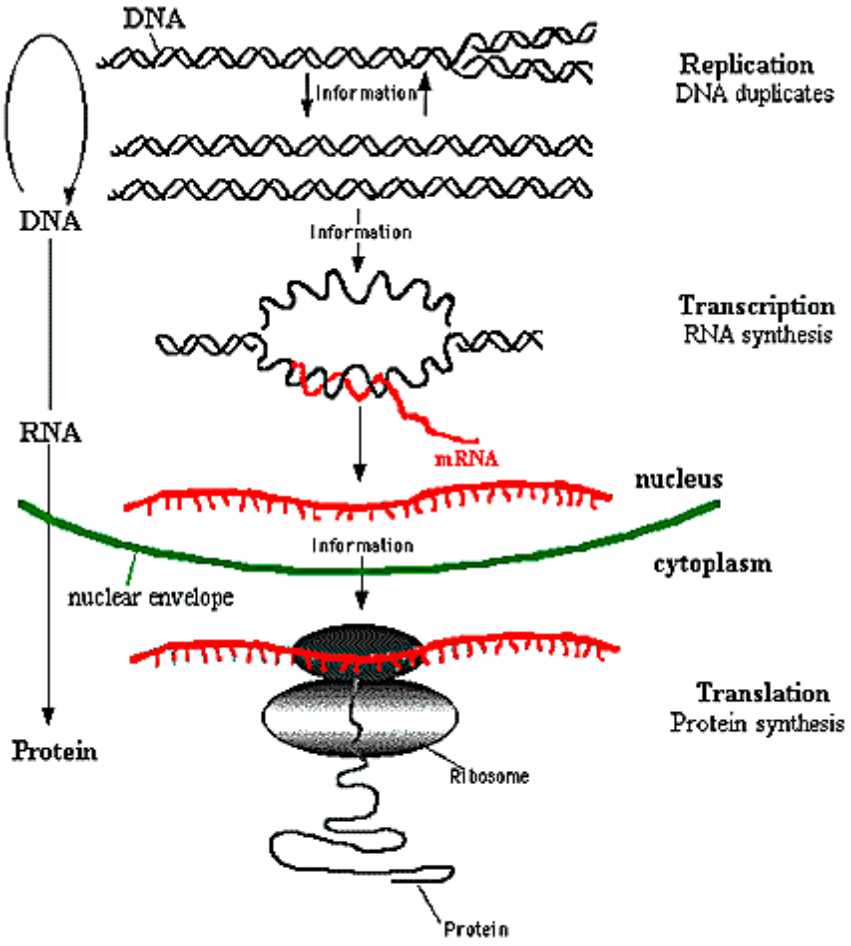
# Overview

- Microarray technology
- Gene expression data sets
- Clustering techniques
- Biclustering
- SAMBA – Statistical Algorithmic Method for Biclustering Analysis

# Central Dogma of Biology

- Central Dogma of Molecular Biology describes the information transfer process that leads from the information encoded in DNA to the proteins in the cell.

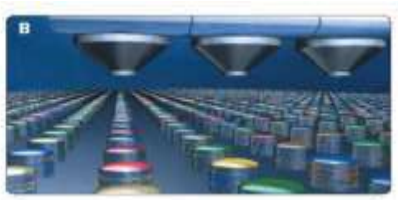
- Three steps are discerned:
  1) Replication
  2) Transcription
  3) Translation



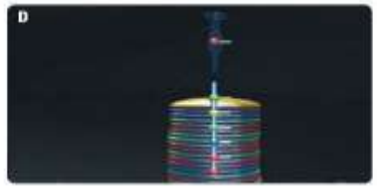
The Central Dogma of Molecular Biology

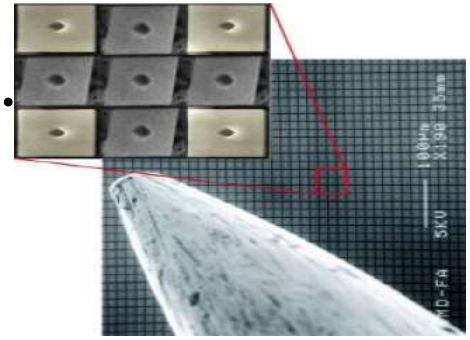
# Microarray technology

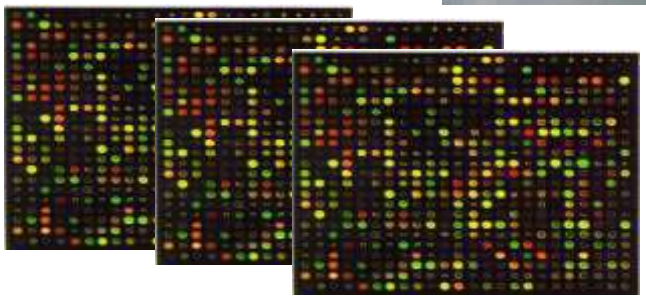
- Conceptually: a measurement device.
- Microarray help understanding of biological process.
- Revolutionize biological research.
- Types of microarrays measure:
  - Gene expression (mRNA, miRNA)
  - DNA copy number
  - Methylation





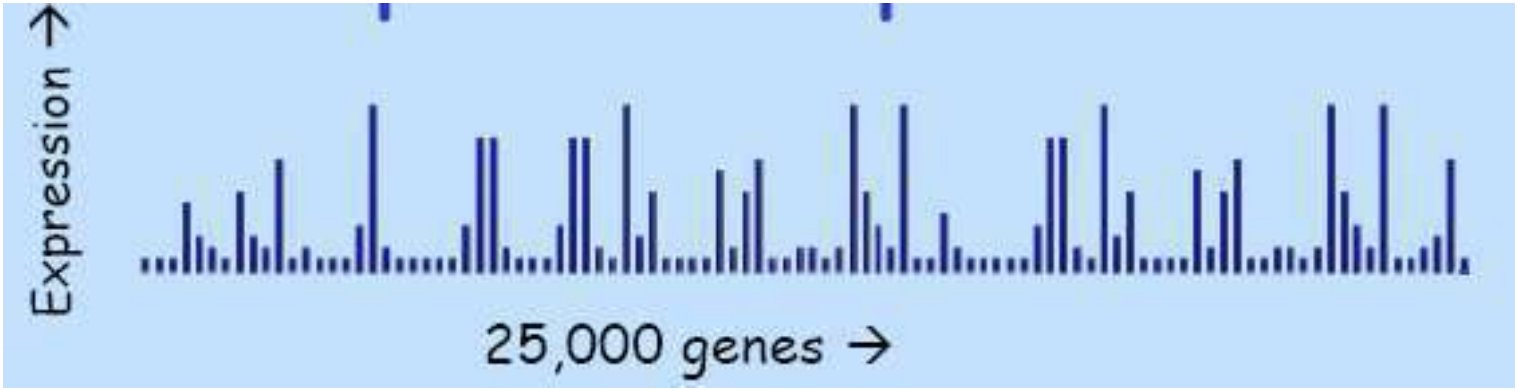





# Gene expression

- Gene expression arrays measure the expression of genes (which genes are expressed and to what extent).

- In fact, it measures mRNA which is related – through the transcription process – to the expression of genes.



# Gene expression data sets

- Row – gene's expression patterns.
- Column – experiment condition's profile.

conditions

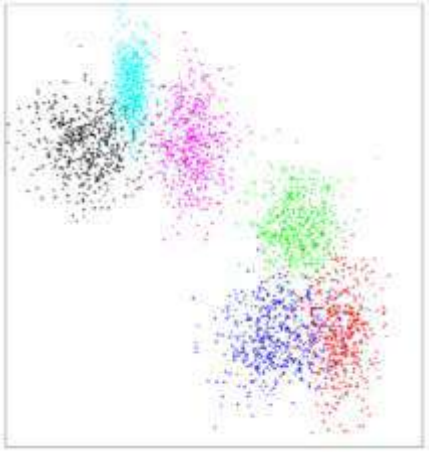
probes

| AffyID | Symbol | t0 | t2 | t4 | t6 | t8 |
|---|---|---|---|---|---|---|
| 1007_s_at | DDR1 | 105.265 | 63.89688 | 112.695 | 150.5448 | 86.05 |
| 1053_at | RFC2 | 30.19 | 18.05 | 31.185 | 47.8044 | 30.04 |
| 121_at | PAX8 | 238.915 | 143.2531 | 234.33 | 377.6472 | 219.735 |
| 1294_at | UBE1L | 119.495 | 53.6125 | 110.175 | 144.7908 | 79.285 |
| 1316_at | THRA | 30.19 | 18.05 | 31.185 | 47.8044 | 29.77 |
| 1431_at | CYP2E1 | 30.19 | 18.05 | 31.185 | 47.8044 | 29.385 |
| 1438_at | EPHB3 | 77.255 | 52.975 | 47.25 | 84.4116 | 49.955 |
| 1487_at | ESRRA | 65.22 | 36.47188 | 78.625 | 115.542 | 73.31 |
| 1494_f_at | CYP2A6 | 58.23 | 30.71563 | 53.73 | 84.3612 | 55.515 |
| 1552256_a | SCARB1 | 116.085 | 58.225 | 120.74 | 167.9496 | 97.03 |
| 1552257_a | KIAA0153 | 75.455 | 38.14375 | 75.485 | 172.536 | 133.72 |
| 1552264_a | MAPK1 | 130.305 | 70.54375 | 161.525 | 219.0132 | 125.225 |
| 1552274_a | PXK | 130.85 | 63.34375 | 62.34 | 56.826 | 31.9 |
| 1552275_s | PXK | 131.01 | 56.99375 | 45.085 | 49.4508 | 30.3 |
| 1552277_a | MGC1733 | 139.87 | 123.425 | 325.305 | 538.104 | 321.925 |
| 1552279_a | MGC9564 | 86.975 | 44.50938 | 71.93 | 115.248 | 78.465 |
| 1552283_s | ZDHHC11 | 30.19 | 18.05 | 31.185 | 47.8044 | 29.385 |
| 1552287_s | AFG3L1 | 30.19 | 18.05 | 31.185 | 47.8044 | 29.385 |
| 1552291_a | FLJ20522 | 30.19 | 18.05 | 31.185 | 47.8044 | 29.385 |
| 1552295_a | SLC39A13 | 397.14 | 178.4313 | 378.815 | 613.7208 | 399.91 |

| ProbeName | log2ratio |
|---|---|
| A_23_P204891 | 1.767202161 |
| A_32_P199884 | 2.831352274 |
| A_24_P143492 | -0.097623193 |
| A_24_P863124 | 0.389514597 |
| A_23_P55897 | -0.277791144 |
| A_32_P18475 | -2.153193648 |
| A_32_P140139 | 0.983097028 |
| A_23_P14105 | 1.675228728 |
| A_23_P4353 | -0.781669369 |
| A_23_P25235 | -0.254786044 |
| A_23_P155688 | 0.669247909 |
| A_23_P204187 | -0.985180564 |
| A_32_P2157 | -0.669712220 |
| A_23_P52697 | 1.763815750 |
| A_23_P360777 | -0.208618717 |
| A_23_P410965 | 0.001498177 |
| A_24_P413126 | -0.362524635 |
| A_32_P91385 | -0.066849545 |
| ... | ... |

# Clustering techniques

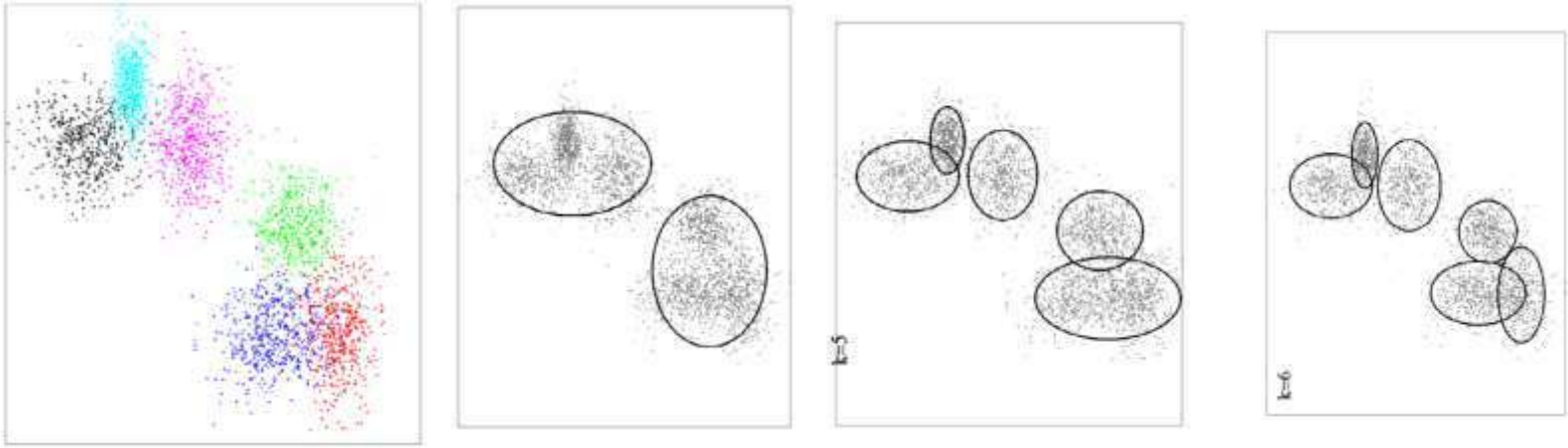
- Clustering - global partition of genes according to common expression pattern across all conditions.
- A set of entities which are alike.
- Ideal cluster compact and isolated.



- Some clustering techniques:
  - K-means
  - Hierarchical clustering
  - SOM

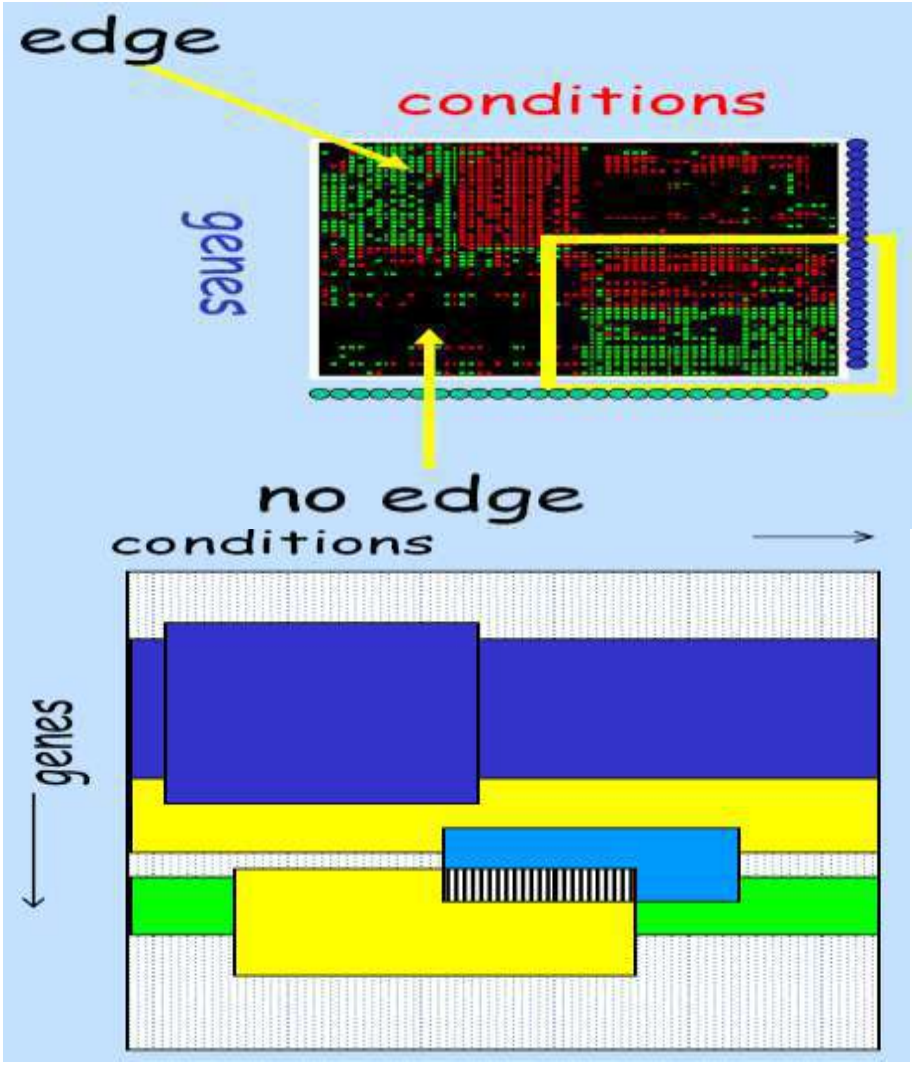
# Clustering objective

- Group elements need to satisfying:
  - Homogeneity – within each cluster elements are highly similar to each other.
  - Separation - different clusters have low similarity between each other.

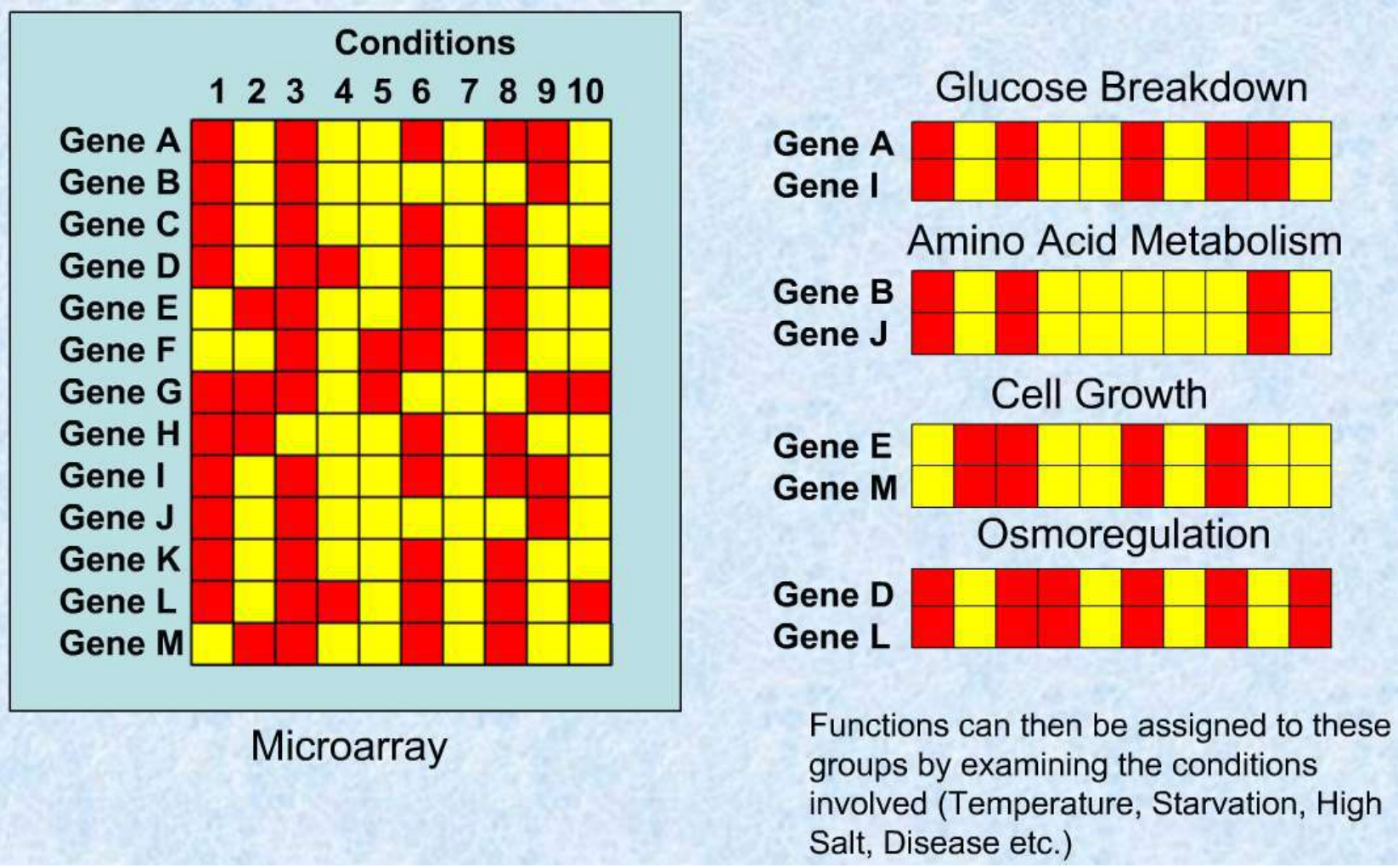


# Bisclustering

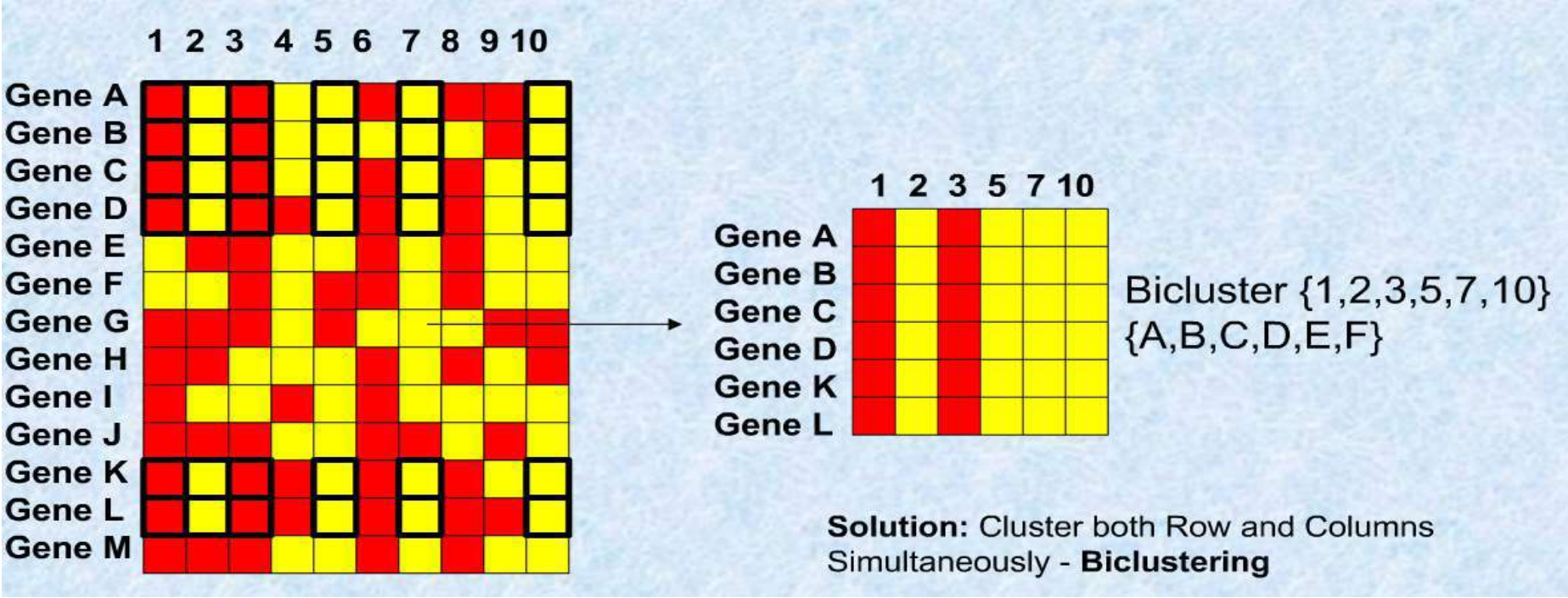
- Standard clustering oversimplified detection of refined local approach.
- Biclustering perform clustering in the two dimension simultaneously.
- Bicluster subset of genes and conditions



# Biclustering vs. Clustering



# Biclustering vs. Clustering

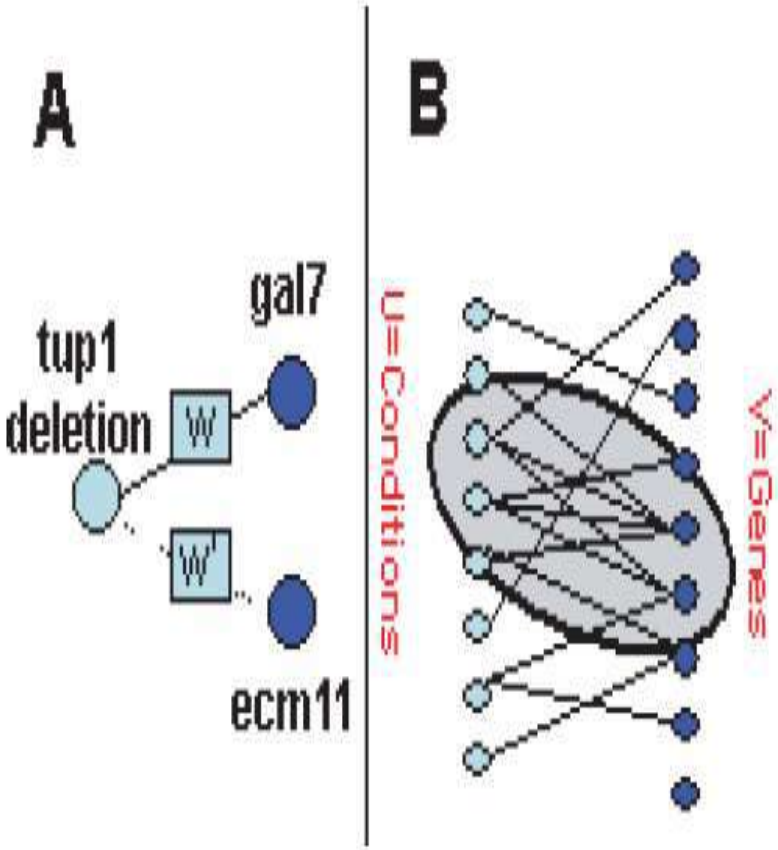



- The problem of searching biclusters is **NP-hard** with the searching space increase exponentially with the object/attributes numbers

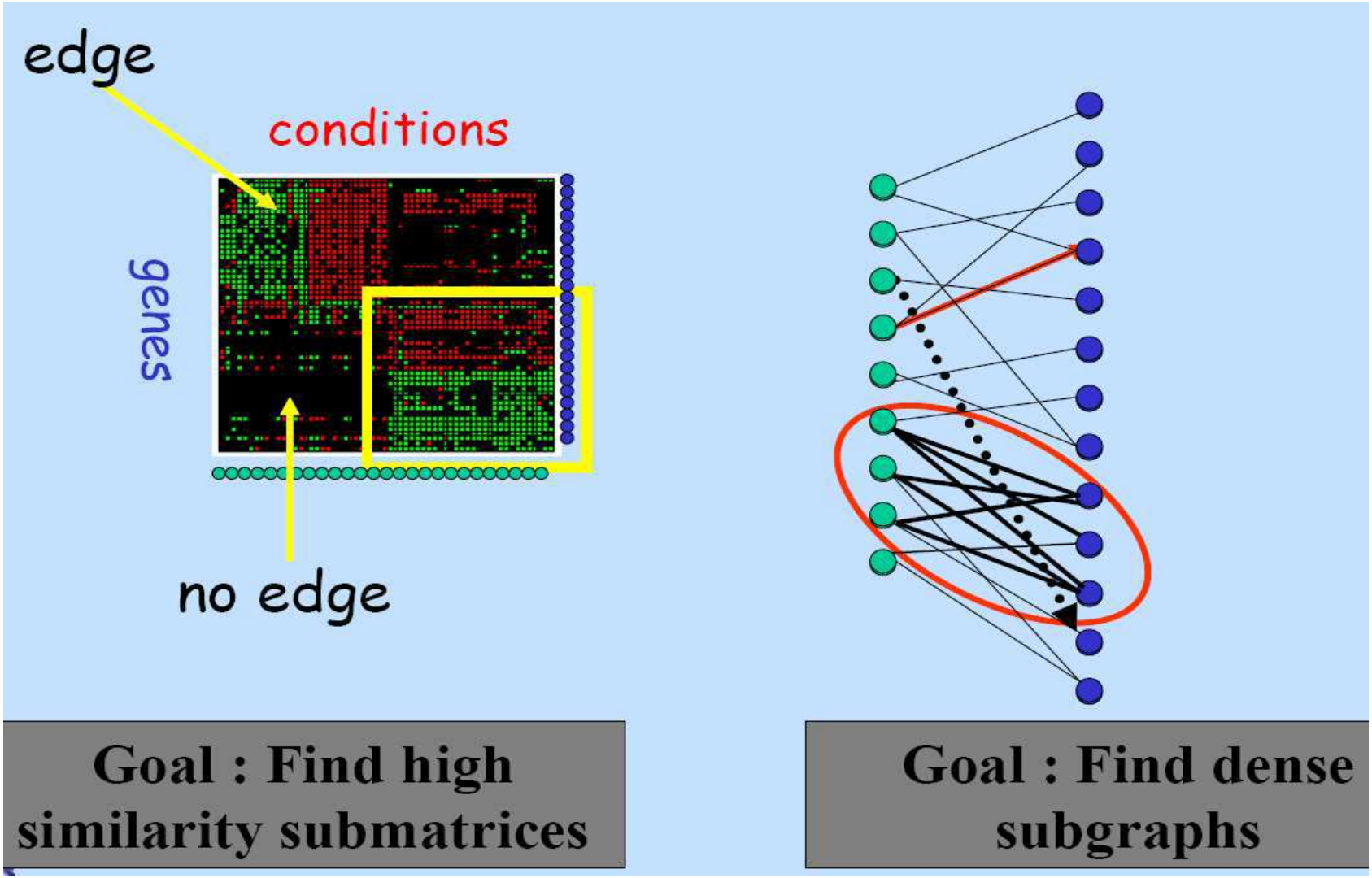
# Bipartite graph

- We develop additive scores that can be decomposed across the edges and non-edges

- **Weight** of bicluster is sum of the weights of **gene-condition** pairs, including edges and non-edges



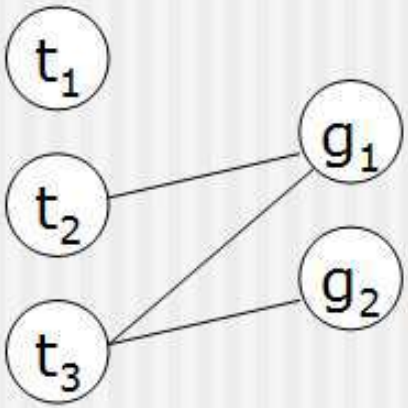
# The SAMBA method



# Simple example

- Example:

| | $t_1$ | $t_2$ | $t_3$ |
|---|---|---|---|
| $g_1$ | 0.8 | 1.5 | 2.6 |
| $g_2$ | 0.4 | 0.7 | 3.2 |



# SAMBA introduction

- The whole dataset forms a bipartite graph G=(U, V, E):
  - U is the set of conditions.
  - V is the set of genes.
  - (u, v) $\in$ E iff v responds in condition u (i.e., the expression level of v changes significantly in u).
- Bipartite subgraph B=(U', V', E') of graph G.

# Statistical data modeling – simple model

- $p = \frac{|E|}{|U||V|}$ - edges occur independently and eqiprobably with density p
- $BT(k', p, n'm')$ - binomial tail, probability of observing k success in n trials, with probability p
- $p(B) = BT(k', p, n'm')$ - probability of observing of observing graph

# Statistical data modeling – simple model

- Goal is to find a subgraph B with lowest p(B)

$$p(B) = \sum_{k' > k} \binom{n'm'}{k'} p^{k'} (1-p)^{n'm'-k'} < 2^{nm} p^{k} (1-p)^{nm-k}$$

- Assuming p<1/2 we obtain the upper bound
- Minimizing $\log p(B)$ is equivalent to finding maximum weighted subgraph of G.
- Edges have positive weight: $(1 + \log p)$
- Non-edges negative weight: $(1 + \log(1-p))$

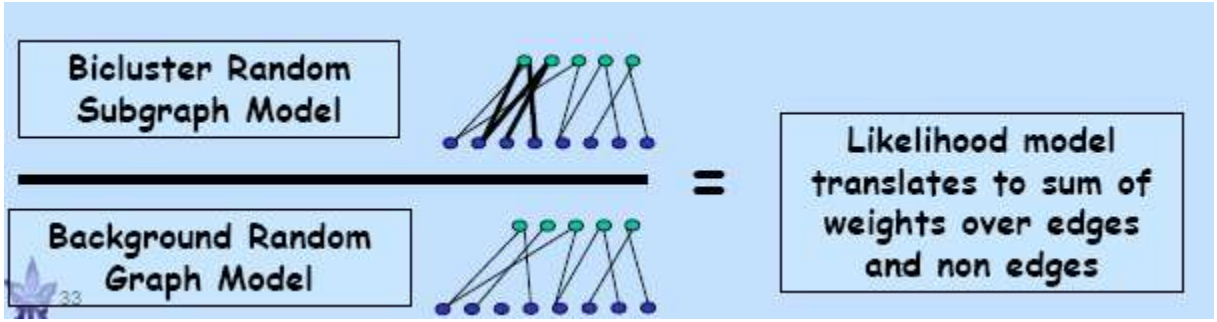
# Problems in simple model

- Not all dense subgraphs are statistically significant.
- Edges incident on nodes of high degree tend to appear in high scoring subgraphs.
- Edges incident on low degree nodes tend to be left out.
- Need a better model accounting for different node characteristics.

# Extended likelihood ratio model

- To overcome this problems we use likelihood ratio to capture the significance of biclusters

- Refined model incorporate behavior of each specific condition and gene



# Extended likelihood ratio model

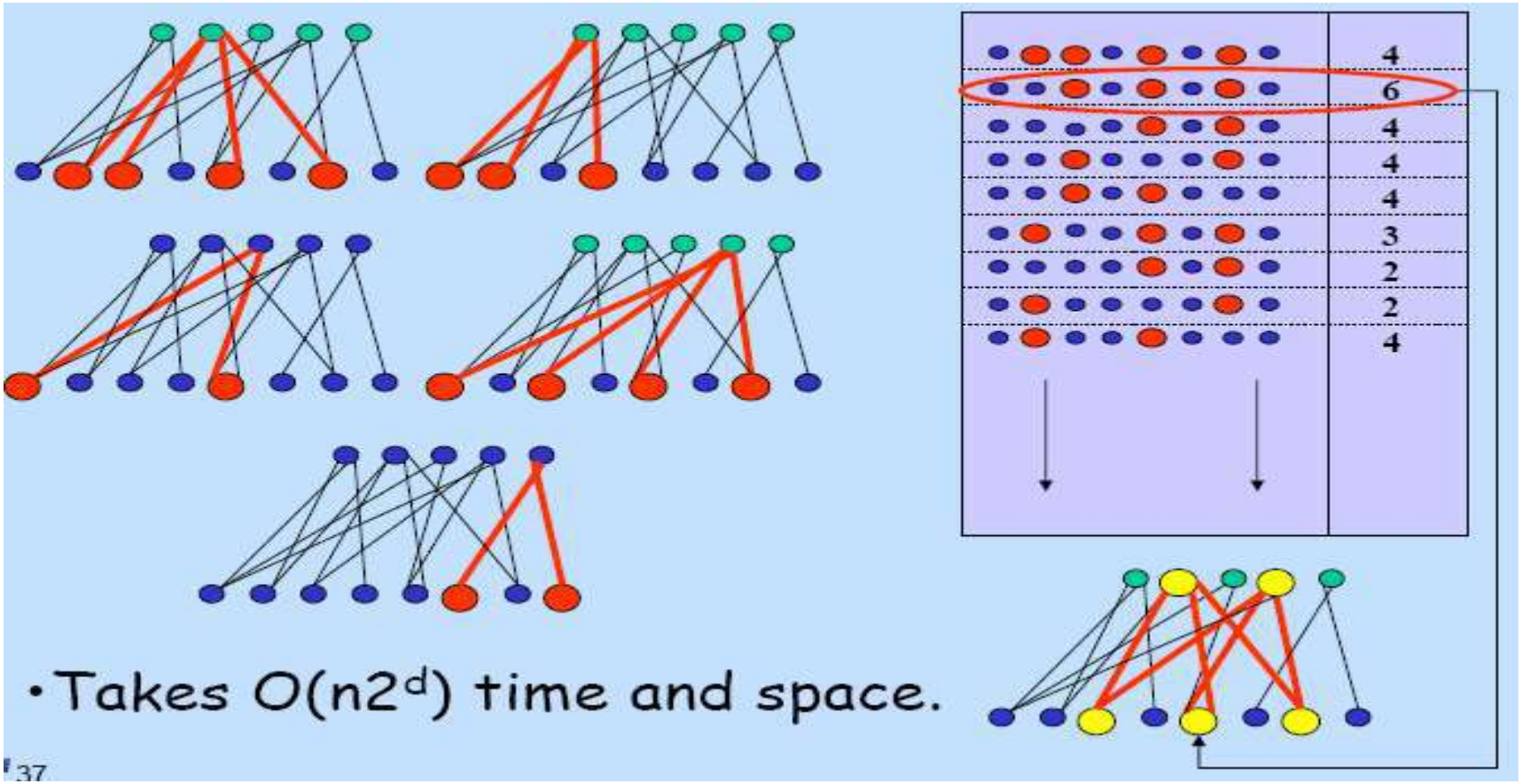
- Edges occurs with constant probability $p_c > max_{(u,v)\in U\times V} p_{u,v}$
- The log-likelihood ratio for B is therefore:

$$\log L(B) = \sum_{(u,v)\in E'} \frac{p_c}{p(u,v)} + \sum_{(u,v)\notin E'} \frac{1-p_c}{1-p(u,v)}$$

- Weights for edges - $log \frac{p_c}{p_{u,v}} > 0$
- Weights for non-edges - $log \frac{1-p_c}{1-p_{u,v}} < 0$

# Finding heaviest bicluster in bipartite graph

- Finding the heaviest bicluster in bipartite graph is NP-hard



# Algorithm maximum bounded bicluster

- Algorithm identify a maximum weighted subgraph of given weighted bipartite graph G.
- d-bounded degree - no more than d edges incident on each gene vertex.
- v - neighborhood of a vertex
- N(v) - be the set of vertices adjacent to v in G.

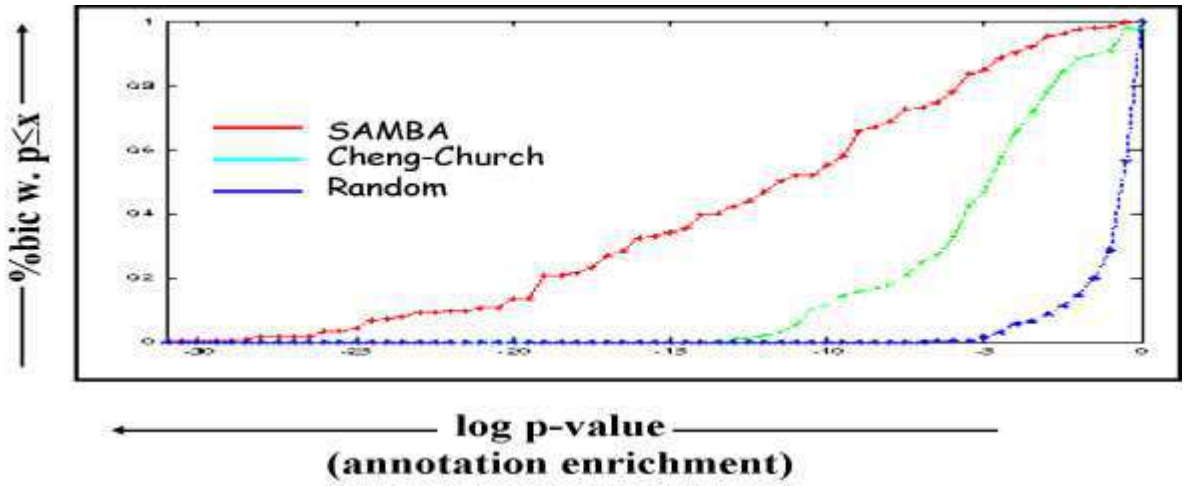
MaxBoundBiClique$(U, V, E, d)$:
Initialize a hash table $weight$; $weight_{best} \leftarrow 0$
**For** all $v \in V$ do
  **For** all $S \subseteq N(v)$ do
    $weight[S] \leftarrow weight[S] + \max\{0, w(S, \{v\})\}$
    **If** $(weight[S] > weight_{best})$
      $U_{best} \leftarrow S$
      $weight_{best} \leftarrow weight[S]$
Compute $V_{best} = \cap_{u \in U_{best}} N(u)$
Output $(U_{best}, V_{best})$

# SAMBA's implementation

- Phase I: Form the bipartite graph and calculate vertex pair weights.
- Phase II: Apply the hashing technique to find the k heaviest bicluster in the graph.
- Phase III: Perform greedy addition/removal of vertices and filter biclusters that are too similar.

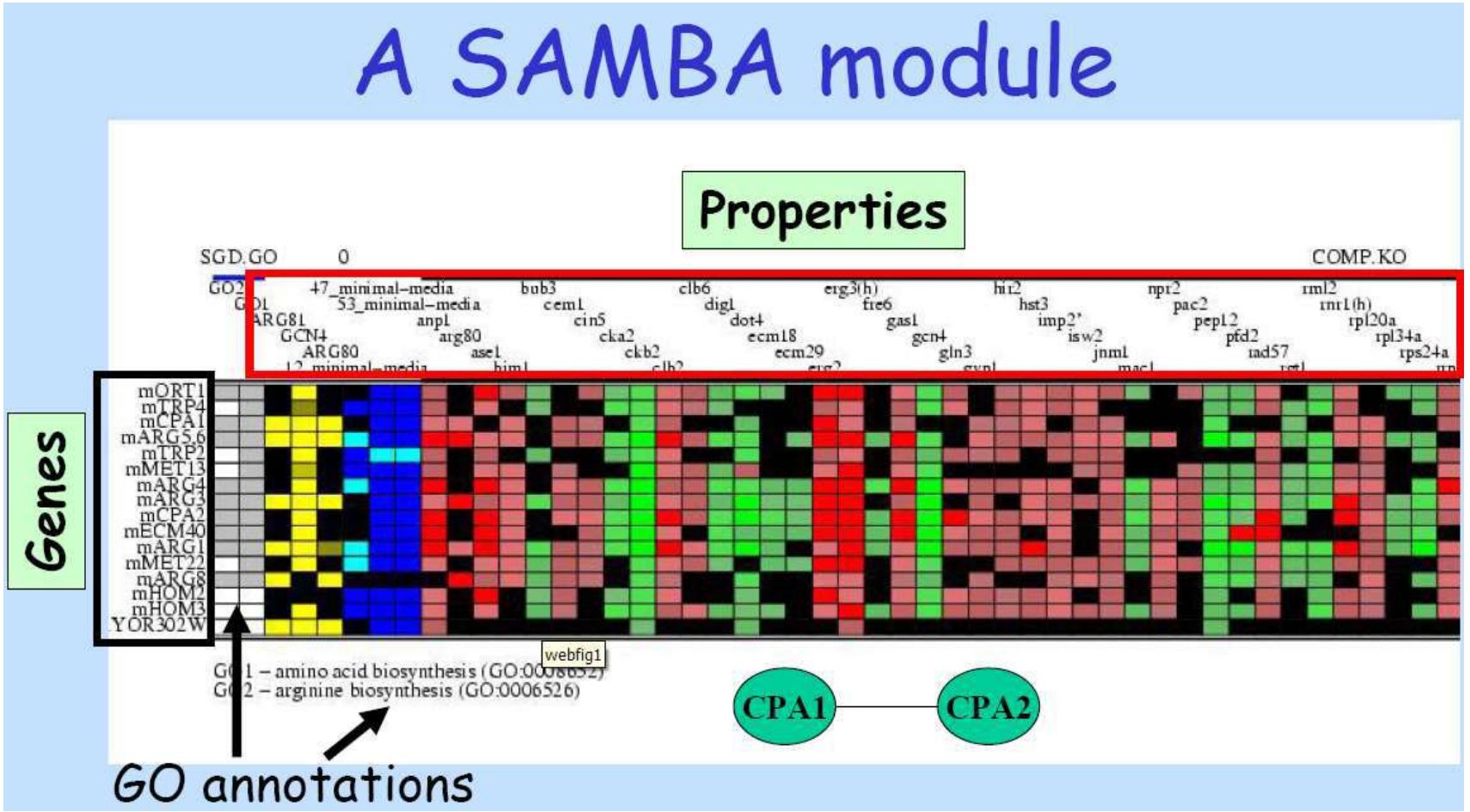
# Biclusters quality

- Applied SAMBA to dataset that contained the expression levels of 4,026 genes over 96 human tissue samples, which are classified into 9 classes of lymphoma and normal ones.

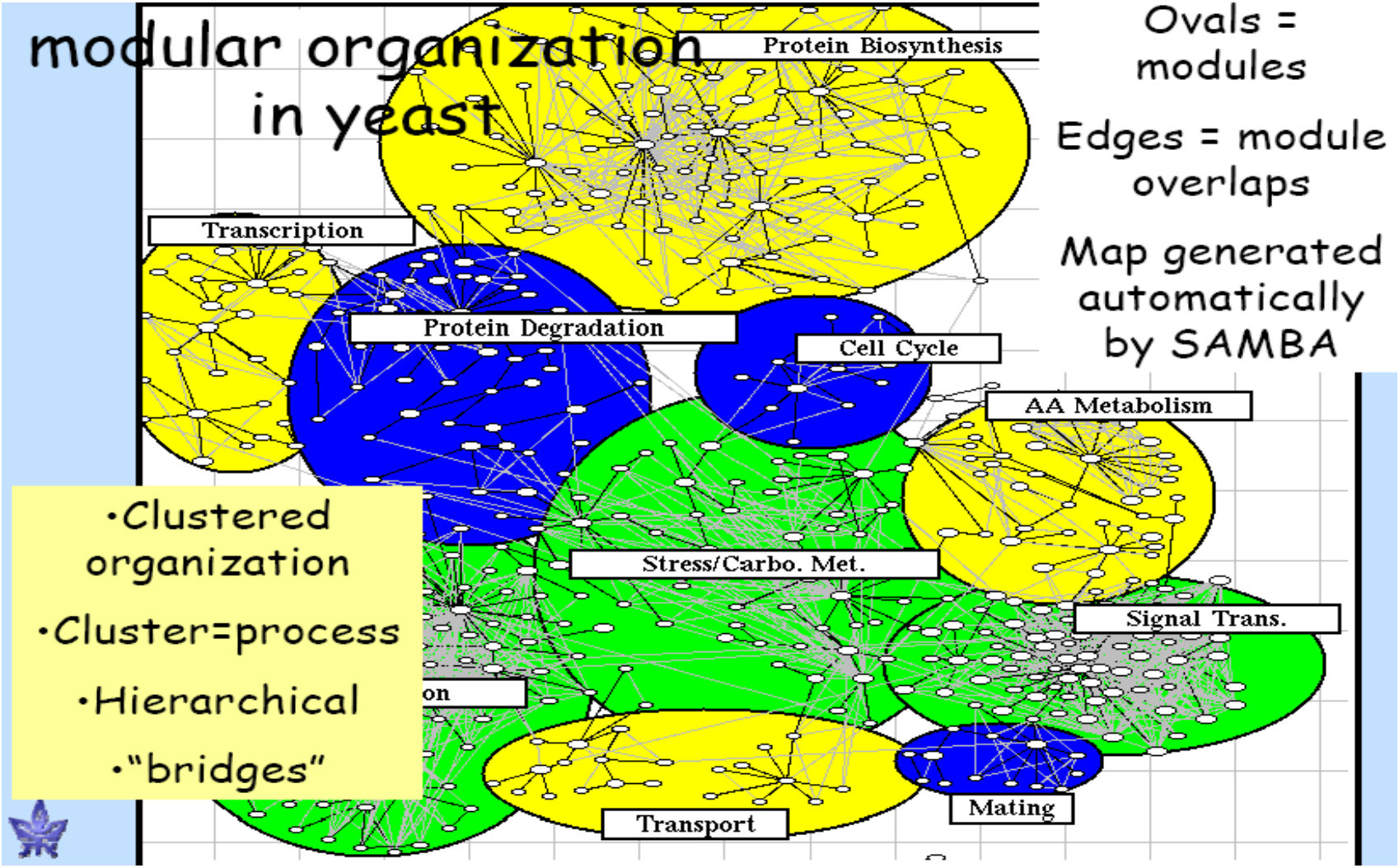



- Correspondence plots for SAMBA, the algorithm of Cheng and Church and random biclusters

# Example I


A SAMBA module
Properties
SGD.GO
0
COMP.KO
47_minimal-media
53_minimal-media
ARG81
GCN4
ARG80
bub3
ceml
cin5
cka2
ckb2
clb6
digl
dot4
ecm18
ecm29
erg3(h)
fre6
gasl
gcn4
gln3
hir2
hst3
imp2'
isw2
jnml
mac1
npr2
pac2
pep12
pfd2
rad57
rml2
rnr1(h)
rpl20a
rpl34a
rps24a
anpl
arg80
ase1
Genes
mORT1
mTRP4
mCPA1
mARG5,6
mTRP2
mMET13
mARG4
mARG3
mCPA2
mECM40
mARG1
mMET22
mARG8
mHOM2
mHOM3
YOR302W
webfig1
G 1 – amino acid biosynthesis (GO:0008652)
G 2 – arginine biosynthesis (GO:0006526)
CPA1
CPA2
GO annotations

# Example II


modular organization in yeast
Ovals = modules
Edges = module overlaps
Map generated automatically by SAMBA
Protein Biosynthesis
Transcription
Protein Degradation
Cell Cycle
AA Metabolism
Stress/Carbo. Met.
Signal Trans.
Transport
Mating
•Clustered organization
•Cluster=process
•Hierarchical
•"bridges"

# Reference


- *Amos Tanay, Roded Sharan Martin Kupiec and Ron Shamir (2003),* ***Reavealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genome wide data.***
- *Amos Tanay, Roded Sharan and Ron Shamir (2002),* ***Discovering statistically significant biclusters in gene expression data.***
- *Roded Sharan (2009), lectures,* ***Analysis of biological networks: Network modules identification.***
- *Roded Sharan, Igor Ulitsky and Ron Shamir (2007),* ***Network based prediction of protein function.***

# Thank you for attention