Model-based clustering and validation techniques for gene expression data

Ka Yee Yeung

Department of Microbiology

University of Washington, Seattle WA

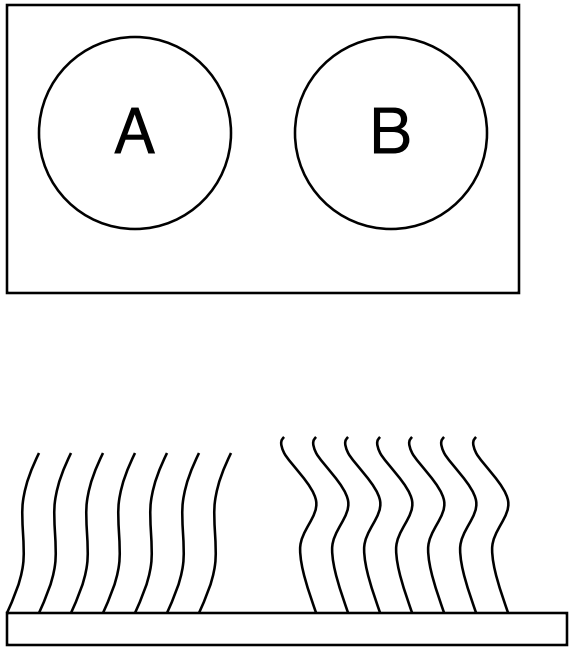
Overview

- Introduction to microarrays
- Clustering 101
- Validating clustering results on microarray data
- Model-based clustering using microarray data
- Co-expression == co-regulation ??

Introduction to Microarrays

DNA Arrays Measure the Concentration of 1000's of Genes Simultaneously

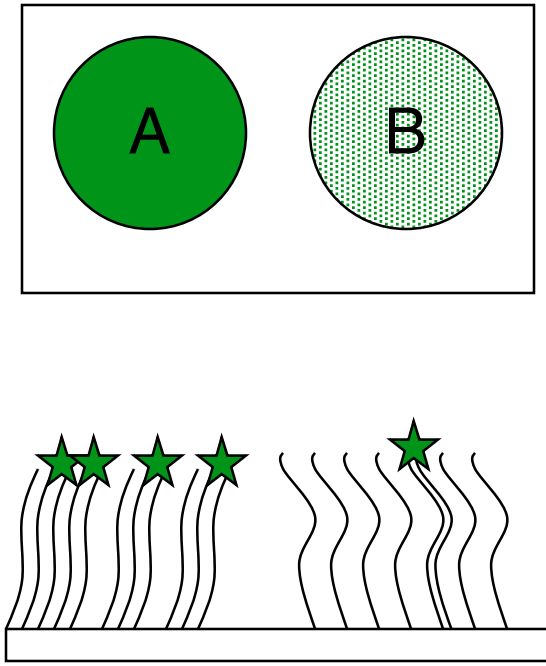
On the surface



In solution

4 copies of gene A,
1 copy of gene B

After Hybridization



Two-color analysis allows for comparative studies to be done

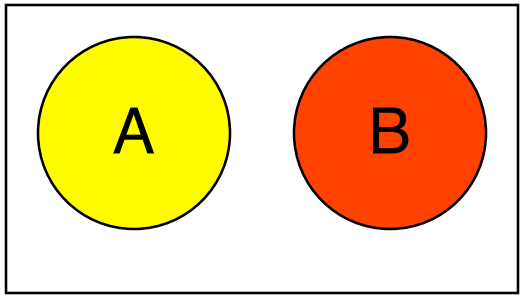
In solution #1

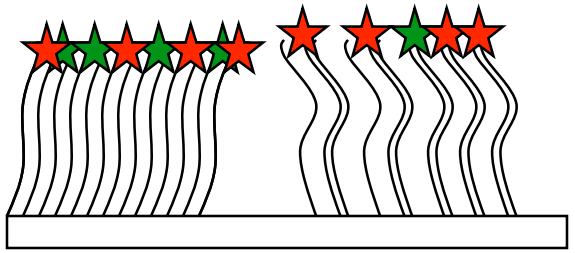
4 copies of gene A,
1 copy of gene B

In solution #2

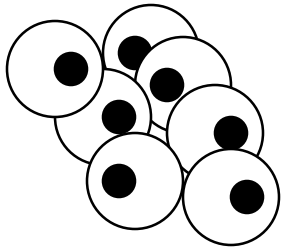
4 copies of gene A,
4 copies of gene B

After Hybridization





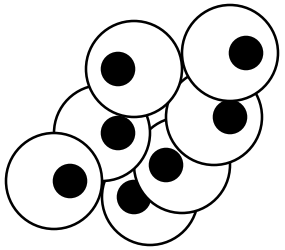
Cell Population #1



Extract mRNA

Make cDNA
Label w/ Green Fluor

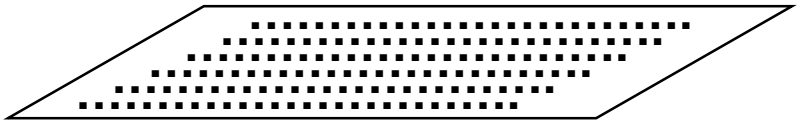
Cell Population #2



Extract mRNA

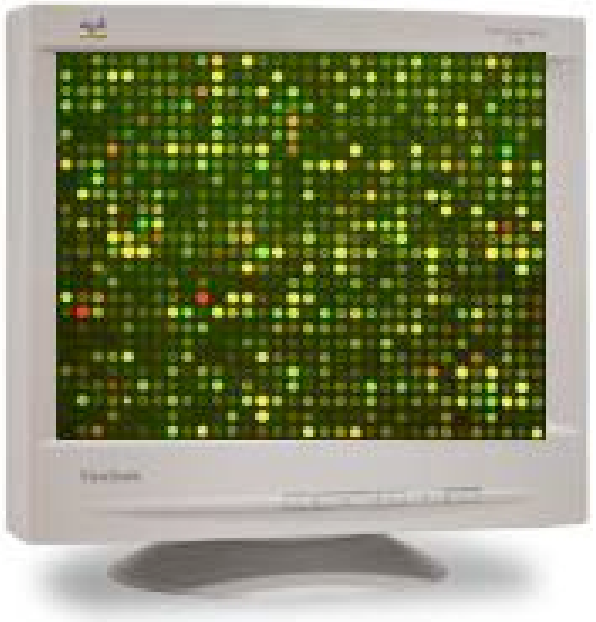
Make cDNA
Label w/ Red Fluor

Co-hybridize



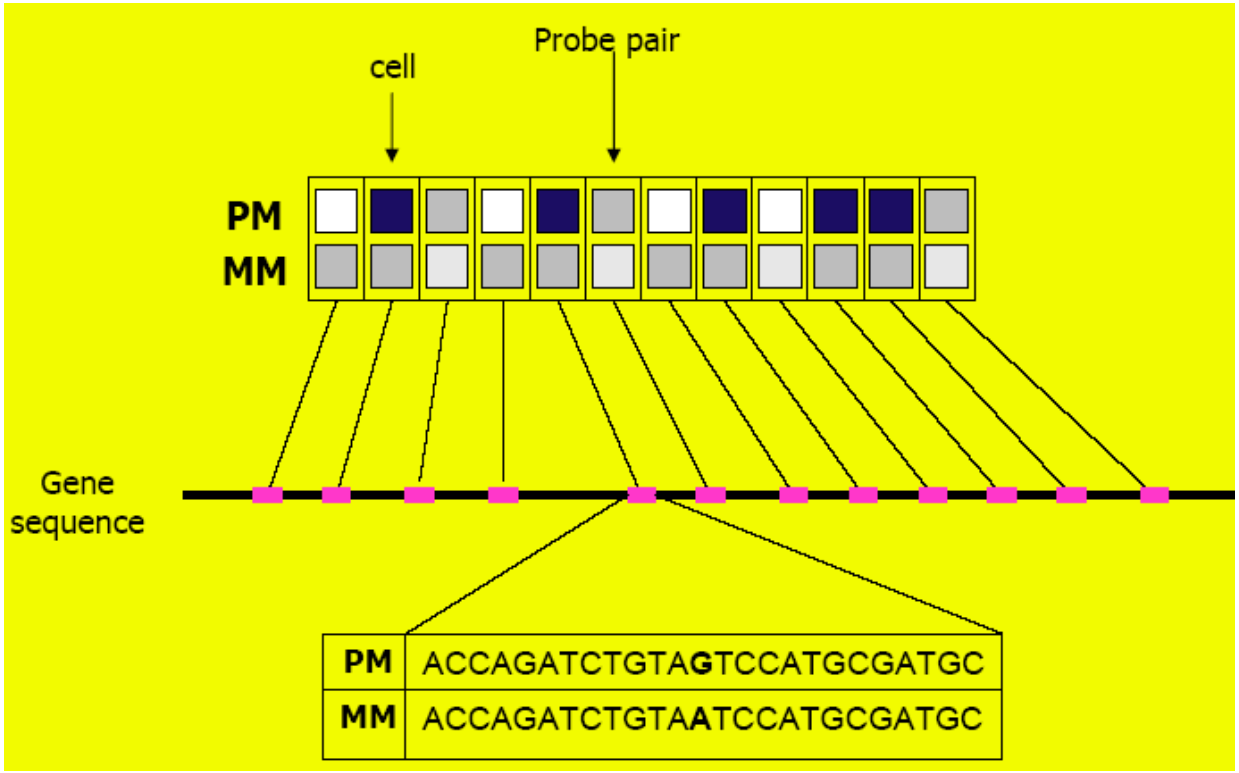
Slide with DNA from
different genes

Scan



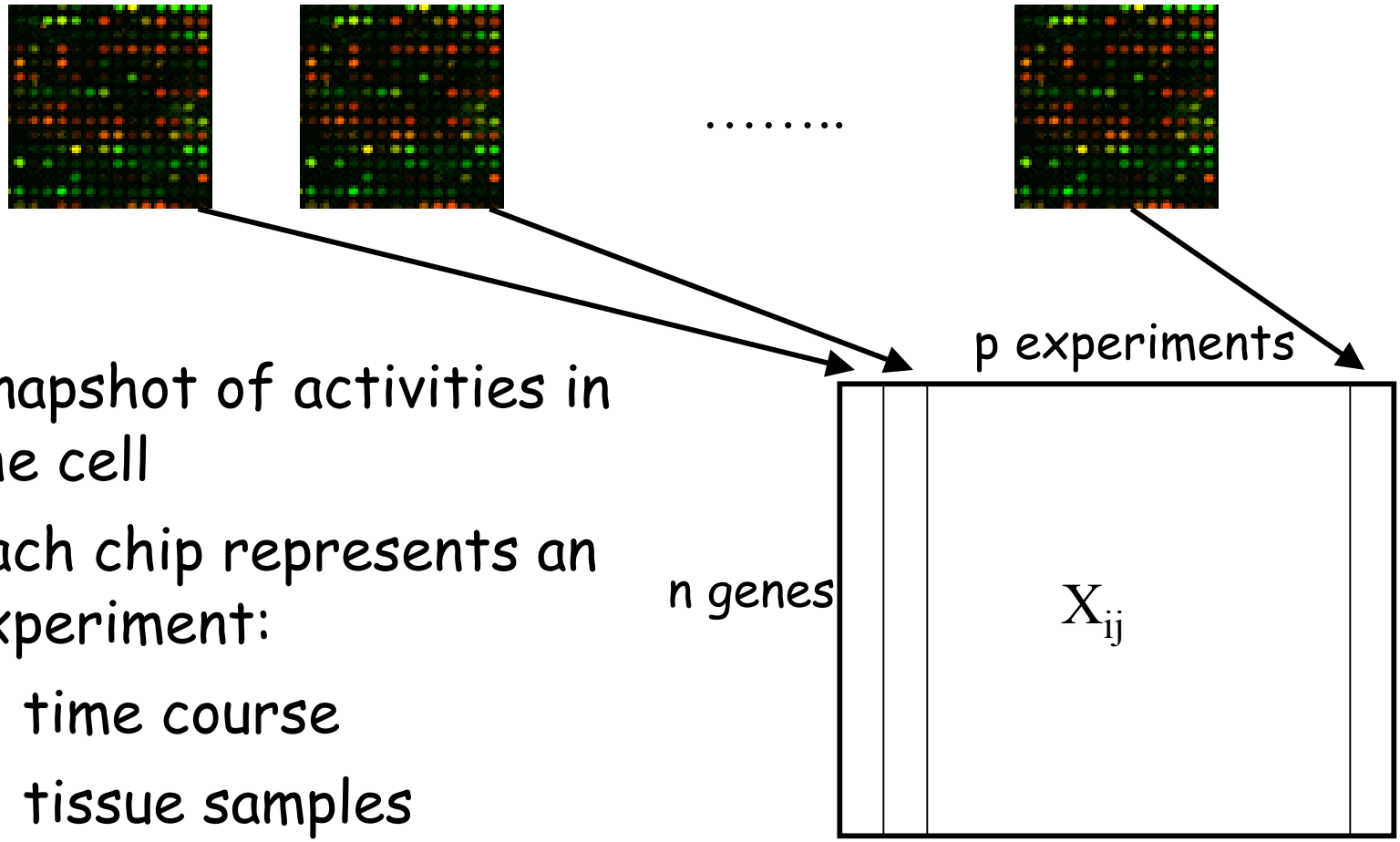
The "Gene Chip" from Affymetrix

Affymetrix chips: oligonucleotide arrays



PM (Perfect Match) vs. MM (mis-match)

A gene expression data set



- Snapshot of activities in the cell
- Each chip represents an experiment:
  - time course
  - tissue samples (normal/cancer)

What is clustering?

- Group *similar* objects together

Applications of clustering gene expression data

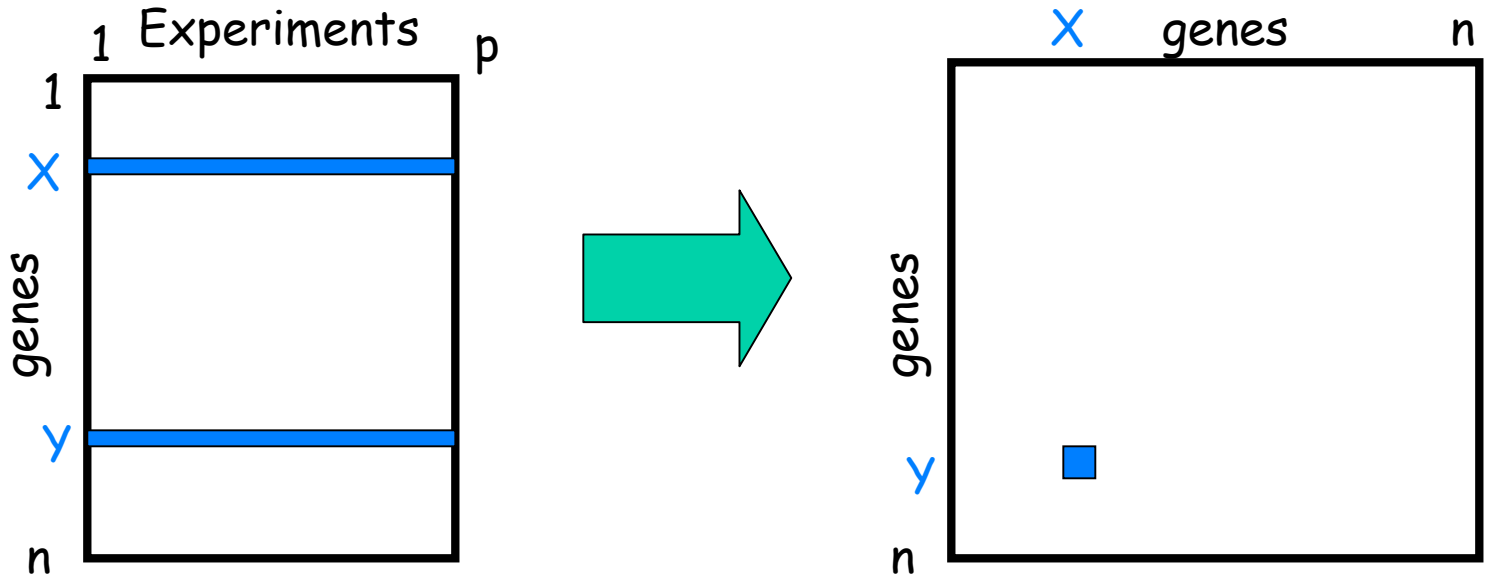
- "Guilt by association"
  - E.g. Cluster the genes ➔ functionally related genes
- Class discovery
  - E.g. Cluster the experiments ➔ discover new subtypes of tissue samples

Clustering 101

What is clustering?

- Group *similar* objects together
- Objects in the same cluster (group) are more similar to each other than objects in different clusters
- Data exploratory tool
- Unsupervised method
  - Do not assume any knowledge of the genes or experiments

How to define similarity?



- Similarity measure:
  - A measure of *pairwise* similarity or dissimilarity
  - Examples:
    - Correlation coefficient
    - Euclidean distance

Similarity measures

- Euclidean distance

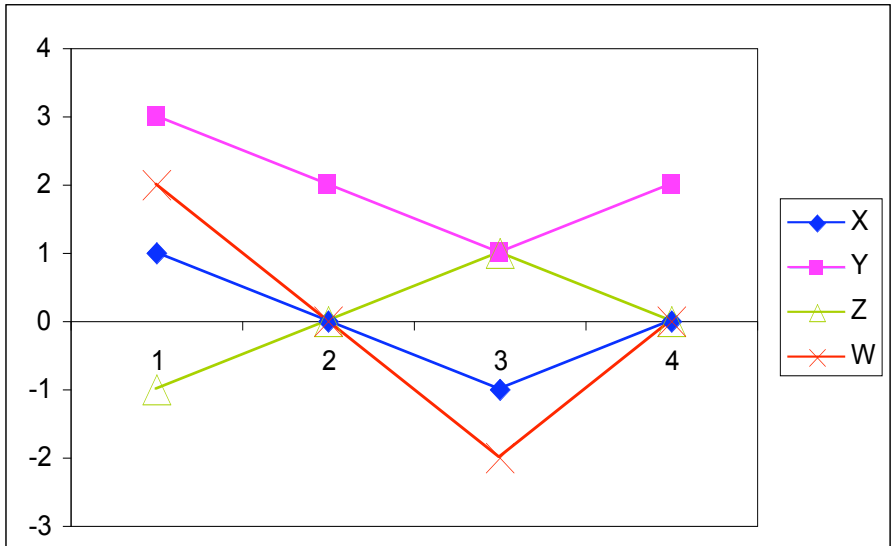
$$\sqrt{\sum_{j=1}^{p} (X[j] - Y[j])^2}$$

- Correlation coefficient

$$\frac{\sum_{j=1}^{p} (X[j] - \overline{X})(Y[j] - \overline{Y})}{\sqrt{\sum_{j=1}^{p} (X[j] - \overline{X})^2 \sum_{j=1}^{p} (Y[j] - \overline{Y})^2}}, \quad where\ \overline{X} = \frac{\sum_{j=1}^{p} X[j]}{p}$$

Example

X	1	0	-1	0
Y	3	2	1	2
Z	-1	0	1	0
W	2	0	-2	0



Correlation (X,Y) = 1 Distance (X,Y) = 4

Correlation (X,Z) = -1 Distance (X,Z) = 2.83

Correlation (X,W) = 1 Distance (X,W) = 1.41

Lessons from the example

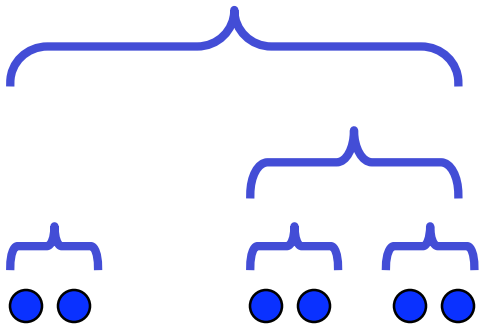
- Correlation – direction only
- Euclidean distance – magnitude & direction

Clustering algorithms

- Inputs:
  - Raw data matrix or similarity matrix
  - Number of clusters or some other parameters
- Hierarchical vs Partitional algorithms

Hierarchical Clustering

[Hartigan 1975]

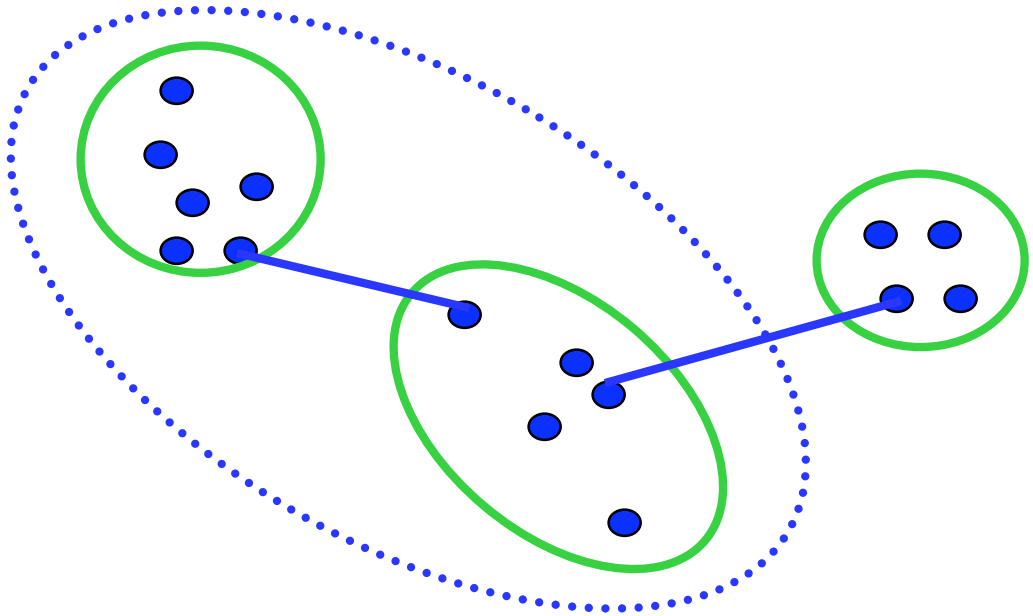


dendrogram

- Agglomerative (bottom-up)
- Algorithm:
  - Initialize: each item a cluster
  - Iterate:
    - select two most similar clusters
    - merge them
  - Halt: when required number of clusters is reached

Hierarchical: Single Link

- cluster similarity = similarity of two most similar members

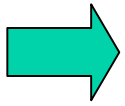


- Potentially long and skinny clusters

+ Fast

Example: single link

	1	2	3	4	5
1	0				
2	2	0			
3	6	3	0		
4	10	9	7	0	
5	9	8	5	4	0

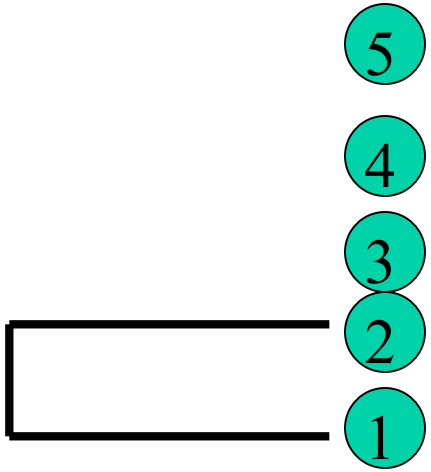


	(1,2)	3	4	5
(1,2)	0			
3	3	0		
4	9	7	0	
5	8	5	4	0

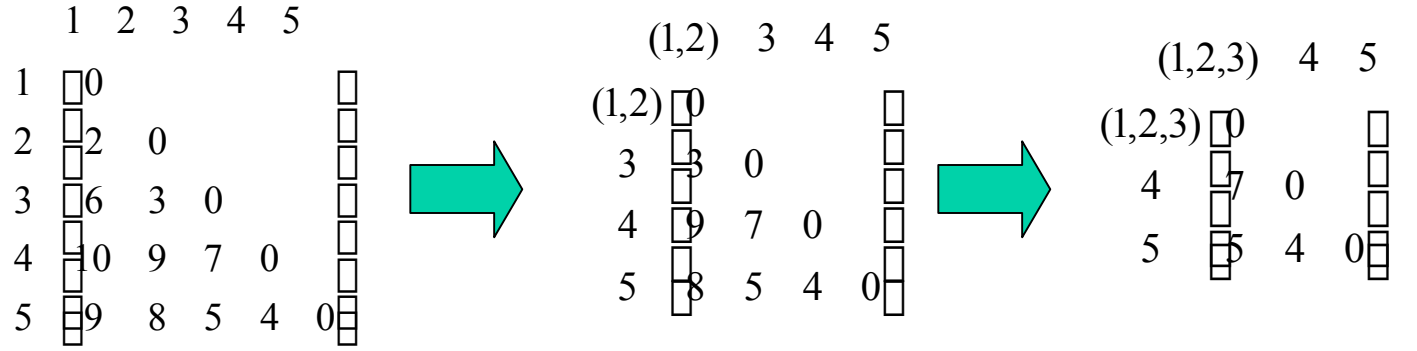
$$d_{(1,2),3} = \min\{ d_{1,3}, d_{2,3}\} = \min\{ 6,3\} = 3$$

$$d_{(1,2),4} = \min\{ d_{1,4}, d_{2,4}\} = \min\{ 10,9\} = 9$$

$$d_{(1,2),5} = \min\{ d_{1,5}, d_{2,5}\} = \min\{ 9,8\} = 8$$

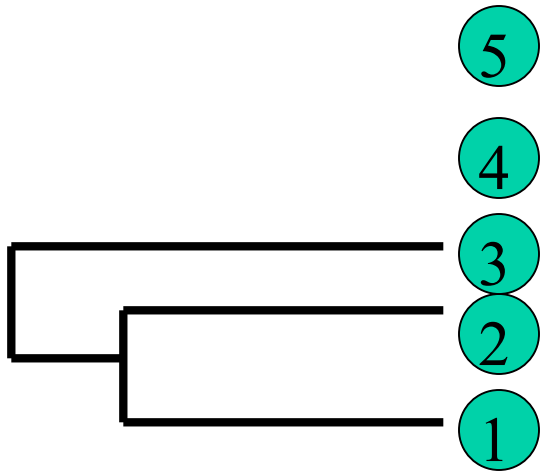


Example: single link

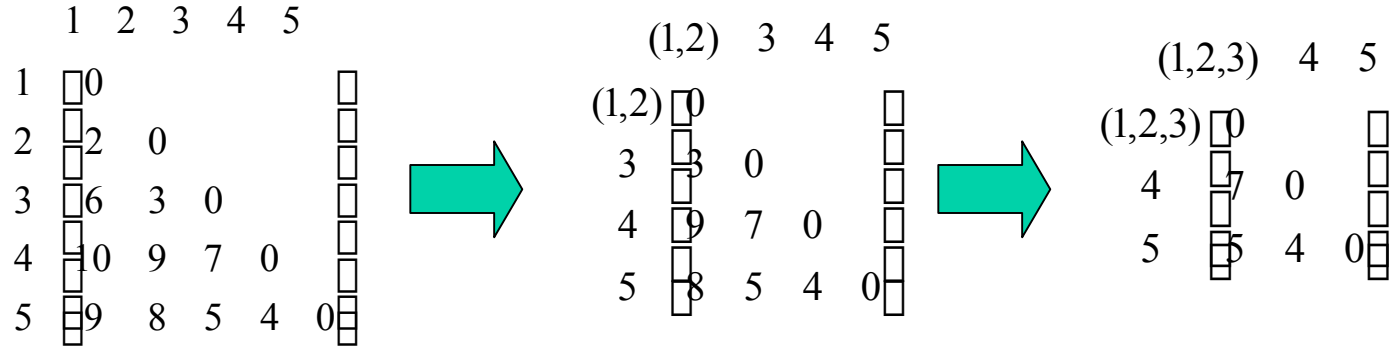


$$d_{(1,2,3),4} = \min\{ d_{(1,2),4}, d_{3,4}\} = \min\{ 9,7\} = 7$$

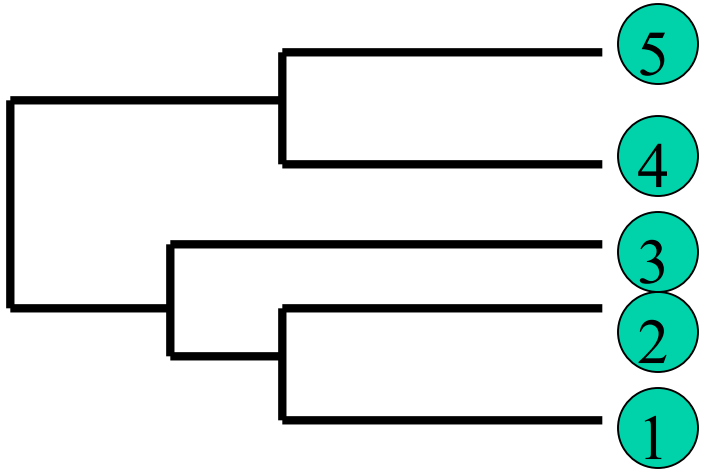
$$d_{(1,2,3),5} = \min\{ d_{(1,2),5}, d_{3,5}\} = \min\{ 8,5\} = 5$$



Example: single link

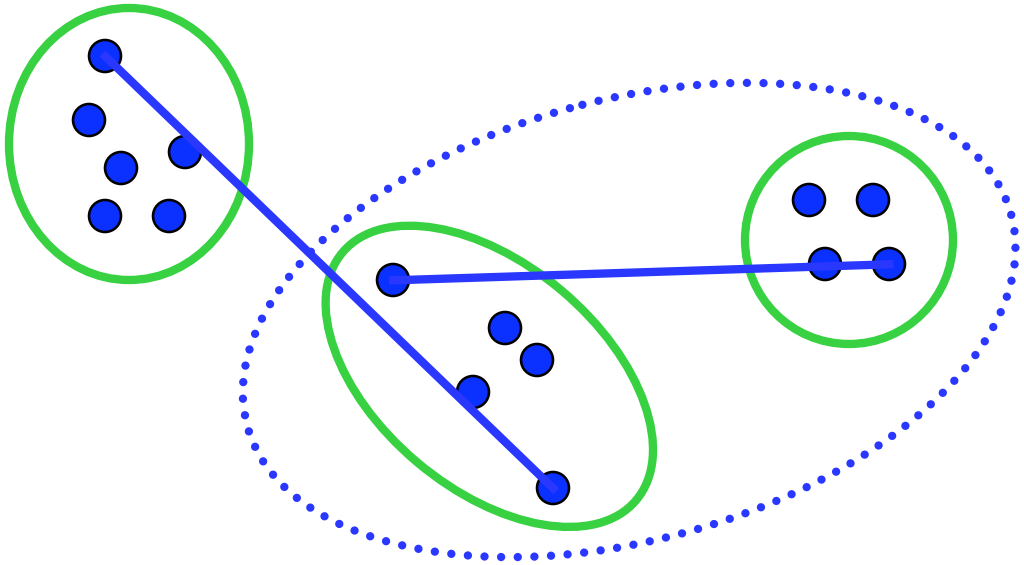


$$d_{(1,2,3),(4,5)} = \min\{ d_{(1,2,3),4}, d_{(1,2,3),5} \} = 5$$



Hierarchical: Complete Link

- cluster similarity = similarity of two least similar members



+ tight clusters

- slow

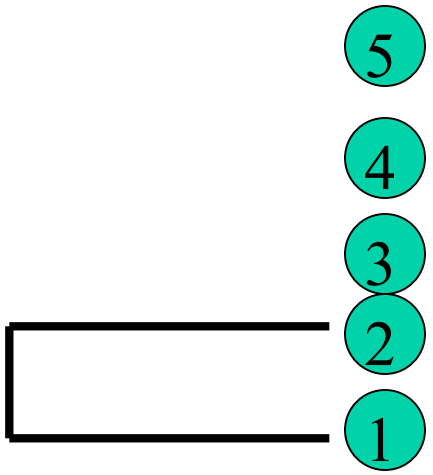
Example: complete link

$$\begin{array}{c|ccccc} & 1 & 2 & 3 & 4 & 5 \\ 1 & 0 & & & & \\ 2 & 2 & 0 & & & \\ 3 & 6 & 3 & 0 & & \\ 4 & 10 & 9 & 7 & 0 & \\ 5 & 9 & 8 & 5 & 4 & 0 \end{array} \Rightarrow \begin{array}{c|cccc} & (1,2) & 3 & 4 & 5 \\ (1,2) & 0 & & & \\ 3 & 6 & 0 & & \\ 4 & 10 & 7 & 0 & \\ 5 & 9 & 5 & 4 & 0 \end{array}$$

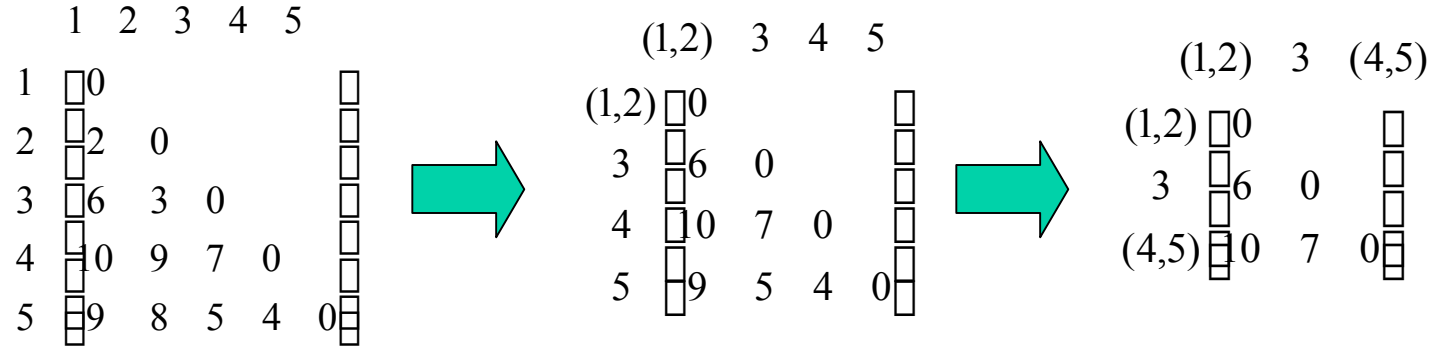
$$d_{(1,2),3} = \max\{d_{1,3}, d_{2,3}\} = \max\{6,3\} = 6$$

$$d_{(1,2),4} = \max\{d_{1,4}, d_{2,4}\} = \max\{10,9\} = 10$$

$$d_{(1,2),5} = \max\{d_{1,5}, d_{2,5}\} = \max\{9,8\} = 9$$

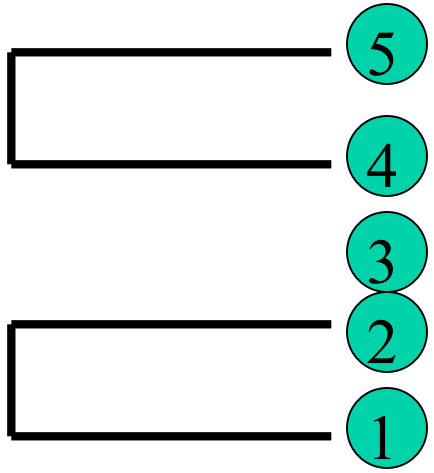


Example: complete link

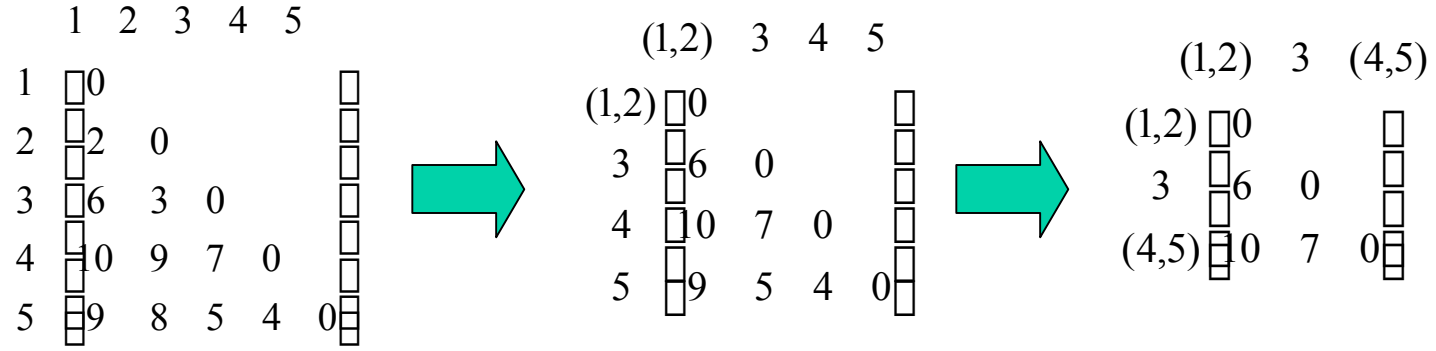


$$d_{(1,2),(4,5)} = \max\{d_{(1,2),4}, d_{(1,2),5}\} = \max\{10,9\} = 10$$

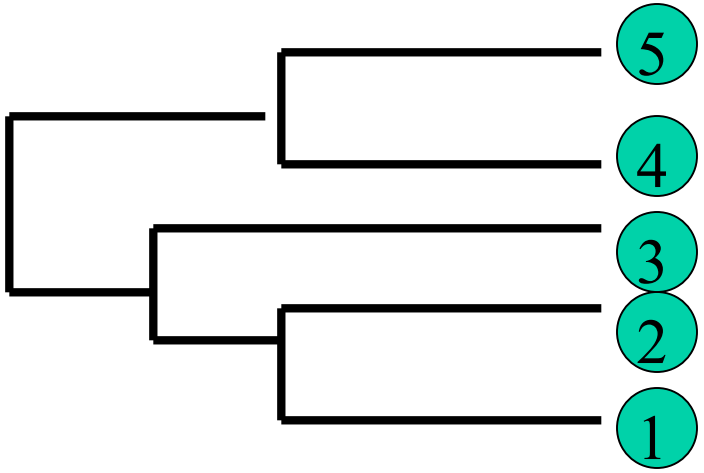
$$d_{3,(4,5)} = \max\{d_{3,4}, d_{3,5}\} = \max\{7,5\} = 7$$



Example: complete link

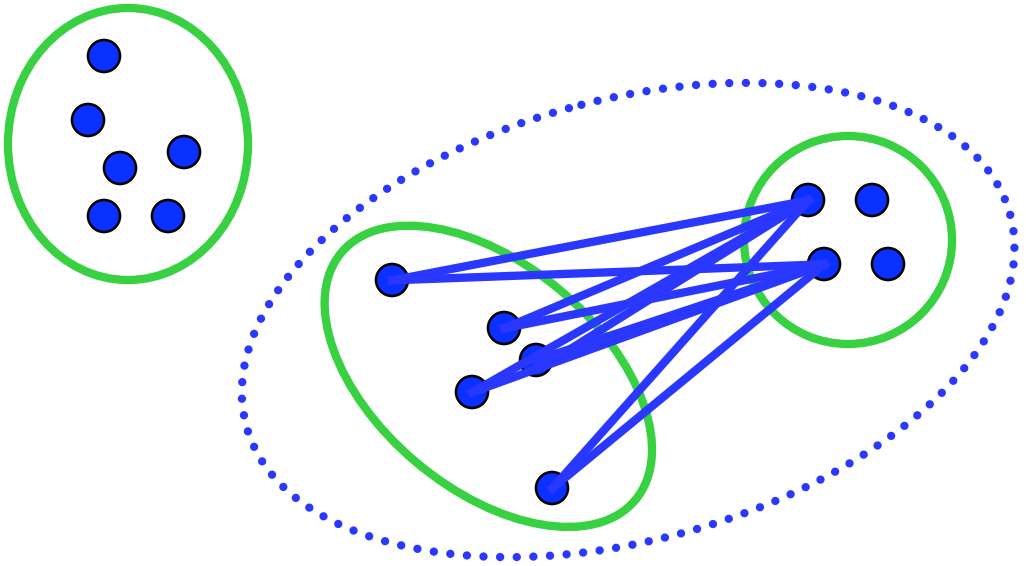


$$d_{(1,2,3),(4,5)} = \max\{d_{(1,2),(4,5)}, d_{3,(4,5)}\} = 10$$



Hierarchical: Average Link

- cluster similarity = average similarity of all pairs

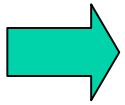


+ tight clusters

- slow

Example: average link

	1	2	3	4	5
1	0				
2	2	0			
3	6	3	0		
4	10	9	7	0	
5	9	8	5	4	0

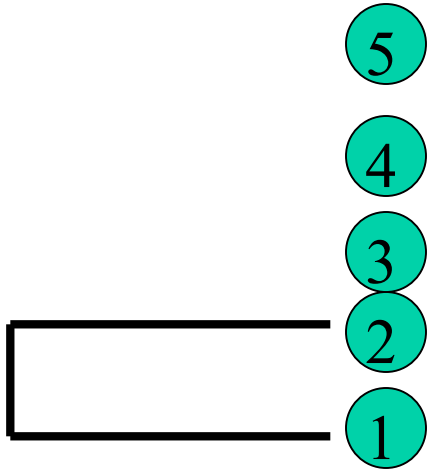


	(1,2)	3	4	5
(1,2)	0			
3	4.5	0		
4	9.5	7	0	
5	8.5	5	4	0

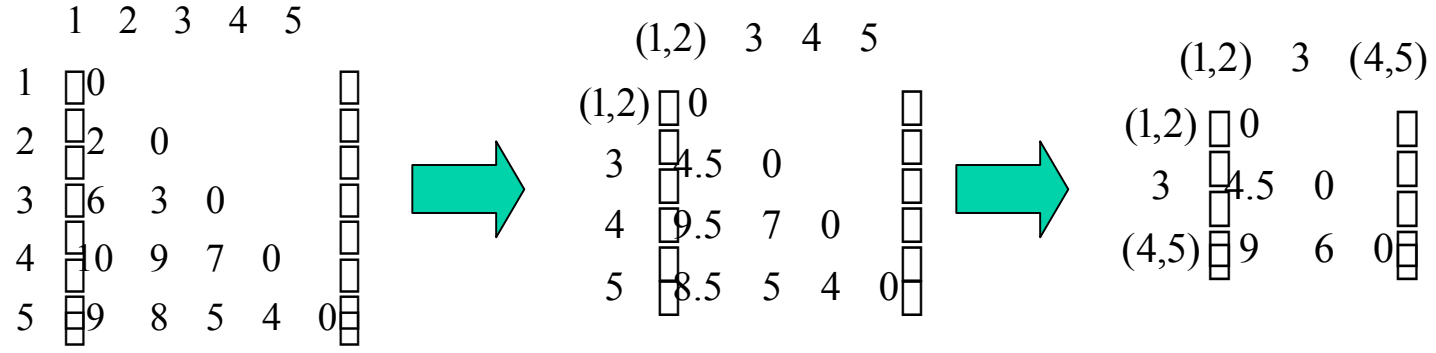
$$d_{(1,2),3} = \frac{1}{2}(d_{1,3} + d_{2,3}) = \frac{6+3}{2} = 4.5$$

$$d_{(1,2),4} = \frac{1}{2}(d_{1,4} + d_{2,4}) = \frac{10+9}{2} = 9.5$$

$$d_{(1,2),5} = \frac{1}{2}(d_{1,5} + d_{2,5}) = \frac{9+8}{2} = 8.5$$

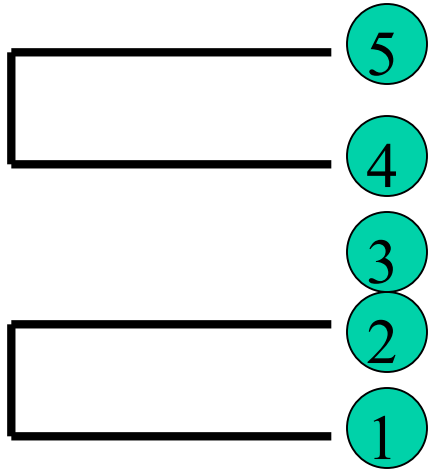


Example: average link



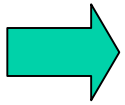
$$d_{(1,2),(4,5)} = \frac{1}{4}(d_{1,4} + d_{1,5} + d_{2,4} + d_{2,5}) = 9$$

$$d_{3,(4,5)} = \frac{1}{2}(d_{3,4} + d_{3,5}) = 6$$

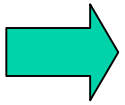


Example: average link

	1	2	3	4	5
1	0				
2	2	0			
3	6	3	0		
4	10	9	7	0	
5	9	8	5	4	0

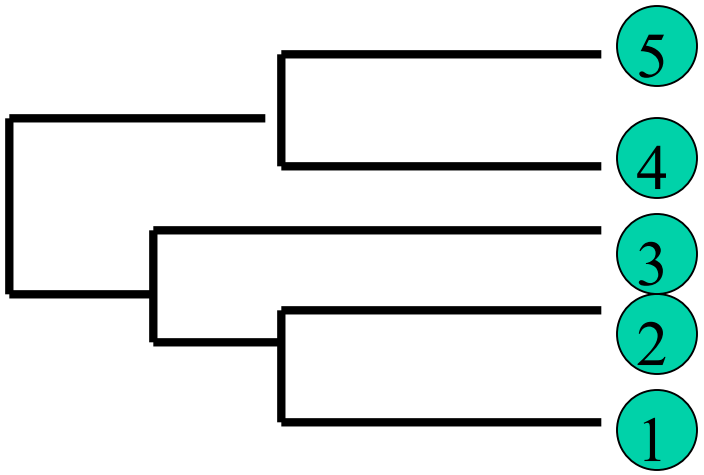


	(1,2)	3	4	5
(1,2)	0			
3	4.5	0		
4	9.5	7	0	
5	8.5	5	4	0



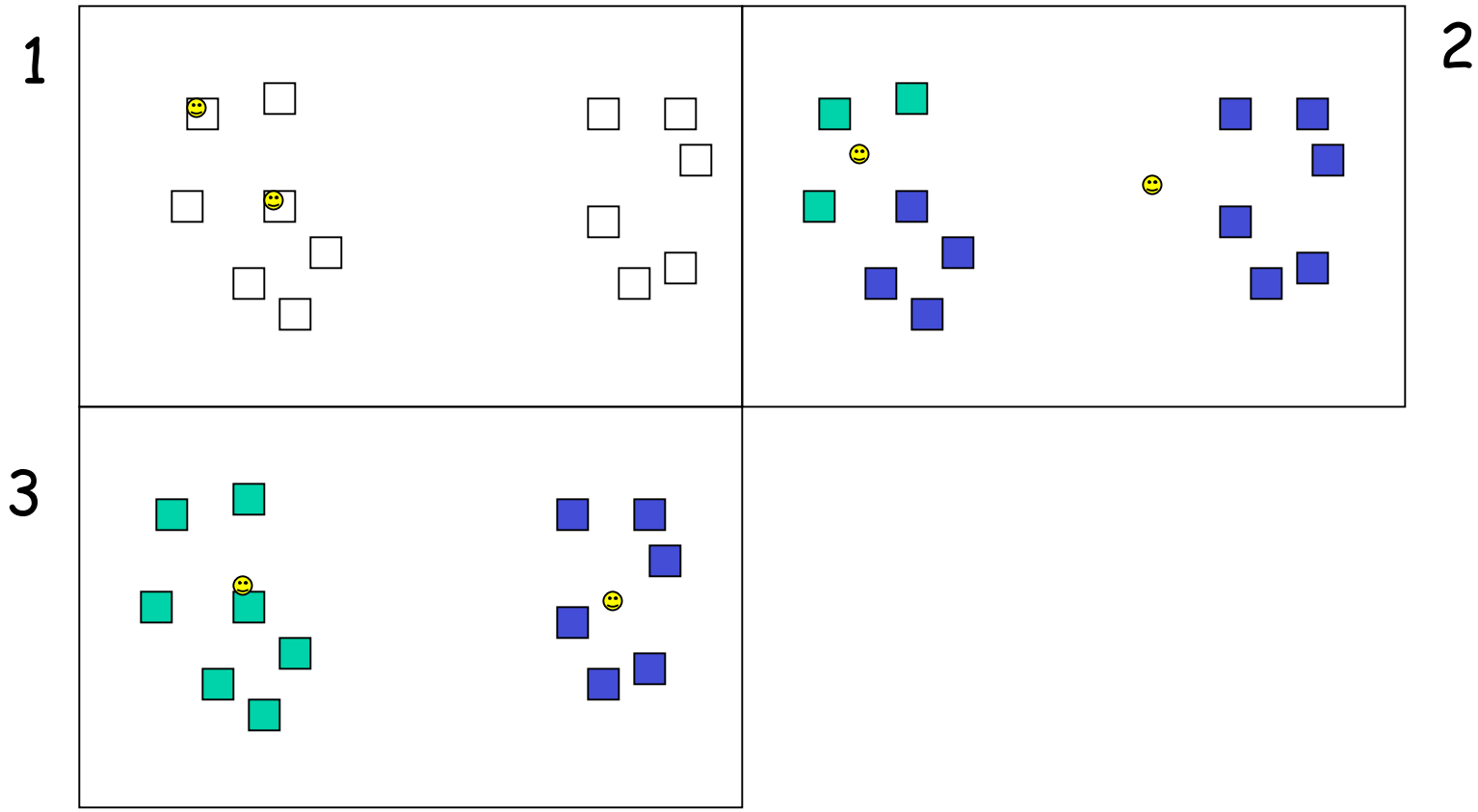
	(1,2)	3	(4,5)
(1,2)	0		
3	4.5	0	
(4,5)	9	6	0

$$d_{(1,2,3),(4,5)} = \frac{1}{6}(d_{1,4} + d_{1,5} + d_{2,4} + d_{2,5} + d_{3,4} + d_{3,5}) = 8$$



Partitional: K-Means

[MacQueen 1965]

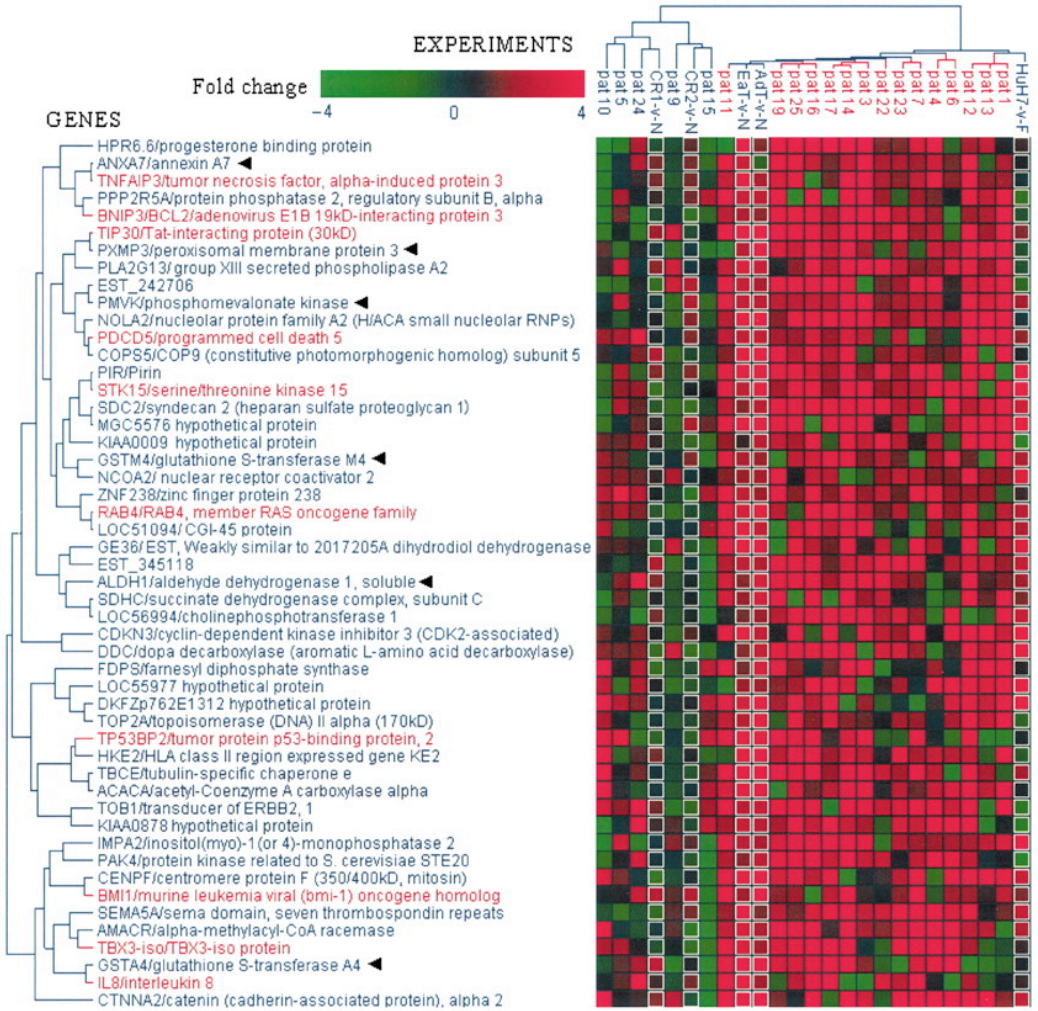


Details of k-means

- Iterate until converge:
  - Assign each data point to the closest centroid
  - Compute new centroid
- Properties:
  - Converge to *local* optimum
  - In practice, converge quickly
  - Tend to produce spherical, equal-sized clusters

2D-clustering

- Cluster both genes and experiments



Summary

- Definition of clustering
- Pairwise similarity:
  - Correlation
  - Euclidean distance
- Clustering algorithms:
  - Hierarchical (single-link, complete-link, average-link)
  - K-means

Clustering microarray data

What has been done?

- Many clustering algorithms have been proposed for gene expression data. For example:
  - Hierarchical clustering algorithms eg. [Eisen *et al* 1998]
  - K-means eg. [Tavazoie *et al.* 1999]
  - Self-organizing maps (SOM) eg. [Tamayo *et al.* 1999]
  - Cluster Affinity Search Technique (CAST) [Ben-Dor, Yakhini 1999]
  - and many others…

Common questions

1. How can I choose between all these clustering methods?
2. Is there a clustering algorithm that works better than the others?
3. How to choose the number of clusters?
4. How often do I get biologically meaningful clusters?
5. How many microarray experiments do I need?

Validating clustering results
[Yeung, Haynor, Ruzzo 2001]

- FOM idea
- Data sets
- Results

ISI most cited paper in Computer Science (Dec 2002)

Validation techniques

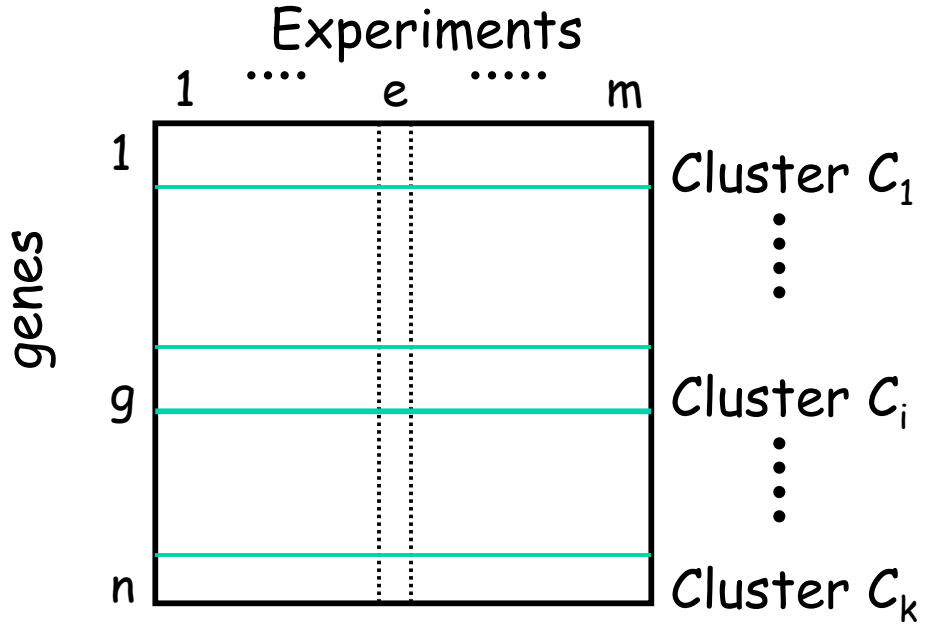
[Jain and Dubes 1988]

- External validation
  - Require external knowledge
- Internal validation
  - Does not require external knowledge
  - Goodness of fit between the data and the clusters

Comparing different heuristic-based algorithms

- Apply a clustering algorithm to all but one experiment
- Use the left-out experiment to check the predictive power of the algorithm

Figure of Merit (FOM)

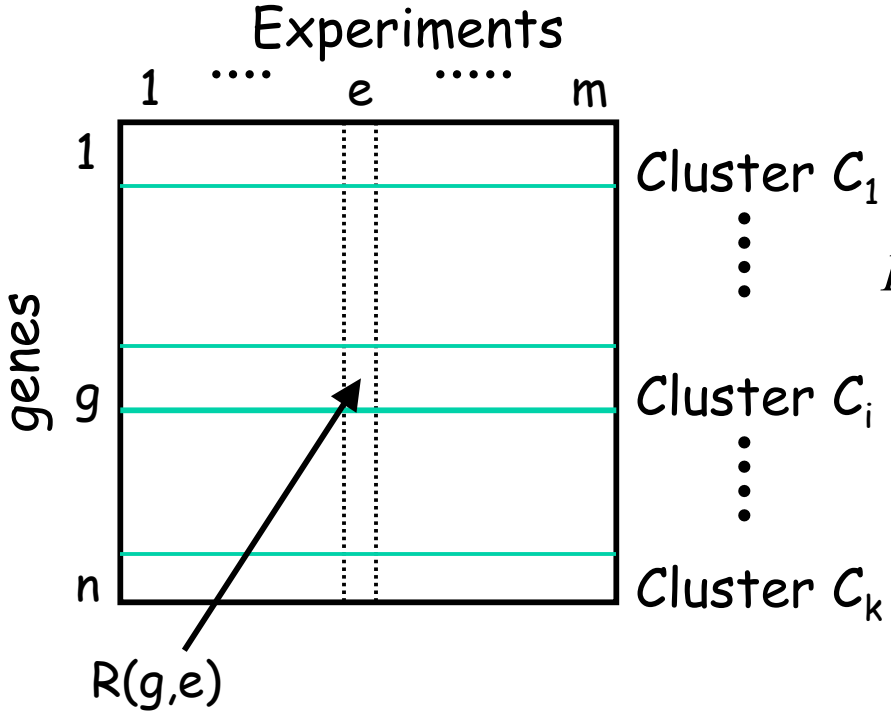


- FOM estimates predictive power:

  measures uniformity of gene expression levels in the left-out condition in clusters formed

- Low FOM => High predictive power

FOM



$$FOM(e,k) = \frac{1}{n} \sum_{i=1}^{k} \sum_{g \in C_i} (R(g,e) - \mu_{C_i}(e))^2$$

$$FOM(k) = \sum_{e=1}^{m} FOM(e,k)$$

$$\text{adjusted FOM(e,k)} = \text{FOM(e,k)} / \frac{\text{n - k}}{\text{n}}$$

Clustering Algorithms

- **Partitional**
  - CAST (Ben-Dor et al. 1999)
  - k-means (Hartigan 1975)
- **Hierarchical**
  - single-link
  - average-link
  - Complete-link
- **Random** (as a control)
  - Randomly assign genes to clusters

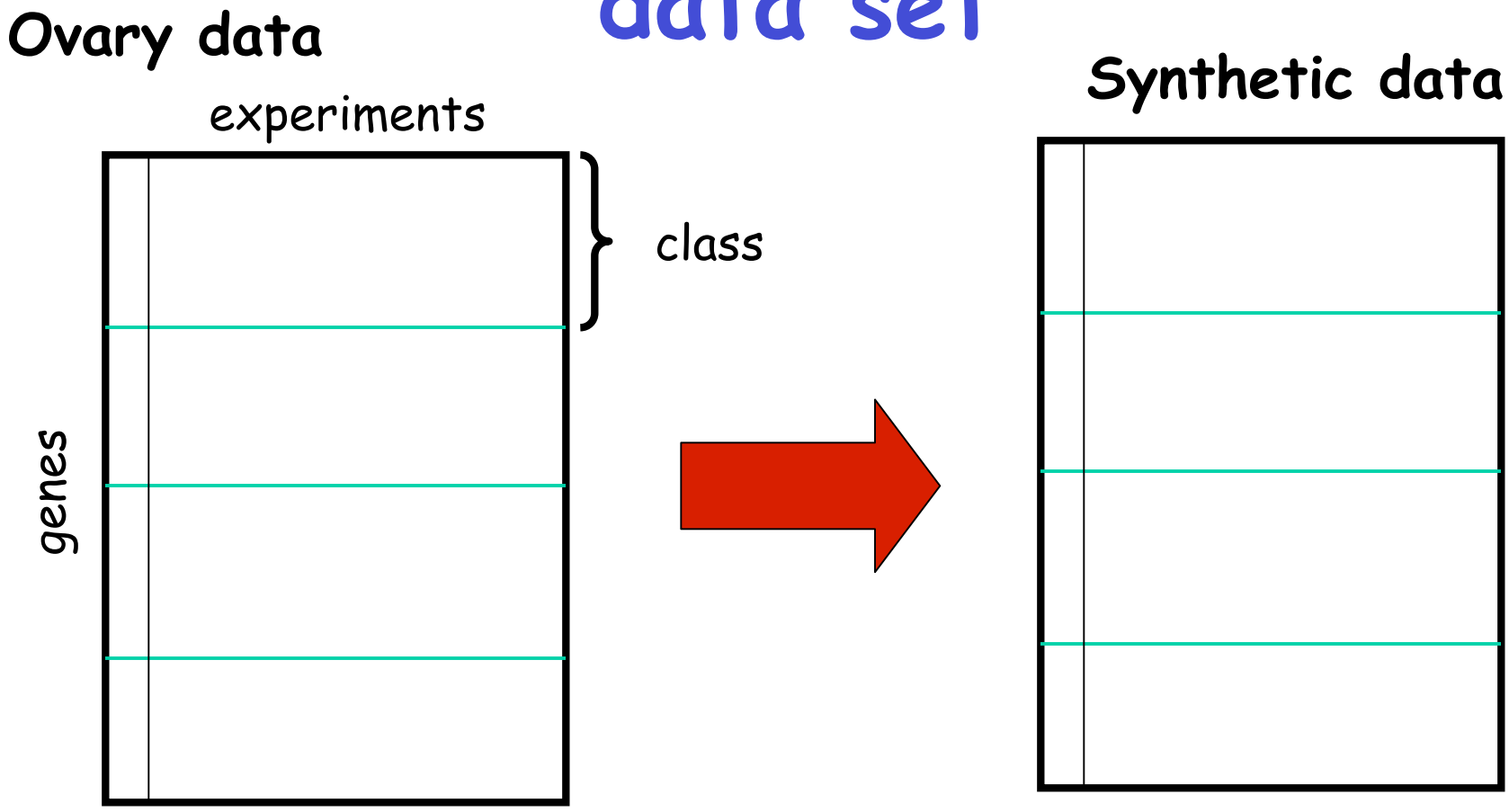
Gene expression data sets

- Ovary data (Michel Schummer, Institute of Systems Biology)
  - Subset of data : 235 clones
    24 experiments (cancer/normal tissue samples)
  - 235 clones correspond to 4 genes (classes)
- Yeast cell cycle data (Cho *et al* 1998)
  - 17 time points
  - Subset of 384 genes correspond to 5 phases of cell cycle

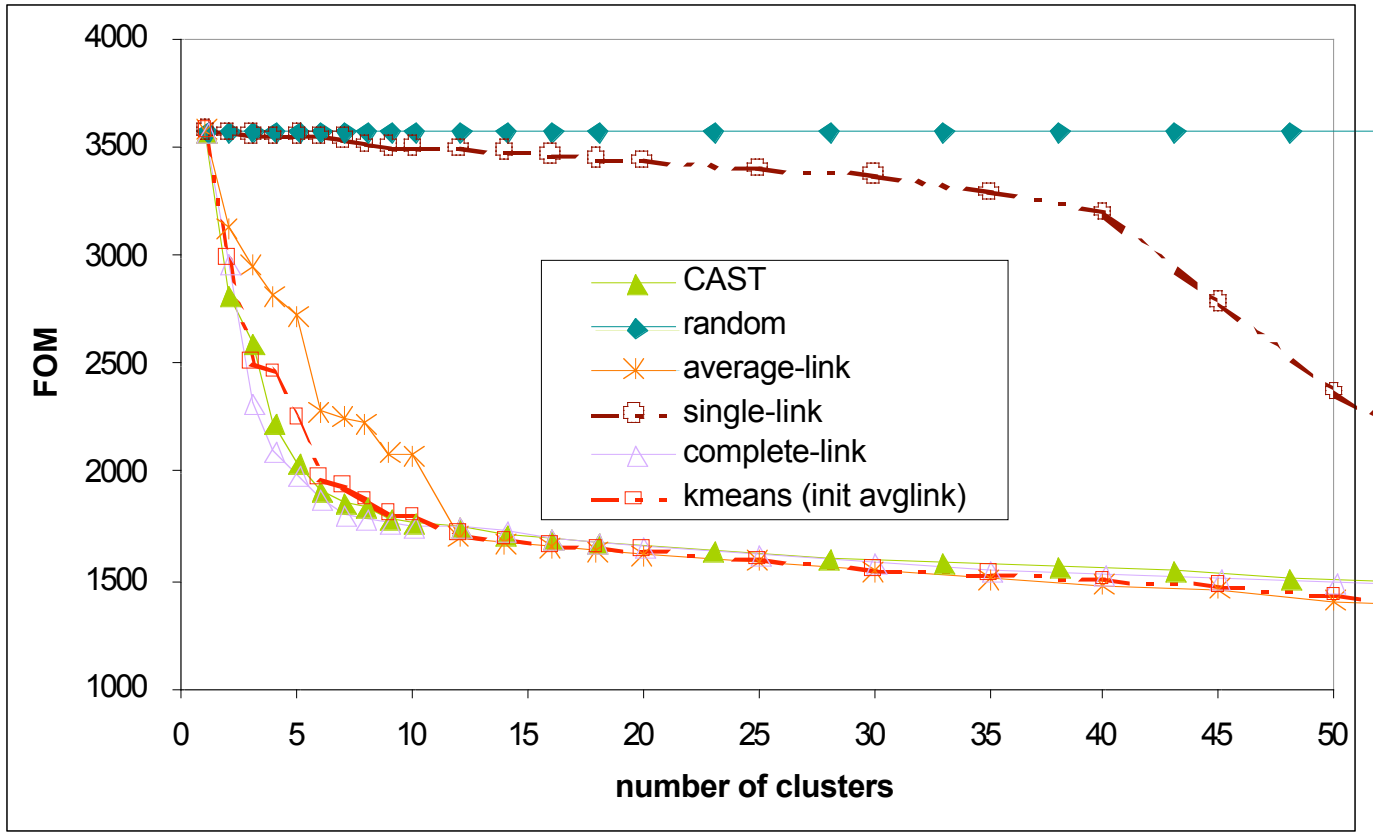
Synthetic data sets

- Mixture of normal distributions based on the ovary data
  - Generate a multivariate normal distributions with the sample covariance matrix and mean vector of each class in the ovary data
- Randomly resampled ovary data
  - For each class, randomly sample the expression levels in each experiment
  - Near diagonal covariance matrix

Randomly resampled synthetic data set

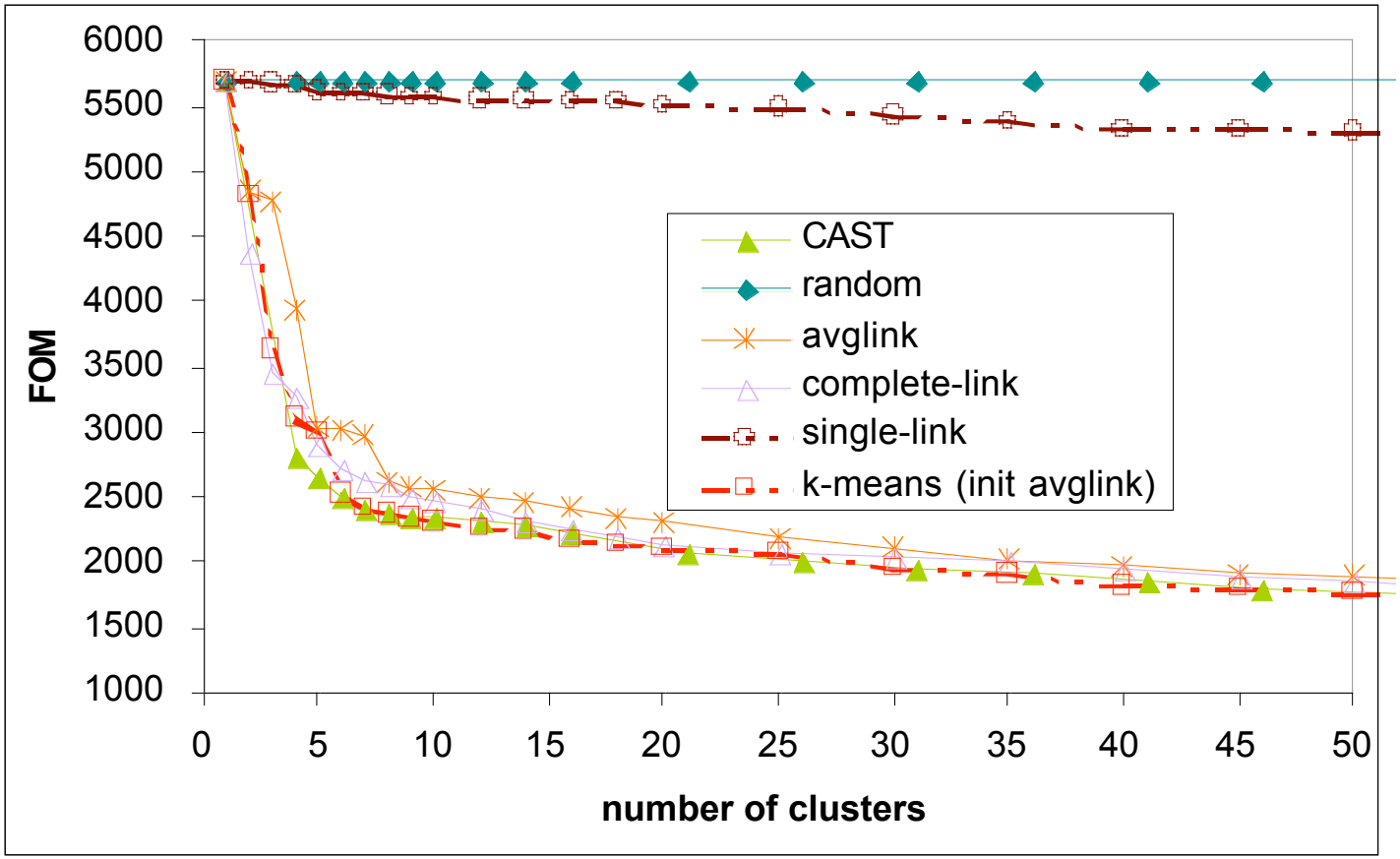


Results: ovary data



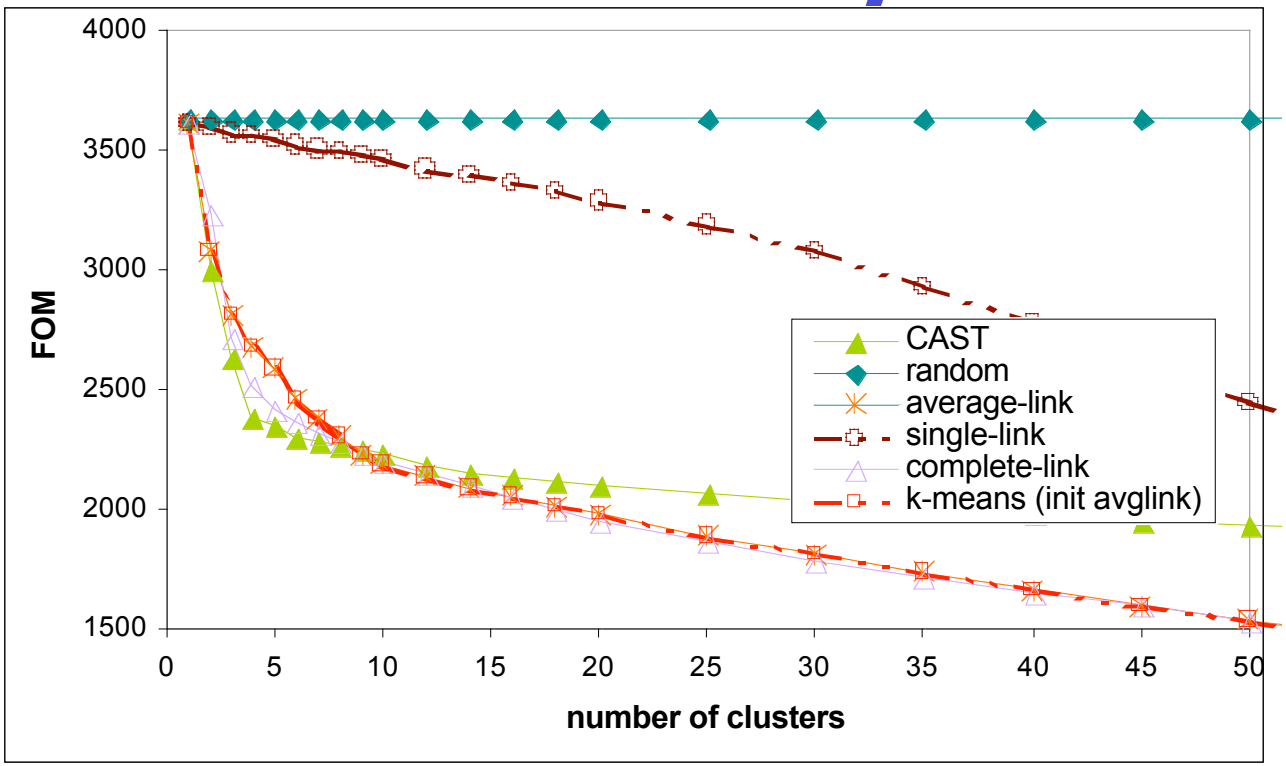
- CAST, k-means and complete-link : best performanace

Results: yeast cell cycle data



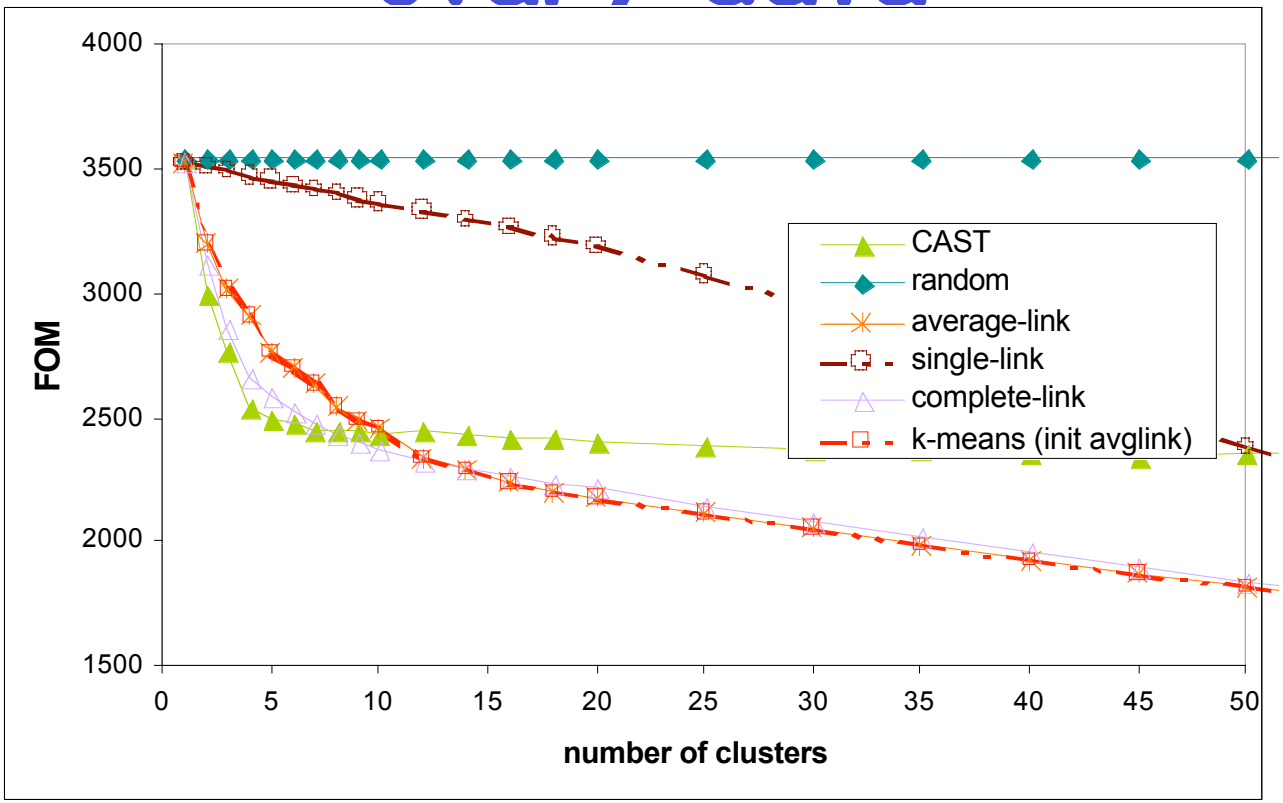
- CAST, k-means: best performance

Results: mixture of normal based on ovary data



- At 4 clusters: CAST lowest FOM

Results: randomly resampled ovary data



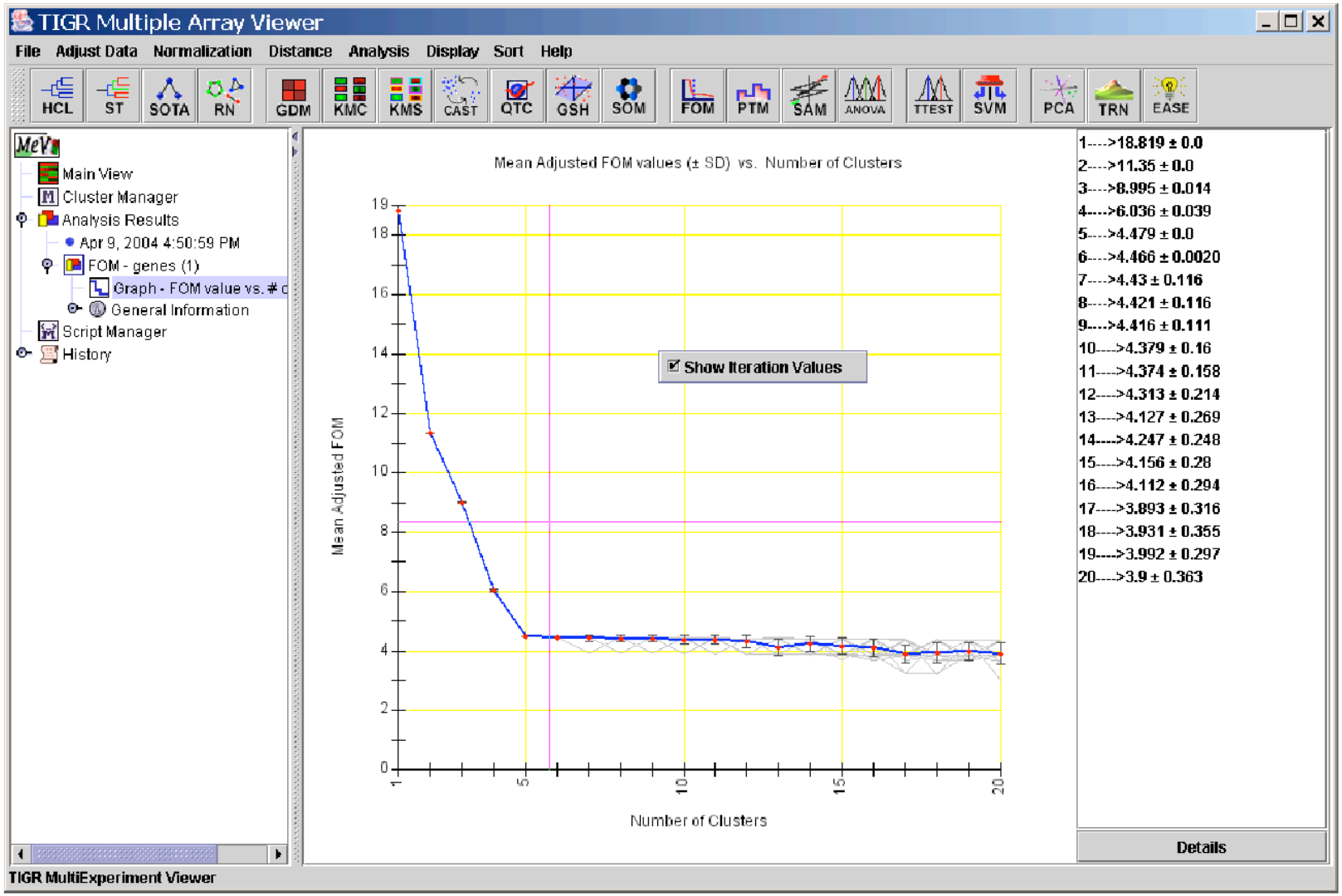
- At 4 clusters: CAST lowest FOM

Summary of Results

- CAST and k-means produce higher quality clusters than the hierarchical algorithms
- Single-link has worse performance among the hierarchical algorithms

Software Implementation

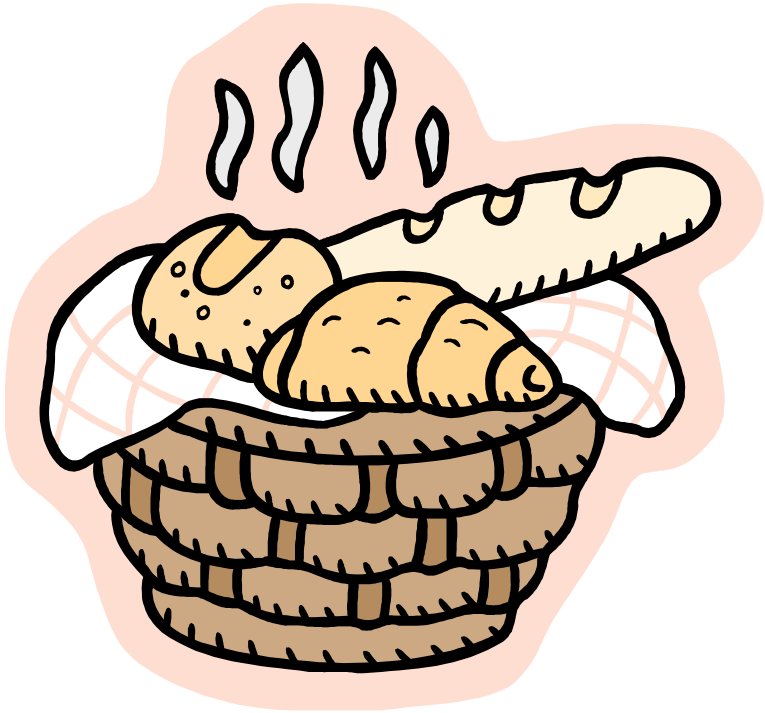
- Command line C code: not very user-friendly at the moment
- Third party implementation: MEV from TIGR
  http://www.tigr.org/software/tm4/mev.html



9.11.1. FOM vs. No. of Clusters graph for KMC algorithm. (1-20 clusters, 20 iterations)

Thank-you's

- FOM work:
  - David Haynor (Radiology, UW)
  - Larry Ruzzo (Computer Science, UW)
- Ovary data: Michel Schummer (Institute of Systems Biology)



Ready for a break?

Overview

- Introduction to microarrays
- Clustering 101
- Validating clustering results on microarray data
- Model-based clustering using microarray data
- Co-expression == co-regulation ??

Common questions

1. How can I choose between all these clustering methods?
2. Is there a clustering algorithm that works better than the others?
3. How to choose the number of clusters?
4. How often do I get biologically meaningful clusters?
5. How many microarray experiments do I need?

Model-based clustering

[Yeung, Fraley, Murua, Raftery, Ruzzo 2001]

- Overview of model-based clustering
- Data sets
- Results
- Summary and Future Work

ISI most cited paper in Computer Science (Jan 2004)

Model-based clustering

- Gaussian mixture model:
  - Assume each cluster is generated by the multivariate normal distribution
  - Each cluster k has parameters :
    - Mean vector: μ_k
    - Covariance matrix: Σ_k
  - Likelihood for the mixture model:

$$L_{mix}(\theta_1,\ldots,\theta_G \mid \mathbf{y}) = \prod_{i=1}^{n} \sum_{k=1}^{G} \tau_k f_k(\mathbf{y}_i \mid \theta_k)$$

- Data transformations & normality assumption

EM algorithm

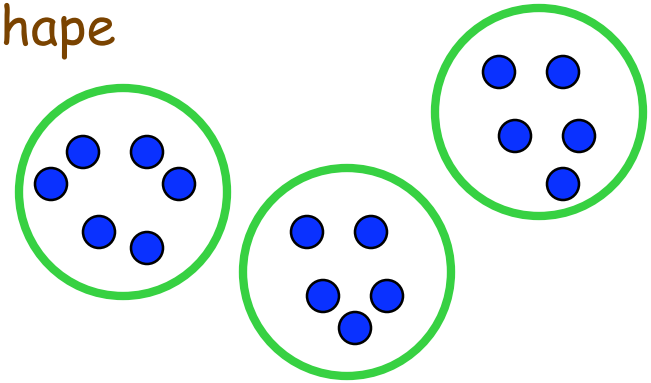
- General appraoch to maximum likelihood
- Iterate between E and M steps:
  - E step: compute the probability of each observation belonging to each cluster using the current parameter estimates
  - M-step: estimate model parameters using the current group membership probabilities

Parameterization of the covariance matrix: $\Sigma_k = \lambda_k D_k A_k D_k^T$ (Banfield & Raftery 1993)

variance: λ_k orientation: D_k shape: A_k

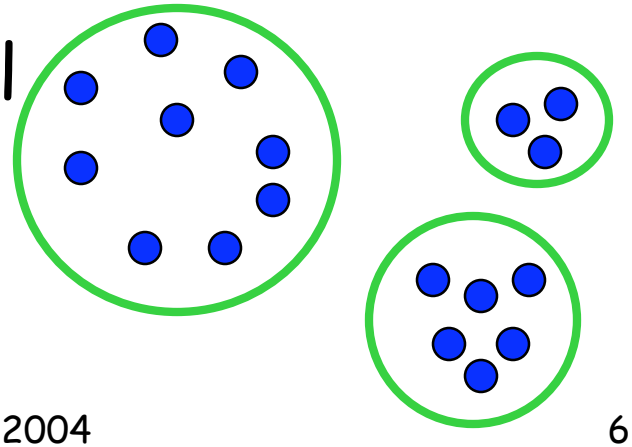
- Equal variance spherical model (EI): ~ kmeans

  $\Sigma_k = \lambda I$



- Unequal variance spherical model (VI):

  $\Sigma_k = \lambda_k I$

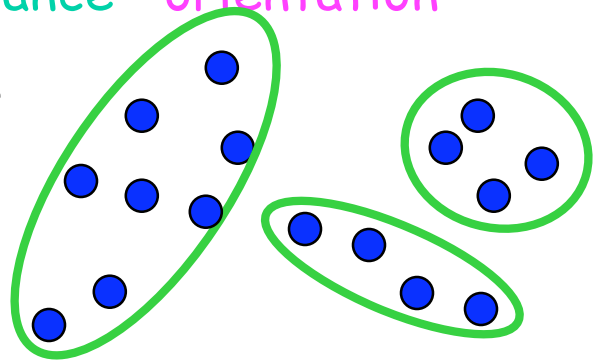


Covariance matrix: $\Sigma_k = \lambda_k D_k A_k D_k^T$

variance: λ_k — orientation: D_k — shape: A_k

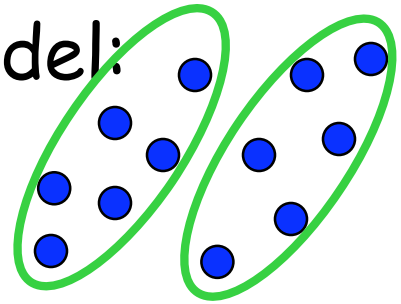
- Unconstrained model (VVV):

  $\Sigma_k = \lambda_k D_k A_k D_k^T$



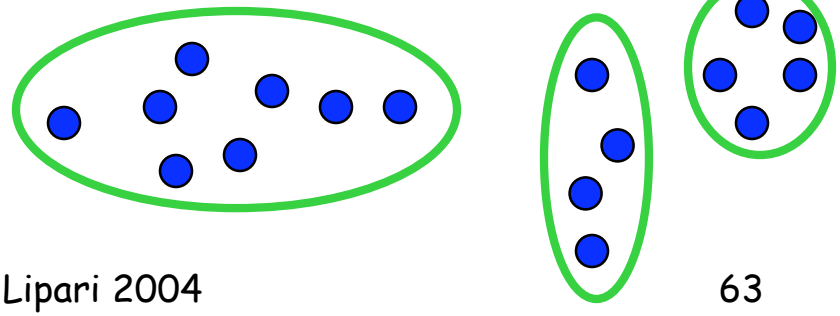
- EEE elliptical model:

  $\Sigma_k = \lambda D A D^T$



- Diagonal model:

  $\Sigma_k = \lambda_k B_k$, where B_k is diagonal with $|B_k| = 1$



Key advantage of the model-based approach: choose the model and the number of clusters

- Bayesian Information Criterion (BIC)
- A large BIC score indicates strong evidence for the corresponding model.

Definition of the BIC score

$$2\log p(D \mid M_k) \approx 2\log p(D \mid \hat{\theta}_k, M_k) - \nu_k \log(n) = BIC_k$$

- The integrated likelihood $p(D|M_k)$ is hard to evaluate,

  where D is the data, M_k is the model.
- BIC is an approximation to log $p(D|M_k)$
- ν_k: number of parameters to be estimated in model M_k

Overall Clustering Approach

For a given range of # clusters (G)

- For each model
  - Apply EM to estimate model parameters and cluster memberships
  - Compute BIC

Our Approach

- Our Goal: To show the model-based approach has superior performance on:
  - Quality of clusters
  - Number of clusters and model chosen (BIC)
- To compare clusters with classes:
  - Adjusted Rand index (Hubert and Arabie 1985)

    High adjusted Rand index ➔ high agreement
- Compare the quality of clusters with a leading heuristic-based algorithm: CAST (Ben-Dor & Yakhini 1999)

Adjusted Rand index

- Compare clusters to classes
- Consider # pairs of objects

	Same cluster	Different cluster
Same class	a	c
Different class	b	d

Example (Adjusted Rand)

	c#1(4)	c#2(5)	c#3(7)	c#4(4)
class#1(2)	2	0	0	0
class#2(3)	0	0	0	3
class#3(5)	1	4	0	0
class#4(10)	1	1	7	1

$$a = \binom{2}{2} + \binom{3}{2} + \binom{4}{2} + \binom{7}{2} = 31$$

$$b = \binom{4}{2} + \binom{5}{2} + \binom{7}{2} + \binom{4}{2} - a = 43 - 31 = 12$$

$$c = \binom{2}{2} + \binom{3}{2} + \binom{5}{2} + \binom{10}{2} - a = 59 - 31 = 28$$

$$d = \binom{20}{2} - a - b - c = 119$$

$$Rand, R = \frac{a+d}{a+d+c+d} = 0.789$$

$$\text{Adjusted Rand} = \frac{R - E(R)}{1 - E(R)} = 0.469$$

Methodology for users:

- **BIC**
- *Not* require classes
- Choose the number of clusters and model

Our evaluation methodology:

- **Adjusted Rand**
- Need classes
- Assess the agreement of clusters to the classes

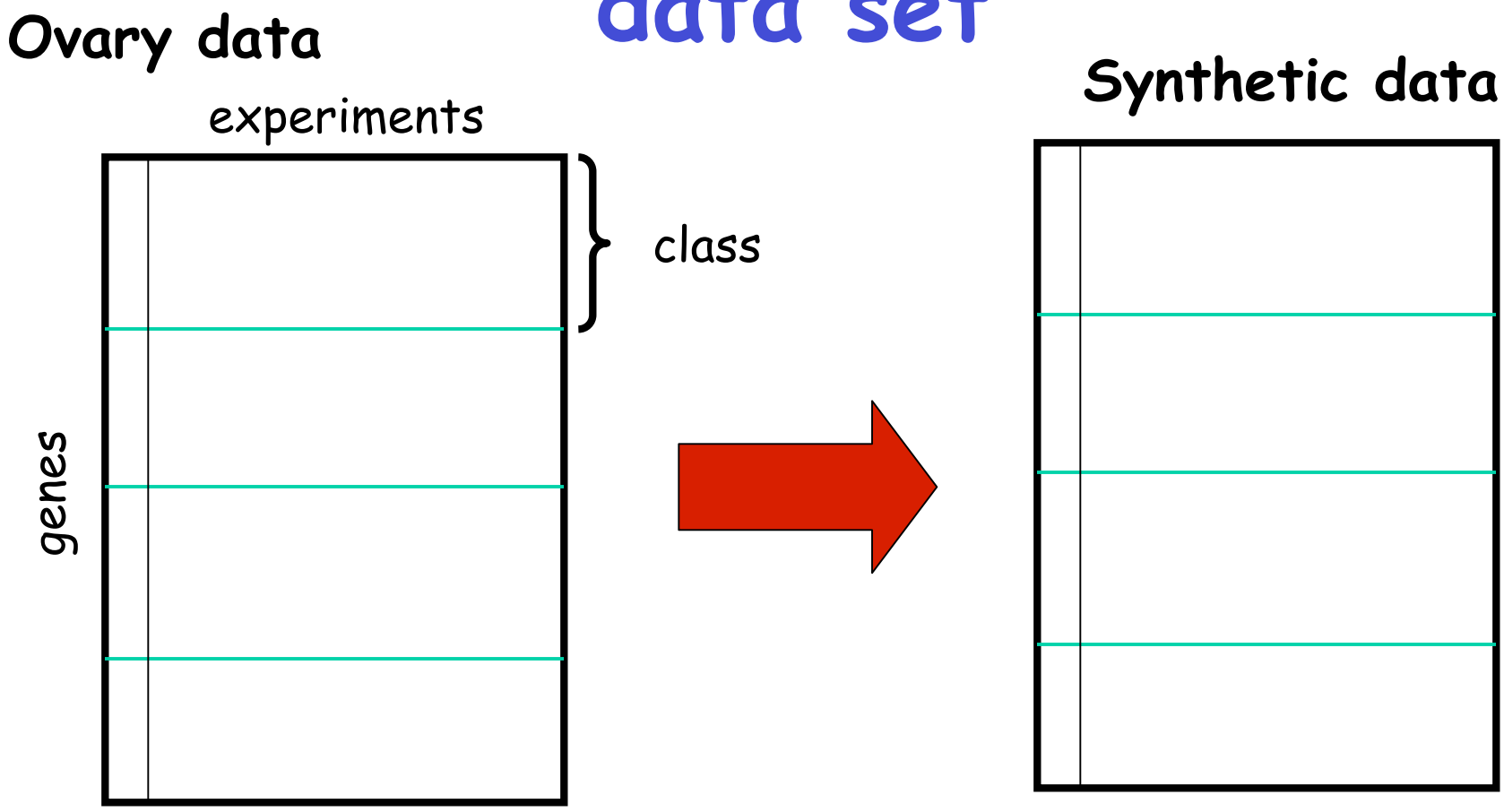
Gene expression data sets

- Ovary data (Michel Schummer, Institute of Systems Biology)
  - Subset of data : 235 clones
    24 experiments (cancer/normal tissue samples)
  - 235 clones correspond to 4 genes (classes)
- Yeast cell cycle data (Cho *et al* 1998)
  - 17 time points
  - Subset of 384 genes correspond to 5 phases of cell cycle

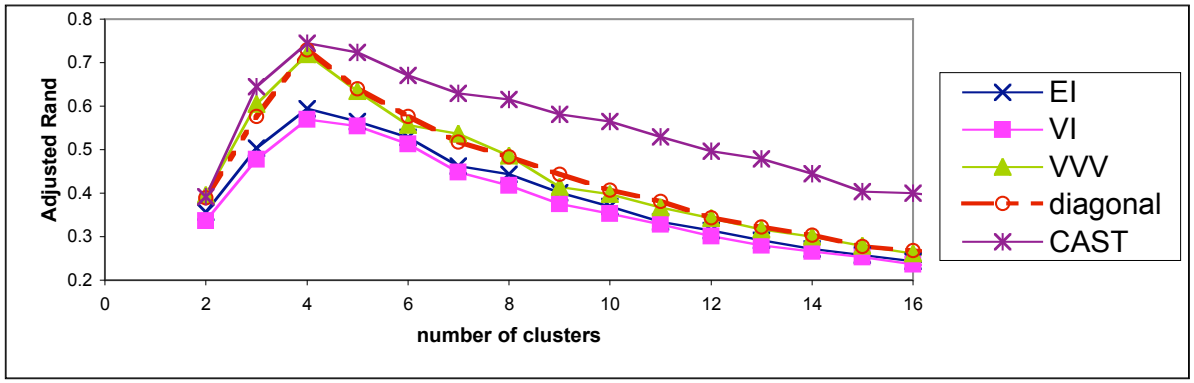
Synthetic data sets

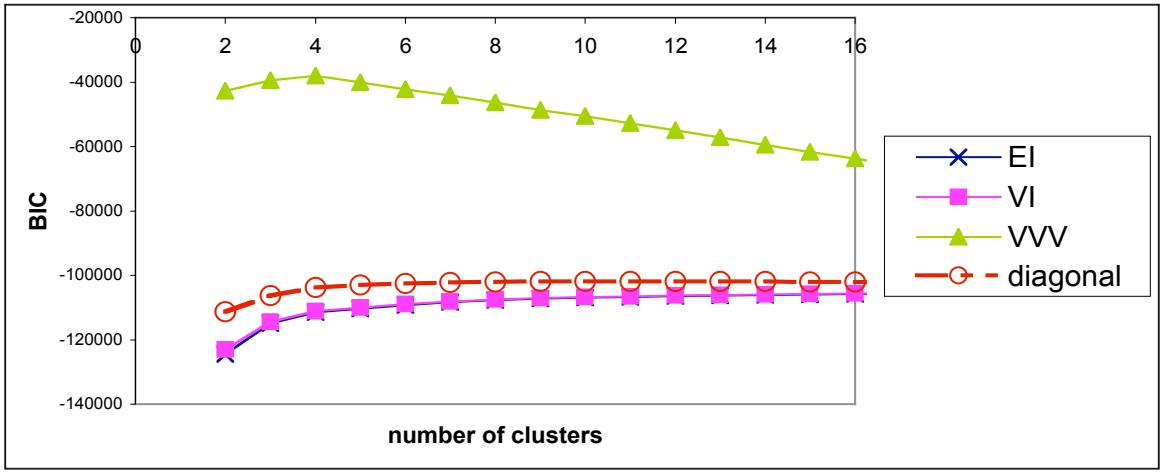
- Mixture of normal distributions based on the ovary data
  - Generate a multivariate normal distributions with the sample covariance matrix and mean vector of each class in the ovary data
- Randomly resampled ovary data
  - For each class, randomly sample the expression levels in each experiment
  - Near diagonal covariance matrix

Randomly resampled synthetic data set



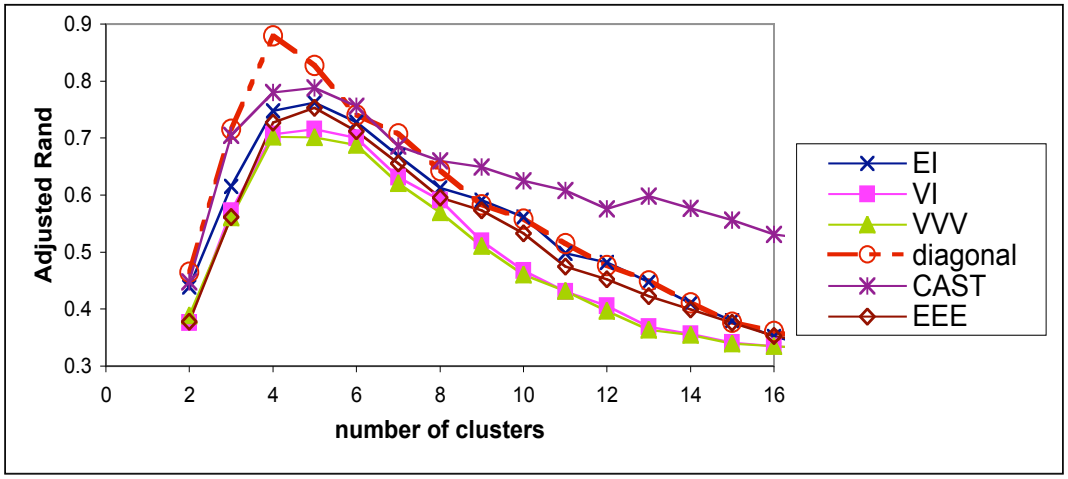
Results: mixture of normal distributions based on ovary data (2350 genes)

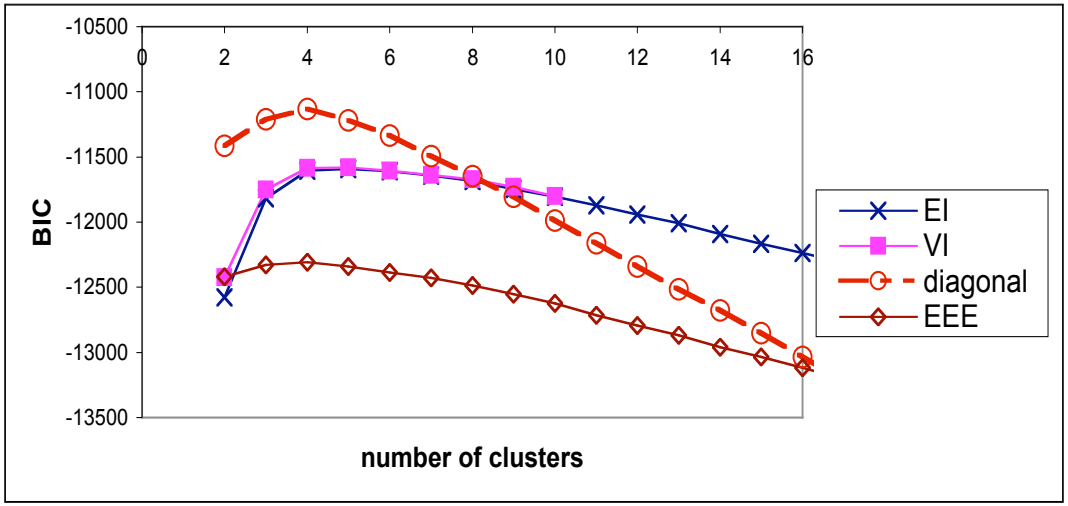




- At 4 clusters, VVV, diagonal, CAST : high adjusted Rand
- BIC selects VVV at 4 clusters.

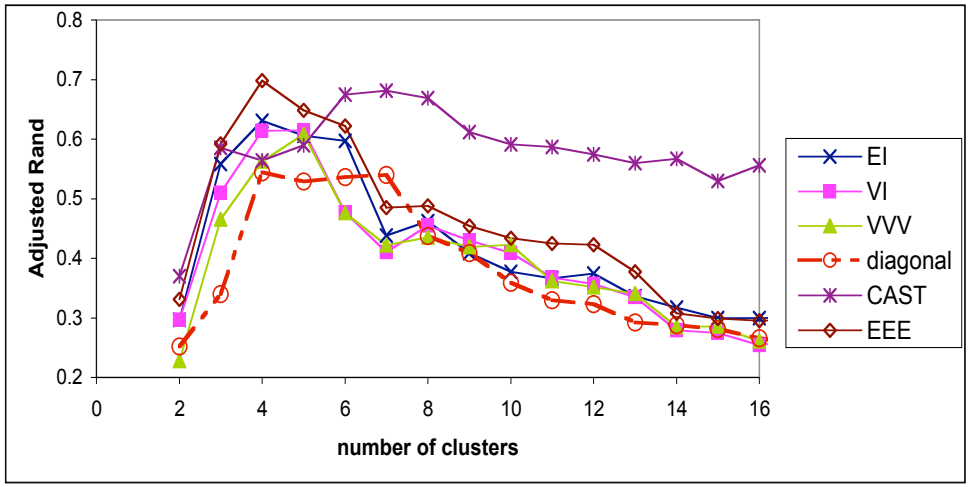
Results: randomly resampled ovary data

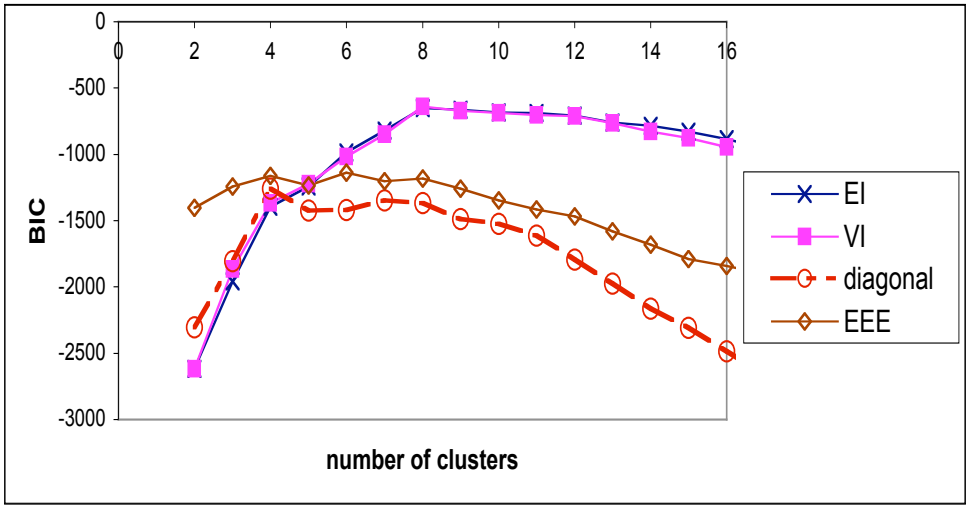




- Diagonal model achieves the max adjusted Rand and BIC score (higher than CAST)
- BIC max at 4 clusters
- Confirms expected result

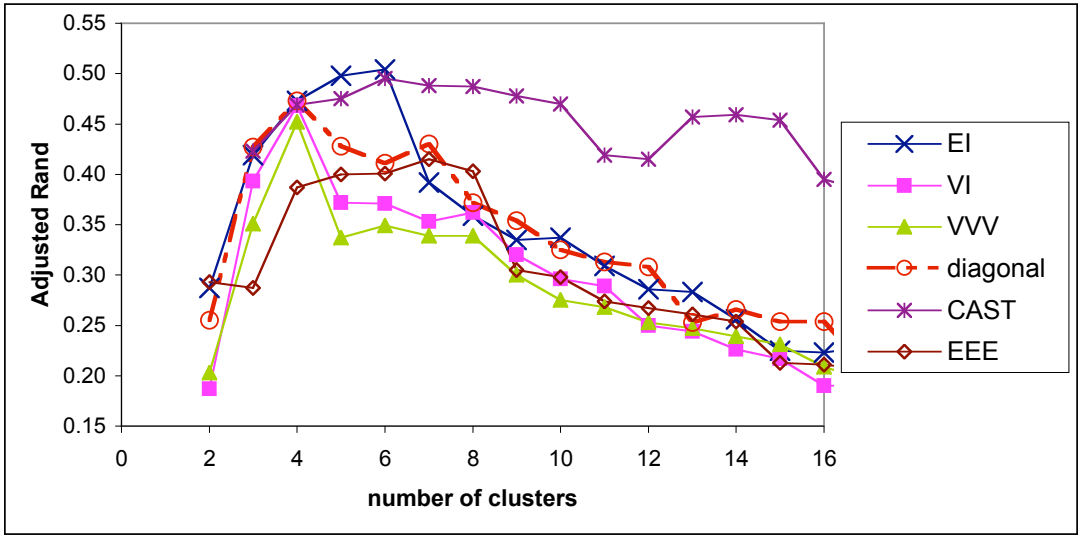
Results: square root ovary data

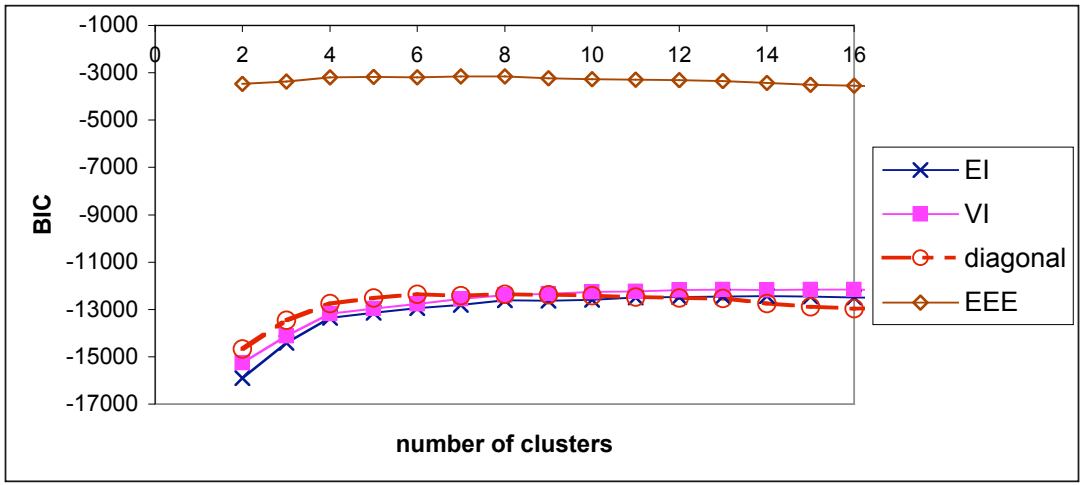




- Adjusted Rand: max at EEE 4 clusters (> CAST)
- BIC analysis:
  - EEE and diagonal models -> first local max at 4 clusters
  - Global max -> VI at 8 clusters

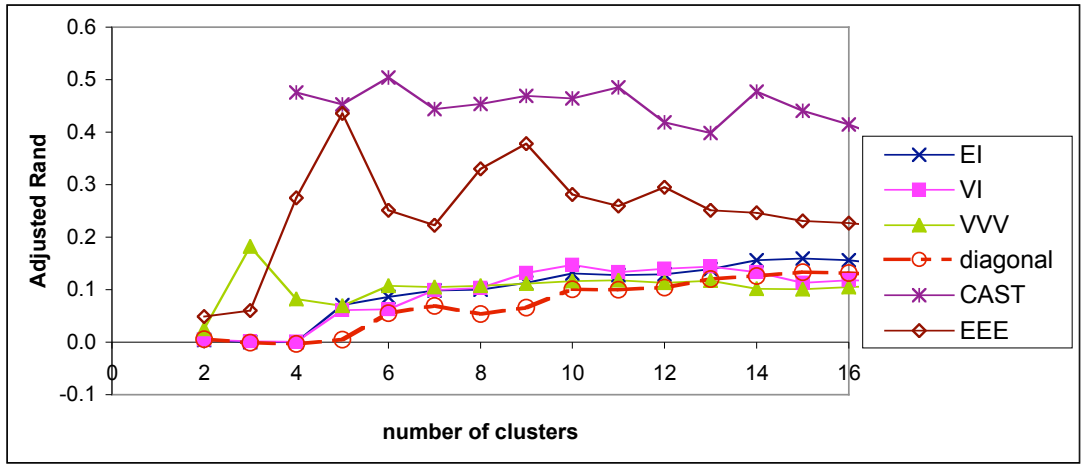
Results: standardized yeast cell cycle data

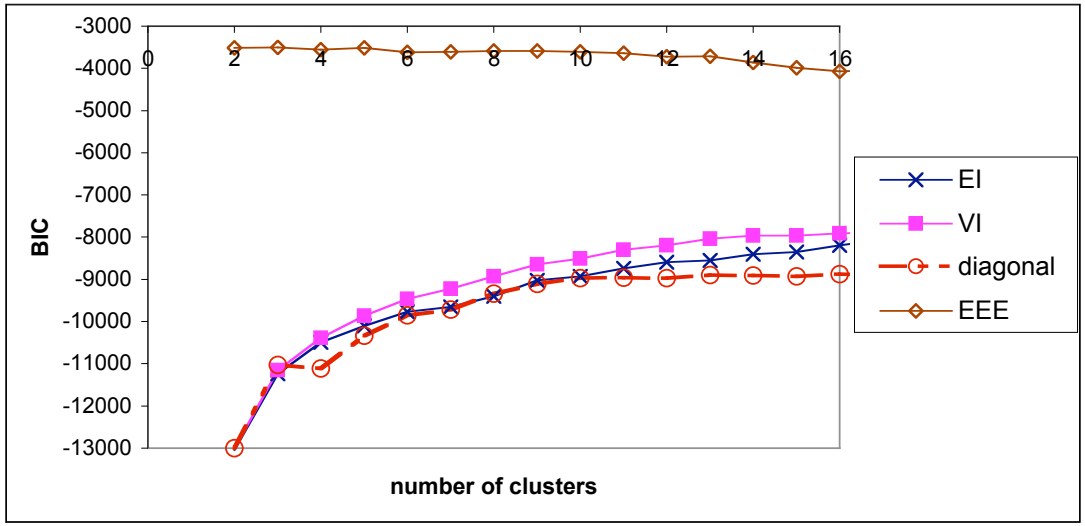




- Adjusted Rand: EI slightly > CAST at 5 clusters.
- BIC: selects EEE at 5 clusters.

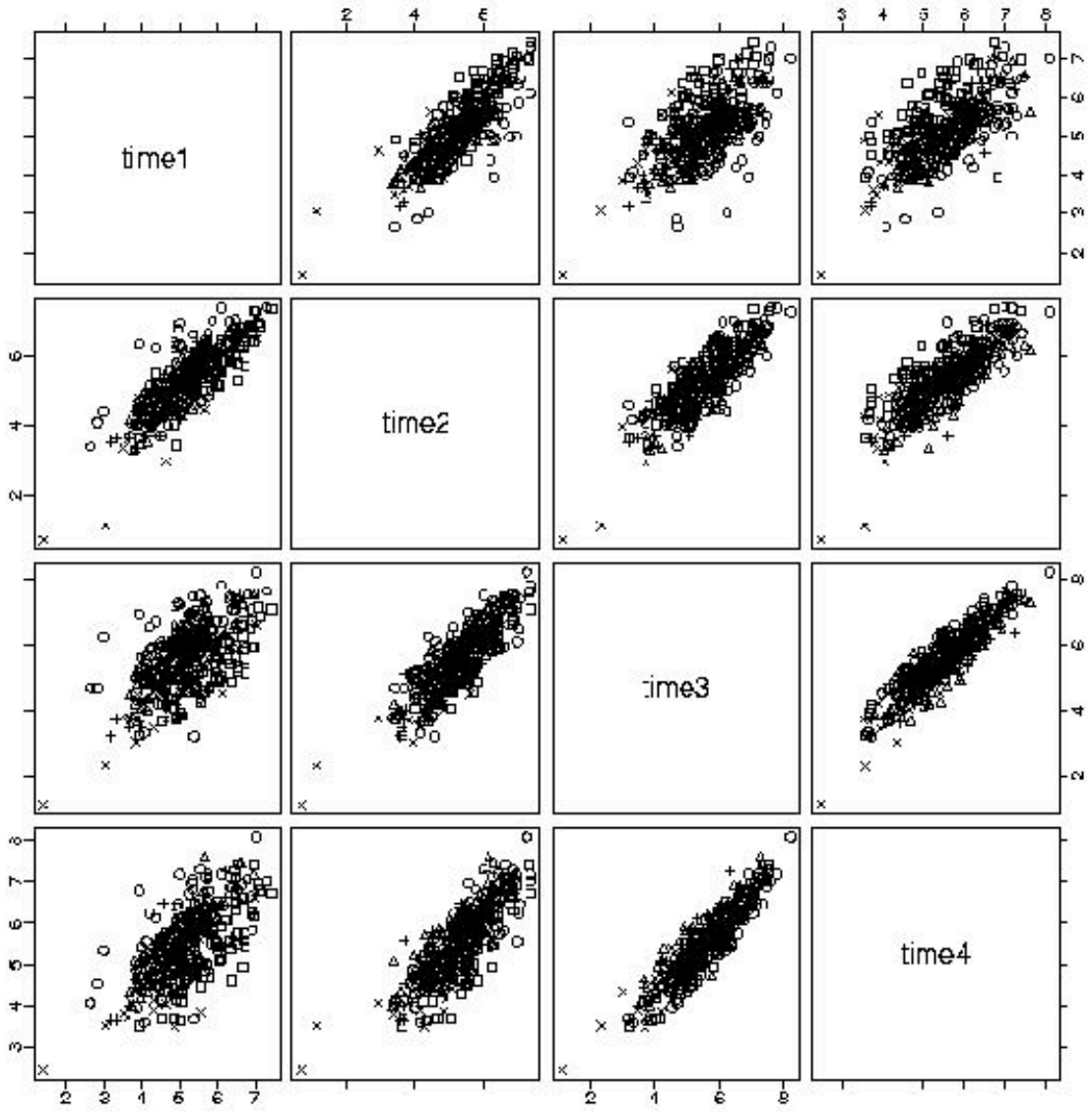
Results: log yeast cell cycle data



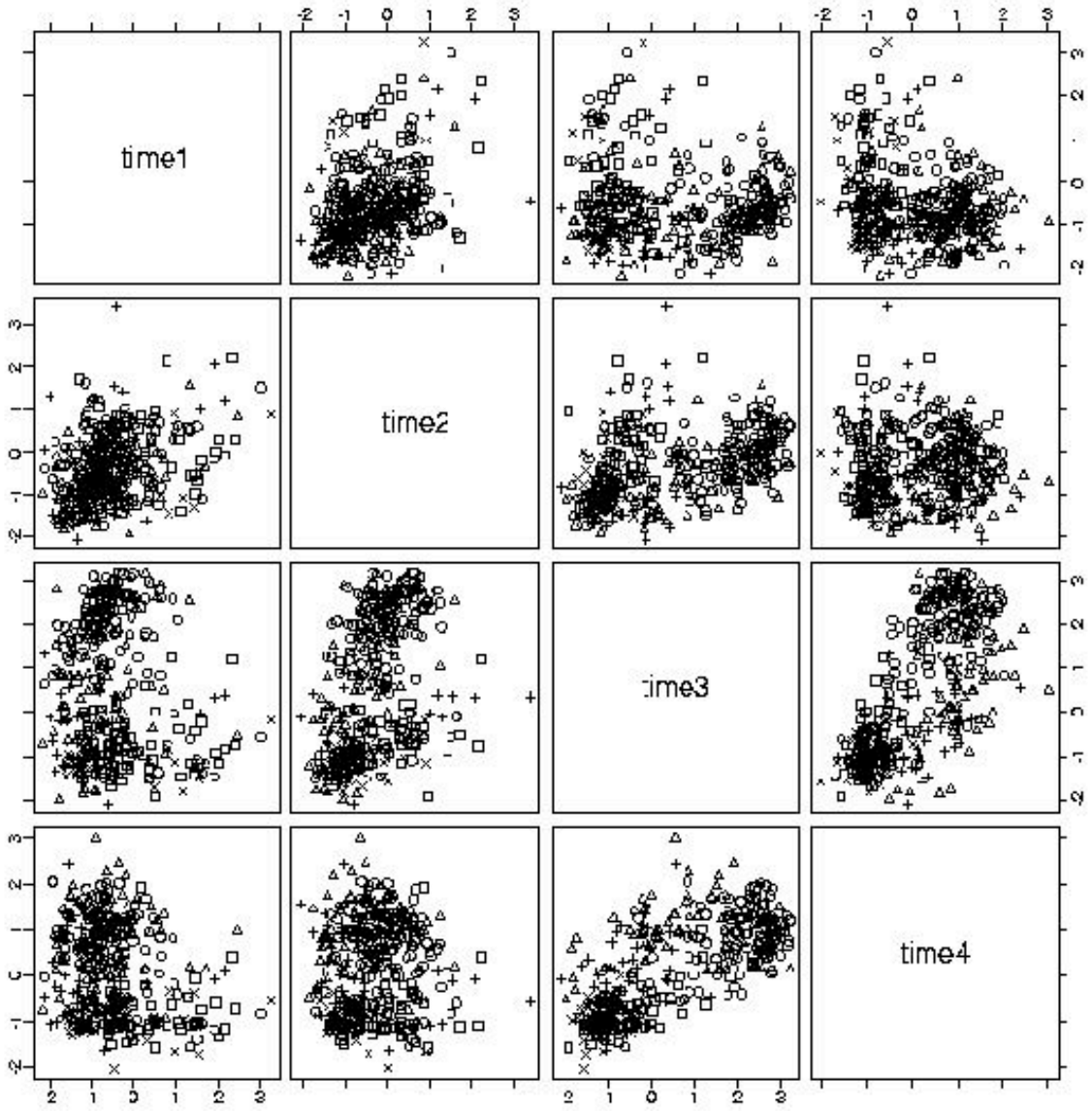


- CAST achieves much higher adjusted Rand indices than most model-based approaches (except EEE).
- BIC scores of EEE much higher than the other models.

log yeast cell cycle data



Standardized yeast cell cycle data



Summary

- Synthetic data sets:
  - With the correct model, the model-based approach excels over CAST
  - BIC selects the right model at the correct number of clusters
- Real expression data sets:
  - Comparable quality of clusters to CAST
  - Advantage: BIC gives a hint to the number of clusters

Software Implementation

- Software: Mclust available in
  - Splus (Chris Fraley and Adrian Raftery)
  - R (Ron Wehrens)
  - Matlab (Angel Martinez and Wendy Martinez)

http://www.stat.washington.edu/mclust/

Future Work

- Custom refinements to the model-based implementation:
  - Design models that incorporate specific information about the experiments,
    eg. Block diagonal covariance matrix
  - Missing data
  - Outliers

Thank-you's

- Model-based work:
  - Chris Fraley (Statistics, UW)
  - Alejandro Murua (Statistics, UW)
  - Adrian Raftery (Statistics, UW)
  - Larry Ruzzo (Computer Science, UW)
- Ovary data:
  - Michel Schummer (Institute of Systems Biology)

Common questions

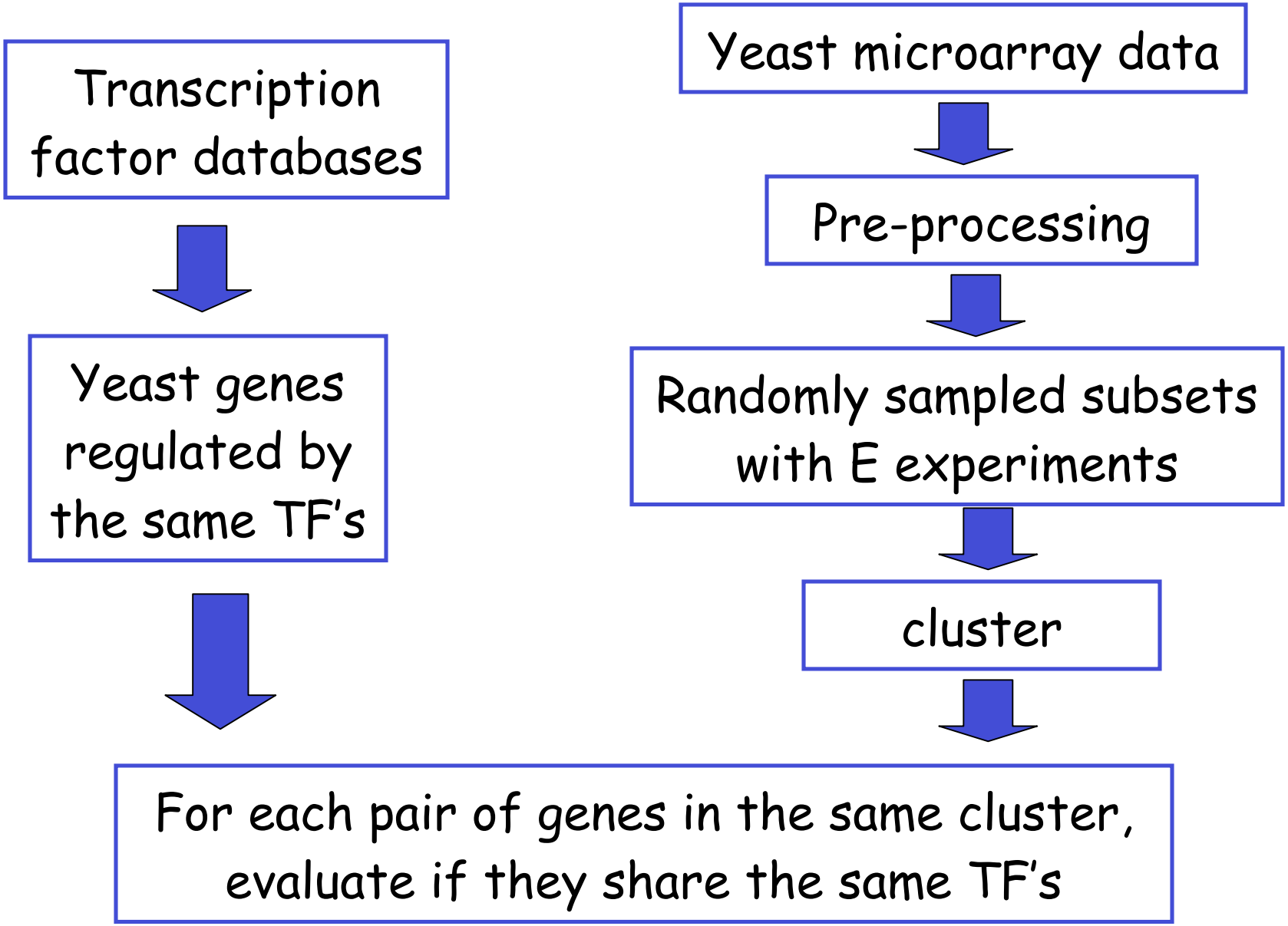
1. How can I choose between all these clustering methods?
2. Is there a clustering algorithm that works better than the others?
3. How to choose the number of clusters?
4. How often do I get biologically meaningful clusters?
5. How many microarray experiments do I need?

From co-expression to co-regulation: How many microarray experiments do we need?

[Yeung, Medvedovic, Bumgarner:
To appear in Genome Biology 2004]

From co-expression to co-regulation

- Motivation:
  - Genes sharing the same transcriptional modules are expected to produce similar expression patterns
  - Cluster analysis is often used to identify genes that have similar expression patterns.
- Questions:
  1. How likely are co-expressed genes regulated by the same transcription factors?
  2. What is the effect of the following factors on the likelihood:
     a. Number of microarray experiments
     b. Clustering algorithm used


Transcription factor databases
Yeast genes regulated by the same TF's
Yeast microarray data
Pre-processing
Randomly sampled subsets with E experiments
cluster
For each pair of genes in the same cluster, evaluate if they share the same TF's

Yeast transcription factors

- SCPD (Saccharomyces Cerevisiae Promoter Database) [zhang *et al.* 1999]
  - List ~235 genes that are regulated by 90 transcription factors (TF's)
- YPD (Yeast Protein Database)
  - Commercial: UW does not have access
  - Appendix of [Lee et al. 2002]
  - List genes regulated by each TF from literature as of Nov 2001
  - List ~584 genes that are regulated by 120 TF's

Comparing YPD and SCPD

	SCPD	YPD	Common
# distinct ORF's	235	584	156
# distinct TF's	108	120	34
# gene-TF interactions	473	1056	119

- SCPD: 41/90 = 46% TF's regulate only 1 gene
- YPD: 17/120 = 14% TF's regulate only 1 gene
- In general, the YPD list contains TF's that regulate a higher # genes

Yeast microarray data

- Rosetta's yeast compendium data [Hughes et al. 2000]
  - 300 knockout 2-color experiments
- Stanford: Gasch *et al.* Data [2000 and 2001]
  - cDNA array data under a variety of environmental stress (eg. heat shock)
  - Total 225 concatenated time course experiments

Evaluation

- For each clustering result
  - Count the number of pairs of genes that belong to the same cluster and share a common TF (*True positive, TP*)

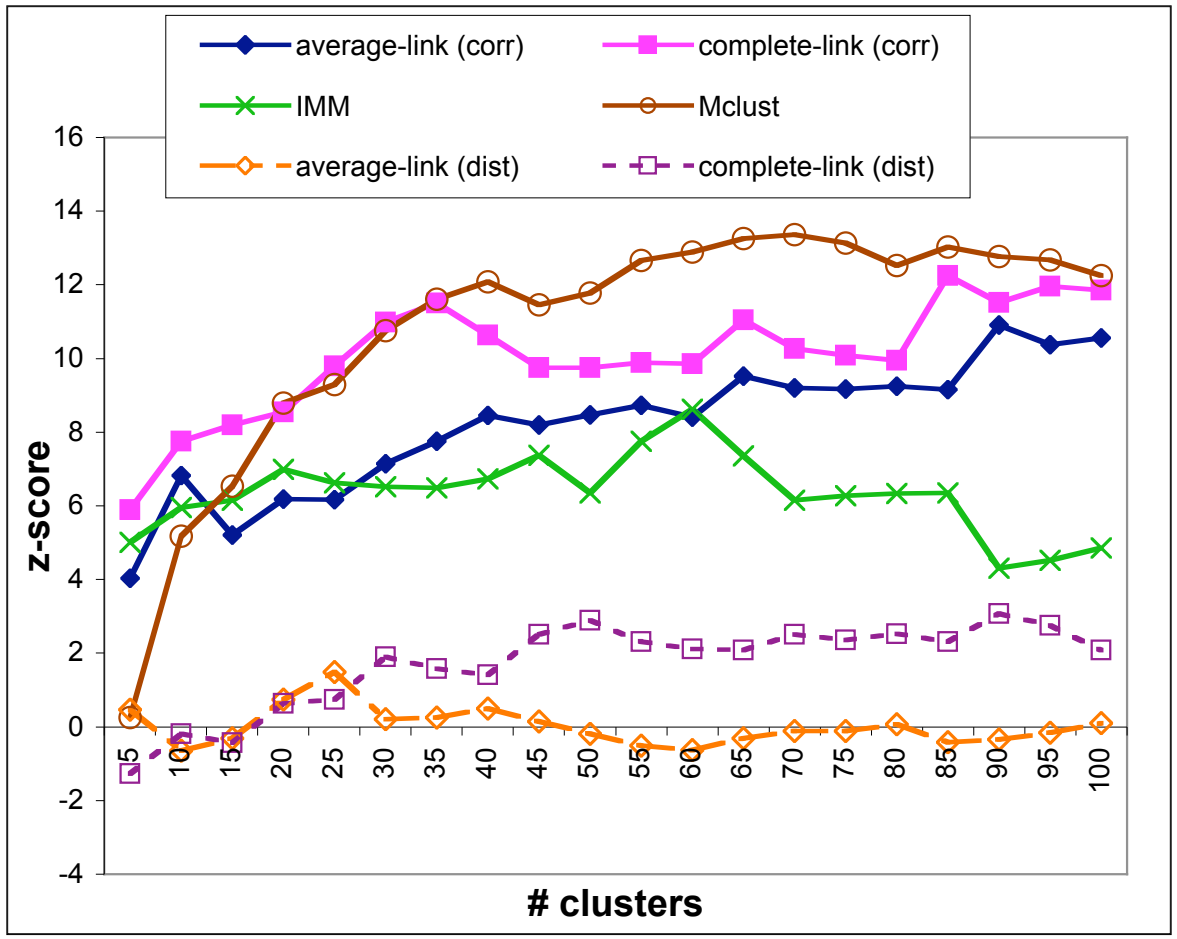
$$TP\,rate = \frac{\#\text{ gene pairs from the same clusters and share at least 1 common TF's}}{\#\text{ gene pairs from the same clusters}}$$

- TP rate may change as a function of # clusters, we compare the TP rate to the TP rate of random partitions:
  - Randomly partition the set of genes 1000 times
  - Distribution of TP rate ~ Normal → mean □ and standard deviation □
  - Z-score = (TP rate - □) / □
  - A high z-score → TP rate is significantly higher than those of random partitions

Results: Compendium data using all experiments

- To compare the performance of different clustering algorithms

compendium data & SCPD (273 E)



MCLUST and complete-link (corr) produced relatively high z-scores

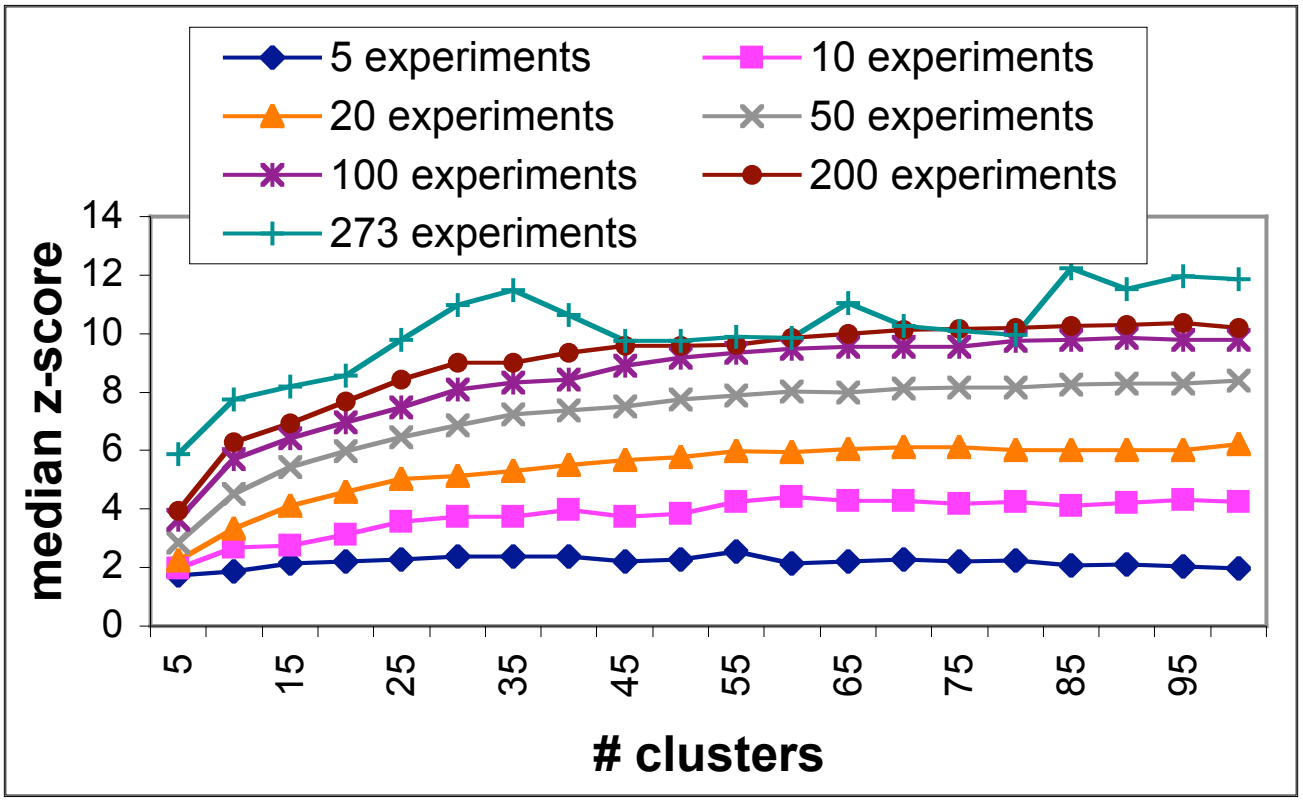
IMM (Infinite Mixture Model-based)

- Infinite mixture model:
  - Each cluster is assumed to follow a multivariate normal distribution
  - do *NOT* assume # clusters
  - Use a Gibbs Sampler to estimate the pairwise probabilities (Pij) for two genes (i,j) to belong to the same cluster
  - To form clusters: cluster Pij with a heuristic clustering algorithm (eg. complete-link)
- Built-in error model
  - Assume the repeated measurements are generated from another multivariate Gaussian distribution.

Results: Compendium data: effect of # experiments

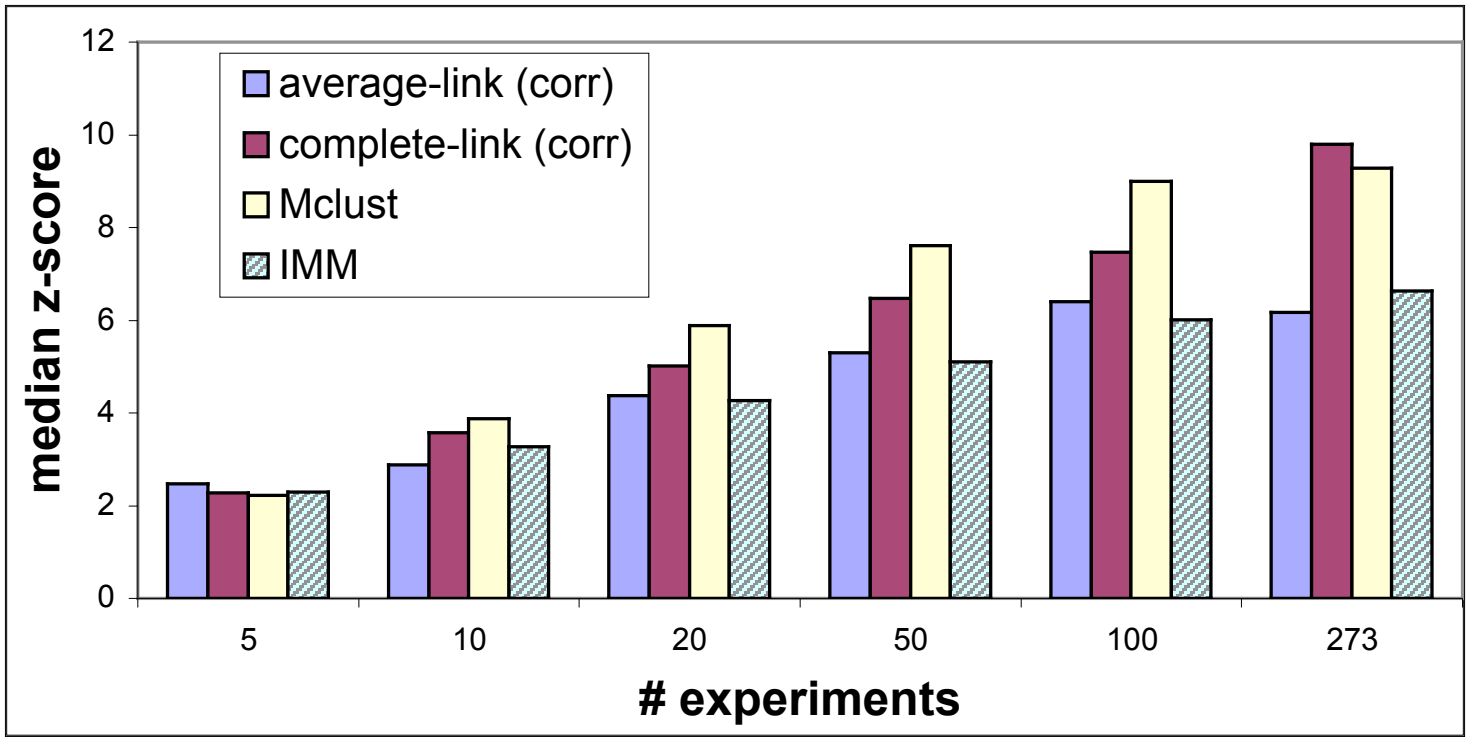
Compendium data & SCPD:

Hierarchical complete-link over a range of # clusters



Observation: Median z-score increases as # experiments increases over different # clusters.

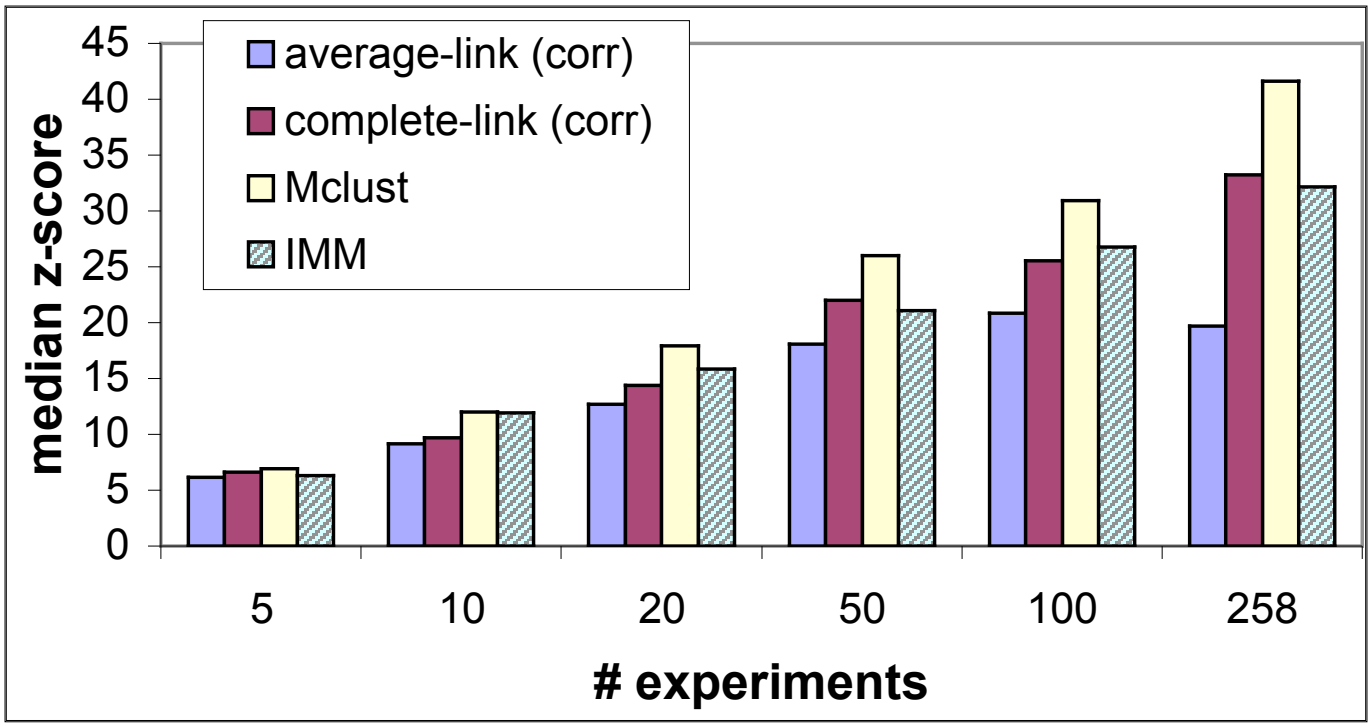
Compendium data & SCPD:
Different clustering algorithms at 25 clusters



- Proportions of co-regulated genes increase as # experiments increases
- Mclust: highest proportions of co-regulated genes

Compendium data & YPD (537G):

Different clustering algorithms at 40 clusters

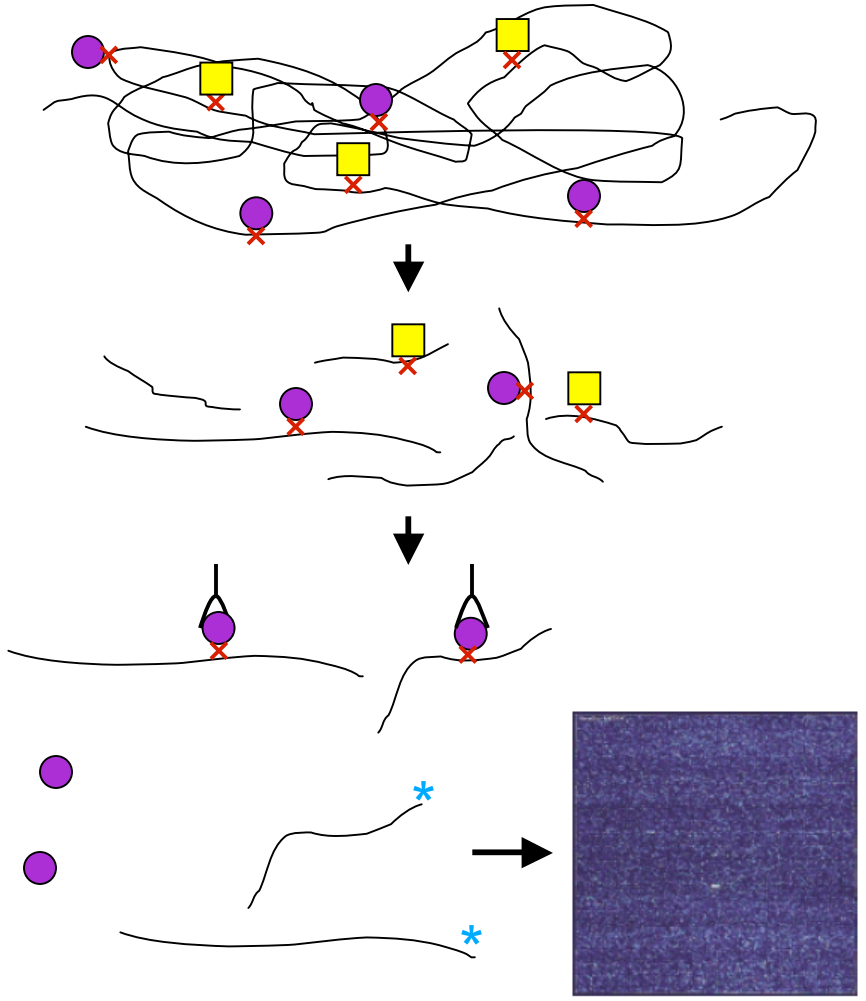


Summary of Results

- More microarray experiments → More likely to find co-regulated genes!!
- SCPD/YPD produces similar results
- Euclidean distance tends to produce relatively low z-score compared to correlation using the same algorithm
- Standardization greatly improves the performance of model-based methods
- Mclust (EI model) produces relatively high z-scores
- IMM doesn't work as well as Mclust. Why??

ChIP-CHIP—the methodology

- Transcription factors are crosslinked to genomic DNA
- DNA is sheared
- Antibodies immunoprecipitate a specific transcription factor
- DNA is un-linked, labeled and used to interrogate arrays



ChIP data:3rd gold standard

[Lee et al. Science 2002]

- Chromatin Immunoprecipitation (ChIP): to detect the binding of TF's of interest to integenic sequences in yeast in vivo
- 106 TF's from YPD (113 TF's in their raw data)
- Adopted error model from [Hughes et al. 2000]
- Raw data (log ratios and p-values for genes/integenic regions to TF's) available
- p-value cutoff = 0.001

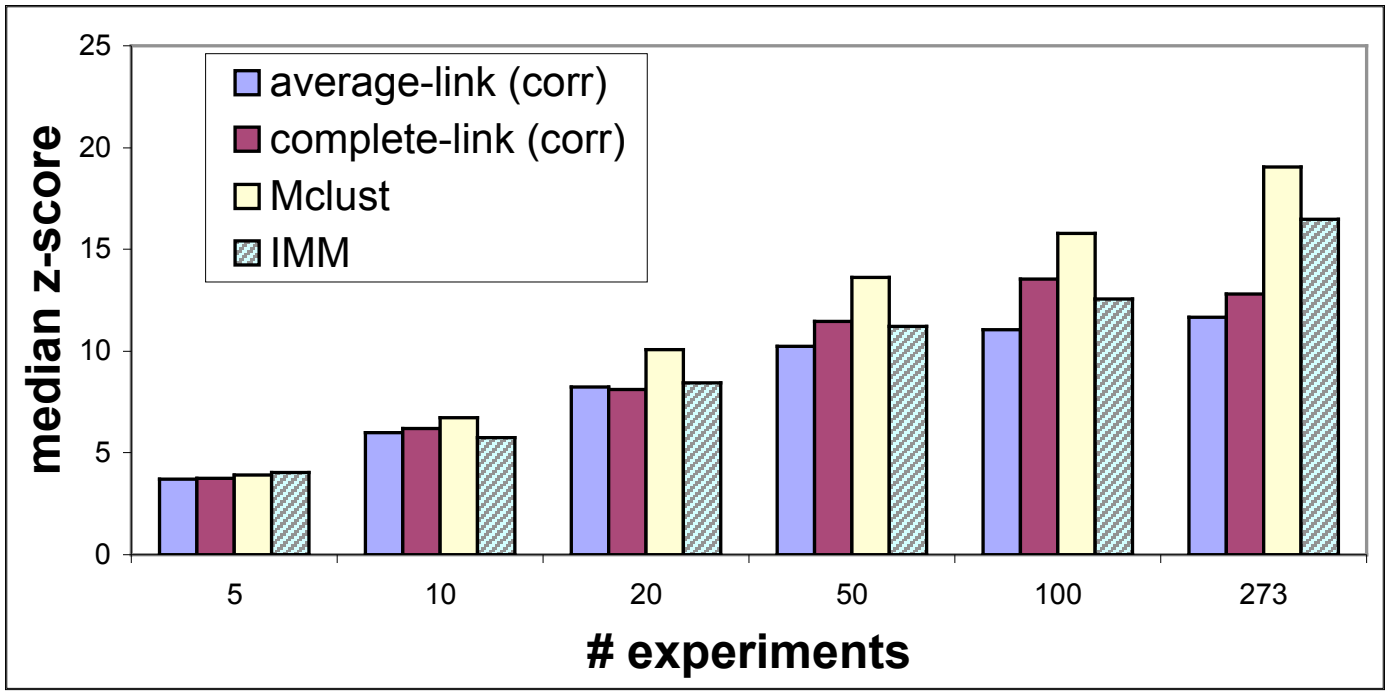
Comparing ChIP data and YPD

- 791 gene-TF interactions from YPD have a common gene and TF in the ChIP data

p-values range	# gene-TF interactions	% gene-TF interactions
0 < x <= 0.001	159	20.10
0.001 < x <= 0.005	37	4.68
0.005 < x <= 0.01	22	2.78
0.01 < x <= 0.05	60	7.59
0.05 < x <= 0.1	40	5.06
0.1 < x <= 0.5	232	29.33
x > 0.5	241	30.47

p-value	TP	FN	total # ChIP interactions	# detected by ChIP but not in YPD	TP %	FN %
0.0010	159	632	642	421	20.10	79.90
0.0050	196	595	1101	775	24.78	75.22
0.0100	218	573	1517	1112	27.56	72.44
0.0500	278	513	3655	2759	35.15	64.85
0.1000	318	473	6226	4722	40.20	59.80

Results: compendium data & ChIP (215 genes)



Very similar results on other datasets as well

Take home message

In order to reliably infer co-regulation from cluster analysis, we need lots of data.

Limitations

- Very naïve assumption of co-regulation: genes sharing at last one common transcription factors
- Yeast data only
- Does not take the information limit of microarray datasets into consideration
- Consider only clustering algorithms in which each gene is assigned to only one cluster
- Our current study does not provide completely quantitative results: how many experiments are sufficient to achieve x% co-regulation?

Thank you's

- Roger Bumgarner (Microbiology, UW)
- Bumgarner Lab, UW:
  - Tanveer Haider, Tao Peng, Mette Peters, Kyle Serikawa, Caimiao Wei
- Ted Young -- Biochemistry, UW
- IMM: Mario Medvedovic -- Univ of Cincinnati
- Mclust: Adrian Raftery + Chris Fraley (Statistics, UW)

Summary

Question	Answer
1. How can I choose between different clustering methods?	FOM: compare any clustering algorithms on any dataset
2. Is there a clustering algorithm that works better than the others? 3. How to choose the number of clusters?	Model-based clustering algorithm: • high cluster quality • estimated number of clusters.
4. How often do I get biologically meaningful clusters? 5. How many experiments do I need?	• More experiments → more likely to find co-regulated genes • Model-based method ☺ • in yeast, ~ 50 experiments

Meet my collaborators and mentors

David Haynor

Mario Medvedovic

Roger Bumgarner

Adrian Raftery

Larry Ruzzo

Key References

- [Yeung, Haynor, Ruzzo 2001] Validating clustering for gene expression data. Bioinformatics 17:309-318
- [Yeung, Fraley, Murua, Raftery, Ruzzo 2001] Model-based clustering and data transformations for gene expresison data. Bioinformatics 17:977-987
- [Yeung, Medvedovic, Bumgarner 2004] From co-expression to co-regulation: how many microarray experiments do we need? To appear in Genome Biology.

http://faculty.washington.edu/kayee/