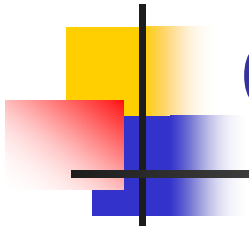
Evaluation and comparison of gene clustering methods in microarray analysis

George C. Tseng
Department of Biostatistics
Department of Human Genetics
University of Pittsburgh



Outline

1. Background
2. Methods
   a) Six gene clustering methods compared
3. Simulation study
   a) Weighted Rand index for comparing two clusterings with noise
   b) Simulation model (hierarchical log-normal model)
   c) Result
4. Real data
   a) Prediction-accuracy plot
   b) Result
5. Conclusion and discussion

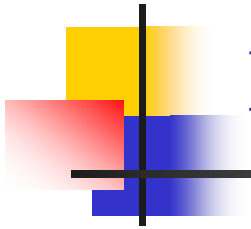
1. Background

Cluster analysis:

Data $X=\{x_i, i=1,\ldots, n\}$, each object $x_i \in R^p$.

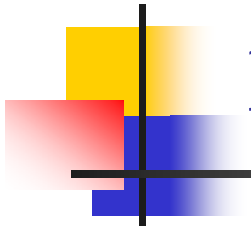
Given a dissimilarity measure $d(x_i, x_j)$, assign the n objects into k disjoint clusters; i.e. $C=\{C_1, \ldots, C_k\}$ and $X = \bigcup_{j=1}^{k} C_j$.

- **Cluster genes** (n: # of genes; p: # of samples): similar expression pattern implies co-regulation.
- **Cluster samples** (n: # of samples; p: # of genes): identify potential sub-classes of disease.
- **Bi-clustering**



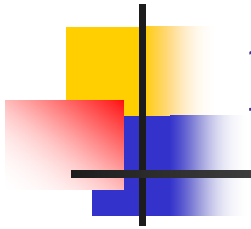
1. Background

- Alternative analysis for detecting regulatory interactions
  - shortest path (Zhou et al., 2002); liquid association (Li 2001)
  - Boolean network; Bayesian network
- Despite the above complimentary choices, gene clustering remains a useful routine in most microarray analysis.



1. Background

- Difficulty of evaluation in clustering:
  - Supervised machine learning (classification): the class labels are known. Performance of different methods evaluated through cross validation.
  - Unsupervised machine learning (clustering): the underlying cluster structure unknown.



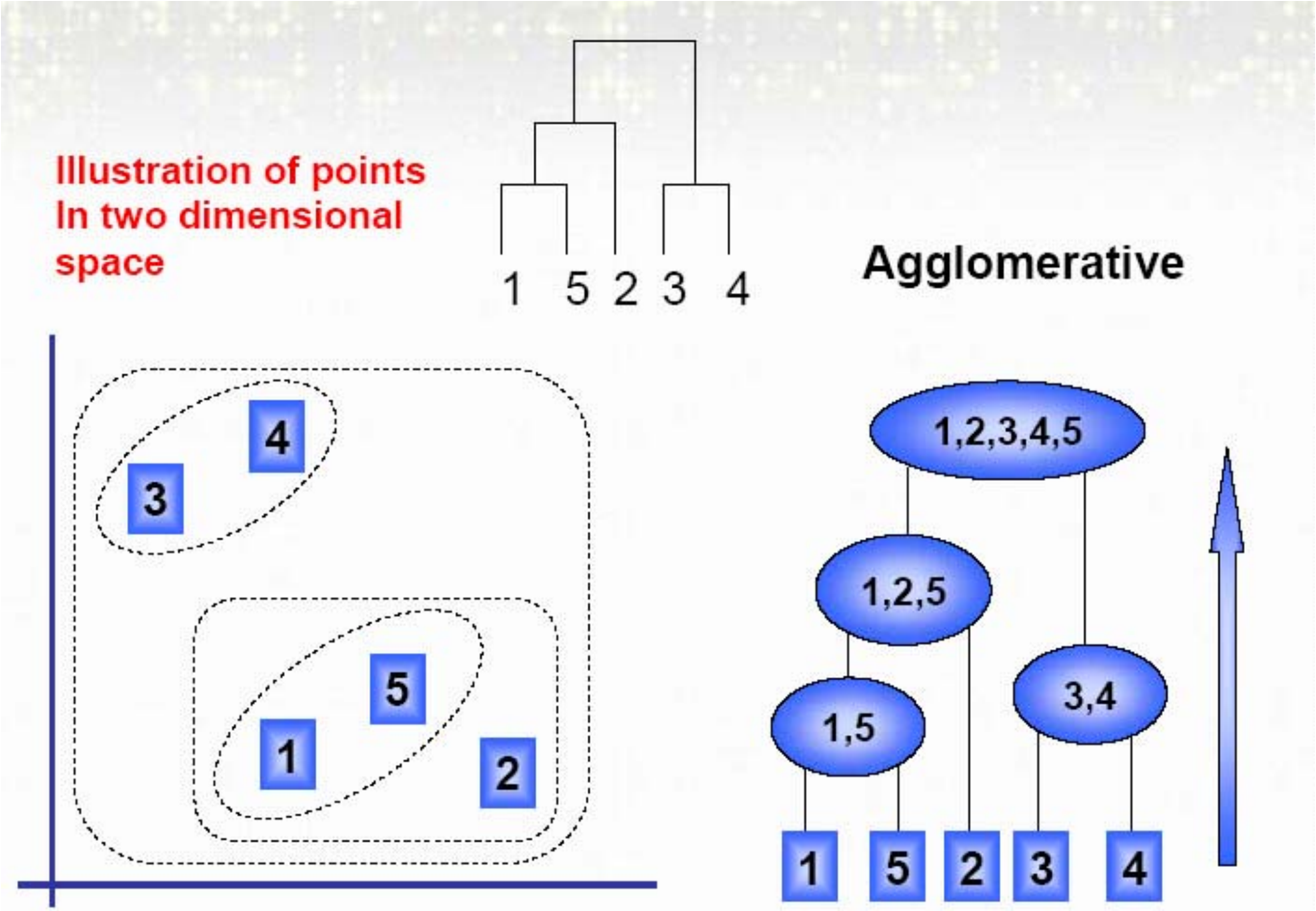
1. Background

- Approaches attempted in this talk:
  - A good simulation model that captures cluster structure in expression profiles.
    -- Underlying true cluster structure is known.
    -- Easy to test robustness of different methods.
  - Validation of clustering results through external biological knowledge.
    -- Evaluate biological relevance of clustering results
    -- Annotated genes are few. Often network-structured instead of cluster-structured.

2.1 Six gene clustering methods

- **Hierarchical clustering** (Eisen *et al.*, 1998)
- **K-means** (MacQueen, 1967; Hartigan, 1975)
- **K-memoids (PAM)**
- **Self-organizing maps (SOM)** (Kohonen, 1990)
- **Model-based clustering** (Fraley and Raftery, 2002; Medvedovic and Sivaganesan, 2002; McLachlan *et al.*, 2002)
- **Tight clustering** (Tseng and Wong, 2005)
- HOPACH (van der Laan and Pollard, 2003)
- Consensus clustering (Monti *et al.*, 2003)
- Fuzzy c-means (Dembele and Kastner, 2003)

2.1 Method 1: hierarchical clustering



Different choices of linkage: single, complete, average, centroid

2.1 Method 2: K-means

Algorithm:

Minimize the within-cluster dispersion to the cluster centers.

$$W(k) = \sum_{j=1}^{k} \sum_{x_i \in C_j} \left\| x_i - \bar{x}^{(j)} \right\|^2$$

$$\bar{x}^{(j)} = \underset{x_i \in C_j}{mean}(x_i), \text{centroids of cluster } j.$$

Note:

1. Points should be in Euclidean space.
2. Optimization performed by iterative relocation algorithms. Local minimum inevitable.
3. k has to be correctly estimated.

2.1 Method 3: K-memoids (PAM)

Algorithm:

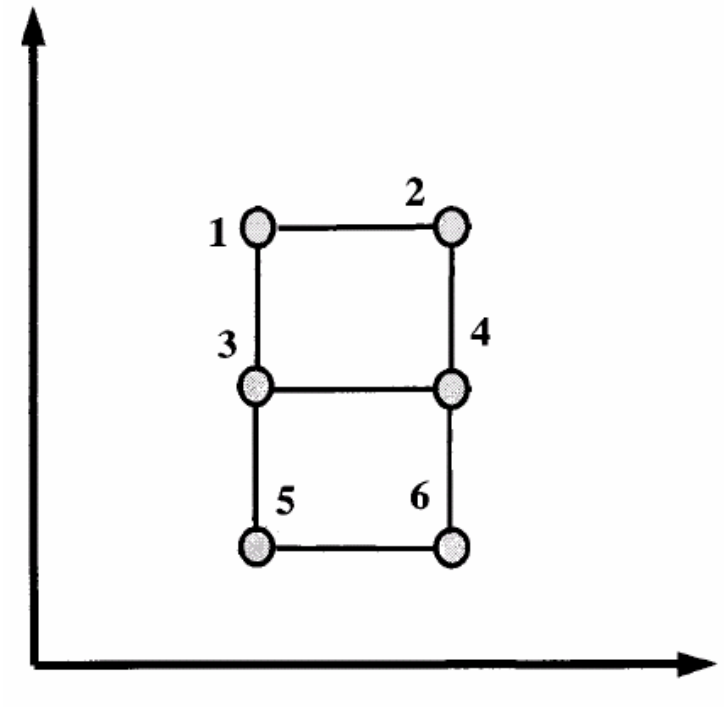
Minimize the within-cluster dispersion to the cluster centers.

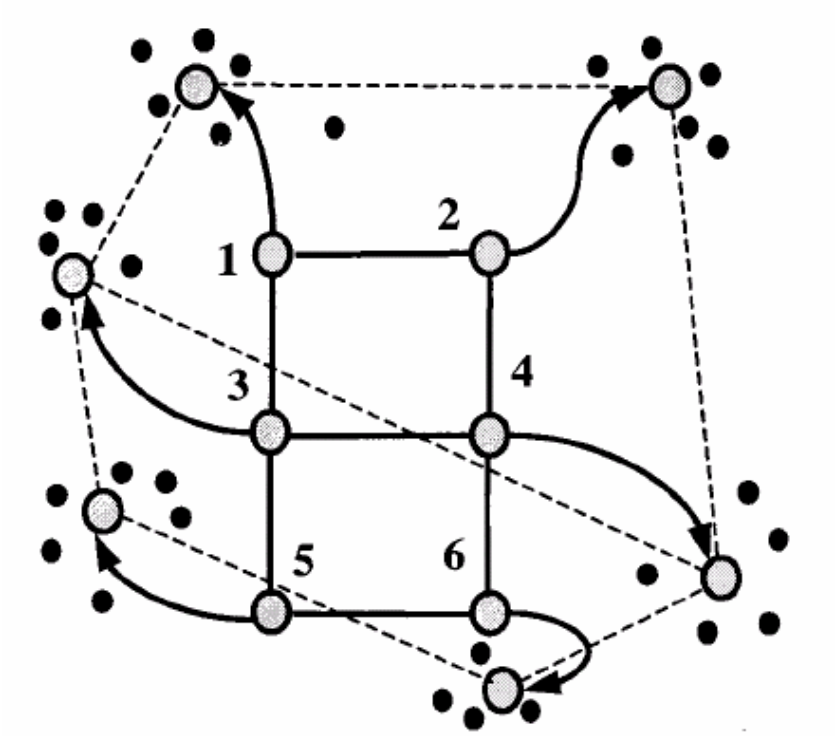
$$W(k) = \sum_{j=1}^{k} \sum_{x_i \in C_j} \left\| x_i - \bar{x}^{(j)} \right\|^2$$

$$\bar{x}^{(j)} = \underset{x_i \in C_j}{median}(x_i), \text{ medoids of cluster } j.$$

2.1 Method 4: SOM

$R^2 \longrightarrow R^p$





2.1 Method 5: model-based clustering

Mixture likelihood:

$$L(\theta, p \mid x) = \prod_{i=1}^{n} \left[\sum_{k=1}^{K} p_k f_k(x_i; \theta_k) \right]$$

Gaussian assumption:

$$f_k(\mathbf{x}_i \mid \mu_k, \Sigma_k) = \frac{\exp\left\{-\frac{1}{2}(\mathbf{x}_i - \mu_k)^T \Sigma_k^{-1} (\mathbf{x}_i - \mu_k)\right\}}{(2\pi)^{\frac{p}{2}} |\Sigma_k|^{\frac{1}{2}}}$$

BIC (model selection): $p(M \mid x) \propto p(x \mid M)$ if $p(M) \propto 1$

$$2 \log p(x|\mathcal{M}) + \text{const.} \approx 2 l_{\mathcal{M}}(x, \hat{\theta}) - m_{\mathcal{M}} log(n) \equiv \text{BIC}$$

$p(x|\mathcal{M})$ is the (integrated) likelihood of the data for the model $\mathcal{M}$.

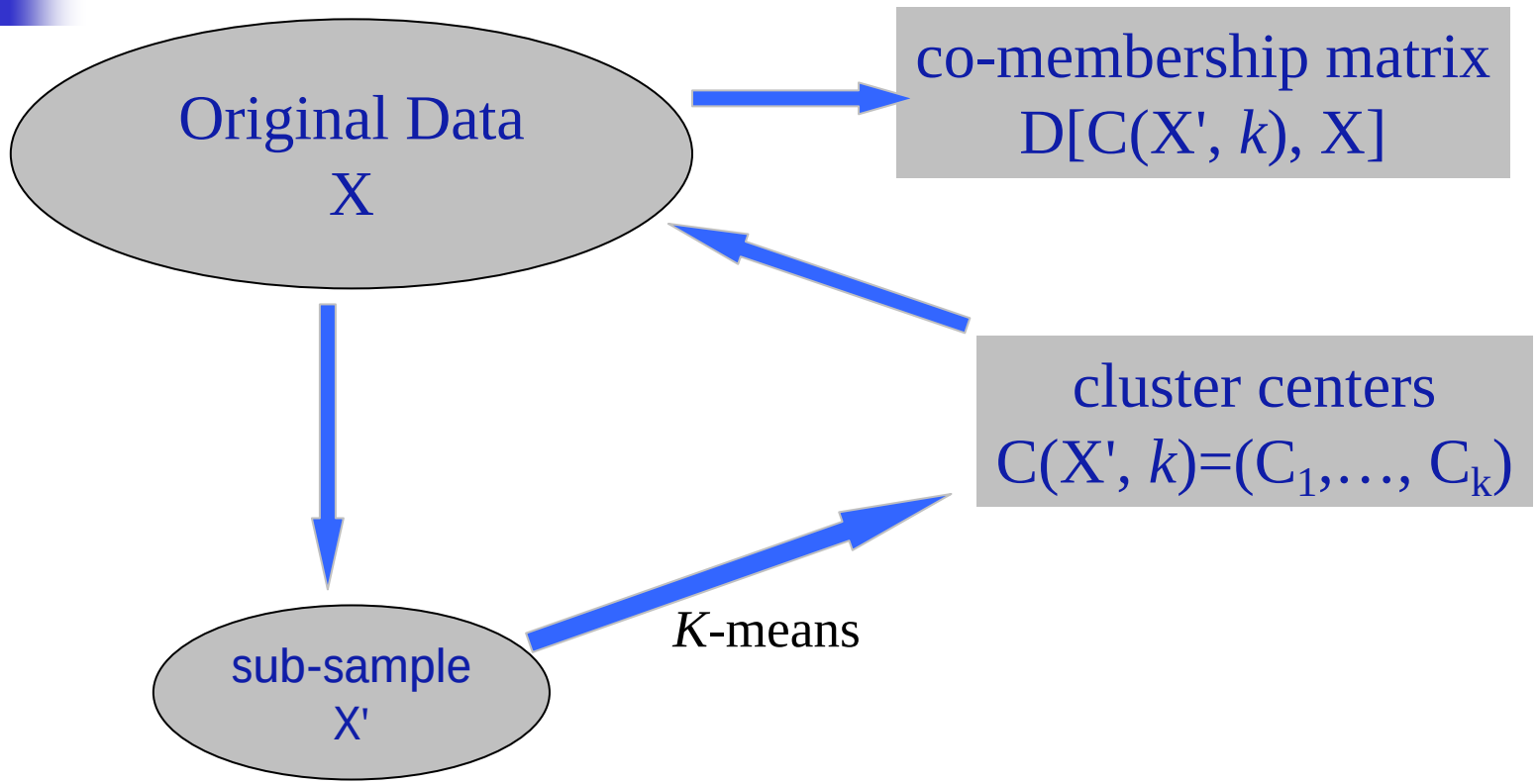
$l_{\mathcal{M}}(x, \hat{\theta})$ is the maximized mixture loglikelihood for the model

$m_{\mathcal{M}}$ is the number of independent parameters to be estimated in the model

Clustering with noise:

$$L(\theta, p \mid x) = \prod_{i=1}^{n} \left[\sum_{k=1}^{K} p_k f_k(x_i; \theta_k) + p_{K+1} u(x_i) \right]$$

2.1 Method 6: tight clustering



- Resampling approach to evaluate whether gene pairs are stably clustered together or by chance.
- Directly identify tight clusters and leave scattered (noise) genes.
- More robust to a wrongly estimated K.

2.1 Clustering with scattered genes

Cluster analysis with noise objects (scattered genes):

Data $X=\{x_i, i=1,\ldots, n\}$, each object $x_i \in R^p$.

Given a dissimilarity measure $d(x_i, x_j)$, assign the n objects into k disjoint clusters and a set of noise without being clustered; i.e. $C=\{C_1, \ldots, C_k\}$ and S where $X = (\bigcup_{j=1}^{k} C_j) \bigcup S$

Allowing clustering with scattered genes is found important in gene clustering.

K-means, *K*-memoids, SOM: No

model-based clustering and tight clustering: Yes

3.1 Similarity of two clusterings

Rand index: (Rand 1971)

C={(a,b,c), (d,e,f)} C'={(a,b), (c,d,e), (f)}

	ab	ac	ad	ae	af	bc	bd	be	bf	cd	ce	cf	de	df	ef	Total
together in both	*												*			2
separate in both			*	*	*		*	*	*			*				7
discordant		*				*				*	*			*	*	6

Rand index: r(C, C') =(2+7)/15=0.6 (percentage of concordance)

1. $1 \geq r(C, C') \geq 0$
2. Clustering methods can be evaluated by $r(C, C_{truth})$ if C_{truth} available.

3.1 Similarity of two clusterings

Adjusted Rand index: (Hubert and Arabie 1985)

	v_1	v_2	…	…	v_C	
u_1	n_{11}	n_{12}	…	…	n_{1C}	$n_{1\bullet}$
u_2	n_{21}	n_{22}	…	…	n_{2C}	$n_{2\bullet}$
u_R	n_{R1}	n_{R2}	…	…	n_{RC}	$n_{R\bullet}$
	$n_{\bullet 1}$	$n_{\bullet 2}$	…	…	$n_{\bullet C}$	n

$$\frac{\text{index - expected index}}{\text{maximum index - expected index}}$$

Adjusted Rand index =

$$\frac{\sum_{i=1}^{R}\sum_{j=1}^{C}\binom{n_{ij}}{2} - \sum_{i=1}^{R}\binom{n_{i\bullet}}{2}\sum_{j=1}^{C}\binom{n_{\bullet j}}{2} \Big/ \binom{n}{2}}{0.5\left[\sum_{i=1}^{R}\binom{n_{i\bullet}}{2} + \sum_{j=1}^{C}\binom{n_{\bullet j}}{2}\right] - 0.5\sum_{i=1}^{R}\binom{n_{i\bullet}}{2}\sum_{j=1}^{C}\binom{n_{\bullet j}}{2}}$$

The adjusted Rand index will take maximum value at 1 and constant expected value 0 (when two clusterings are totally independent)

3.1 Similarity of two clusterings

- What if we have two clusterings with possible sets of scattered genes?

	v_1	v_2	…	…	v_C	v_{noise}	
u_1	n_{11}	n_{12}	…	…	n_{1C}	$n_{1(C+1)}$	$n_{1\bullet}$
u_2	n_{21}	n_{22}	…	…	n_{2C}	$n_{2(C+1)}$	$n_{2\bullet}$
u_R	n_{R1}	n_{R2}	…	…	n_{RC}	$n_{R(C+1)}$	$n_{R\bullet}$
u_{noise}	$n_{(R+1)1}$	$n_{(R+1)2}$	…	…	$n_{(R+1)C}$	$n_{(R+1)(C+1)}$	$n_{(R+1)\bullet}$
	$n_{\bullet 1}$	$n_{\bullet 2}$	…	…	$n_{\bullet C}$	$n_{\bullet(C+1)}$	n

3.1 Similarity of two clusterings

Proposal 1: View the sets of scattered genes as a regular cluster.

$$Rand_1^*(R,C) = \frac{\sum_{i=1}^{R+1}\sum_{j=1}^{C+1}\binom{n_{ij}}{2} - \sum_{i=1}^{R+1}\binom{n_{i\bullet}}{2}\sum_{j=1}^{C+1}\binom{n_{\bullet j}}{2} \Big/ \binom{n}{2}}{0.5\left[\sum_{i=1}^{R+1}\binom{n_{i\bullet}}{2} + \sum_{j=1}^{C+1}\binom{n_{\bullet j}}{2}\right] - 0.5\sum_{i=1}^{R+1}\binom{n_{i\bullet}}{2}\sum_{j=1}^{C+1}\binom{n_{\bullet j}}{2}}$$

It is immediately seen that this index is in favor of methods generating clusters with scattered genes.

3.1 Similarity of two clusterings

Proposal 2: Completely ignore the sets of scattered genes.

$$Rand_2^*(R,C) = \frac{\sum_{i=1}^{R}\sum_{j=1}^{C}\binom{n_{ij}}{2} - \sum_{i=1}^{R}\binom{\tilde{n}_{i\bullet}}{2}\sum_{j=1}^{C}\binom{\tilde{n}_{\bullet j}}{2} \Big/ \binom{\tilde{n}}{2}}{0.5\left[\sum_{i=1}^{R}\binom{\tilde{n}_{i\bullet}}{2} + \sum_{j=1}^{C}\binom{\tilde{n}_{\bullet j}}{2}\right] - 0.5\sum_{i=1}^{R}\binom{\tilde{n}_{i\bullet}}{2}\sum_{j=1}^{C}\binom{\tilde{n}_{\bullet j}}{2}}$$

$$\tilde{n}_{i\bullet} = n_{i\bullet} - n_{i(C+1)} \quad \tilde{n}_{\bullet j} = n_{\bullet j} - n_{(R+1)j} \quad \tilde{n} = n - \sum_{\{i=R+1 \text{ or } j=C+1\}} n_{ij}$$

This index is in favor of methods generating clusters without scattered genes.

3.1 Similarity of two clusterings

A weighted proposal:

$$Rand^{*}(R,C) = \lambda.Rand_1^{*}(R,C) + (1-\lambda).Rand_2^{*}(R,C)$$

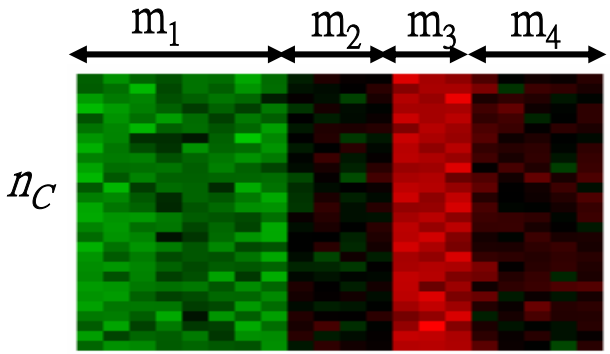
$$?? \quad \lambda = \left|u_{R+1}\right| / n = n_{(R+1)\bullet} / n \quad ??$$

$$\lambda = \left|u_{R+1} \cup v_{C+1}\right| / n = \left(n_{(R+1)\bullet} + n_{\bullet(C+1)} - n_{(R+1)(C+1)}\right) / n$$

3.2 Simulation model

Base model:

$$n_c \sim 4 \times Poisson(\lambda), \quad \sum m_i = 50, m_i > 2$$



$$\log(T_j^{(c)}) \sim N(\log(\mu_i^{(c)}), \sigma_S^2) \,\&\, \log(\mu_i^{(c)}) \sim N(\mu_c, \sigma_c^2)$$

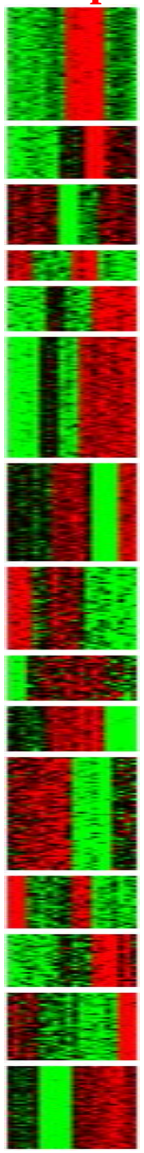
$$x_{lj} \sim N(\log(T_j^{(c)}), \sigma), l = 1,2,\ldots, n_c,\ j = 1,2,\ldots 50$$

$$\mu_c = 6, \sigma_c = 1, \sigma_S = 0.1, \sigma = 0.1 \,\&\, \lambda = 10$$

- **The simulated data well resembles structure of real data by visualization.**

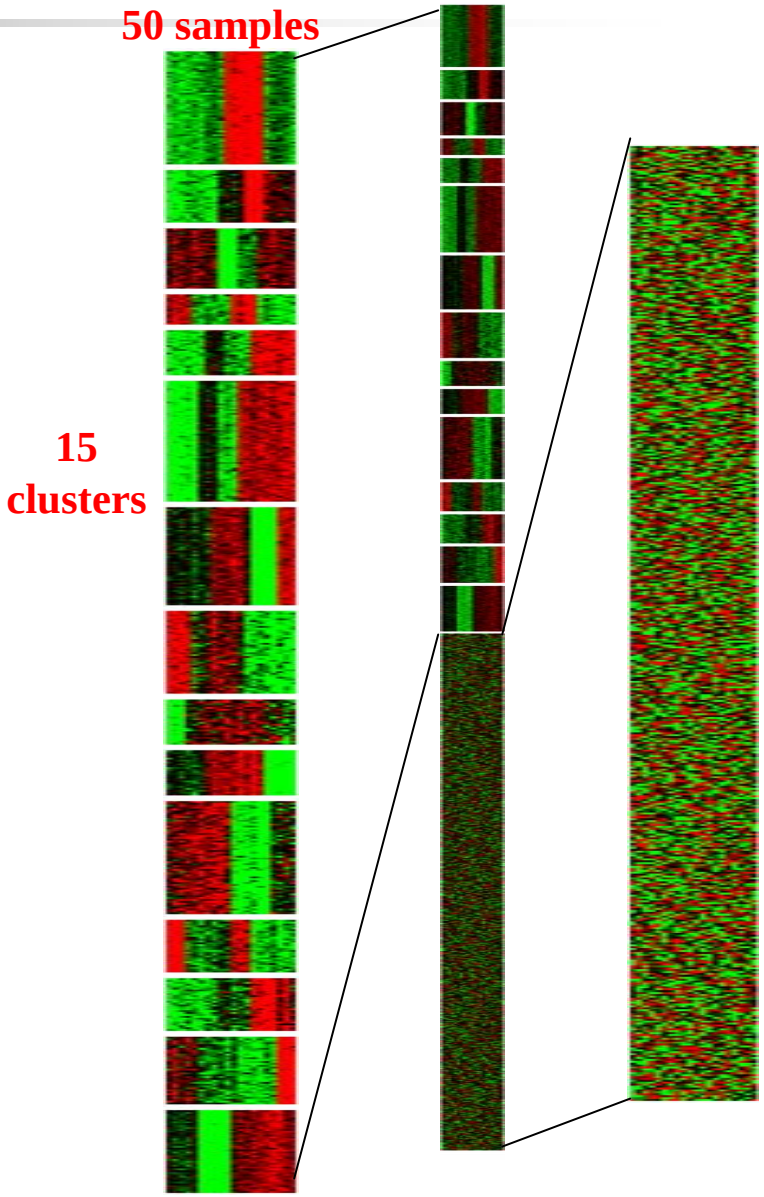
50 samples

15 clusters



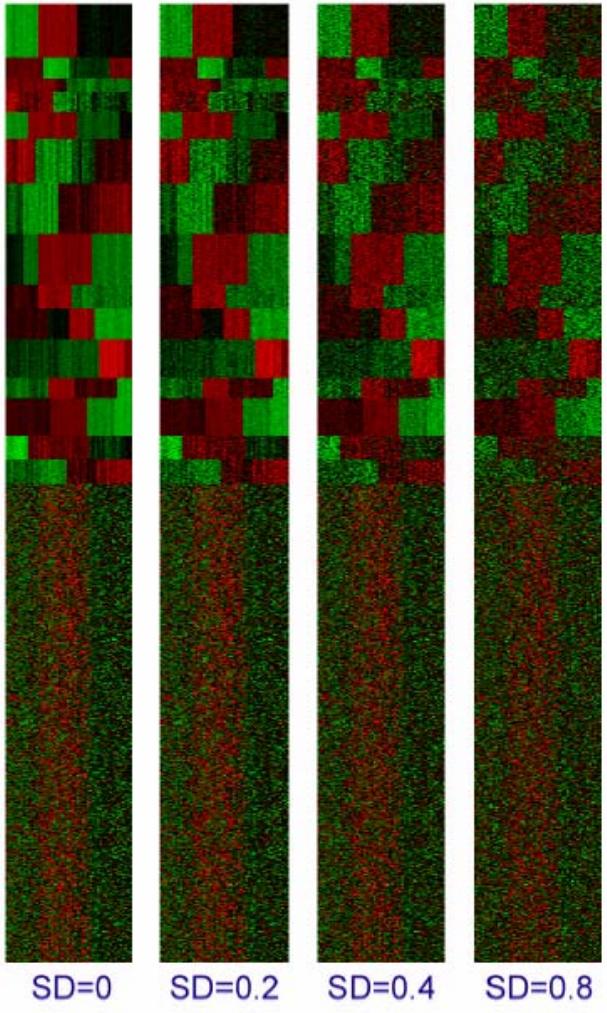
3.2 Simulation model

- Two types of perturbation
  - Type I: add randomly simulated scattered genes to the original data sets.



3.2 Simulation model

- Two types of perturbation
  - Type I: add randomly simulated scattered genes to the original data sets.
  - Type II: add random errors to the base model.
  - Type III: add random errors and then add scattered genes.



3.2 Simulation model

	0	0.05	0.1	0.2	0.4	0.8	1.2
0	×	×	×	×	×	×	×
5%	×						
10%	×						
20%	×						
60%	×						
100%	×	×	×	×	×	×	×
200%	×	×	×	×	×	×	×

Type II

Type I

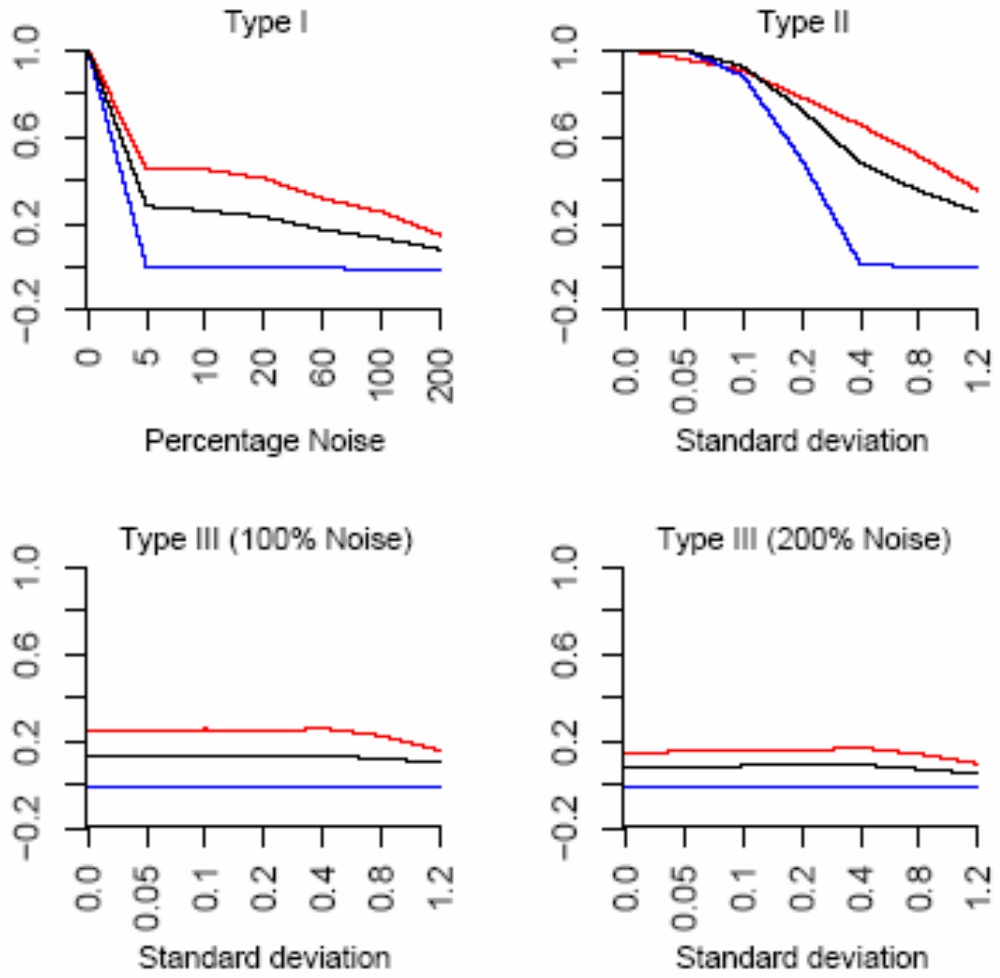
Type III

25 data sets each replicated 100 times constitute a total of 2500 simulated data sets for our analysis

3.3 Result: simulation study

Hierarchical clustering (selection of linkage)

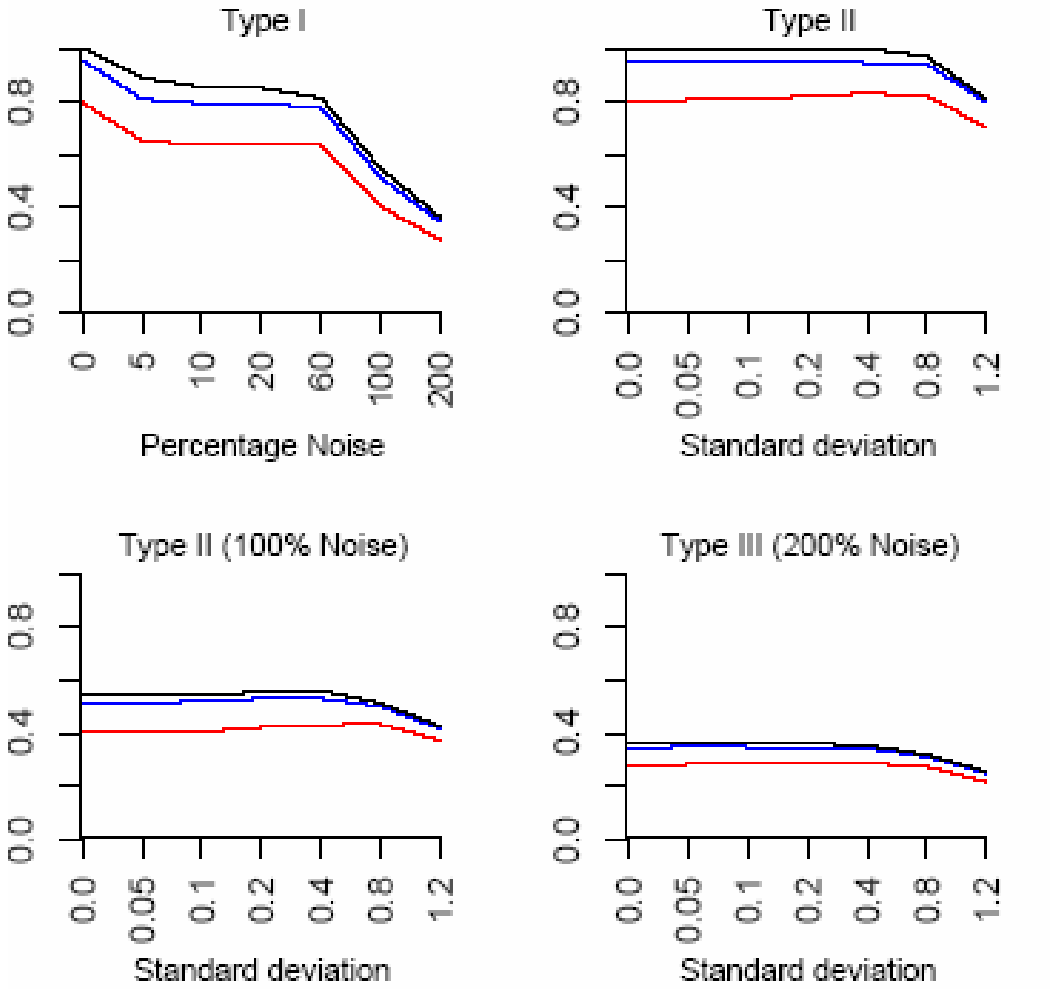
Weighted Rand Index for hierarchical clustering with single (blue), complete (red) and average (black) linkage on simulated data.



3.3 Result: simulation study

K-means (selection of # of random initial)

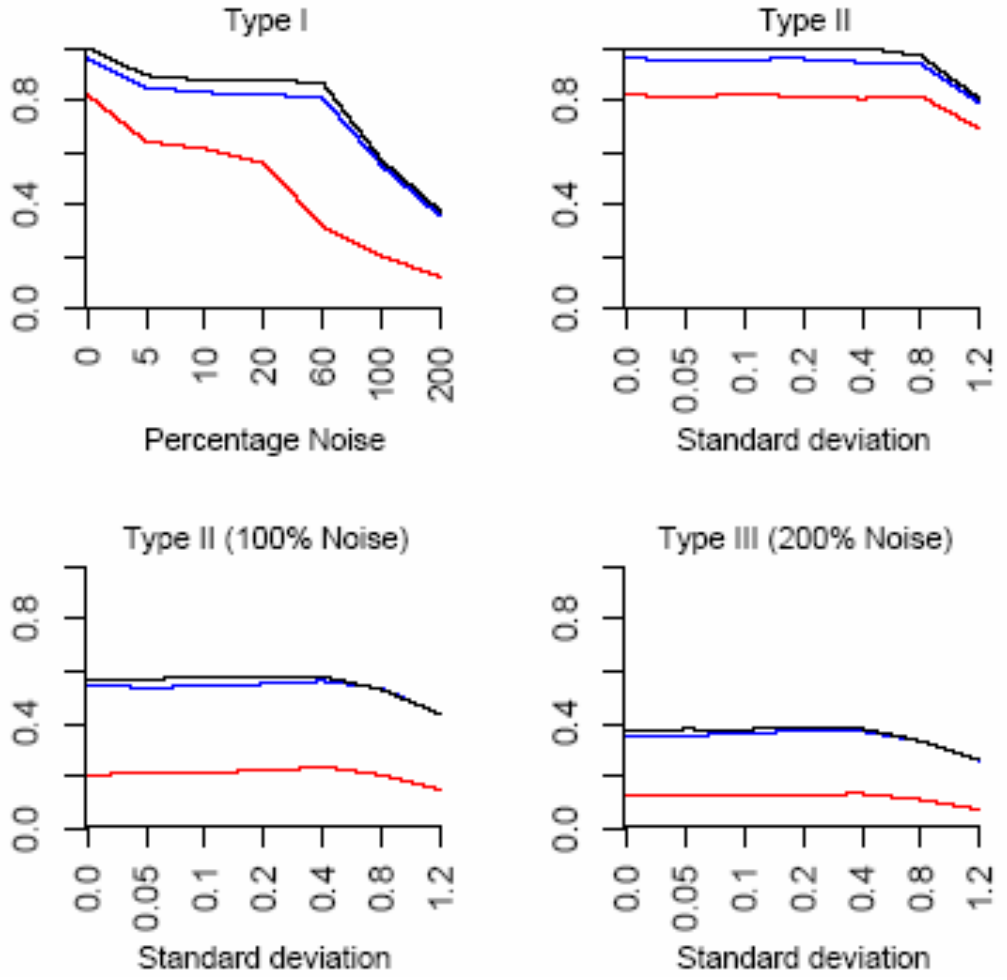
Weighted Rand index of K-means with 1 (red), 100 (blue) and 1000 (black) random initial values on simulated data.



3.3 Result: simulation study

PAM (selection of # of random initial)

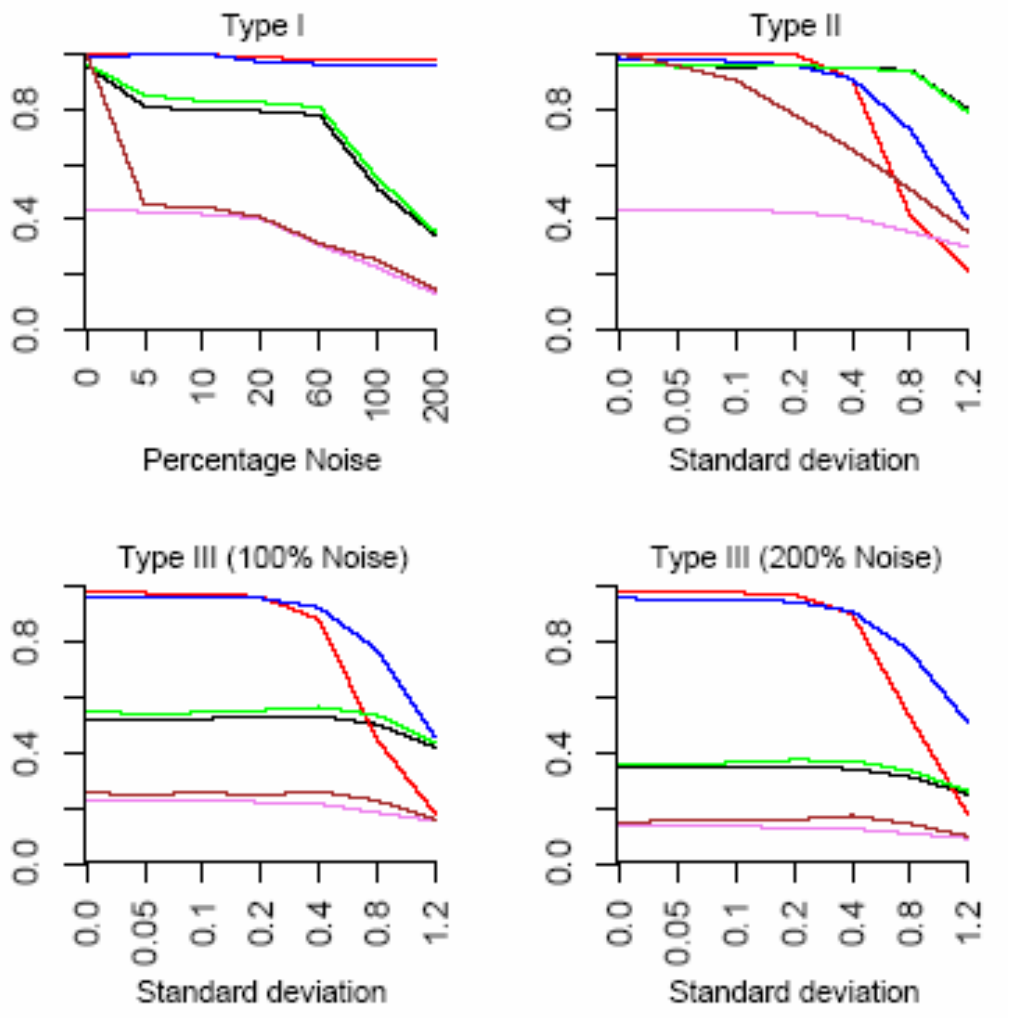
Weighted Rand index of PAM with 1 (red), 100 (blue) and 1000 (black) random initial values on simulated data.



3.3 Result: simulation study

Compare six methods

Weighted Rand index for SOM (violet), hierarchical (brown), K-means (black), PAM (green), model based clustering (red), tight clustering (blue) based on simulated data sets.

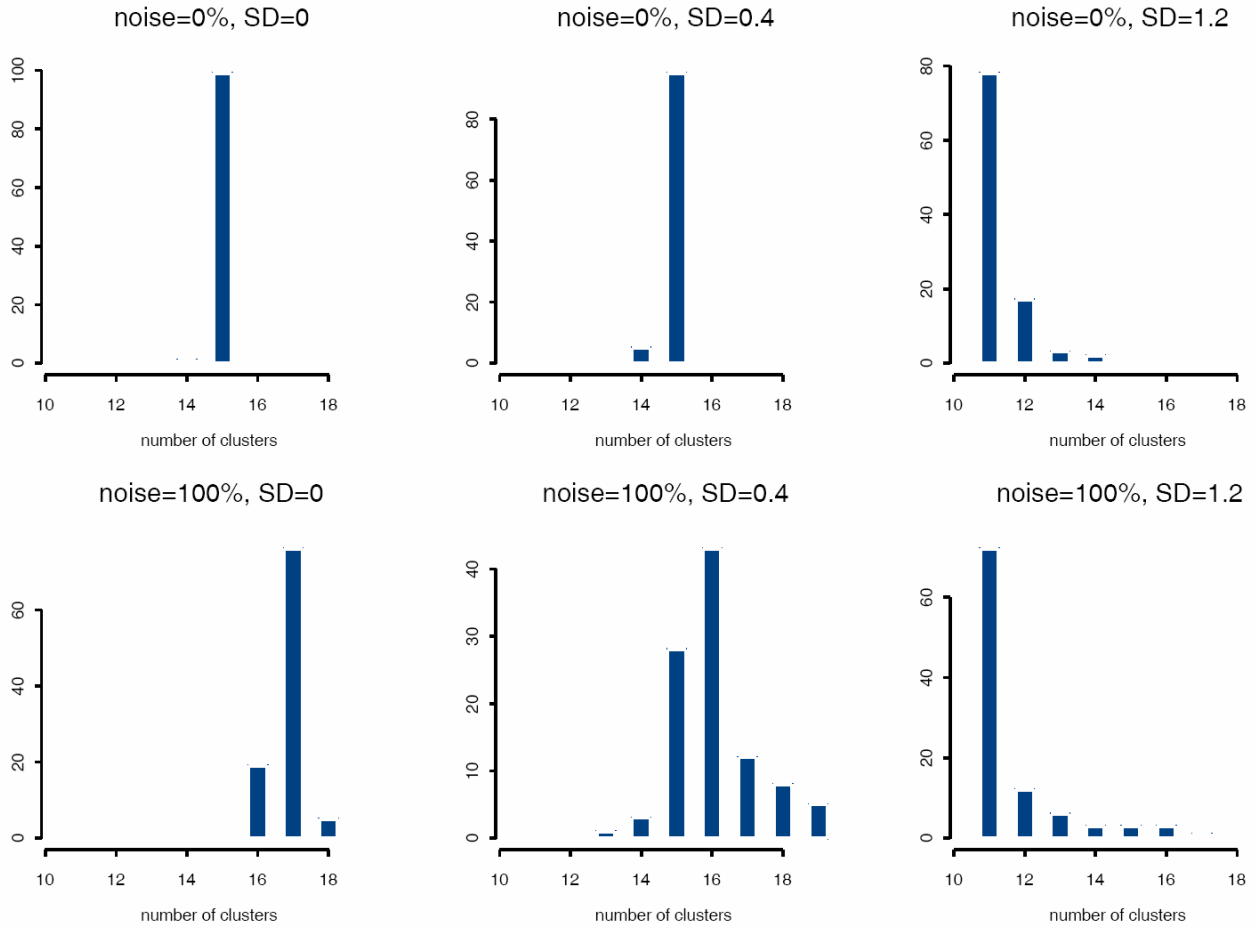


Conclusion:
SOM < Hierarchical clustering << *K*-means ≈ PAM << model-based clustering <≈ tight clustering

3.3 Result: simulation study

In the above study, the correct K=15 is given to all the methods.

Histogram of selected K by model-based clustering



3.3 Result: simulation study

Two reasons of incorrect estimation of K

- BIC is an approximate measure of the conditional density of data given model.

$$2\log p(x|\mathcal{M}) + \text{const.} \approx 2l_{\mathcal{M}}(x,\hat{\theta}) - m_{\mathcal{M}}log(n) \equiv \text{BIC}$$

- Local optimum often obtained in the maximum likelihood.

4.1 Prediction-accuracy plot

Yeast cell-cycle data

- Yeast cell cycle data (Spellman et al. 1998) contains 6179 genes. 1663 genes were retained for analysis. Dropped genes with more than 20% missing values and SD values at log-2 scale less than 0.4. Missing values for the genes retained for analysis were imputed by KNN algorithm.

- 104 genes that are cell cycle regulated in yeast have been identified by traditional methods (Paul T. Spellman, http://genome-7www.stanford.edu/cellcycle/data/rawdata/KnownGenes.doc of these 87 were found in our preprocessed data set.

  Among 1663 genes analyzed, 87 genes were annotated in six functional category.

4.1 Prediction-accuracy plot

87 cell-cycle related genes

Cell cycle period	Genes
M/G1 Boundary (F_1)	AGA1 ASH1 CDC46 CDC47 CDC6 CHS1 CLN3 CTS1 EGT2 FUS1 MFA2 PCL2 PCL9 RME1 SIC1 SST2 STE2 SWI4 TEC1
Late G1, SCB regulated (F_2)	CLN1 CLN2 CSD2 CHS3 FKS1 CWH53 GAS1 HO KAR4 KRE6 MNN1 PCL1 PSA1 SWE1 TIP1 VAN2 GOG5
Late G1, MCB regulated (F_3)	ASF1 ASF2 CDC21 CDC45 CDC8 CDC9 CLB5 CLB6 DBF4 DPB2 DPB3 GIC2 MCD1 MSH2 MSH6 NIK1 HSL1 PDS1 PMS1 POL1 POL12 POL2 POL3 CDC2 POL30 PRI1 PRI2 RAD17 RAD27 RAD51 RAD54 RFA1 RFA2 RFA3 RNR1 RNR3 SPC110 NUF1 SPC42 SPK1 SRS2 HPR5 UNG1
S-phase (F_4)	HHT1 HHT2 HHF1 HHF2 HTA1 HTA2 HTB1 HTB2
S/G2-phase (F_5)	CDC14 CIK1 CLB3 CLB4 CWP1 CWP2 KAR3 NUM1 TIR1
G2/M-phase (F_6)	ACE2 ASE1 CDC20 CDC5 CLB1 CLB2 DBF2 FAR1 KIN3 MOB1 YRO2 (MST1) MRH1 (MST2) SED1 SPO12 SWI5

4.1 Prediction-accuracy plot

Functional prediction from clustering:

A common use of clustering is functional prediction of novel genes. If a cluster has exceptionally high occurrences of a certain gene annotation *F*, all genes in this cluster are predicted to the functional category *F*.

Tabulation of clustering result and gene annotation

	F_1	F_2	F_3	F_4	F_5	F_6	F_{noise}
v_1	0(1)	0(1)	0(1)	0(1)	0(1)	0(1)	137
v_2	0(1)	0(1)	0(1)	0(1)	0(1)	0(1)	143
v_3	2(0.52)	**6(0.0005)**	**23(0)**	0(1)	0(1)	0(1)	136
v_4	0(1)	4(0.013)	5(0.074)	**8(1.7E-9)**	**5(6.5E-5)**	0(1)	115
v_5	0(1)	1(0.34291)	0(1)	0(1)	0(1)	0(1)	56
v_{noise}	15	1	0	0	2	15	989

4.1 Prediction-accuracy plot

Null (hypergeometric) distribution:

G genes in the genome (*G*=1663) are analyzed; Functional category "*F*" (Six functional categories). In a cluster of size C, *h* genes are found to be in a functional category "*F*" with *m* genes, then p-value (i.e. the probability of observing *h* or more annotated genes in the cluster is calculated as (Tavazoie et al. 1999):

$$P[X \geq h] = 1 - \sum_{i=0}^{h-1} \frac{\binom{C}{i}\binom{G-C}{m-i}}{\binom{G}{m}}$$

All genes in the cluster is annotated as category *F*, if the p-value is less than a threshold level δ.

4.1 Prediction-accuracy plot

Given K and p-value threshold δ,

Predictions made: $PM_K(\delta) = \sum_{i=1}^{K} \sum_{\{j: p_{ij} < \delta\}} n_{i\bullet}$

Verified Predictions: $VP_K(\delta) = \sum_{i=1}^{K} vp_{Ki} = \sum_{i=1}^{K} \sum_{\{j: p_{ij} < \delta\}} n_{ij}$

Accuracy: $A_K(\delta) = VP_K(\delta) / PM_K(\delta).$

$A_5(0.01) = (6+23+8+5)/(2\times167+2\times137) = 6.9\%$

	F_1	F_2	F_3	F_4	F_5	F_6	F_{noise}
v_1	0(1)	0(1)	0(1)	0(1)	0(1)	0(1)	137
v_2	0(1)	0(1)	0(1)	0(1)	0(1)	0(1)	143
v_3	2(0.52)	**6(0.0005)**	**23(0)**	0(1)	0(1)	0(1)	136
v_4	0(1)	4(0.013)	5(0.074)	**8(1.7E-9)**	**5(6.5E-5)**	0(1)	115
v_5	0(1)	1(0.34291)	0(1)	0(1)	0(1)	0(1)	56
v_{noise}	15	1	0	0	2	15	989

4.1 Prediction-accuracy plot

In practice, it's almost impossible to estimate the correct *K* in a microarray data. Instead of attempting to estimate *K*, we evaluate through a pooled analysis using *K*=5~30 and compute

$$A(\delta) = VP(\delta) / PM(\delta) = \sum_K VP_K(\delta) / \sum_K PM_K(\delta)$$

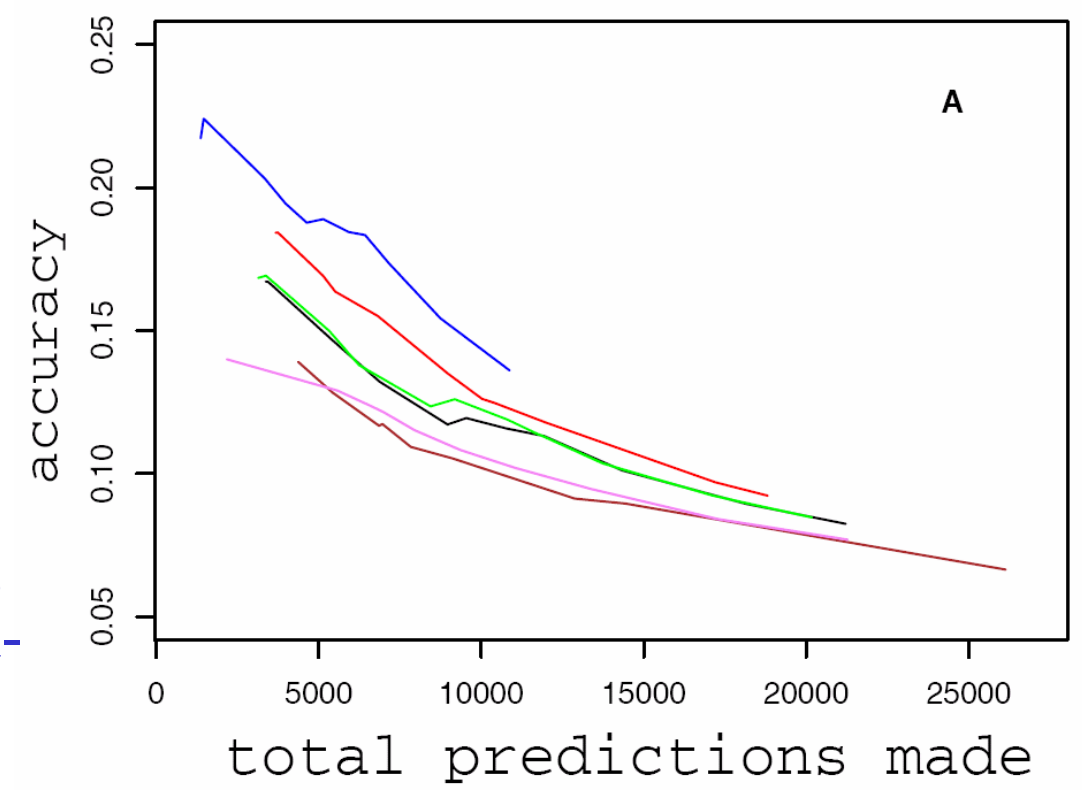
By varying δ, we can draw a prediction-accuracy plot (i.e. PM(δ) on x-axis and A(δ) on y-axis).

In general, a more stringent (small) δ makes fewer predictions (fewer PM) and better accuracy (higher A).

4.2 Result: real data 1

Compare six methods in yeast cell-cycle data

Prediction accuracy plot for SOM (violet), hierarchical (brown), K-means (black), PAM (green), model-based clustering (red), tight clustering (blue) based on yeast cell-cycle data.

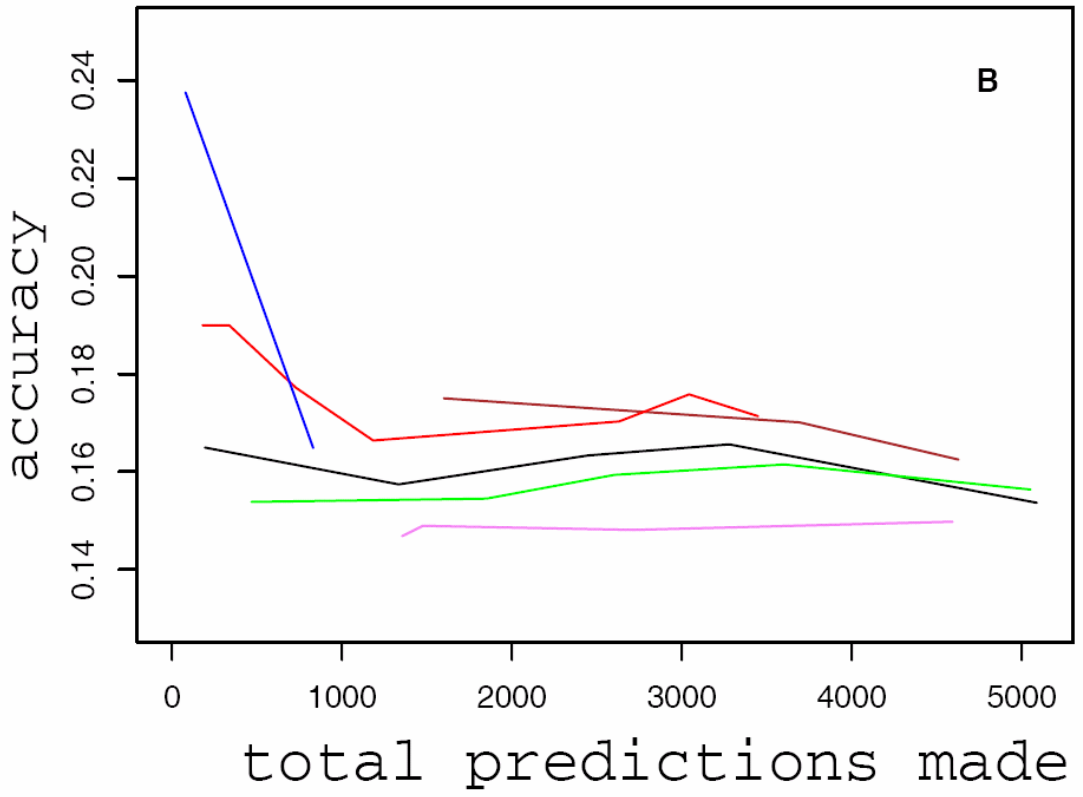


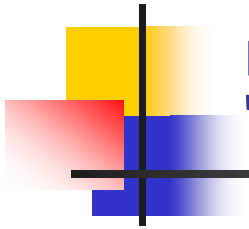
Conclusion:
SOM ≈ Hierarchical clustering << *K*-means ≈ PAM << model-based clustering << tight clustering

4.2 Result: real data 2

Compare six methods in yeast environmental change data (Causton *et al.*, 2001)

Prediction accuracy plot for SOM (violet), hierarchical (brown), K-means (black), PAM (green), model-based clustering (red), tight clustering (blue) based on yeast environmental change data.

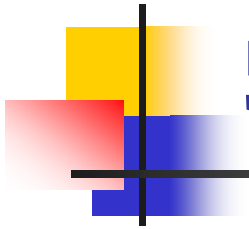




5. Conclusion

Methods:

- Weighted Rand index to measure similarity of two clusterings with scattered genes.
- Simulation model to generate cluster structure and examine robustness of methods.
- Prediction-accuracy plot to evaluate performance with external biology annotation (without estimate K).



5. Conclusion

- SOM <≈ Hierarchical clustering << *K*-means ≈ PAM << model-based clustering <≈ tight clustering
- SOM and hierarchical clustering
  - Better visualization; popular methods
  - Generate clusters less biologically relevant
- K-means and PAM
  - Adequate performance but do not allow scattered genes.

5. Conclusion

- Model-based clustering and tight clustering always among the best.
- Model selection (BIC) in model-based clustering can be difficult and harm the result in microarray data.
- The resampling procedure in tight clustering provides better robustness.

Acknowledgement

- Anbupalam Thalamuthu

  (visiting scholar, Dept. of Human Genetics, Pitt)
- Indranil Mukhopadhyay

  (visiting scholar, Dept. of Human Genetics, Pitt)
- Xiaojing Zheng

  (student, Dept. of Human Genetics, Pitt)

THANK YOU