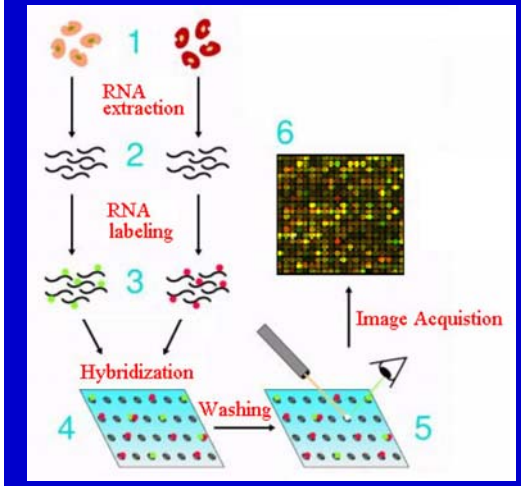
ICARCV 2004

A Comparative Review of Gene Clustering in Expression Profile

George C. Tseng
Dept. of Biostatistics/Human Genetics
University of Pittsburgh
12/07/2004

Microarray Introduction



Background:
- **Proteins play an important role in the body.**
- **DNA => mRNA => protein**

Goal:
- **Simultaneously measure mRNA abundance (gene expression) of thousands of genes.**

Data Matrix

Experimental design → Image analysis → Normalization

samples → (10~500 samples)

genes ↓ (10K~30K genes)

row.names	chromosome	sample1	sample2	sample3	sample4
		time0	time3	time5	time7
96669_at	8.00	0.00	-0.10	0.02	-0.27
100877_at	6.00	-0.22	0.59	0.20	0.46
93490_at	15.00	0.00	0.02	-0.02	-0.07
100978_at	8.00	-0.40	0.03	-0.18	-0.02
103516_at	15.00	NA	NA	NA	NA
160378_at	19.00	0.16	0.41	0.76	0.57
99670_at	19.00	-0.11	0.11	0.04	-0.12
98569_at	2.00	NA	0.11	0.35	0.37
93794_at	8.00	0.01	0.45	-0.02	0.12

Clustering

Goal: Given a dissimilarity measure, *n* points are grouped into *k* clusters based on their similarity.

General problems encountered:
1. Which dissimilarity measure to use?
2. How many # of clusters, *k*?
3. How to assign the points into clusters?

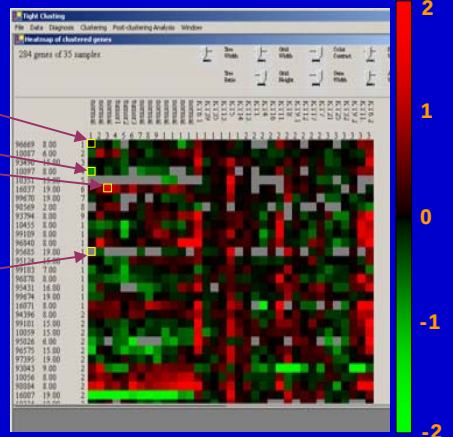
Clustering genes : similar gene expression pattern may imply co-regulation/network.
Clustering samples: identify potential sub-classes of disease

Heat Map (Data visualization)

Positive: gradient green
Negative: gradient red

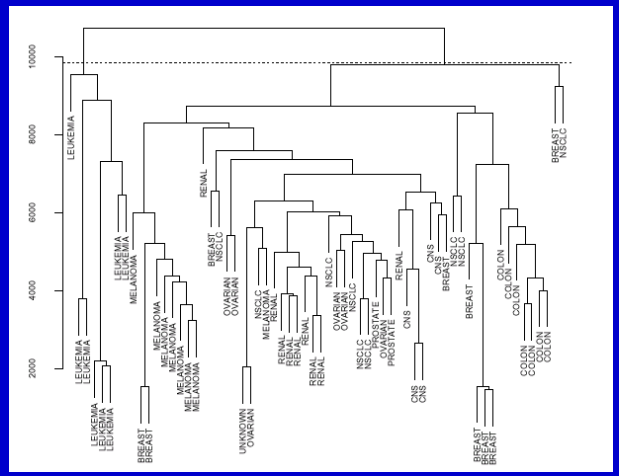
0: black
Missing: grey

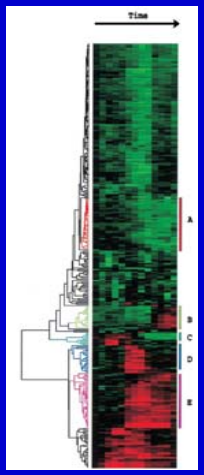
row.names	chromosome	sampl1	sample2	sample3	sample4	sample5
		time0	time3	time5	time7	time10
96669_at	8.00	0.00	0.10	0.02	-0.27	NA
100877_at	6.00	-0.22	0.59	0.29	0.46	-0.21
93490_at	15.00	0.00	0.02	-0.02	-0.07	-0.24
100978_at	8.00	-0.40	0.03	-0.18	-0.02	-0.26
103516_at	15.00	NA	NA	NA	NA	NA
160378_at	19.00	0.16	0.41	0.76	-0.57	0.24
99670_at	19.00	-0.11	0.11	0.04	-0.12	NA
98569_at	2.00	NA	0.11	0.35	0.37	NA
93794_at	8.00	0.01	0.45	-0.02	0.12	0.08
104554_at	8.00	-0.22	-0.35	-0.28	-0.12	-0.61
99109_at	8.00	-0.09	-0.14	0.21	0.06	-0.50
96840_at	8.00	0.07	0.84	0.26	0.15	0.05
95685_at	19.00	NA	-0.22	NA	NA	NA
95124_i_at	15.00	0.08	0.05	-0.30	-0.17	-0.31
99103_at	7.00	-0.33	-0.03	0.02	-0.14	-0.50
96878_at	8.00	-0.16	-0.15	-0.28	-0.33	-0.46
95431_at	16.00	-0.32	0.29	-0.03	-0.03	NA
99674_at	19.00	-0.21	-0.18	0.15	0.15	0.14
160714_at	8.00	0.39	0.28	0.30	0.38	0.03



Method 1: Hierarchical Clustering

- Iteratively merge the nearest node pairs.
- Bottom-up agglomerative merging method





Method 2: K-means & K-memoids

Procedures:

Step 1: estimate the number of clusters, k.

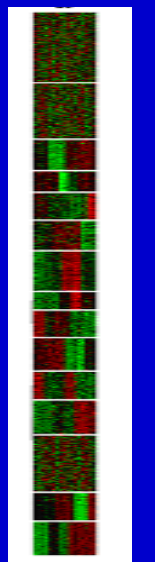
Step 2: minimize the within-cluster dispersion to the cluster centers.

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

$M(C_j)$: cluster mean (K-means)
cluster median (K-memoids)

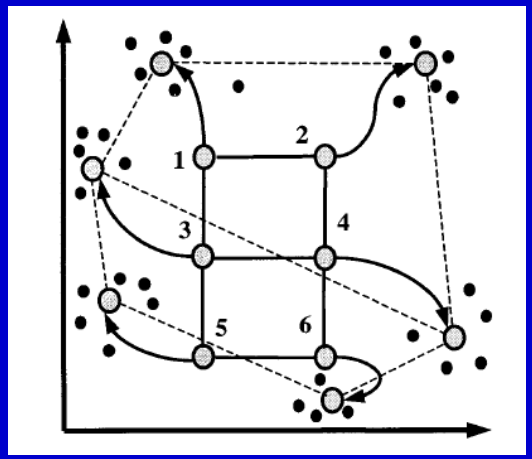
Note:

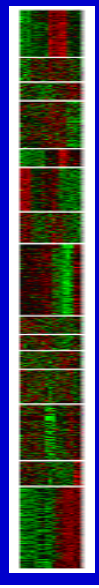
1. Points should be in Euclidean space.
2. Optimization performed by iterative relocation algorithms. Local minimum inevitable.
3. k has to be correctly estimated.



Self-organizing Maps (SOM)

- Similar to K-means.
- Cluster formation is restricted to 2D geometry.

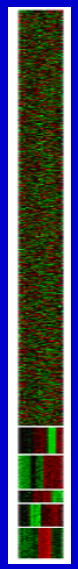




CLICK

- Graph-theoretical techniques to find tight “kernels”, sets of 5-7 tightly clustered genes.
- Several heuristic procedures are then used to expand the kernels into full clustering.

(R. Sharan & R. Shamir, 2003)



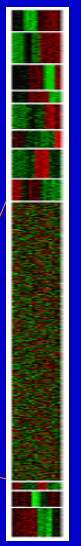
Model-based Clustering

Fraley and Raftery (1998) applied a Gaussian mixture model.

$$\mathcal{L}_M(\theta_1, \ldots, \theta_G; \tau_1, \ldots, \tau_G \mid \mathbf{x}) = \prod_{i=1}^{n} \sum_{k=1}^{G} \tau_k f_k(\mathbf{x}_i \mid \theta_k),$$

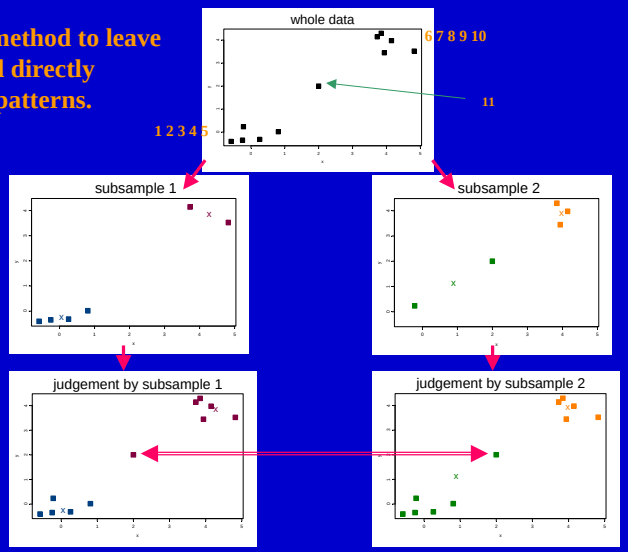
(1) Bayesian Information Criterion (BIC) for determining *k* and the complexity of the covariance matrix.
(2) Can also model scattered genes.

Scattered (noisy) genes



Tight Clustering

A re-sampling method to leave noisy points and directly recognize tight patterns.



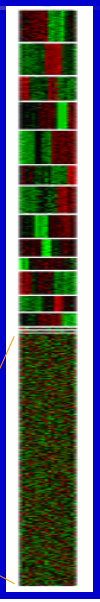
Tight Clustering

For sequence k>k0,

1. Identify sets of genes that are constantly clustered together under sub-sample clustering judgment. Consider the top q sets for each k.
2. Stop when for (k, k+1), two sets are nearly identical. Take the set corresponding to (k+1) as a tight cluster. Remove the identified cluster from the data.
3. Set k0=k0-1. Continue the procedure to find the next tight cluster.

(GC Tseng & W Wong, 2004)

Scattered (noisy) genes



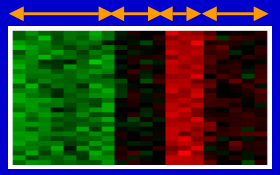
Comparison of methods

	Advantage	Disadvantage
Hierarchical clustering	▪Intuitive algorithm ▪Good interpretability ▪Do not need to estimate # of clusters	▪Very vulnerable to outliers ▪Tree not unique; gene closer not necessarily more similar ▪Hard to read when tree is big
K-means	▪Simplified Gaussian mixture model ▪Normally get nice clusters	▪ Local minimum ▪ Estimating # of clusters
SOM	▪Clusters has interpretation on 2D geometry (more interpretable)	▪The algorithm very heuristic ▪Solution sub-optimal due to 2D geometry restriction
Model-based clustering	▪Flexibility on cluster structure ▪Rigorous statistical inference	▪Model selection usually difficult ▪Local minimum problem
Tight clustering	▪Allow genes not being clustered; only produce tight clusters ▪Ease the problem of accurate estimation of # of clusters ▪Biologically more meaningful	▪Slower computation when data large

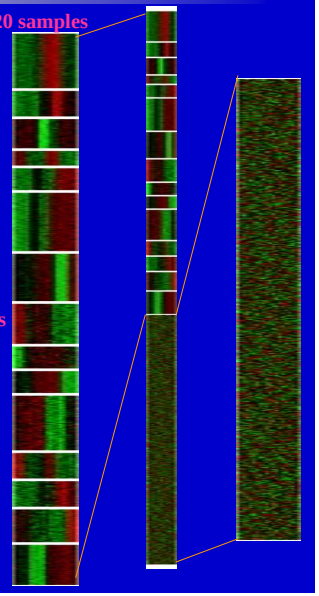
Comparison (simulation)

Simulation:

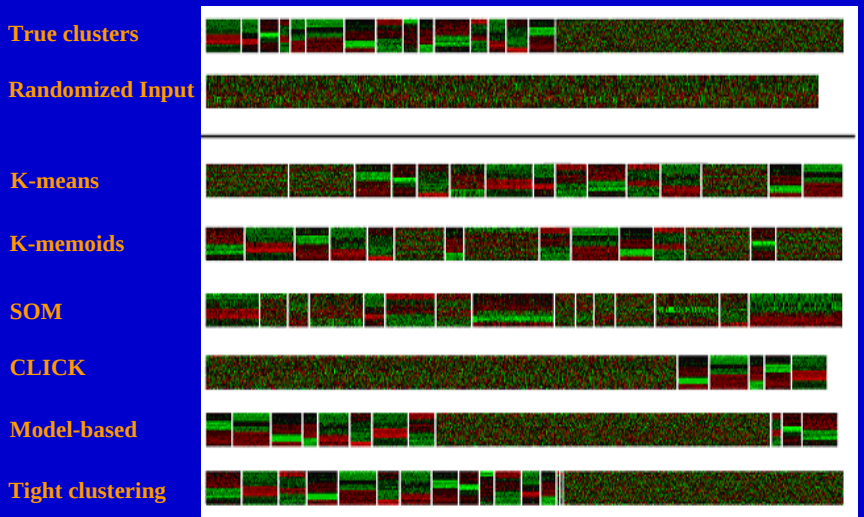
- 20 time-course samples for each gene.
- In each cluster, four groups of samples with similar intensity.



- Individual sample and gene variation are added.
- # of genes in each cluster ~ Poisson(10)
- Scattered (noise) genes are added.
- The simulated data well assembles real data by visualization.

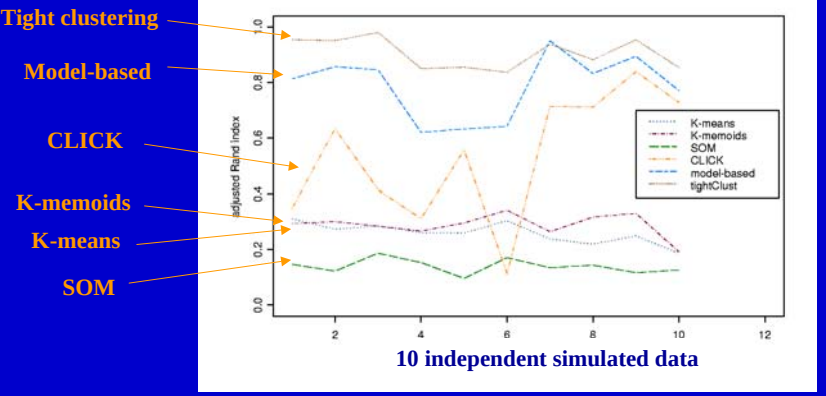


Comparison (simulation)



Comparison (simulation)

- Adjusted Rand index: a measure of similarity of two clustering;
- Compare each clustering result to the underlying true clusters. Obtain the adjusted Rand index (the higher the better).



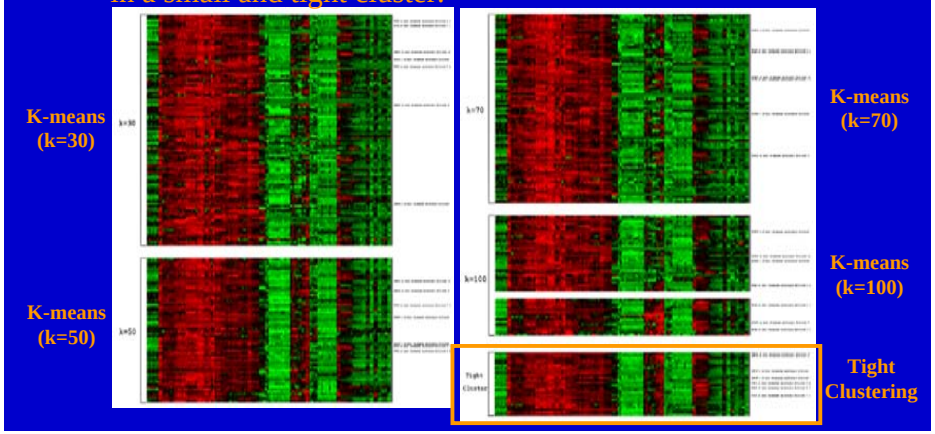
Comparison (stem cell data)

- Mouse embryonic experiment: oligonucleotide array (U74Av2 mouse array from Affymetrix) containing probe sets for about 10,000 mouse genes.

- Totally 126 samples. Half of them are from different stages of mouse embryonic development. The remaining half is a diverse collection of samples from various tissues, including several types of adult stem cells.

Comparison (stem cell data)

- 7 MCM genes (mini-chromosome maintenance genes) are known to tightly co-regulated.
- A good clustering method should assign them together in a small and tight cluster.



Comparison (stem cell data)

k	*K*-means	*K*-memoids	SOM	Model-based	Tight Clustering
30	7/96	7/65	7/203	EII 20 7/41	7/26
50	7/60	7/60	6/49		
70	7/77	7/133	5/74		
100	4/31, 3/15	7/38	5/100		

(# of MCM genes / #of genes in the cluster)

Acknowledgement

- Wing Wong's lab (Stanford): stem cell data
- Jean Yang (UCSF): discussion & reviewing
- Haiyan Huang (UC Berkeley): discussion & reviewing
- Xuewen Chen (U of Kansas): help proposing and organizing this bioinformatics session.