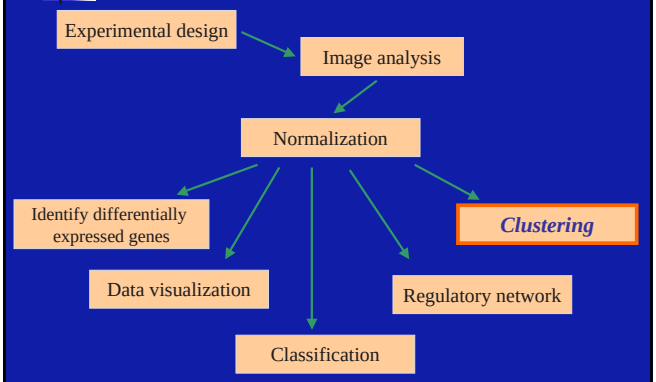
Tight Clustering: a method for extracting stable and tight patterns in expression profiles

George C. Tseng
Dept. of Biostatistics & Human Genetics
University of Pittsburgh

Statistical issues in microarray analysis

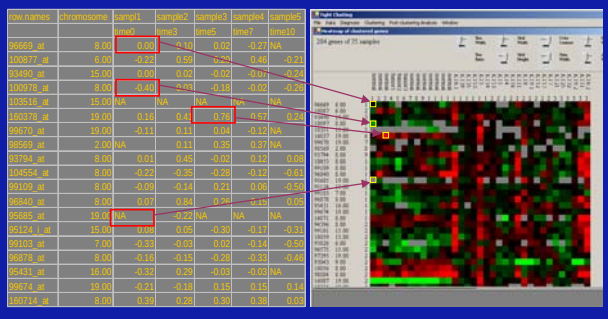


Data matrix

Data: $X=\{x_{ij}\}_{n\times d}$, an n (genes) × d (samples) matrix.

row.names	chromosome	sample1	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	0.20	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

Heatmap (data visualization)



Why clustering:

- **Cluster genes**: similar expression pattern implies co-regulation.

  Although many sophisticated methods for detecting regulatory interactions (e.g. Shortest-path and Liquid Association), cluster analysis remains a useful routine in array analysis.

  Subsequent analysis:

  • Identify novel genes participating in known cellular process

  • Enrichment of particular Gene Ontology (GO) terms in clusters

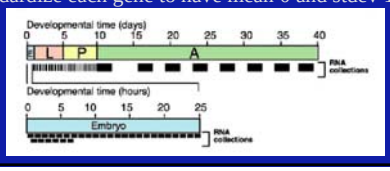
  • Motif finding in clusters

- **Cluster samples**: identify potential sub-classes of disease

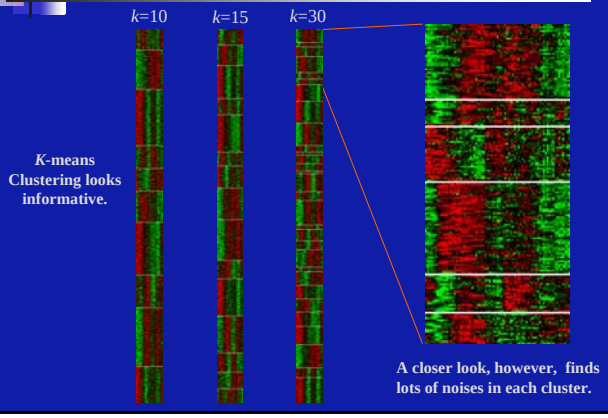
Clustering in microarray: an example

Gene expression during the life cycle of Drosophila melanogaster. (2002) *Science* 297:2270-2275

- 4028 genes monitored. Reference sample is pooled from all samples.
- 66 sequential time points spanning embryonic (E), larval (L), pupal (P) and adult (A) periods.
- Filter genes without significant pattern (1100 genes) and standardize each gene to have mean 0 and stdev 1.

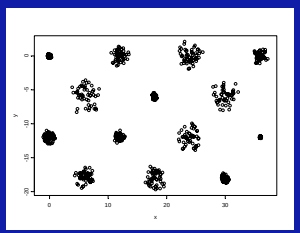


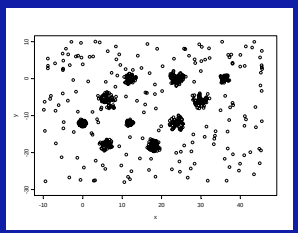
Example: Data from life cycle of Drosophila melanogaster. (2002) *Science* 297:2270-2275



Main challenges for clustering in microarray

Challenge 1: Lots of scattered genes. i.e. genes not belonging to any tight cluster of biological function.





Main challenges for clustering in microarray

Challenge 2: Microarray is an exploratory tool to guide further biological experiments

Hypothesis driven: hypothesis => experimental data.

Data driven: high-throughput experiment => data mining => hypothesis => further validation experiment

⇒ Important to provide the most informative clusters instead of lots of loose clusters (reduce false positives).

Current Methods

Dimension reduction and data visualization:

- Principle Component Analysis (PCA) (Alter 2000)
- Multi-Dimensional Scaling (MDS)

Clustering methods

- Hierarchical Clustering (Eisen 1998)
- *K*-means (Hartigan 1975)
- *K*-memoids
- Self-Organizing Map (SOM) (Tamayo 1999)
- CLICK (Ron Shamir 2001)
- Model-based approach (Fraley and Raftery 1998)

Model-based approach

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

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

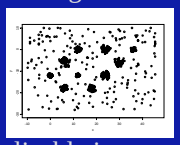
(1)EM algorithm to maximize the classification likelihood.

(2) Bayesian Information Criterion (BIC) for determining *k* and the complexity of the covariance matrix.

Model-based approach

Advantage:

- A sound probabilistic model for inference: model selection and estimation
- Can easily extend to model scattered genes



Problems:

- Local minimum
- Model selection is usually inapplicable in array data; BIC is approximate

K-means clustering

Procedures:

Step 1: estimate the number of clusters, *k*.

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

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

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.

K-means clustering

K-means is a special case of model-based approach.

Problems:

- Local minimum
- Does not allow scattered genes
- Estimation of number of clusters *k*

Estimate the number of clusters *k*:

Milligan & Cooper(1985) compared 30 published rules.

1. Calinski & Harabasz (1974)

$$\max CH(k) = \frac{B(k)/(k-1)}{W(k)/(n-k)}$$

2. Hartigan (1975)

$$H(k) = [\frac{W(k)}{W(k+1)} - 1](n-k-1)$$, Stop when H(k)<10

3. Tibshirani, Walther & Hastie (2000)

$$\max Gap_n(k) = E_n^*(\log(W(k))) - \log(W(k))$$

4. Tibshirani et al(2001), Dudoit & Fridlyand(2002) Prediction-based resampling approach.

Hierarchical clustering

Hierarchical clustering

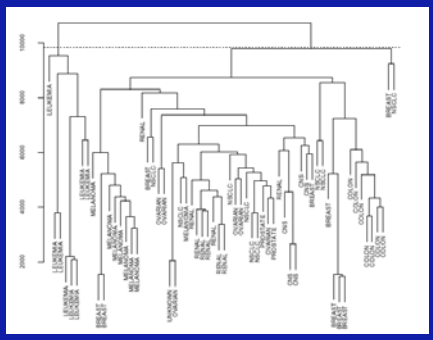
Iteratively agglomerate nearest nodes to form bottom-up tree.

Single Linkage: shortest distance between points in the two nodes.

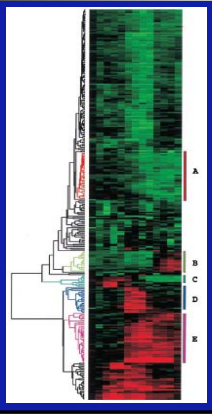
Complete Linkage: largest distance between points in the two nodes.

Note: Clusters can be obtained by cutting the hierarchical tree.

Hierarchical clustering



Example of hierarchical clustering



Eisen et al 1998

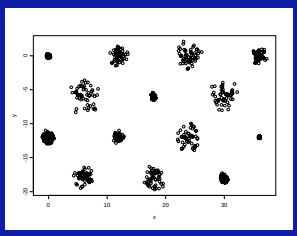
Other Methods

Current methods aim to find tight clusters:

1. CLICK: graph-theoretical techniques to find tight "kernels". Several heuristic procedures then used to expand the kernels into full clustering.
2. Committee algorithm: similar idea to find tight "committees" and then expand to full clustering.

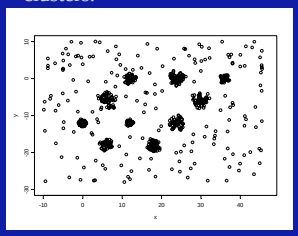
Traditional:

- Estimate the number of clusters, *k*. (except for hierarchical clustering)
- Perform clustering through assigning all genes into clusters.

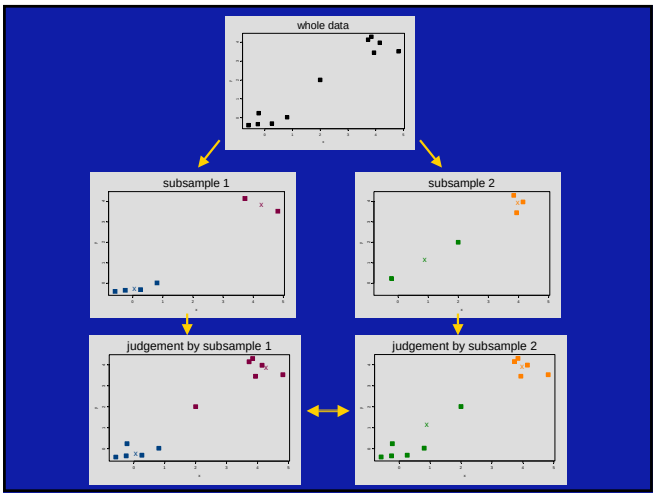


Tight Clustering:

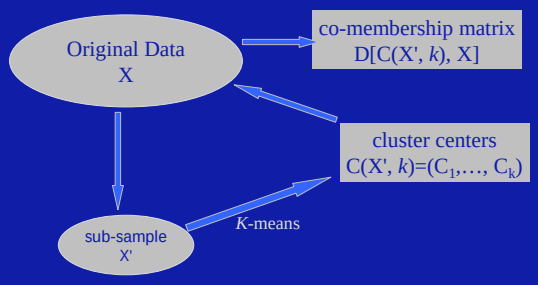
- Directly identify informative, tight and stable clusters with reasonable size, say, 20~60 genes.
- Need not estimate *k* !!
- Need not assign all genes into clusters.



Tight Clustering



Algorithm(Tight Clustering)



Algorithm(Tight Clustering)

- $X=\{x_{ij}\}_{n\times d}$: data to be clustered.
- $X'=\{x'_{ij}\}_{n/2\times d}$: random sub-sample
- $C(X', k)=(C_1, C_2,\ldots, C_k)$: the cluster centers obtained from clustering X' into *k* clusters.
- $D[C(X', k), X]$: an $n\times n$ matrix denoting co-membership relations of X classified by $C(X', k)$. (Tibshirani 2001)

  $D[C(X', k), X]_{ij}$ =1 if *i* and *j* in the same cluster.
  =0 o.w.

- $s(V_i, V_j)=\frac{|V_i \cap V_j|}{|V_i \cup V_j|}$: a measure of similarity of two sets of genes

Algorithm(Tight Clustering)

Algorithm 1 (when fixing k):

1. Fix k. Random sub-sampling $X^{(1)}, \ldots, X^{(B)}$. Define the average co-membership matrix to be
$\overline{D} = \text{mean}\left(D[C(X^{(1)}, k), X], \ldots, D[C(X^{(B)}, k), X]\right)$.

Note:

a. $\overline{D}_{ij} = 1 \Rightarrow$ i and j always clustered together in each sub-sampling judgment.

b. $\overline{D}_{ij} = 0 \Rightarrow$ i and j never clustered together in each sub-sampling judgment.

c. $\overline{D}_{ii} = 1 \quad \forall\, i$

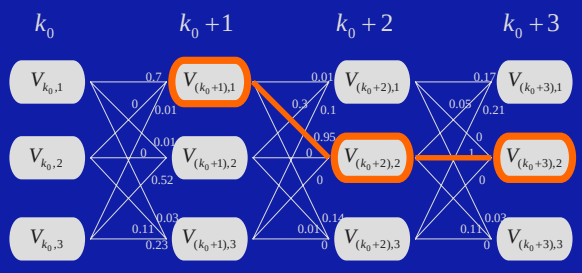
Algorithm(Tight Clustering)

Algorithm 1 (when fixing k): (cont'd)

2. Search for a large set of points $V = \{v_1, \ldots, v_m\} \subseteq \{1, \ldots, n\}$ such that $\overline{D}_{v_i v_j} \geq 1 - \alpha \quad \forall\, i, j$ α close to 0. Sets with this property are candidates of tight clusters. Order sets with this property by their size to obtain $V_{k1}, V_{k2}, \ldots$

Algorithm(Tight Clustering)

Tight Clustering Algorithm:



Algorithm(Tight Clustering)

Tight Clustering Algorithm:

1. Start with a suitable k_0. Search for consecutive k's and choose the top 3 clusters for each k.

$$\{V_{k_0 1}, V_{k_0 2}, V_{k_0 3}\}, \{V_{(k_0+1)1}, V_{(k_0+1)2}, V_{(k_0+1)3}\}, \ldots$$

2. Stop when

$$s(V_{k'l}, V_{(k'+1)m}) \geq \beta, \quad s(V_{(k'+1)m}, V_{(k'+2)n}) \geq \beta$$

$$k' \geq k_0, \quad l, m, n \in \{1,2,3\}, \ \beta \text{ close to } 1$$

Select $V_{(k'+1)m}$ to be the tightest cluster.

Algorithm(Tight Clustering)

Tight Clustering Algorithm: (cont'd)

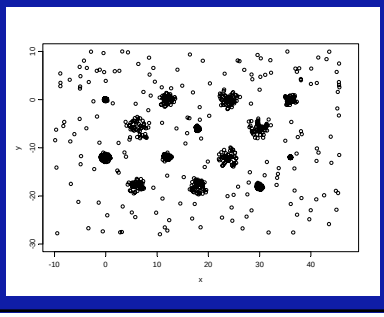
3. Identify the tightest cluster and remove it from the whole data.
4. Decrease k_0 by 1. Repeat 1.~3. to identify the next tight cluster.

Remark: α, β and k_0 determines the tightness and size of resulting clusters.

Simulation

A simple simulation on 2-D:

14 clusters normally distributed (50 points each) plus 175 sporadic points. Stdev=0.1, 0.2, …, 1.4.



Simulation

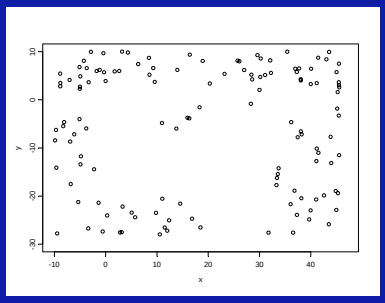
Tight clustering on simulated data:

$\alpha = 0, \beta = 0.7, B = 10,$ $k_0 = 10, 20, 25$ and 40

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	remain
truth	50	50	50	50	50	50	50	50	50	50	50	50	50	50		175
alpha 0 beta 0.7																
k_0=10	58	59	59	78	72	60										489
k_0=20	59	56	55	53	57	53	53	52	52	52	56	51	51	51	12	112
k_0=25	55	56	53	56	53	53	52	55	51	51	51	50	50	50	9	130
k_0=40	52	51	51	52	51	51	51	50	26	25	22	50	18	17	30	278

Simulation

$k_0 = 25, \alpha = 0, \beta = 0.7, B = 10$



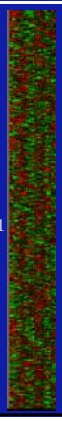
Example 1: Data from life cycle of Drosophila melanogaster. (2002) *Science* 297:2270-2275

Tight Clustering $\alpha = 0.1, \beta = 0.6, k_0 = 15$

69 75 28 30 34 49

33 28 15 20 58

661

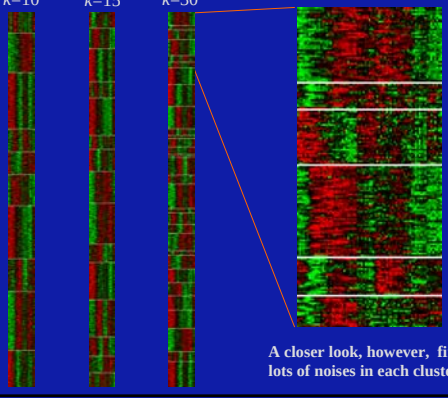


11 clusters and 661 remaining scattered genes

Example 1: Data from life cycle of Drosophila melanogaster. (2002) *Science* 297:2270-2275

k=10 *k*=15 *k*=30

***K*-means Clustering looks informative.**



A closer look, however, finds lots of noises in each cluster.

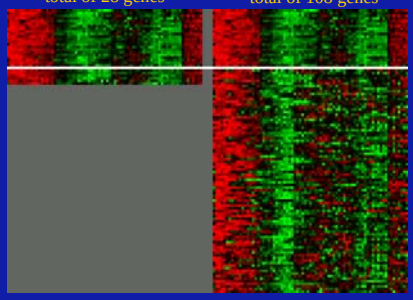
Example 1: Data from life cycle of Drosophila melanogaster. (2002) *Science* 297:2270-2275

Comparison: a corresponding cluster of K-means & Tight Clustering

Tight Clustering *K*-means clustering

total of 28 genes total of 108 genes

22 common genes



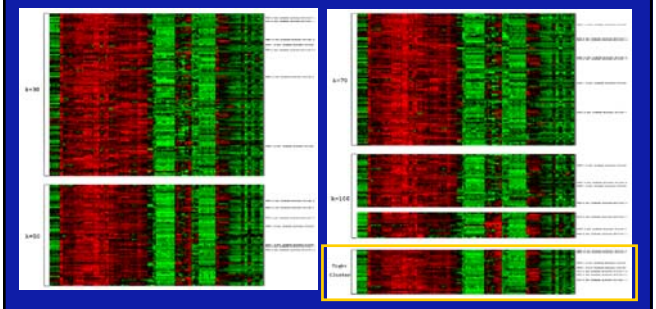
Mean sq. distance: 15.49 39.80

Example 2: Mouse embryonic experiment

- 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.

Example 2: Mouse embryonic experiment

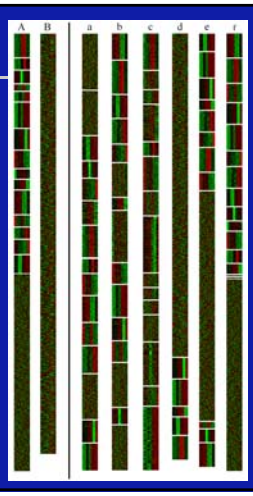
Comparison of various K-means and tight clustering:



Example 3: Simulated data

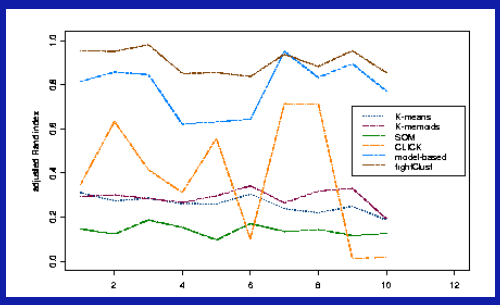
A. simulated gene expression of 15 clusters and 500 scattered genes.
B. Randomly permuted from A.

a. K-means
b. K-memoid
c. SOM
d. CLICK
e. Model-based clustering
f. Tight clustering



Example 3: Simulated data

Adjusted Rand index is a measure to compare similarity of two clustering results. We compare clustering results from each method to the underlying truth.

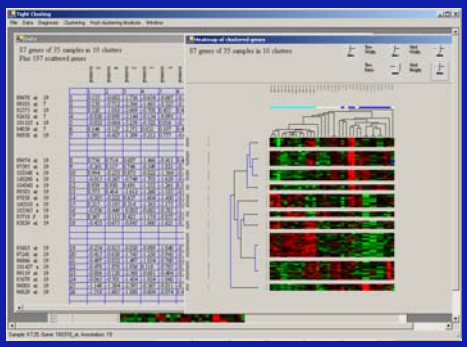


Ongoing developments

- Theoretical foundation for re-sampling approach.
- Multi-resolution tight clustering.
- Extend the idea to bi-clustering.
- Incorporating multiple tight clustering results.
- Other general and fundamental problems in clustering.

tightClust: a software for Tight Clustering

http://www.pitt.edu/~ctseng/tightClust.html



Acknowledgement:

Harvard:
Wing H. Wong (Department of Statistics)

Inputs from:
Chen Li (Department of Biostatistics)
Ryung Kim
Richard Zhong