CS681: Advanced Topics in Computational Biology

Week 2, Lecture 1

Can Alkan

EA224

calkan@cs.bilkent.edu.tr

http://www.cs.bilkent.edu.tr/~calkan/teaching/cs681/

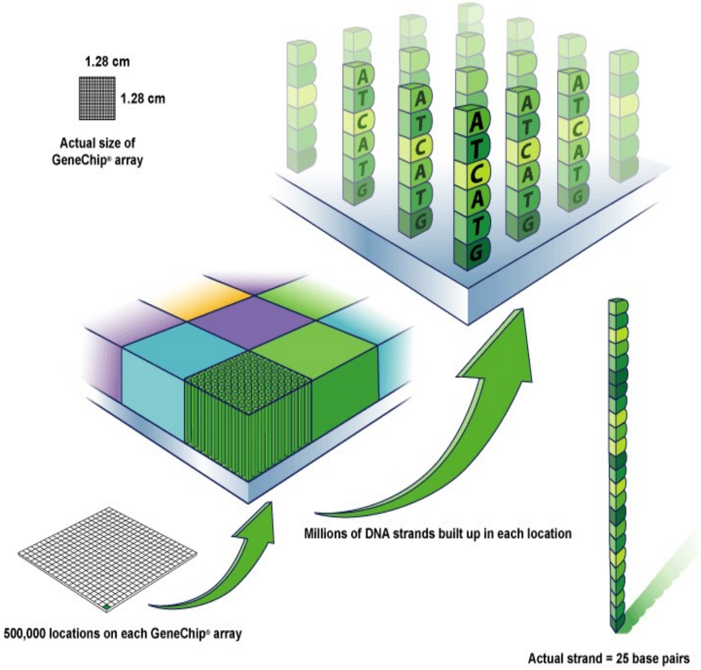
Microarrays

- Targeted approach for:
  - ❑ SNP / indel detection/genotyping
    - Screen for mutations that cause disease
  - ❑ Gene expression profiling
    - Which genes are expressed in which tissue?
    - Which genes are expressed “together”
    - Gene regulation (chromatin immunoprecipitation)
  - ❑ Fusion gene profiling
  - ❑ Alternative splicing
  - ❑ CNV discovery & genotyping
  - ❑ ….
- 50K to 4.3M probes per chip

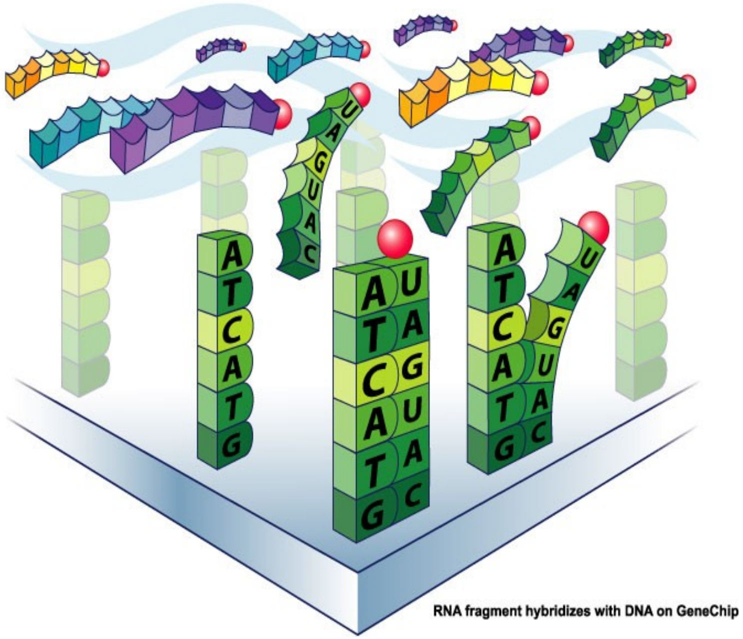
Microarray experiments

- Produce DNA library
  - If working on RNA, then make cDNA from mRNA
- Attach phosphor (marker) to DNA/cDNA
- Different color phosphors are available to compare many samples at once
- Hybridize DNA/cDNA over the micro array
- Scan the microarray with a phosphor-illuminating laser
- Illumination reveals hybridization
- Scan microarray multiple times for the different color phosphor's

DNA Microarray


1.28 cm
1.28 cm
Actual size of GeneChip® array
A T C A T G
Millions of DNA strands built up in each location
500,000 locations on each GeneChip® array
Actual strand = 25 base pairs

RNA fragments with fluorescent tags from sample to be tested



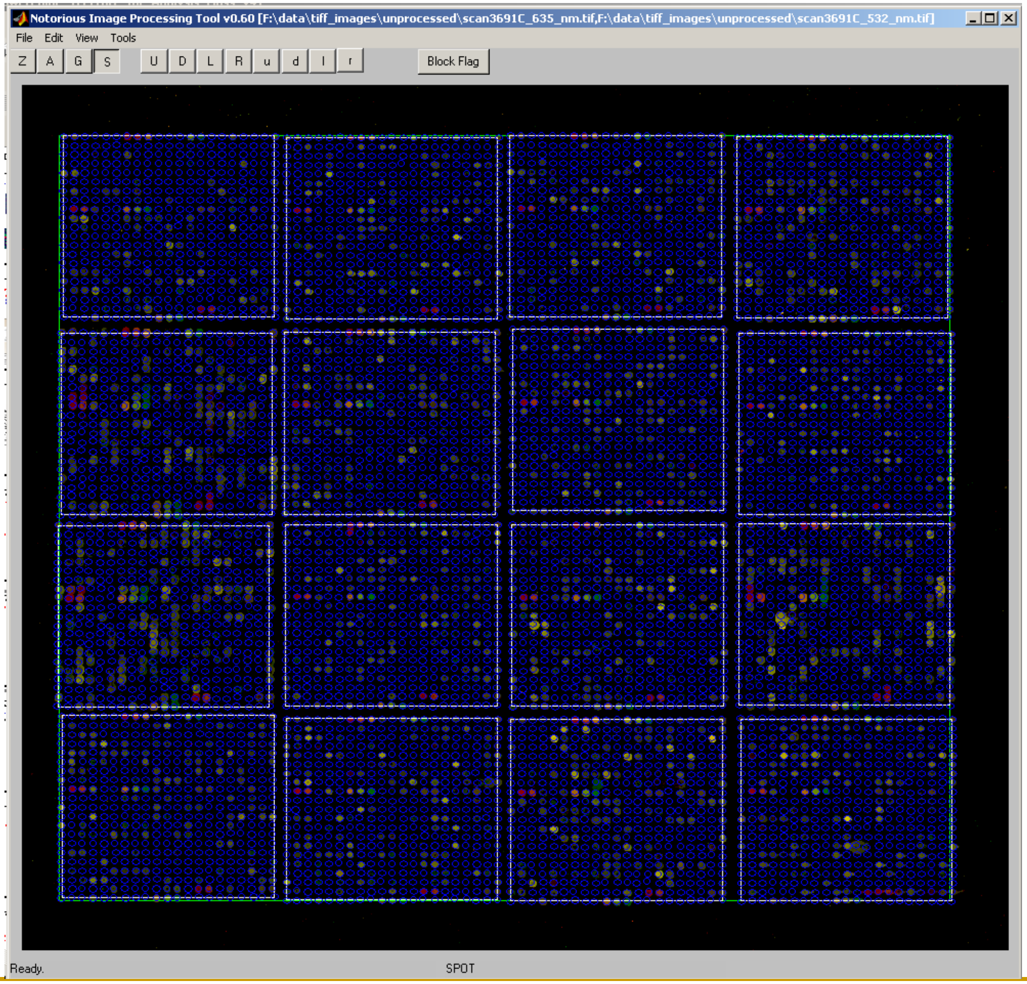
Millions of DNA strands build up on each location.

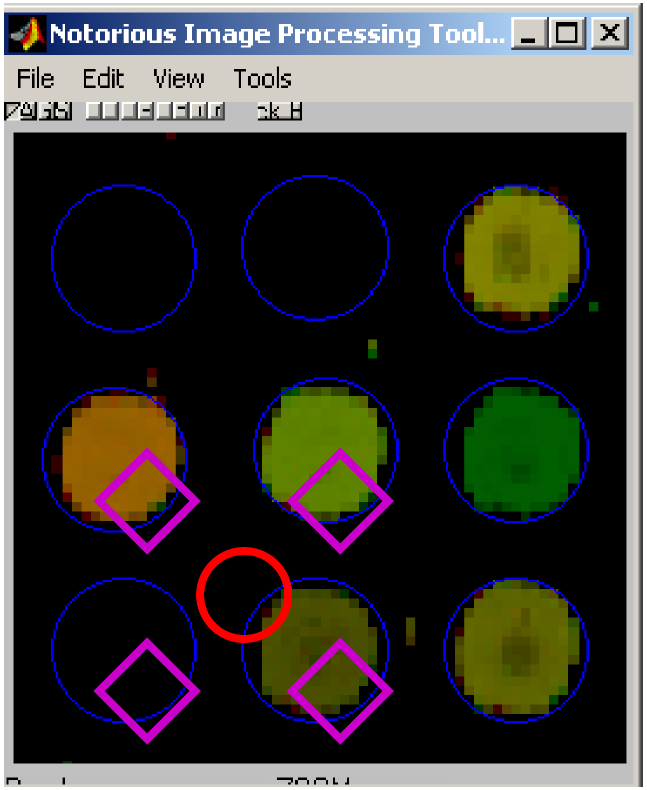
Tagged probes become hybridized to the DNA chip’s microarray.

http://www.affymetrix.com/corporate/media/image_library/image_library_1.affx

Image processing and normalization: what is microarray data?

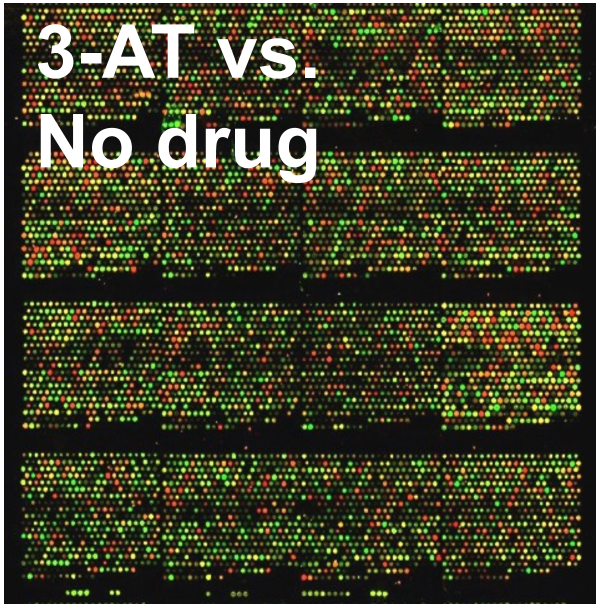
Microarray data is summary information from image files that come out of the scanner.
Image processing: line up grids, flag bad spots, quantify.

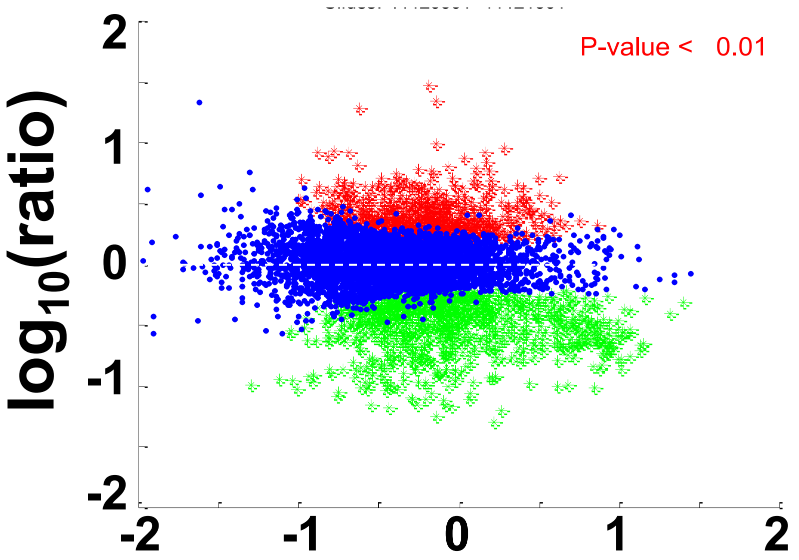

Notorious Image Processing Tool v0.60 [F:\data\tiff_images\unprocessed\scan3691C_635_nm.tif,F:\data\tiff_images\unprocessed\scan3691C_532_nm.tif]
File
Edit
View
Tools
Z
A
G
S
U
D
L
R
u
d
l
r
Block Flag
Ready.
SPOT

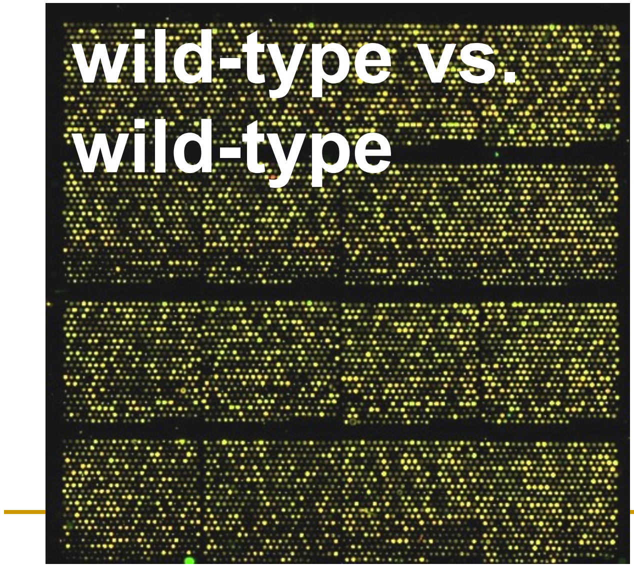


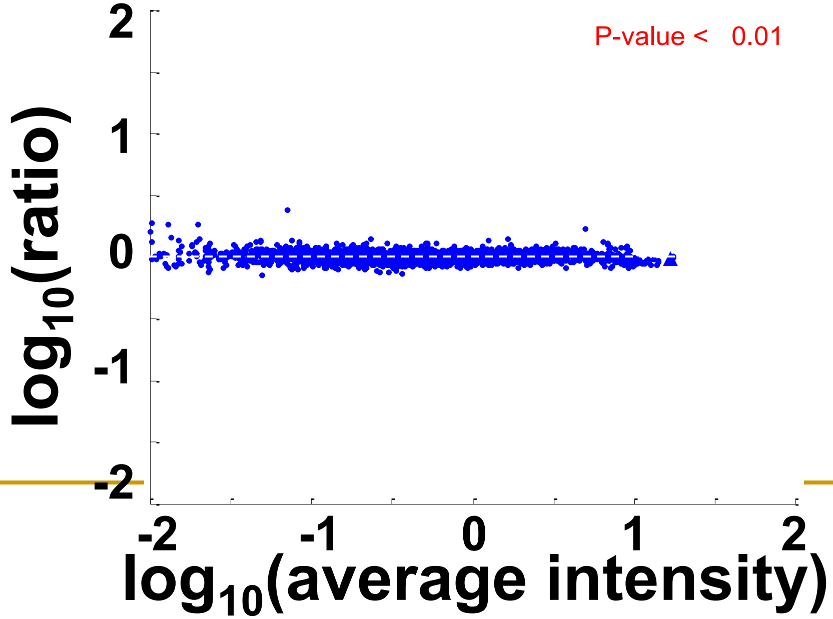
Segmentation & clustering algorithms

Data


3-AT vs.
No drug


P-value < 0.01
$\log_{10}$(ratio)
2
1
0
-1
-2
-2
-1
0
1
2


wild-type vs.
wild-type


P-value < 0.01
$\log_{10}$(ratio)
2
1
0
-1
-2
-2
-1
0
1
2
$\log_{10}$(average intensity)

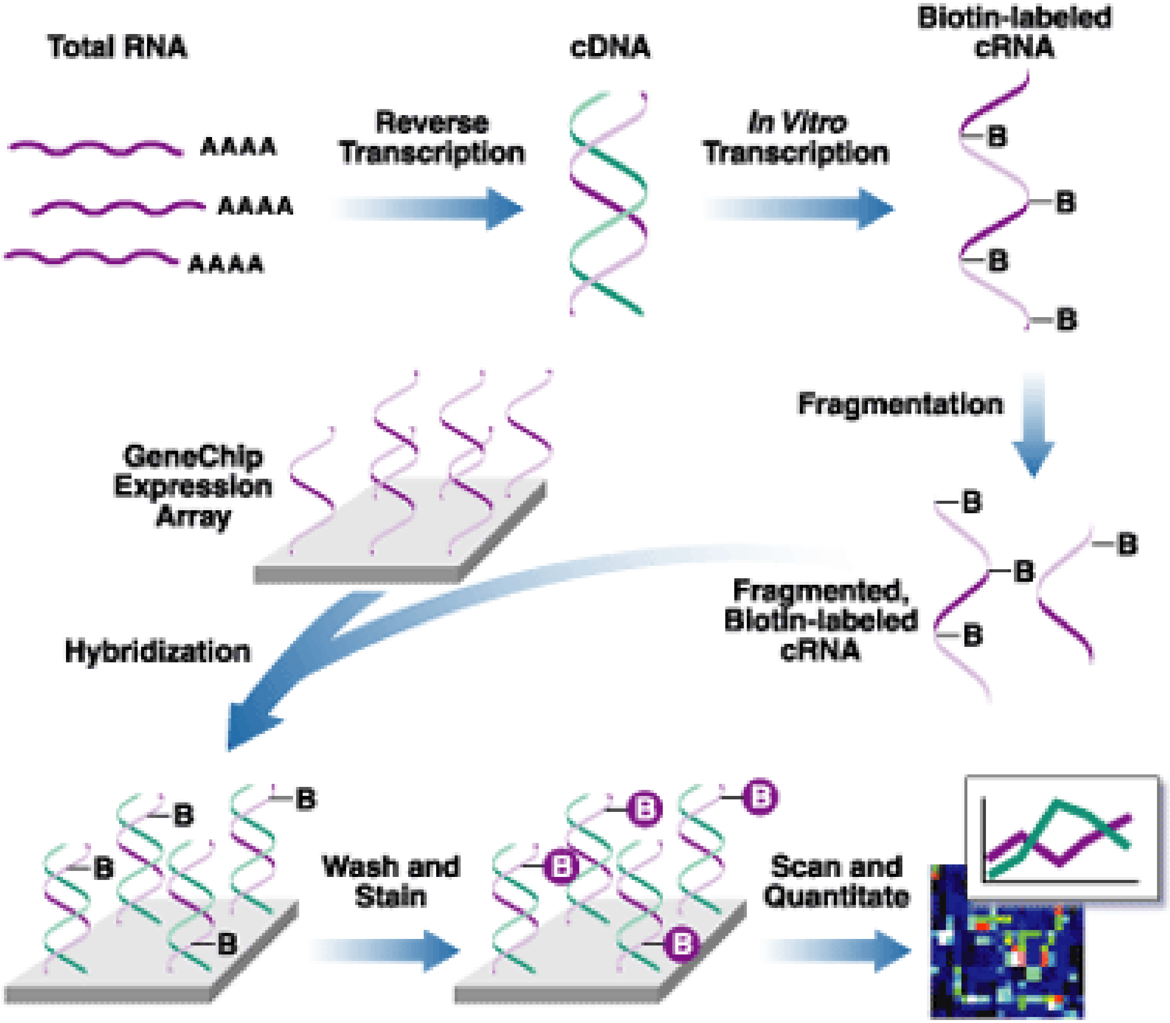
Microarray Vendors

- Illumina
  - Omni5 chip – 1000 Genomes: 4.3M markers
- Agilent
- NimbleGen
- Affymetrix
- All similar principles; different markers
- Custom designs can be made

Using Microarrays (SNP genotyping)

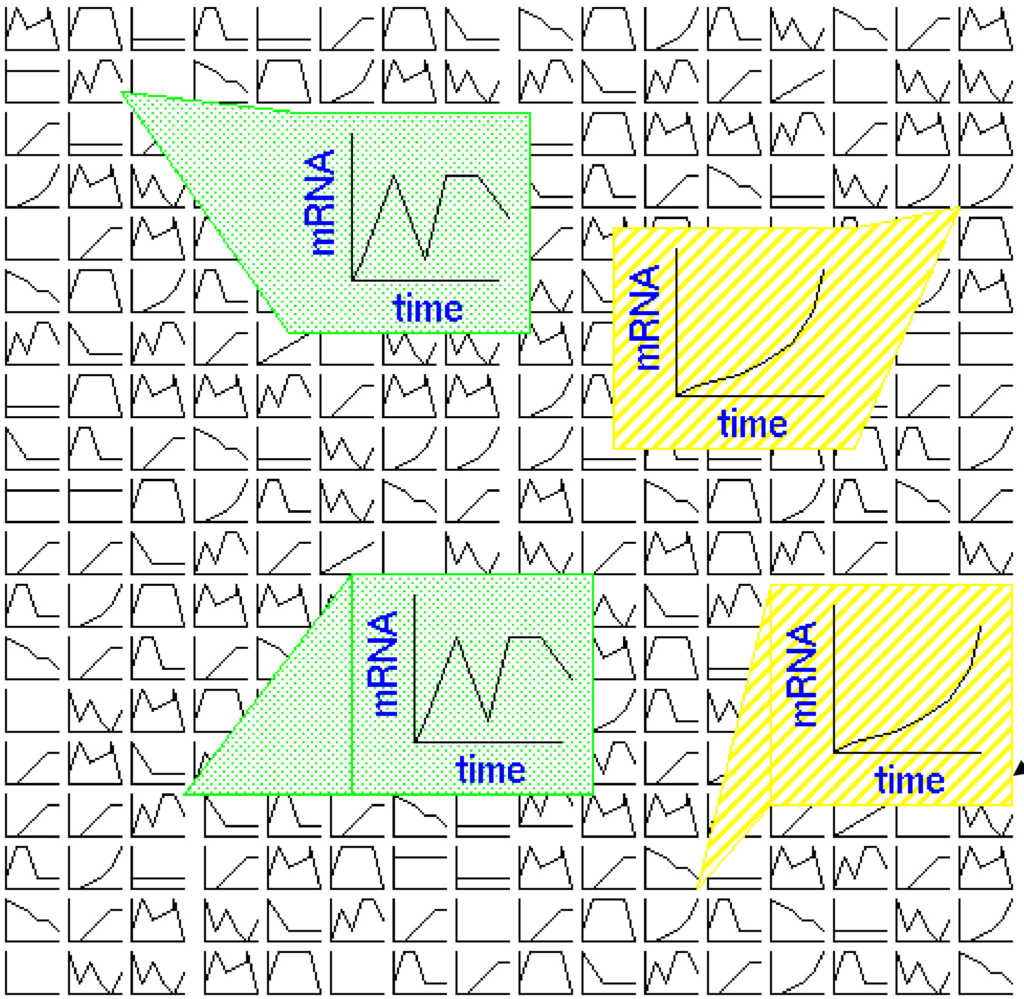
- Microarrays designed with oligonucleotides that harbor “target” SNPs.
- Comprehensively and rapidly study single nucleotide polymorphisms in human genomes
- Current SNP arrays feature 2 million genetic markers
- Analysis based on image processing and statistical methods

Microarray Experiments (gene expression)



www.affymetrix.com

Using Microarrays (gene expression)

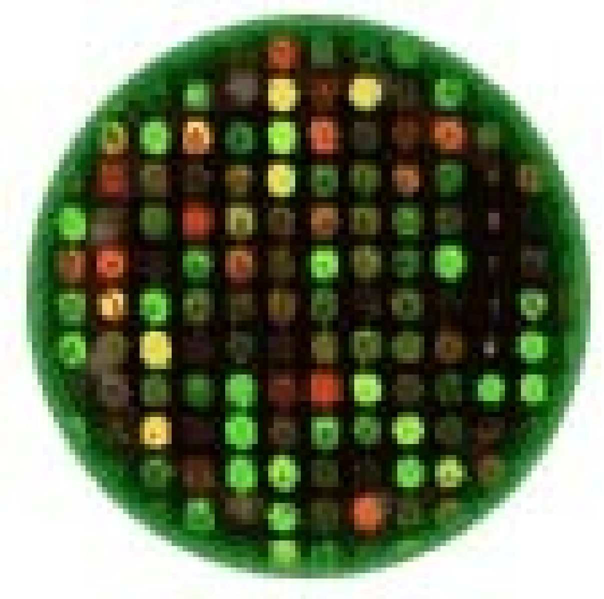


- Track the sample over a period of time to see gene expression over time
- Track two different samples under the same conditions to see the difference in gene expressions

Each box represents one gene's expression over time

Using Microarrays (cont'd)

- **Green**: expressed only from control
- **Red**: expressed only from experimental cell
- **Yellow**: equally expressed in both samples
- **Black**: NOT expressed in either control or experimental cells



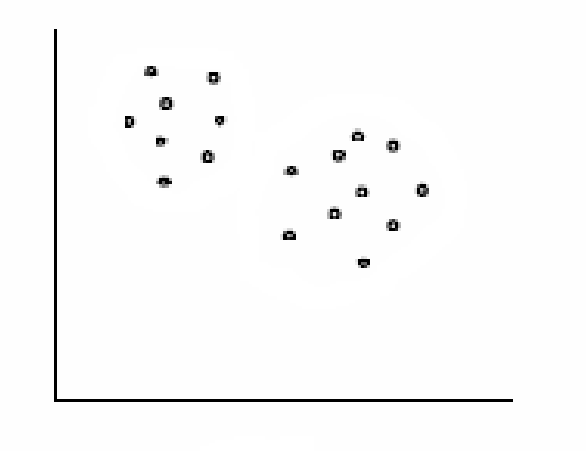
Clustering algorithms

- Clustering can be used for:
  - Primary analysis: cluster signals in microarray image to
    - Merge real signals from the same molecule
    - Separate real signals from noise
  - Secondary analysis:
    - Grouping probes: which probes are hybridized together?
      - Good for probes that might be repetitive in the genome/transcriptome
    - Gene expression: which genes are expressed together?
  - Many other bioinformatic applications exist

Homogeneity and Separation Principles

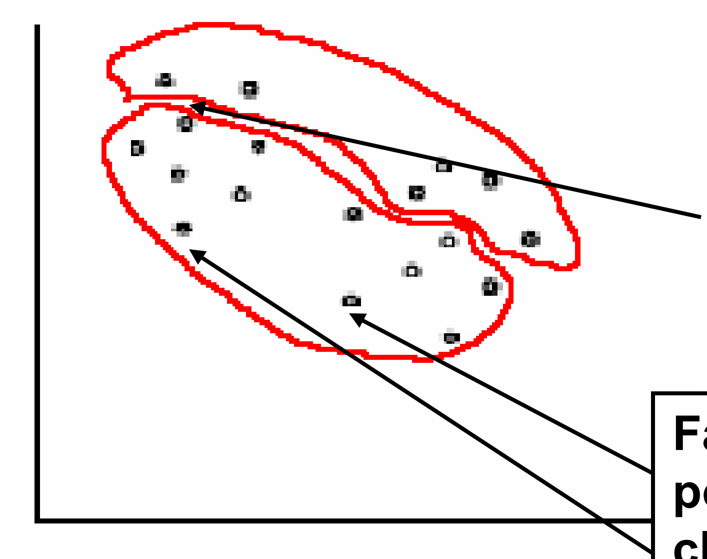
- **Homogeneity:** Elements within a cluster are close to each other
- **Separation:** Elements in different clusters are further apart from each other
- …clustering is not an easy task!

Given these points a clustering algorithm might make two distinct clusters as follows



Bad Clustering

This clustering violates both Homogeneity and Separation principles

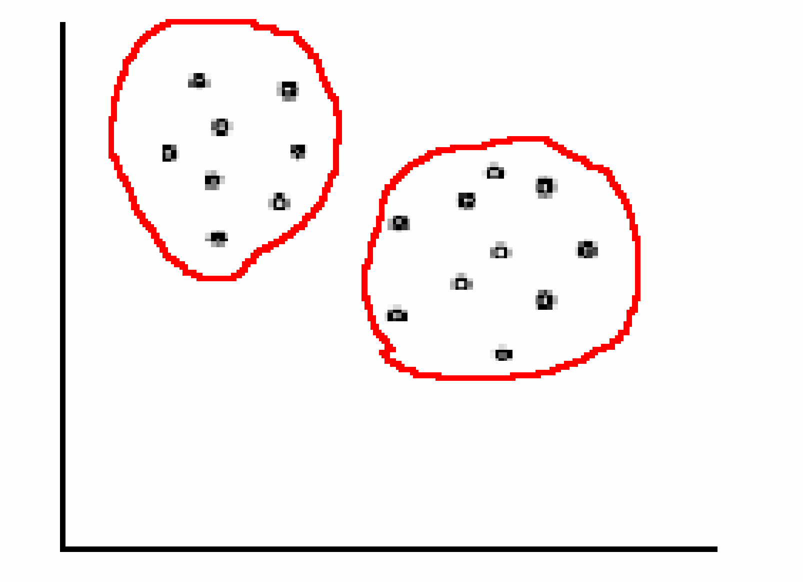


Close distances from points in separate clusters

Far distances from points in the same cluster

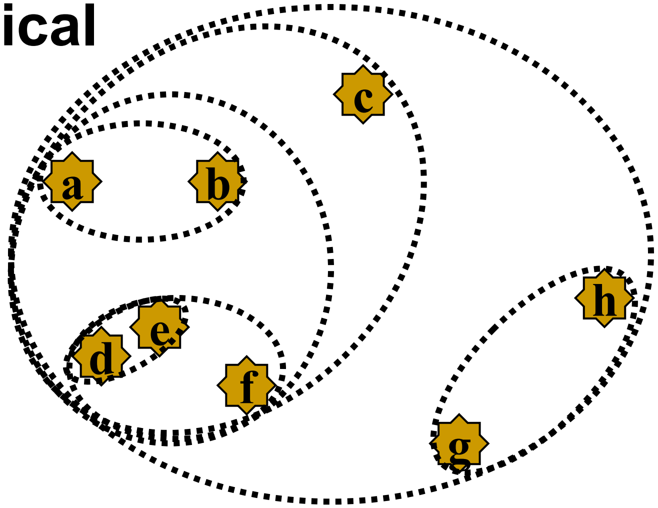
Good Clustering

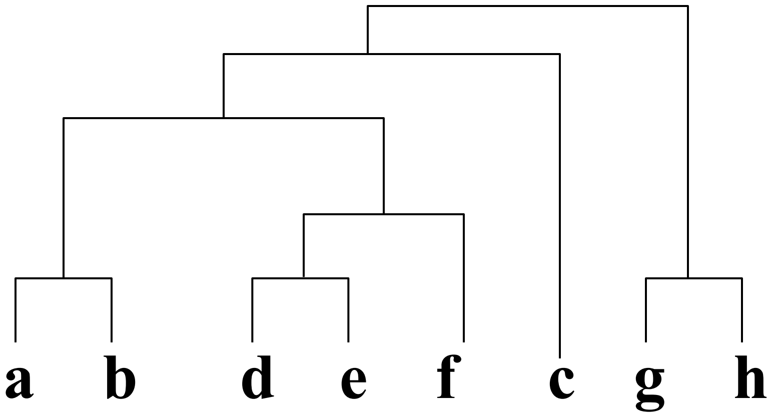
This clustering satisfies both Homogeneity and Separation principles



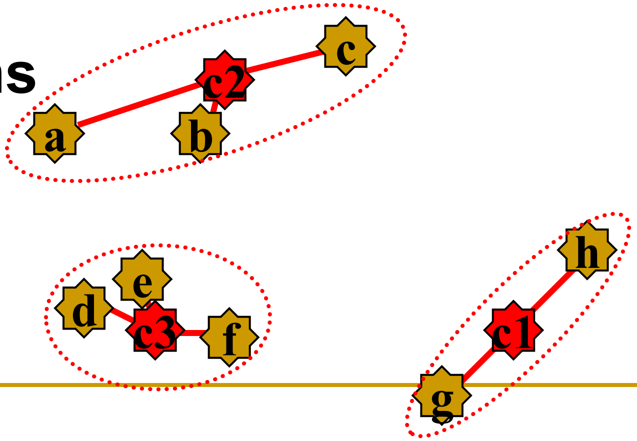
Clustering Algorithms

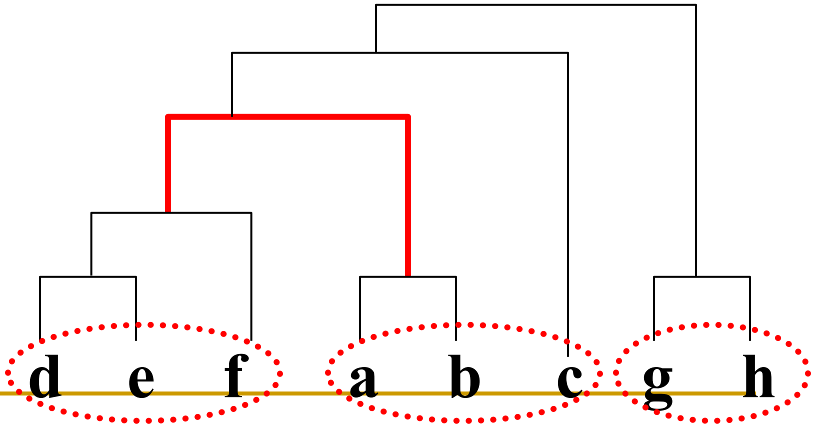
- **Hierarchical**





- **K-means**

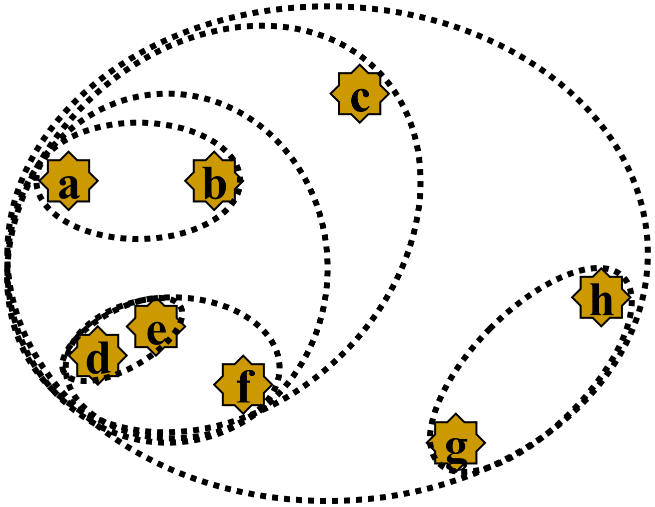




slide credits: M. Kellis

Hierarchical clustering

- Bottom-up algorithm:
  - Initialization: each point in a separate cluster
- At each step:
  - Choose the pair of closest clusters
  - Merge
- The exact behavior of the algorithm depends on how we define the distance CD(X,Y) between clusters X and Y
- Avoids the problem of specifying the number of clusters

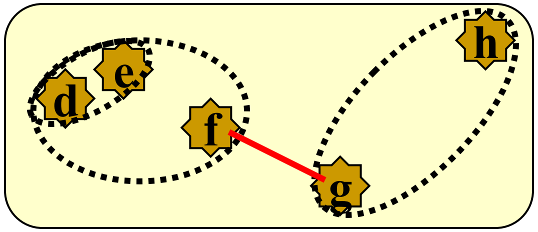


slide credits: M. Kellis

Distance between clusters

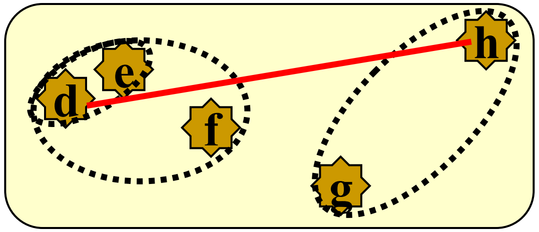
- $CD(X,Y)=\min_{x \in X,\, y \in Y} D(x,y)$

  Single-link method



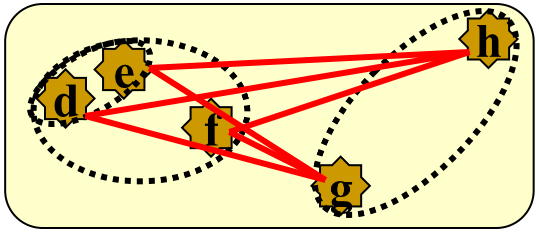
- $CD(X,Y)=\max_{x \in X,\, y \in Y} D(x,y)$

  Complete-link method



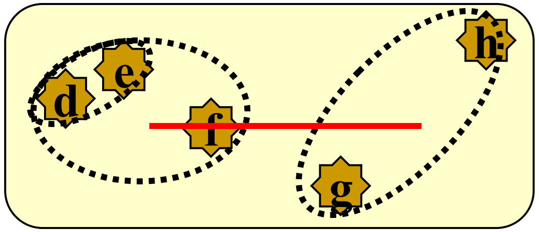
- $CD(X,Y)=\mathrm{avg}_{x \in X,\, y \in Y} D(x,y)$

  Average-link method



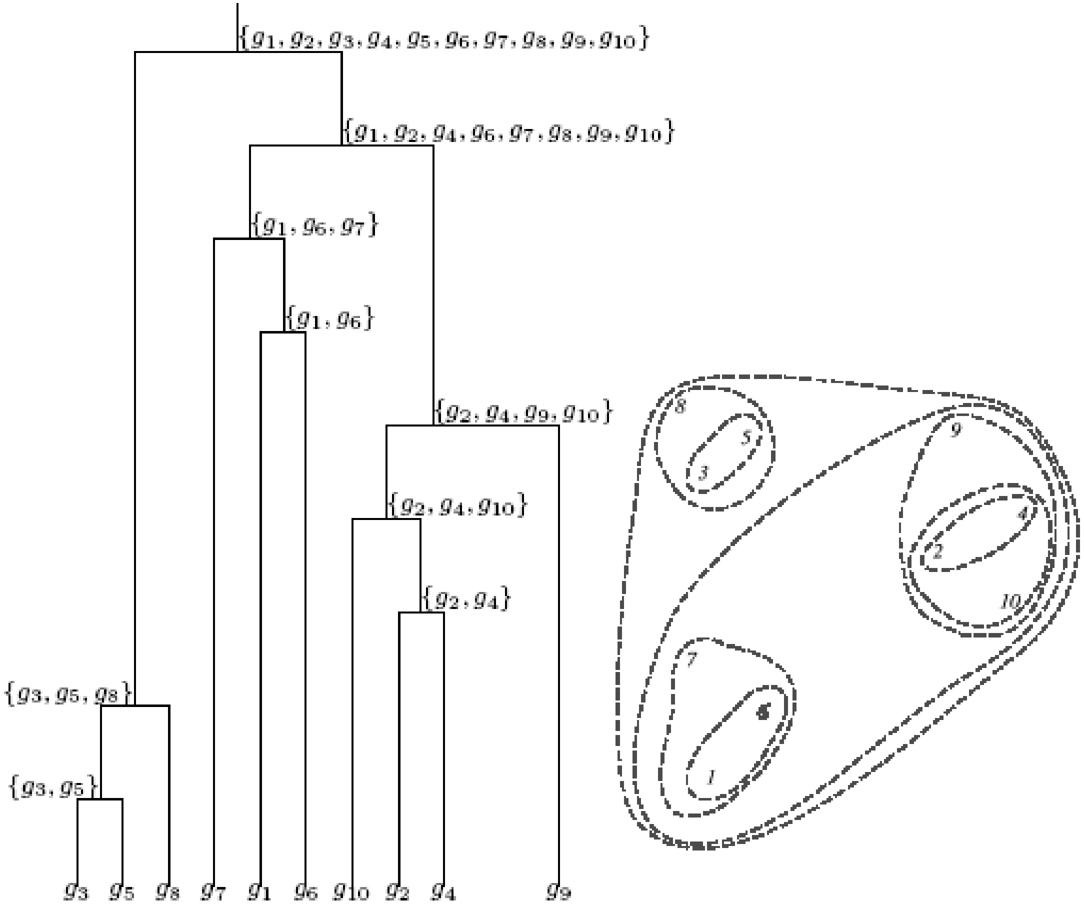
- $CD(X,Y)=D(\ \mathrm{avg}(X)\ ,\ \mathrm{avg}(Y)\ )$

  Centroid method

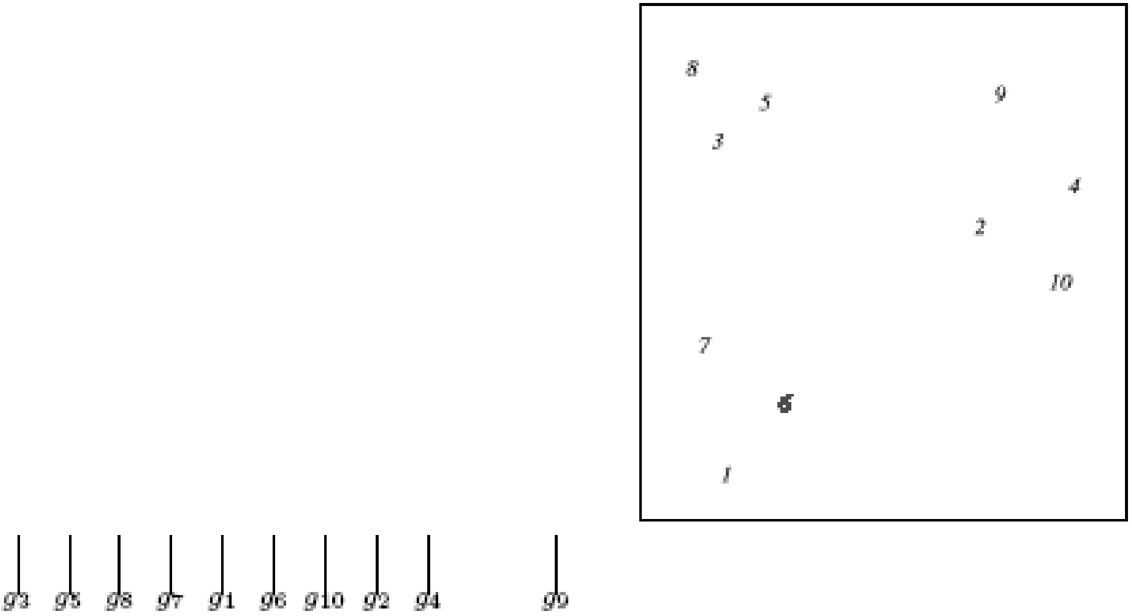


slide credits: M. Kellis

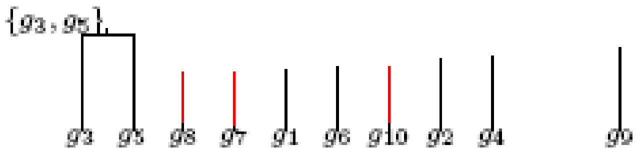
Hierarchical Clustering

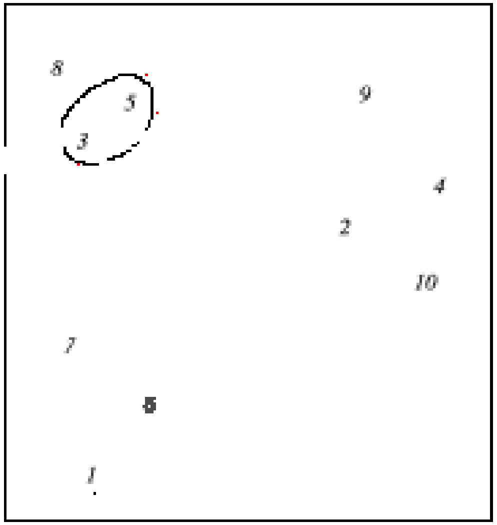

$\{g_1, g_2, g_3, g_4, g_5, g_6, g_7, g_8, g_9, g_{10}\}$
$\{g_1, g_2, g_4, g_6, g_7, g_8, g_9, g_{10}\}$
$\{g_1, g_6, g_7\}$
$\{g_1, g_6\}$
$\{g_2, g_4, g_9, g_{10}\}$
$\{g_2, g_4, g_{10}\}$
$\{g_2, g_4\}$
$\{g_3, g_5, g_8\}$
$\{g_3, g_5\}$
g_3 g_5 g_8 g_7 g_1 g_6 g_{10} g_2 g_4 g_9
8
5
3
9
4
2
10
7
6
1

Hierarchical Clustering: Example

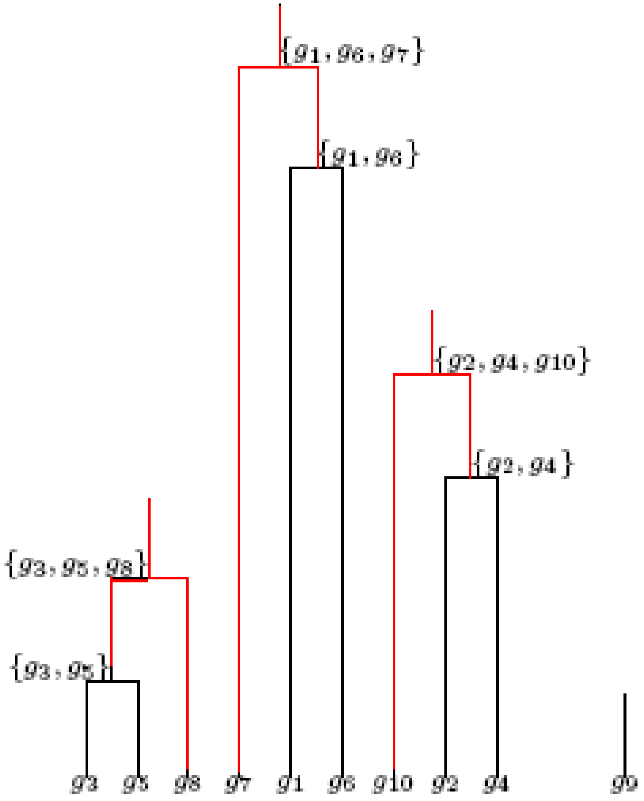


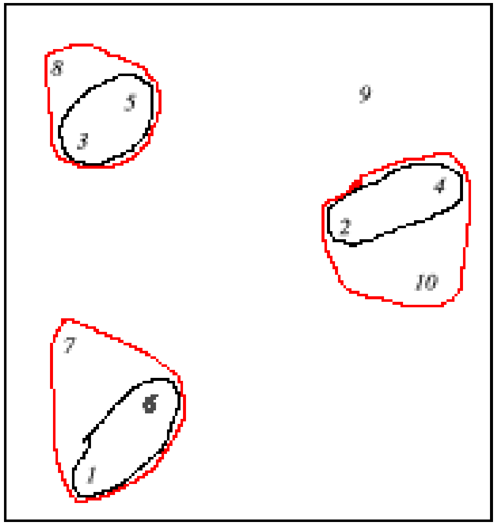
Hierarchical Clustering: Example



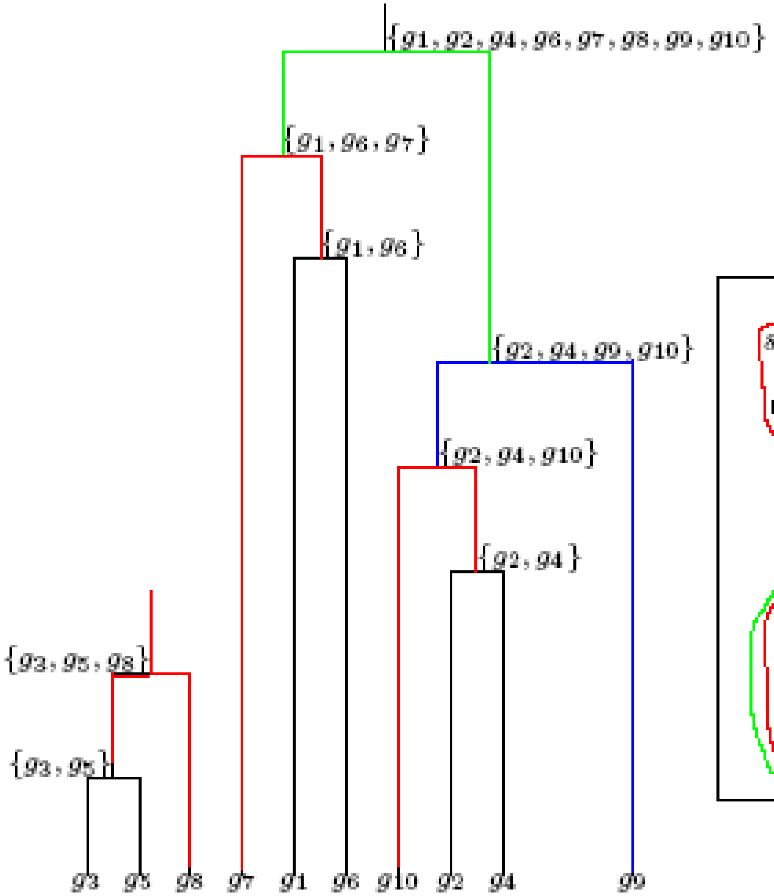


Hierarchical Clustering: Example

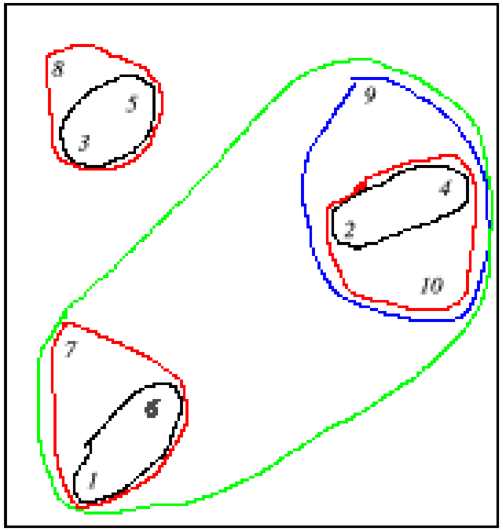




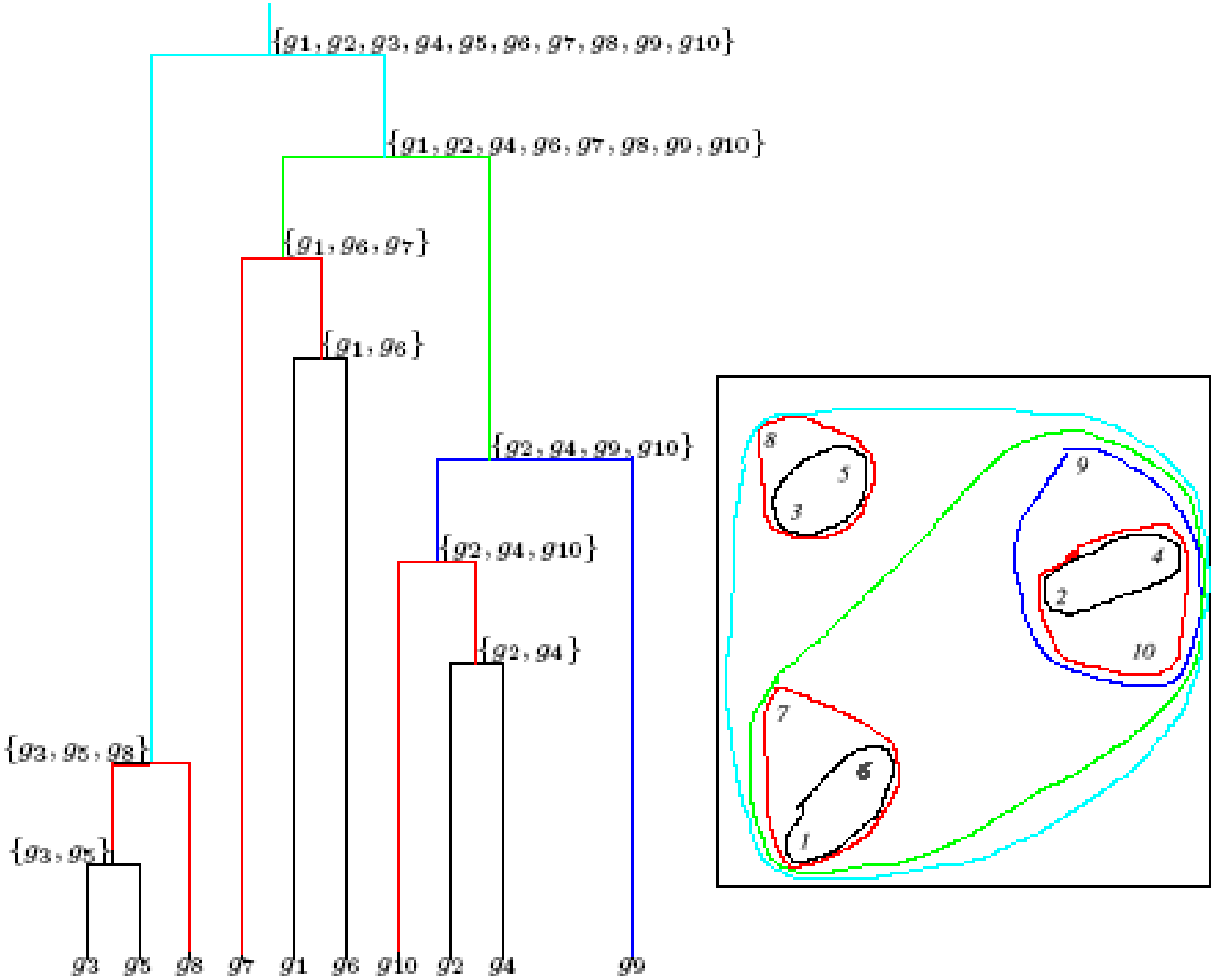
Hierarchical Clustering: Example



$\{g_1, g_2, g_4, g_6, g_7, g_8, g_9, g_{10}\}$
$\{g_1, g_6, g_7\}$
$\{g_1, g_6\}$
$\{g_2, g_4, g_9, g_{10}\}$
$\{g_2, g_4, g_{10}\}$
$\{g_2, g_4\}$
$\{g_3, g_5, g_8\}$
$\{g_3, g_5\}$
g_3 g_5 g_8 g_7 g_1 g_6 g_{10} g_2 g_4 g_9



Hierarchical Clustering: Example



Hierarchical Clustering Algorithm

```
Hierarchical Clustering ($d$ , $n$)
  Form $n$ clusters each with one element
  Construct a graph $T$ by assigning one vertex to each cluster
  while there is more than one cluster
    Find the two closest clusters $C_1$ and $C_2$
    Merge $C_1$ and $C_2$ into new cluster $C$ with $/C_1/ + /C_2/$ elements
    Compute distance from $C$ to all other clusters
    Add a new vertex $C$ to $T$ and connect to vertices $C_1$ and $C_2$
    Remove rows and columns of $d$ corresponding to $C_1$ and $C_2$
    Add a row and column to $d$ corrsponding to the new cluster $C$
  return $T$
```

The algorithm takes a *nxn* distance matrix *d* of pairwise distances between points as an input.

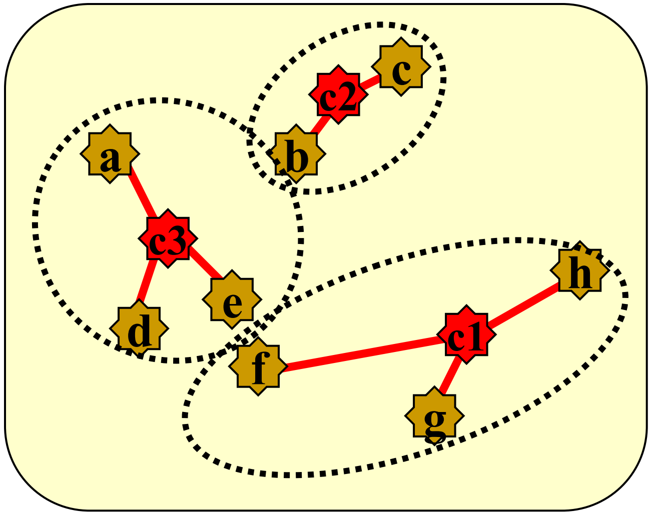
Hierarchical Clustering Algorithm

```
Hierarchical Clustering ($d$ , $n$)
  Form $n$ clusters each with one element
  Construct a graph $T$ by assigning one vertex to each cluster
  while there is more than one cluster
    Find the two closest clusters $C_1$ and $C_2$
    Merge $C_1$ and $C_2$ into new cluster $C$ with $/C_1/ + /C_2/$ elements
    Compute distance from $C$ to all other clusters
    Add a new vertex $C$ to $T$ and connect to vertices $C_1$ and $C_2$
    Remove rows and columns of $d$ corresponding to $C_1$ and $C_2$
    Add a row and column to $d$ corrsponding to the new cluster $C$
  return $T$
```

Different ways to define distances between clusters may lead to different clusterings

K-Means Clustering Algorithm

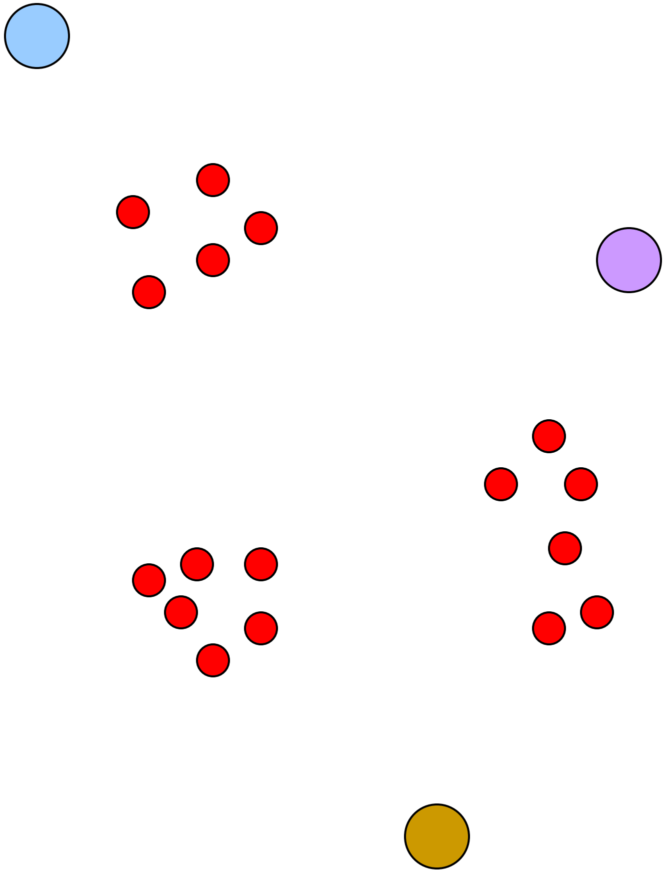
- Each cluster **X_i** has a center **c_i**
- Define the clustering cost criterion
- $\mathrm{COST}(X_1,\ldots X_k) = \sum_{X_i} \sum_{x \in X_i} |x - c_i|^2$
- Algorithm tries to find clusters $X_1 \ldots X_k$ and centers $c_1 \ldots c_k$ that minimize COST
- K-means algorithm:
  - Initialize centers
  - Repeat:
    - Compute best clusters for given centers
    - → Attach each point to the closest center
    - Compute best centers for given clusters
    - → Choose the centroid of points in cluster
  - Until the changes in COST are "small"



slide credits: M. Kellis

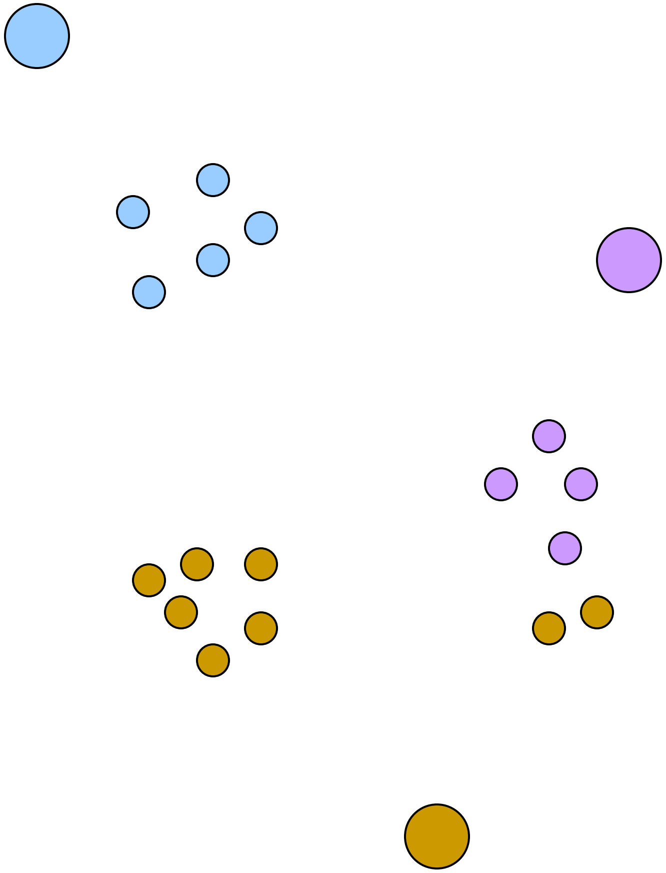
K-Means Algorithm

- Randomly Initialize Clusters



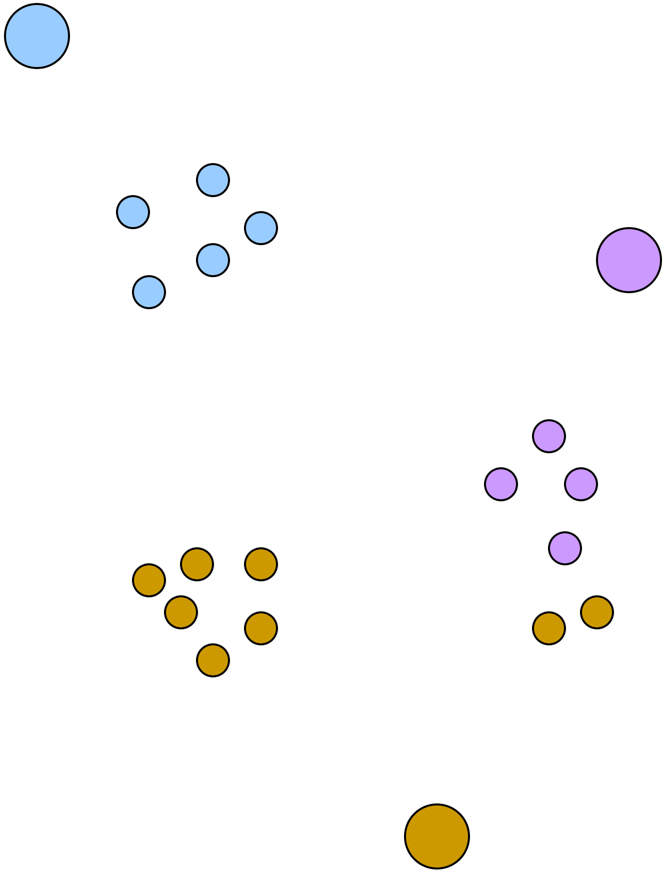
K-Means Algorithm

- Assign data points to nearest clusters



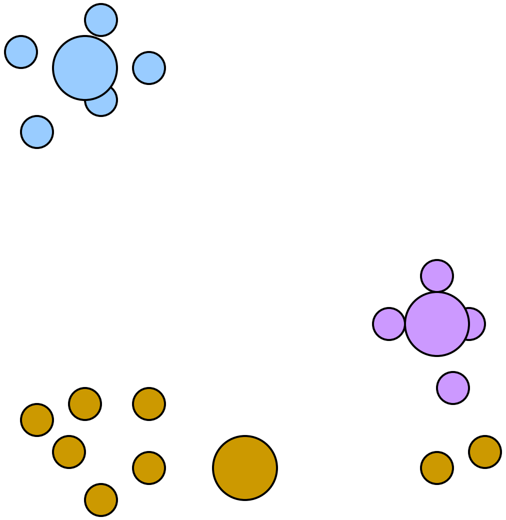
K-Means Algorithm

- Recalculate Clusters



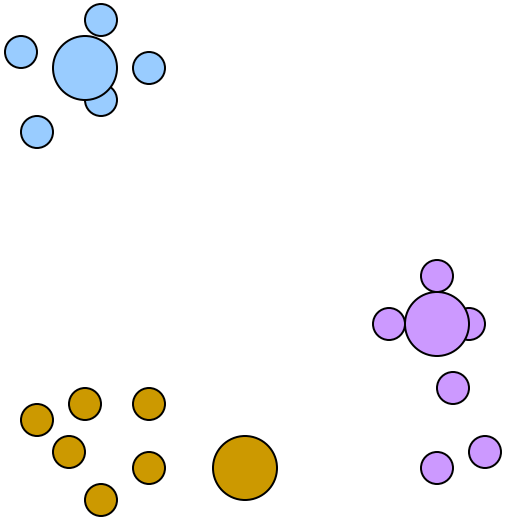
K-Means Algorithm

- Recalculate Clusters



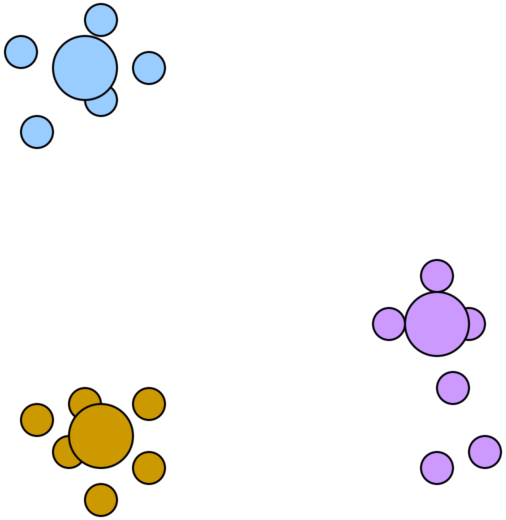
K-Means Algorithm

- Repeat



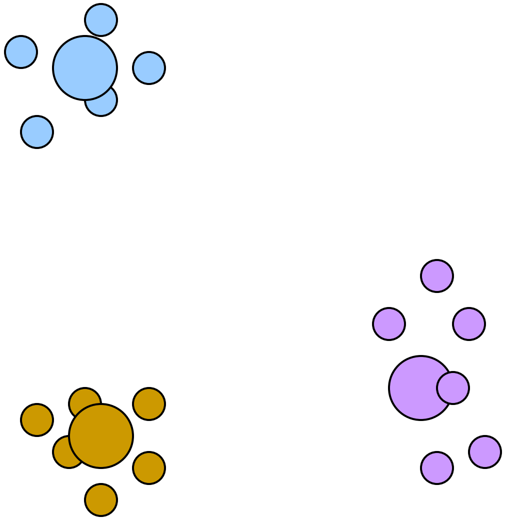
K-Means Algorithm

- Repeat



K-Means Algorithm

- Repeat … until convergence



Time: O(KNM) per iteration

N: #genes
M: #conditions

K-Means Greedy Algorithm

```
ProgressiveGreedyK-Means($k$)
Select an arbitrary partition $P$ into $k$ clusters
while forever
  $bestChange \leftarrow 0$
  for every cluster $C$
    for every element $i$ not in $C$
      if moving $i$ to cluster $C$ reduces its clustering cost
        if $(cost(P) - cost(P_{i \rightarrow C}) > bestChange$
          $bestChange \leftarrow cost(P) - cost(P_{i \rightarrow C})$
          $i^* \leftarrow I$
          $C^* \leftarrow C$
  if $bestChange > 0$
    Change partition $P$ by moving $i^*$ to $C^*$
  else
    return $P$
```

Clustering: Gene ontology (GO)

- Catalogue for genes, gene products, gene annotations across all species
- Clustered genes with respect to biological processes they were involved in
- Single gene can appear in multiple processes

GO-Biological Process categories

		# annotated genes (mouse)
Very Broad	metabolism	1548
	development	2341
Broad	vision	163
	CNS development	137
	eye morphogenesis	21
Mid-level	ATP biosynthesis	36
	pigment metabolism	25
	striated muscle contraction	33
Narrow	eye pigment metabolism	3
	insulin secretion	4