♠ Clustering

- Clustering is applied to multivariate data
  - Gene $i, i = 1, \ldots, p$
  - Expression level $x_{i,j}$ on array j
- The data matrix

$$\mathbf{X} = [x_{ij}] = \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,n} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{p,1} & x_{p,2} & \cdots & x_{p,n} \end{pmatrix}$$

- Columns are snapshots of gene expression
- Normalization of rows or columns (to make them comparable)
- Statistics used for similarity – comparing “distances” between genes
  - Euclidean distance between gene i and gene k

$$d(i,k) = \sqrt{\sum_{j=1}^{n} (x_{i,j} - x_{k,j})^2}$$

  - Correlation $r_{i,k}$ between two genes
- Similar ideas to compare samples/microarrays

♠ K-Means clustering

- Partitions the data into unrelated clusters
- Shuffles observations from cluster to cluster to improve similarity within clusters
- Fast and makes efficient use of computer memory
- The final clustering depends on the initial partition
- Number of clusters remains constant and must be specified

♠ Hierarchical clustering

- Conceptually, hierarchical clustering recursively partitions the data into a tree like structure
- As usually implemented, clusters are agglomerated pairwise
- Hierarchical agglomeration
  Single linkage
  Average linkage
  Complete linkage
- Number of clusters can be assessed retrospectively

♠ You can make your own clustering algorithm

- Example: consolidation of hierarchical clusters
  Perform hierarchical clustering using average linkage
  Run the k-means algorithm using hierarchical clusters as the initial state

♠ Clustering in general

- The higher the dimension of the data the more sparse it is
- Well defined clusters may not or may not exist in the data
- No sensible clustering method is necessarily “right” - appropriate methods to use are data dependent - Try more than one clustering method
- Clustering is very much over-used in gene expression analysis. Often the use in a given application is quite aimless ...

♠ Clustering for Metagenes

- We use k-means clustering just as a device to
  - Reduce dimension (e.g., from 20,000 genes to 3-500 metagenes), and then
  - expecting that a dominant principal component/singular factor within each of the resulting, relatively small clusters will adequately summarise any underlying common pattern
- The idea is that multiple patterns (”pathways”) are generating the complex patterns of variation and covariation among genes, but life is not 20,000 dimensional. The clustering and then averaging within clusters aims to reduce dimension and improve signal resolution
- Good tools for clustering include the Matlab and R/Splus clustering routines (generally limited to smaller numbers of variables)
- We have found Gavin Sherlock’s *xcluster* software easy - look for Gavin Sherlock’s Stanford web site
  http://genetics.stanford.edu/˜sherlock/cluster.html
  We have found this easy to use and robust for k-means; it does other forms of clustering too. The web site has a nice series of tutorial examples.
- Michael Eisen’s cluster software, and the associated treeview software for hierarchical clustering, is taylored to gene expression data and quite widely used (look for Eisen at the Berkeley web site). Look also at the Cluster & treeview web sites:
  http://bonsai.ims.u-tokyo.ac.jp/mdehoon/software/cluster/software.htm
  http://sourceforge.net/projects/jtreeview/
- Other tools are mentioned and linked on the Bioconductor web site in various contexts:
  www.bioconductor.org