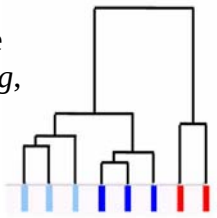
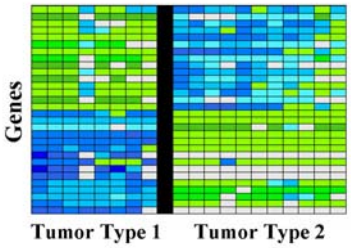
BINF 636: Lecture 9: Clustering: How Do They Make and Interpret Those Dendrograms and Heat Maps; Differences Between Unsupervised Clustering and Classification.

Description: Clustering, for the purpose of this lecture, is the exploratory partitioning of a set of data points into subgroups (clusters) such that members of each subgroup are relatively similar to each other and members of distinct clusters are relatively dissimilar. For example, one might have gene expression profiles from a set of samples of a particular type of tumor and wish to see if the samples separate out into distinct subgroups. In this case one could be looking to uncover evidence of previously unknown subtypes, or one might wish to see if the results of clustering the gene expression profiles are consistent with classification by histopathology.



In this class we will describe how dendrograms, such as the example to the right, are constructed using *hierarchical agglomerative clustering*, where one starts with each of the data points as an individual cluster, and in successive steps combines the pair of clusters that are "closest" to each other into one new cluster. This requires specifying a distance measure between data points and between clusters. Each clustering step reduces by one the number of existing clusters until at the end of the final step there is one cluster containing all the data points. If one has ordered the data points along a line so that at each step the clusters that are joined together are adjacent to each other, one can draw a corresponding diagram (dendrogram) where the heights of the vertical lines reflect the distance between the pair of clusters joined at each stage of the procedure. If one has, e.g., microarray data from a set of tumor samples, one can cluster both the tissue sample gene expression profiles and the expression profiles of the genes across the tissue samples, thus determining a corresponding ordering of the tumor samples and of the genes. One can then color code each rectangle representing the expression level of one gene in one tumor sample, producing a *heat map* such as the example to the right.



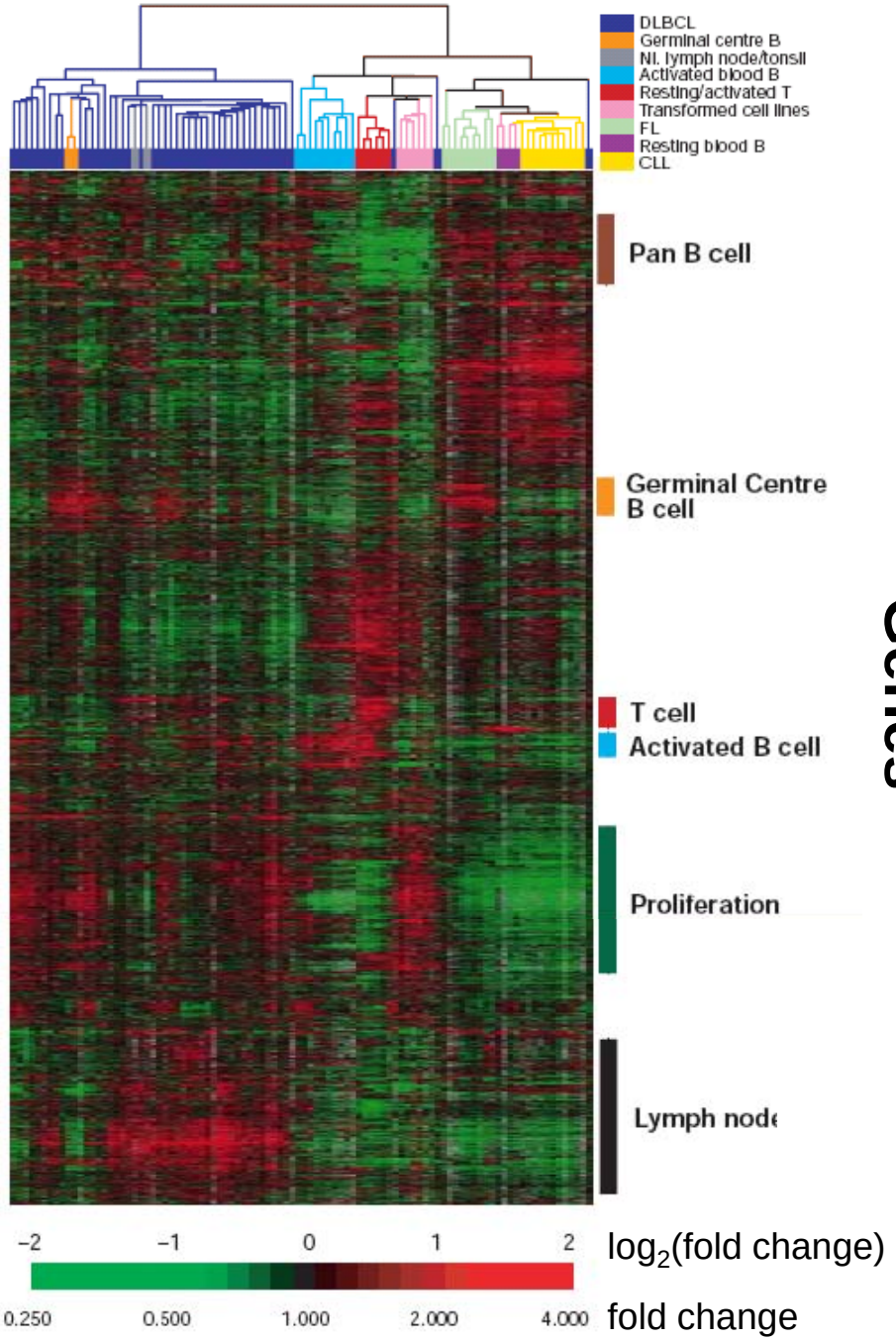
We will describe how these procedures are carried out, and how the resulting hierarchies of clusters can depend on the specifications for distances between data points and between clusters. The type of changes in appearance that may occur in a dendrogram in response to small changes in the data points will also be illustrated. An individual dendrogram is essentially a one-dimensional ordering of a data set, in contrast with two or three dimensional visualizations that can be obtained by principal component analysis (PCA) which is the subject of Lecture 10. The difference between exploratory (unsupervised) clustering and classification will be noted, along with the importance of proper validation of classification methods.

Alan E. Berger, Ph.D., JHBMC Lowe Family Genomics Core, Johns Hopkins University School of Medicine, aberger9@jhmi.edu (410) 550-5089

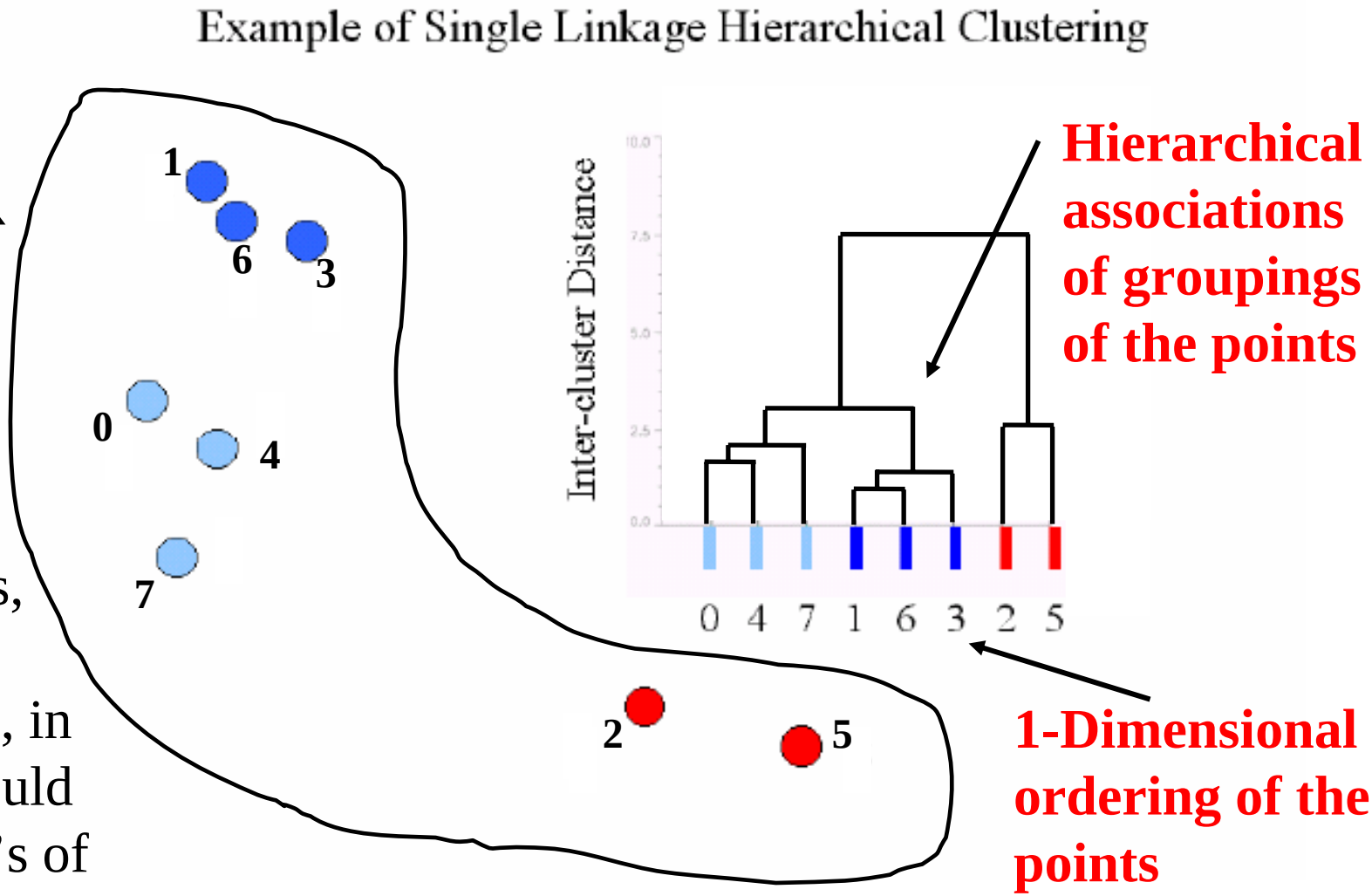
Clustering example; modified version of Figure 1 of **A. A. Alizadeh et al.,** Distinct Types of Diffuse Large B-Cell Lymphoma Identified by Gene Expression Profiling, Nature 403, 3 Feb. 2000, pp. 503-511.

Centroid clustering on log of fold changes of measured expression levels with Pearson correlation similarity for tissue samples (columns) and cos(angle) similarity for genes (rows). Fold changes are ratios of mRNA expression level in tissue sample relative to mRNA level in reference pool. The values in each row (gene) were median centered before the clustering / heat map plot.

Heat map color code for coloring of each [tissue sample × gene] rectangle is at the bottom of the figure



Information from a Clustering



Points being clustered, each point is an n-tuple of numbers, here have (x,y) pairs, in general, could have 1000's of components for each point, e.g., microarray data

Hierarchical associations of groupings of the points

1-Dimensional ordering of the points

Each point has form $P = (x_1, x_2, \ldots, x_n)$

Clustering: How Do They Make Those Dendrograms and Heat Maps – Outline

- Definition of unsupervised clustering
- Dendrogram construction by hierarchical agglomerative clustering given specified inter-cluster and inter-point distance measures
- Uniqueness of the dendrogram if an unambiguous choice of left/right ordering is specified for each join of two clusters in the dendrogram construction
- Dependence of the clustering dendrogram on the definition of inter-cluster distance.
- Additional examples, Heat map construction
- Briefly noting other methods for clustering and data visualization
- The difference between exploratory and supervised clustering

Some References for Clustering

[1] R. O. Duda, P. E. Hart and D. G. Stork, *Pattern Classification*, Second Edition, John Wiley & Sons, 2001, Chapter 10 – Unsupervised Learning and Clustering

[2] R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis*, Fourth Edition, Prentice Hall, 1998, Chapter 12 – Clustering, Distance Methods, and Ordination

[3] J. Quackenbush, Computational Analysis of Microarray Data, *Nature Reviews Genetics*, 2 (2001), pp. 418-427.

[4] R. Simon, M. D. Radmacher, K. Dobbin and L. M. McShane, Pitfalls in the Use of DNA Microarray Data for Diagnostic and Prognostic Classification, *Journal of the National Cancer Institute*, 95 (2003), pp. 14-18. <exploratory class discovery, supervised classification, proper use of cross-validation> see also, M. West et al., Predicting the Clinical Status of Human Breast Cancer by Using Gene Expression Profiles, *PNAS*, 98 (2001), pp. 11462-11467.

[5] http://en.wikipedia.org/wiki/Data_clustering

[6] M. B. Eisen, P. T. Spellman, P. O. Brown and D. Botstein, Cluster Analysis and Display of Genome-wide Expression Patterns, *PNAS*, 95 (1998), pp. 14863-14868

[7] Software for performing a variety of clustering methods is available in <with usual disclaimers>, e.g., R (open source) & open source R programs (see in particular the Bioconductor suite of software), and in general data analysis software such as MATLAB and IDL, statistics packages (SAS etc.),

commercial packages for microarray analysis (such as Partek, GeneSpring, GeneSifter, JMP Genomics, and the MATLAB Bioinformatics Toolbox),

free software such as Cluster (http://rana.lbl.gov/EisenSoftware.htm), and GenePattern (http://www.broad.mit.edu/cancer/software/genepattern/),

software available at NIH, including the Mathematical and Statistical Computing Lab toolbox of scripts that complement and interface with the JMP statistics package (http://abs.cit.nih.gov/MSCLtoolbox/), BRB Array

Tools (http://linus.nci.nih.gov/BRB-ArrayTools.html) and
mAdb (http://nciarray.nci.nih.gov/).

[8] M. Zvelebil and J. O. Baum, *Understanding Bioinformatics*, Garland Science, NY, 2008, Chapter 16 – Clustering Methods and Statistics

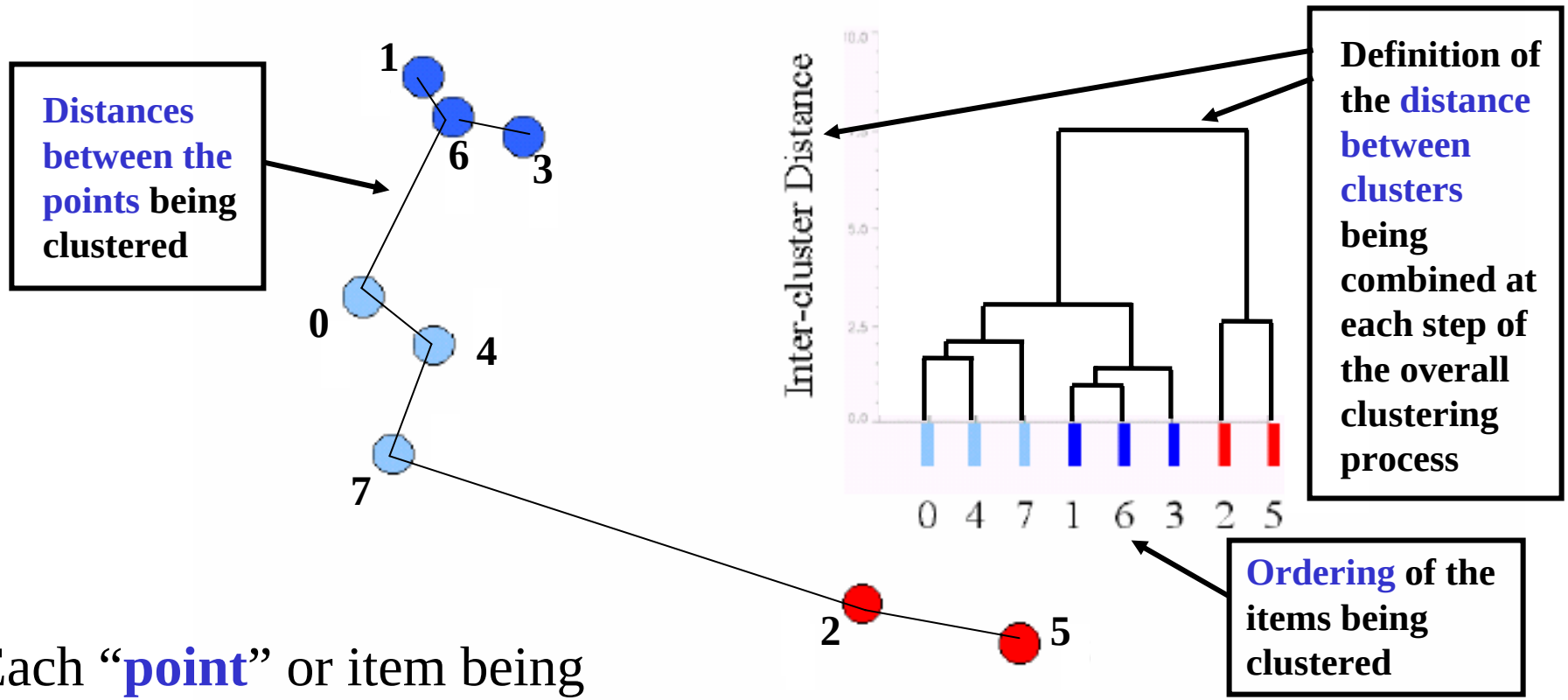
[9] D. Stekel, *Microarray Bioinformatics*, Cambridge University Press, Cambridge, 2003, Chapter 8 – Analysis of Relationships Between Genes, Tissues or Treatments

CLUSTERING

(Without Training Data: Unsupervised Clustering)
Exploratory separation of data into groups of points. Points in distinct groups to be more different than points within one group. ***Discovery of classes***. Information about the data may be used to evaluate the results but is **NOT** used in doing the clustering.

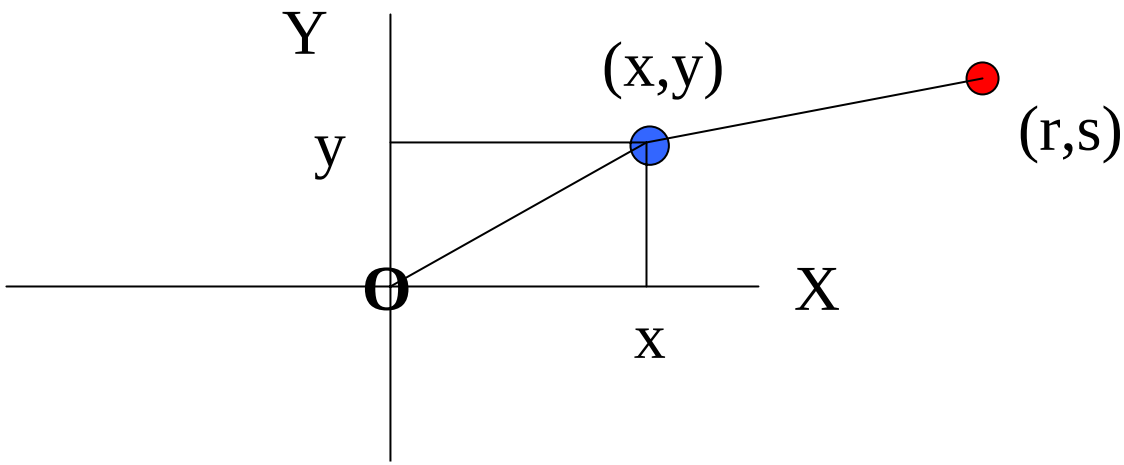
(With Training Data: Supervised Clustering)
Accurate classification of new data points. Verification of accuracy of the training data. ***Discovery of additional classes***. Information about the training data **is** used to do the clustering/classification.

Components of the Clustering Process



Each "**point**" or item being clustered could, e.g., be a set of gene expression values from a tumor sample. A **cluster** is a subset of the items being clustered. Start with each individual point as a cluster, and successively combine the **closest pair** of clusters into one new cluster. End with one cluster.

Standard (Euclidean) Distance between two points

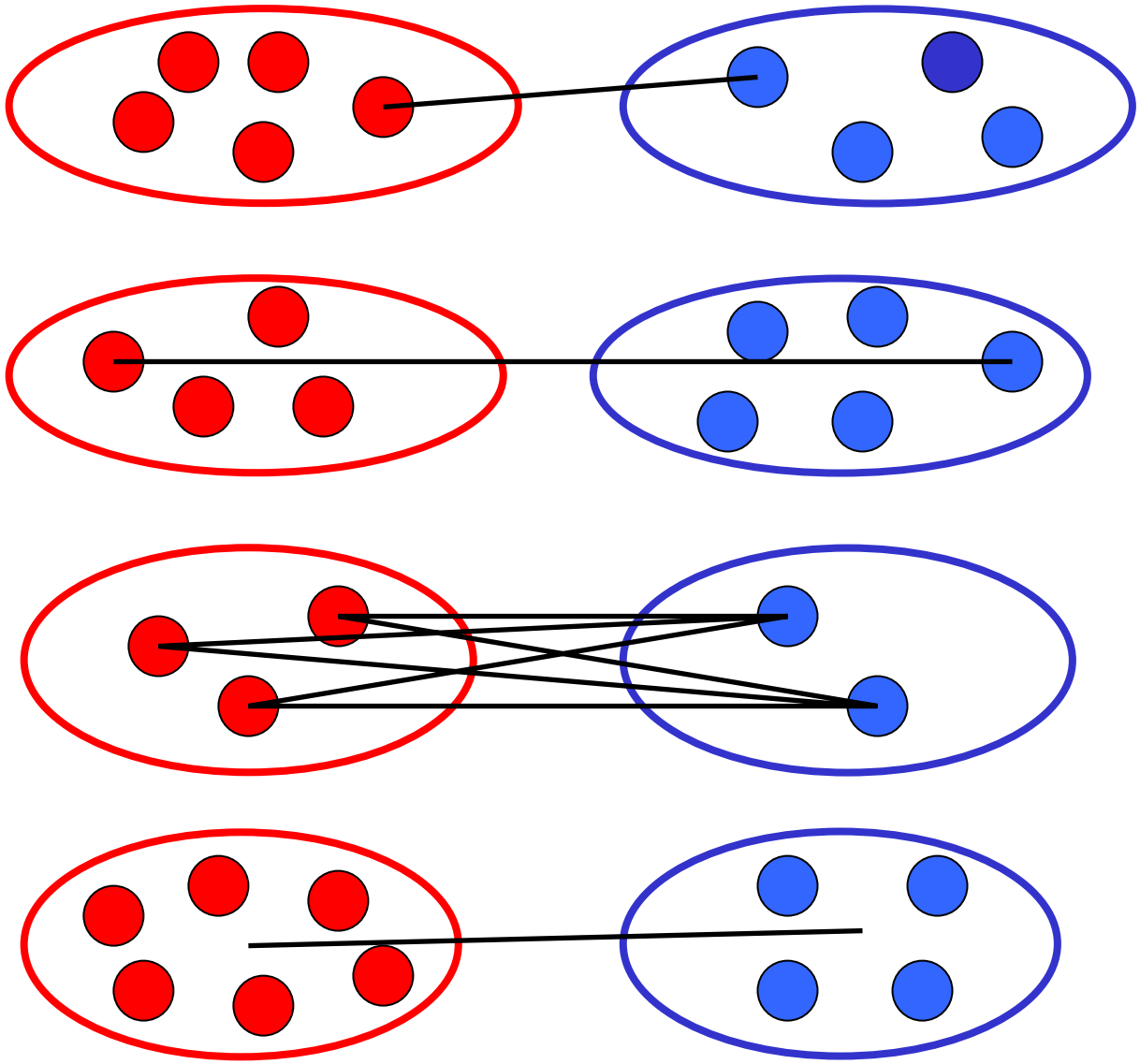


Distance^2 between (x,y) and (r,s) is $(x-r)^2 + (y-s)^2$

Distance^2 between $(p_1,\ldots,p_n)$ and $(q_1,\ldots,q_n)$ is:

$$(p_1-q_1)^2 + \cdots + (p_n-q_n)^2$$

Definitions of Inter-cluster Distances



Single Linkage
min

Complete Linkage
max

Average Linkage

Centroid

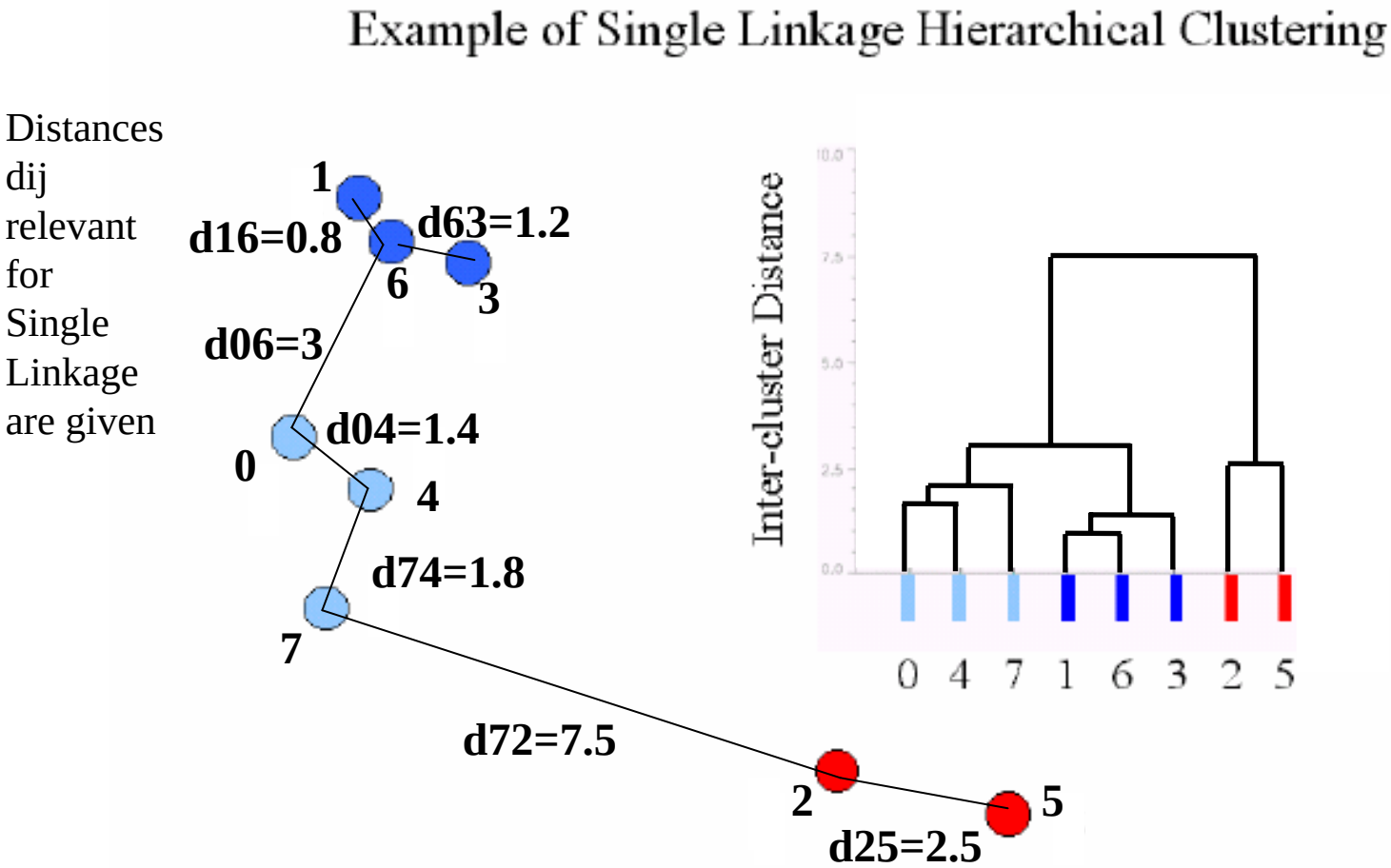
Hierarchical Agglomerative Clustering

1. Make choice of inter-cluster distance (and specify the distances (dissimilarities) between points).

2. Start with each point as a singleton cluster.

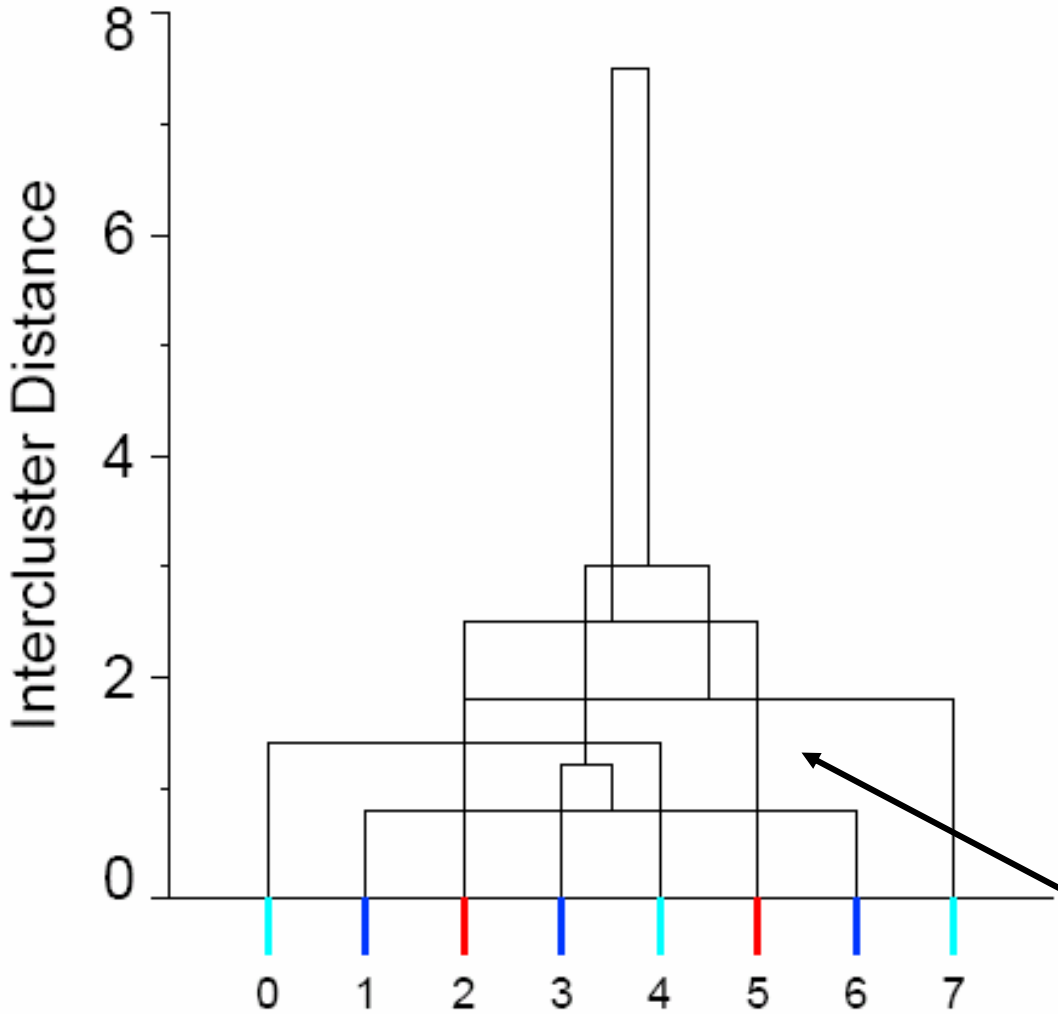
3. At each step, join the pair of clusters that have the smallest distance between them. Draw vertical line from top of each joined cluster up to height = distance between them, connect with horizontal line. Top of new joined cluster is midway between them.

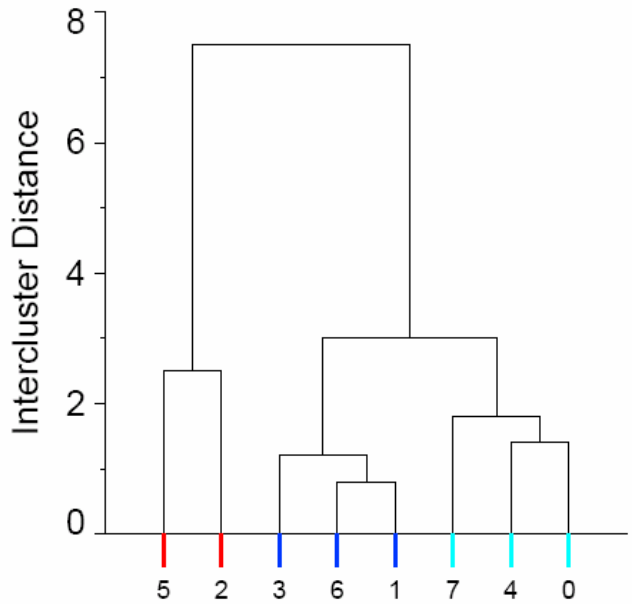
4. To avoid crossed lines, must have ordered the points so that at each step, joined clusters are next to each other (get unique dendrogram if specify rule for left-right order at each join)



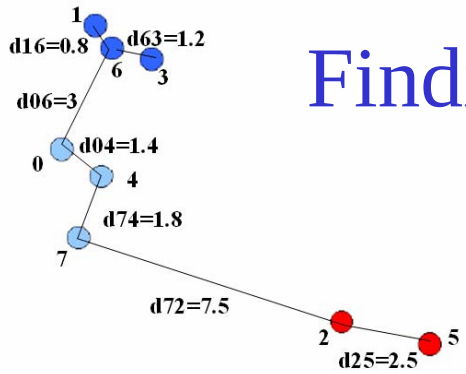
3. At each step, join the pair of clusters that have the smallest distance between them. Draw vertical line from top of each joined cluster up to height = distance between them, connect with horizontal line. Top of new joined cluster is midway between them.

Consequences of Improper Ordering of the Points



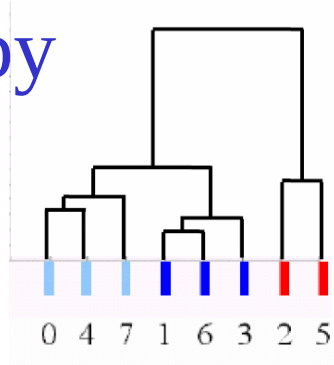


Totally Uninformative



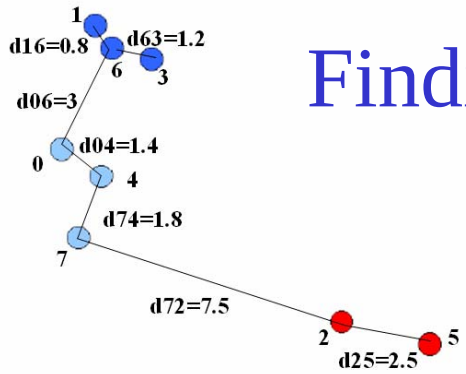
Finding an Admissible Ordering by "Left Sliding"

Single Linkage Example



starting with the ordering 0 1 2 3 4 5 6 7

first merge is 1-6 so form 0 16 2 3 4 5 7
next merge is 16-3 so form 0 163 2 4 5 7
next merge is 0-4 so form 04 163 2 5 7
next merge is 04-7 so form 047 163 2 5
next merge is 2-5 so form 047 163 25
next merge is 047-163 so form 047163 25
last merge is 047163-25 so corresponding ordering of the points is 0 4 7 1 6 3 2 5: with this ordering there will be no "crisscrossing" of lines when draw the dendrogram
Exercise: try starting with: 5 3 7 6 4 1 2 0

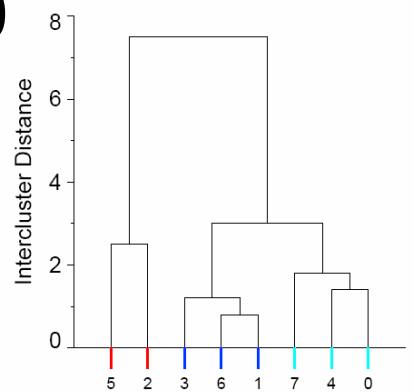


Finding an Admissible Ordering by "Left Sliding"

Single Linkage Example

starting with the ordering 5 3 7 6 4 1 2 0

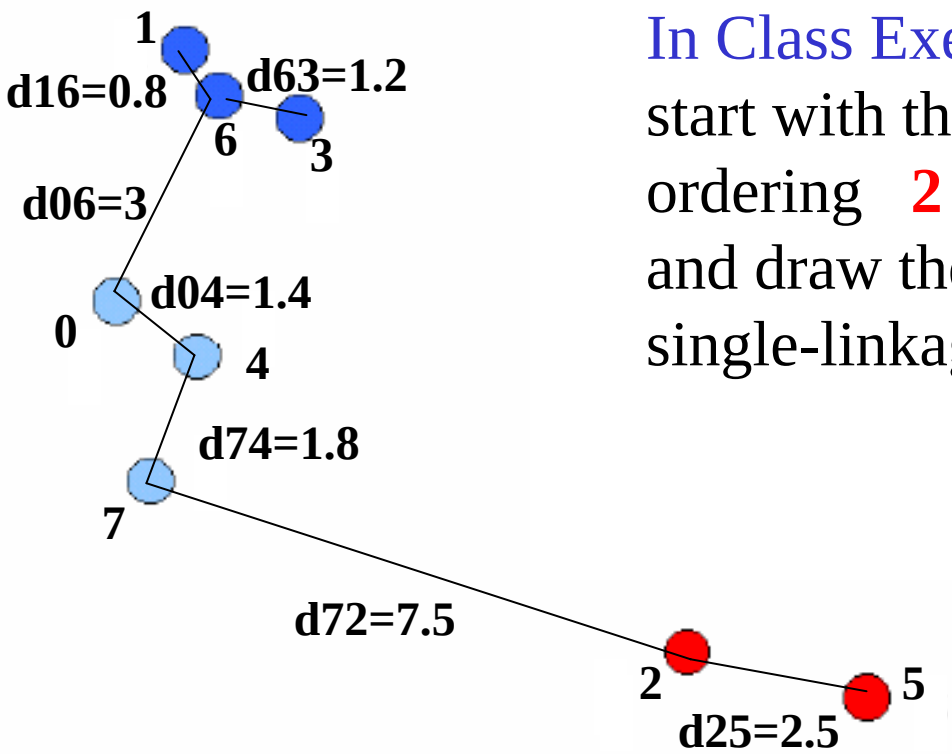
first merge is 1-6 so form 5 3 7 61 4 2 0
next merge is 16-3 so form 5 361 7 4 2 0
next merge is 0-4 so form 5 361 7 40 2
next merge is 04-7 so form 5 361 740 2
next merge is 2-5 so form 52 361 740
next merge is 047-163 so form 52 361740



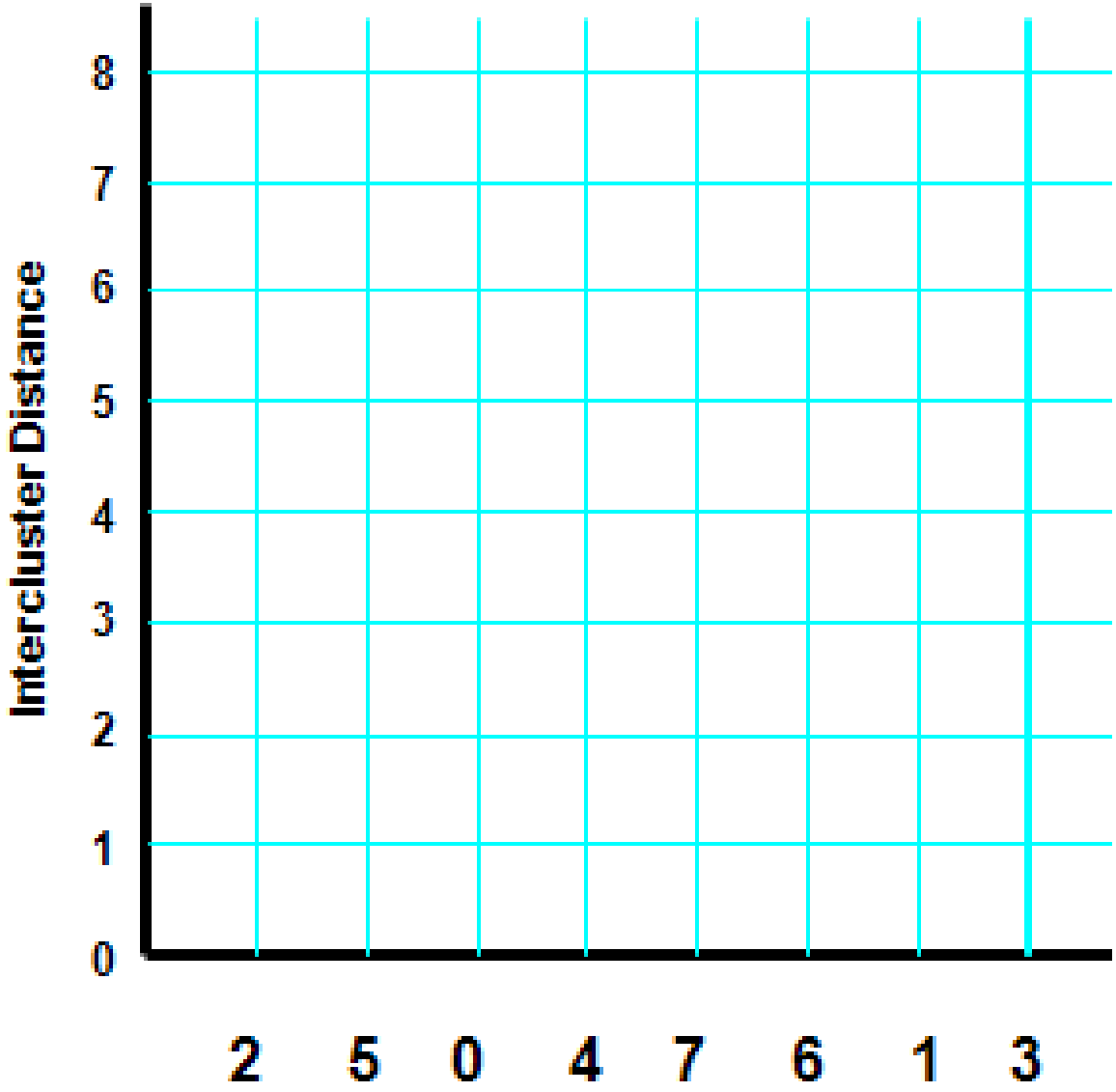
last merge is 52-361740 so corresponding ordering of the points is 5 2 3 6 1 7 4 0: with this ordering there will be no "crisscrossing" of lines when draw the dendrogram

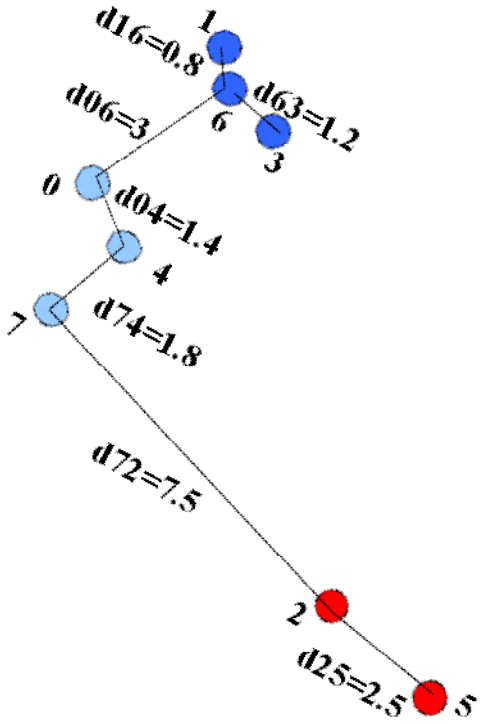
Example of Single Linkage Hierarchical Clustering

Distances dij relevant for Single Linkage are given

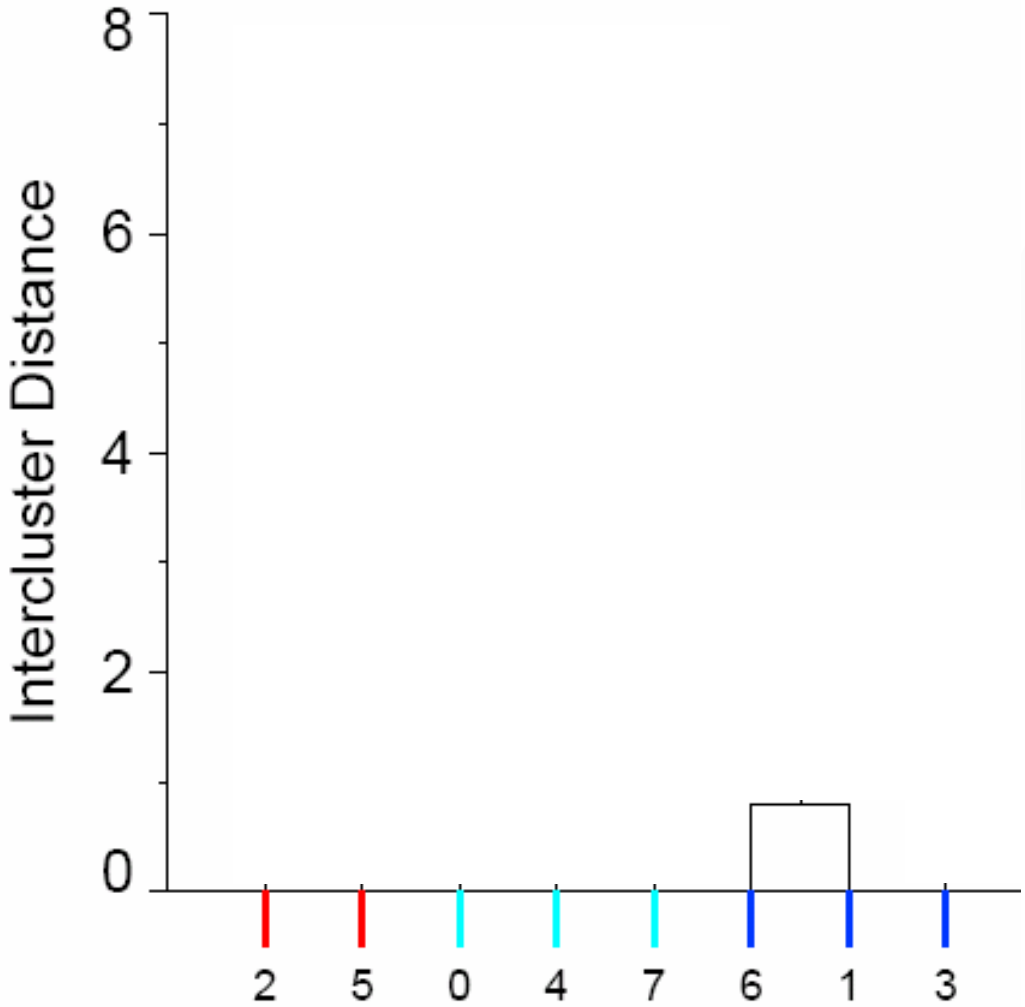


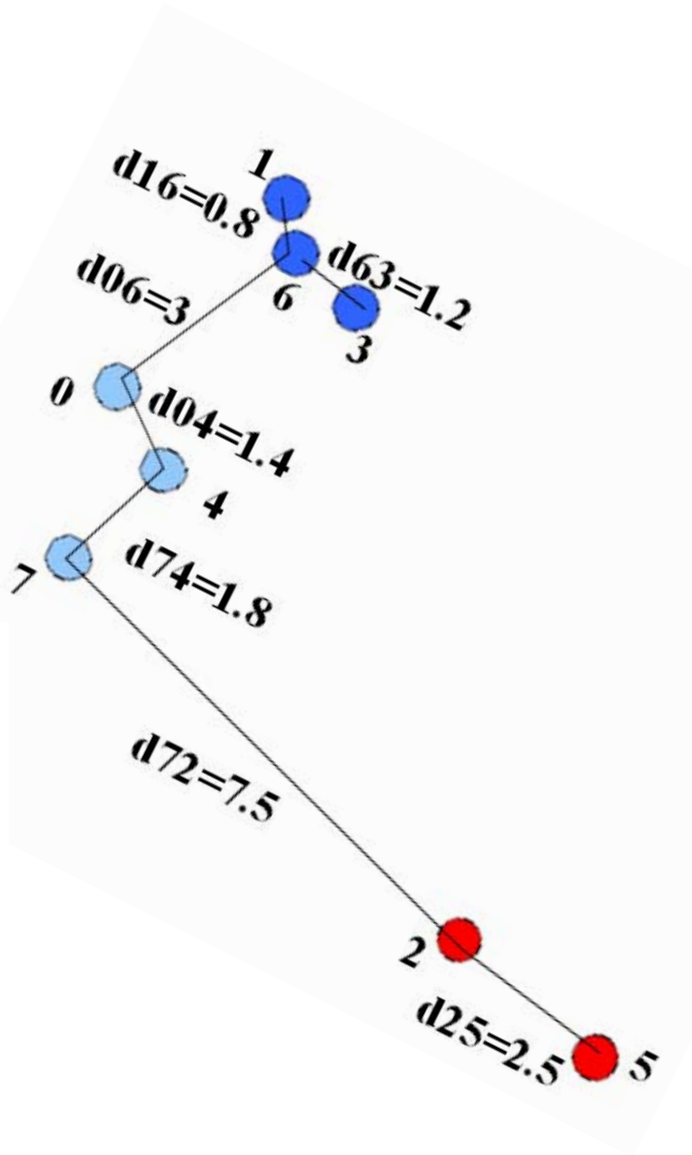
In Class Exercise:
start with the admissible ordering **2 5 0 4 7 6 1 3**
and draw the resulting single-linkage dendrogram

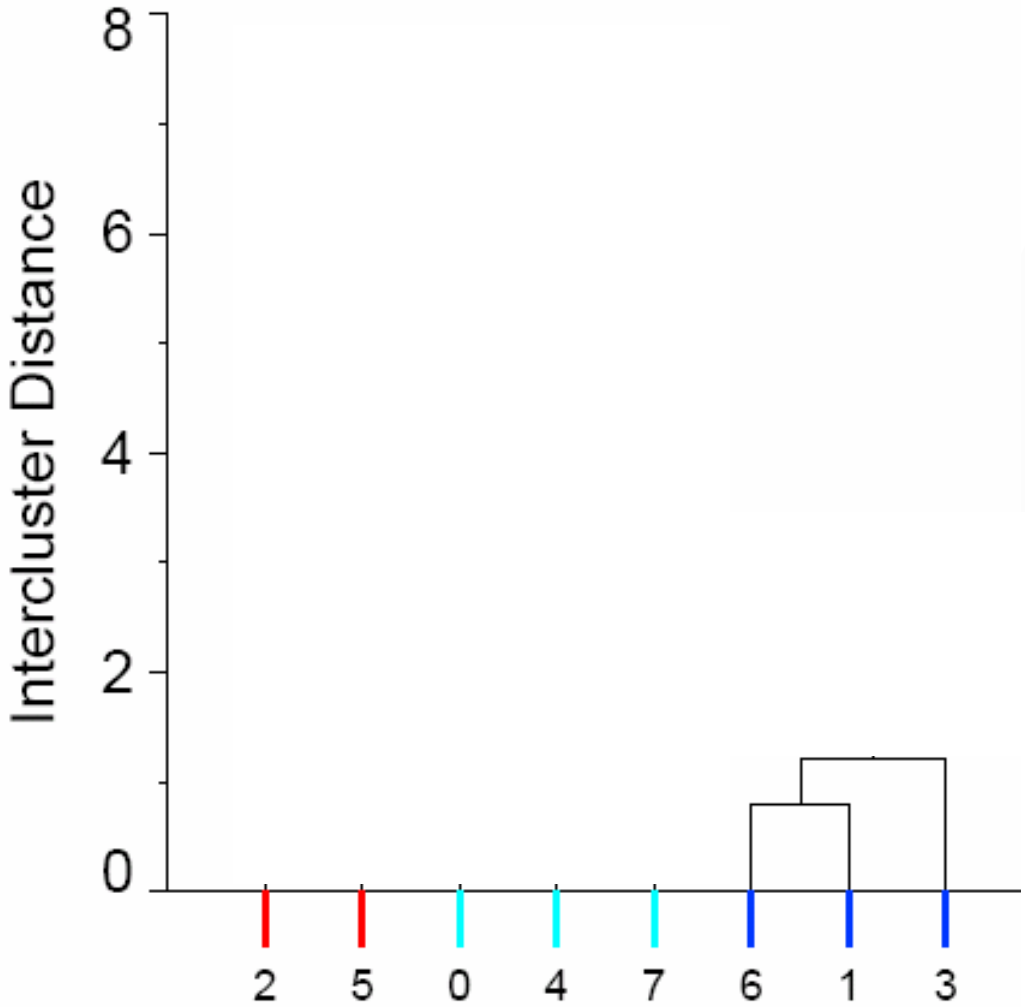


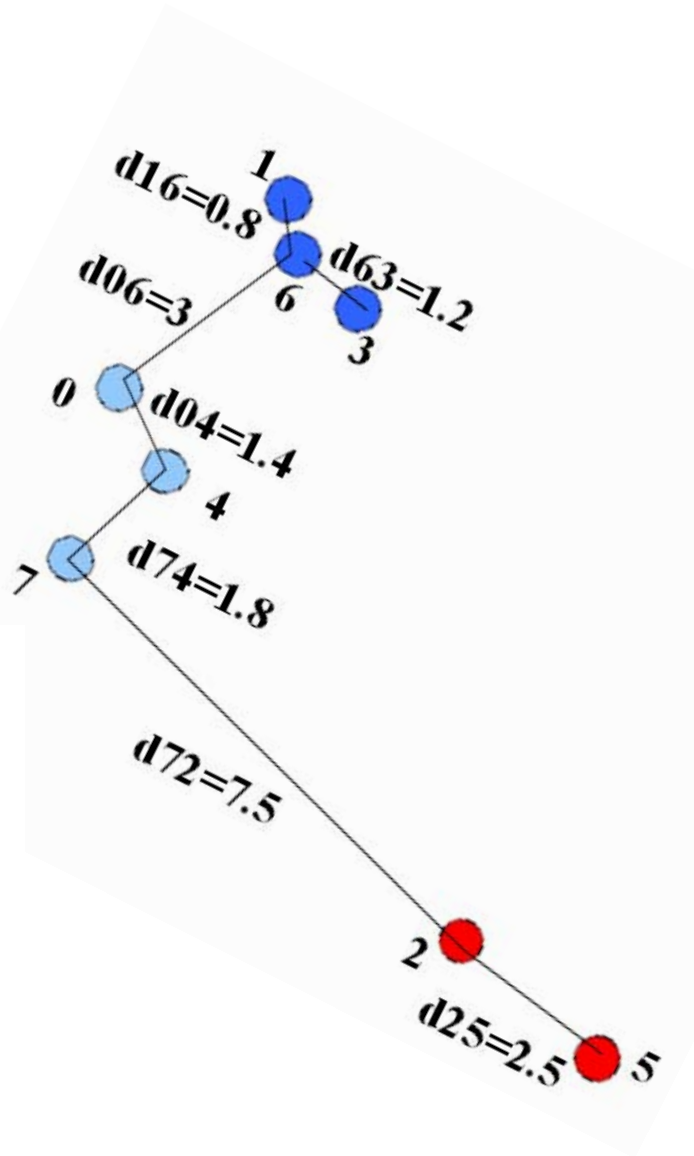


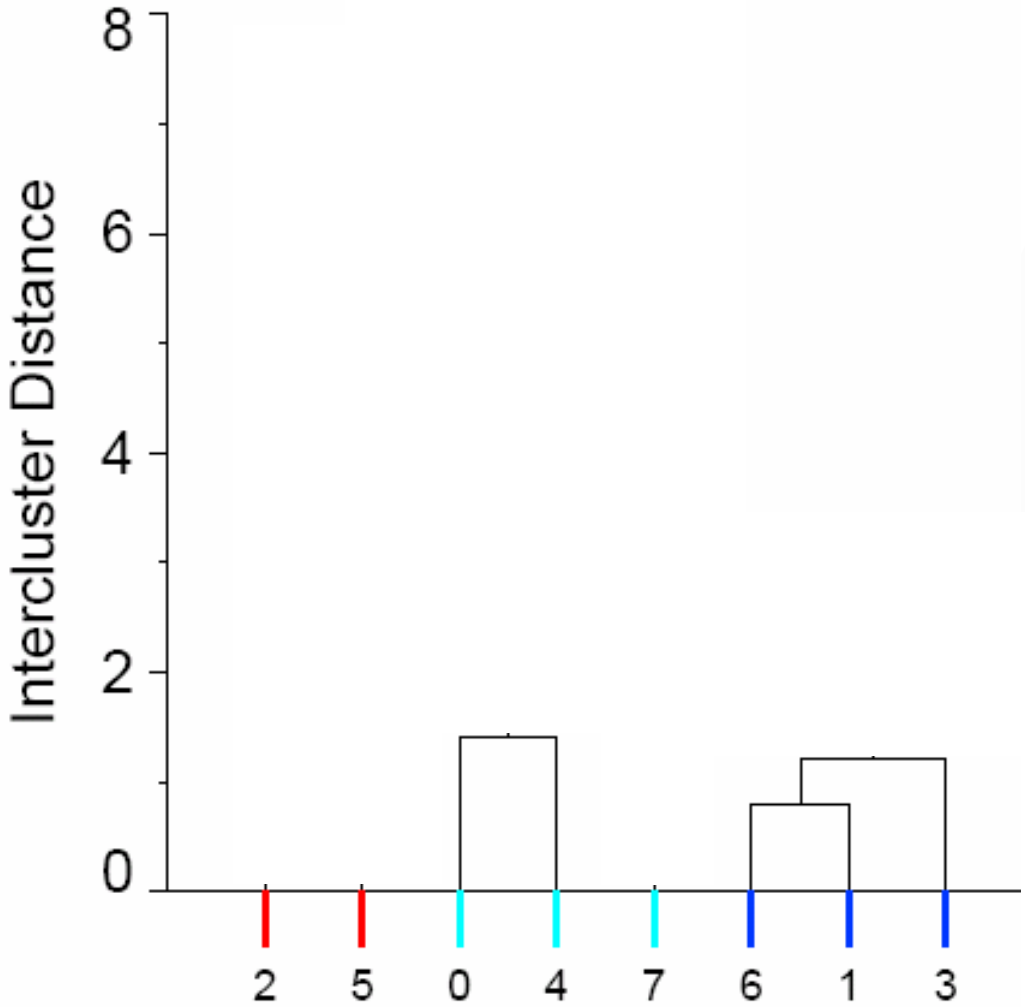
3. At each step, join the pair of clusters that have the smallest distance between them. Draw vertical line from top of each joined cluster up to height = distance between them, connect with horizontal line. Top of new joined cluster is midway between them.

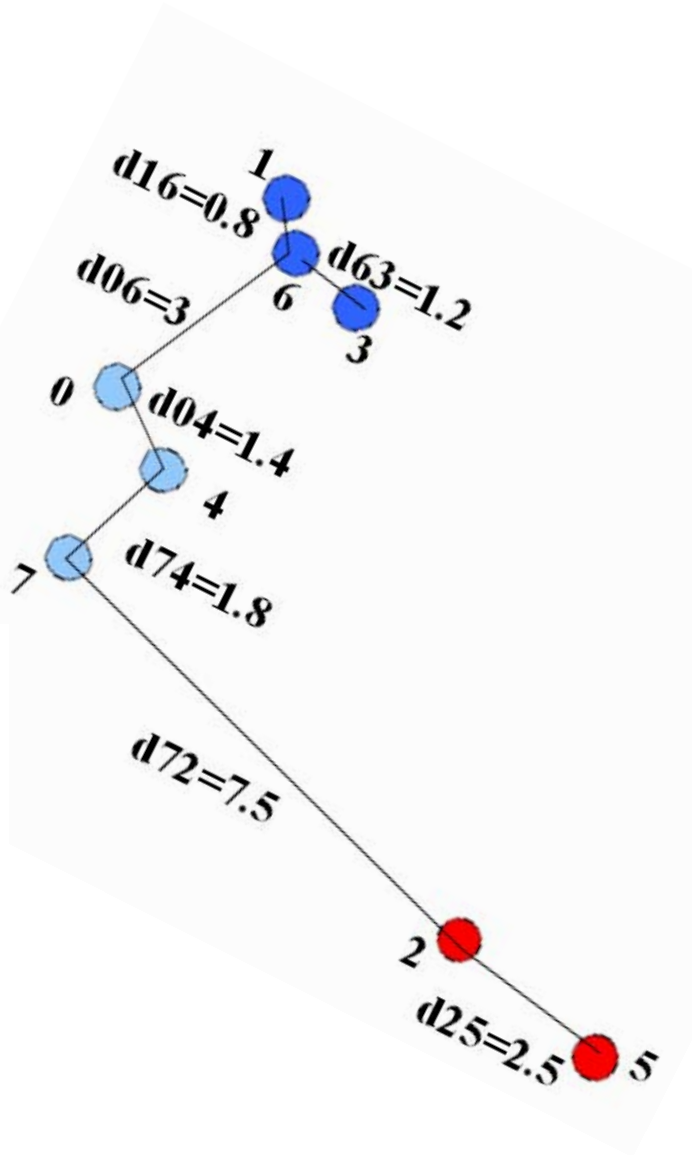

Intercluster Distance
8
6
4
2
0
2
5
0
4
7
6
1
3

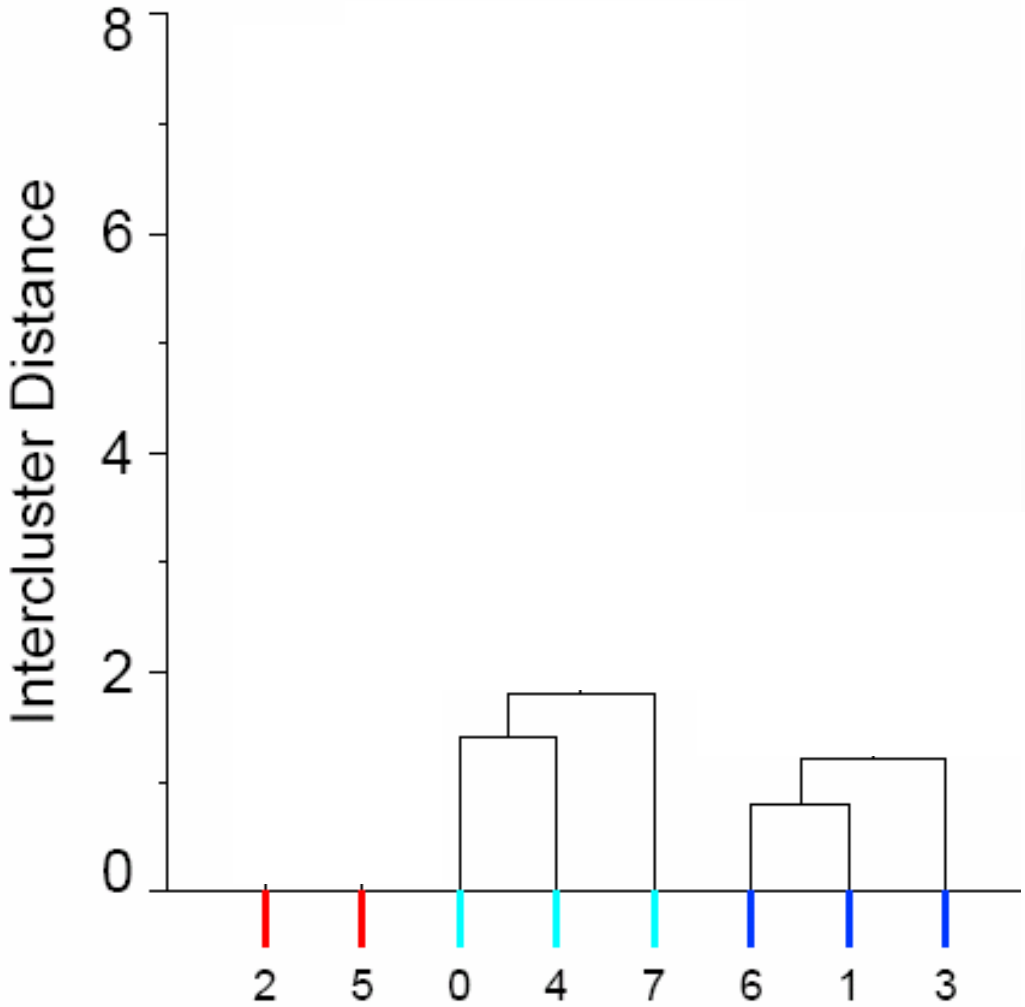

1
d16=0.8
d63=1.2
6
3
d06=3
0
d04=1.4
4
7
d74=1.8
d72=7.5
2
d25=2.5
5

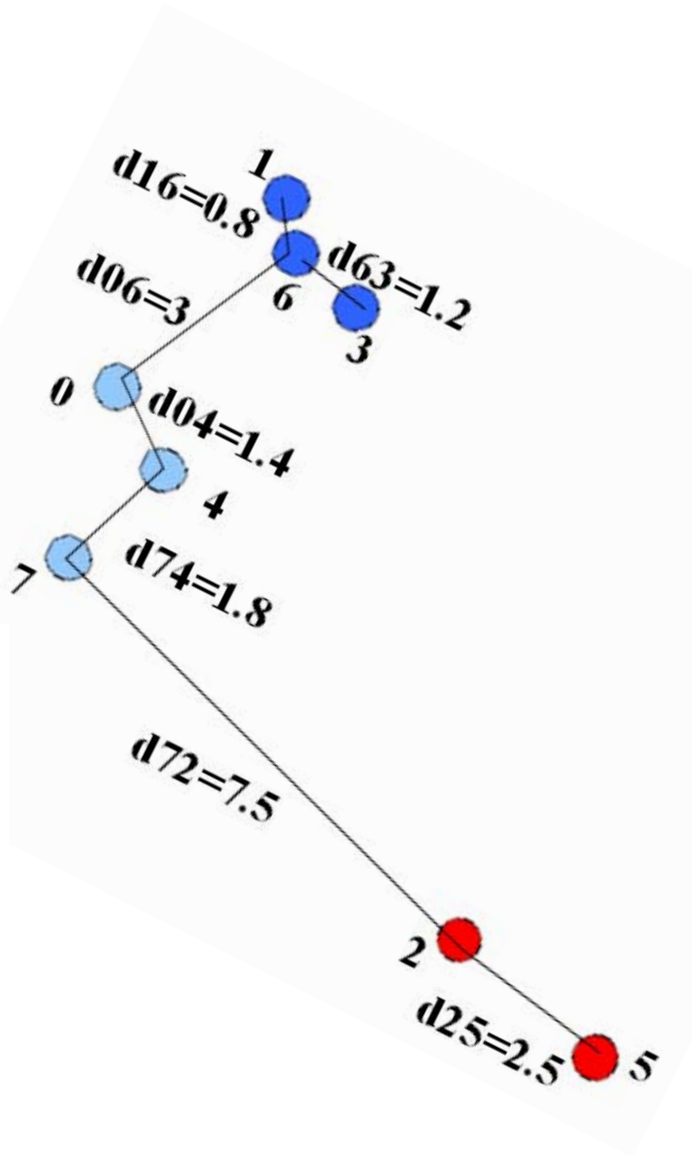

Intercluster Distance
8
6
4
2
0
2
5
0
4
7
6
1
3

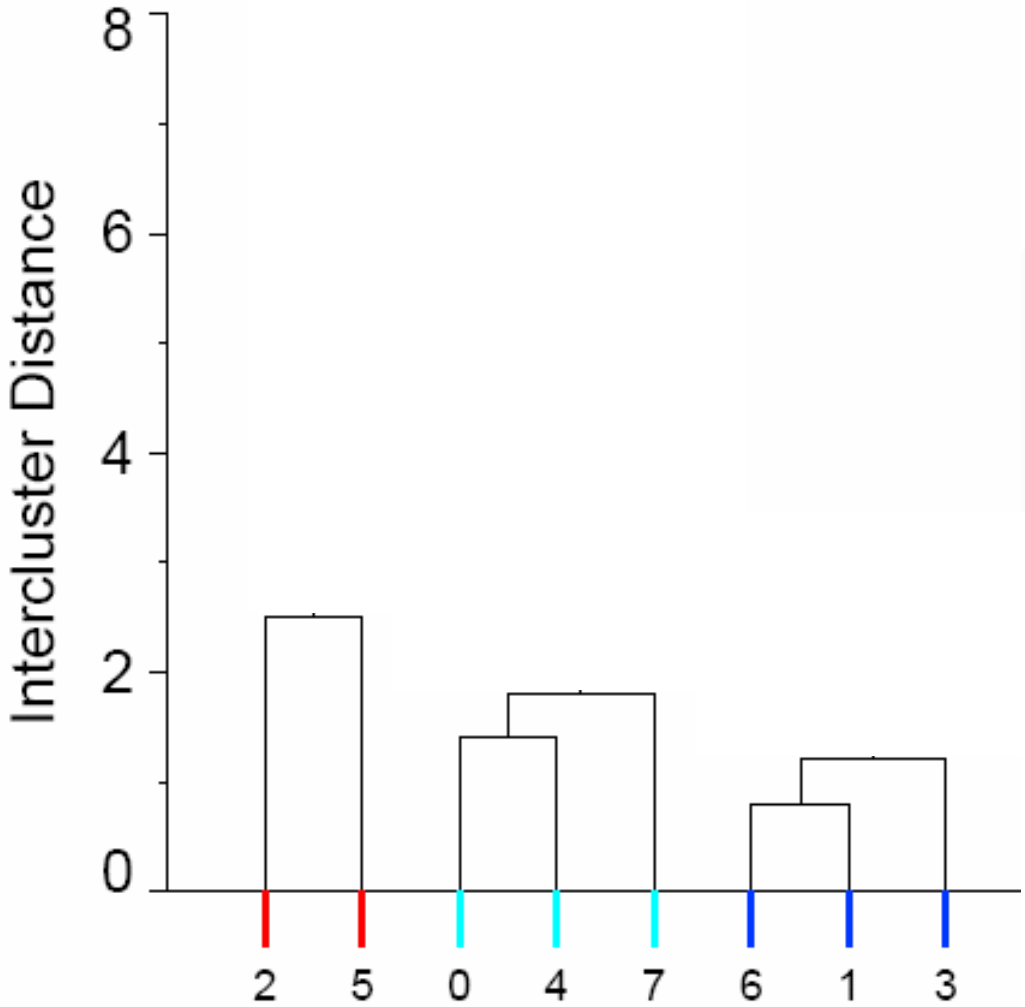

1
d16=0.8
d63=1.2
6
3
d06=3
0
d04=1.4
4
7
d74=1.8
d72=7.5
2
d25=2.5
5

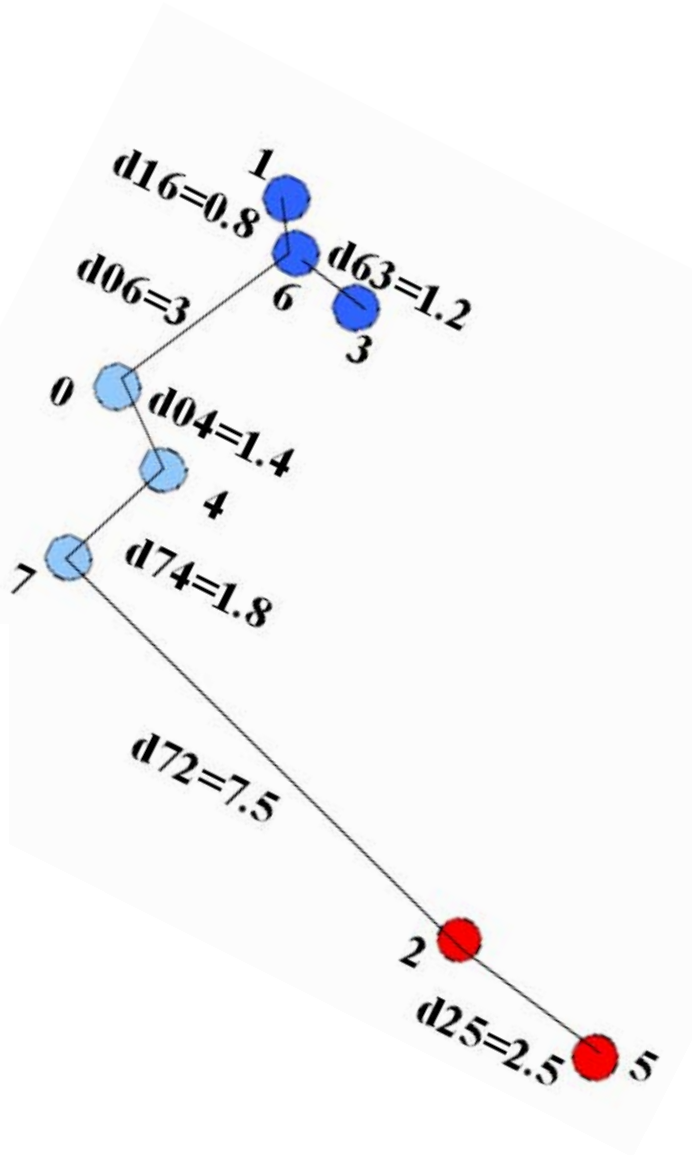

Intercluster Distance
8
6
4
2
0
2
5
0
4
7
6
1
3

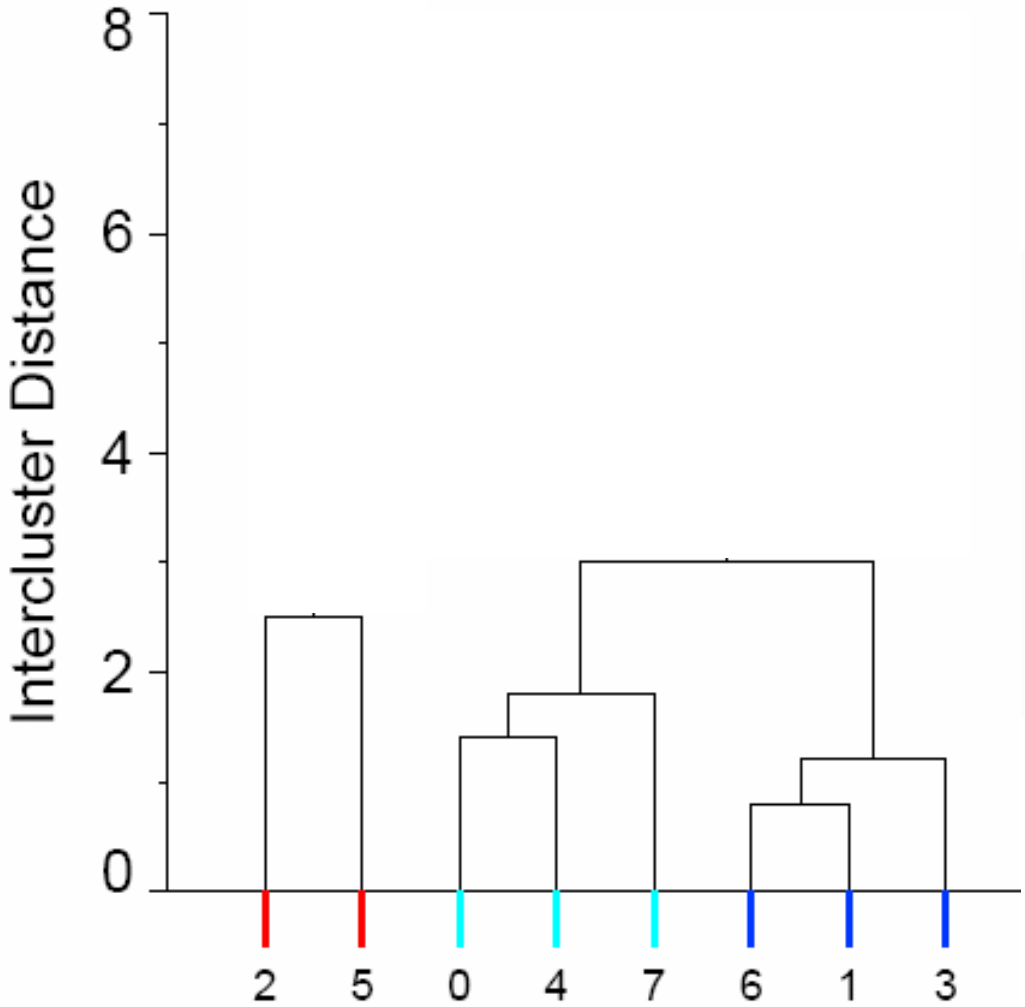

1
d16=0.8
6
d63=1.2
3
d06=3
0
d04=1.4
4
7
d74=1.8
d72=7.5
2
d25=2.5
5

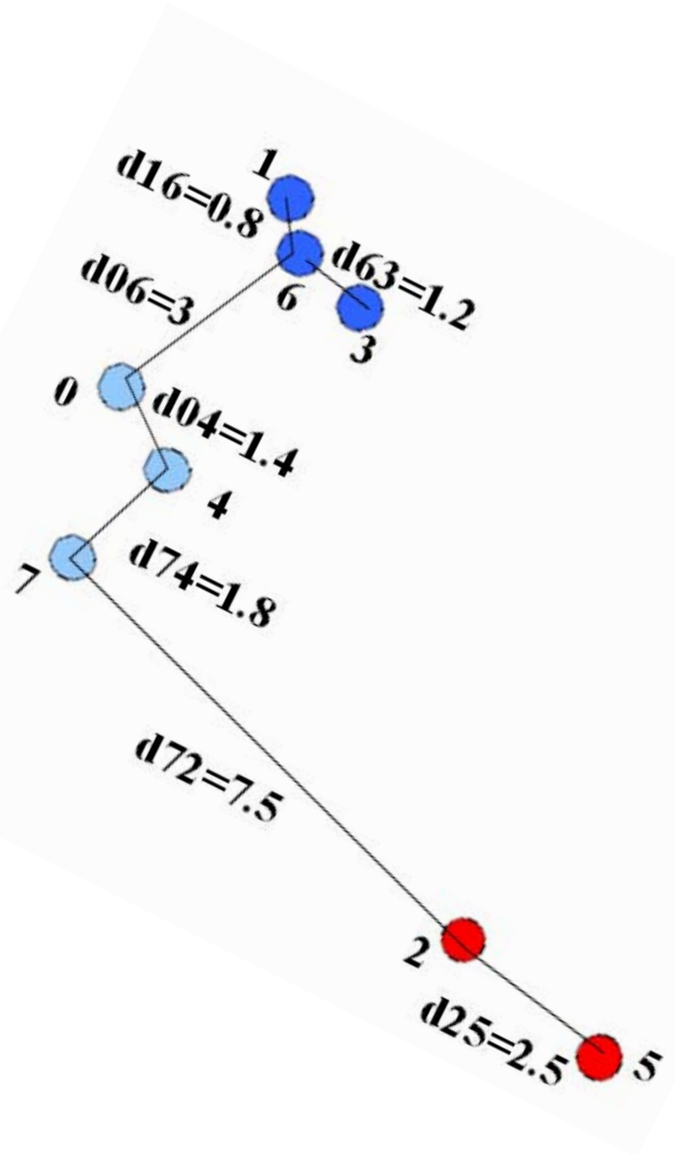

Intercluster Distance
8
6
4
2
0
2
5
0
4
7
6
1
3

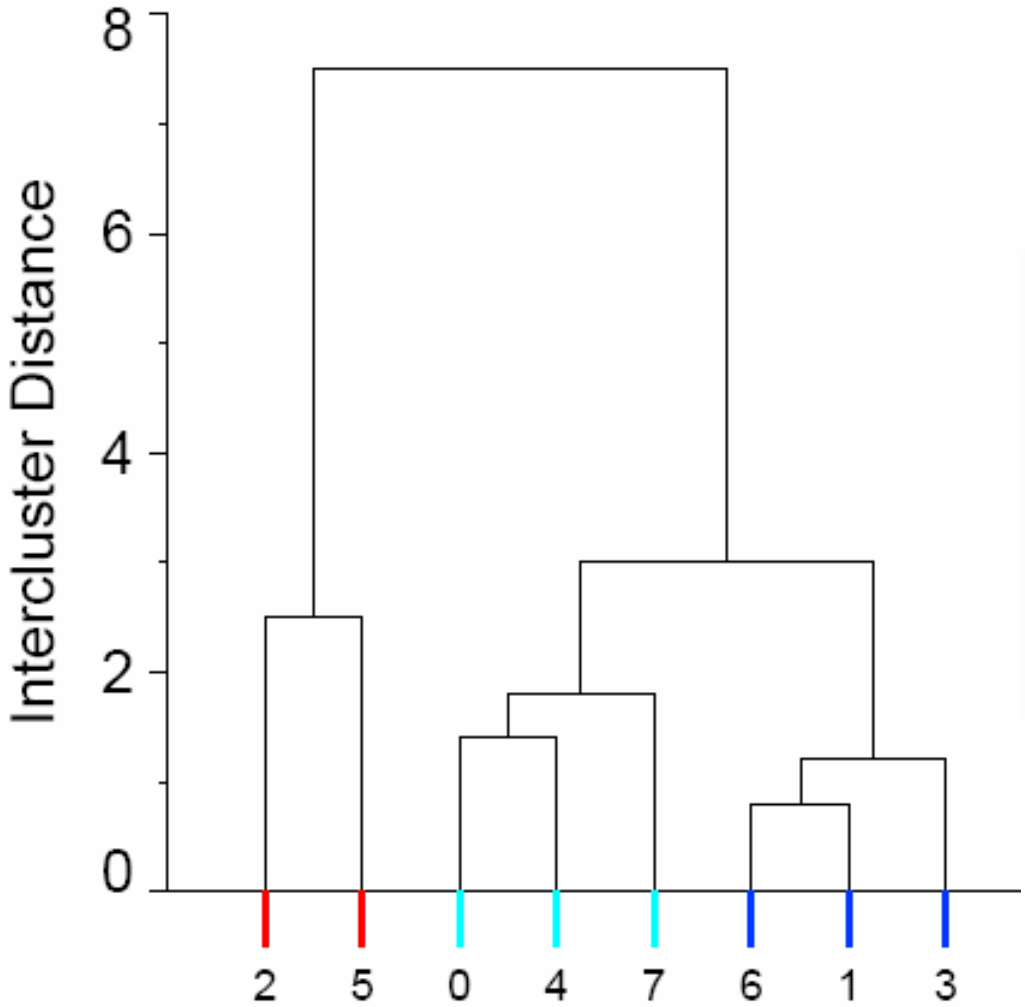

1
d16=0.8
d63=1.2
6
3
d06=3
0
d04=1.4
4
7
d74=1.8
d72=7.5
2
d25=2.5
5

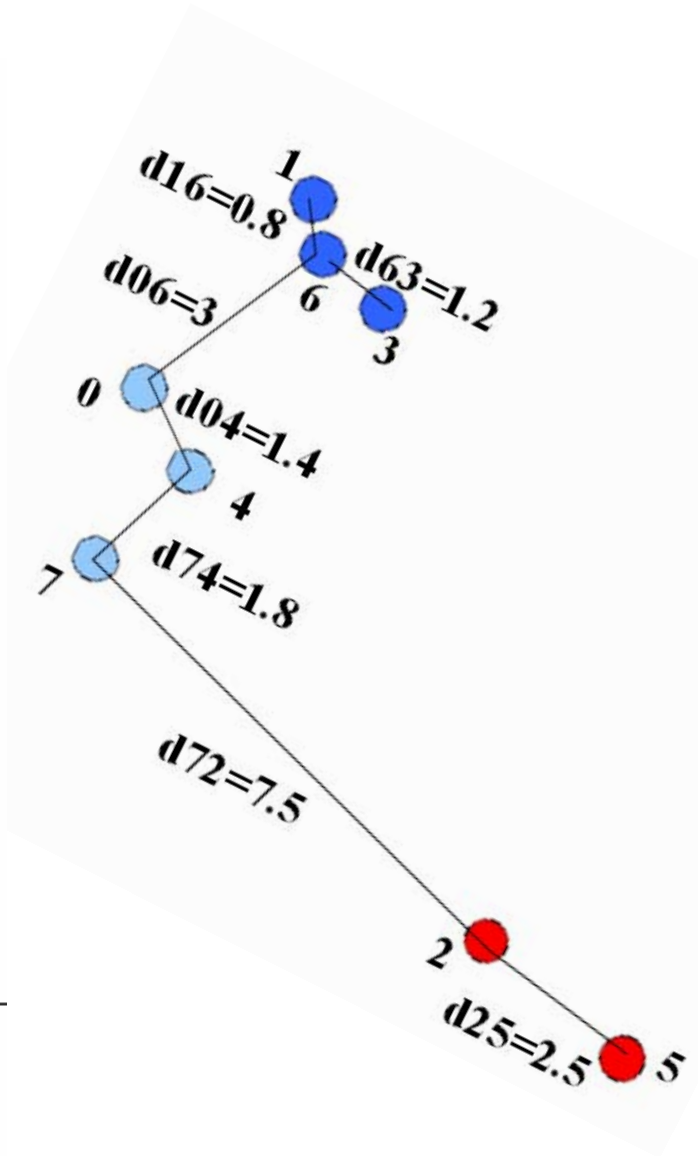

Intercluster Distance
8
6
4
2
0
2
5
0
4
7
6
1
3

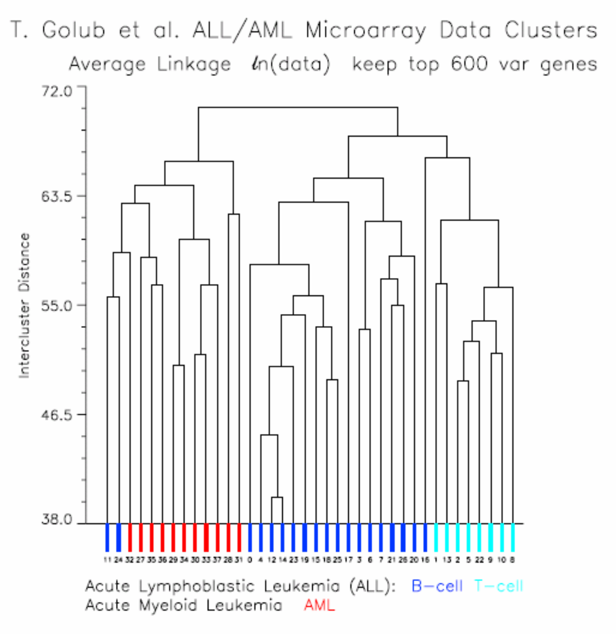

1
d16=0.8
d63=1.2
6
3
d06=3
0
d04=1.4
4
7
d74=1.8
d72=7.5
2
d25=2.5
5

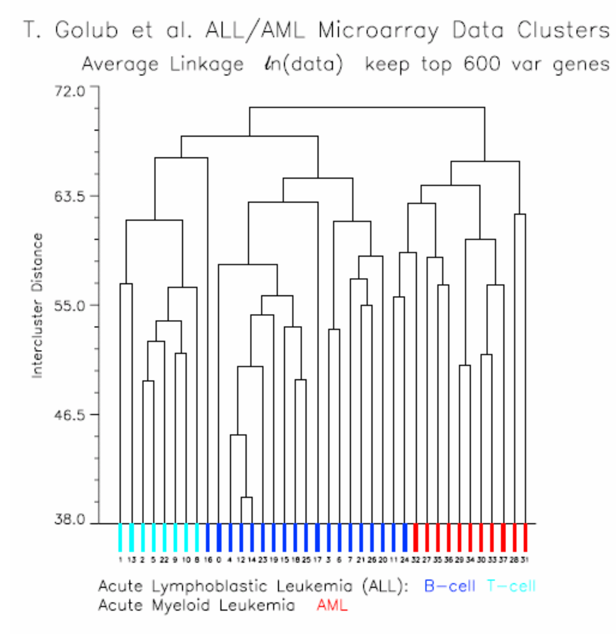

Intercluster Distance
8
6
4
2
0
2
5
0
4
7
6
1
3

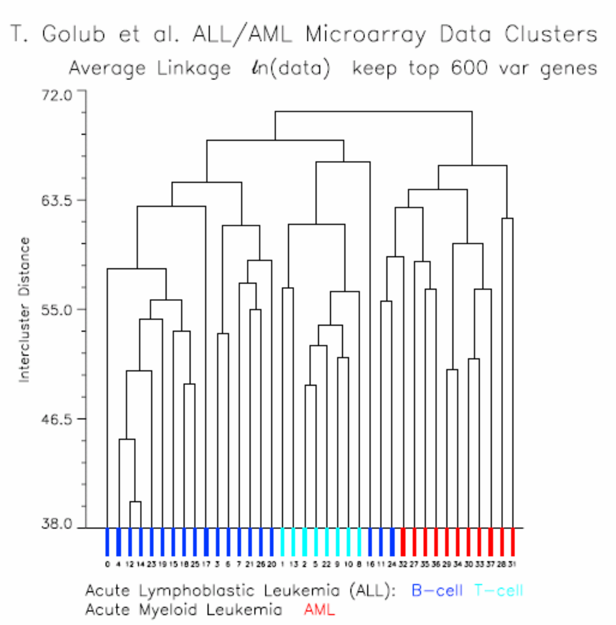

1
d16=0.8
6
d63=1.2
3
d06=3
0
d04=1.4
4
7
d74=1.8
d72=7.5
2
d25=2.5
5

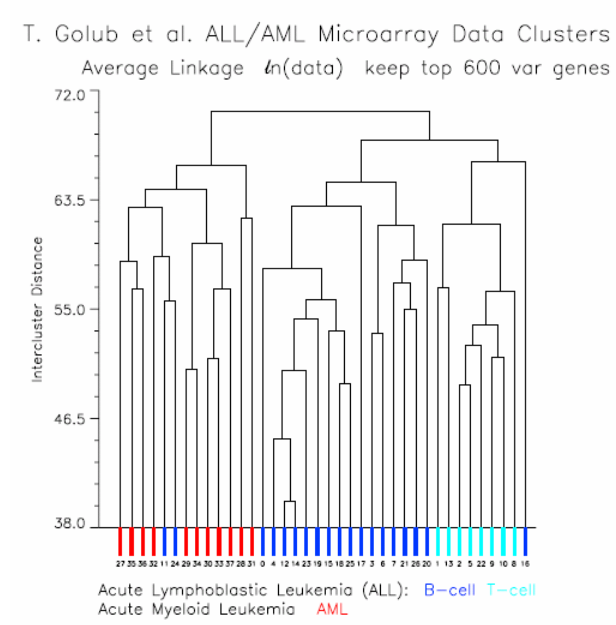

Intercluster Distance
8
6
4
2
0
2
5
0
4
7
6
1
3


1
d16=0.8
6
d63=1.2
3
d06=3
0
d04=1.4
4
7
d74=1.8
d72=7.5
2
d25=2.5
5


T. Golub et al. ALL/AML Microarray Data Clusters
Average Linkage ln(data) keep top 600 var genes
72.0
63.5
55.0
46.5
38.0
Intercluster Distance
Acute Lymphoblastic Leukemia (ALL): B−cell T−cell
Acute Myeloid Leukemia AML


T. Golub et al. ALL/AML Microarray Data Clusters
Average Linkage ln(data) keep top 600 var genes
72.0
63.5
55.0
46.5
38.0
Intercluster Distance
Acute Lymphoblastic Leukemia (ALL): B−cell T−cell
Acute Myeloid Leukemia AML


T. Golub et al. ALL/AML Microarray Data Clusters
Average Linkage ln(data) keep top 600 var genes
72.0
63.5
55.0
46.5
38.0
Intercluster Distance
Acute Lymphoblastic Leukemia (ALL): B−cell T−cell
Acute Myeloid Leukemia AML

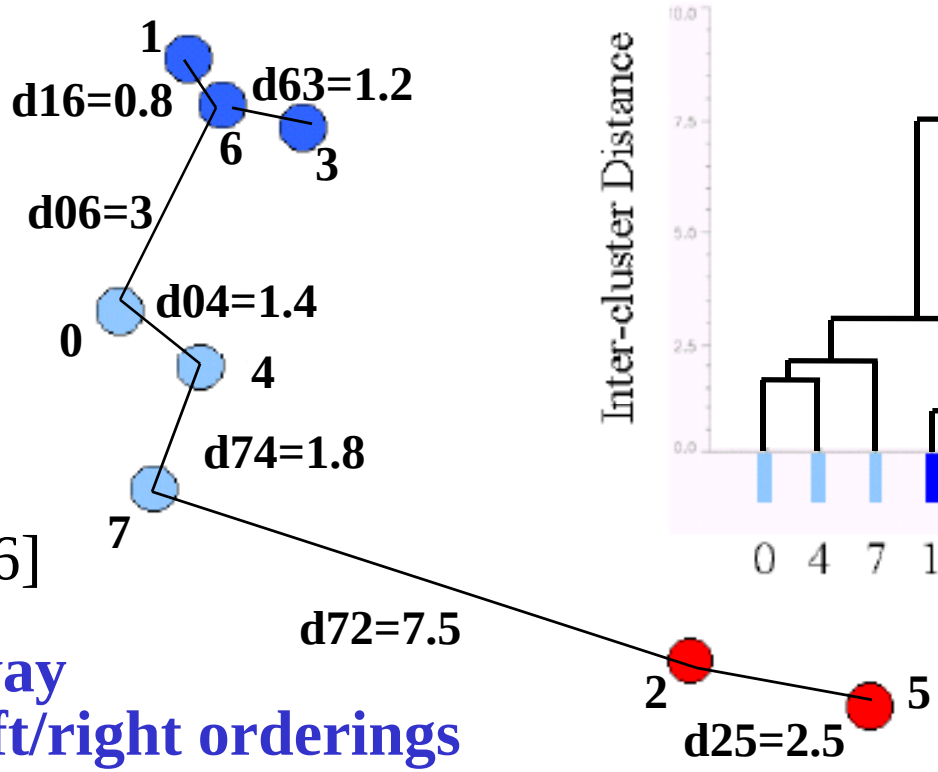

T. Golub et al. ALL/AML Microarray Data Clusters
Average Linkage ln(data) keep top 600 var genes
72.0
63.5
55.0
46.5
38.0
Intercluster Distance
Acute Lymphoblastic Leukemia (ALL): B−cell T−cell
Acute Myeloid Leukemia AML

Clustering: How Do They Make Those Dendrograms and Heat Maps – Outline

- Definition of unsupervised clustering
- Dendrogram construction by hierarchical agglomerative clustering given specified inter-cluster and inter-point distance measures
- Uniqueness of the dendrogram if an unambiguous choice of left/right ordering is specified for each join of two clusters in the dendrogram construction
- Dependence of the clustering dendrogram on the definition of inter-cluster distance.
- Additional examples, Heat map construction
- Briefly noting other methods for clustering and data visualization
- The difference between exploratory and supervised clustering

Example of Single Linkage Hierarchical Clustering

Distances dij relevant for Single Linkage are given



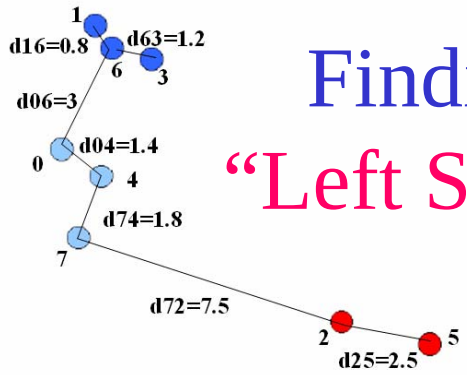
cf. Eisen et al. [6]

Example of a way of specifying left/right orderings

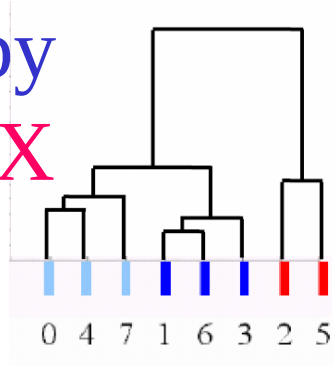
Examples of Point Rankings

left 0 7 1 4 6 2 3 5 (Gene X) At each join of two subclusters, place the one
left 5 2 7 4 0 3 6 1 (Gene Y) containing the "leftmost" point number on the left.
left 5 2 3 6 1 7 4 0 (Gene Z) Different orderings give rearranged dendrograms.
left 1 2 3 0 4 7 5 6 (Gene G) Could also use average gene expression (Eisen et al.)

Dendrogram is indep. of original order given unique left/right orderings



Finding an Admissible Ordering by "Left Sliding" Left/Right by Gene X



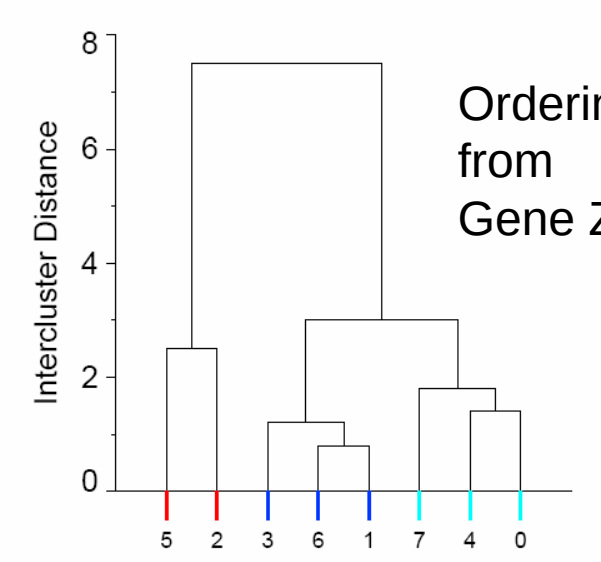
Single Linkage Example

starting with the ordering 6 5 7 4 3 2 1 0

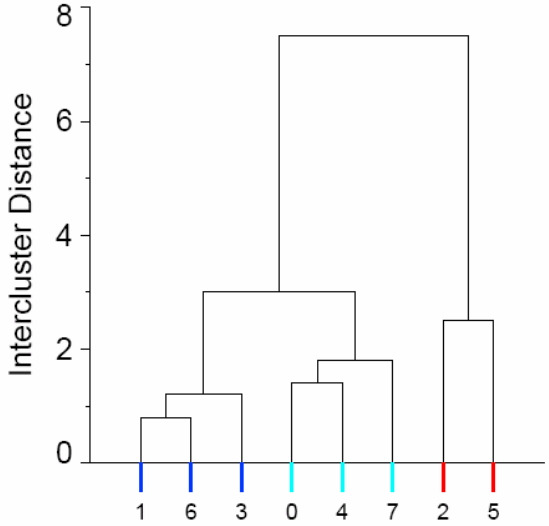
Gene X left / right point ranking = 0 7 1 4 6 2 3 5

first merge is 1-6 so form	16 5 7 4 3 2 0
next merge is 16-3 so form	163 5 7 4 2 0
next merge is 0-4 so form	163 5 7 04 2
next merge is 04-7 so form	163 5 047 2
next merge is 2-5 so form	163 25 047
next merge is 047-163 so form	047163 25
last merge is 047163-25 so get	**0 4 7 1 6 3 2 5**

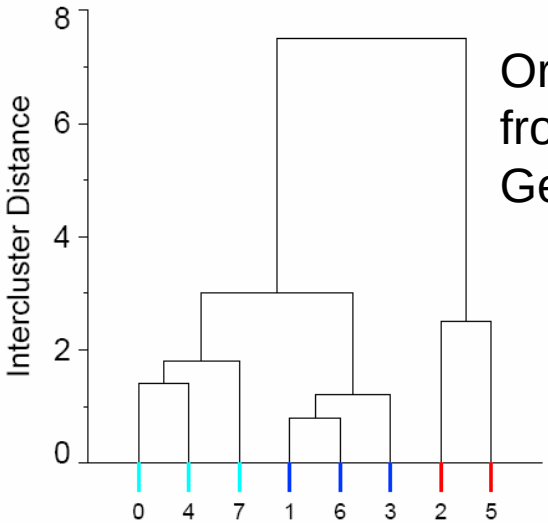
Exercise: try starting with: 3 5 0 2 6 7 1 4



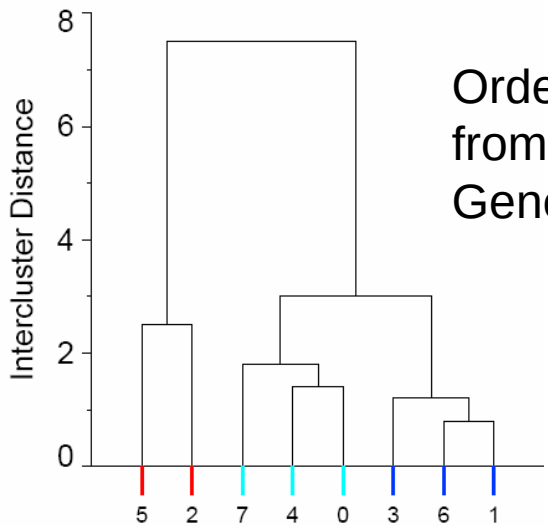
Ordering from Gene Z



Ordering from Gene G



Ordering from Gene X



Ordering from Gene Y

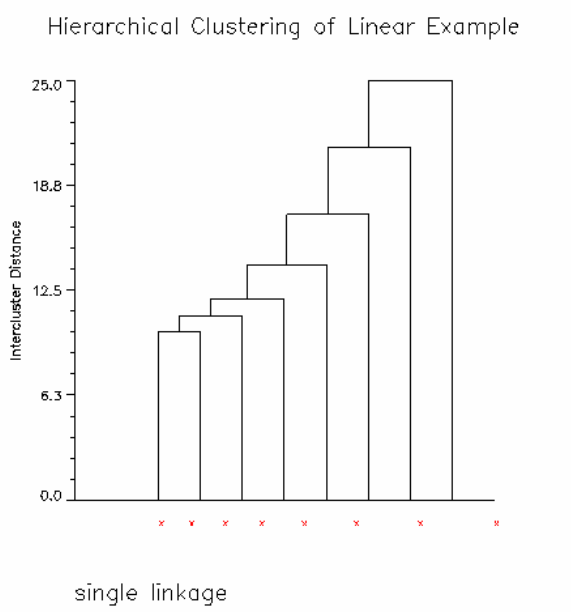
Clustering: How Do They Make Those Dendrograms and Heat Maps – Outline

- Definition of unsupervised clustering
- Dendrogram construction by hierarchical agglomerative clustering given specified inter-cluster and inter-point distance measures
- Uniqueness of the dendrogram if an unambiguous choice of left/right ordering is specified for each join of two clusters in the dendrogram construction
- Dependence of the clustering dendrogram on the definition of inter-cluster distance.
- Additional examples, Heat map construction
- Briefly noting other methods for clustering and data visualization
- The difference between exploratory and supervised clustering

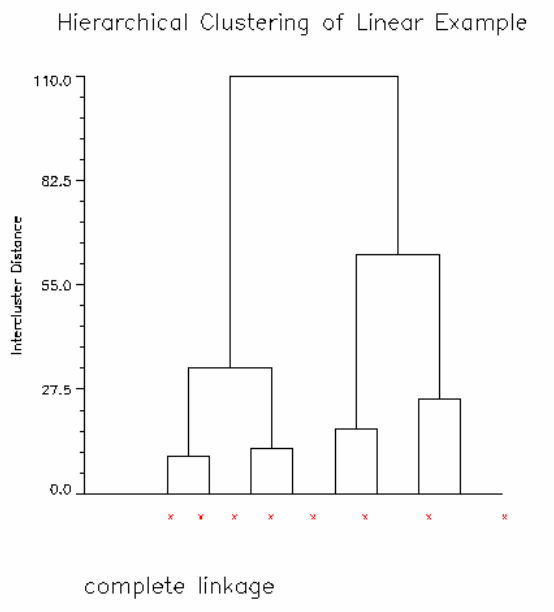
Data for 1-Dimensional Clustering Example

v1--v2---v3----v4-----v5------v6--------v7----------v8 points on line

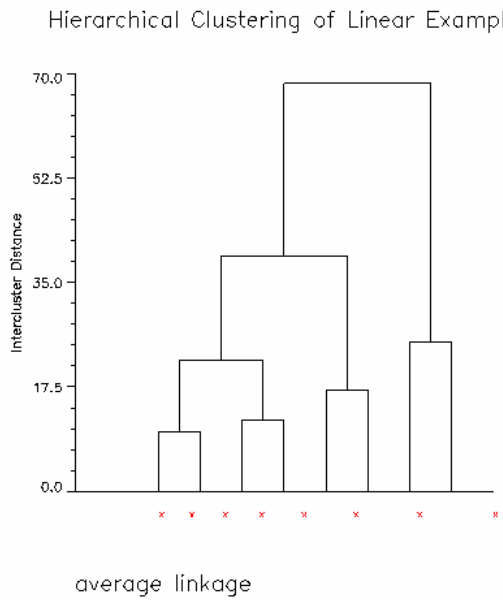
10 11 12 14 17 21 25 distances(V_{i+1},V_i)



single linkage



complete linkage



average linkage

Warning: hierarchical agglomerative clustering programs will always return dendrograms, even for random data

should use any information available about data to judge usefulness of clustering, as well as compactness and separation of clusters; use, e.g., PCA to visualize data and clusters.

Correlation Distances

Suppose each point is a vector of log(fold changes), e.g., log(treated gene expression level / control expression level).

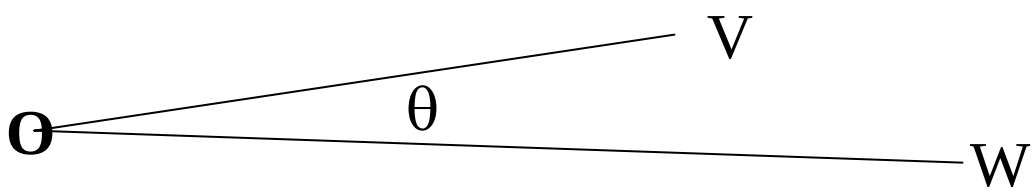
Two such vectors V, W are well (positively) correlated if when an entry in V is > 0, the corresponding entry in W is > 0, and when V_j is < 0 then W_j is < 0.

Two such vectors V, W are highly negatively correlated if when an entry in V is > 0, the corresponding entry in W is < 0, and when V_j is < 0 then W_j is > 0.

Two common choices of correlation distance are

d(V,W) = 1 – cos(angle between the vectors V and W)

d(V,W) = 1 – | cos(angle between the vectors V and W) |



Clustering: How Do They Make Those Dendrograms and Heat Maps – Outline

- Definition of unsupervised clustering
- Dendrogram construction by hierarchical agglomerative clustering given specified inter-cluster and inter-point distance measures
- Uniqueness of the dendrogram if an unambiguous choice of left/right ordering is specified for each join of two clusters in the dendrogram construction
- Dependence of the clustering dendrogram on the definition of inter-cluster distance.
- Additional examples, Heat map construction
- Briefly noting other methods for clustering and data visualization
- The difference between exploratory and supervised clustering

Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring

T. R. Golub,[1,2*†] D. K. Slonim,[1†] P. Tamayo,[1] C. Huard,[1] M. Gaasenbeek,[1] J. P. Mesirov,[1] H. Coller,[1] M. L. Loh,[2] J. R. Downing,[3] M. A. Caligiuri,[4] C. D. Bloomfield,[4] E. S. Lander[1,5*]

Although cancer classification has improved over the past 30 years, there has been no general approach for identifying new cancer classes (class discovery) or for assigning tumors to known classes (class prediction). Here, a generic approach to cancer classification based on gene expression monitoring by DNA microarrays is described and applied to human acute leukemias as a test case. A class discovery procedure automatically discovered the distinction between acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) without previous knowledge of these classes. An automatically derived class predictor was able to determine the class of new leukemia cases. The results demonstrate the feasibility of cancer classification based solely on gene expression monitoring and suggest a general strategy for discovering and predicting cancer classes for other types of cancer, independent of previous biological knowledge.

The challenge of cancer treatment has been to target specific therapies to pathogenetically distinct tumor types, to maximize efficacy and minimize toxicity. Improvements in cancer classification have thus been central to advances in cancer treatment. Cancer classification has been based primarily on morphological appearance of the tumor, but this has serious limitations. Tumors with similar histopathological appearance can follow significantly different clinical courses and show different responses to therapy. In a few cases, such clinical heterogeneity has been explained by dividing morphologically similar tumors into subtypes with distinct pathogeneses. Key examples include the subdivision of acute leukemias, non-Hodgkin's lymphomas, and childhood "small round blue cell tumors" [tumors with variable response to chemotherapy (*1*) that are now molecularly subclassified into neuroblastomas, rhabdomyosarcoma, Ewing's sarcoma, and other types (*2*)]. For many more tumors, important subclasses are likely to exist but have yet to be defined by molecular markers. For example, prostate cancers of identical grade can have widely variable clinical courses, from indolence over decades to explosive growth causing rapid patient death. Cancer classification has been difficult in part because it has historically relied on specific biological insights, rather than systematic and unbiased approaches for recognizing tumor subtypes. Here we describe such an approach based on global gene expression analysis.

We divided cancer classification into two challenges: class discovery and class prediction. Class discovery refers to defining previously unrecognized tumor subtypes. Class prediction refers to the assignment of particular tumor samples to already-defined classes, which could reflect current states or future outcomes.

We chose acute leukemias as a test case. Classification of acute leukemias began with the observation of variability in clinical outcome (*3*) and subtle differences in nuclear morphology (*4*). Enzyme-based histochemical analyses were introduced in the 1960s to demonstrate that some leukemias were periodic acid-Schiff positive, whereas others were myeloperoxidase positive (*5*). This provided the first basis for classification of acute leukemias into those arising from lymphoid precursors (acute lymphoblastic leukemia, ALL) or from myeloid precursors (acute myeloid leukemia, AML). This classification was further solidified by the development in the 1970s of antibodies recognizing either lymphoid or myeloid cell surface molecules (*6*). Most recently, particular subtypes of acute leukemia have been found to be associated with specific chromosomal translocations—for example, the t(12;21)(p13;q22) translocation occurs in 25% of patients with ALL, whereas the t(8;21)(q22;q22) occurs in 15% of patients with AML (*7*).

Although the distinction between AML and ALL has been well established, no single test is currently sufficient to establish the diagnosis. Rather, current clinical practice involves an experienced hematopathologist's interpretation of the tumor's morphology, histochemistry, immunophenotyping, and cytogenetic analysis, each performed in a separate, highly specialized laboratory. Although usually accurate, leukemia classification remains imperfect and errors do occur.

Distinguishing ALL from AML is critical for successful treatment; chemotherapy regimens for ALL generally contain corticosteroids, vincristine, methotrexate, and L-asparaginase, whereas most AML regimens rely on a backbone of daunorubicin and cytarabine (*8*). Although remissions can be achieved using ALL therapy for AML (and vice versa), cure rates are markedly diminished, and unwarranted toxicities are encountered.

We set out to develop a more systematic approach to cancer classification based on the simultaneous expression monitoring of thousands of genes using DNA microarrays (*9*). It has been suggested (*10*) that such microarrays could provide a tool for cancer classification. Microarray studies to date (*11*), however, have primarily been descriptive rather than analytical and have focused on cell culture rather than primary patient material, in which genetic noise might obscure an underlying reproducible expression pattern.

We began with class prediction: How could one use an initial collection of samples belonging to known classes (such as AML and ALL) to create a "class predictor" to classify new, unknown samples? We developed an analytical method and first tested it on distinctions that are easily made at the morphological level, such as distinguishing normal kidney from renal cell carcinoma (*12*). We then turned to the more challenging problem of distinguishing acute leukemias, whose appearance is highly similar.

Our initial leukemia data set consisted of 38 bone marrow samples (27 ALL, 11 AML) obtained from acute leukemia patients at the time of diagnosis (*13*). RNA prepared from bone marrow mononuclear cells was hybridized to high-density oligonucleotide microarrays, produced by Affymetrix and containing probes for 6817 human genes (*14*). For each gene, we obtained a quantitative expression level. Samples were subjected to a priori quality control standards regarding the amount of labeled RNA and the quality of the scanned microarray image (*15*).

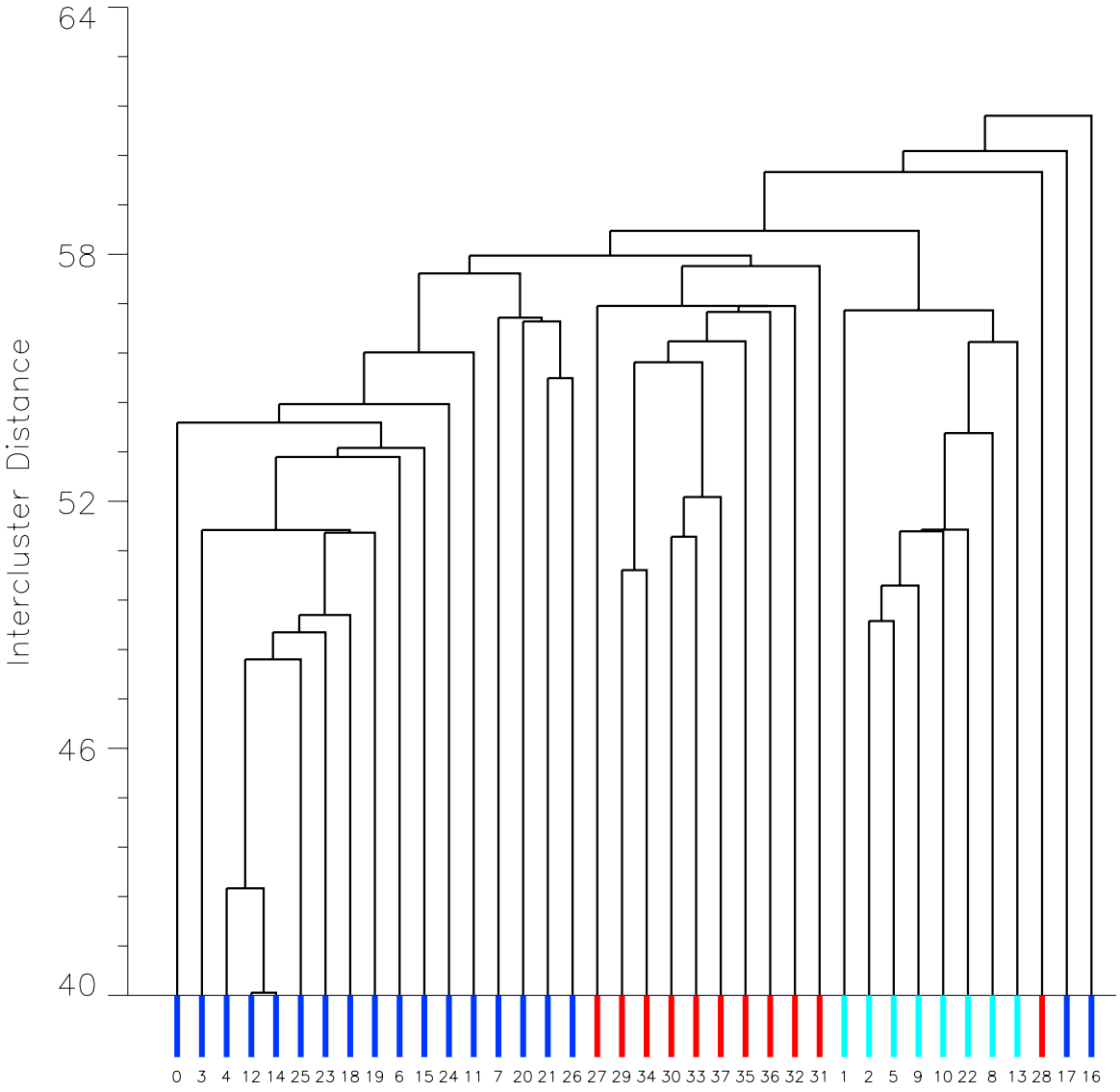
The first issue was to explore whether

[1]Whitehead Institute/Massachusetts Institute of Technology Center for Genome Research, Cambridge, MA 02139, USA. [2]Dana-Farber Cancer Institute and Harvard Medical School, Boston, MA 02115, USA. [3]St. Jude Children's Research Hospital, Memphis, TN 38105, USA. [4]Comprehensive Cancer Center and Cancer and Leukemia Group B, Ohio State University, Columbus, OH 43210, USA. [5]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02142, USA.

*To whom correspondence should be addressed. E-mail: golub@genome.wi.mit.edu; lander@genome.wi.mit.edu.

†These authors contributed equally to this work.

T. Golub et al. ALL/AML Microarray Data Clusters



Single Linkage ℓn(data) keep top 600 var genes

Acute Lymphoblastic Leukemia (ALL): B−cell T−cell

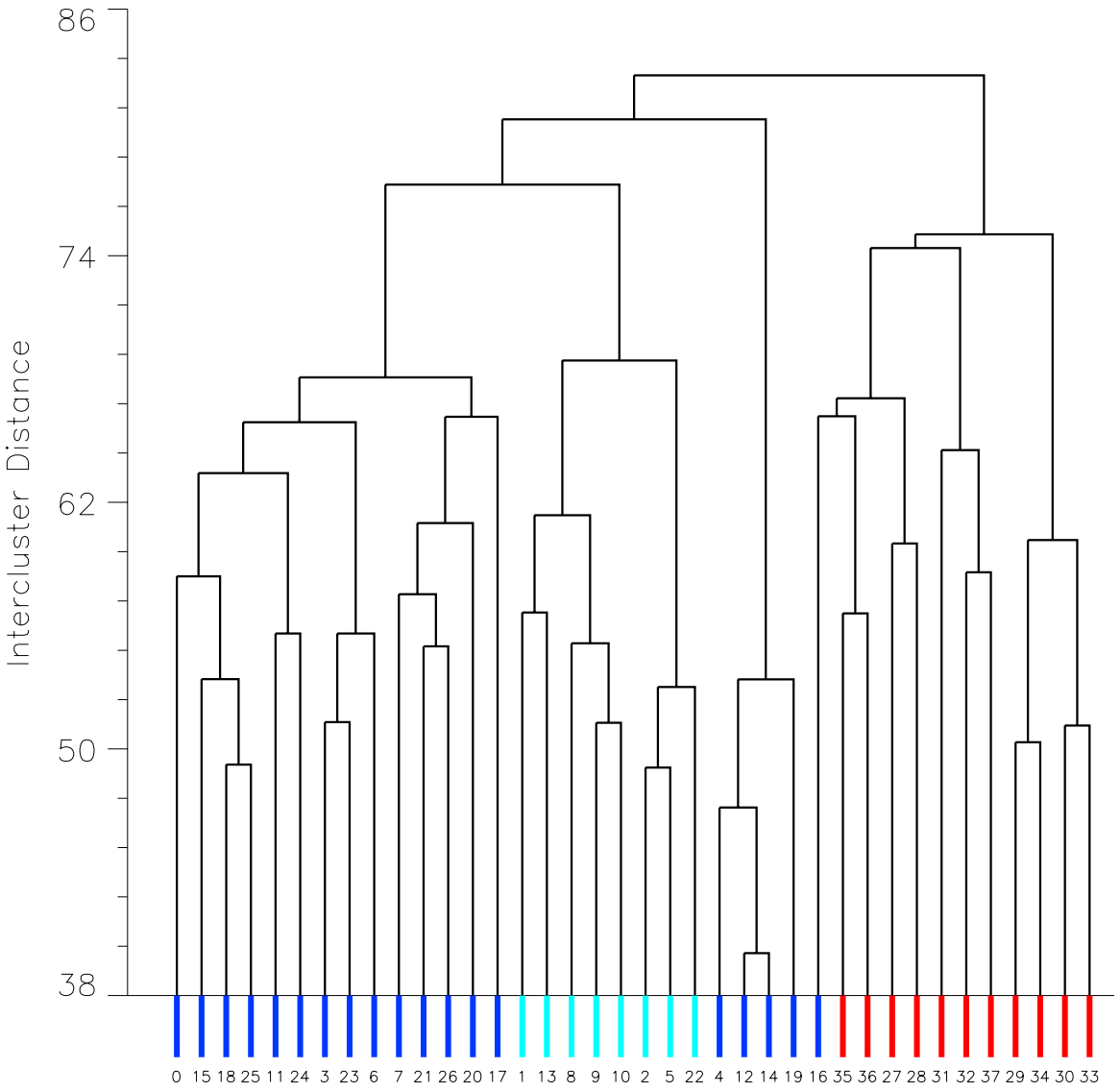
Acute Myeloid Leukemia AML

nnlines =

plotted Tue Oct 08 02:08:08 2002 plothierallaml3n.pro−−Oct 8, 2002

C:\berger\biochem\golubdata\allaml1trainfullploth3n−single.ps

T. Golub et al. ALL/AML Microarray Data Clusters



Complete Linkage ln(data) keep top 600 var genes

Acute Lymphoblastic Leukemia (ALL): B-cell T-cell

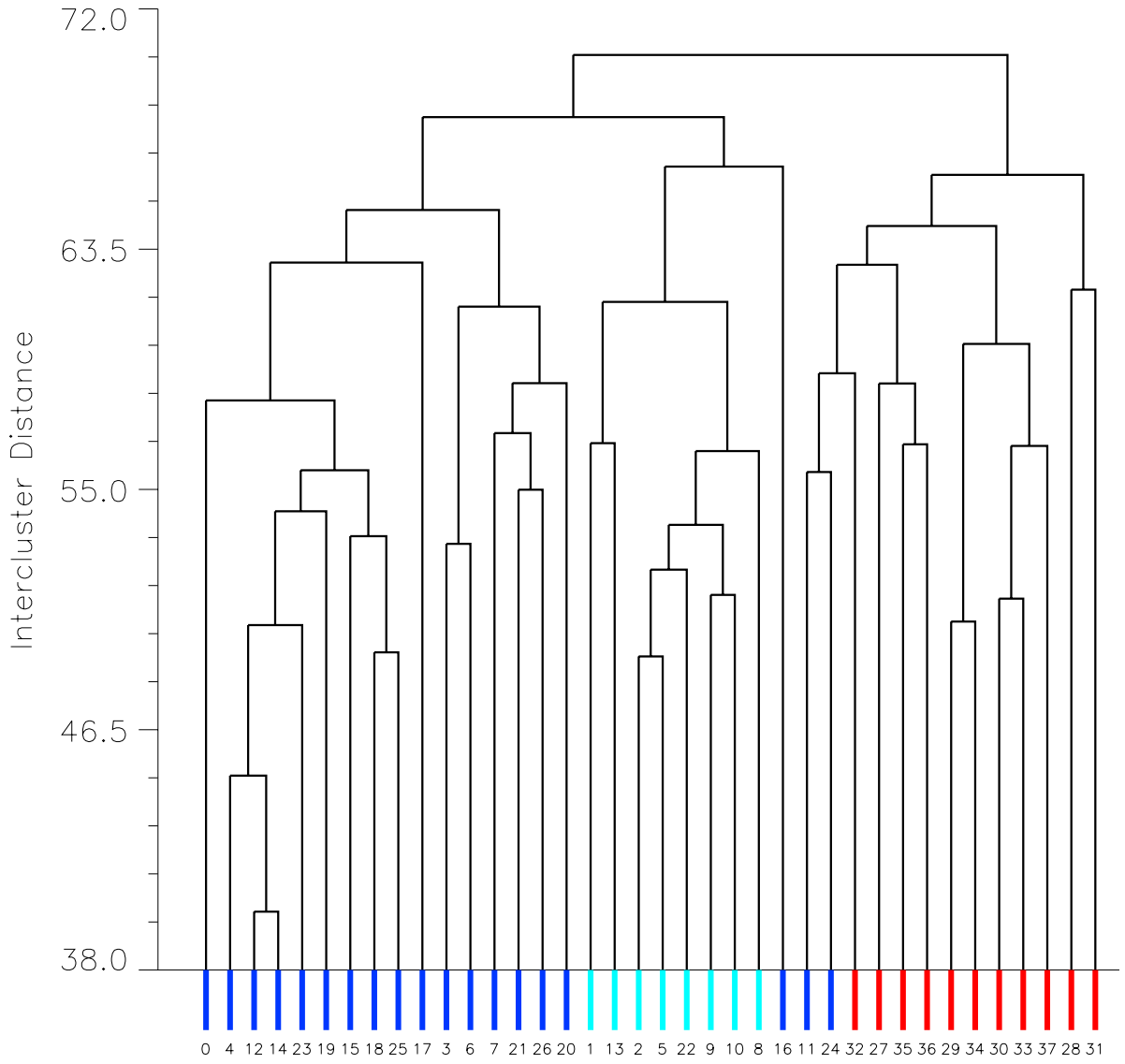
Acute Myeloid Leukemia AML

nnlines =

plotted Tue Oct 08 02:24:11 2002 plothierallaml3n.pro--Oct 8, 2002

C:\berger\biochem\golubdata\allaml1trainfullploth3n-complete.ps

T. Golub et al. ALL/AML Microarray Data Clusters



Average Linkage ln(data) keep top 600 var genes

Acute Lymphoblastic Leukemia (ALL): B−cell T−cell

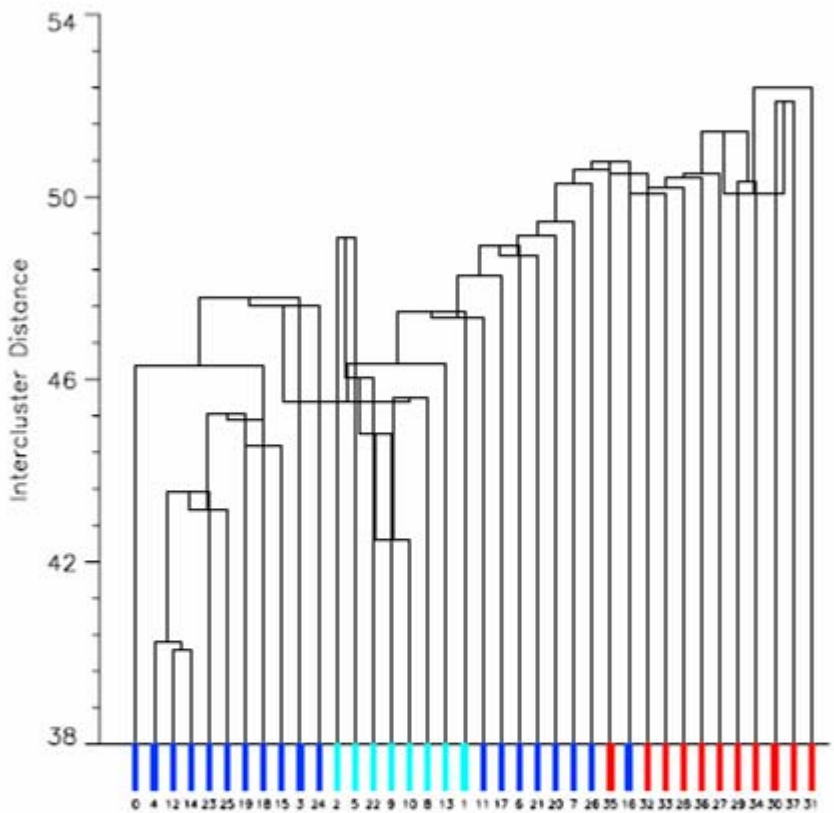
Acute Myeloid Leukemia AML

nnlines =

plotted Tue Oct 08 02:20:45 2002 plothierallaml3n.pro−−Oct 8, 2002

C:\berger\biochem\golubdata\allaml1trainfullploth3n−average.ps

T. Golub et al. ALL/AML Microarray Data Clusters



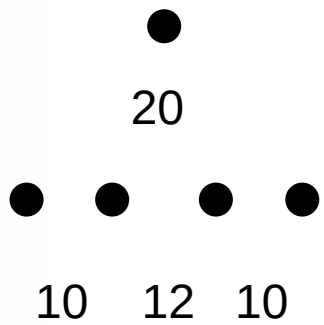
Centroid *ln*(data) keep top 600 var genes

Acute Lymphoblastic Leukemia (ALL): B-cell T-cell

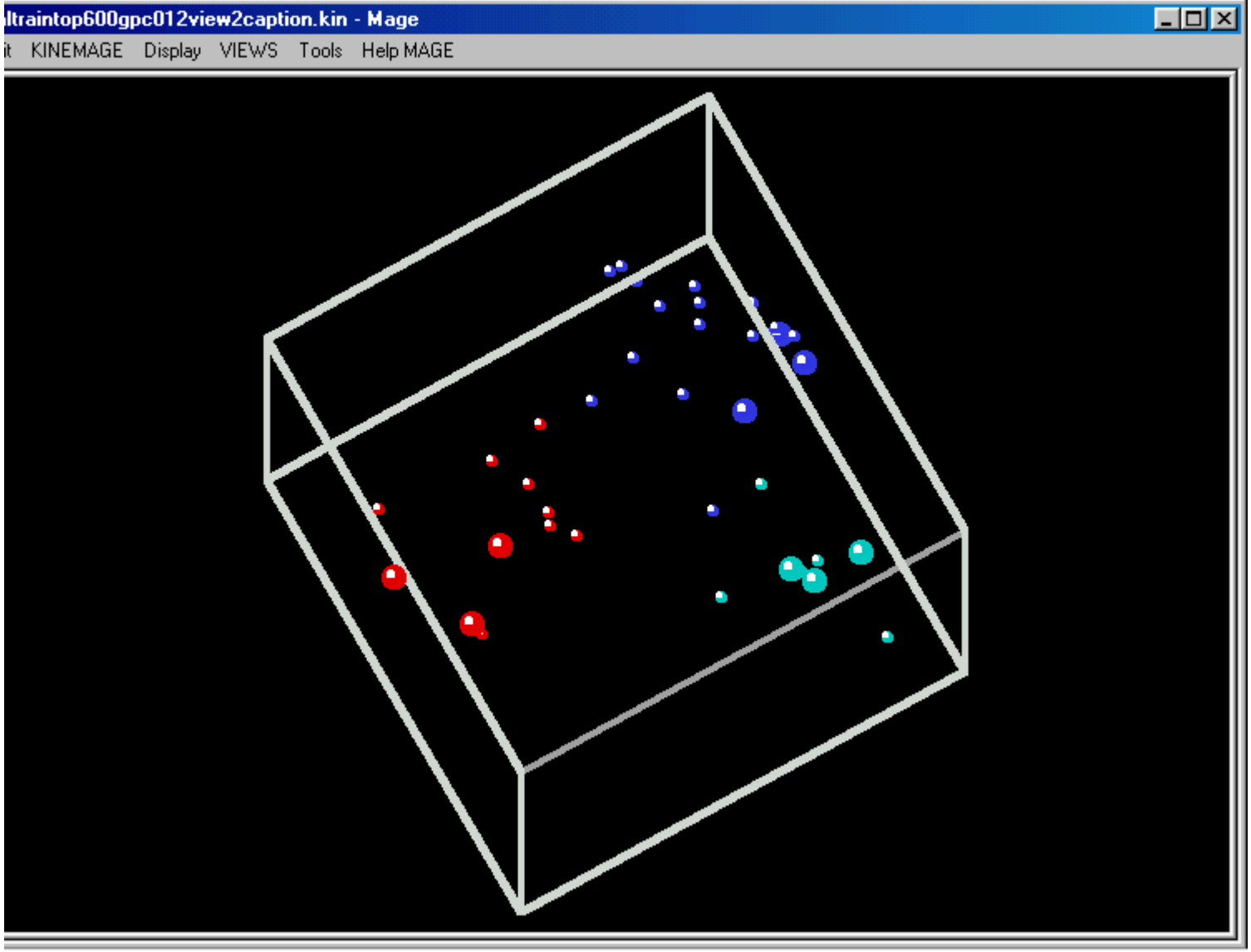
Acute Myeloid Leukemia AML

Note sometimes the heights at which combine successive subclusters can **decrease** when using the **centroid** set distance.

This **never** happens when using single, complete or average linkage.

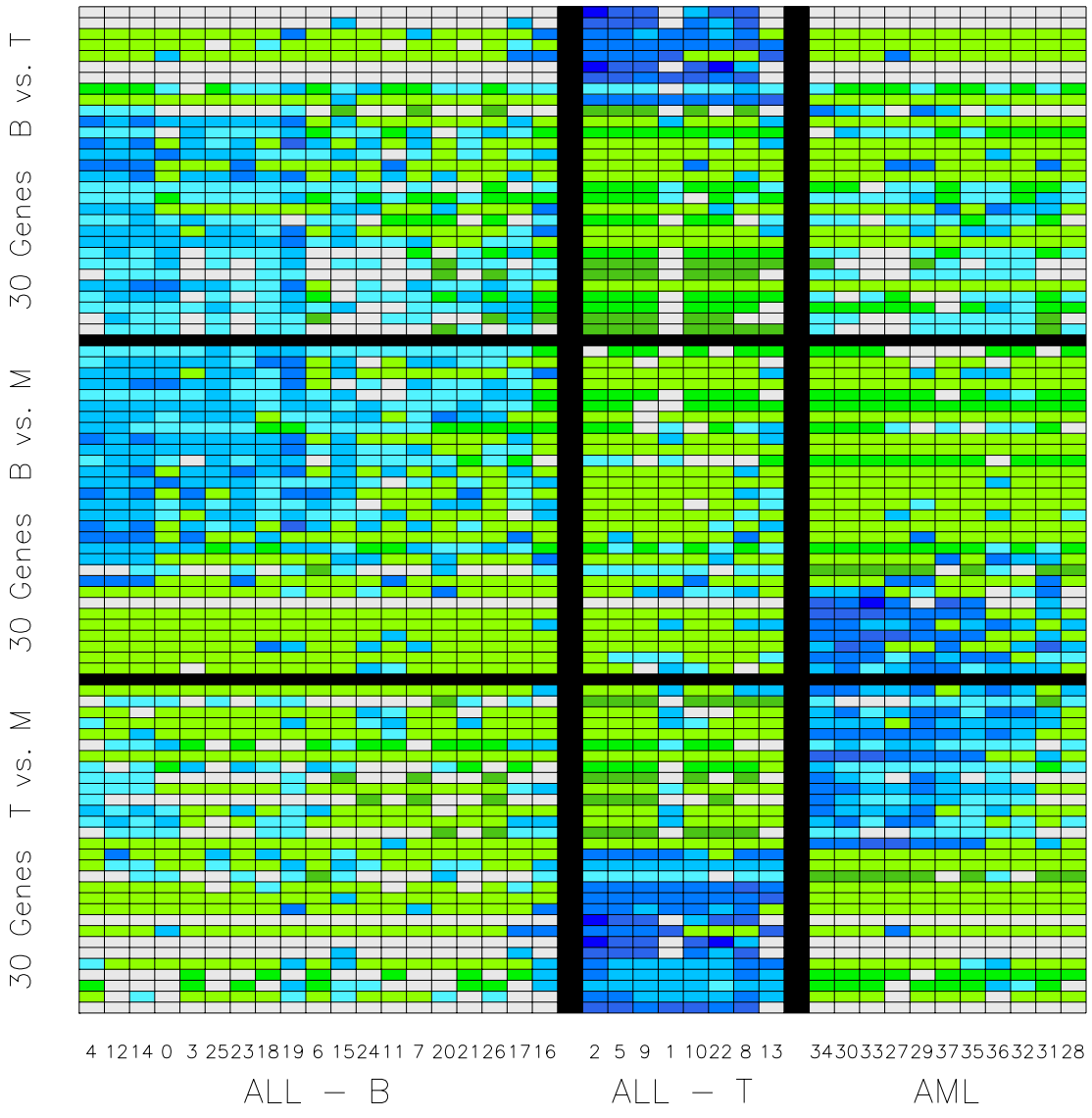


Successive centroid distances are 10, 10, 22, 20 (note sqrt(521) ≅ 22.825)

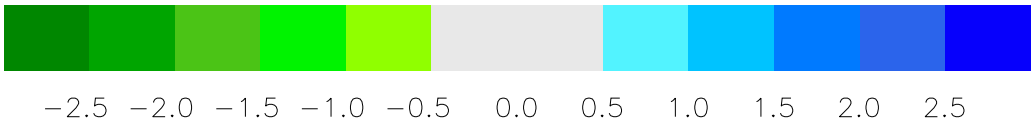


PCA of Golub training data, using 600 top variance genes

Expression Levels for Genes Classifying Tumor Type



Color scale for z ranges of expression for each gene (σ units)



T. R. Golub et al. Acute Leukemia Data (Science, V286, 15 Oct 1999)

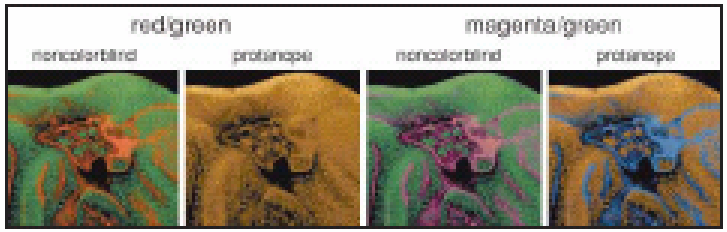
ALL – B is B–cell acute lymphoblastic leukemia

ALL – T is T–cell acute lymphoblastic leukemia

AML is acute myeloid leukemia

C:\berger\biochem\golubdata\allaml3x30genes3ps.ps

Breaking the Color Barrier



Conventional and magenta-tinged slides as seen by the colorblind.

Biologists usually use red and green stains to depict structures inside cells. But that's hard on researchers with the most common form of colorblindness; one in 12 Caucasian and one in 20 Asian males, for example, can't tell the two hues apart. Now help is on the way from a pair of colorblind Japanese scientists, who have persuaded Japan's leading molecular biology journal to start printing images the colorblind can interpret.

Kei Ito, a *Drosophila* neuroscientist at the University of Tokyo, and Masataka Okabe of the National Institute of Genetics in Mishima say that with software it's easy to convert the reds into magentas, which contain enough blue to be seen by the colorblind (see jfly.nibb.ac.jp/html/color_blind).

Months of proselytizing paid off last summer, when the editors of *Saibo Kogaku* (*Cell Technology*) agreed to make their journal "color-barrier-free." The method "opens up a whole new world that was previously hidden," says colorblind geneticist Cahir O'Kane of Cambridge University.

Some U.S. journals may follow suit, says Ito. Everyone would benefit from the changes, he notes: There's a good chance that one of the reviewers of the next paper you submit will be colorblind.

page from **Science** addressing choice of colors distinguishable to those with the most common form of colorblindness

Proc. Natl. Acad. Sci. USA
Vol. 96, pp. 6745–6750, June 1999
Cell Biology

Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays

U. Alon*†, N. Barkai*†, D. A. Notterman*, K. Gish‡, S. Ybarra‡, D. Mack‡, and A. J. Levine*§

Departments of *Molecular Biology and †Physics, Princeton University, Princeton, NJ 08540; and ‡EOS Biotechnology, 225A Gateway Boulevard, South San Francisco, CA 94080

Contributed by A. J. Levine, April 13, 1999

ABSTRACT **Oligonucleotide arrays can provide a broad picture of the state of the cell, by monitoring the expression level of thousands of genes at the same time. It is of interest to develop techniques for extracting useful information from the resulting data sets. Here we report the application of a two-way clustering method for analyzing a data set consisting of the expression patterns of different cell types. Gene expression in 40 tumor and 22 normal colon tissue samples was analyzed with an Affymetrix oligonucleotide array complementary to more than 6,500 human genes. An efficient two-way clustering algorithm was applied to both the genes and the tissues, revealing broad coherent patterns that suggest a high degree of organization underlying gene expression in these tissues. Coregulated families of genes clustered together, as demonstrated for the ribosomal proteins. Clustering also separated cancerous from noncancerous tissue and cell lines from *in vivo* tissues on the basis of subtle distributed patterns of genes even when expression of individual genes varied only slightly between the tissues. Two-way clustering thus may be of use both in classifying genes into functional groups and in classifying tissues based on gene expression.**

Recently introduced experimental techniques based on oligonucleotide or cDNA arrays now allow the expression level of thousands of genes to be monitored in parallel (1–9). To use the full potential of such experiments, it is important to develop the ability to process and extract useful information from large gene expression data sets. Elegant methods recently have been applied to analyze gene expression data sets that are comprised of a time course of expression levels. Examples of such time-course experiments include following a developmental process or changes as the cell undergoes a perturbation such as a shift in growth conditions. The analysis methods were based on clustering of genes according to similarity in their temporal expression (5, 6, 9–11). Such clustering has been demonstrated to identify functionally related families of genes, both in yeast and human cell lines (5, 6, 9, 11). Other methods have been proposed for analyzing time-course gene expression data, attempting to model underlying genetic circuits (12, 13).

Here we report the application of methods for analyzing data sets comprised of snapshots of the expression pattern of different cell types, rather than detailed time-course data. The data set used is composed of 40 colon tumor samples and 22 normal colon tissue samples, analyzed with an Affymetrix oligonucleotide array (8) complementary to more than 6,500 human genes and expressed sequence tags (ESTs) (14). We focus here on generally applicable analysis methods; a more detailed discussion of the cancer-specific biology associated with this study will be presented elsewhere (D.A.N. and A.J.L., unpublished work). The correlation in expression levels across different tissue samples is demonstrated to help identify genes that regulate each other or have similar cellular function. To detect large groups of related genes and tissues we applied two-way clustering, an effective technique for detecting patterns in data sets (see e.g., refs. 15 and 16). The main result is that an efficient clustering algorithm revealed broad, coherent patterns of genes whose expression is correlated, suggesting a high degree of organization underlying gene expression in these tissues. It is demonstrated, for the case of ribosomal proteins, that clustering can classify genes into coregulated families. It is further demonstrated that tissue types (e.g., cancerous and noncancerous samples) can be separated on the basis of subtle distributed patterns of genes, which individually vary only slightly between the tissues. Two-way clustering thus may be of use both in classifying genes into functional groups and in classifying tissues based on their gene expression similarity.

MATERIALS AND METHODS

Tissues and Hybridization to Affymetrix Oligonucleotide Arrays. Colon adenocarcinoma specimens (snap-frozen in liquid nitrogen within 20 min of removal) were collected from patients (D.A.N. and A.J.L., unpublished work). From some of these patients, paired normal colon tissue also was obtained. Cell lines used (EB and EB-1) have been described (17). RNA was extracted and hybridized to the array as described (1, 8).

Treatment of Raw Data from Affymetrix Oligonucleotide Arrays. The Affymetrix Hum6000 array contains about 65,000 features, each containing $\approx 10^7$ strands of a DNA 25-mer oligonucleotide (8). Sequences from about 3,200 full-length human cDNAs and 3,400 ESTs that have some similarity to other eukaryotic genes are represented on a set of four chips. In the following, we refer to either a full-length gene or an EST that is represented on the chip as EST. Each EST is represented on the array by about 20 feature pairs. Each feature contains a 25-bp sequence, which is either a perfect match (PM) to the EST, or a single central-base mismatch (MM). The hybridization signal fluctuates between different features that represent different 25-mer oligonucleotide segments of the same EST. This fluctuation presumably reflects the variation in hybridization kinetics of different sequences, as well as the presence of nonspecific hybridization by background RNAs. Some of the features display a hybridization signal that is many times stronger than their neighbors ($\approx$4% of the intensities are $>$3 SD away from the mean for their EST). These outliers appear with roughly equal incidence in PM or MM features. If not filtered out, outliers contribute significantly to the reading of the average intensity of the gene. Because most features

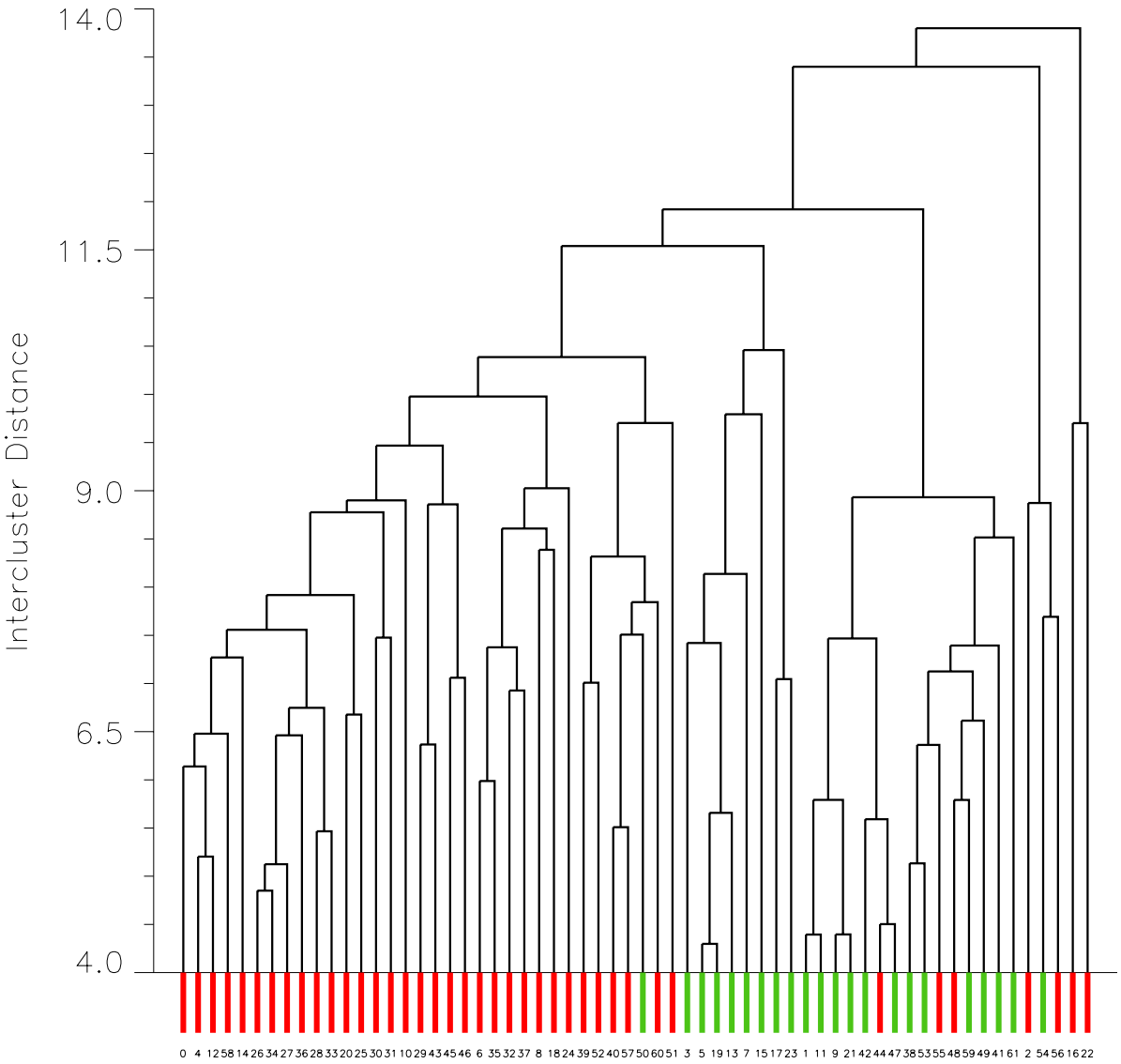
The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

PNAS is available online at www.pnas.org.

Abbreviation: EST, expressed sequence tag.

§To whom reprint requests should be sent at present address: President's Office, Rockefeller University, 1230 York Avenue, New York, NY 10021. e-mail: ajlevine@rockvax.rockefeller.edu.

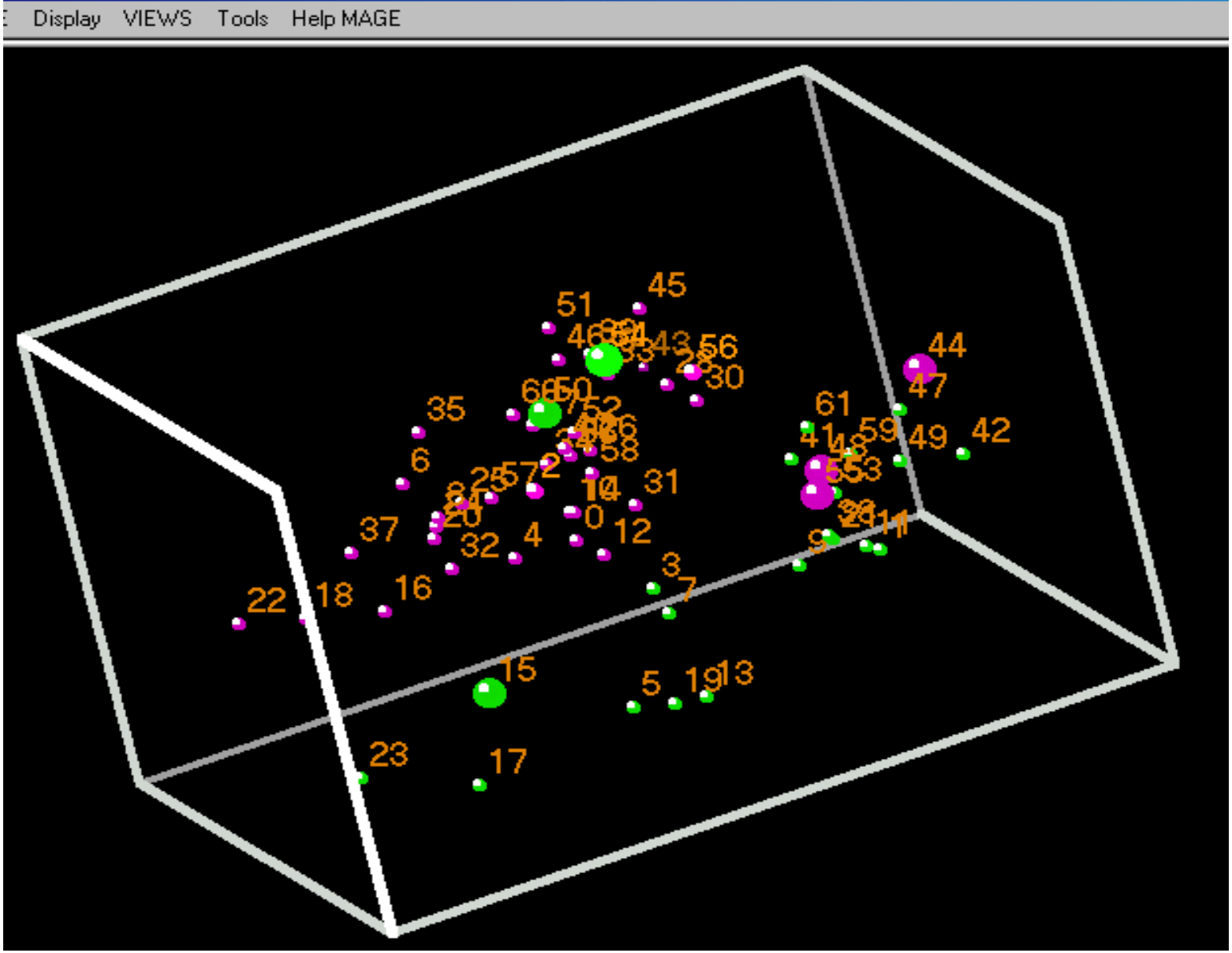
U. Alon et al. Colon Microarray Data Clusters



Average Linkage ln(data) keep top 50 var genes

Colon Tissue Samples: Normal Cancer

plotted Thu Jun 09 22:34:28 2005 plothieralon3nnotrue.pro--July 11, 2003
C:\berger\biochem\alondata\alonallploth3nnotrue-average-nk50PCAclass.ps



Leave-One-Out Crossvalidation for Alon et al. Data:
Misclassified points (3 tumor ● , 3 normal ●) are larger size

Clustering: How Do They Make Those Dendrograms and Heat Maps – Outline

- Definition of unsupervised clustering
- Dendrogram construction by hierarchical agglomerative clustering given specified inter-cluster and inter-point distance measures
- Uniqueness of the dendrogram if an unambiguous choice of left/right ordering is specified for each join of two clusters in the dendrogram construction
- Dependence of the clustering dendrogram on the definition of inter-cluster distance.
- Additional examples, Heat map construction
- Briefly noting other methods for clustering and data visualization
- The difference between exploratory and supervised clustering

K-Means Clustering

1. Choose number of clusters
2. Choose initial cluster centers
3. Assign each point to the cluster whose center is closest
4. Redefine each cluster center as center of mass of all the points assigned to that cluster
5. Repeat 3. & 4. until clusters stabilize

Part of the Anderson Fisher Iris Data Set

5.1	3.5	1.4	0.2	1
4.9	3.0	1.4	0.2	1
4.7	3.2	1.3	0.2	1
4.6	3.1	1.5	0.2	1
5.0	3.6	1.4	0.2	1
5.4	3.9	1.7	0.4	1
4.6	3.4	1.4	0.3	1
5.0	3.4	1.5	0.2	1
.....				
7.0	3.2	4.7	1.4	2
6.4	3.2	4.5	1.5	2
6.9	3.1	4.9	1.5	2
5.5	2.3	4.0	1.3	2
6.5	2.8	4.6	1.5	2
5.7	2.8	4.5	1.3	2
6.3	3.3	4.7	1.6	2
4.9	2.4	3.3	1.0	2
.....				
6.3	3.3	6.0	2.5	3
5.8	2.7	5.1	1.9	3
7.1	3.0	5.9	2.1	3
6.3	2.9	5.6	1.8	3
6.5	3.0	5.8	2.2	3
7.6	3.0	6.6	2.1	3
4.9	2.5	4.5	1.7	3
7.3	2.9	6.3	1.8	3
.....				

column 5 1 = Setosa
2 = Versicolor
3 = Verginica

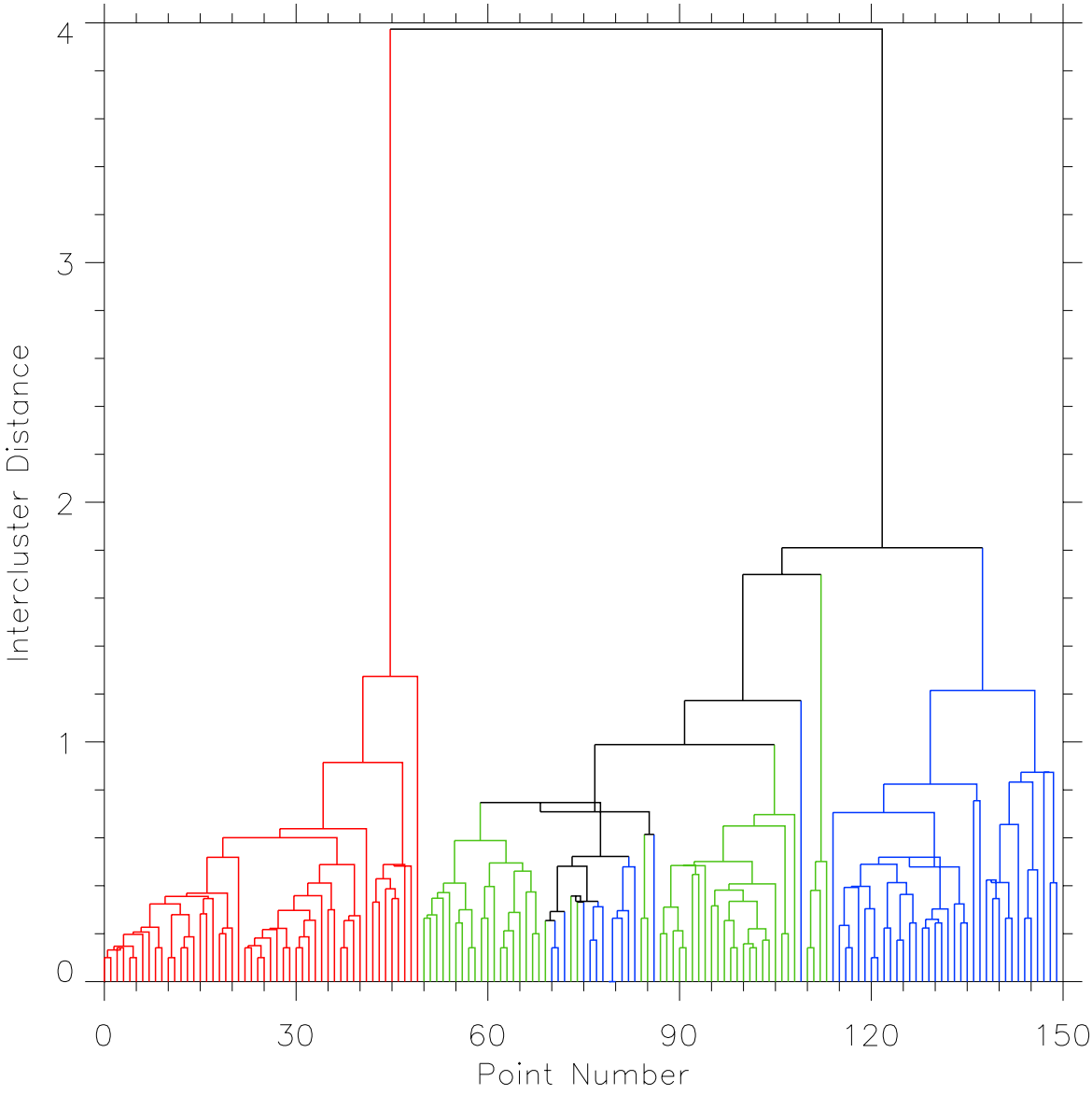
full data set is
50 samples of each iris flower species
(data from R. A. Johnson and D. W. Wichern)

each row is the data from one flower

columns 1, 2, 3, 4 are measured properties of each flower (sepal length, sepal width, petal length, petal width)

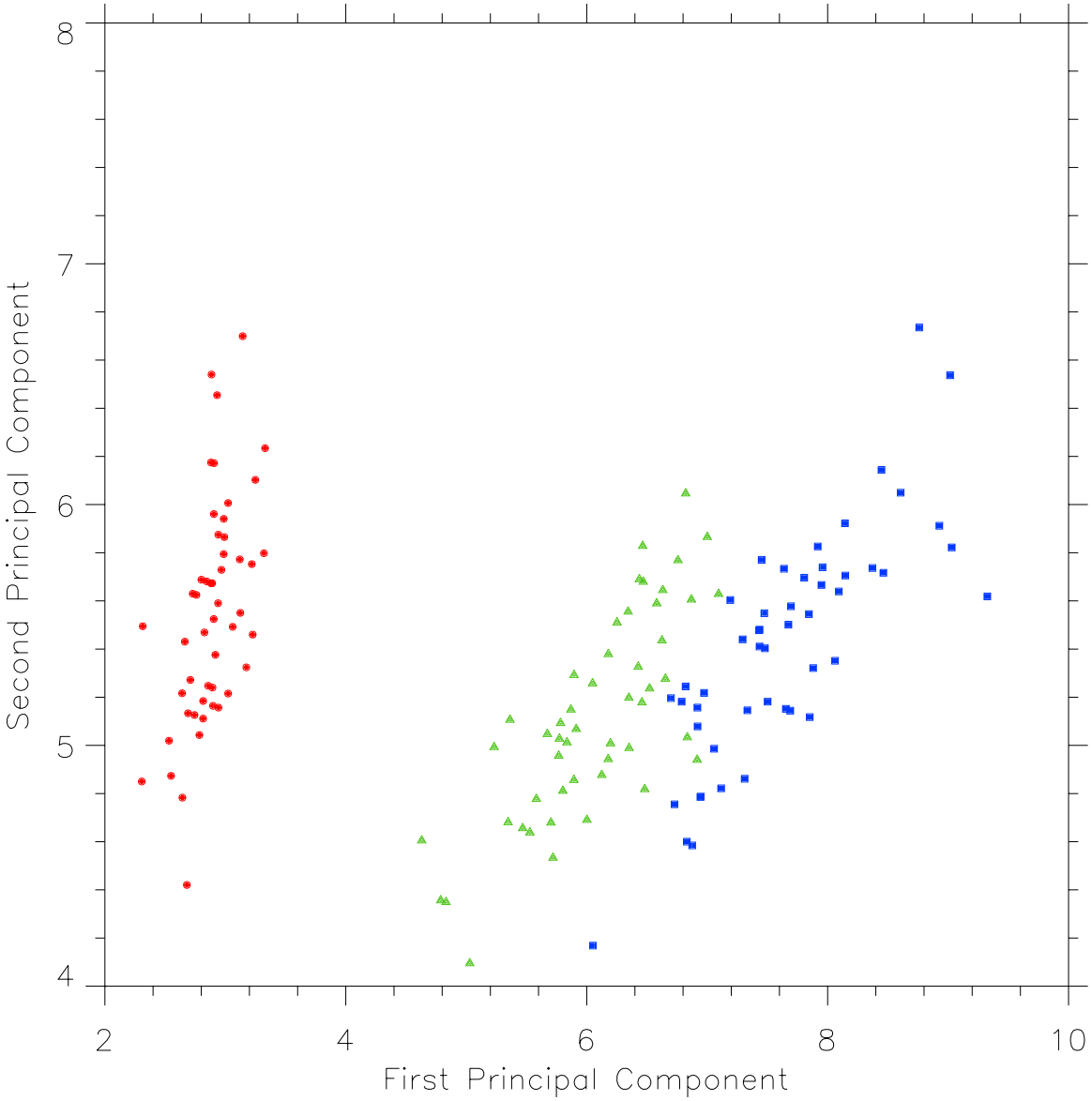
Botany definitions: the *calyx* is the outermost group of floral parts, usually green; *sepals* are the individual leaves or parts of the calyx

Centroid Hierarchical Clustering of Anderson Iris Data



plotted Wed May 29 12:21:09 2002 cenhieriris.pro--May 29, 2002 cenhiermay29

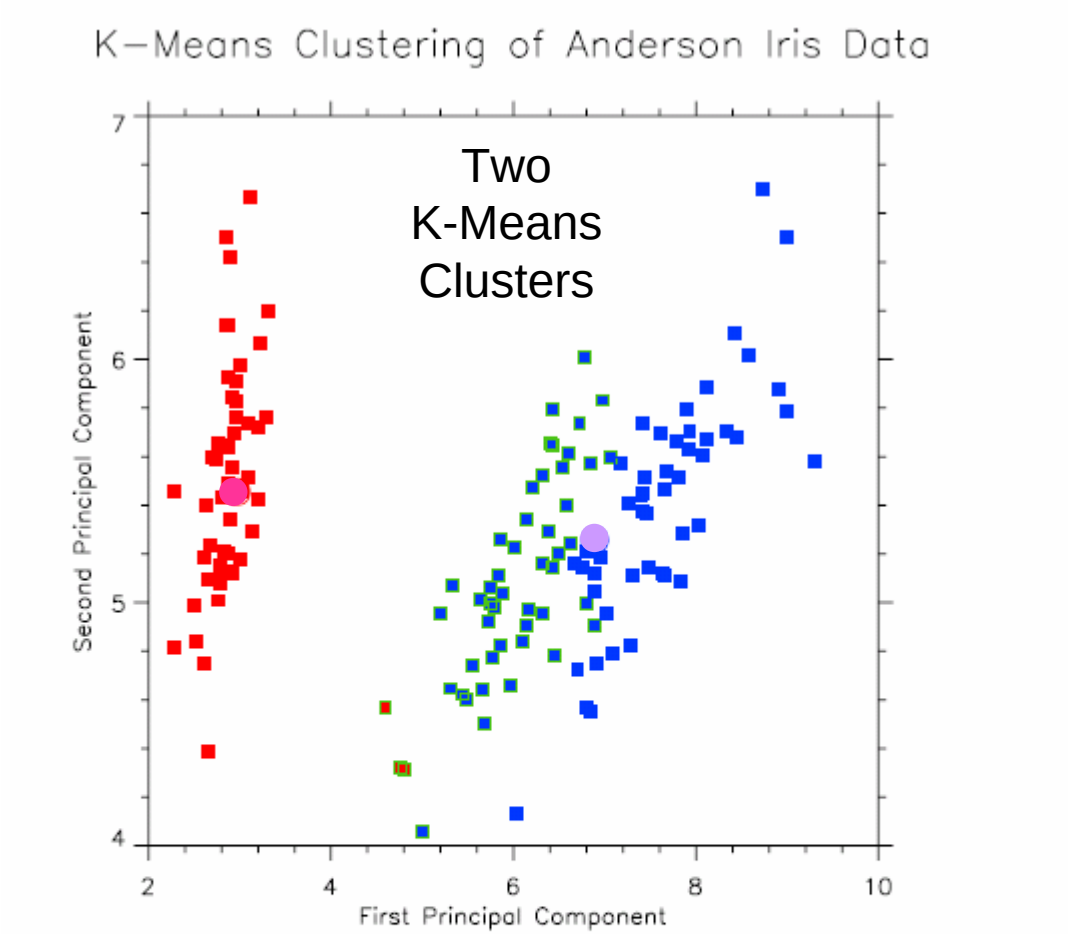
Principal Component Plot of Anderson Iris Data



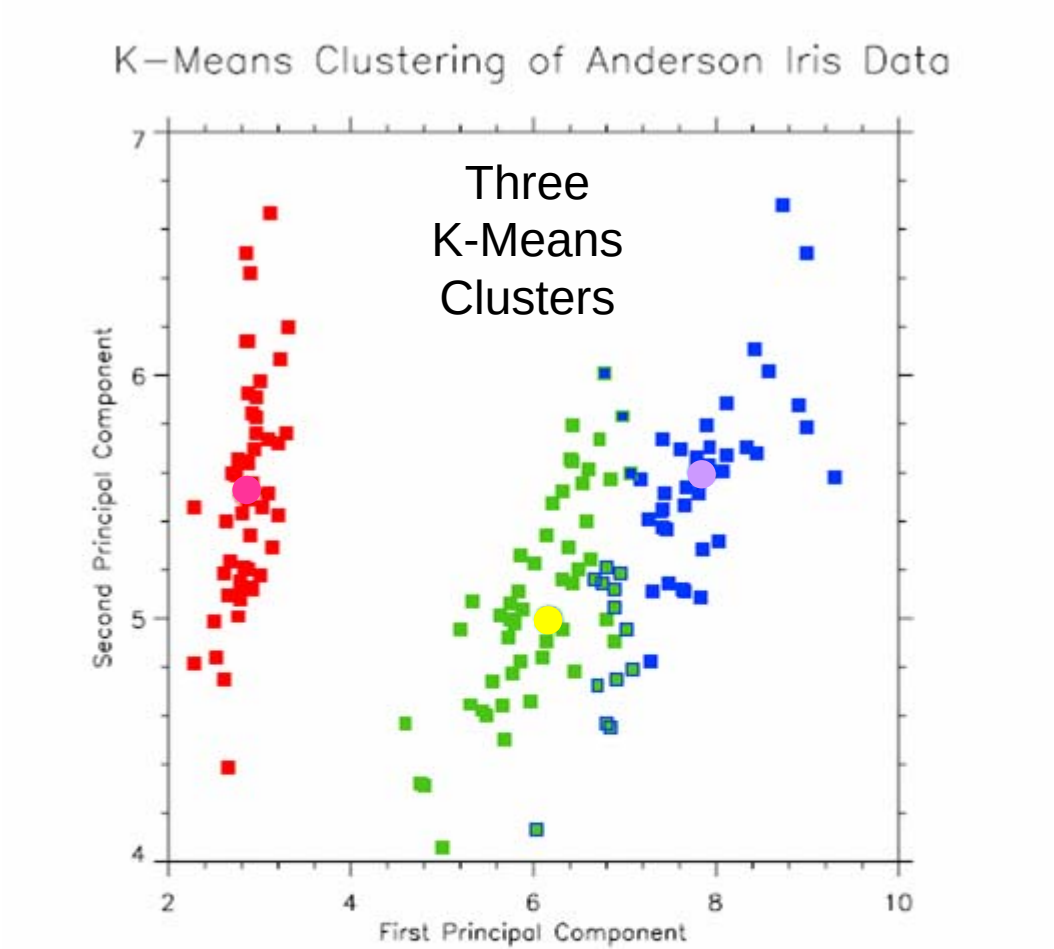
cov evalues = 4.228 0.243 0.078 0.024

Fri Sep 21 12:36:00 2001 irispc1.pro September 21, 2001 irispc1cov.ps

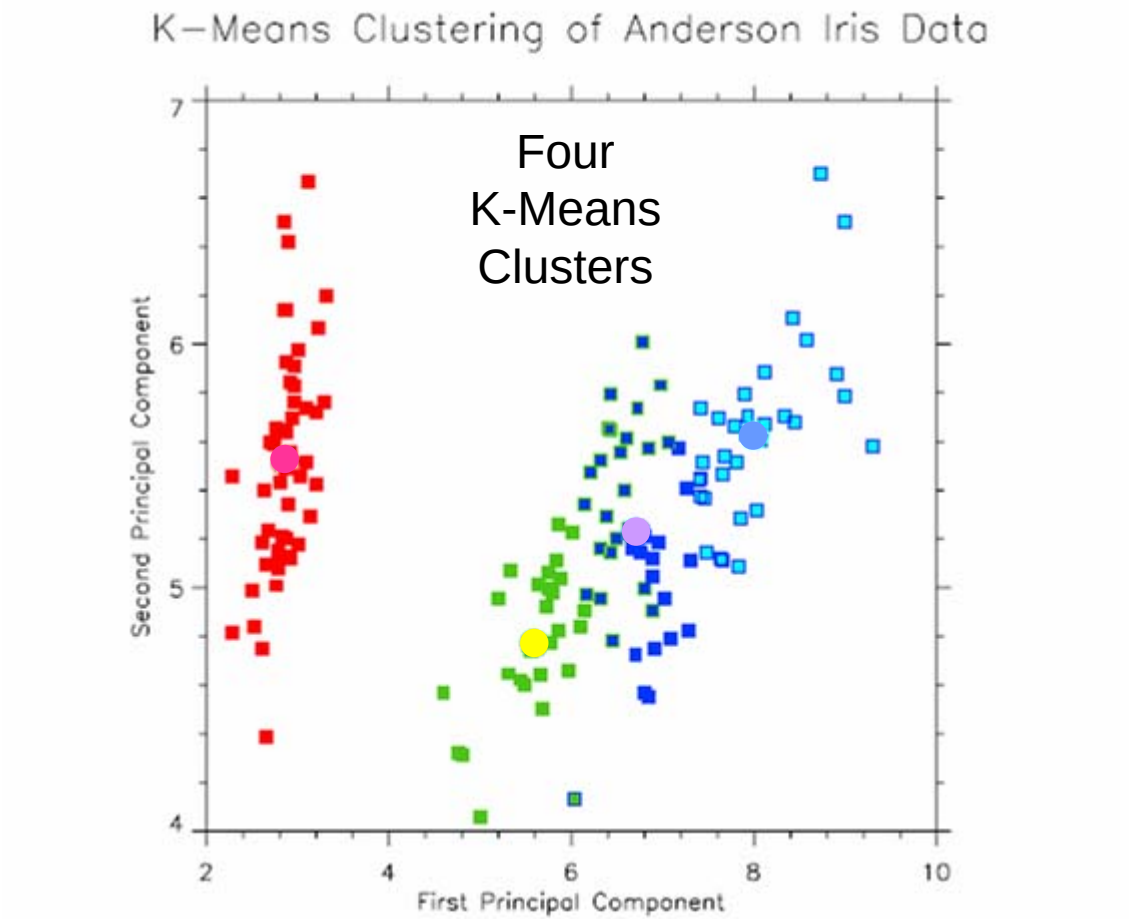
data from Table 11.5 of Johnson and Wichern 1998



The color of the perimeter of each square designates it's correct group; the color of the inside of each square gives its K-Means cluster



The color of the perimeter of each square designates it's correct group; the color of the inside of each square gives its K-Means cluster



The color of the perimeter of each square designates it's correct group; the color of the inside of each square gives its K-Means cluster

Clustering: How Do They Make Those Dendrograms and Heat Maps – Outline

- Definition of unsupervised clustering
- Dendrogram construction by hierarchical agglomerative clustering given specified inter-cluster and inter-point distance measures
- Uniqueness of the dendrogram if an unambiguous choice of left/right ordering is specified for each join of two clusters in the dendrogram construction
- Dependence of the clustering dendrogram on the definition of inter-cluster distance.
- Additional examples, Heat map construction
- Briefly noting other methods for clustering and data visualization
- The difference between exploratory and supervised clustering

CLUSTERING

(Without Training Data: Unsupervised Clustering)
Exploratory separation of data into groups of points. Points in distinct groups to be more different than points within one group. ***Discovery of classes***. Information about the data is used to evaluate the results but is **NOT** used in doing the clustering.

(With Training Data: Supervised Clustering)
Accurate classification of new data points. Verification of accuracy of the training data. ***Discovery of additional classes***. Information about the data **is** used in doing the clustering/classification.

t-like statistic for selecting genes to be used in a classifier

$$\tau(g) \equiv |m_1 - m_2| \,/\, (V_1/N_1 + V_2/N_2 + \delta)^{1/2}$$

m_i = mean of gene g expression levels over training group i

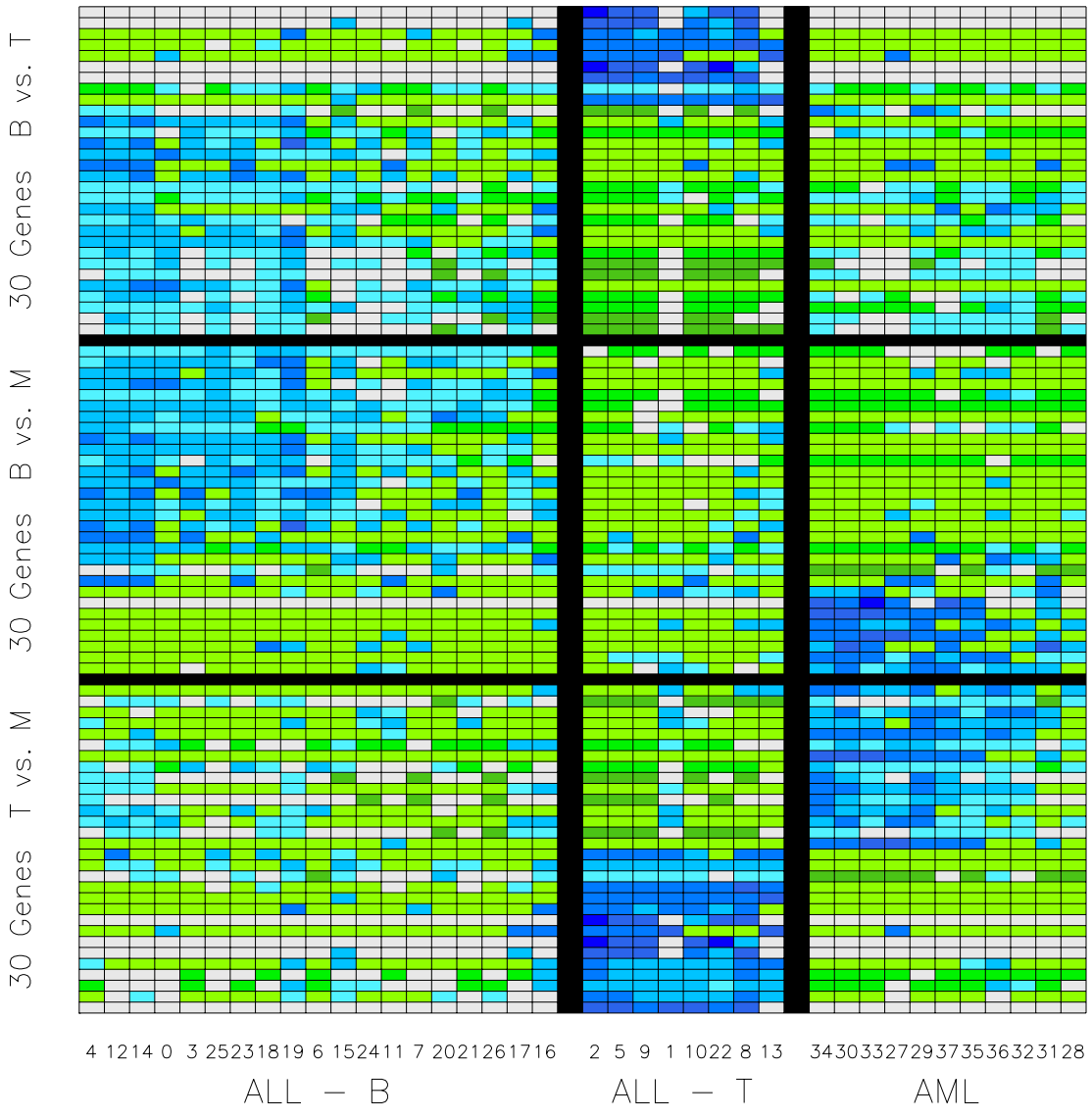
V_i = variance of gene g expression levels over training group i

N_i = number of samples in training group i

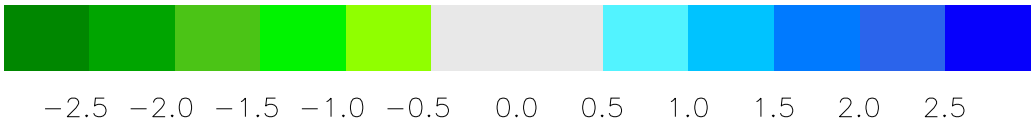
δ = a small constant to prevent division by (nearly) 0

Genes with large τ are more likely to be useful discriminators

Expression Levels for Genes Classifying Tumor Type



Color scale for z ranges of expression for each gene (σ units)



T. R. Golub et al. Acute Leukemia Data (Science, V286, 15 Oct 1999)

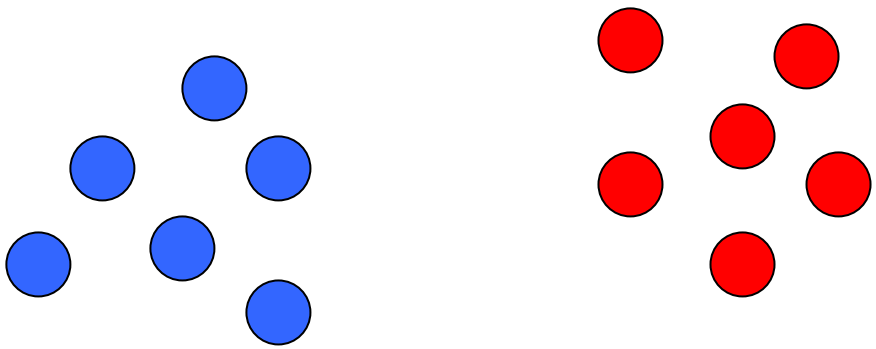
ALL – B is B–cell acute lymphoblastic leukemia

ALL – T is T–cell acute lymphoblastic leukemia

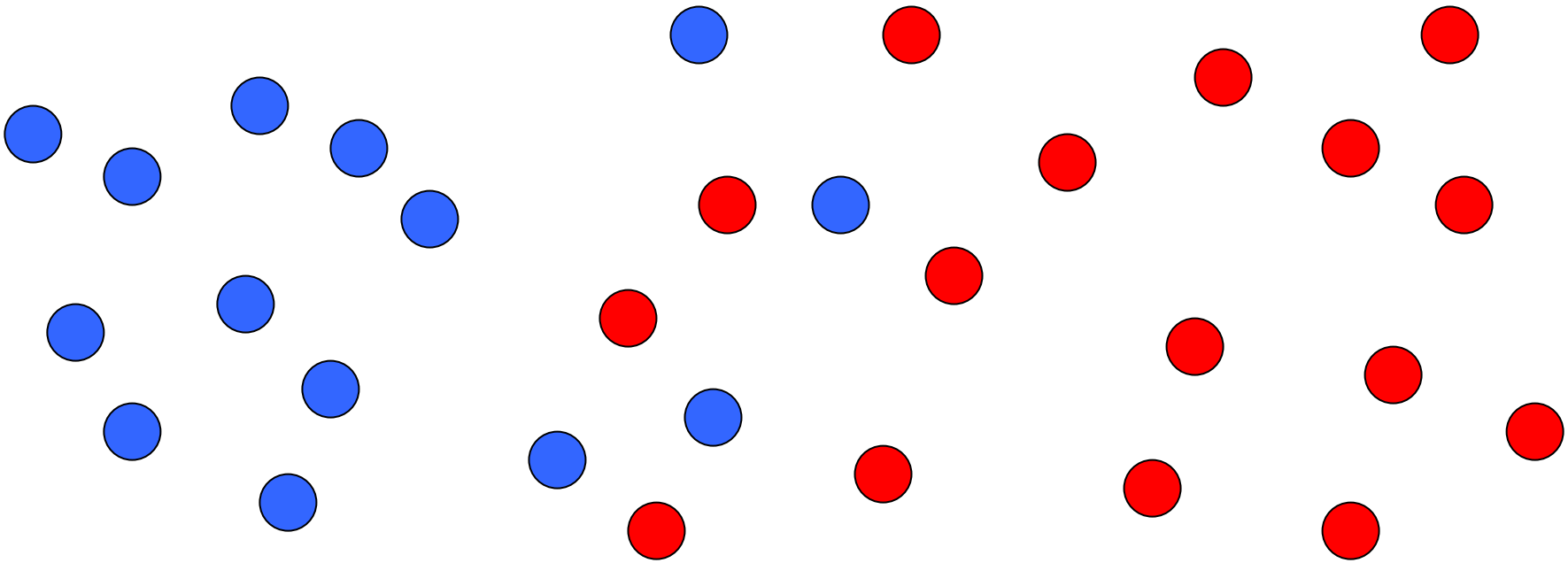
AML is acute myeloid leukemia

C:\berger\biochem\golubdata\allaml3x30genes3ps.ps

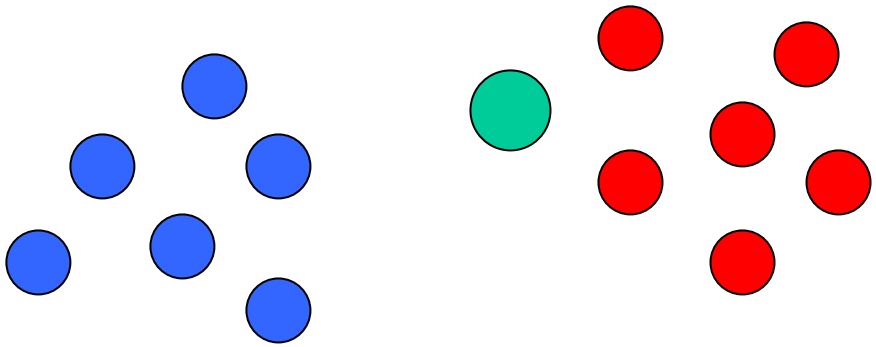
Variance Matters



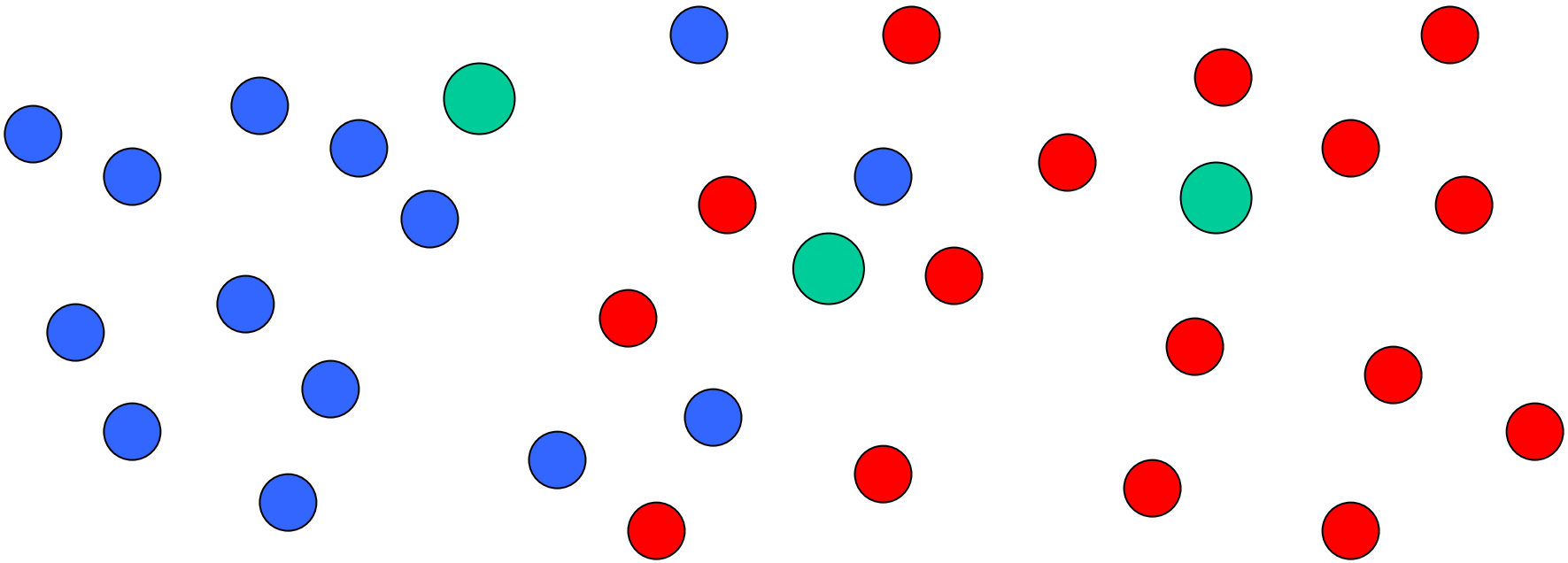
Top situation is better as classifier even though the red and blue group centroids are closer - because variances are smaller



K-Nearest Neighbor Classifier



The class assigned to a "new" data point ● is that of its nearest neighbor <or weighted vote of K nearest neighbors>



Must Do Sound Validation of Any Proposed Classification Method
(use cross-validation, test on independent data, see, e.g.,

R. Simon, M. D. Radmacher, K. Dobbin and L. M. McShane, Pitfalls in the Use of DNA Microarray Data for Diagnostic and Prognostic Classification, *Journal of the National Cancer Institute*, 95 (2003), pp. 14-18. <exploratory class discovery, supervised classification, proper use of cross-validation>

Clustering: How Do They Make Those Dendrograms and Heat Maps – Summary

- Dendrogram construction by hierarchical agglomerative clustering given specified inter-cluster and inter-point distance measures, and a proper ordering of the points
- Uniqueness of the dendrogram if an unambiguous choice of left/right ordering is specified for each join of two clusters in the dendrogram construction
- Dependence of the clustering dendrogram on the definition of inter-cluster distance.
- Heat map construction
- Other methods for clustering and data visualization (k-means, PCA)
- The difference between exploratory and supervised clustering

Hierarchical Methods for **clustering** a set of n points:

Suppose one has *n* "points" one wants to hierarchically cluster --
"draw the hierarchical dendrogram for"
in order to get an exploratory look to see if there
appears to be naturally occurring subgroups.
the points could be expression patterns of *n* genes across
a set of experiments, or the points could be the expression
patterns of *n* patient samples.
In general, one then has *n* vectors V1,...,Vj,...,Vn
each of length, say, m,
so each point Vj has m components:
column(V1j, V2j, ..., Vij, ..., Vmj)
(I tend to think of vectors as column vectors, but that is
just personal preference).
Thus each Vj could be the expression of a given gene across
m experiments, or the expression level of a given tissue
sample measured for *m* genes. If one considers a customary
"summary" of a set of microarray experiments -- the matrix M
of expression levels where the entry in row *g*, column *e*
is the (suitably normalized and often transformed by taking log)
expression level of gene *g* in
experiment (or sample) *e,* then the vectors Vj are either the rows
or the columns of M, depending on whether one wishes to cluster
genes or samples.

To carry out the clustering, one wants to form successive
hierarchical groups (clusters),
combining 2 subsets of points at each step of the process.
To do this, one needs to specify **TWO distance measures**;

1. a **distance measure d(V,W) between any two vectors** V and W
of size m, e.g., the Euclidean distance
d(V,W) = sqrt((V1-W1)^2 + ... + (Vm - Wm)^2)

or the absolute value distance
d(V,W) = |V1-W1| + ... + |Vm - Wm|

2. given a particular choice of distance d in item 1. above,
one also needs to specify a **distance measure**
between any two sets A and B of vectors.
common set distances are:

single linkage --- the distance between set A and set B of points
is the *minimum* of the distances d(a,b) where a ranges over vectors
in the set A and b ranges over vectors in the set B

complete linkage --- the distance between set A and set B of points
is the *maximum* of the distances d(a,b) where a ranges over vectors
in the set A and b ranges over vectors in the set B

average linkage --- the distance between set A and set B of points
is the *average* of the distances d(a,b) where a ranges over vectors
in the set A and b ranges over vectors in the set B

simple example:
given "vectors" with only one component (points on a line)

with (Euclidean) distances as indicated <picture not to scale>

```
v1--v2---v3----v4-----v5------v6--------v7----------v8   points on line
  10   11    12     14     17       21         25   distances(Vi+1,Vi)
```

single linkage produces successive joinings
v1 with v2 <distance 10>
v3 with {v1,v2} <distance 11>
v4 with {v1,v2,v3} <distance 12>
...
v8 with {v1,v2,v3,v4,v5,v6,v7} <distance 25>
illustrating the so-called "chaining" effect often
associated with single linkage unless there are well
separated groups of points

complete linkage produces successive joinings
v1 with v2 <distance 10>
but now the distance from {v1,v2} to v3 is 21, not 11,
so the next join
is {v3,v4},
then {v5,v6}
then {v7,v8},
then {v1,v2, v3,v4} <distance 33 between {v1,v2} and {v3,v4}>
then (by narrow margin) {v5,v6,v7,v8}
<distance 63 between {v5,v6} & {v7,v8}>
         (vs distance of 64 between {v1,v2,v3,v4} and {v5,v6})
then (lastly) the join of the latter two size 4 subsets,
so one gets apparent structure where arguably in this
case there really isn't much.
by nature of using *maximum* inter-set distance,
complete linkage tends to force formation of a larger number
of more compact clusters.

average linkage produces joinings
{v1,v2}
then {v3,v4},
then {v5,v6}
then {v7,v8},
then {v1,v2, v3,v4} as before,
but here the next join is
{v1,v2,v3,v4} with {v5,v6}
<average distance (with some arithmetic) is 41.625>
(average distance between {v5,v6} and {v7,v8} is 42)
and so the last join is
{v1,v2,v3,v4,v5,v6} with {v7,v8}
so here (and arguably in general)
average linkage gives clustering somewhat between
single linkage and complete linkage.
**note the "unstable" nature of the clustering results
depending on choice of inter-set distance measure and
specific point locations**.

the dendrograms are drawn by arranging the point numbers equally
spaced along the x-axis, in an ordering compatible with the
order in which the subsets are joined. At each step of the
process (initially there are *n* "clusters" - the individual points)
each subset has associated with it an x-axis location and a height

(heights are initially 0). If subsets A and B are joined, then
I. a vertical line is drawn at the x location of A
   from y=height of A to y=d(A,B),
II. a vertical line is drawn at the x location of B
    from y=height of B to y=d(A,B), and
III. a horizontal line is drawn between x=x location of A
     and x=x location of B at the height y=d(A,B)
IV. the x location of the new cluster A union B is the
    average of the x location of A and the x location of B,
    and the height of the new cluster is d(A,B).

WARNING: clustering algorithms produce clusters,
whether or not there is real structure to the points.
feed any computer program that does clustering vectors from a
random number generator and it will return a dendrogram!
So one always needs to verify whether apparent
clustering results make biological sense.

last -- one may consider "correlation distance."
One might, for example, be interested in whether two genes
have responses that either both go up (or down) in tandem,
or that go in opposite directions in each experiment (as
might be the case if they were in "competing"
pathways (when 1 is up the other is downregulated)).
Consider the case when are working with vectors whose components
are log(expression level in "treatment"/expression
                                level in control),
so the "null case" is a value of 0.
Thus if two genes are expressed in a coordinated fashion,
the dot product of their vectors should have "large" magnitude;
  V dot W = V.W is defined as V1*W1 + ... + Vm*Wm
if V and W go in "opposite directions" then V.W would be
negative with large magnitude. If in some cases both are
up (or both down regulated) while in other experiments one
is up while the other is down, there will be
cancellation within the sums in the dot product.
Here let's take the view that we are interested
in whether both go up or both go down (positive correlation)
or always go oppositely (one is up while other is down) which
is negative correlation, so rather than weighing the magnitude of individual
gene expression levels, we'll **normalize** each vector by multiplying
each component by the same constant (v ---> cV) so the
length of each expression ratio vector is now 1
(scale so V1^2 + ...+ Vm^2 = 1 for each V)
before doing the dot products.
Then V.W will always be between -1 and 1 so a
standard way of defining a "correlation distance" is to set
d(V,W) = 1 - [absolute value of(V.W)]
(when correlation is high, distance is low)