Hierarchical clustering for gene expression data analysis

Giorgio Valentini

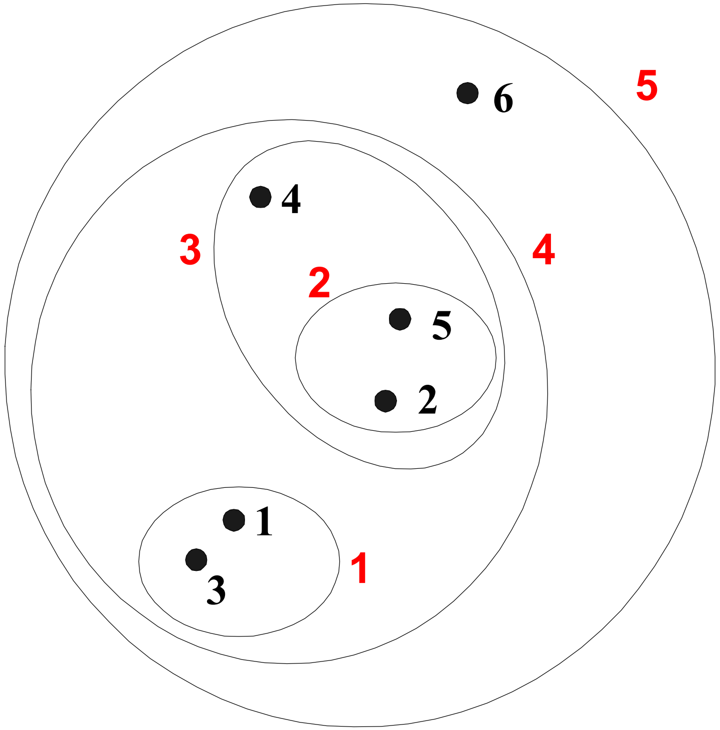
e-mail: valentini@dsi.unimi.it

Dipartimento di Scienze dell'Informazione
Università degli Studi di Milano

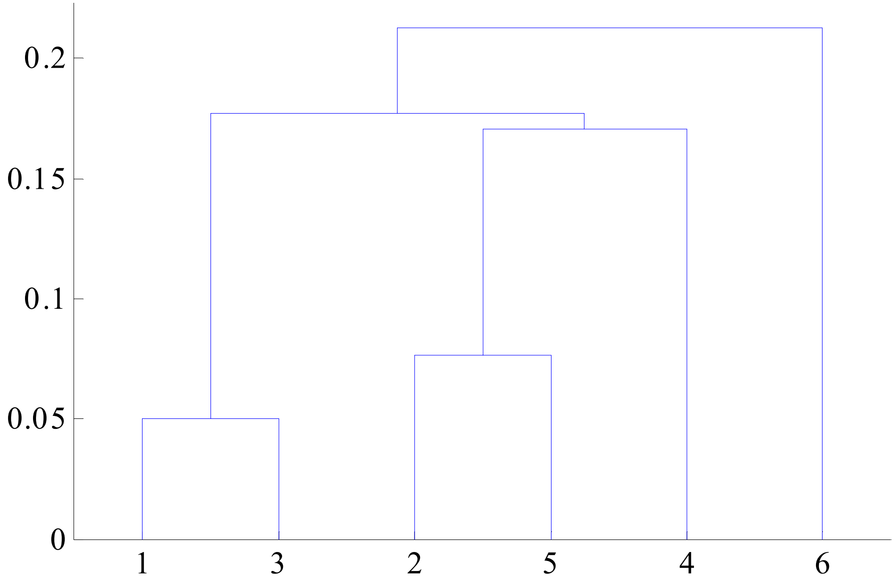
Clustering of Microarray Data

1. *Clustering of gene expression profiles* (rows) => discovery of co-regulated and functionally related genes(or unrelated genes: different clusters)
2. *Clustering of samples* (columns) => identification of sub-types of related samples
3. *Two-way clustering* => combined sample clustering with gene clustering to identify which genes are the most important forsample clustering

Hierarchical Clustering

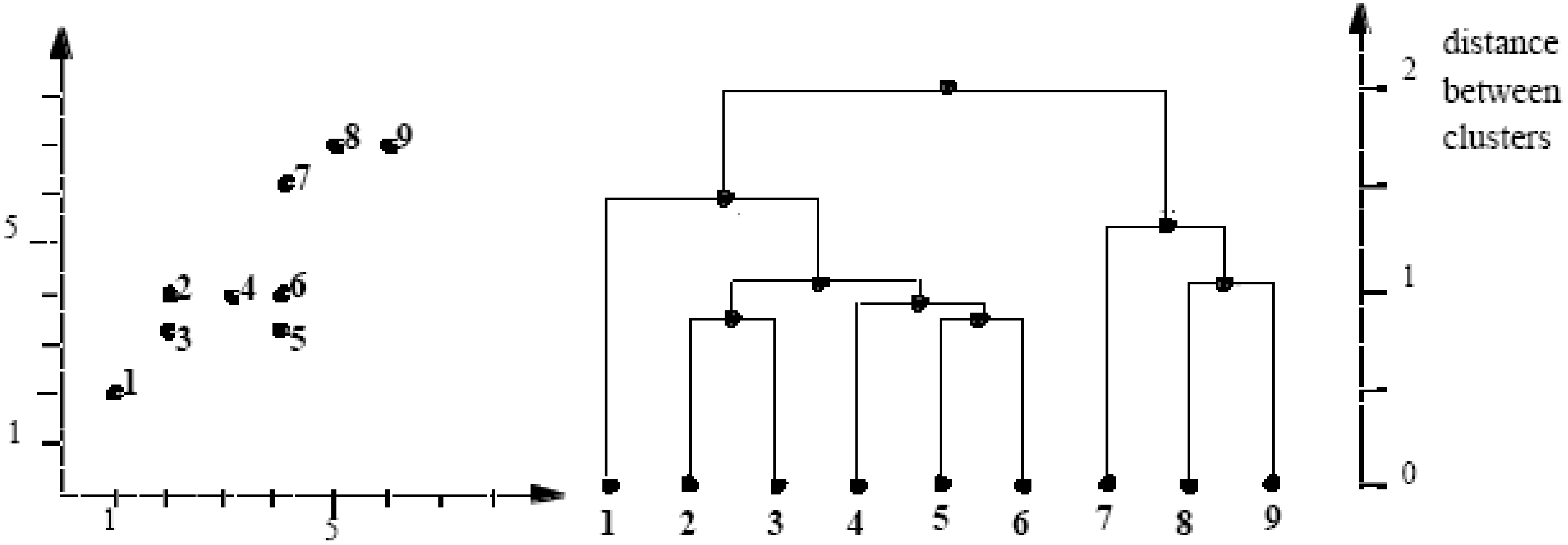


Hierarchical Clustering



Dendrogram

Dendrograms



- The *root* represents the whole data set
- A *leaf* represents a single object in the data set
- An *internal node* represent the union of all objects in its sub-tree
- The *height* of an internal node represents the distance between its two child nodes

Hierarchical Clustering

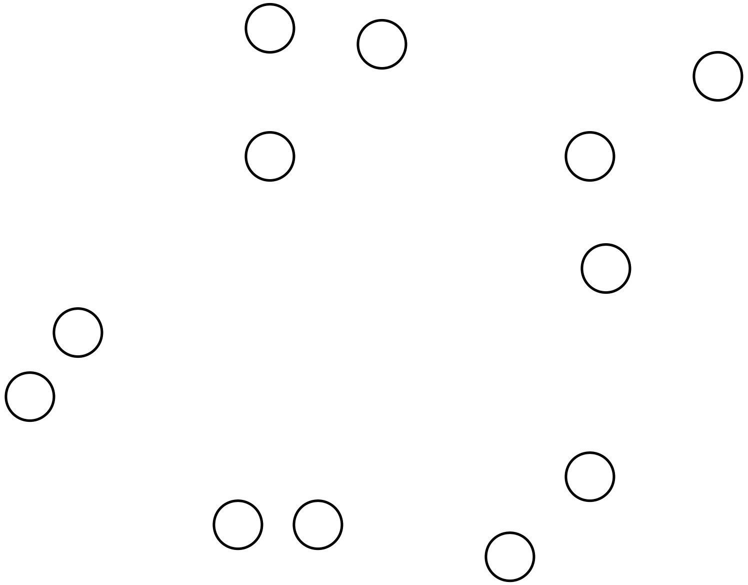
- Two main types of hierarchical clustering.
  - **Agglomerative**:
    - Start with the points as individual clusters
    - At each step, merge the closest pair of clusters.
    - Until only one cluster (or k clusters) left
    - This requires defining the notion of cluster proximity.
  - **Divisive:**
    - Start with one, all-inclusive cluster
    - At each step, split a cluster
    - Until each cluster contains a point (or there are k clusters)
    - Need to decide which cluster to split at each step.

Basic Agglomerative Hierarchical Clustering Algorithm

```
1. Initially, each object forms its own cluster
2. Compute all pairwise distances between the initial clusters (objects)
repeat
    3. Merge the closest pair (A, B) in the set of the current clusters into a new cluster $C = A \cup B$
    4. Remove A and B from the set of current clusters; insert C into the set of current clusters
    5. Determine the distance between the new cluster C and all other clusters in the set of current clusters
until only a single cluster remains
```

Agglomerative Hierarchical Clustering: Starting Situation

- For agglomerative hierarchical clustering we start with clusters of individual points and a proximity matrix.



	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						

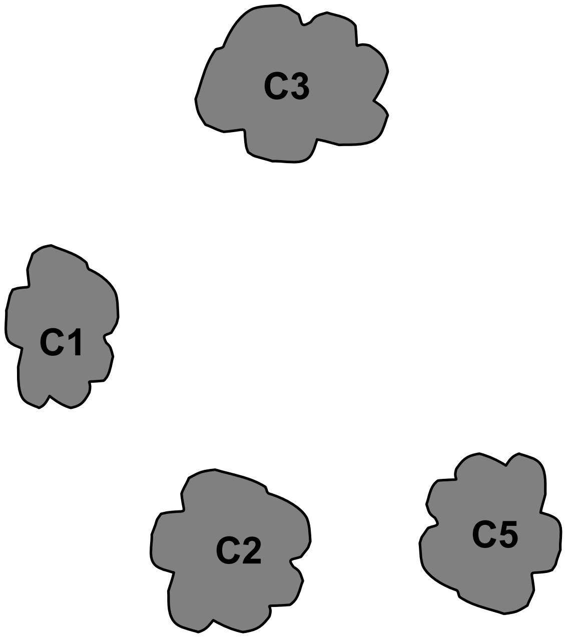
.

.

Proximity Matrix

Agglomerative Hierarchical Clustering: Intermediate Situation

- After some merging steps, we have some clusters.

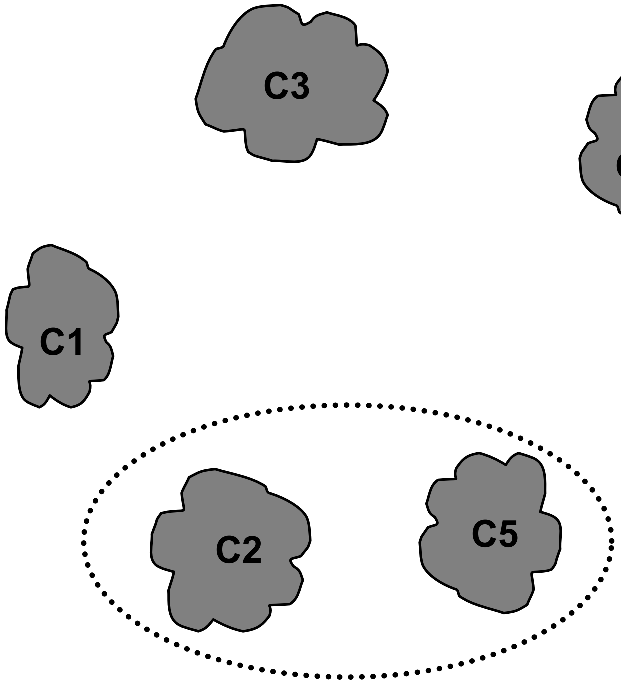


	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Proximity Matrix

Agglomerative Hierarchical Clustering: Intermediate Situation

- We want to merge the two closest clusters (C2 and C5) and update the proximity matrix.

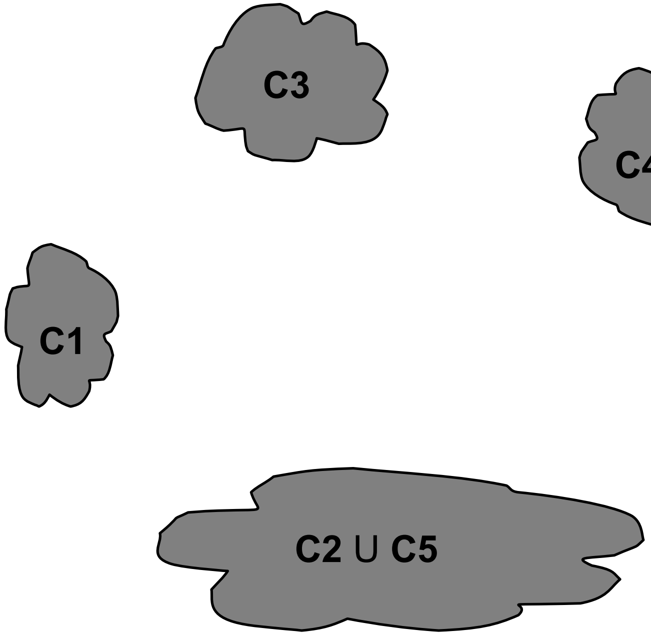


	C1	C2	C3	C4	C5
C1		▨			▨
C2	▨	▨	▨	▨	▨
C3		▨			▨
C4		▨			▨
C5	▨	▨	▨	▨	▨

Proximity Matrix

Agglomerative Hierarchical Clustering: after Merging

- The question is “How do we update the proximity matrix?”



	C1	C2 ∪ C5	C3	C4
C1		?		
C2 ∪ C5	?	?	?	?
C3		?		
C4		?		

Distance Matrix

•Key operation is the computation of the distance of two clusters.

•Different approaches to defining the distance between clusters distinguishes the different algorithms

Inter-cluster distances

- Four widely used ways of defining the **inter-cluster distance**, i.e., the distance between two separate clusters C_i and C_j, are
  - o **single linkage method** (nearest neighbor):

    $$d(C_i, C_j) = \min_{x\in Ci, y\in Cj}\{\, d(x, y) \,\}$$

  - o **complete linkage method** (furthest neighbor):

    $$d(C_i, C_j) = \max_{x\in Ci, y\in Cj}\{\, d(x, y) \,\}$$

  - o **average linkage method** (unweighted pair-group average): $d(C_i, C_j) = avg_{x\in Ci, y\in Cj}\{\, d(x, y) \,\}$

  - o **centroid linkage method** (distance between cluster centroids c_i and c_j): $d(C_i, C_j) = d(c_i, c_j)$

Single linkage (minimum distance) method

- Distance (dissimilarity) of two clusters is based on the two most similar (closest) points in the different clusters C_i and C_j :

$$d(C_i, C_j) = \min_{x \in C_i, y \in C_j} \{ d(x, y) \}$$

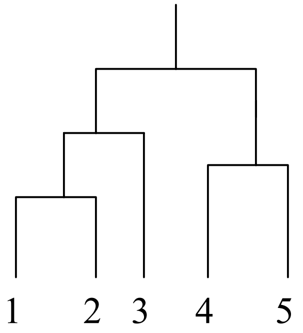
–Determined by one pair of points, i.e., by one link in the proximity graph.

–Can handle non-elliptical shapes.

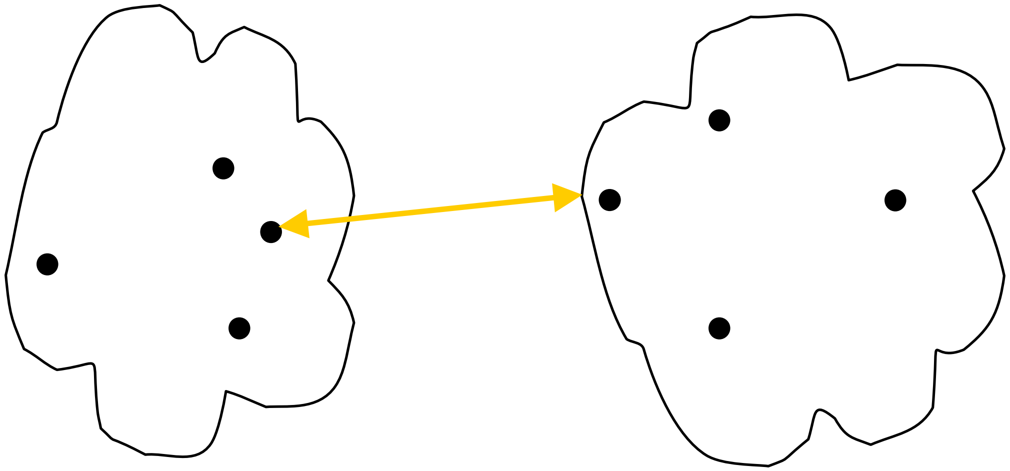
–Sensitive to noise and outliers.

Similarity matrix

	I1	I2	I3	I4	I5
I1	1.00	0.90	0.10	0.65	0.20
I2	0.90	1.00	0.70	0.60	0.50
I3	0.10	0.70	1.00	0.40	0.30
I4	0.65	0.60	0.40	1.00	0.80
I5	0.20	0.50	0.30	0.80	1.00

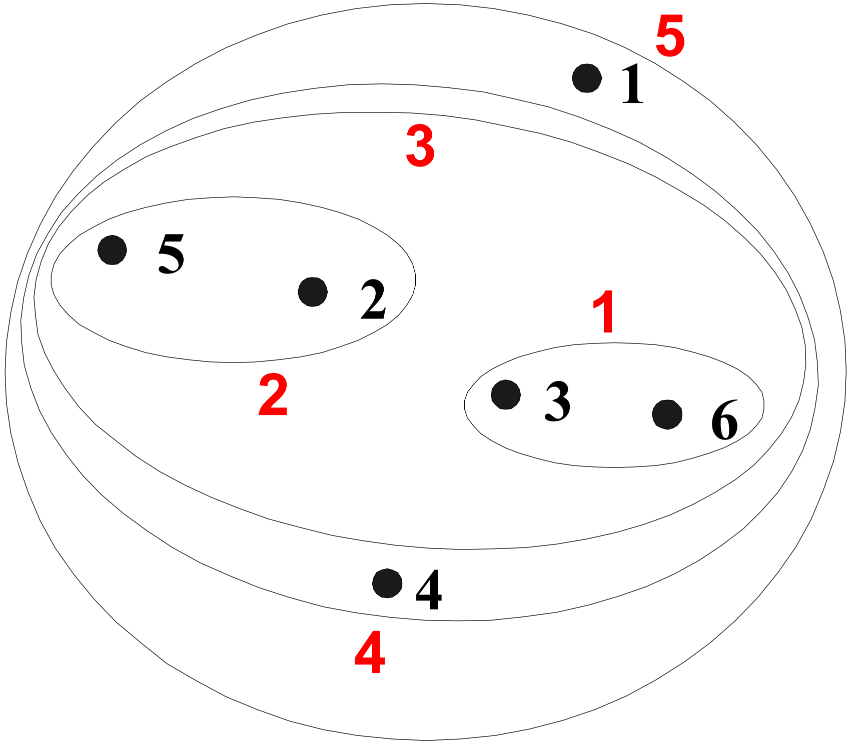


Single linkage

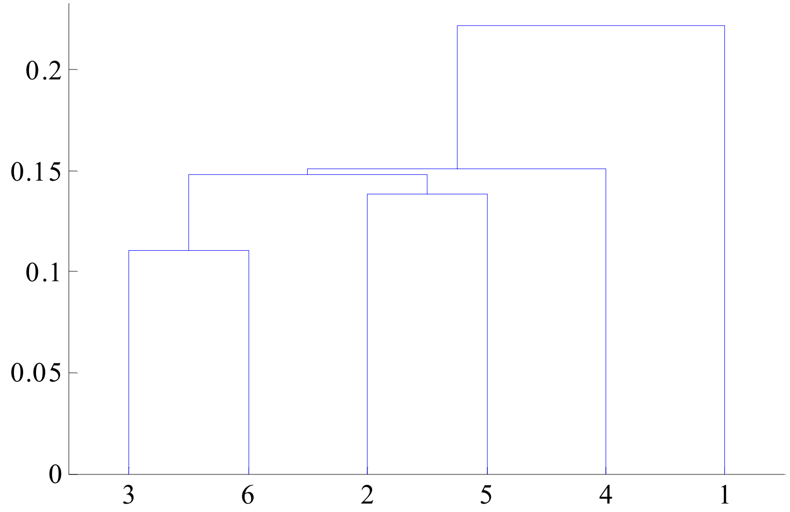


$$d(C_i, C_j) = \min_{x \in C_i, y \in C_j} \{ d(x, y) \}$$

Hierarchical Clustering: minimum distance

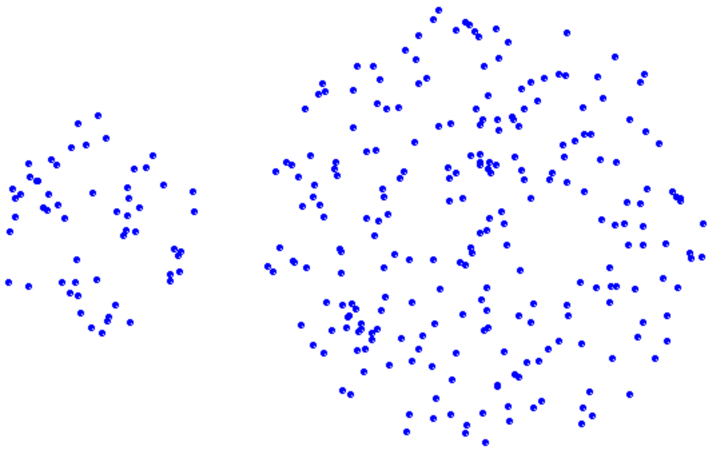


Nested Clusters

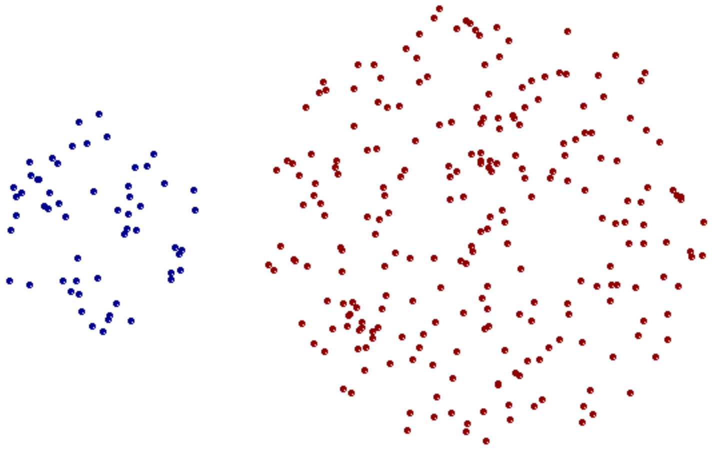


Dendrogram

Strength of minimum distance

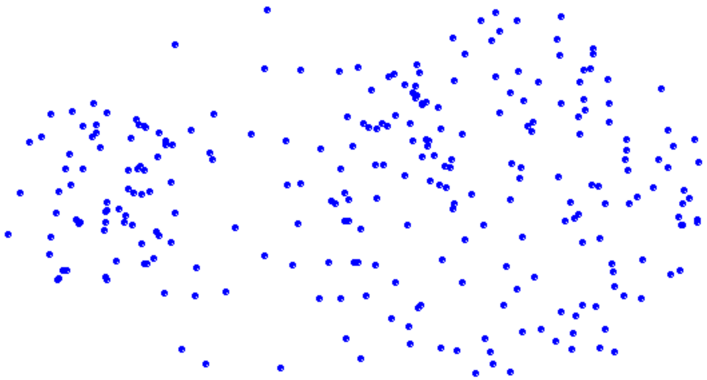


Original Points

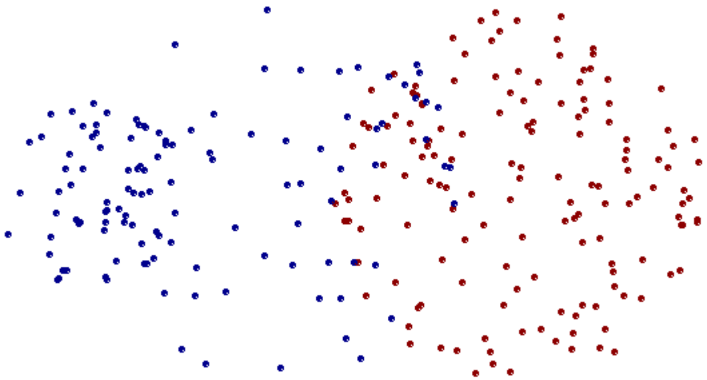


Two Clusters

Limitation of minimum distance



Original Points



Two Clusters

Complete Linkage (minimum distance) method

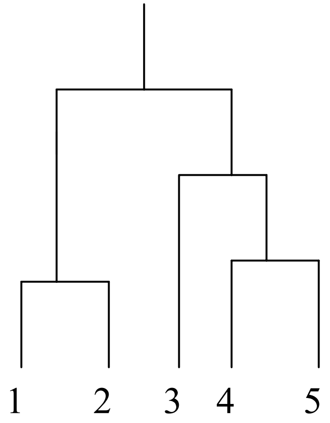
- Distance of two clusters is based on the two least similar (most distant) points in the different clusters C_{i} and C_j:

$$d(C_i, C_j) = \max_{x\in Ci, y\in Cj} \{ d(x, y) \}$$

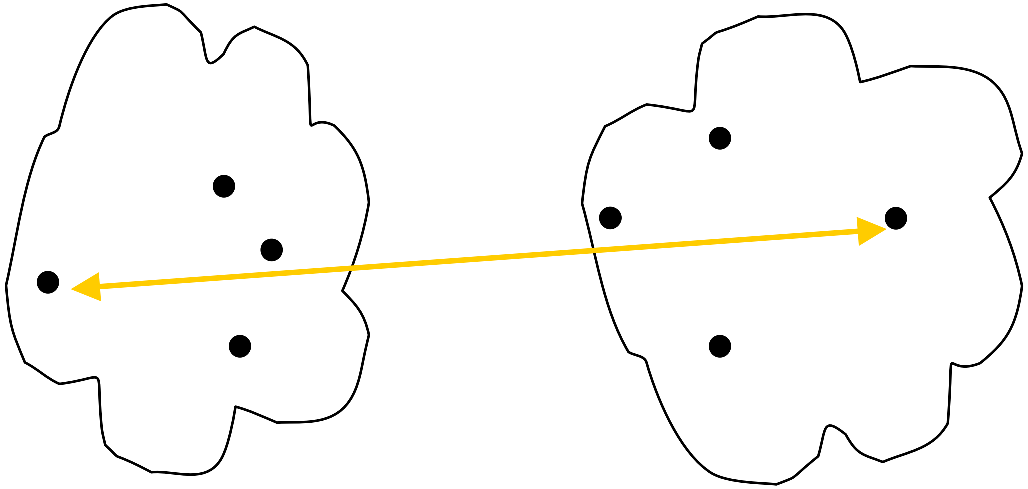
  - –Determined by all pairs of points in the two clusters.
  - –Tends to break large clusters.
  - –Less susceptible to noise and outliers.

Similarity matrix

	I1	I2	I3	I4	I5
I1	1.00	0.90	0.10	0.65	0.20
I2	0.90	1.00	0.70	0.60	0.50
I3	0.10	0.70	1.00	0.40	0.30
I4	0.65	0.60	0.40	1.00	0.80
I5	0.20	0.50	0.30	0.80	1.00



Complete linkage



$$d(C_i, C_j) = \max_{x \in C_i, y \in C_j} \{ d(x,y) \}$$

Cluster Similarity: maximum distance or Complete Linkage

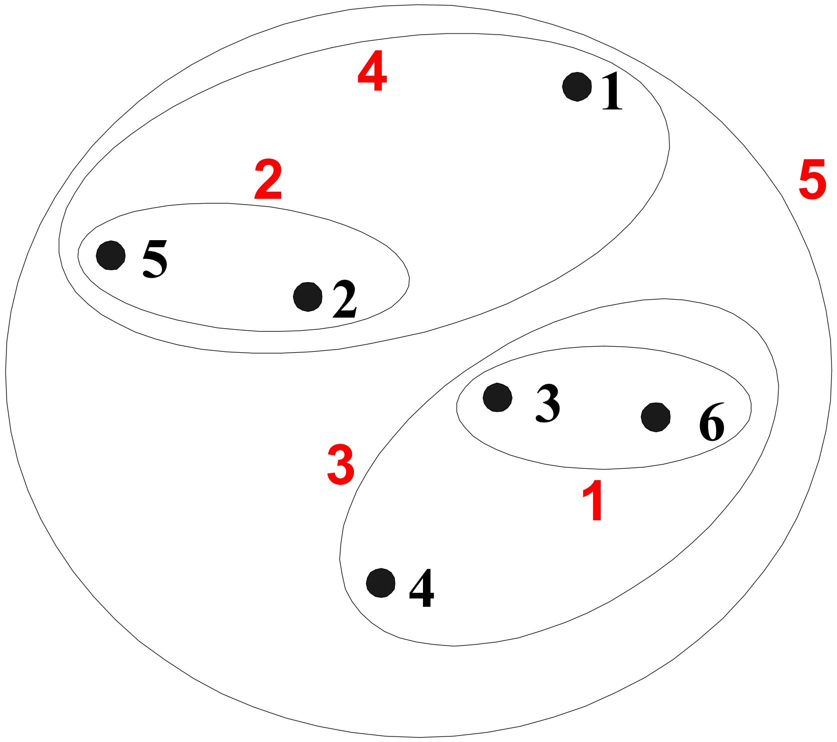
- Similarity of two clusters is based on the two most distant points in the different clusters.

- Tends to break large clusters.

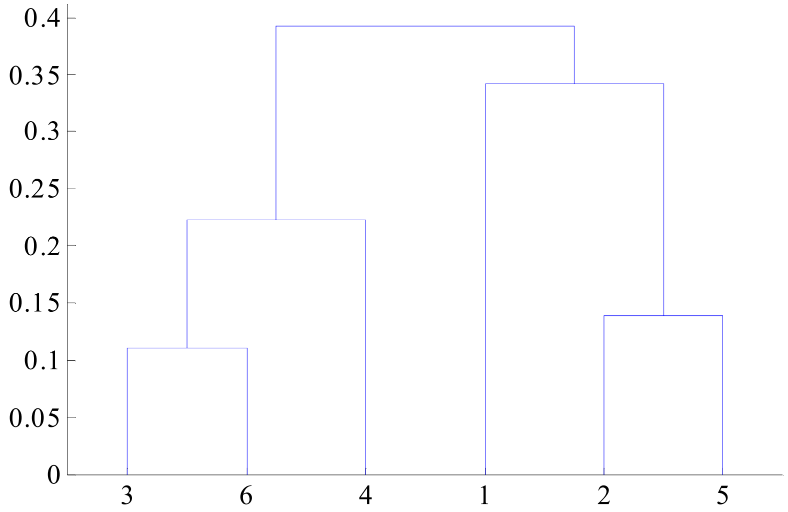
- Less susceptible to noise and outliers.

- Biased towards globular clusters.

Hierarchical Clustering: maximum distance

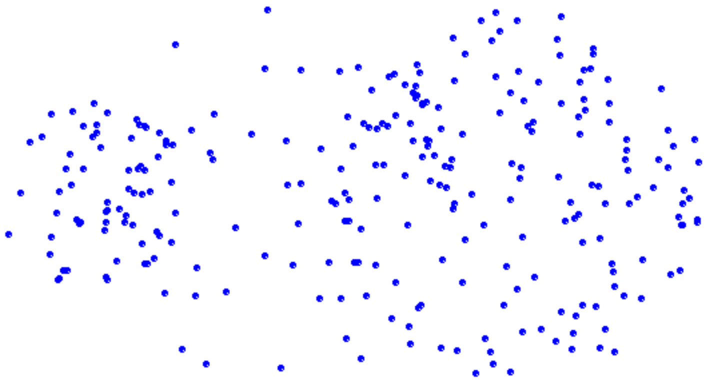


Nested Clusters

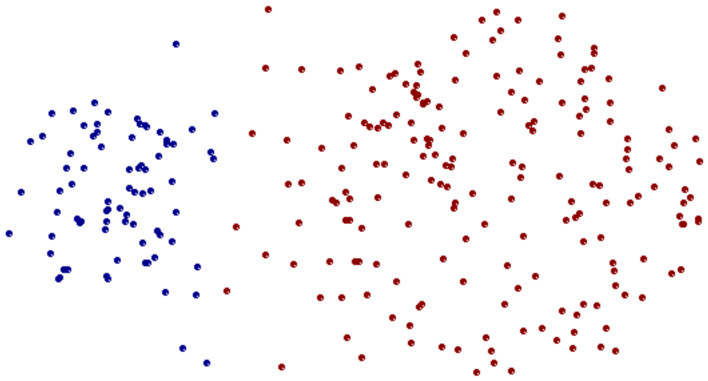


Dendrogram

Strength of maximum distance

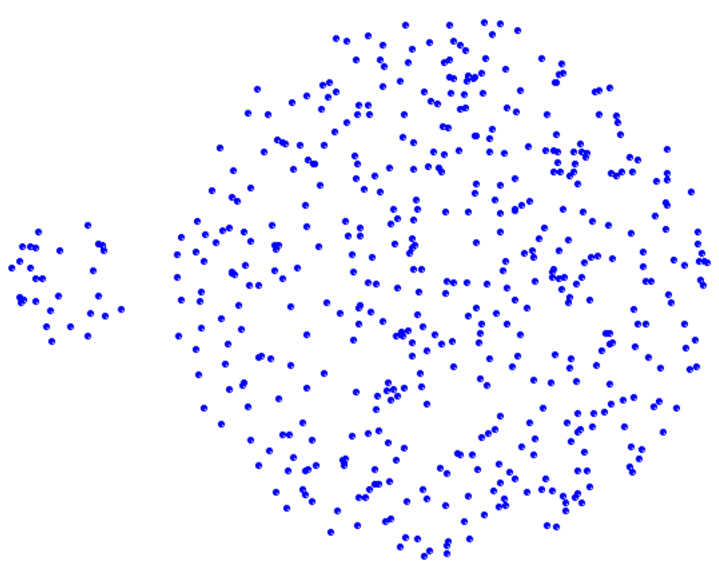


Original Points

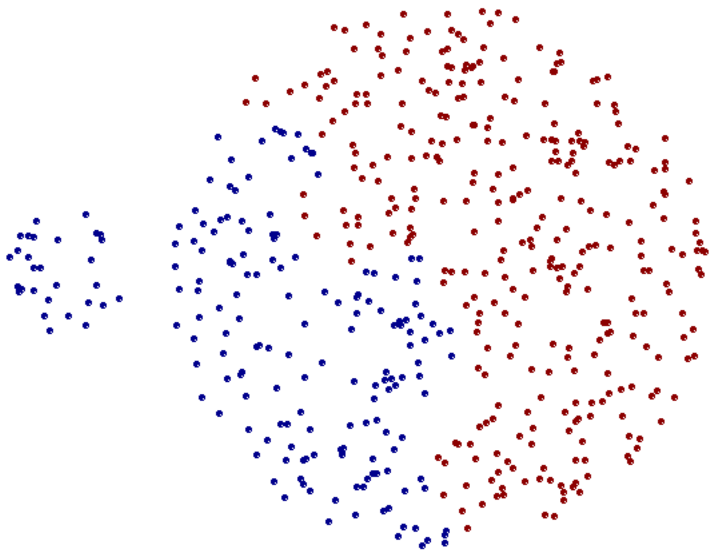


Two Clusters

Limitations of maximum distance



Original Points



Two Clusters

Average linkage (average distance) method

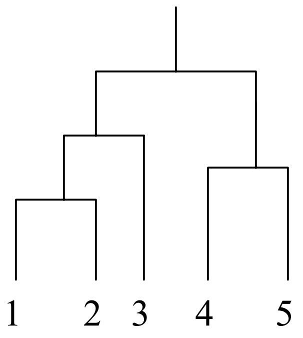
- Distance of two clusters is the average of pairwise distances between points in the two clusters C_{i} and C_j:

$$d(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{x \in C_i} \sum_{y \in C_j} d(x, y)$$

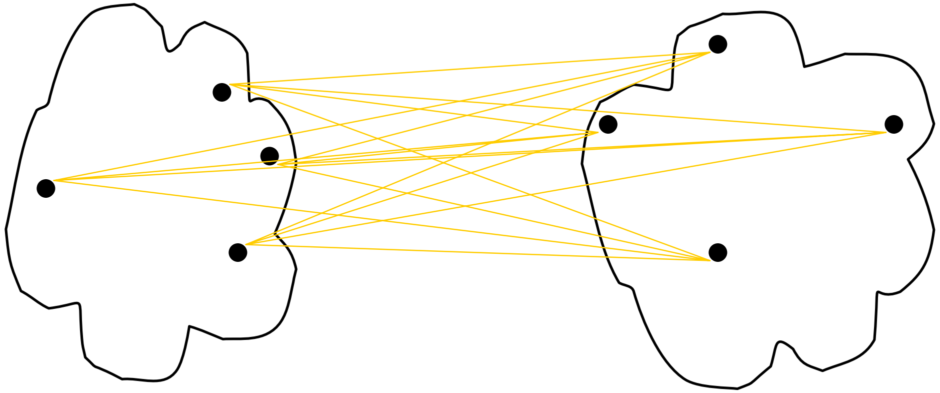
- Compromise between Single and Complete Link.
- Need to use average connectivity for scalability since total connectivity favors large clusters.
- Less susceptible to noise and outliers.
- Biased towards globular clusters.

Similarity matrix

	I1	I2	I3	I4	I5
I1	1.00	0.90	0.10	0.65	0.20
I2	0.90	1.00	0.70	0.60	0.50
I3	0.10	0.70	1.00	0.40	0.30
I4	0.65	0.60	0.40	1.00	0.80
I5	0.20	0.50	0.30	0.80	1.00

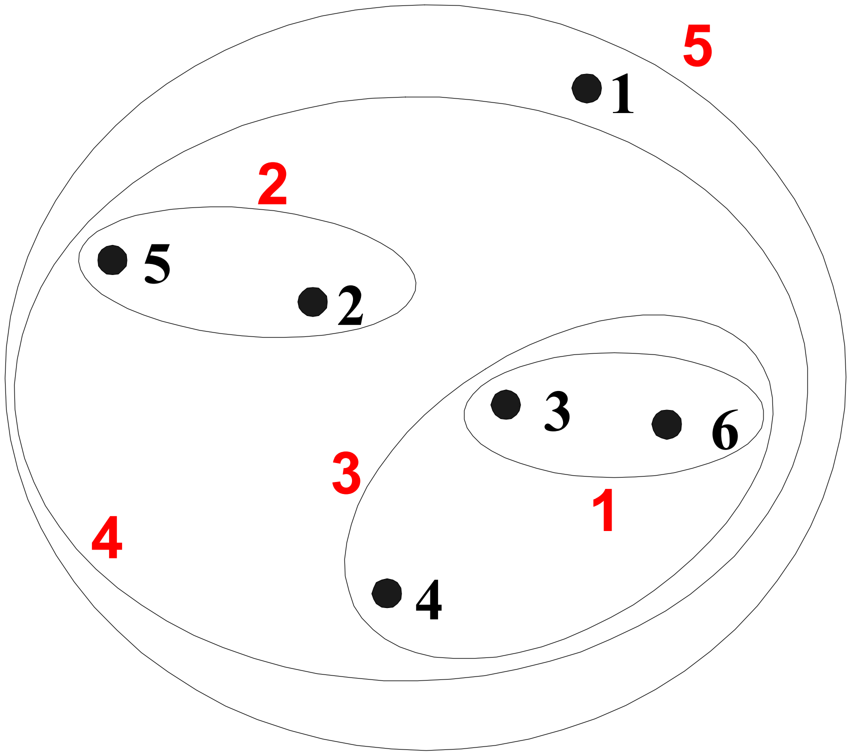


Average linkage

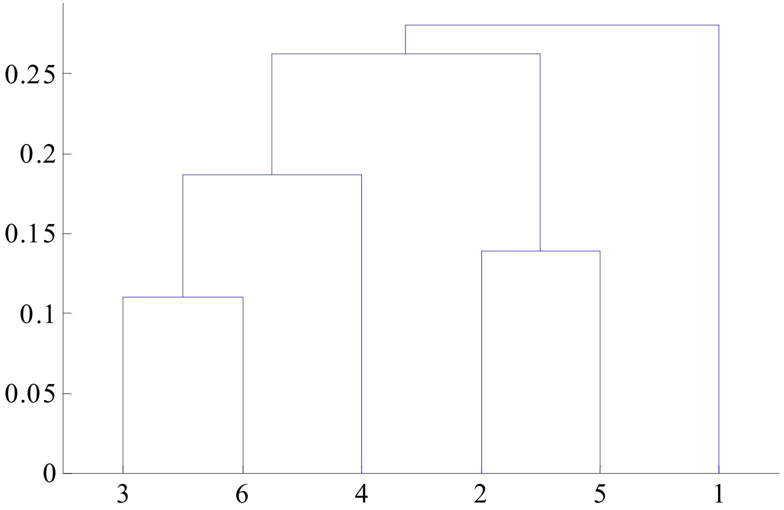


$$d(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{x \in C_i} \sum_{y \in C_j} d(x, y)$$

Hierarchical Clustering: Average distance



Nested Clusters



Dendrogram

Centroid linkage (centroid distance) method

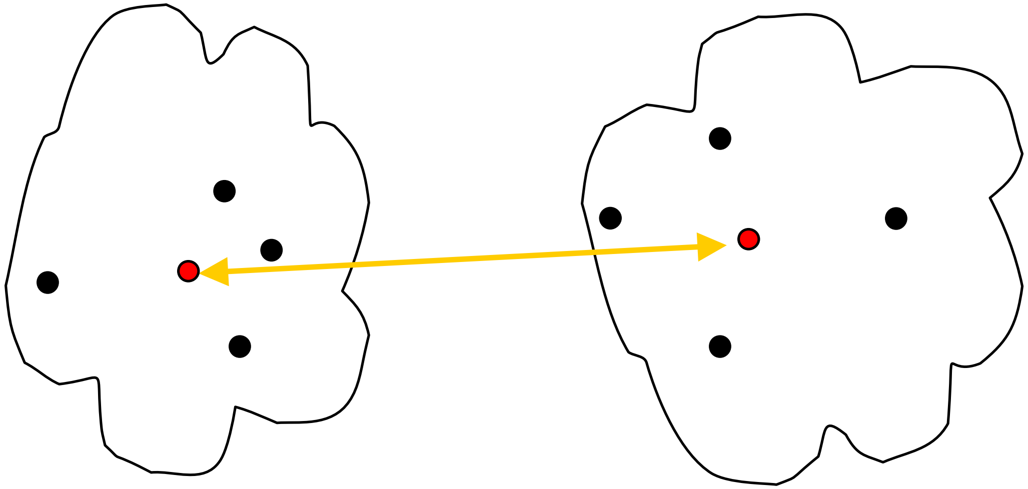
- Distance of two clusters is distance of the two centroids c_i and c_j of the two clusters C_i and C_j:

$$d(C_i, C_j) = d(c_i, c_j)$$

$$c_i = \frac{1}{|C_i|} \sum_{x \in C_i} x \qquad c_j = \frac{1}{|C_j|} \sum_{x \in C_j} x$$

- Compromise between Single and Complete Link.
- Less computationally intensive with respect to average linkage.

Centroid linkage



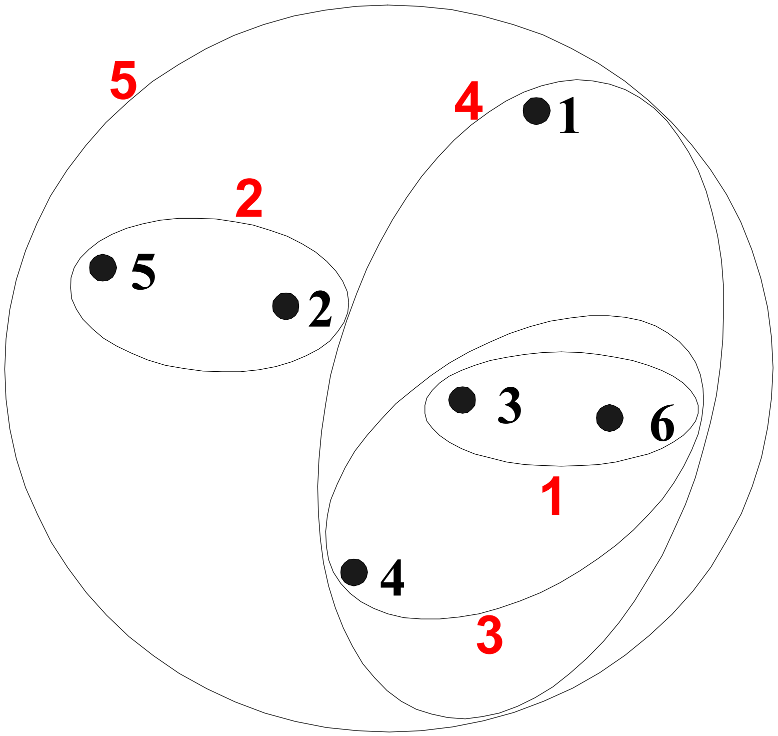
$$d(C_i, C_j) = d(c_i, c_j)$$

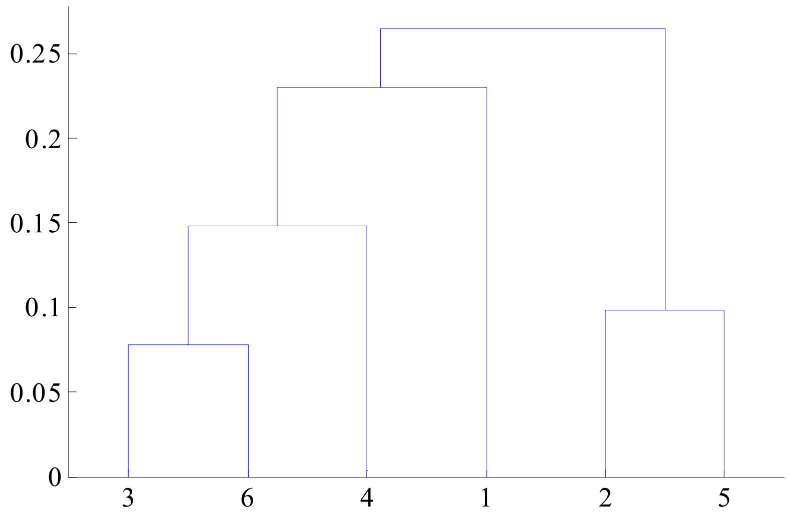
$$c_i = \frac{1}{|C_i|} \sum_{x \in C_i} x \qquad c_j = \frac{1}{|C_j|} \sum_{x \in C_j} x$$

Cluster Similarity: Ward's Method

- Similarity of two clusters is based on the increase in squared error when two clusters are merged.
  - Similar to group average if distance between points is distance squared.
- Less susceptible to noise and outliers.
- Biased towards globular clusters.
- Hierarchical analogue of K-means
  - But Ward's method does not correspond to a local minimum
  - Can be used to initialize K-means

Hierarchical Clustering: Ward's method

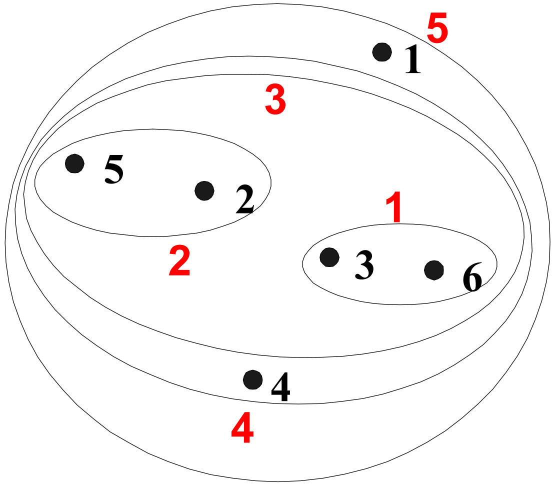




Nested Clusters

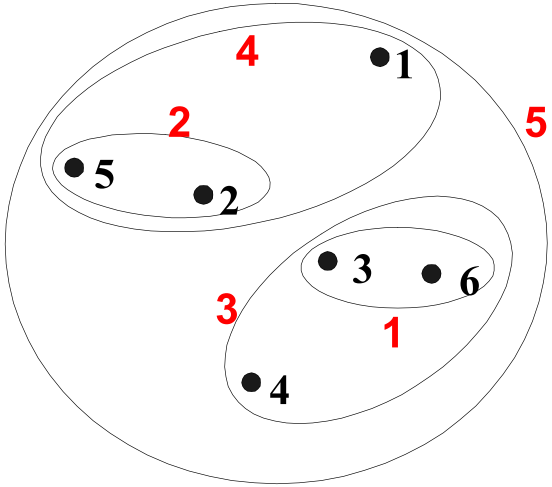
Dendrogram

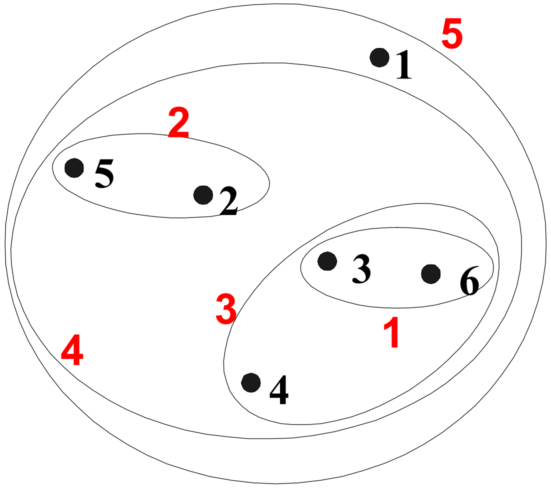
Hierarchical Clustering: comparison



MIN

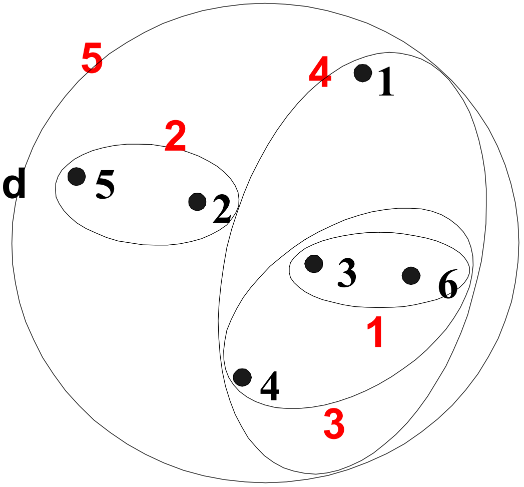
MAX





Average

Ward's Method



Comparison of minimum, maximum, average and centroid distance

Minimum distance

- When d min is used to measure distance between clusters, the algorithm is called the nearest-neighbor or single- linkage clustering algorithm
- If the algorithm is allowed to run until only one cluster remains, the result is a minimum spanning tree (MST)
- This algorithm favors elongated classes

Maximum distance

- When d max is used to measure distance between clusters, the algorithm is called the farthest-neighbor or complete- linkage clustering algorithm
- From a graph- theoretic point of view, each cluster constitutes a complete sub- graph
- This algorithm favors compact classes

Average and centroid distance

- The minimum and maximum distance are extremely sensitive to outliers since their measurement of between- cluster distance involves minima or maxima
- The average and centroid distance approaches are more robust to outliers
- Of the two, the centroid distance is computationally more attractive
- Notice that the average distance approach involves the computation of $|C_i||C_j|$ distances for each pair of clusters

Hierarchical Clustering: Time and Space requirements

- $O(N^2)$ space since it uses the proximity matrix.
  - N is the number of points.

- $O(N^3)$ time in many cases.
  - There are N steps and at each step the size, N^2, proximity matrix must be updated and searched.
  - By being careful, the complexity can be reduced to $O(N^2 \log(N))$ time for some approaches.

Hierarchical Clustering: problems and limitations

- Once a decision is made to combine two clusters, it cannot be undone.
- No objective function is directly minimized.
- Different schemes have problems with one or more of the following:
  - Sensitivity to noise and outliers.
  - Difficulty handling different sized clusters and convex shapes.
  - Breaking large clusters.

Advantages and disadvantages of Hierarchical clustering

Advantages

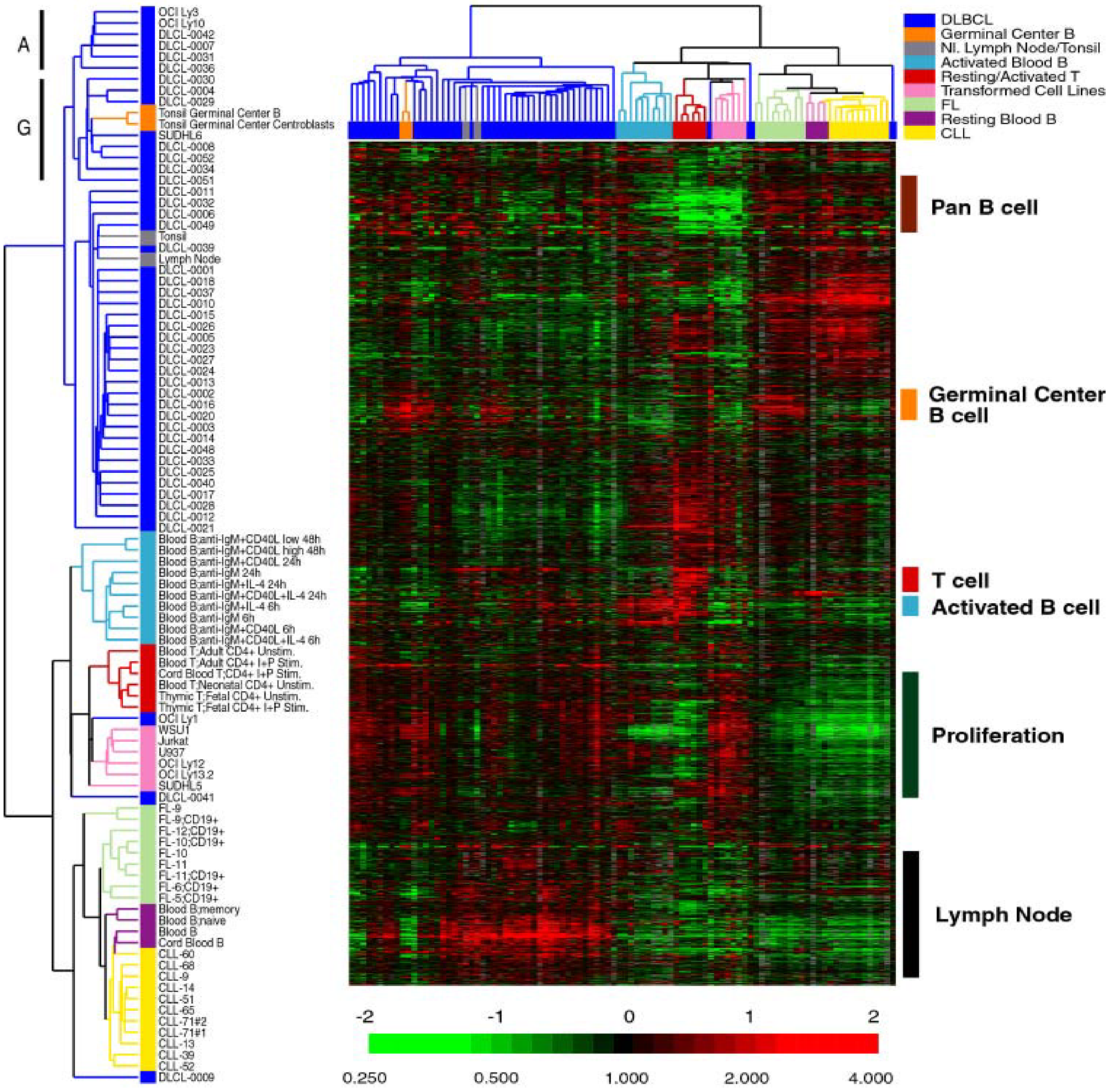
- Does not require the number of clusters to be known in advance
- No input parameters (besides the choice of the (dis)similarity)
- Computes a complete hierarchy of clusters
- Good result visualizations integrated into the methods

Disadvantages

- May not scale well: runtime for the standard methods: $O(n^2 \log n)$
- No explicit clusters: a “flat” partition can be derived afterwards (e.g. via a cut through the dendrogram or termination condition in the construction)
- No automatic discovering of “optimal clusters”

Hierarchical clustering of tissues and genes:

Alizadeh et al. 2000, Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling, Nature 403:3.


A
G
OCI Ly3
OCI Ly10
DLCL-0042
DLCL-0007
DLCL-0031
DLCL-0036
DLCL-0030
DLCL-0004
DLCL-0029
Tonsil Germinal Center B
Tonsil Germinal Center Centroblasts
SUDHL6
DLCL-0008
DLCL-0052
DLCL-0034
DLCL-0051
DLCL-0011
DLCL-0032
DLCL-0006
DLCL-0049
Tonsil
DLCL-0039
Lymph Node
DLCL-0001
DLCL-0018
DLCL-0037
DLCL-0010
DLCL-0015
DLCL-0026
DLCL-0005
DLCL-0023
DLCL-0027
DLCL-0024
DLCL-0013
DLCL-0002
DLCL-0016
DLCL-0020
DLCL-0003
DLCL-0014
DLCL-0048
DLCL-0033
DLCL-0025
DLCL-0040
DLCL-0017
DLCL-0028
DLCL-0012
DLCL-0021
Blood B;anti-IgM+CD40L low 48h
Blood B;anti-IgM+CD40L high 48h
Blood B;anti-IgM+CD40L 24h
Blood B;anti-IgM 24h
Blood B;anti-IgM+IL-4 24h
Blood B;anti-IgM+CD40L+IL-4 24h
Blood B;anti-IgM+IL-4 6h
Blood B;anti-IgM 6h
Blood B;anti-IgM+CD40L 6h
Blood B;anti-IgM+CD40L+IL-4 6h
Blood T;Adult CD4+ Unstim.
Blood T;Adult CD4+ I+P Stim.
Cord Blood T;CD4+ I+P Stim.
Blood T;Neonatal CD4+ Unstim.
Thymic T;Fetal CD4+ Unstim.
Thymic T;Fetal CD4+ I+P Stim.
OCI Ly1
WSU1
Jurkat
U937
OCI Ly12
OCI Ly13.2
SUDHL5
DLCL-0041
FL-9
FL-9;CD19+
FL-12;CD19+
FL-10;CD19+
FL-10
FL-11
FL-11;CD19+
FL-6;CD19+
FL-5;CD19+
Blood B;memory
Blood B;naive
Blood B
Cord Blood B
CLL-60
CLL-68
CLL-9
CLL-14
CLL-51
CLL-65
CLL-71#2
CLL-71#1
CLL-13
CLL-39
CLL-52
DLCL-0009
DLBCL
Germinal Center B
Nl. Lymph Node/Tonsil
Activated Blood B
Resting/Activated T
Transformed Cell Lines
FL
Resting Blood B
CLL
Pan B cell
Germinal Center B cell
T cell
Activated B cell
Proliferation
Lymph Node