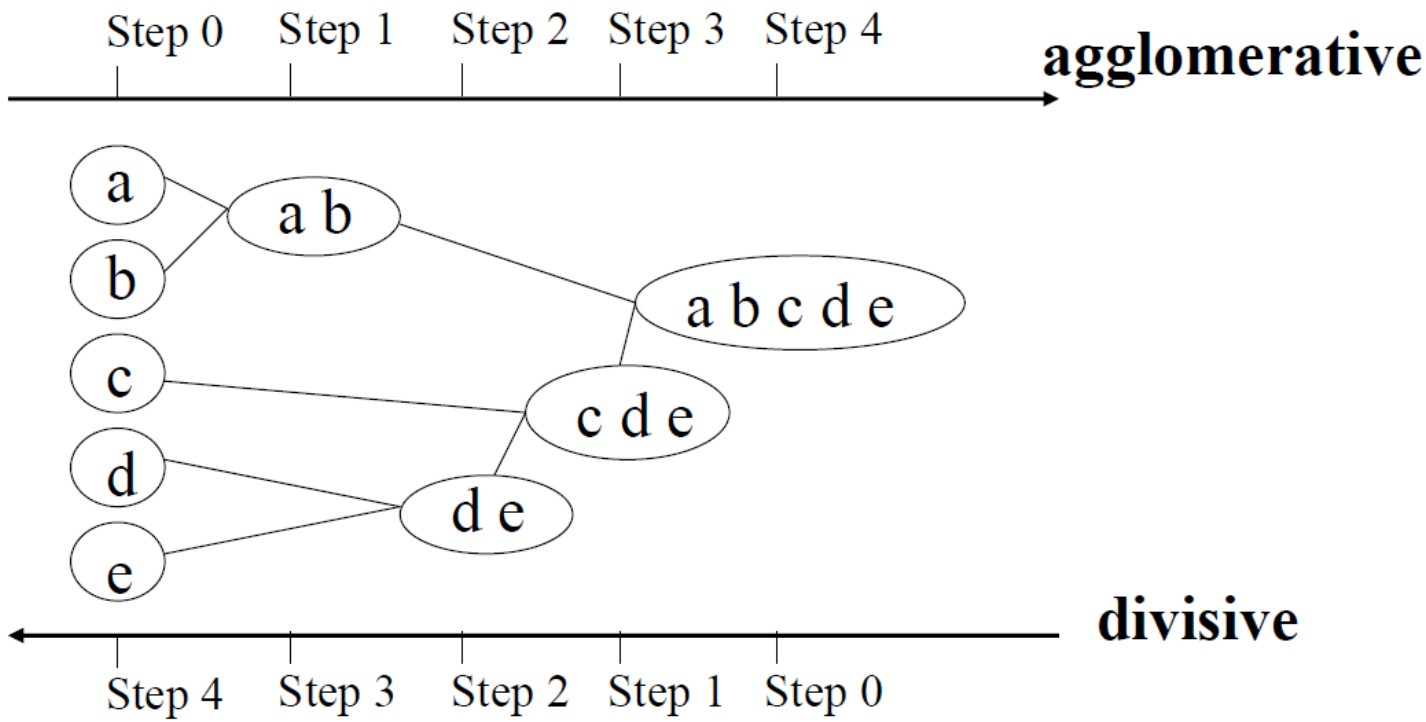
Hierarchical clustering

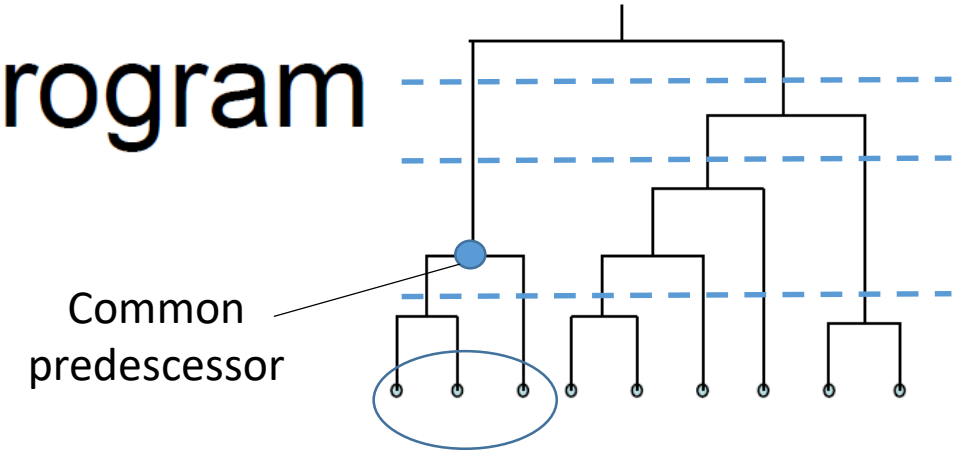
First dendrogram : R. Ling, 1973

Hierarchical Clustering

- Group data objects into a tree of clusters

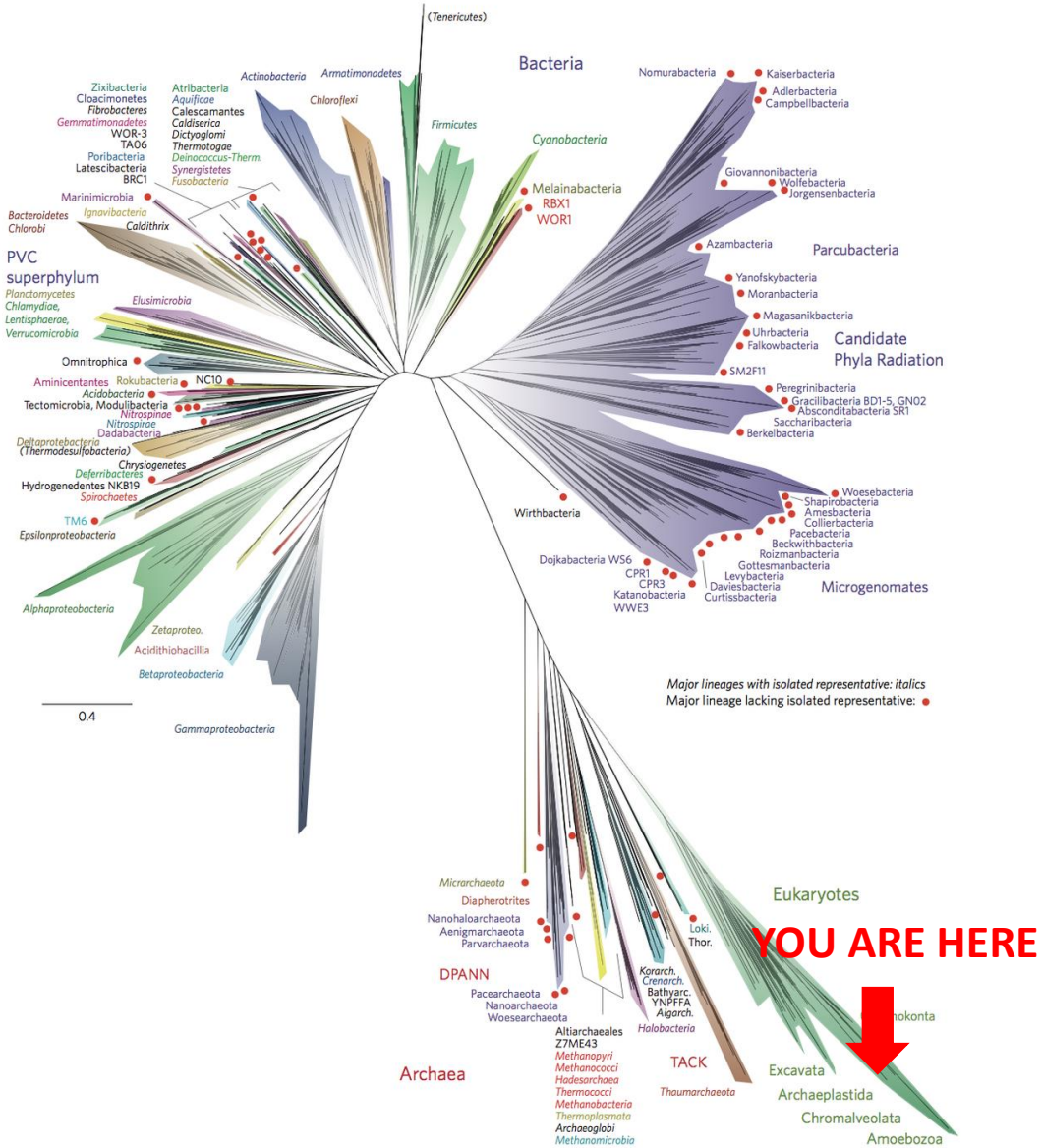


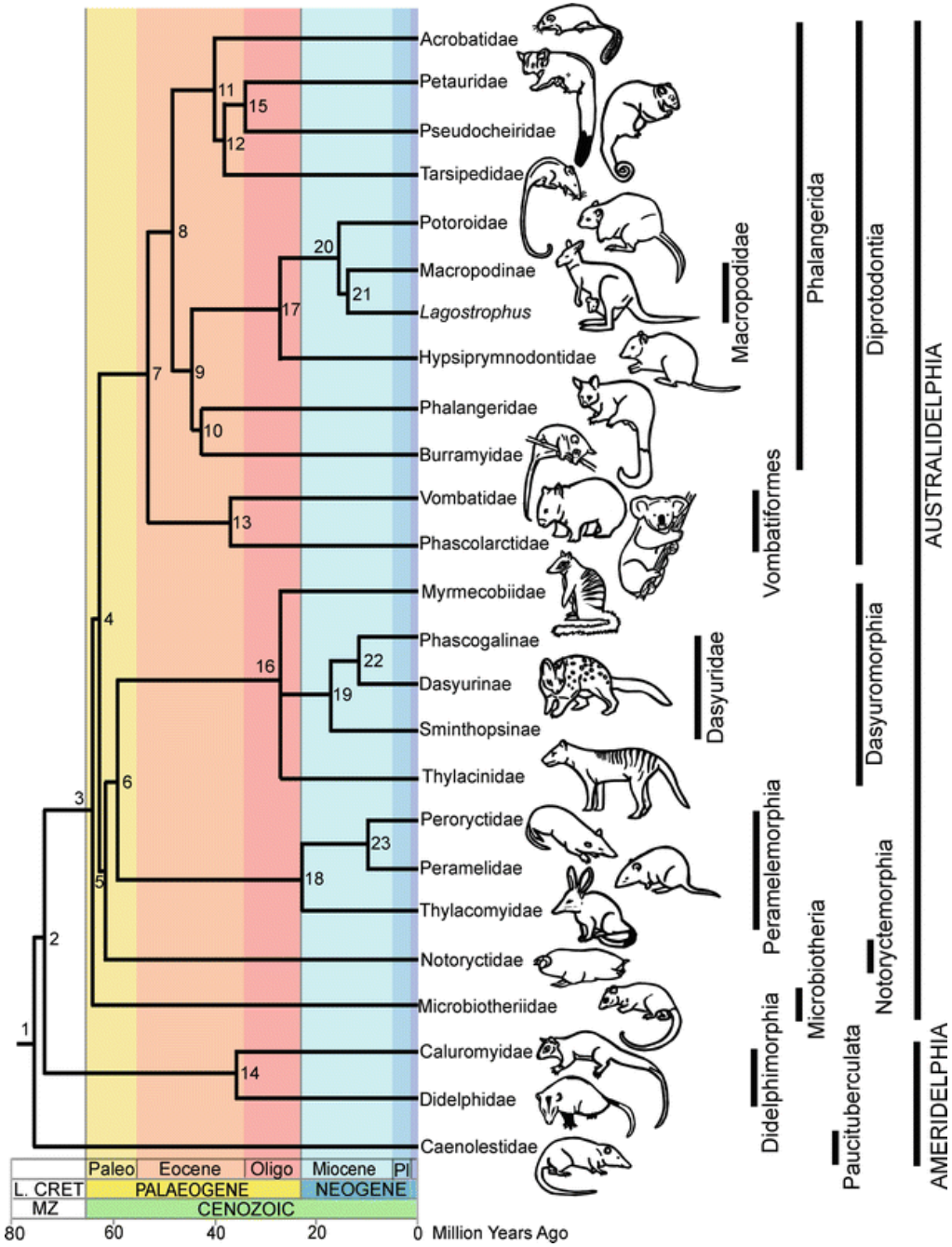
Dendrogram



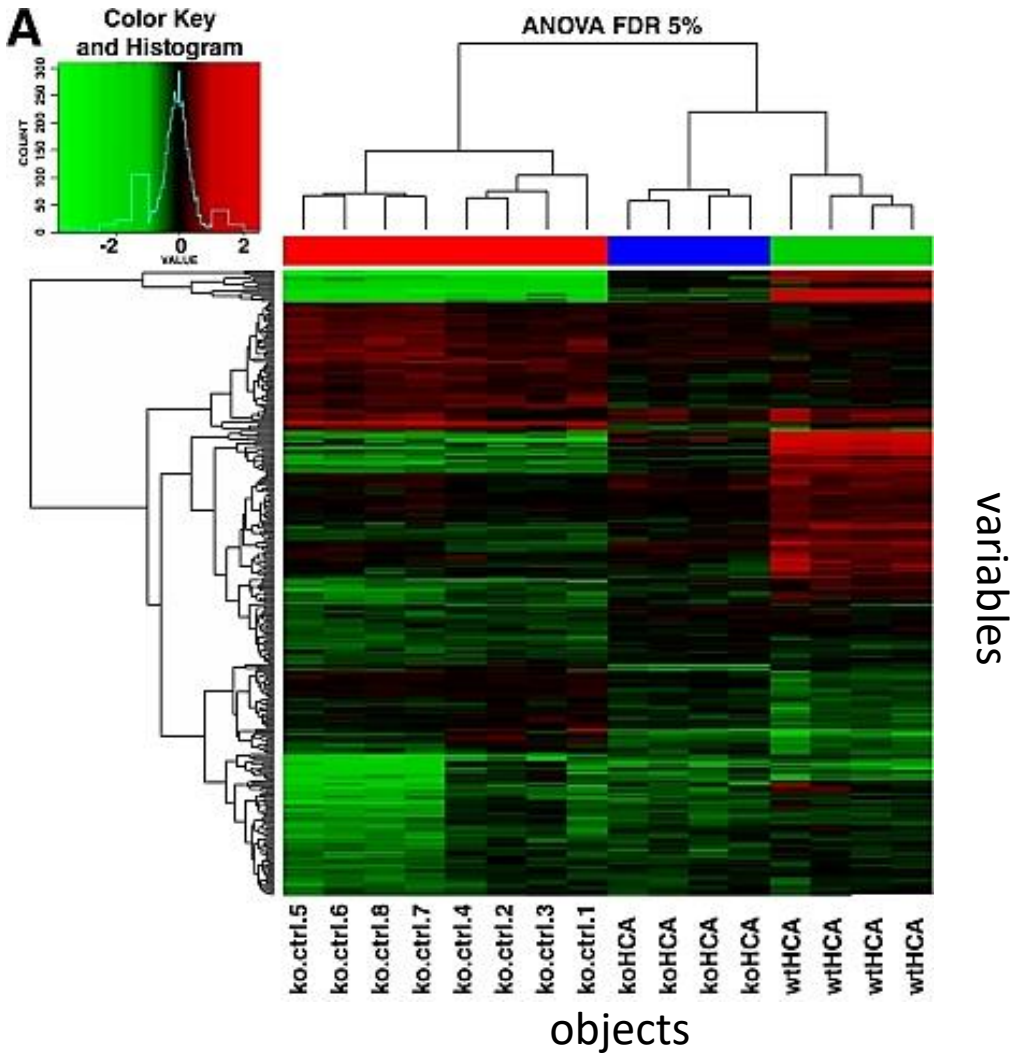
- Show how to merge clusters hierarchically
- Decompose data objects into a multi-level nested partitioning (a tree of clusters)
- A clustering of the data objects: cutting the dendrogram at the desired level
  - Each connected component forms a cluster

Example of dendrogram: tree of life





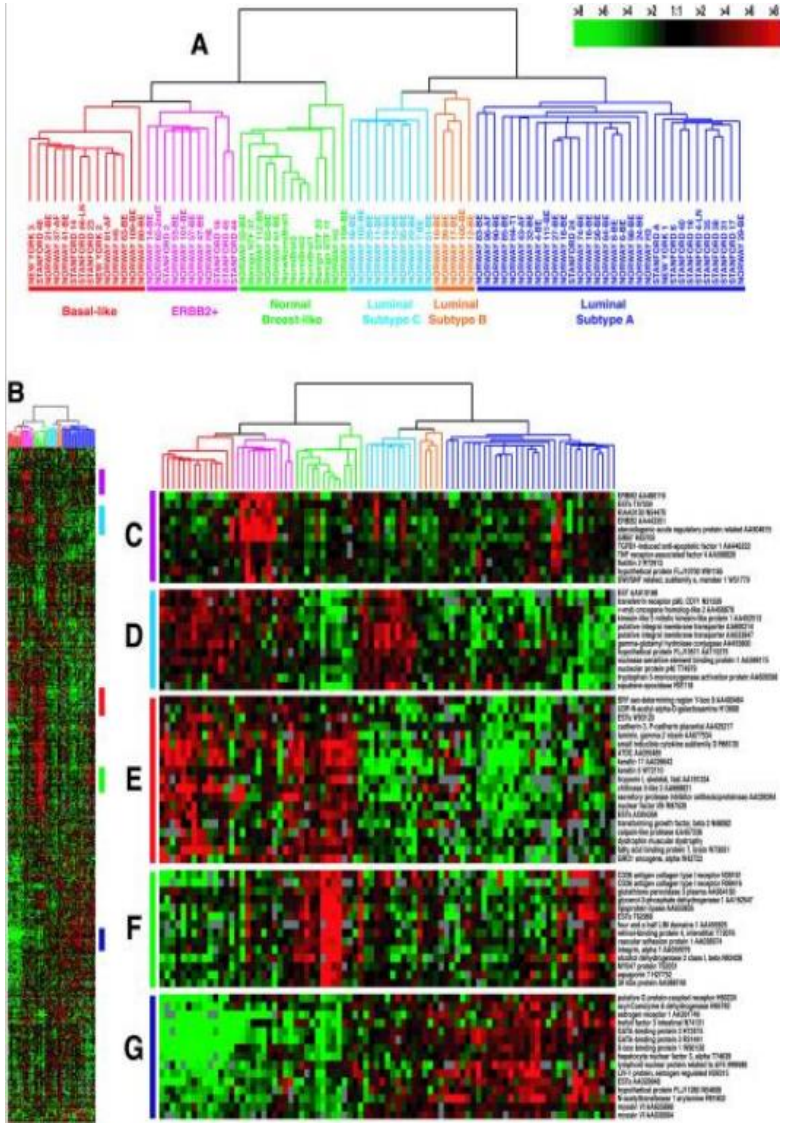
Heatmap representation of data



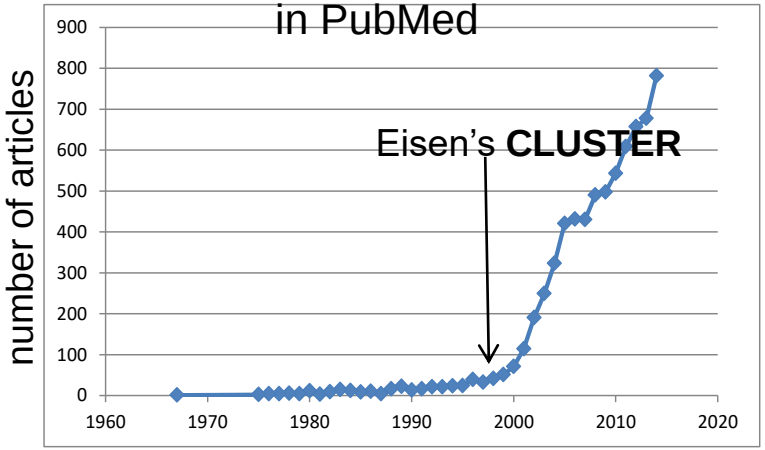
Dendrogram + Heatmap = killer application in life sciences!*

* In 1990-2010s, lost in popularity in the last years

Hierarchical clustering for studying cancer



Sorlie, PNAS 2001



Agglomerative algorithm(s)

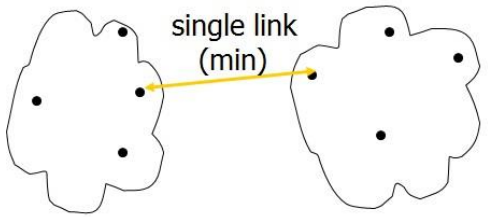
- Initially, each object is a cluster
- Step-by-step **merging of the closest clusters**, until all objects form a single cluster

– One needs to define:

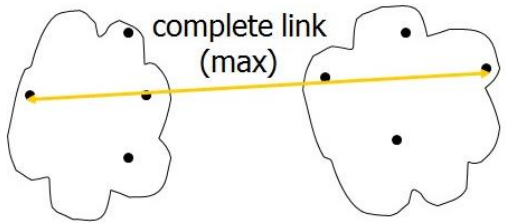
Distance metrics between data points

Distance metrics between groups of data points

Dissimilarity (‘distance’) between clusters

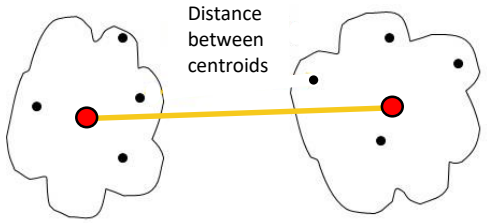


$$d_{\min}(C_i, C_j) = \min_{p \in C_i, q \in C_j} d(p, q)$$

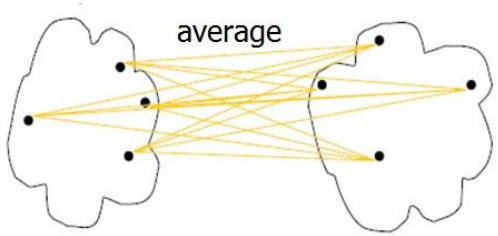


$$d_{\max}(C_i, C_j) = \max_{p \in C_i, q \in C_j} d(p, q)$$

Special case!
Full data matrix is needed!

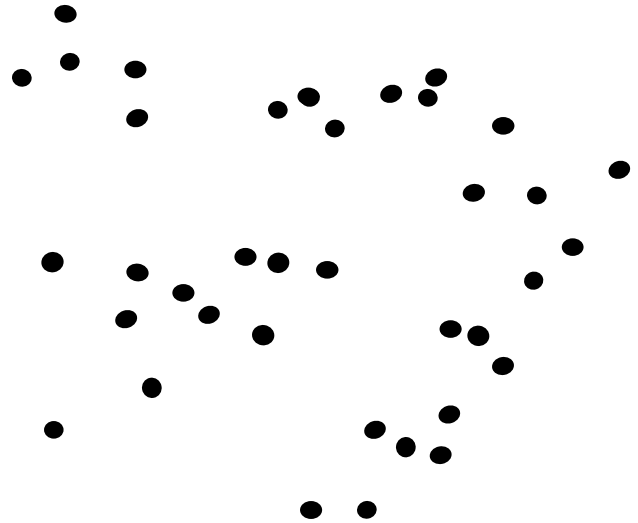


$$d_{mean}(C_i, C_j) = d(m_i, m_j)$$



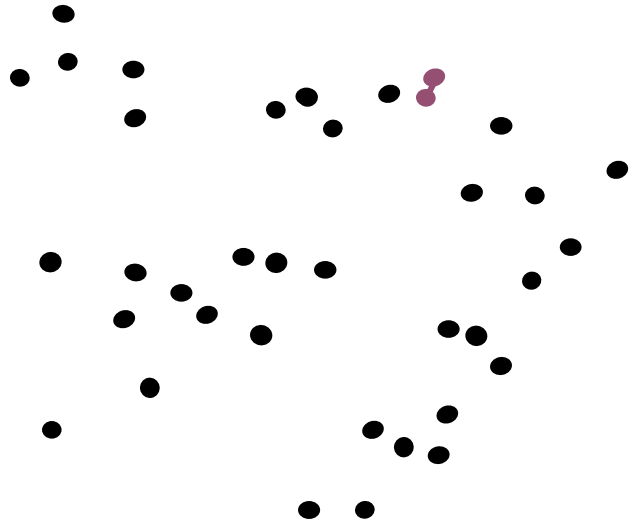
$$d_{avg}(C_i, C_j) = \frac{1}{n_i n_j} \sum_{p \in C_i} \sum_{q \in C_j} d(p, q)$$

Example: single linkage, Euclidean distance



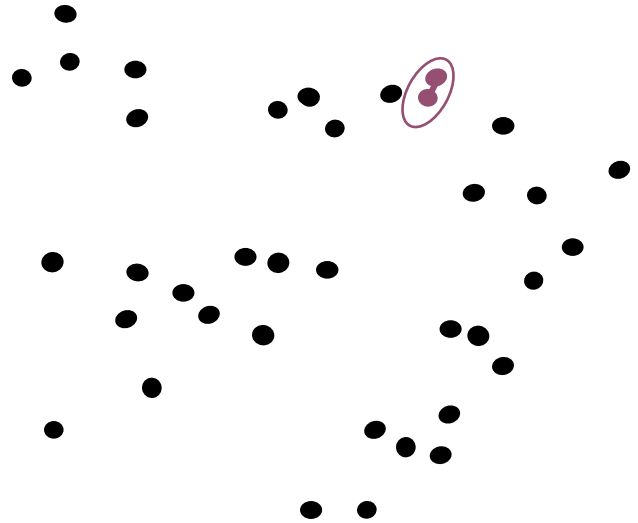
1. Say “Every point is its own cluster”

Example: single linkage, Euclidean distance



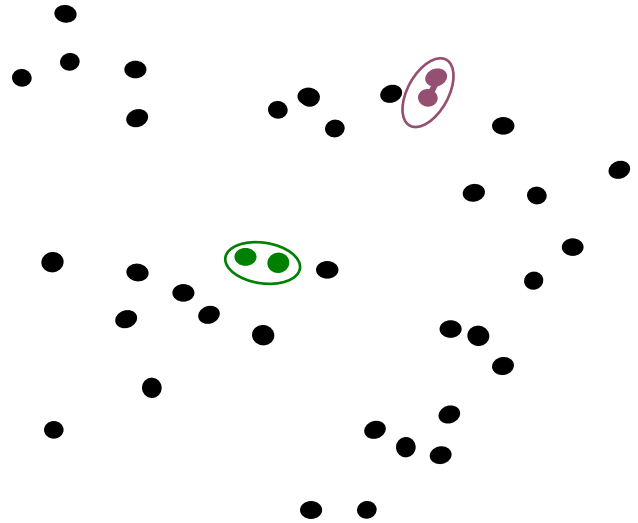
1. Say “Every point is its own cluster”
2. Find “most similar” pair of clusters

Example: single linkage, Euclidean distance



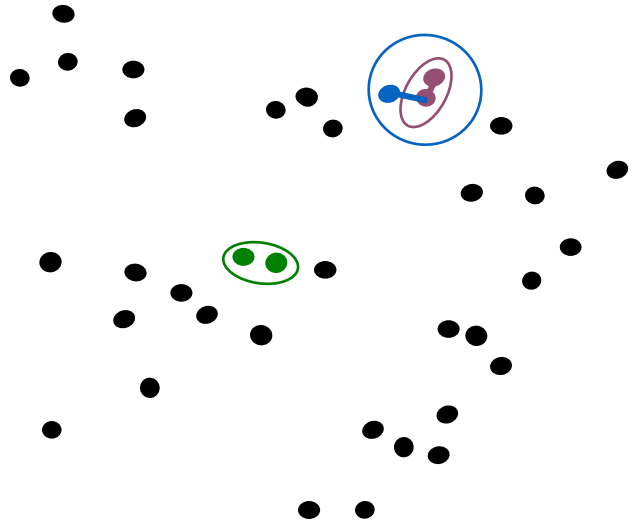
1. Say “Every point is its own cluster”
2. Find “most similar” pair of clusters
3. Merge it into a parent cluster

Example: single linkage, Euclidean distance

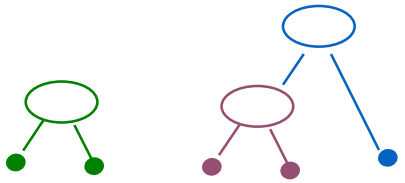


1. Say "Every point is its own cluster"
2. Find "most similar" pair of clusters
3. Merge it into a parent cluster
4. Repeat

Example: single linkage, Euclidean distance



1. Say "Every point is its own cluster"
2. Find "most similar" pair of clusters
3. Merge it into a parent cluster
4. Repeat

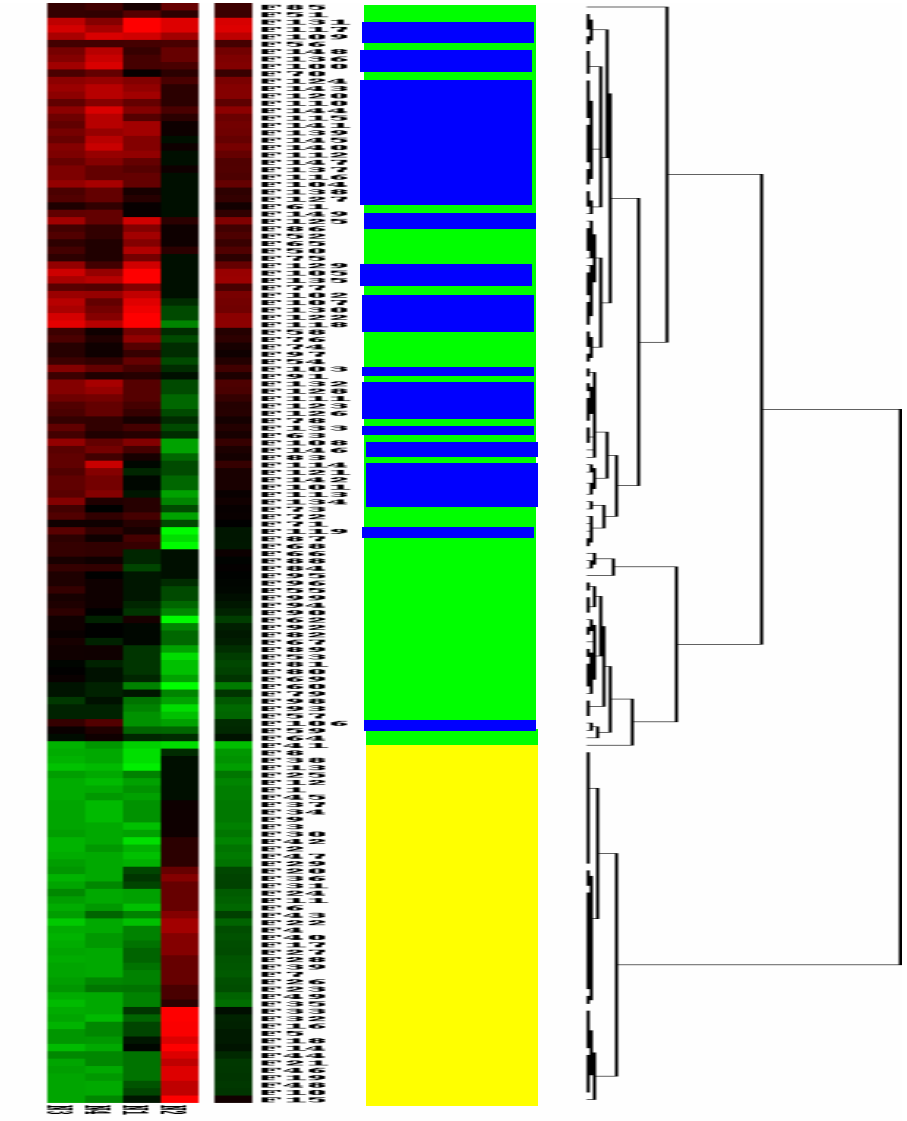


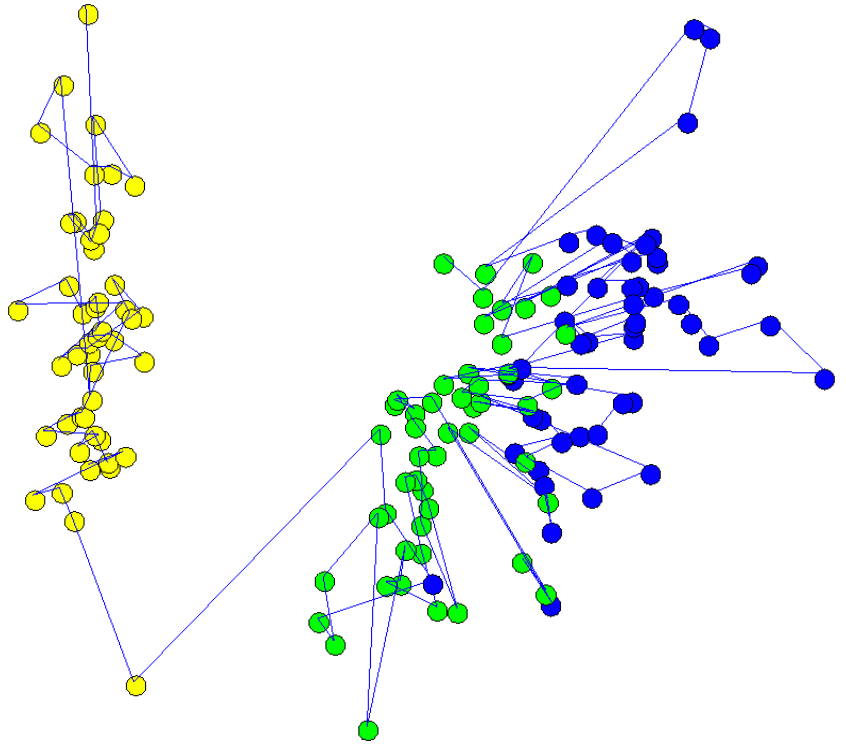
Pros and cons of hierarchical clustering

- It’s nice that you get a hierarchy instead of an amorphous collection of groups
- If you want k groups, just cut the (k-1) longest links
- Clusters can have complex shapes
- Can work with any dissimilarity measure

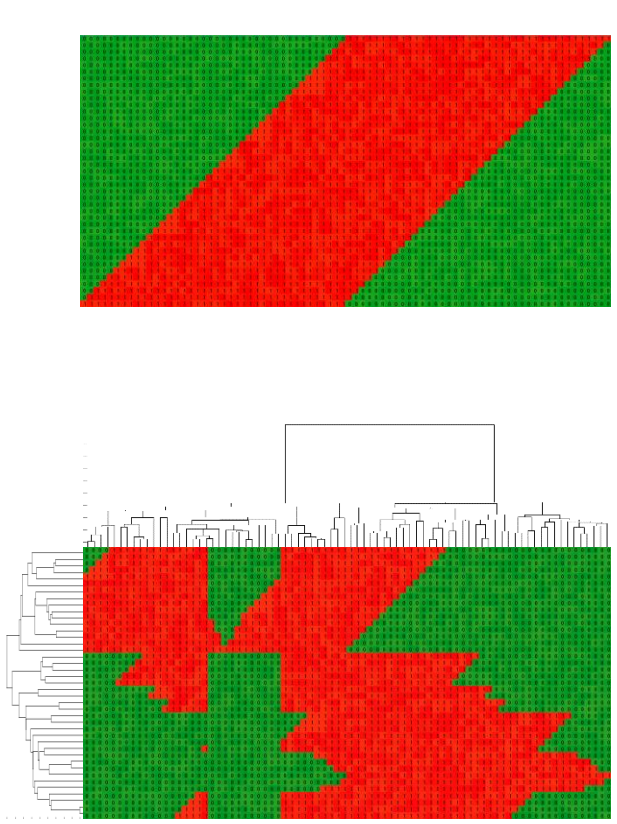
- There’s no real statistical or information-theoretic foundation to this
- Do not scale well: $O(n^2)$ or even $O(m^2n^2)$
- Uses complete distance matrix – challenge with memory
- Problem with representation of the dendrogram (leaves order)
- Might be unstable

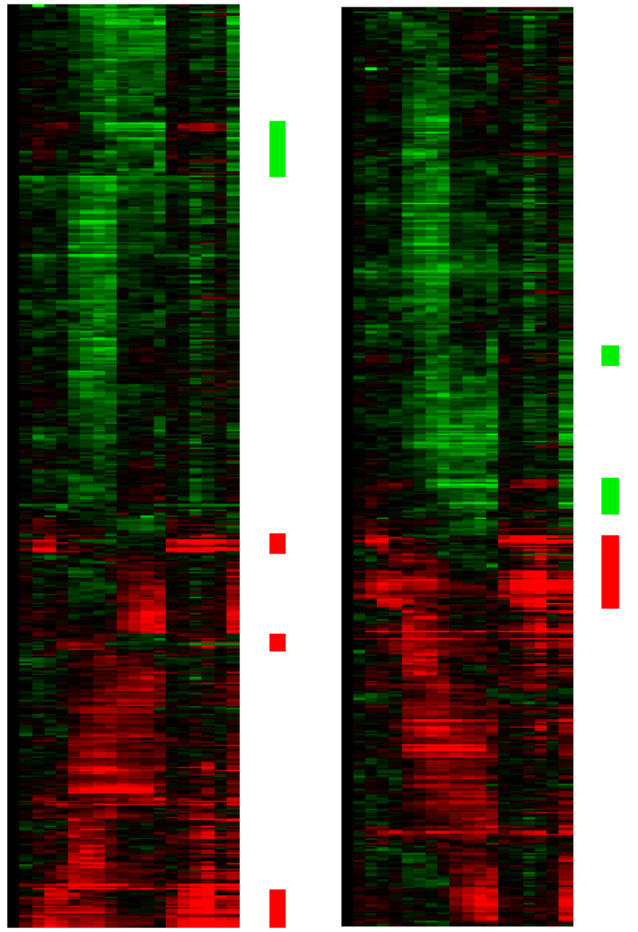
Problem of leaves ordering is ill-posed





Hierarchical clustering and dendrograms: leaves order





(b) Eisen's (a) Optimal

Biedl et al, 2001; Bar-Joseph et al., 2003

Hierarchical clustering and dendrograms: cluster instability

- Hierarchical clustering results can be very sensitive to a random removal of a small percentage of points