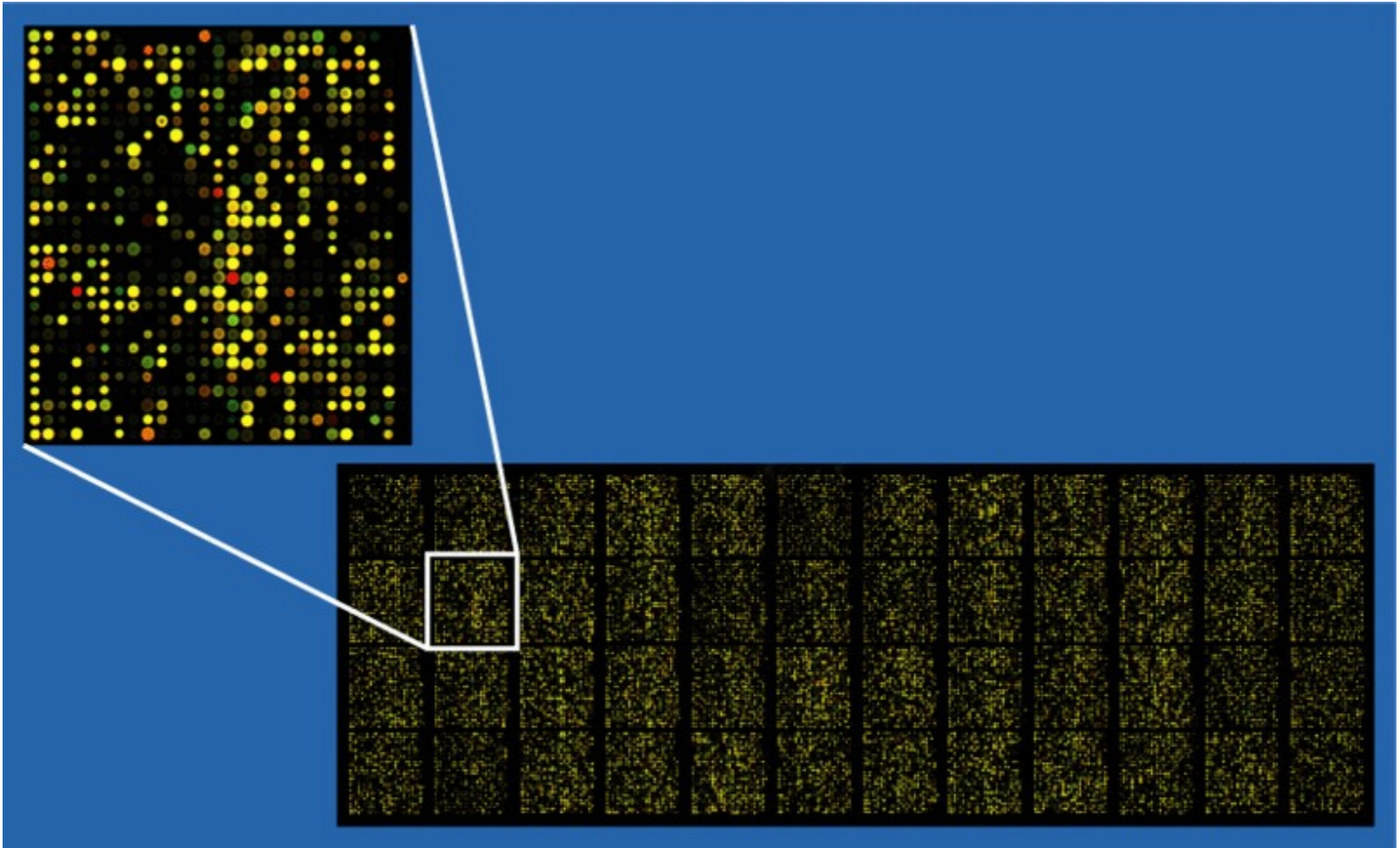


# CMSC423: Bioinformatic Algorithms, Databases and Tools

Data clustering

# Why data clustering?



What does this mean?

# Data clustering...

>F4BT0V001CZSIM rank=0000138 x=1110.0 y=2700.0 length=57  
ACTGCTCTCATGCTGCCTCCCGTAGGAGTGCCTCCCTGAGCCAGGATCAAACGTCTG

>F4BT0V001BBJQS rank=0000155 x=424.0 y=1826.0 length=47  
ACTGACTGCATGCTGCCTCCCGTAGGAGTGCCTCCCTGCGCCATCAA

>F4BT0V001EDG35 rank=0000182 x=1676.0 y=2387.0 length=44  
ACTGACTGCATGCTGCCTCCCGTAGGAGTCGCCGTCCCTCGACNC

>F4BT0V001D2HQQ rank=0000196 x=1551.0 y=1984.0 length=42  
ACTGACTGCATGCTGCCTCCCGTAGGAGTGCCGTCCCTCGAC

>F4BT0V001CM392 rank=0000206 x=966.0 y=1240.0 length=82  
AANCAGCTCTCATGCTCGCCCTGACTTGGCATGTGTTAAGCCTGTAGGCTAGCGTTCATC  
CCTGAGCCAGGATCAAACCTCTG

>F4BT0V001EIMFX rank=0000250 x=1735.0 y=907.0 length=46  
ACTGACTGCATGCTGCCTCCCGTAGGAGTGTGCGGCCATCAGACTG

>F4BT0V001ENDKR rank=0000262 x=1789.0 y=1513.0 length=56  
GACACTGTCATGCTGCCTCCCGTAGGAGTGCCTCCCTGAGCCAGGATCAAACCTCTG

>F4BT0V001D91MI rank=0000288 x=1637.0 y=2088.0 length=56  
ACTGCTCTCATGCTGCCTCCCGTAGGAGTGCCTCCCTGAGCCAGGATCAAACCTCTG

>F4BT0V001D0Y5G rank=0000341 x=1534.0 y=866.0 length=75  
GTCTGTGACATGCTGCCTCCCGTAGGAGTCTACACAAGTTGTGGCCCAGAACCACTGAGC  
CAGGATCAAACCTCTG

>F4BT0V001EMLE1 rank=0000365 x=1780.0 y=1883.0 length=84  
ACTGACTGCATGCTGCCTCCCGTAGGAGTGCCTCCCTGCGCCATCAATGCTGCATGCTGC  
TCCCTGAGCCAGGATCAAACCTCTG

# Data clustering...

- Given a collection of data-points can we identify any patterns?
- Data-points:
  - DNA sequences
  - Gene expression levels
  - Organism abundances in an environment
  - Vitals
- Patterns:
  - do certain points group together?

# Example

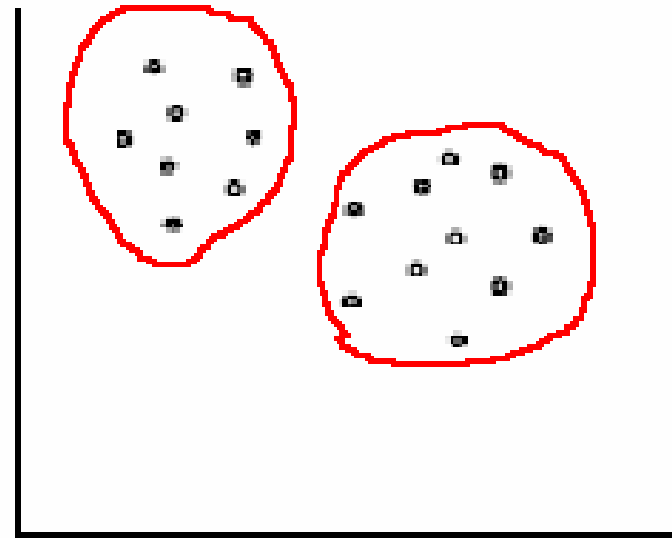
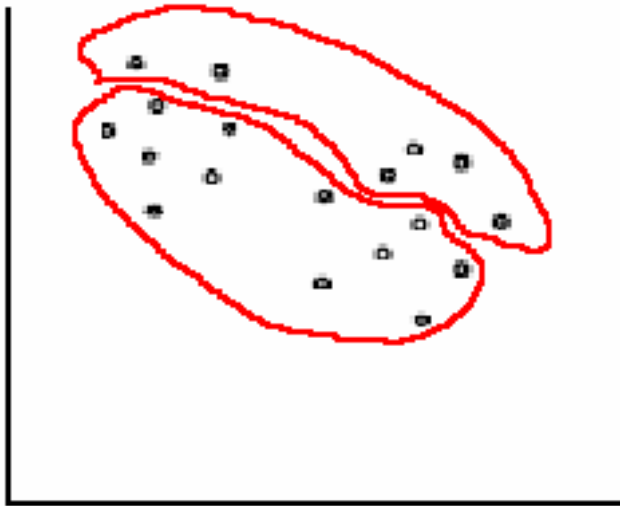
- 16S rRNA sequences from gut microbiome
- How many types of organisms are there?
- Assumption: organisms that are similar have similar 16S sequences
- Operational Taxonomic Unit – sequences within 2% distance from each other
- Next step: given OTU abundances for a number of patients, do the patients cluster by age/disease/etc?

# Types of clustering algorithms

- Agglomerative
  - Start with single observations
  - Group similar observations into the same cluster
- Divisive
  - All datapoints start in the same cluster
  - Iteratively divide cluster until you find good clustering
- Hierarchical
  - Build a tree – leaves are datapoints, internal nodes represent clusters

# Measures of goodness of clustering

- Homogeneity
  - All points in a cluster must be similar
- Separation
  - Points in different clusters are dissimilar



# Microarray (gene abundance) clustering

- Each gene can be viewed as an array of numbers
  - expression of gene at different time-points
  - expression of gene in different conditions (normal, variants of a disease, etc.)
- Each time-point or tissue sample can also be viewed as an array of numbers
  - expression levels for all genes
- Basic idea: cluster genes and/or samples to highlight genes involved in disease



# Hierarchical clustering

- Need: definition of distance between data-points (e.g. individual genes).
- Some measures:
  - Euclidean distance  $D(x, y) = \sqrt{\sum_i (x_i - y_i)^2}$
  - Manhattan distance  $D(x, y) = \sum_i |x_i - y_i|$
  - Pearson correlation  $D(x, y) = \frac{E[(x - \mu_x)(y - \mu_y)]}{\sigma_x \sigma_y}$
  - Angle between vectors (centered Pearson correlation)
- Clustering algorithm
  - group together data-points that are most similar
  - repeat...

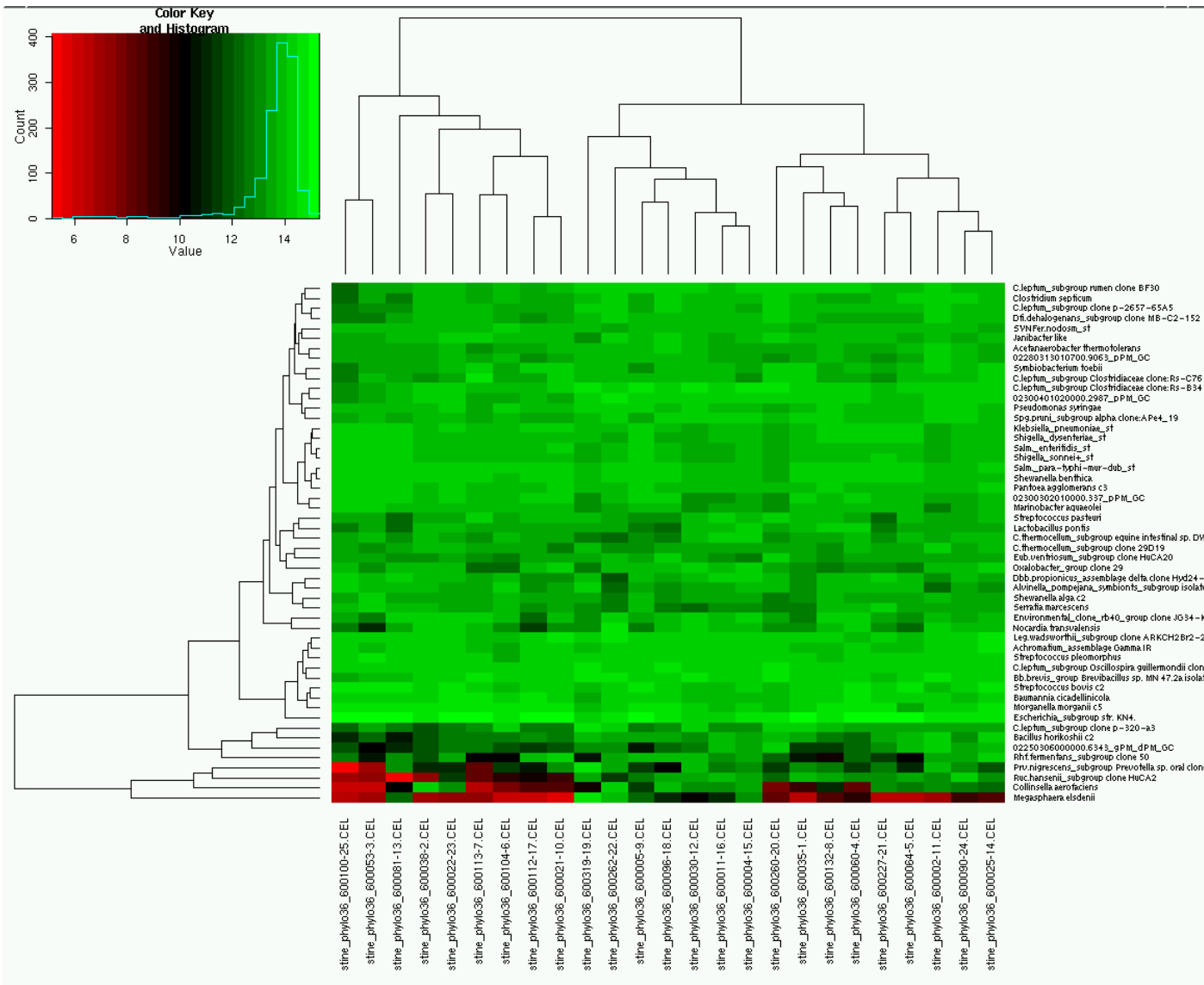
# Hierarchical clustering

- Key element: how do you compute distance between two clusters, or a point and a cluster ?
- UPGMA/average neighbor (average linkage)
  - average distance between all genes in the two clusters
- Furthest neighbor (complete linkage)
  - largest distance between all genes in clusters
- Nearest neighbor (single linkage)
  - smallest distance between all genes in clusters
- Ward's distance
  - inter-cluster distance is variance of inter-gene distances

# Hierarchical clustering...cont

- Irrespective of distance choice, algorithm is the same
  1. compute inter-gene/cluster distances
  2. join together pair of genes/clusters with smallest distance
  3. recompute distances to include the newly created cluster
  4. repeat until all points in one cluster
- Output of program is a tree
- Cluster sets – defined by “cut” nodes – any subset of internal tree nodes defines a set of clusters – the sets of leaves in the corresponding subtrees
- Choice of cut can be tricky – usually problem-specific
- Note: all this can be done easily in R

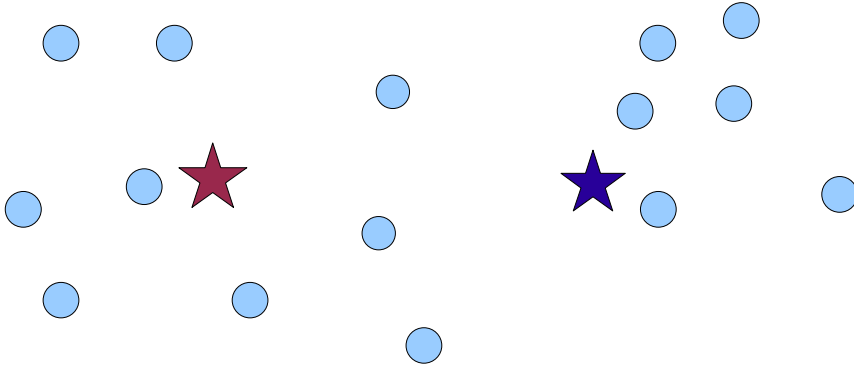
# Example: gut microbiome in children



# k-means clustering

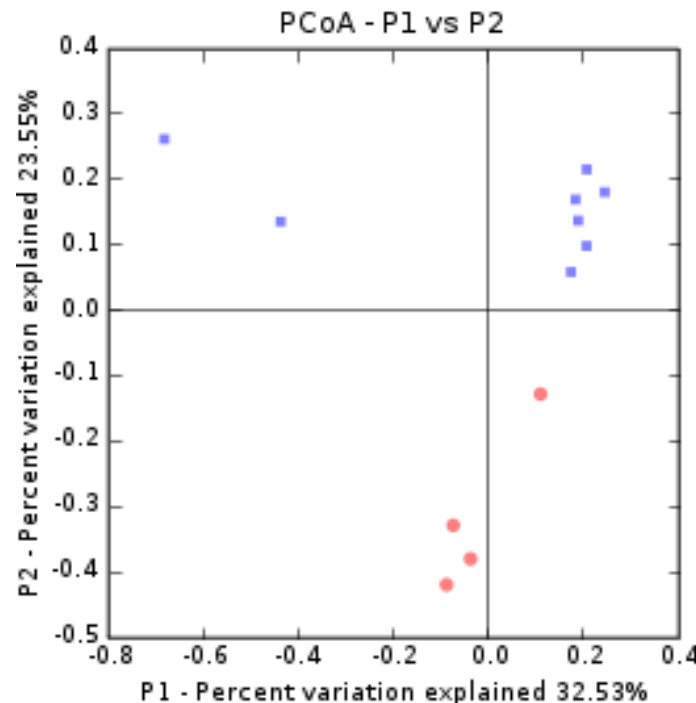
- Goal: split data into exactly  $k$  clusters
- Basic algorithm:
  - Create  $k$  arbitrary clusters - pick  $k$  points as cluster centers and assign each other point to the closest center
  - Re-compute the center of each cluster
  - Re-assign points to clusters
  - Repeat
- Another approach: pick a point at and see if moving it to a different cluster will improve the quality of the overall solution. Repeat!

# K-means clustering... $K=2$



# Other clustering approaches

- Principal component analysis
  - Identify a direction (vector  $V$ ) such that the projection of data on  $V$  has maximum variance (first principal component)
  - repeat (vector  $V' \neq V$  such that project of data on  $V'$  has maximum variance)
  - Usually plot the first 2 or 3 principal components

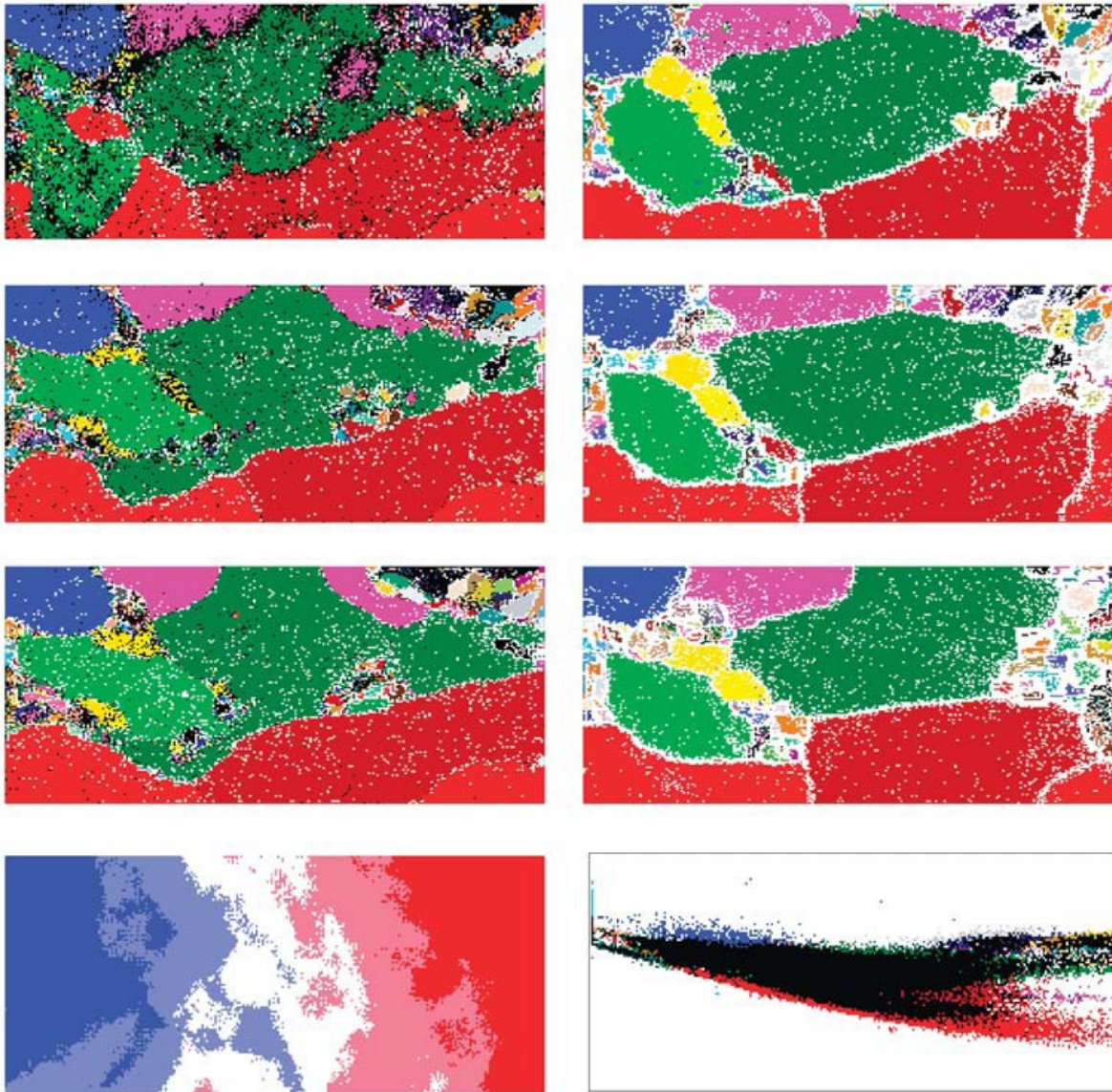


# Other clustering approaches

- Self-organizing maps
  - Neural-network based approach
  - Output layer of network are points in a low-dimensional space
- Graph theoretic
  - Points are connected by edges representing strength of "connection" (e.g. similarity or dissimilarity)
  - Pick clusters such that number of "similar" edges spanning boundaries is minimized, or number of "dissimilar" edges within each cluster is minimized
- Markov chain clustering
  - basic idea – a random walk through a graph will stay within a local strongly connected region



# Self organizing map of genomes



[http://www.jamstec.go.jp/esc/esc/publication/journal/jes\\_vol.6/pdf/JES6\\_22-Abe.pdf](http://www.jamstec.go.jp/esc/esc/publication/journal/jes_vol.6/pdf/JES6_22-Abe.pdf)

# Questions

- Compare nearest, farthest, average, and Ward's cluster distances in terms of how compact the resulting clusters might be.