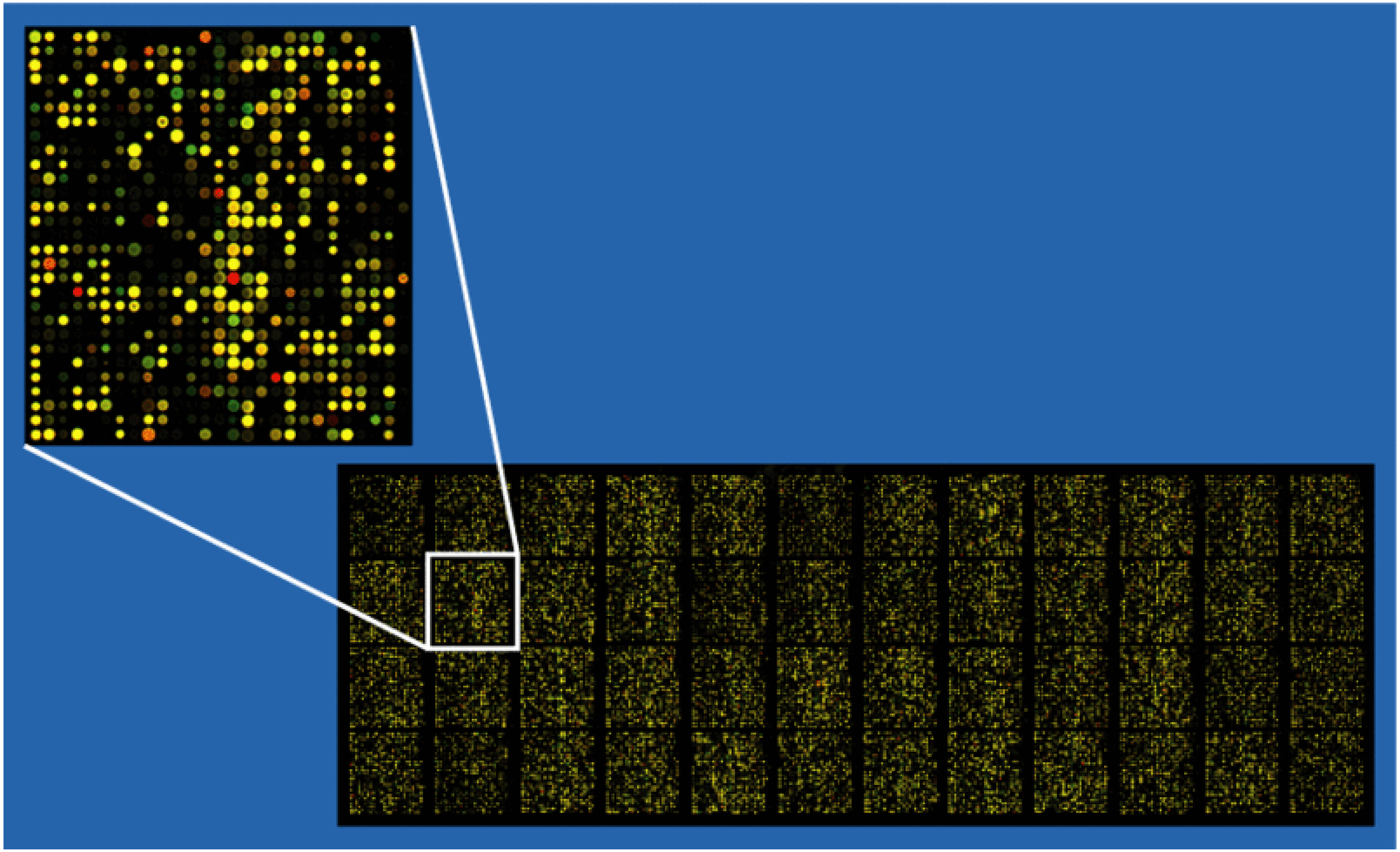
CMSC423: Bioinformatic Algorithms, Databases and Tools

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
ACTGACTGCATGCTGCCTCCCGTAGGAGTCGCCGTCCTCGACNC
>F4BT0V001D2HQQ rank=0000196 x=1551.0 y=1984.0 length=42
ACTGACTGCATGCTGCCTCCCGTAGGAGTGCCGTCCCTCGAC
>F4BT0V001CM392 rank=0000206 x=966.0 y=1240.0 length=82
AANCAGCTCTCATGCTCGCCCTGACTTGGCATGTGTTAAGCCTGTAGGCTAGCGTTCATC
CCTGAGCCAGGATCAAACTCTG
>F4BT0V001EIMFX rank=0000250 x=1735.0 y=907.0 length=46
ACTGACTGCATGCTGCCTCCCGTAGGAGTGTCGCGCCATCAGACTG
>F4BT0V001ENDKR rank=0000262 x=1789.0 y=1513.0 length=56
GACACTGTCATGCTGCCTCCCGTAGGAGTGCCTCCCTGAGCCAGGATCAAACTCTG
>F4BT0V001D91MI rank=0000288 x=1637.0 y=2088.0 length=56
ACTGCTCTCATGCTGCCTCCCGTAGGAGTGCCTCCCTGAGCCAGGATCAAACTCTG
>F4BT0V001D0Y5G rank=0000341 x=1534.0 y=866.0 length=75
GTCTGTGACATGCTGCCTCCCGTAGGAGTCTACACAAGTTGTGGCCCAGAACCACTGAGC
CAGGATCAAACTCTG
>F4BT0V001EMLE1 rank=0000365 x=1780.0 y=1883.0 length=84
ACTGACTGCATGCTGCCTCCCGTAGGAGTGCCTCCCTGCGCCATCAATGCTGCATGCTGC
TCCCTGAGCCAGGATCAAACTCTG
```

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

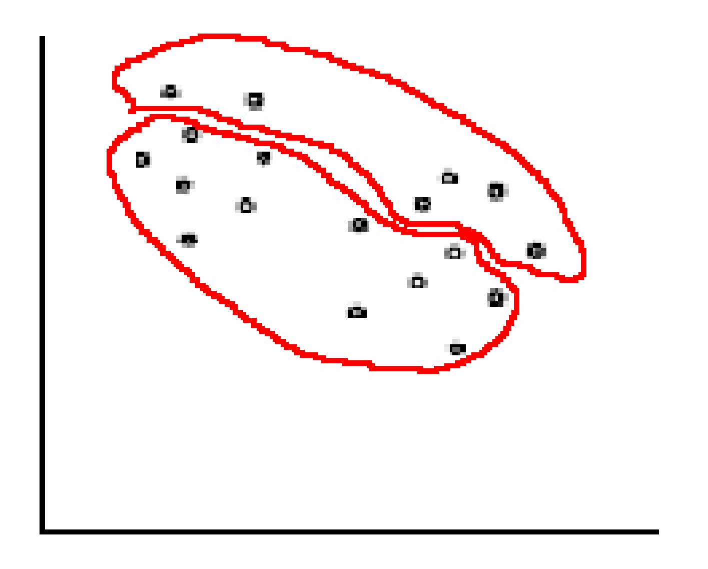
- Patterns:
  - do certain points group together?

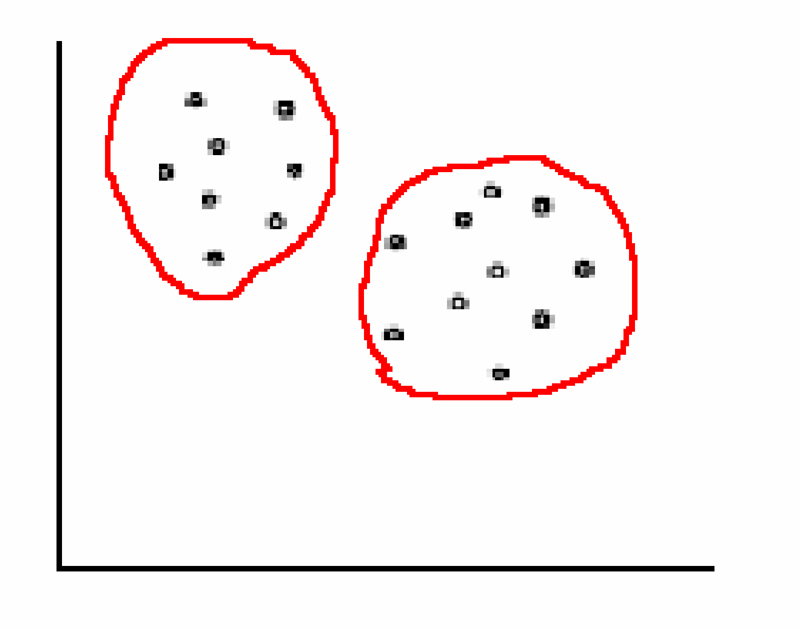
Types of clustering algorithms

- 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

The good clustering principle

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





Some issues with clustering

- Good clustering principle may not be achievable

- Finding the optimal clustering is usually NP-hard

- In how many ways can you partition n points into 2 clusters?

k-center clustering

- Pick k centers
- For each point, select the nearest center
- Find the set of k centers that minimizes the maximum distance between any point and its nearest center

- How many centers can there be?
- For k = 1, how can you pick the center?

Farthest-first clustering

- Pick a point – first center
- Pick the farthest point from it – second center
- repeat until k centers found

- Can you prove that solution is at most twice as bad as optimal?

Properties of distance

- Distance is Euclidean distance
- It is a metric – satisfies triangle inequality
- This property helps prove 2-approximation
- Note: Euclidean is not important – farthest distance works with any metric distance

k-means clustering

- Instead of min-max, use squared error – average distance from points to corresponding centers

- For k = 1, how do you pick center?

k-means clustering – Lloyd's algorithm

- 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...visual

https://www.naftaliharris.com/blog/visualizing-k-means-clustering/

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...

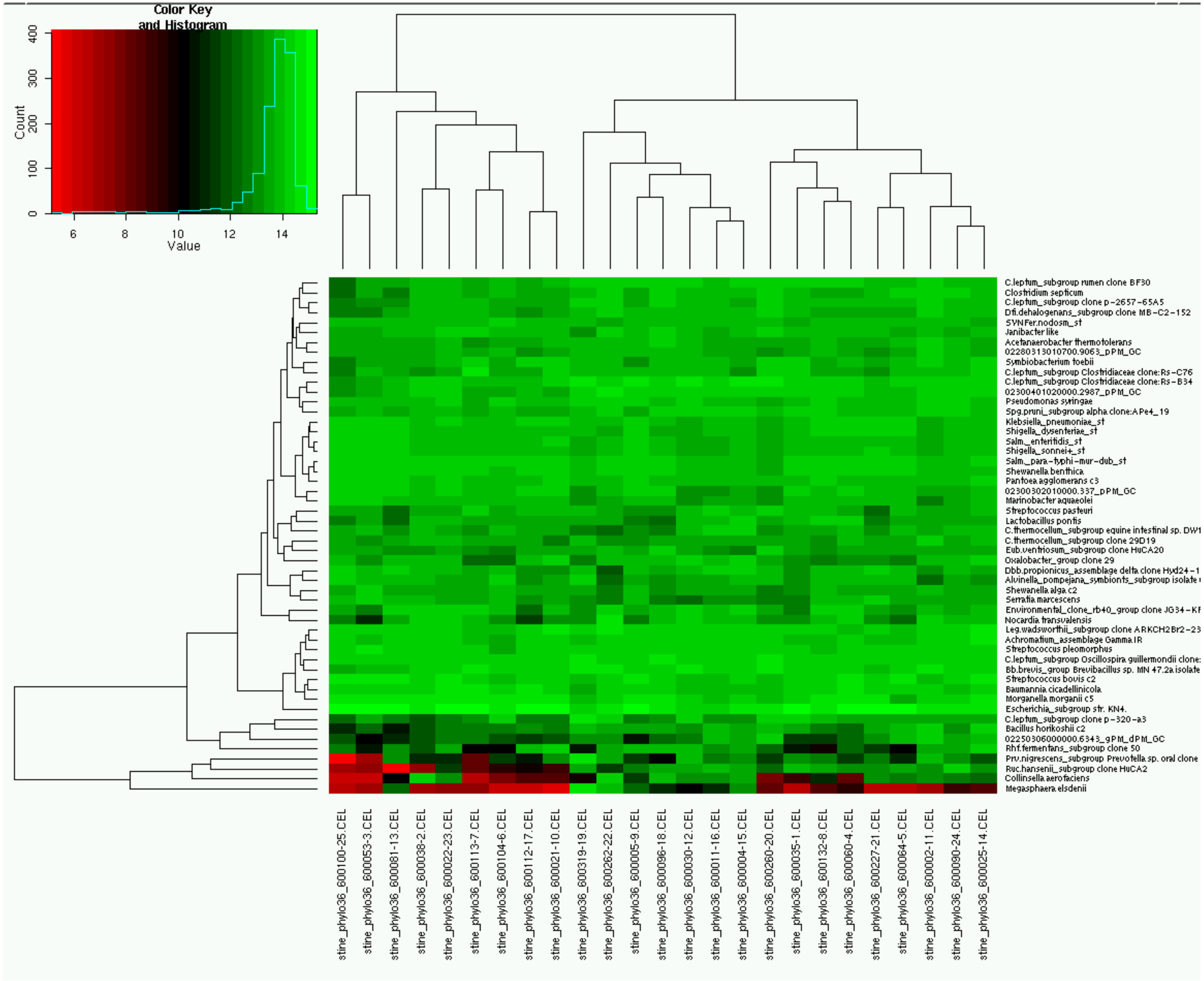
Hierarchical clustering

- 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

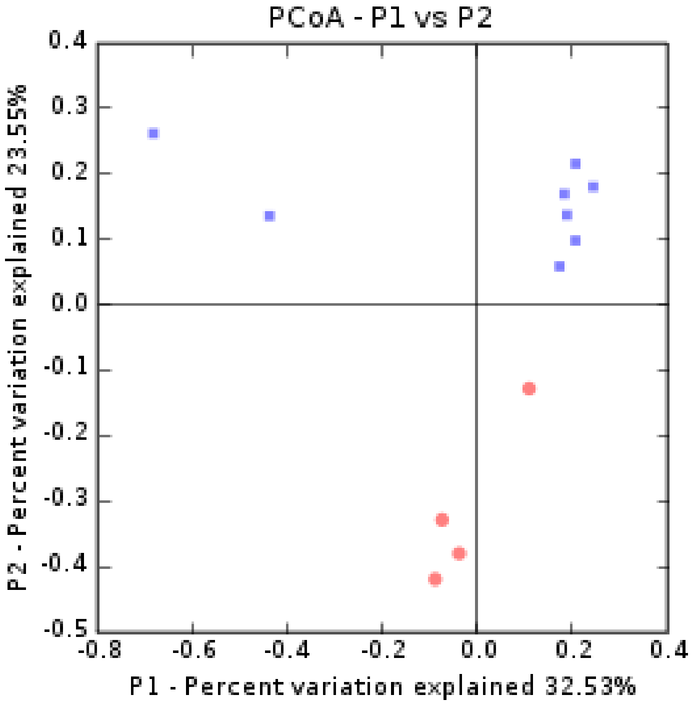
- Irrespective of distance choice, algorithm is the same
  1. compute inter-gene/cluster distances
  2. join together pair of genes/clusters with smallest distance
  3. recompute distances to include the newly created cluster
  4. repeat until all points in one cluster
- Output of program is a tree
- 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

Example: gut microbiome in children



Other clustering approaches

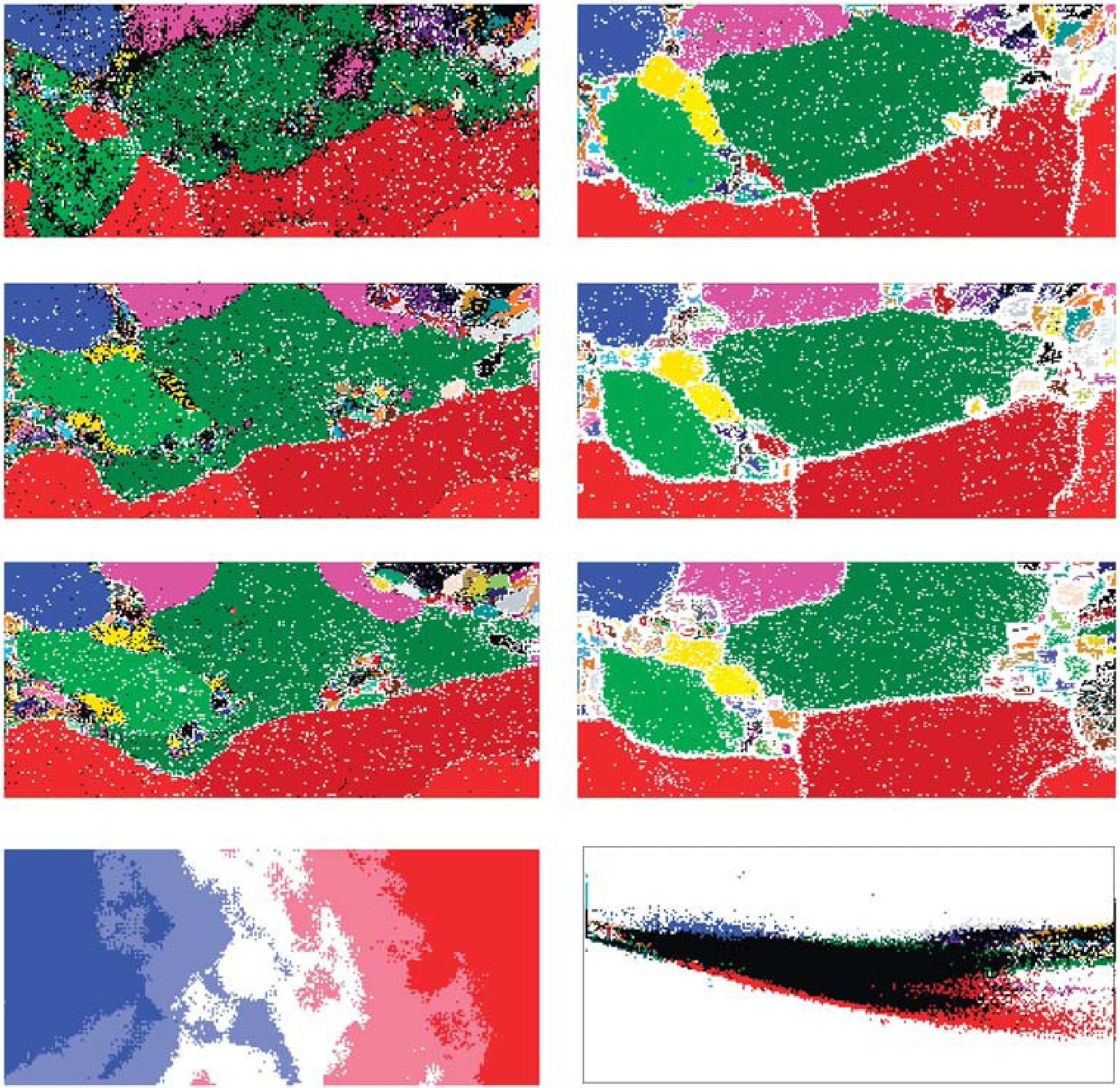
- Principal component analysis
  - Identify a direction (vector V) such that the projection of data on V has maximum variance (first principal component)
  - repeat (vector V' != V such that project of data on V' has maximum variance)
  - Usually plot the first 2 or 3 principal components



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