COSC 348:
Computing for Bioinformatics

Lecture 15:
Phylogenetic tree inference by clustering

Lubica Benuskova
Prepared according to the notes of Dr. Richard O'Keefe

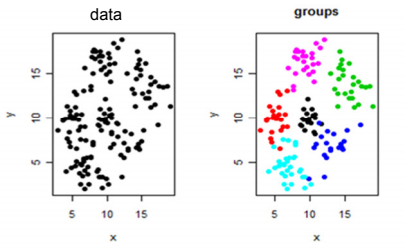
http://www.cs.otago.ac.nz/cosc348/

Three major approaches to phylogeny

- Parsimony methods.
  - Step 1: define a **cost** measure for phylogenies.
  - Step 2: announce that the best tree(s) is(are) the one(s) having the least cost.
- **Distance-based methods.**
  - Step 1: define a distance measure reflecting how different each OTU is from each other OTU.
  - Step 2: use a **clustering algorithm** to fit a tree to the mutual distances.
- Maximum likelihood.
  - Step 1: define a probability model for evolutionary change.
  - Step 2: announce that the best tree(s) is(are) the one(s) which ascribes the highest probability to the observed outcome.

Definition of clustering

- Clustering means automatically dividing a collection of objects into clusters (groups), such that objects in the same cluster are *similar* in some sense.



Cluster analysis

- No assumptions about the number of groups, or the group structure are made.
- Objects are either vectors (i.e. arrays) of numbers or DNA/RNA/protein sequences.
- Vectors of values are represented as points in *N*-dimensional space where *N* is the vector dimension (= size of an array of characters).
- A quantitative scale (metric) is used to measure the closeness or similarity between vectors of values or between bio-sequences. Grouping is done on the basis of similarities or dissimilarities between objects.

Types of clustering

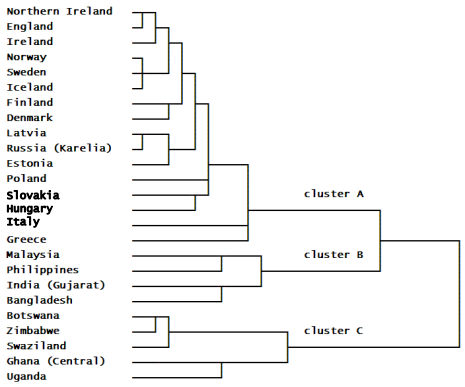
- If the clusters form a tree, we have **hierarchical** clustering; if they form a partition, we have **partitional** clustering; if they overlap, we have **fuzzy** clustering.
- Clustering algorithms can work either "bottom-up" or "top-down":
  - "Bottom-up" algorithms begin with elements and merge them into successively larger clusters.
  - "Top-down" algorithms begin with the whole set and proceed to divide it into successively smaller clusters.

More on types of clustering

- A **hierarchy** is an arrangement of objects in which the objects are represented as being "above," "below," or "at the same level as" one another.
- A **partition** of a set *X* is a division of *X* into *K* non-overlapping and non-empty partitions (blocks, groups, cells) that cover all of *X*.
- Fuzzy sets are sets whose elements have degrees of membership. Fuzzy sets were introduced by Lotfi A. Zadeh (1965) as an extension of the classical notion of a set. In classical set theory, an element either belongs or does not belong to the set. By contrast, an element of a fuzzy set is described with the aid of a membership function valued in the real unit interval [0, 1].

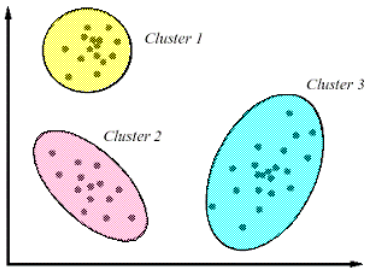
Hierarchical clustering: dendrogram

- Example: branches illustrate the distances in km between countries.



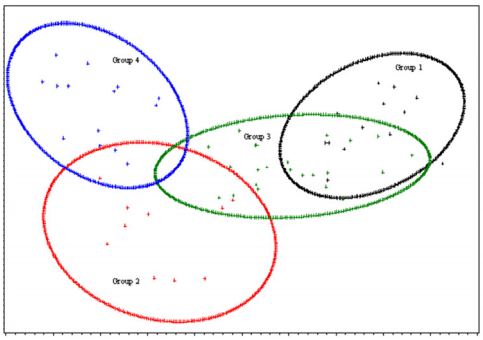
Partitional clustering

- Here is an example of partitional clustering. Note: the same data can be clustered into *different number of clusters* based on the clustering method and distance measure we use (more on this later):



Fuzzy clustering

- Membership of an object to a fuzzy cluster is described with the aid of a membership function valued in the real unit interval [0, 1]. Thus each object belongs to each cluster to a certain level.



Metric = distance or similarity measure

- Crucial step in most clustering is to select a distance or similarity measure, i.e. the **metric**, which will determine how the similarity of two elements is calculated.
- Distance or similarity measure will influence the shape of the clusters, as some elements may be close to one another according to one distance and farther away according to another.
- The choice depends on the measurement scale, and our heuristic knowledge. We can experiment with different measures to see, which one yields better results. The metric is always symmetric.
- Let's have two vectors (arrays) of character values a and b. Vector $a = (a_1, a_2, \ldots, a_n)$ and vector $b = (b_1, b_2, \ldots, b_n)$, etc.

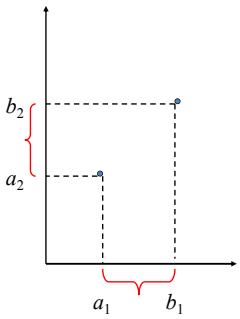
Manhattan distance

- ***Manhattan distance***: distance between two points or vectors is the sum of the absolute differences of their coordinates:

$$d_M(a,b) = |a_1 - b_1| + \ldots + |a_n - b_n|$$

- As an example let $a = (a_1, a_2)$ and $b = (b_1, b_2)$. Then:

$$d_M(a,b) = |a_1 - b_1| + |a_2 - b_2|$$



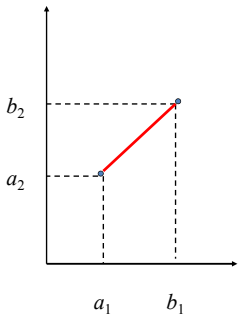
Euclidean distance

- ***Euclidean distance*** is the distance between two points in the Euclidean space:

$$d_E(a,b) = \sqrt{(a_1 - b_1)^2 + \ldots + (a_n - b_n)^2}$$

- As an example let $a = (a_1, a_2)$ and $b = (b_1, b_2)$. Then (recall the Pythagoras theorem):

$$d_E(a,b) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2}$$



Hamming distance

- ***Hamming distance*** between two strings of equal length is the number of positions at which the corresponding symbols are different, i.e.:

$$d_H(a,b) = \sum_{k=1}^{N} count(a_k \neq b_k)$$

- Hamming distance measures the minimum number of substitutions required to change one string into the other, i.e. the number of changes that transformed one string into the other.

- In phylogeny, Hamming distance corresponds to the *Fitch cost*.

Distance measure: inverse of the alignment score

- In case of protein / DNA / RNA sequences we obtain the alignment score using the protein substitution matrices and DNA/RNA scoring matrices, respectively.

- The *alignment score* is the sum of substitution scores and gap penalties. The alignment score evaluates goodness of alignment.

- Thus, the distance measure between two sequences can be proportional to the *inverse* of the alignment score.
  - E.g., distance between two identical sequences will be equal to zero.
  - Nonzero distances will be scaled according to the inverse of the alignment score such that the sequences, in which each character is different will have the maximal distance and all others will have the distance in the interval [0, max_distance].

Linkage criteria

- Linkage criteria determines which vectors in the clusters are taken into account to judge the distance between the clusters themselves.

- The so-called ***linkage criteria*** specifies the (dis)similarity of clusters as a function of the pair-wise distances of vectors in the clusters.

- When we use the agglomerative bottom-up clustering, we use the linkage criteria to merge the clusters together.

- When we use *partitioning* with chosen metric, then after partioning into *K* clusters, we will use the same metric to hierarchically organise the clusters using the linkage criteria.

Three main linkage criteria

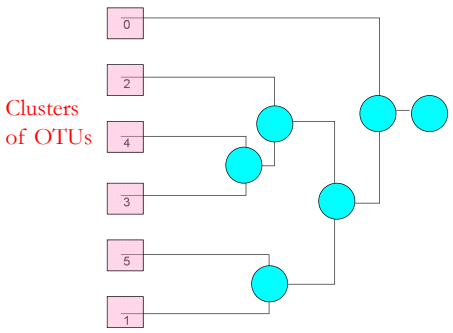
- The linkage criteria determines the distance between clusters of vectors as a function of the pairwise distances between vectors.

- Let *d* is the chosen metric. Three commonly used linkage criteria between two clusters A and B are:

Criterium	Formula
Maximum linkage clustering	$\max\{d(a,b): a \in A, b \in B\}$
Minimum linkage clustering	$\min\{d(a,b): a \in A, b \in B\}$
Mean or average linkage clustering	$\frac{1}{\lvert A \rvert \lvert B \rvert} \sum_{a \in A} \sum_{b \in B} d(a,b)$

Hierarchical clustering: final tree

- We build the tree "bottom-up" from the closest clusters. The lengths of branches are proportional to distances between clusters.



Clusters of OTUs

We do not have to worry about HTUs. They are inserted between clusters of OTUs.

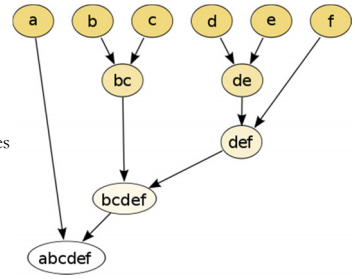
Vectors of character values for *Caminalcules*

- We will have 29 vectors (arrays) of $N = 85$ character values.
- We can have binary, nominal or ordinal values.
- We can treat 'x' as another value of the character.

```
Vector a: 11x0x1xx0120xx1x015001xxx0xx0x1xxxxxxx0000x1212200x...
Vector b: 11x0x01x0100xx2111411x101xxx0x2xxxx211100110x10x00x...
Vector c: 11x1101x0100xx0x01401x110xxx0x2xxxx0xx000110x10x00x...
Vector d: 11x1001x0100xx0x01401x101xxx0x2xxxx1xx000110x10x00x...
Vector e: 11x0x01x0100xx1x01411x100xxx0x2xxxx20x100110x10x00x...
Vector f: 1010x11x0100xx1x014001xxx0xx1101xx2xxx2010x0x0xx00x...
Vector g: 1x0x20x10x0xx0x00xxxxxxxxxxxx0011xxxx2010x0x0xx00xx...
Vector h: 11x0x21x0100xx0x010002xxx0xx0x0011xxxx2010x0x0xx00x...
etc.....
```

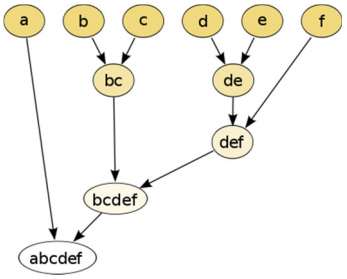
Agglomerative hierarchical clustering

- We can do hierarchical clustering, starting from raw data, i.e. vectors (arrays) of values, here denoted by letters **a, b, c, d, e, f**.
- Thus, the top row of nodes represent vectors of character values i.e., a = (a_1, a_2, … , a_N), etc.
- Remaining nodes represent the clusters to which the data belong.
  - E.g. "**bc**" denotes the cluster (union) of vectors **b** and **c**, etc.
- The length of an arrow represents the distance between clusters.



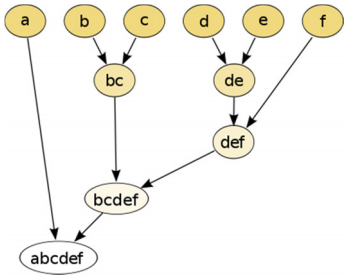
Agglomerative hierarchical clustering

- First we construct the N x N matrix **D** contains all distances between vectors *d*(i,j).
- The clusters are assigned sequence numbers 0, 1,..., (n−1) and L(m) is the level of the *m*th clustering.
- A cluster with sequence number m is denoted (m) and the distance between clusters (r) and (s) is denoted *d*[(r),(s)].
- The C code for algorithm is here: http://www.cs.otago.ac.nz/cosc348/phylo/phylo5.htm



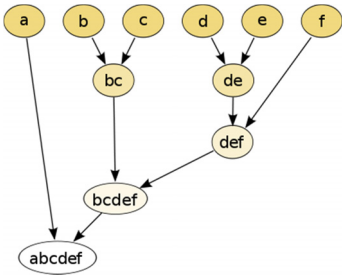
Agglomerative hierarchical clustering

1. Begin with the disjoint clustering having sequence number m = 0. (At this level, each vector is one cluster.)
2. Find the most similar pair of clusters in the current clustering, say pair (r), (s), according to *d*[(r),(s)] = min *d*[(i),(j)] where the minimum is over all pairs of clusters in the current clustering and is defined by the chosen linkage criterion.



Agglomerative hierarchical clustering

3. Increment the sequence number: m = m + 1. Merge clusters (r) and (s) into a single cluster to form the next clustering level m.
4. Update the distance matrix, **D**, by deleting the rows and columns corresponding to clusters (r) and (s) and adding a row and column corresponding to the newly formed cluster.
5. If all vectors are in one cluster, stop. Else, go to step 2.



Caveats of distance-based methods

- Cluster analysis is a statistical area of great practical importance. We can cluster archaeological artefacts, documents, diseases, signals, images, societies, anything where we can define some kind of distance or similarity.
- Unfortunately, clustering techniques don't have the mathematical objectivity that parsimony and maximal likelihood have. They are designed for grouping things by perceived similarity (phenetic) rather than evolutionary descent (cladistic). This counts against them.
- There is one great argument for distance-based methods: we can actually afford to use them. They can cope with hundreds of OTUs.