Phylogenetic tree building methods

method of building trees	type of data	
	distances	DNA sequences or other characters
clustering algorithm	→ UPGMA → neighbor-joining tree	
optimality criterion	→ minimum evolution tree	parsimony maximum likelihood Bayesian analysis

Cluster analysis

A cluster is a group of objects that within a larger group have neither random nor regular occurrence and their mutual distance or dissimilarity is less than distance or dissimilarity with objects belonging to other clusters.

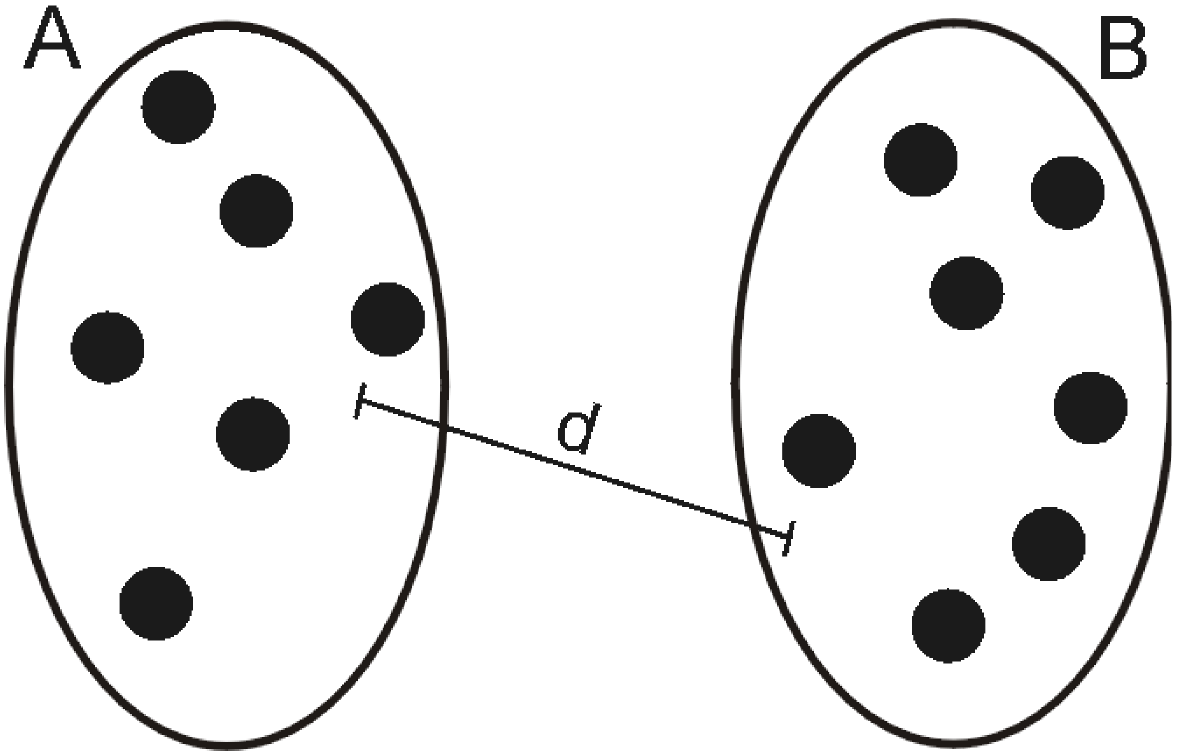
The center of gravity (centroid) of a cluster is a hypothetical (not necessarily existing) element, the coordinates of which in character space are given by the average values of the coordinates of individual objects.

Cluster analysis

According to:
cluster formation: agglomerative methods - divisive methods
cluster arrangement: hierarchical methods - non-hierarchical methods
cluster overlap: non-overlapping or overlapping clusters (fuzzy clustering)
clustering procedure: sequential methods - simultaneous

Methods SAHN clustering methods:
(a) methods based on minimizing the distance between clusters (b) methods based on optimizing cluster homogeneity according to a certain criterion

Single linkage, the nearest neighbor method

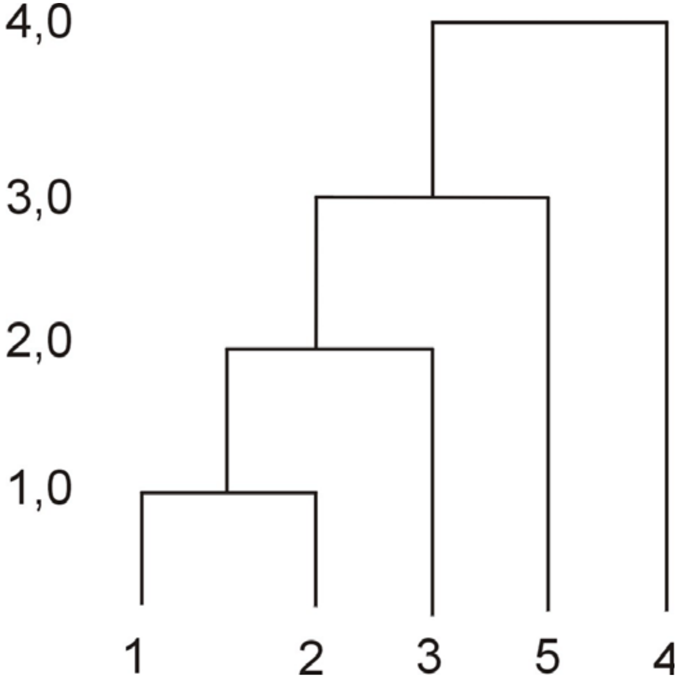


$$D_1 = \begin{array}{cccccc} & 1 & 2 & 3 & 4 & 5 \\ 1 & 0{,}0 & 1{,}0 & 7{,}0 & 4{,}0 & 12{,}0 \\ 2 & 1{,}0 & 0{,}0 & 2{,}0 & 5{,}0 & 9{,}0 \\ 3 & 7{,}0 & 2{,}0 & 0{,}0 & 8{,}0 & 3{,}0 \\ 4 & 4{,}0 & 5{,}0 & 8{,}0 & 0{,}0 & 6{,}0 \\ 5 & 12{,}0 & 9{,}0 & 3{,}0 & 6{,}0 & 0{,}0 \end{array}$$

$$d_{(1,\,2)3} = \min\{d_{1,\,3}, d_{2,\,3}\} = d_{2,\,3} = 2{,}0$$

$$d_{(1,\,2)4} = \min\{d_{1,\,4}, d_{2,\,4}\} = d_{1,\,4} = 4{,}0$$

$$d_{(1,\,2)5} = \min\{d_{1,\,5}, d_{2,\,5}\} = d_{2,\,5} = 9{,}0$$

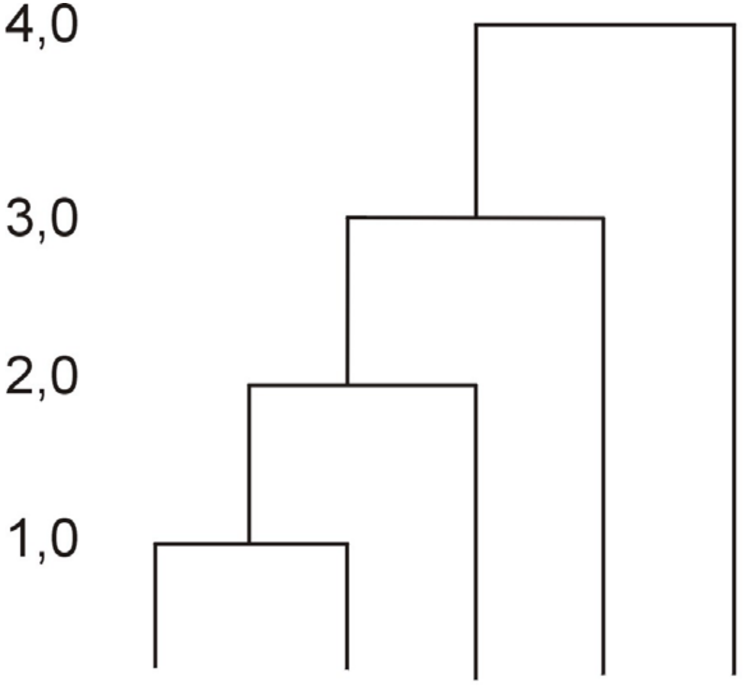


$$D_2 = \begin{array}{ccccc} & (1,\,2) & 3 & 4 & 5 \\ (1,\,2) & 0{,}0 & 2{,}0 & 4{,}0 & 9{,}0 \\ 3 & 2{,}0 & 0{,}0 & 8{,}0 & 3{,}0 \\ 4 & 4{,}0 & 8{,}0 & 0{,}0 & 6{,}0 \\ 5 & 9{,}0 & 3{,}0 & 6{,}0 & 0{,}0 \end{array}$$

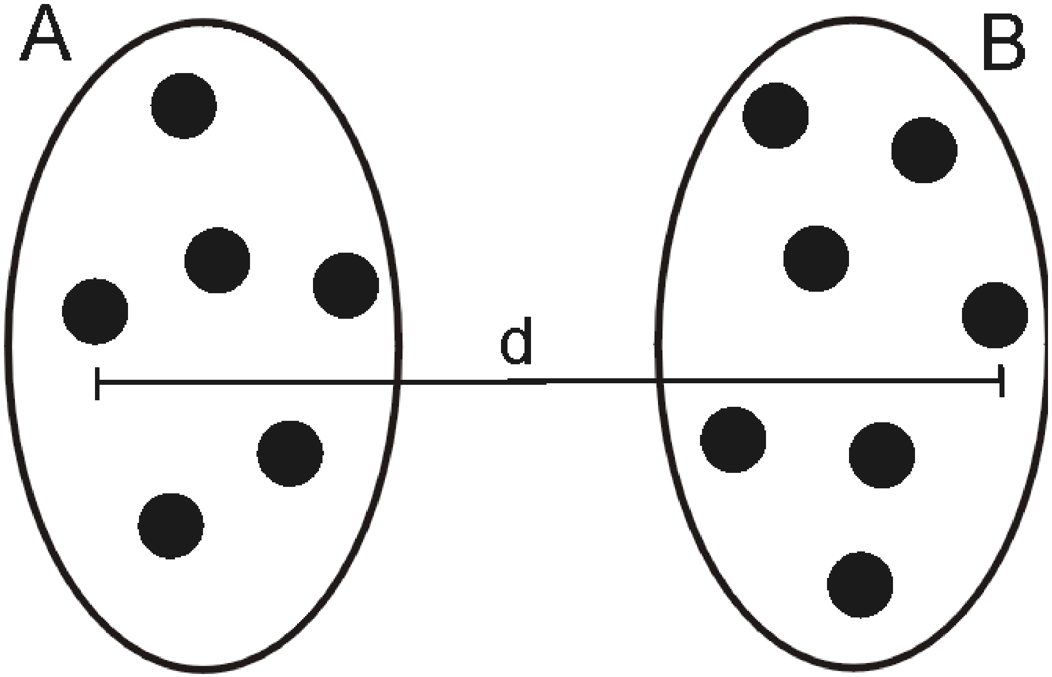
$$d_{(1,\,2,\,3)4} = \min\{d_{(1,\,2)\,4}, d_{3,\,4}\} = d_{(1,\,2)\,4} = 4{,}0$$

$$d_{(1,\,2,\,3)5} = \min\{d_{(1,\,2)\,5}, d_{3,\,5}\} = d_{3,\,5} = 3{,}0$$

$$D_3 = \begin{array}{cccc} & (1,\,2,\,3) & 4 & 5 \\ (1,\,2,\,3) & 0{,}0 & 4{,}0 & 3{,}0 \\ 4 & 4{,}0 & 0{,}0 & 6{,}0 \\ 5 & 3{,}0 & 6{,}0 & 0{,}0 \end{array}$$



Complete linkage, the furthest neighbor method



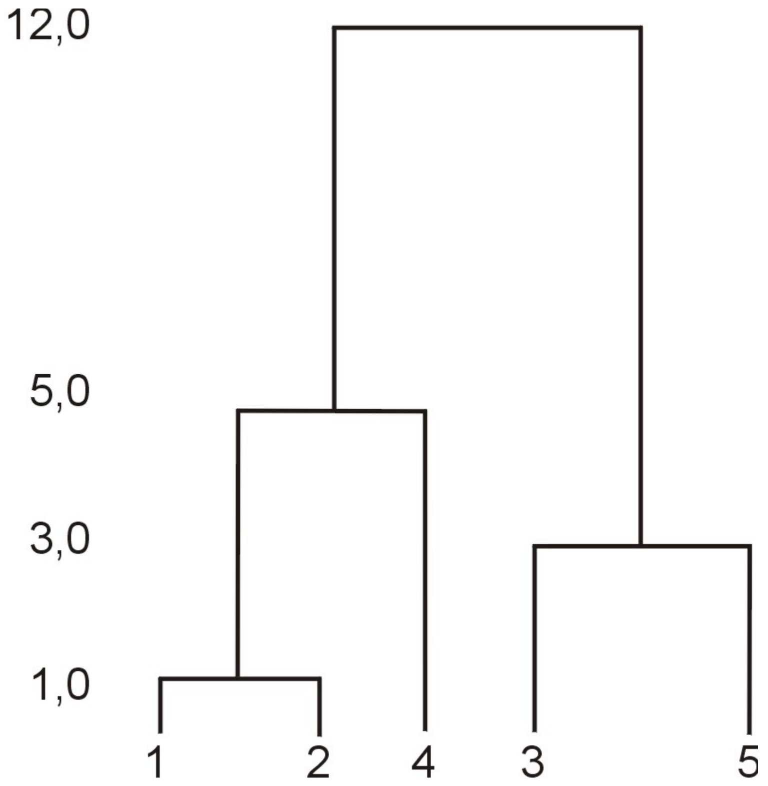
$$D_1 = \begin{array}{cccccc} & 1 & 2 & 3 & 4 & 5 \\ 1 & 0{,}0 & 1{,}0 & 7{,}0 & 4{,}0 & 12{,}0 \\ 2 & 1{,}0 & 0{,}0 & 2{,}0 & 5{,}0 & 9{,}0 \\ 3 & 7{,}0 & 2{,}0 & 0{,}0 & 8{,}0 & 3{,}0 \\ 4 & 4{,}0 & 5{,}0 & 8{,}0 & 0{,}0 & 6{,}0 \\ 5 & 12{,}0 & 9{,}0 & 3{,}0 & 6{,}0 & 0{,}0 \end{array}$$

$$d_{(1,\ 2)3} = \max\ \{d_{1,\ 3}, d_{2,\ 3}\} = d_{1,\ 3} = 7{,}0$$

$$d_{(1,\ 2)4} = \max\ \{d_{1,\ 4}, d_{2,\ 4}\} = d_{2,\ 4} = 5{,}0$$

$$d_{(1,\ 2)5} = \max\ \{d_{1,\ 5}, d_{2,\ 5}\} = d_{1,\ 5} = 12{,}0$$

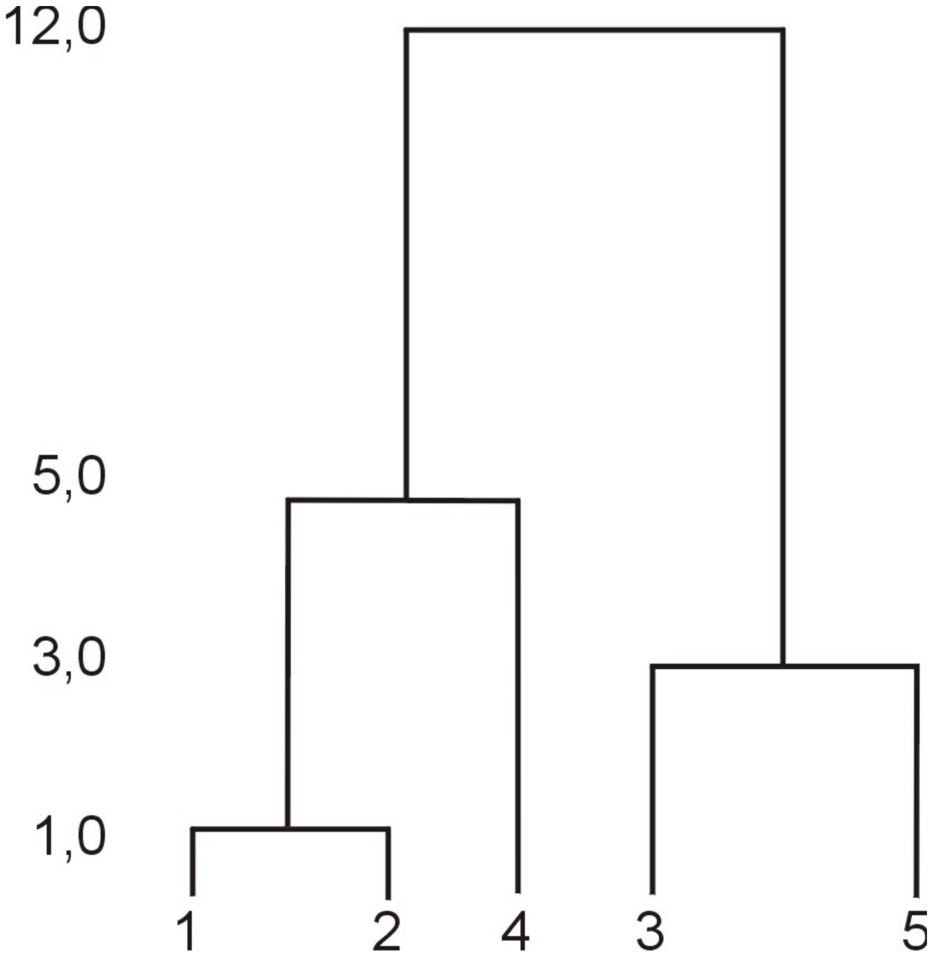
$$D_2 = \begin{array}{ccccc} & (1,\ 2) & 3 & 4 & 5 \\ (1,\ 2) & 0{,}0 & 7{,}0 & 5{,}0 & 12{,}0 \\ 3 & 7{,}0 & 0{,}0 & 8{,}0 & 3{,}0 \\ 4 & 5{,}0 & 8{,}0 & 0{,}0 & 6{,}0 \\ 5 & 12{,}0 & 3{,}0 & 6{,}0 & 0{,}0 \end{array}$$



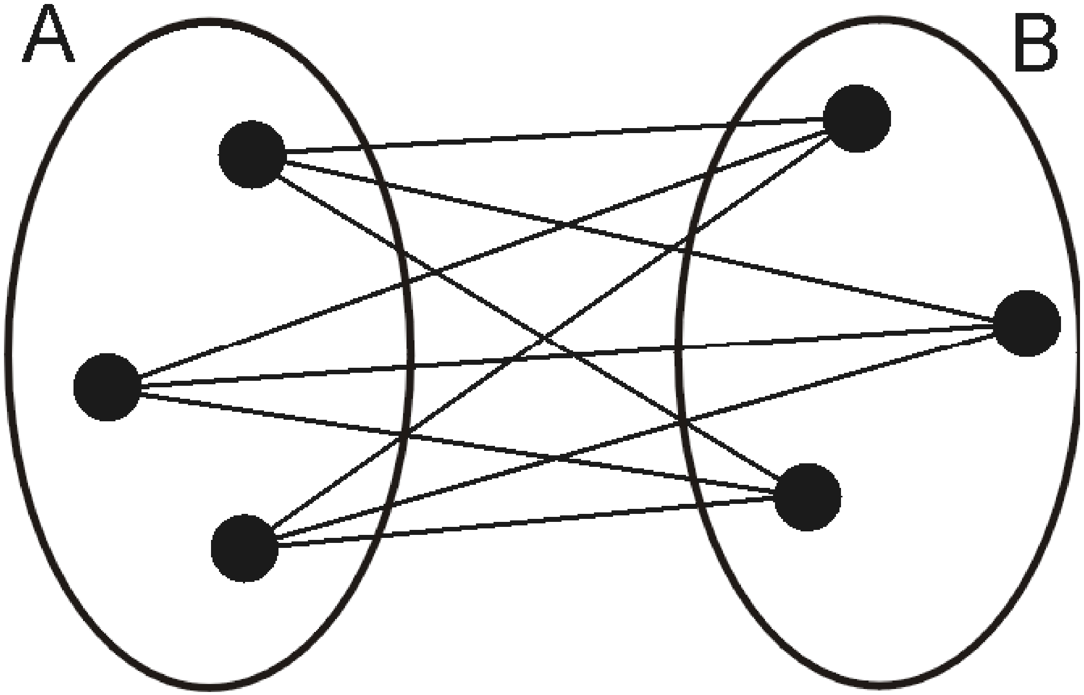
$$d_{(1,\ 2)(3,\ 5)} = \max\{d_{(1,\ 2)\ 3}, d_{(1,\ 2)\ 5}\} = d_{(1,2),\ 5} = 12{,}0$$

$$d_{(3,\ 5)4} = \max\{d_{3,\ 4}, d_{3,\ 5}\} = d_{3,\ 4} = 8{,}0$$

$$D_3 = \begin{array}{cccc} & (1,\ 2) & (3,\ 5) & 4 \\ (1,\ 2) & 0{,}0 & 12{,}0 & 5{,}0 \\ (3,\ 5) & 12{,}0 & 0{,}0 & 8{,}0 \\ 4 & 5{,}0 & 8{,}0 & 0{,}0 \end{array}$$



Average linkage, UPGMA – unweighted pair-group method using arithmetic averages



$$D_1 = \begin{array}{c|ccccc} & 1 & 2 & 3 & 4 & 5 \\ \hline 1 & 0{,}0 & 1{,}0 & 7{,}0 & 4{,}0 & 12{,}0 \\ 2 & 1{,}0 & 0{,}0 & 2{,}0 & 5{,}0 & 9{,}0 \\ 3 & 7{,}0 & 2{,}0 & 0{,}0 & 8{,}0 & 3{,}0 \\ 4 & 4{,}0 & 5{,}0 & 8{,}0 & 0{,}0 & 6{,}0 \\ 5 & 12{,}0 & 9{,}0 & 3{,}0 & 6{,}0 & 0{,}0 \end{array}$$

$d_{(1, 2)3} = 1/2\ (d_{1, 3} + d_{2, 3}) = 4{,}5$

$d_{(1, 2)4} = 1/2\ (d_{1, 4} + d_{2, 4}) = 4{,}5$

$d_{(1, 2)5} = 1/2\ (d_{1, 5} + d_{2, 5}) = 10{,}5$

$$D_2 = \begin{array}{c|cccc} & (1, 2) & 3 & 4 & 5 \\ \hline (1, 2) & 0{,}0 & 4{,}5 & 4{,}5 & 10{,}5 \\ 3 & 4{,}5 & 0{,}0 & 8{,}0 & 3{,}0 \\ 4 & 4{,}5 & 8{,}0 & 0{,}0 & 6{,}0 \\ 5 & 10{,}5 & 3{,}0 & 6{,}0 & 0{,}0 \end{array}$$

Neighbor-joining method

The method is based on genetic distance, which e.g. when evaluating AFLP data, it depends on the number of matching bands in the respective samples being compared. When using DNA sequence data, the genetic distance is calculated differently.

It is to some extent related to clustering methods. The procedure for calculating the distance of the formed clusters from the remaining objects is similar to the average distance method.

However, the analogy is not complete, because the "neighboring objects" are not the ones that are closest to each other, but the ones which result in the shortest possible dendrogram (tree). These dendrograms consist of nodes connected by internodes and branches.

Genetic distances for AFLP data

Coefficient of Nei & Li (1979): $NL_{xy} = 1 – (2\ N_{xy} / N_x + N_y)$

kde

N_{xy} = number of bands (fragments) common to samples *x* and *y*

N_x = total number of bands (fragments) present in sample *x*

N_y = total number of bands (fragments) present in sample *y*

Example:

sample *x*: 1010100011

sample *y*: 1010111101

$N_x = 5$; $N_y = 7$; $N_{xy} = 4$

$NL_{xy} = 1 – (2 * 4 / 5 + 7) = 0,333$

Coefficient of Link et al. (1995):

$L_{xy} = (N_x' + N_y') / (N_x' + N_y' + N_{xy})$

where

N_{xy} = number of bands (fragments) common to samples *x* and *y*

N_x' = the number of bands (fragments) present in sample *x* but absent in sample *y*

N_y' = the number of bands (fragments) present in sample *y* but absent in sample *x*

Example:

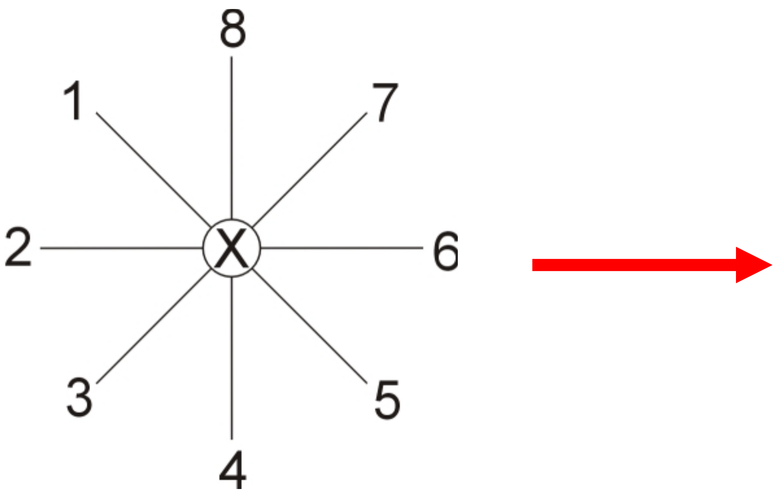
sample *x*: 1010100011

sample *y*: 1010111101

$N_x' = 1$; $N_y' = 3$; $N_{xy} = 4$

$L_{xy} = 1+3 / 1+3+4 = 0{,}5$

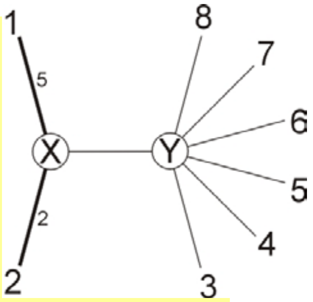
OTU	1	2	3	4	5	6	7
2	7						
3	8	5					
4	11	8	5				
5	13	10	7	8			
6	16	13	10	11	5		
7	13	10	7	8	6	9	
8	17	14	11	12	10	13	8



The total length of the dendrogram S is calculated according to the following formula (the formula is given for the pair of objects 1 and 2, in other cases the procedure is analogous, changing the values „*i* = 3“, „*k* = 3“ a „3 ≤ *i* < *j*“, which are specifically designed to exclude objects 1 and 2):

$$S_{12} = L_{XY} + (L_{1X} + L_{2X}) + \sum_{i=3}^{N} L_{iY} = \frac{1}{2(N-2)}\sum_{k=3}^{N}(D_{1k} + D_{2k}) + \frac{1}{2}D_{12} + \frac{1}{N-2}\sum_{3\le i<j} D_{ij}$$

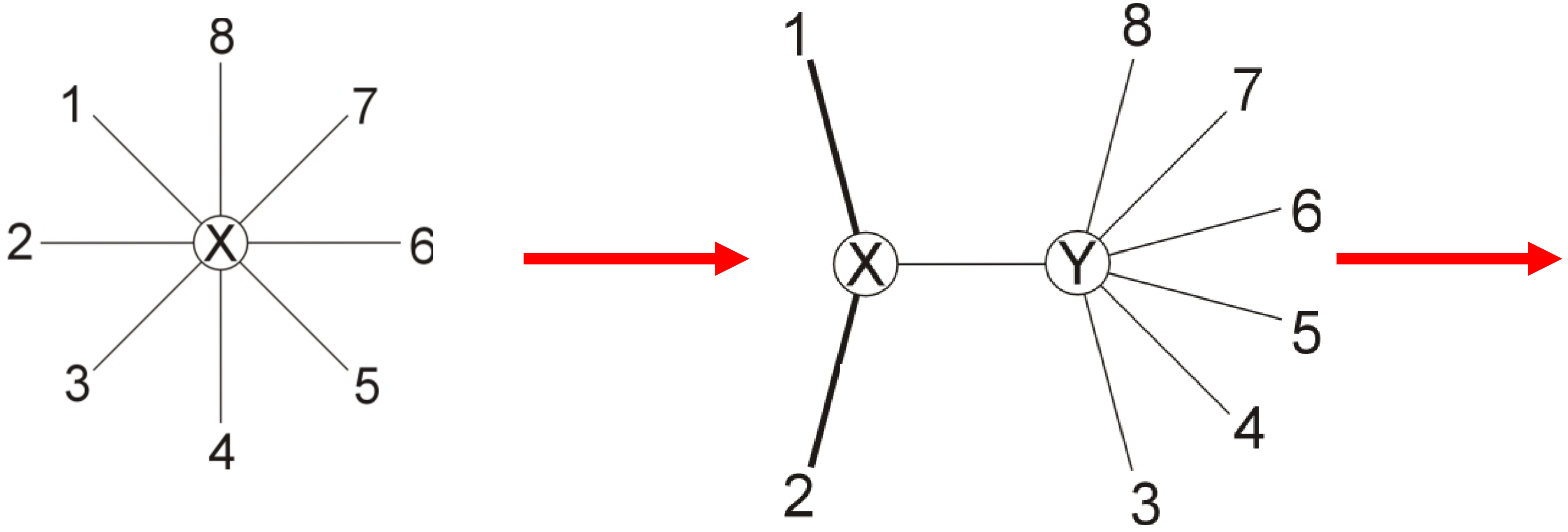
OTU	1	2	3	4	5	6	
2	7						
3	8	5					
4	11	8	5				
5	13	10	7	8			
6	16	13	10	11	5		
7	13	10	7	8	6	9	
8	17	14	11	12	10	13	8



$$S_{12} = \frac{1}{2(8-2)}(8+5+11+8+13+10+16+13+13+10+17+14) + \frac{7}{2} +$$

$$\frac{1}{8-2}(5+7+10+7+11+8+11+8+12+5+6+10+9+13+8) = 36{,}67$$

OTU	1	2	3	4	5	6	7
2	**36.67**						
3	38.33	38.33					
4	39.00	39.00	38.67				
5	40.33	40.33	40.00	39.67			
6	40.33	40.33	40.00	39.67	37.00		
7	40.17	40.17	39.83	39.50	38.83	38.83	
8	40.17	40.17	39.83	39.50	38.83	38.83	37.67



The distances L_{1X} and L_{2X} are calculated according to the formulas:

$$L_{1X} = \frac{D_{12} + D_{1Z} - D_{2Z}}{2} \quad L_{2X} = \frac{D_{12} + D_{2Z} - D_{1Z}}{2} \text{ kde } D_{1Z} = \frac{\sum_{i=3}^{N} D_{1i}}{N-2} \text{ a } D_{2Z} = \frac{\sum_{i=3}^{N} D_{2i}}{N-2}$$

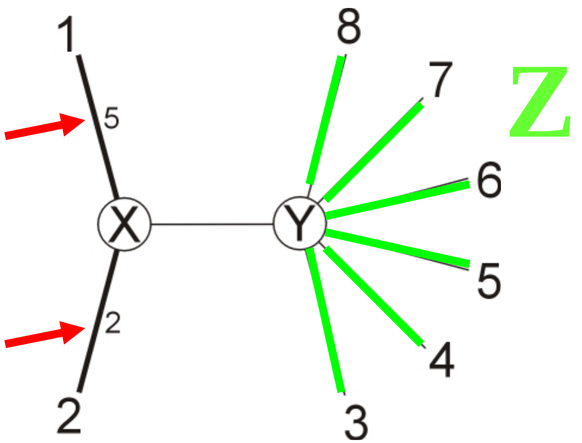
OTU	1	2	3	4	5	6	7
2	7						
3	8	5					
4	11	8	5				
5	13	10	7	8			
6	16	13	10	11	5		
7	13	10	7	8	6	9	
8	17	14	11	12	10	13	8

For the case of objects 1 and 2:

$$D_{2Z} = \frac{5+8+10+13+10+14}{8-2} = 10$$

$$D_{1Z} = \frac{8+11+13+16+13+17}{8-2} = 13$$

$$L_{1X} = \frac{7+13-10}{2} = 5 \quad L_{2X} = \frac{7+10-13}{2} = 2$$

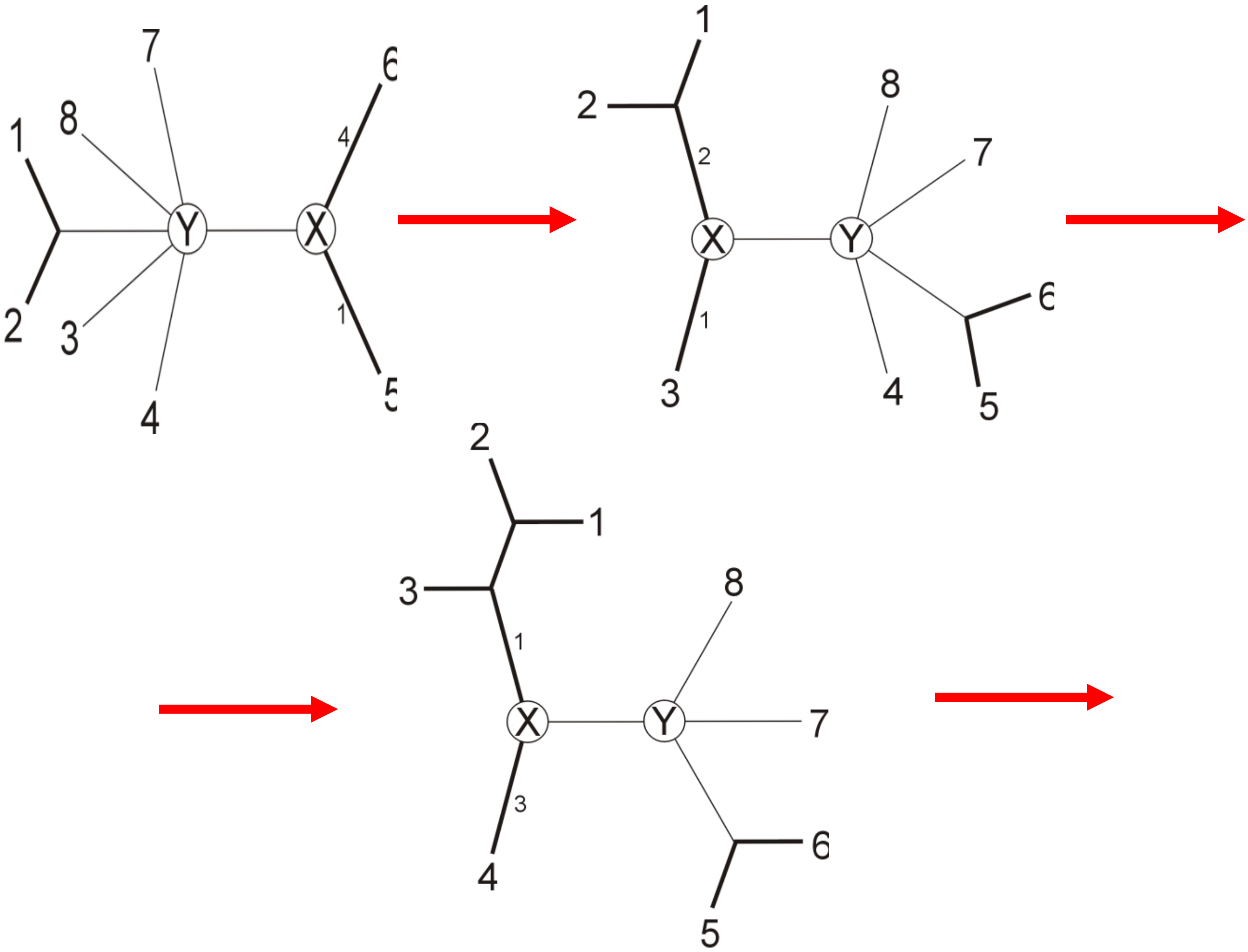


In the next cycle, the procedure is similar, except that objects 1 and 2 are considered as one object (or cluster). The distance of cluster 1 and 2 from other objects is calculated similarly to the method of average linkage (UPGMA) cluster analysis:

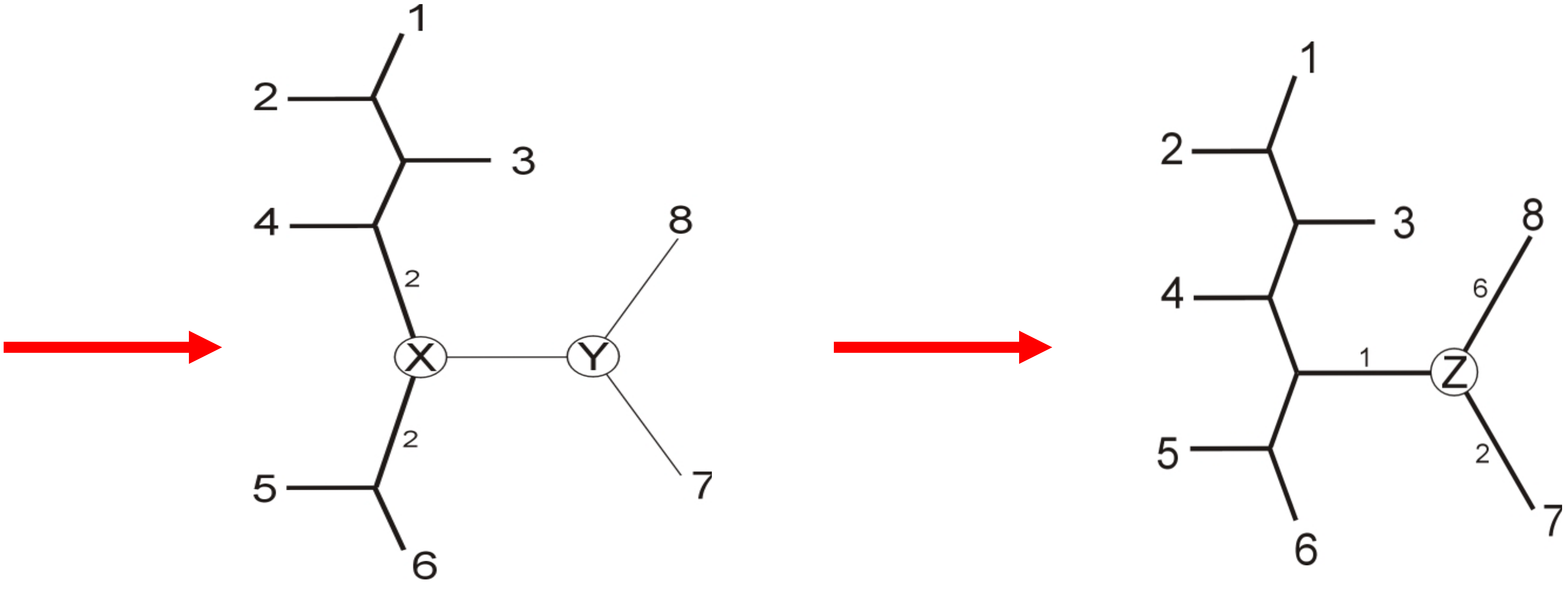
$$D_{(1-2)j} = \frac{D_{1j} + D_{2j}}{2}, \text{ where } 3 \le j \le \text{N}.$$

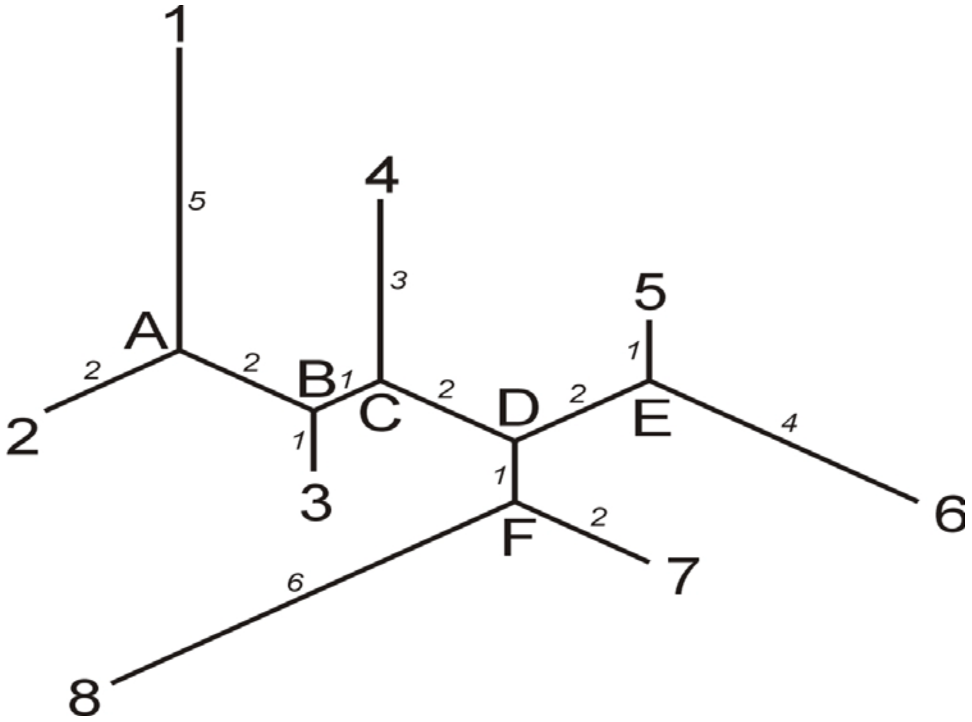
Matrix of S_{ij} values in the second cycle (objects 5 and 6 were selected as neighboring objects):

OTU	**1-2**	**3**	**4**	**5**	**6**	**7**
3	31.50					
4	32.30	32.30				
5	33.90	33.90	33.70			
6	33.90	33.90	33.70	**31.30**		
7	33.70	33.70	33.50	33.10	33.10	
8	33.70	33.70	33.50	33.10	33.10	31.90



7
6
8
1
4
Y
X
2
3
1
4
5
1
2
8
2
7
X
Y
1
6
3
4
5
2
1
3
8
1
X
Y
7
3
6
4
5





Unrooted tree

Bootstrap

From characters we create variations of the *n*th class from *n* elements with repetition (where *n* is the number of characters) - "resampling with replacement" (i.e., we create new replicated data files)

Each replicated data file is analyzed

We are looking for clusters that are repeated in the trees from individual analyzes

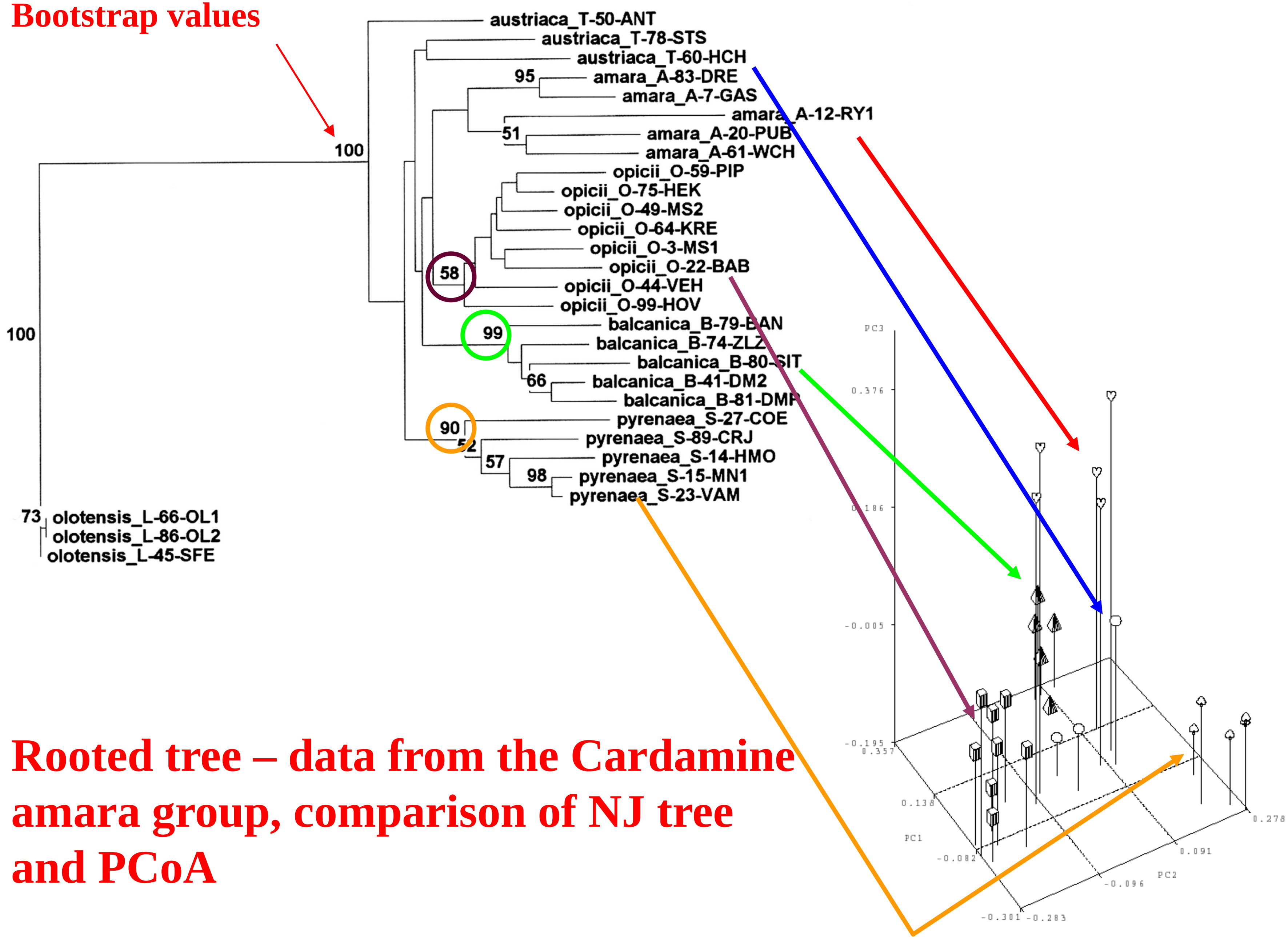
original matrix of data

	Characters							
Taxa	1	2	3	4	5	6	7	8
A	R	R	Y	Y	Y	Y	Y	Y
B	R	R	Y	Y	Y	Y	Y	Y
C	Y	Y	Y	Y	Y	R	R	R
D	Y	Y	R	R	R	R	R	R
Outgp	R	R	R	R	R	R	R	R

replicated matrix of data

	Characters							
Taxa	1	2	2	5	5	6	6	8
A	R	R	R	Y	Y	Y	Y	Y
B	R	R	R	Y	Y	Y	Y	Y
C	Y	Y	Y	Y	Y	R	R	R
D	Y	Y	Y	R	R	R	R	R
Outgp	R	R	R	R	R	R	R	R

Bootstrap values



Rooted tree – data from the Cardamine amara group, comparison of NJ tree and PCoA

It is also possible to evaluate DNA sequences by the neighbor joining method

In this case, we calculate the genetic distance in two ways:

(1) simple distance = number of different nucleotide positions / total number of nucleotide positions (especially in the case of a high substitution rate, it can significantly underestimate the number of substitutions)

(2) distance calculated based on substitution models

Substitution models

Substitution models:

- Jukes Cantor, Kimura, Tajima & Nei, Transversion analyses

- General time reversable model (GTR): includes a different probability for each type of change

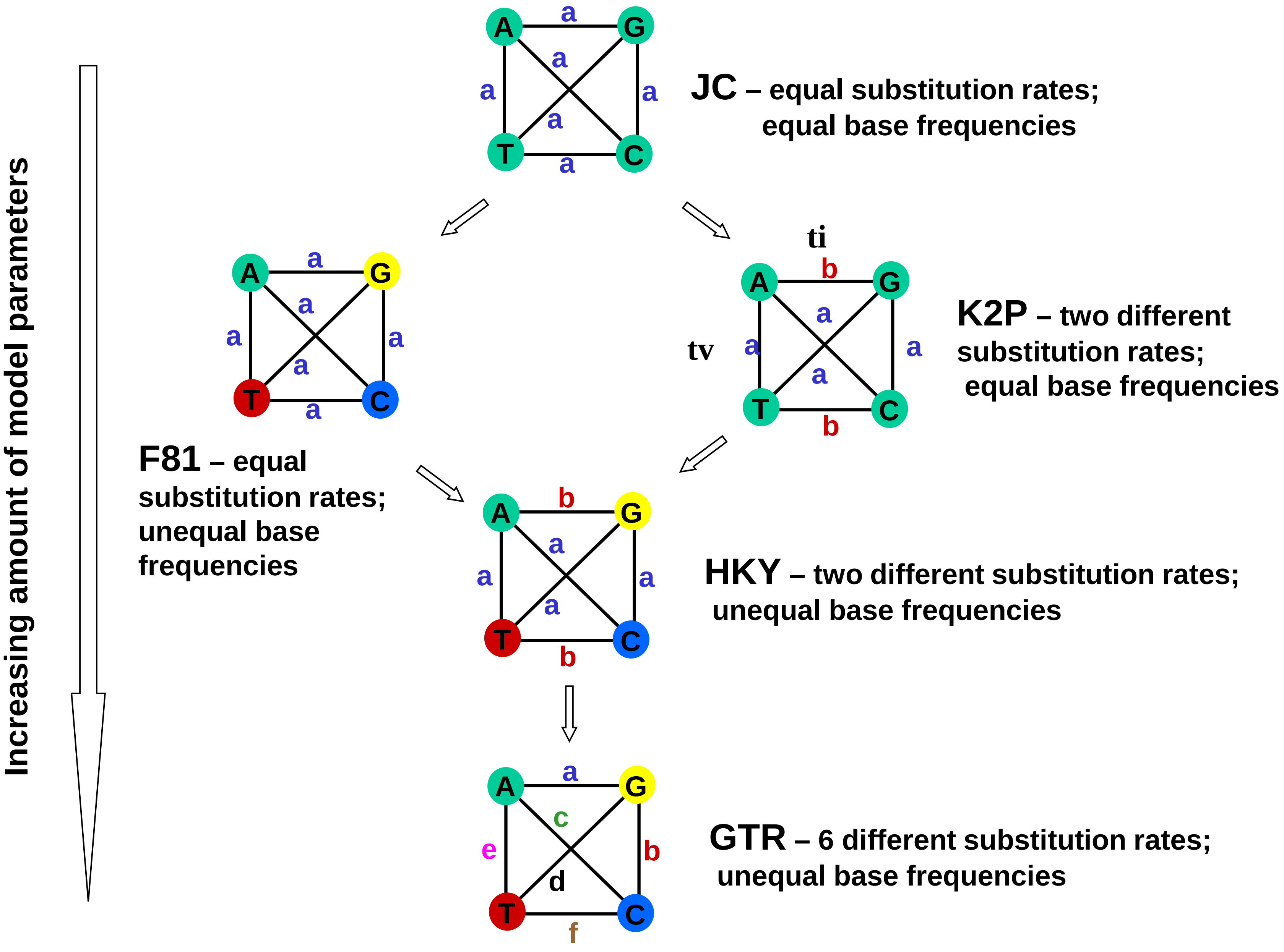
- LogDet / Paralinear distance model: allows to take into account different base frequencies in different sequences

All of these models include correction for multiple substitutions at the same position

All (except Logdet/paralinear distances) can be modified to include a gamma correction for site rate heterogeneity

A. Wilmotte, Uni Liège

Substitution models



Increasing amount of model parameters

JC – equal substitution rates; equal base frequencies

F81 – equal substitution rates; unequal base frequencies

K2P – two different substitution rates; equal base frequencies

HKY – two different substitution rates; unequal base frequencies

GTR – 6 different substitution rates; unequal base frequencies

Gamma distances

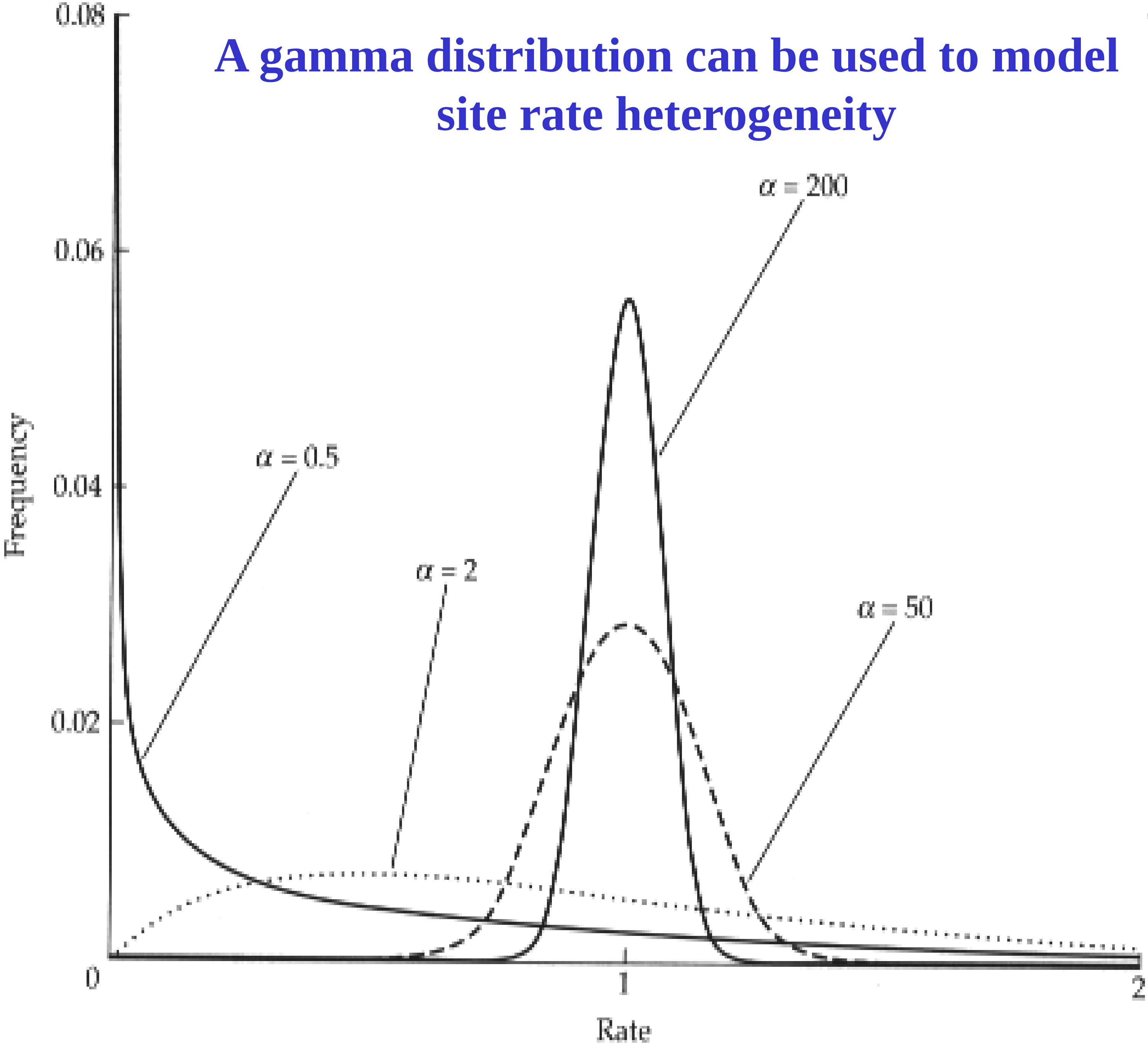
The distance measures usually start from the assumption that the rate of nucleotide substitution is the same for all nucleotide sites.

However, in real sequences, this assumption rarely holds.

Different studies suggest that the rate of nucleotide substitution varies approximately according to the gamma distribution (see Uzzell and Corbin, 1971; Jin and Nei, 1990; Nei, 1991).

This gamma distribution is specified by a parameter alpha which is the square of the inverse of the coefficient of variation of substitution rate (Nei, 1991).

A gamma distribution can be used to model site rate heterogeneity



Jukes & Cantor model: d_{xy} **= - (3/4) *ln* (1- 4/3 D)**

- d_{xy} = distance between sequence x and sequence y expressed as the number of changes per site
- (note d_{xy} = r/n where r is number of replacements and n is the total number of sites. This assumes all sites can vary and when unvaried sites are present in two sequences it will underestimate the amount of change which has occurred at variable sites)
- D = is the observed proportion of nucleotides which differ between two sequences (fractional dissimilarity)
- *l*n = natural log function to correct for superimposed substitutions
- The 3/4 and 4/3 terms reflect that there are four types of nucleotides and three ways in which a second nucleotide may not match a first - with all types of change being equally likely (i.e. unrelated sequences should be 25% identical by chance alone)

A. Wilmotte, Uni Liège

The natural logarithm ln is used to correct for superimposed changes at the same site

- If two sequences are 95% identical they are different at 5% or 0.05 (D) of sites thus:

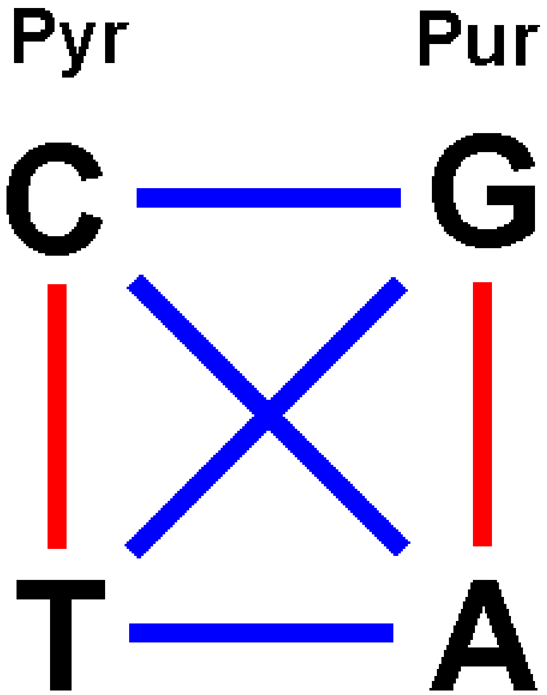
$$d_{xy} = -3/4 \ln [1-(4/3 \text{ x } 0.05)] = 0.0517$$

- Note that the observed dissimilarity 0.05 increases only slightly to an estimated 0.0517 - this makes sense because in two very similar sequences one would expect very few changes to have been superimposed at the same site in the short time since the sequences diverged apart
- However, if two sequences are only 50% identical they are different at 50% or 0.50 (D) of sites thus:

$$d_{xy} = -3/4 \ln [1-(4/3 \text{ x } 0.5)] = 0.824$$

- For dissimilar sequences, which may diverged apart a long time ago, the use of *l*n infers that a much larger number of superimposed changes have occurred at the same site

A. Wilmotte, Uni Liège



Transitions (Ts) are interchanges between pyrimidines (C - T), or between purines (A - G)

Transversions (Tv) are interchanges beween purines & pyrimidines.

A. Wilmotte, Uni Liège

Kimura - two parameter model

VKimura (1980) provided a method for inferring evolutionary distance in which transitions and transversions are treated separately:

$$d_{AB} = -\frac{1}{2}\ln\left[(1-2P-Q)\sqrt{1-2Q}\right]$$

where P is the fraction of sequence positions differing by a transition and Q is the fraction of sequence positions differing by a transversion.

There are twice as many kinds of transversions as transitions

K transition bias = [Ti] / [Tv]

Should be close to **0.5**

However, Tv Tv are rare for close comparisons, more common for distant relationships

K > 6 is given for close comparisons

A. Wilmotte, Uni Liège

Tajima & Nei

In the general correction of Tajima and Nei (1984), the evolutionary distance is estimated by:

$$d_{AB} = -b \ln\left(1 - \frac{1}{b} f_{AB}\right) \quad \text{where} \quad b = 1 - \sum_{i \in N} f_i^2$$

and fi is the frequency of the i-th type of nucleotide belonging to the set of possible nucleotide types N (= A, G, C, U or T) in the sequences being compared. This equation holds for the model of nucleotide substitutions with equal substitution rates between different nucleotides and does NOT take into account unequal rates of substitution among different nucleotide pairs (Tajima and Nei, 1984). In TREECON, the computed base composition is the average for all the sequences analyzed (as suggested in Swofford et al., 1996). If the frequencies are 0.25 for all four nucleotides, this equation equals the one of Jukes and Cantor.

A. Wilmotte, Uni Liège

Transversion analysis

Sometimes, it can be interesting to estimate the evolutionary distance on the basis of transversions only (see e.g. Woese et al., 1991; Van de Peer et al., 1996b). The evolutionary distance is then estimated by (Tajima and Nei, 1984; Swofford et al., 1996):

$$d_{AB} = -b \ln\left(1 - \frac{1}{b} Q\right)$$

where Q is the fraction of transversions and

$$b = 1 - \left[\left(f_A + f_G\right)^2 + \left(f_C + f_{(T,U)}\right)^2\right]$$

and fA + fG being the fraction of purines, and fC + fU or fT being the fraction of pyrimidines, computed over the complete alignment

logDet/paralinear distance method

- LogDet/paralinear distances was designed to deal with unequal base frequencies in each pairwise sequence comparison - thus it allows base compositions to vary over the tree!
- This distinguishes it from the GTR distance model which takes the average base composition and applies it to all comparisons
- LogDet/paralinear distances assume all sites can vary - thus it is important to remove those sites which cannot change (termed invariable by PAUP) the proportion of such sites is typically slightly smaller than the observed number of constant sites and is estimated using ML
- Invariable sites are removed according to the base composition of constant sites (rather than the base composition of all sites - which may be different) in order to preserve the correct base frequencies among remaining constant sites

A. Wilmotte, Uni Liège

logDet/Paralinear Distances

$$\mathbf{d_{xy} = -ln\ (det\ F_{xy})}$$

- d_{xy} = estimated distance between sequence x and sequence y
- *l*n = natural log function to correct for superimposed substitutions
- F_{xy} = 4 x 4 (there are four bases in DNA) divergence matrix for seq X & Y - this matrix summarises the relative frequencies of bases in a given pairwise comparison
- det = is the determinant (a unique mathematical value) of the matrix

A. Wilmotte, Uni Liège

LogDet – example for two sequences A and B

		Sequence B			
		a	c	g	t
Sequence A	a	224	5	24	8
	c	3	149	1	16
	g	24	5	230	4
	t	5	19	8	175

- For sequences A and B, over 900 sequence positions, this matrix summarises pairwise site by site comparisons
- The matrix Fxy expresses this data as the proportions (e.g. 224/900 = 0.249) of sites:

$$\text{Fxy} = \begin{array}{ccccc} & a & c & g & t \\ a & .249 & .006 & .027 & .009 \\ c & .003 & .166 & .001 & .018 \\ g & .027 & .006 & .256 & .004 \\ t & .006 & .021 & .009 & .194 \end{array}$$

- Dxy = -*l*n [det Fxy] = -*l*n [.002] = 6.216 (the logDet distance between sequences A and B)

A. Wilmotte, Uni Liège

logDet/paralinear distance: advantages

- Very good for situations where base compositions vary significantly between sequences
- Even when base compositions do not appear to vary the LogDet / paralinear distances model performs at least as well as other distance methods
- A drawback is that it assumes sites evolve identically and rates are equal for all sites
- However, a correction whereby a proportion of invariable sites are removed prior to analysis appears to work very well in simulations

A. Wilmotte, Uni Liège

Distance methods - advantages

- Fast - suitable for analysing data sets which are too large for ML

- A large number of models are available with many parameters - improves estimation of distances

Distances - disadvantages

- Information is lost - given only the distances, it is impossible to derive the original sequences
- Only through character based analyses (ML, parsimony) can the most informative positions be inferred (e.g. signature analysis of 16S rRNA)
- Generally outperformed by Maximum likelihood methods in choosing the correct tree in computer simulations (but logDet is better in some situations)

A. Wilmotte, Uni Liège

Heterogeneity of nucleotide change rates at different positions can be a problem

- Occurs when different sites in a molecule evolve at different rates due to different functional constraints

- Many models (Jukes Cantor, Parsimony, LogDet, some ML models) assume all sites can vary and all evolve at the same rate

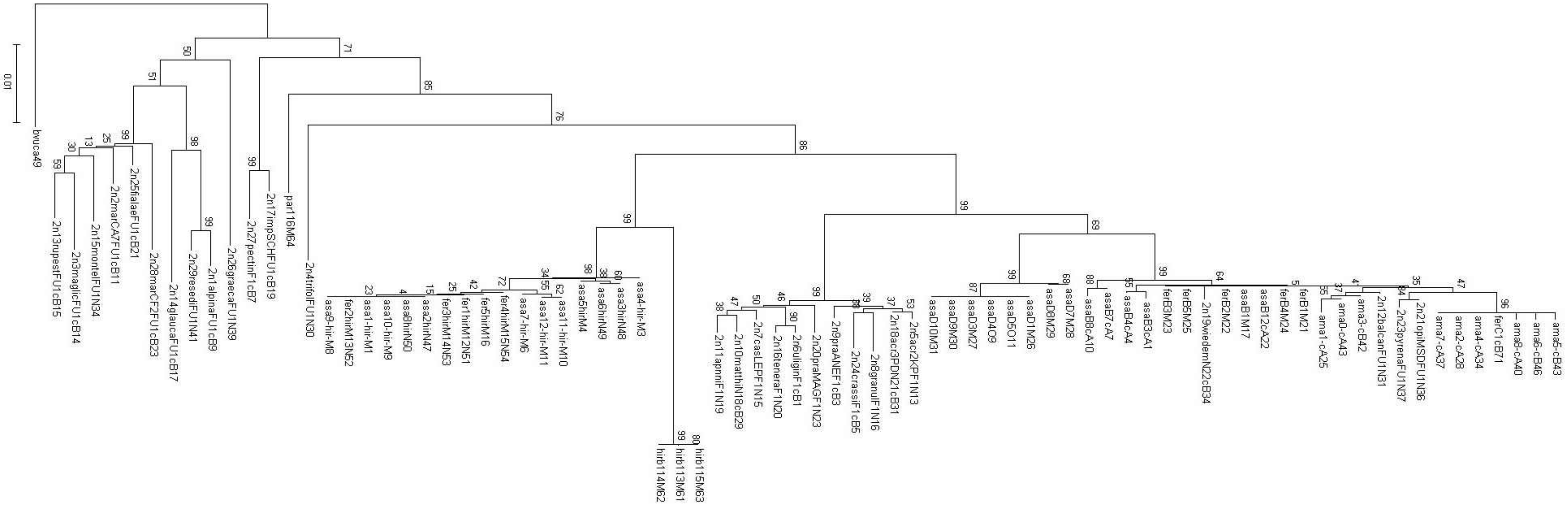
- This underestimates the amount of change that has occurred - and thus distances between sequences - leading to incorrect trees

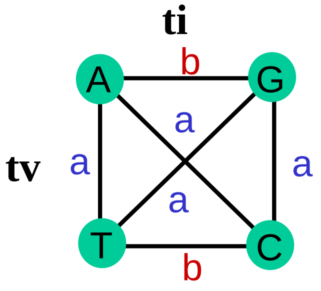
A. Wilmotte, Uni Liège

Heterogeneity of nucleotide change rates at different positions can be a problem

- Can include a gamma correction for site rate heterogeneity - if model allows this (many do - PAUP* has many of the most useful)

- Or edit the data to remove sites which are constant across the alignment (i.e. the slowest evolving), or those sites which are evolving more quickly than others





K2P – dve rôzne rýchlosti substitúcie; rovnaké frekvencie báz

Cardamine asarifolia a príbuzné druhy

Neighbor joining, Kimura 2 parameter, 1000 bootstraps

Minimum Evolution

For each possible alternative tree one can estimate the length of each branch from the estimated pairwise distances between taxa and then compute the sum (S) of all branch length estimates.

The minimum evolution criterion is to choose the tree with the smallest S value.

A. Wilmotte, Uni Liège