



Distance-based phylogenetic methods

“It is not the strongest of the species that survives, nor the most intelligent that survives. It is the one that is the most adaptable to change.” (C. R. Darwin)

Table of contents

- ✦ History of molecular phylogenetics
- ✦ Advantages of molecular phylogenies
- ✦ Phylogenetic trees
- ✦ Clustering methods and distance matrices
- ✦ Multiple sequence alignments

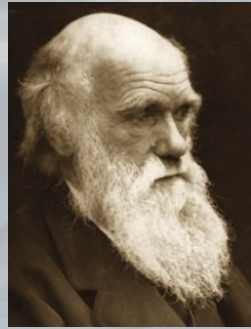
Introduction – 1

- ✦ The classification of organisms according to their species is the result of the phylogenetic reconstruction of their evolutionary history, an analysis that is now primarily conducted at the molecular level, based on the comparison of nucleotide and/or amino acid sequences
- ✦ **Molecular phylogeny**, also used for the study of the evolution of specific families of genes and proteins, is an analytical method established in the early '90s, rapidly grown thanks to the advances in the molecular biology and bioinformatic fields

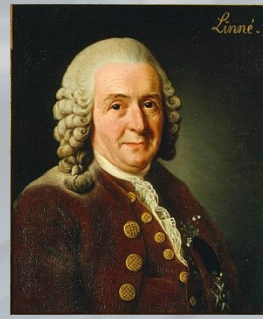
Introduction – 2

- ✦ The different types of molecular data are in fact a kind of historical document, which contains the traces of the basic steps in the evolution of a gene
- ✦ Furthermore, the events characteristic of the evolution of genes (substitutions, insertions, deletions and rearrangements) can be used also to resolve questions about the evolutionary history and relationships among entire species
- ✦ Molecular phylogeny is an important tool for the protein structure analysis, for the biodiversity conservation and for the epidemic control

History of molecular phylogenetics

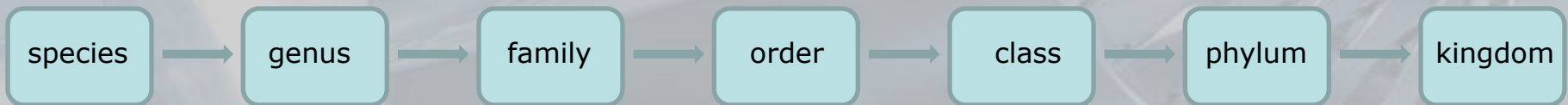


- ✦ **Taxonomy** deals with classification and naming of living organisms; it is used as a tool within the science of systematics
- ✦ **Taxonomists** began classifying and grouping living organisms long before the code of life and evolution was suspected to be written in their genomes
- ✦ Based on anatomy and physiology, taxonomy has produced remarkable insights, especially after that **Darwin's** ideas (1809–1882) showed how the system proposed by **Linnaeus** (1707–1778), for classifying organisms, actually reflected the evolutionary relationships among them



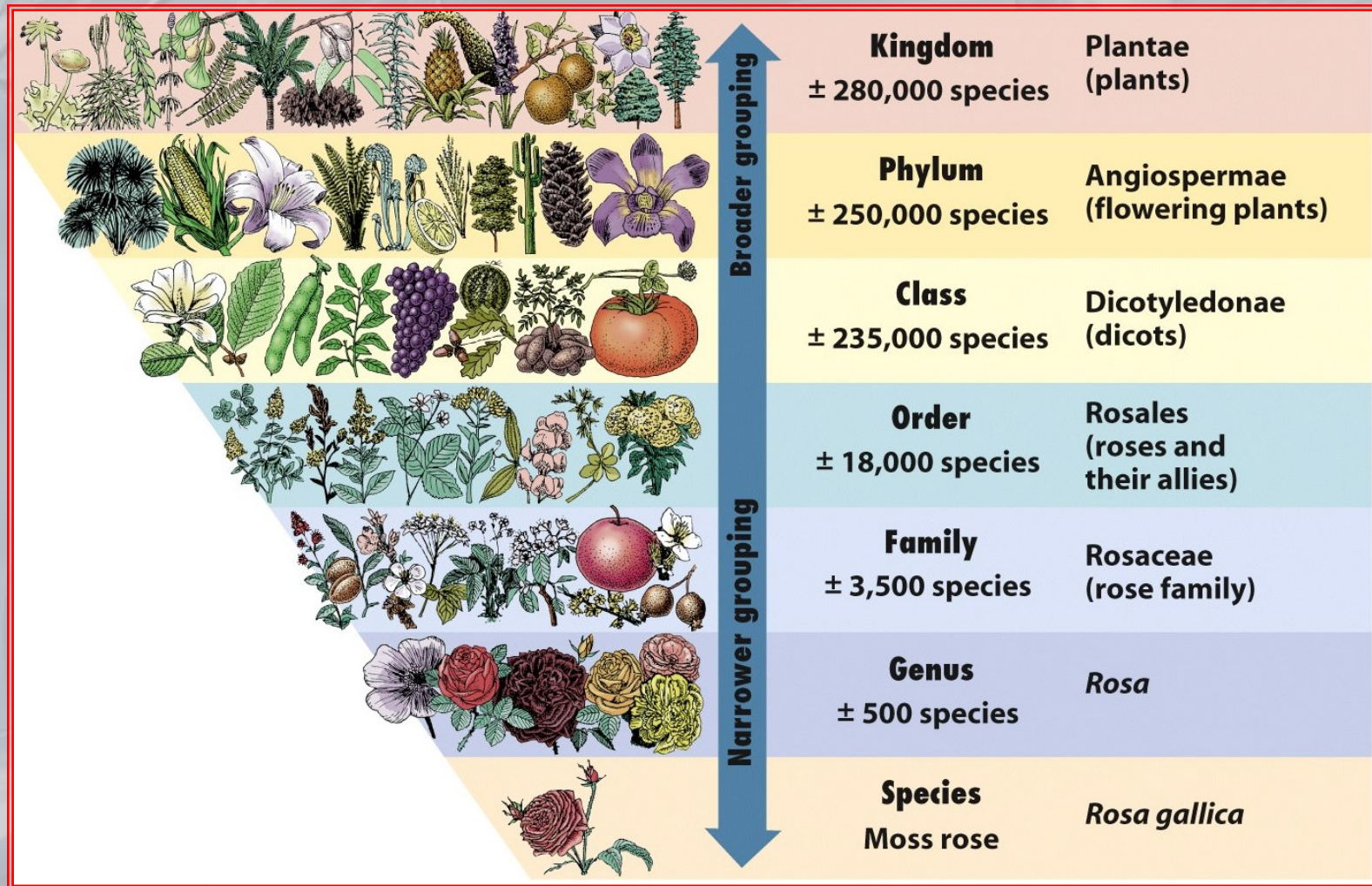
The Linnaeus taxonomic system

- **SPECIES:** This is the smallest category and includes organisms that share many features; organisms belonging to the same species can mate and have a fertile offspring
- **GENUS:** It includes species very similar to each other, such as donkey and horse or cat and lynx; in the case of mating, they can have only an infertile offspring
- **FAMILY:** It includes different genera that have some common characteristics; for instance, cat, lynx and lion belong to the same family (**Felidae**)
- **ORDER:** It includes many families who have common physical characteristics, such as the type of teeth; for example, dog and lion are very different, but both belong to the same order (**Carnivora**)
- **CLASS:** It includes many orders, with some common characteristics; for example, dog and horse, although different, belong to the same class of **Mammalia**
- **PHYLUM:** It includes many classes related to each other (mammals, birds, reptiles, amphibians and fishes all belong to the phylum of **Chordata**, collecting organisms that possess an internal support structure or a notochord)
- **KINGDOM:** It is the largest grouping that includes very different phyla
- Linnaeus grouped all living beings into two kingdoms: **the animal and the vegetable kingdoms**



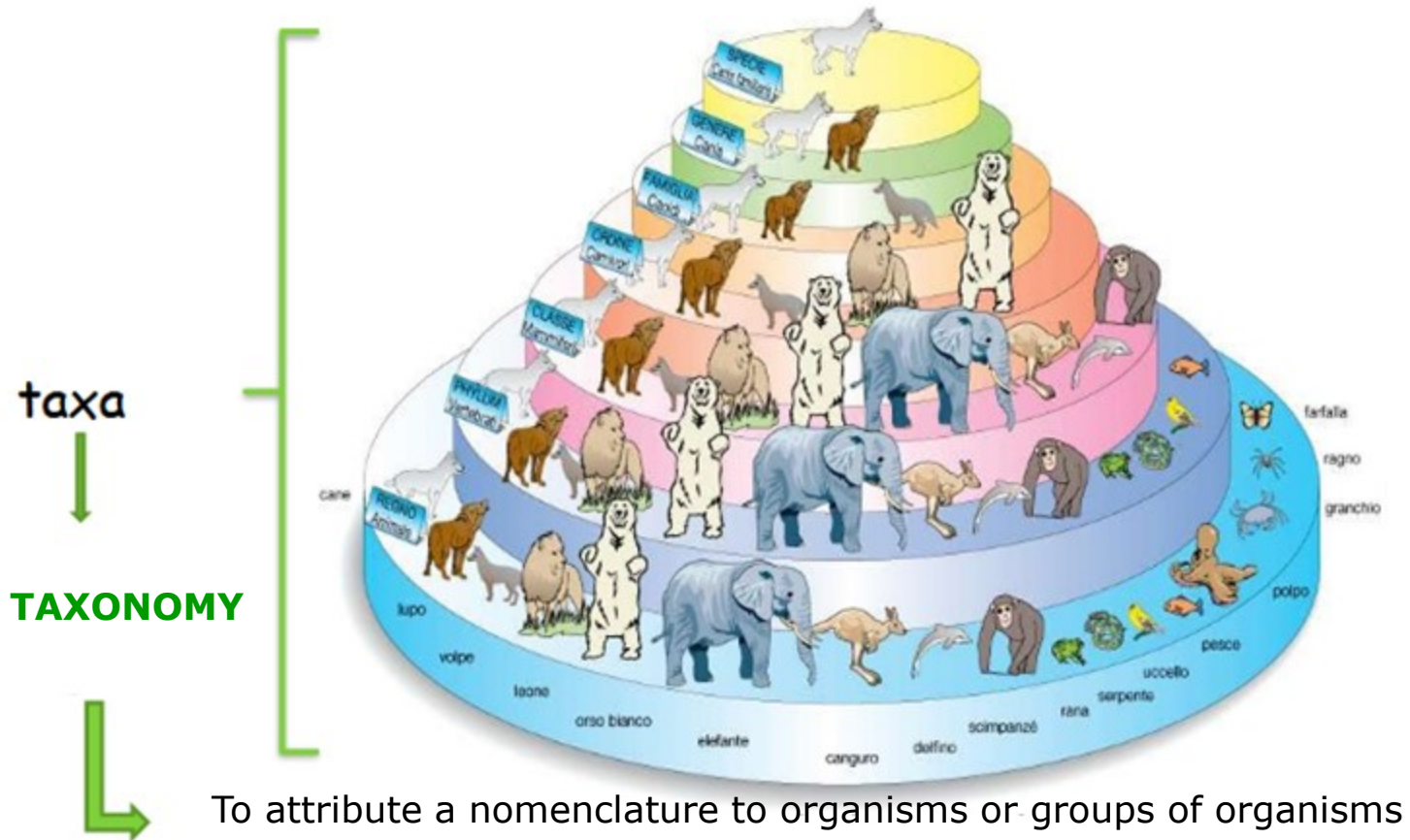
The Linnaeus taxonomic system

Example 1



The Linnaeus taxonomic system

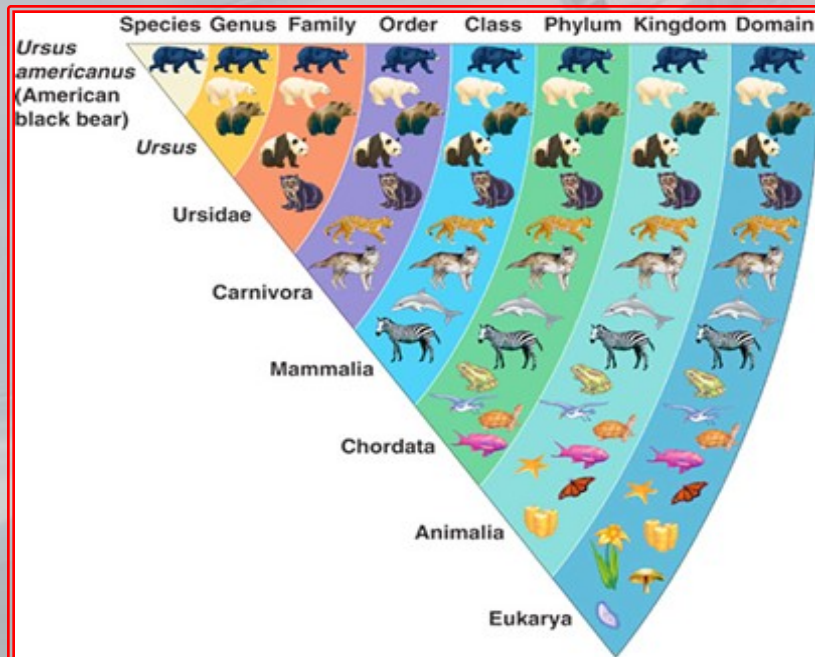
Example 2



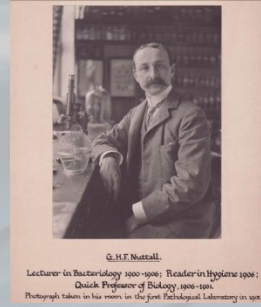
The Linnaeus taxonomic system

Example 3

Kingdom: Animalia
Phylum: Chordata
Class: Mammalia
Order: Artiodactyla
Family: Giraffidae
Genus: Giraffa
Species: Giraffe camelopardalis



History of molecular phylogenetics (cont.)



- ✦ The insights of Linnaeus and Darwin made it possible important applications, such as the development of new crops and the discovery of treatments against infectious diseases but, above all, they provided the awareness that all the living organisms – on the planet – share a single common ancestor
- ✦ Therefore, the study of similarities and differences at the molecular level seemed a natural addition to the tools commonly used by taxonomists, particularly after that G. H. F. Nuttall (1862–1937) showed that the intensity of the immune response, generated in an organism inoculated with the blood of another organism, is directly related to their evolutionary correlation (1902–1904)

History of molecular phylogenetics (cont.)

- ✦ Through these experiments, Nuttall examined the relationships among hundreds of living beings and concluded, for example, that humans and apes share a more recent ancestor with respect to the other primates
- ✦ Antibodies and their ability to interact with other molecules have until recently been used as a phylogenetic screening tool with organisms for which little nucleotide or protein data was available
- ✦ However, molecular data have been collected and have been used extensively for phylogenetic researches only after 1950

History of molecular phylogenetics (cont.)

- ✦ The protein electrophoresis technique permitted the separation of proteins and their comparison, according to their surface characteristics, such as size and charge
- ✦ Also, the protein sequencing became possible (since mid '60s), which was able to get the full amino acid sequence of many essential proteins
- ➡ A large amount of measurable molecular parameters with the possibility to go beyond morphological similarities

History of molecular phylogenetics (cont.)

- ✦ The speed at which the denatured genomes could hybridize provided some hints on the existing relationships between phylogenetically related organisms
- ✦ After that, since the early '70s, when genomic information has become available, first in the form of **restriction maps** (that describe the relative arrangement of the various sites recognized by restriction enzymes on the DNA sequence), and then as full DNA sequences, many mathematically rigorous approaches were developed, useful to molecular biologists
- ➡ It became possible to assign statistical confidence to phylogenetic groupings and it became also relatively easy to formulate testable hypotheses on evolutionary processes

History of molecular phylogenetics (cont.)

- ✦ Today, DNA data are much more abundant than any other form of molecular information
 - The traditional taxonomic approaches, based on morphological characteristics, continue to provide additional information to evolutionary studies, as well as paleontological records offer some clues to the time scan with which organisms differ and evolve
 - Techniques such as PCR, and NGS (*Next Generation Sequencing*) however, are the actual frontier of research, to answer the most salient questions about the history and the mutual relationships among all the living organisms on the planet

Advantages of molecular phylogenies – 1

- ✦ Since the evolution corresponds to a genetic change, genetic relationships are of primary importance in deciphering the evolutionary relationships
 - Hypothesis: Organisms with a high degree of molecular similarity are phylogenetically closer than those that show a lot of dissimilarities
- ✦ Before that molecular biology tools were able to provide data useful for the molecular phylogenetic analysis, taxonomists were forced to rely on the comparison of **phenotypes** (outward appearance of an organism) to infer their **genotypes** (the set of genes that encode for their aspect)
 - Similar phenotypes \Rightarrow similar genes that encode for the given phenotypes
 - Different phenotypes \Rightarrow different genetic code

Advantages of molecular phylogenies – 2

- ✦ Originally, in the phenotype examination, the most obvious anatomical features were considered; subsequently, also behavioral, ultrastructural and biochemical characteristics were taken into account

- Ultrastructure is the architecture of cells that is visible at higher magnifications than found on a standard optical light microscope; such cellular structures as organelles, which allow the cell to function properly within its specified environment, can be examined at the ultrastructural level
- Construction of morphological evolutionary trees still in use both for plants and animals

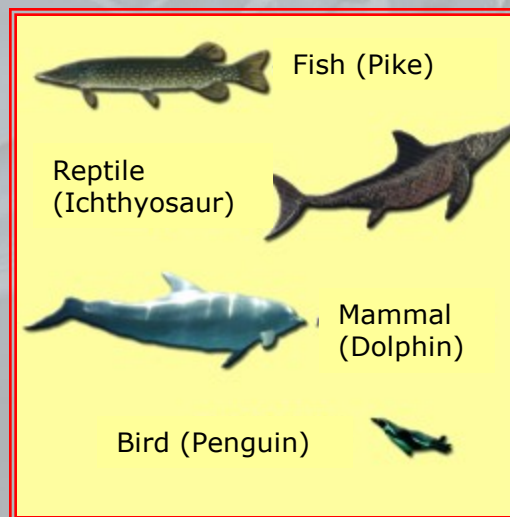
✦ Limitations

- Similar phenotypes can appear into phylogenetically distant organisms, due to **convergent evolution**, when two or more species, related to the same type of environment, develop morphological characters suitable for their habitat (at the same time, or even during very long periods of time)
- Difficulties in the selection of phenotypic information
- Difficulties in the study of phenotypic characteristics that can be used for comparisons among “distant” species

Advantages of molecular phylogenies – 3

+ Examples

- The hydrodynamic shape of the body, with paddled limbs and a bilobed backend has evolved at least four times during the history of the Earth: in fishes, in ichthyosaurs (reptiles), in dolphins (mammals) and in penguins (birds)
- Bacteria show a few easily observable features, even with a microscopic analysis
- What phenotypic characteristics can we select to compare bacteria, worms and mammals, so different from each other?

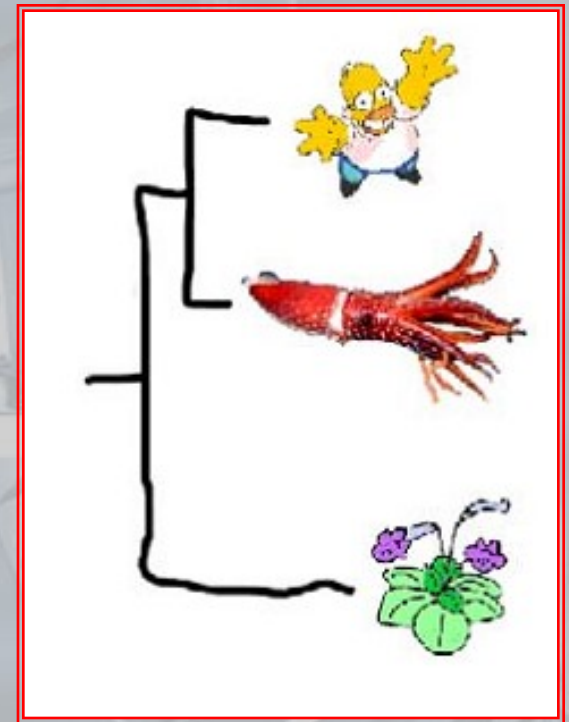


Advantages of molecular phylogenies – 4

- ✦ The analyses that are based on the nucleotide or protein sequences do not have these limitations, since many homologous molecules are essential for all living organisms (e.g.: 5s and 16s rRNAs)
- ✦ Even if the relative speed of molecular evolution can vary from one lineage to another (and the divergence times inferred from molecular analyses should therefore be treated with caution), molecular approaches for generating phylogenies are extremely reliable
 - Probably they are actually the most reliable methods, even in presence of alternative data (e.g., morphological), because molecular data are less sensitive to exogenous factors
 - ➡ In such cases where differences were found between morphological and molecular phylogenies, one can observe the effects of natural selection on phenotypic products

Phylogenetic trees – 1

- **Phylogenetic tree:** A graphical representation of the evolutionary relationships among three or more genes or species
- Through phylogenetic trees, it is possible not only to express the parental relationships within a set of data, but also to establish their time of divergence and the nature of their common ancestors



Phylogenetic trees – 2

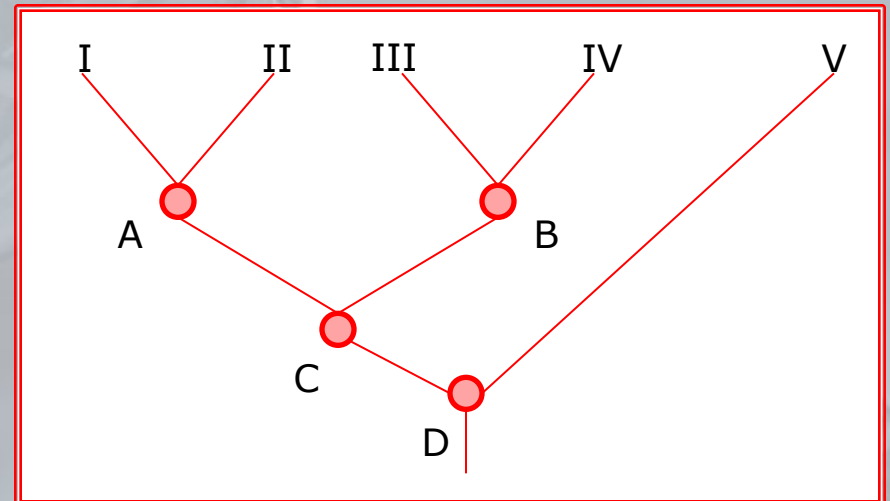
- ✦ Also known as **dendrograms**, in phylogenetic trees each node represents a distinct taxon
- ✦ **Taxon** (pl. taxa): A taxonomic unit, named or not, i.e. a population, or a group of populations of organisms, which are usually inferred to be phylogenetically related and which have characters in common able to differentiate the unit (e.g. a geographic population, a genus, a family, an order) from other such units
- ✦ **Terminal nodes** correspond to a gene or to an organism for which empirical data are available, while **internal nodes** represent a common ancestor, hypothetical or inferred, that gave rise to two independent lineages at some point in the past

Phylogenetic trees – 3

✦ Example

- Nodes I, II, III, IV and V are terminal nodes, that represent known organisms, for which sequential data are observable
- Internal nodes A, B, C and D represent inferred ancestors, for which no data are available
- An alternative representation of the tree is the [Newick format](#):

`((I, II), (III, IV)), V)`



Phylogenetic trees – 4

- ✦ Almost always, internal nodes have only two lineages, and they are, therefore, said to be **bifurcated**
- ✦ Nevertheless, also multiple lineages are possible, which give rise to **multifurcation**
- ✦ Multifurcated nodes can be interpreted in two ways:
 - An ancestral population gave rise simultaneously to three or more independent lineages
 - There have been two or more bifurcations at “almost” the same time in the past, but the small amount of available data makes it impossible to distinguish the order in which they occurred

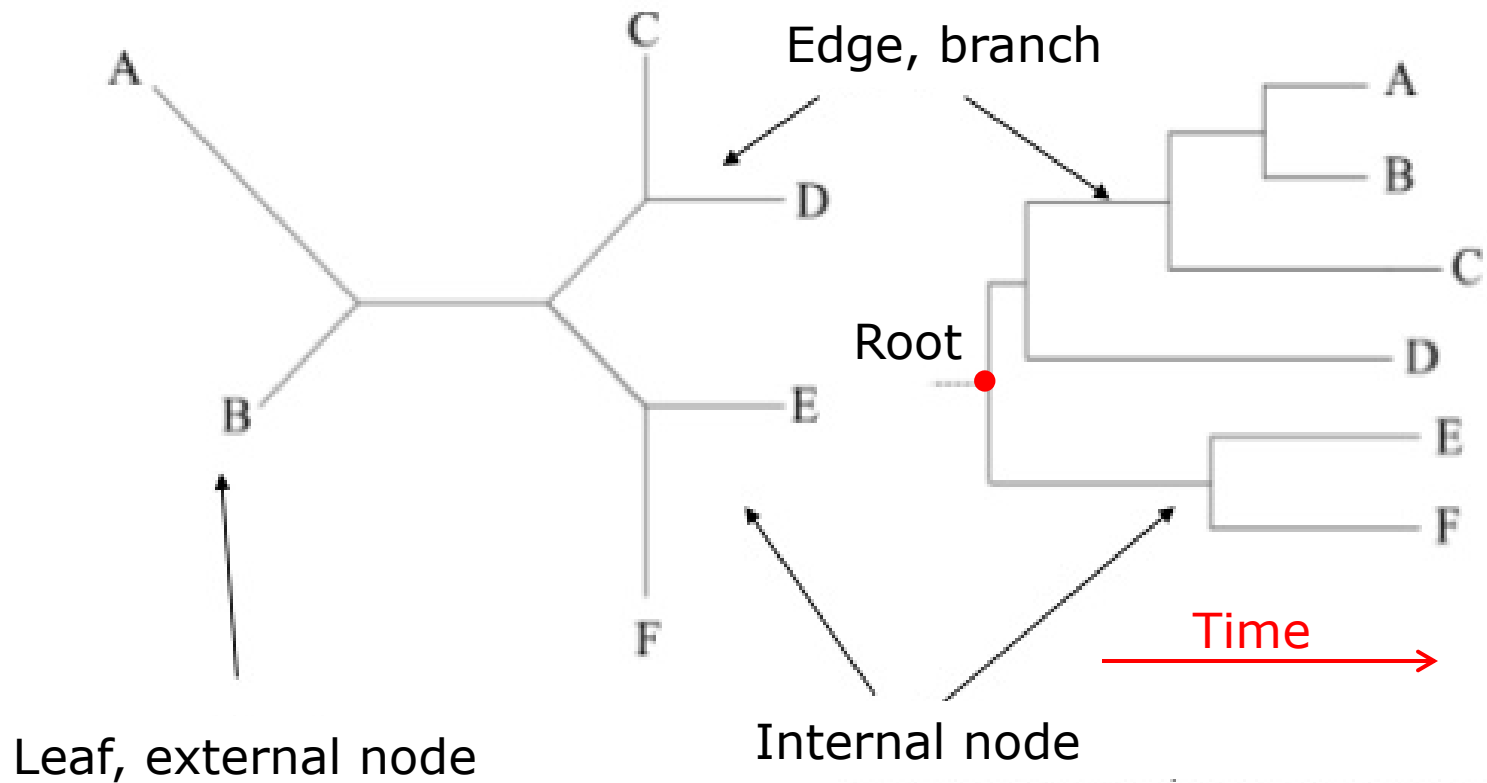
Phylogenetic trees – 5

- ✦ If the ramifications in a phylogenetic tree can be used to give information on the way in which evolutionary events occurred, the length of the branches can be employed to measure how much data diverge
 - **Scaled trees**, in which the arc lengths are proportional to the difference between pairs of adjacent nodes
 - ✕ **Additives trees**, in which the sum of the lengths of the branches connecting any two nodes is a representation of their accumulated differences
 - **Non-scaled trees**, in which all the terminal nodes are on the same level; only their relationships may be argued, without an estimation of their “distance”

Phylogenetic trees – 6

- ✦ Another important distinction may be established between phylogenetic trees that are able to infer a common ancestor, and the direction of evolution, and those which cannot
- ✦ In **rooted trees**, a single node is defined as the unique ancestor and an evolutionary path exists from it to any other node in the tree
- ✦ **Unrooted trees** specify only the existence of relations between adjacent nodes, but do not provide any information about the direction in which evolution took place
 - A root can be assigned to an unrooted tree using an **outer group**, i.e. a species that was previously divided from the other species represented in the tree
 - **Example:** In the case of men and gorillas, when the baboons are used as the outer group, the root of the tree can be placed somewhere along the branch that connects baboons to the common ancestor of men and gorillas

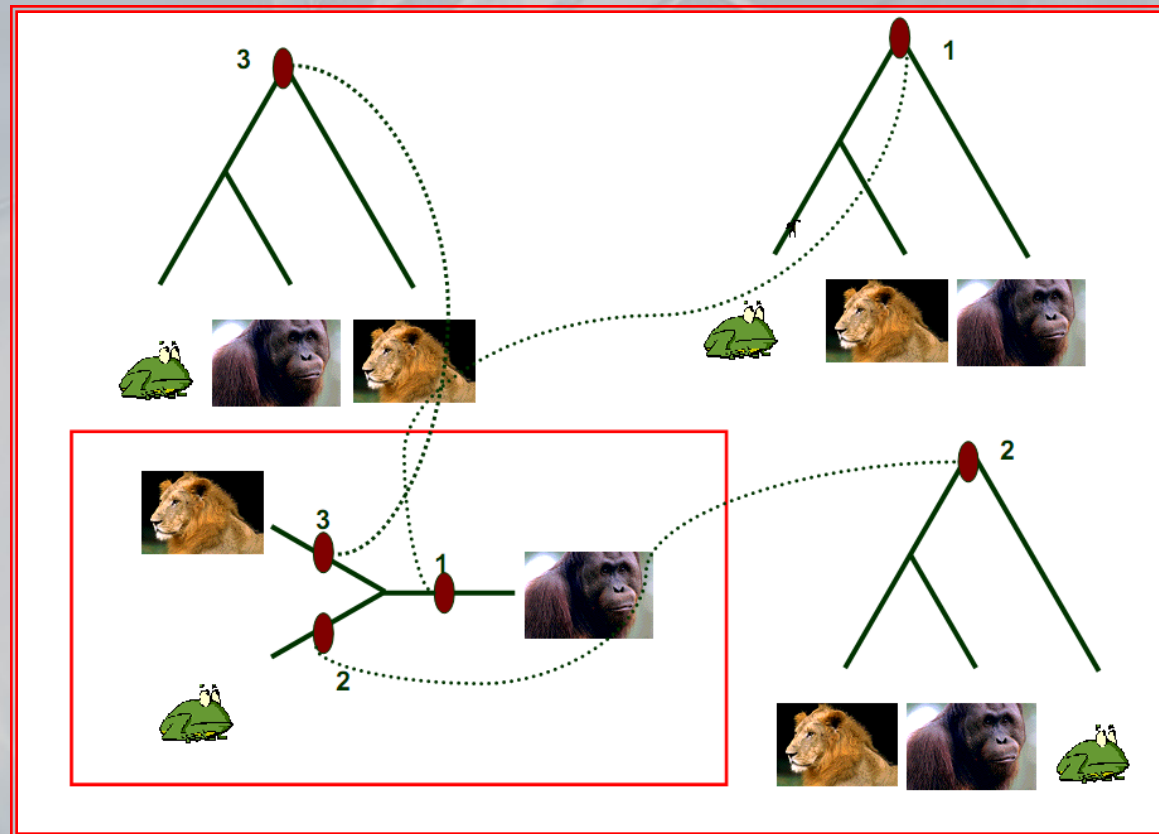
Phylogenetic trees – 7



Unrooted vs. Rooted trees

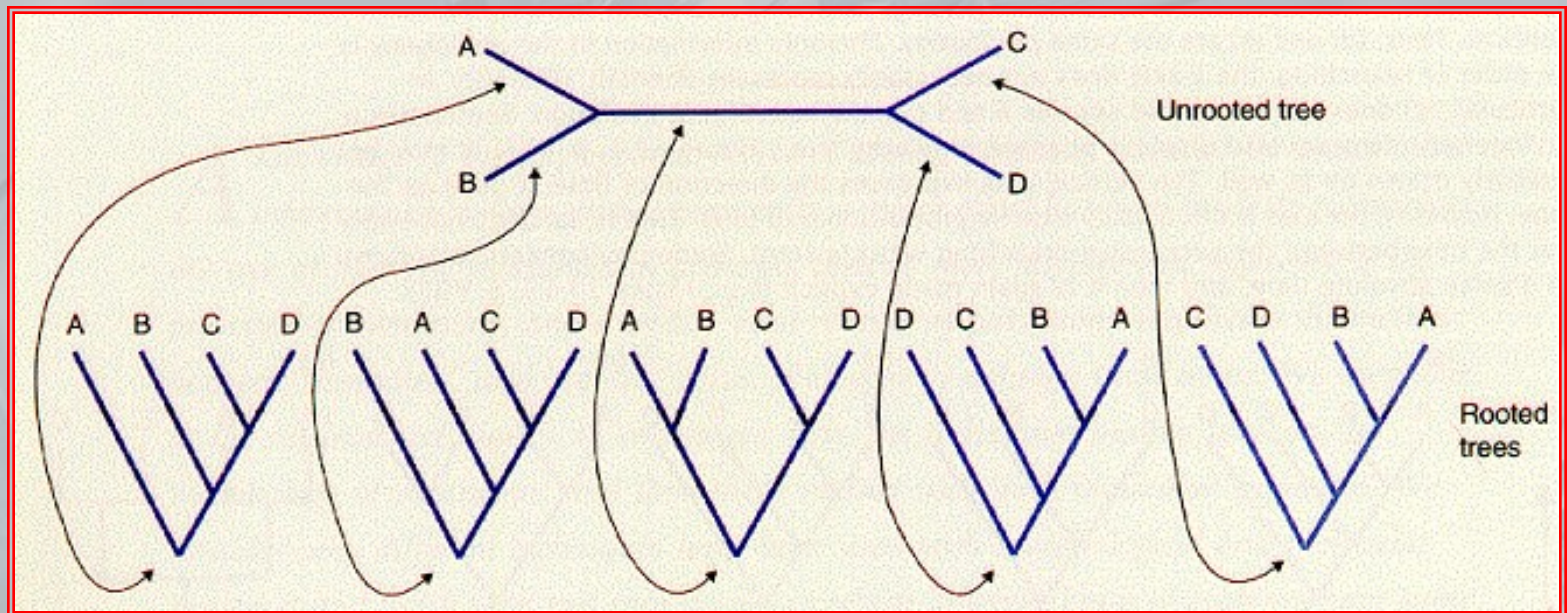
Phylogenetic trees – 8

- ✦ In a situation where only three species are considered, three rooted trees and only one unrooted tree can be drawn



Phylogenetic trees – 9

- More generally, for each unrooted tree, there are $2s-3$ rooted trees, where s is the number of taxonomic units
 - $2s-3$ corresponds to the number of branches of the unrooted tree



Phylogenetic trees – 10

• For any s :

• $N_R = (2s-3)!/[2^{s-2}(s-2)!]$

• $N_U = (2s-5)!/[2^{s-3}(s-3)!]$

Number of species	Number of rooted trees	Number of unrooted trees
2	1	1
3	3	1
4	15	3
5	105	15
10	34459425	2027025
15	213458046767875	7905853580625
20	8200794532637891599375	221643095476699771875

Phylogenetic trees – 11

- ✦ Not even the fastest computers can cope with such a computational explosion, in order to assess the relative quality of all the possible trees for more than a few hundred sequences or species
 - The exhaustive search is totally unfeasible
 - We should try to focus only on those trees that, most likely, can reflect the actual relationships among the various sets of data
 - ▀ However, only one of these trees describes the “true” evolutionary path followed by the considered genes or species

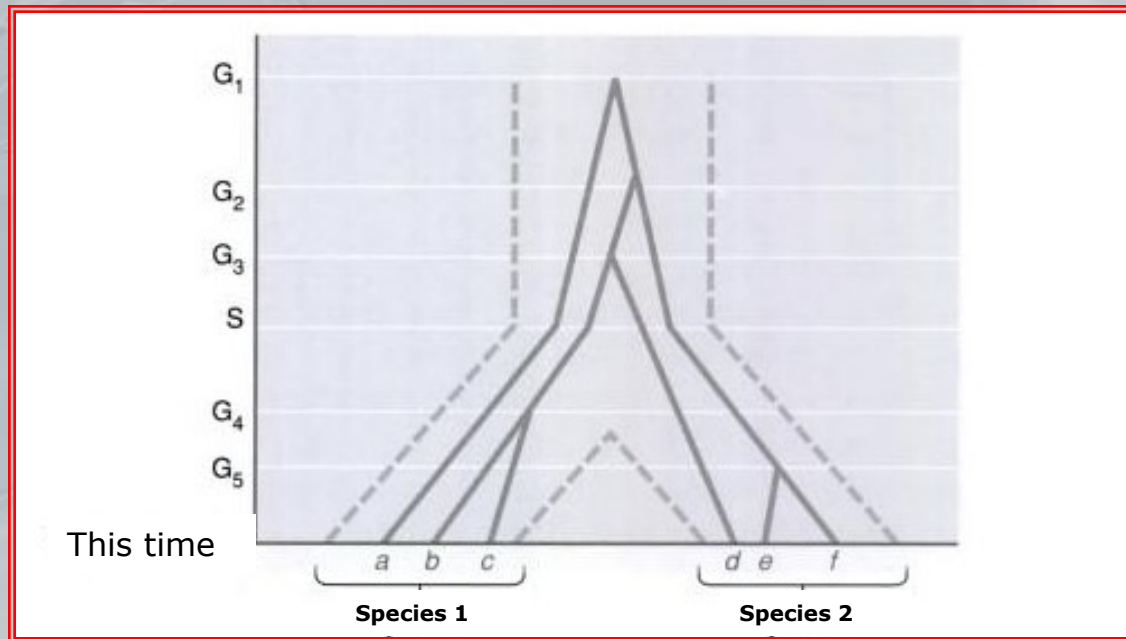
Trees of genes versus trees of species – 1

- ✦ Phylogenetic trees based on the observed divergence among homologous genes are called **trees of genes** (to be distinguished from trees of species)
 - They represent the evolutionary history of a gene, not necessarily that of the species in which it is found
- ✦ **Trees of species** are obtained from the analysis of data coming from multiple genes
 - **Example:** About a hundred of different genes have been used to generate a phylogenetic tree describing the evolution of plant species
 - Trees of species are important, since the evolution occurs at the *population* level, and cannot be studied with respect to single individuals

Trees of genes versus trees of species – 2

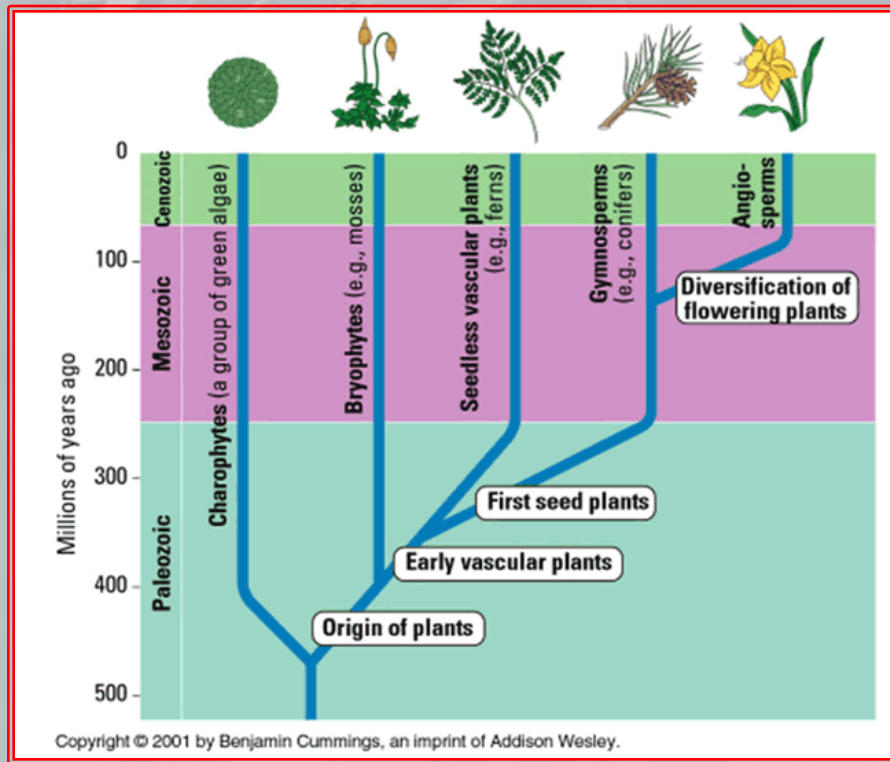
- ✦ Differences at the gene level typically occur before (or even after) a population divides, which happens when (two) new species emerge
- ✦ The difference between trees of genes and species tends to become particularly important when we consider *loci* whose diversity within populations is beneficial, such as the human leukocyte antigen HLA *locus*
 - Using only HLA alleles to determine a tree of species, many men would be grouped with gorillas, because the origin of the HLA polymorphism predates the speciation of primates

Trees of genes versus trees of species – 3

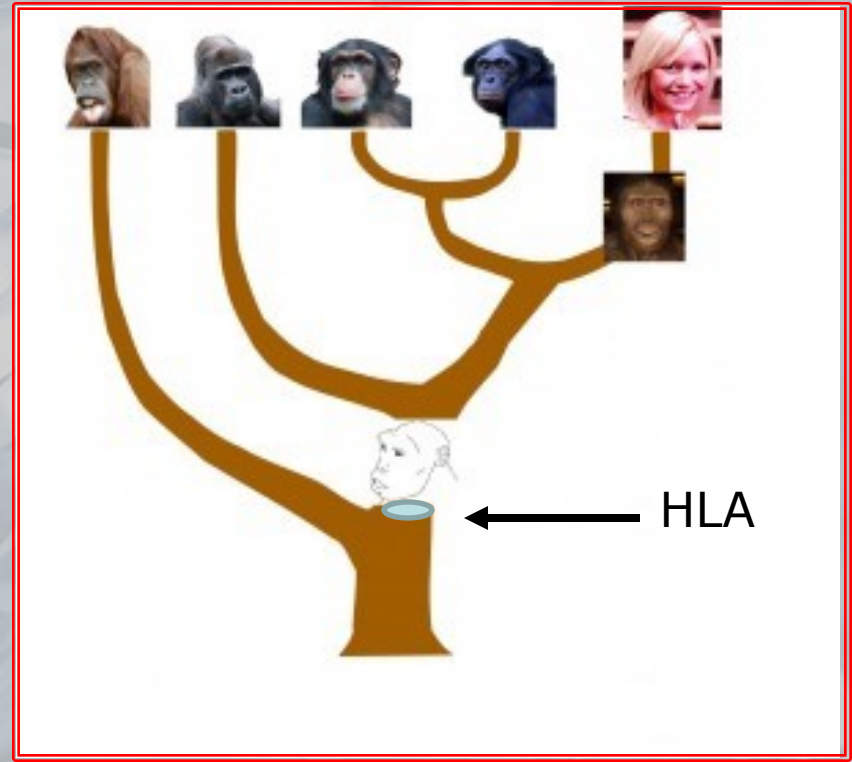


When considering a single gene, individuals may appear phylogenetically closer to members of other species than to their own: The genetic divergence events (from G_1 to G_5) occur both before and after the event of speciation (S); the organism with allele d , although being a member of species 2, would seem to be closest to the individuals of species 1, based on the considered *locus*

Trees of genes versus trees of species – 4



Phylogenetic tree of plant species



Phylogenetic tree of primates and divergence of the HLA gene

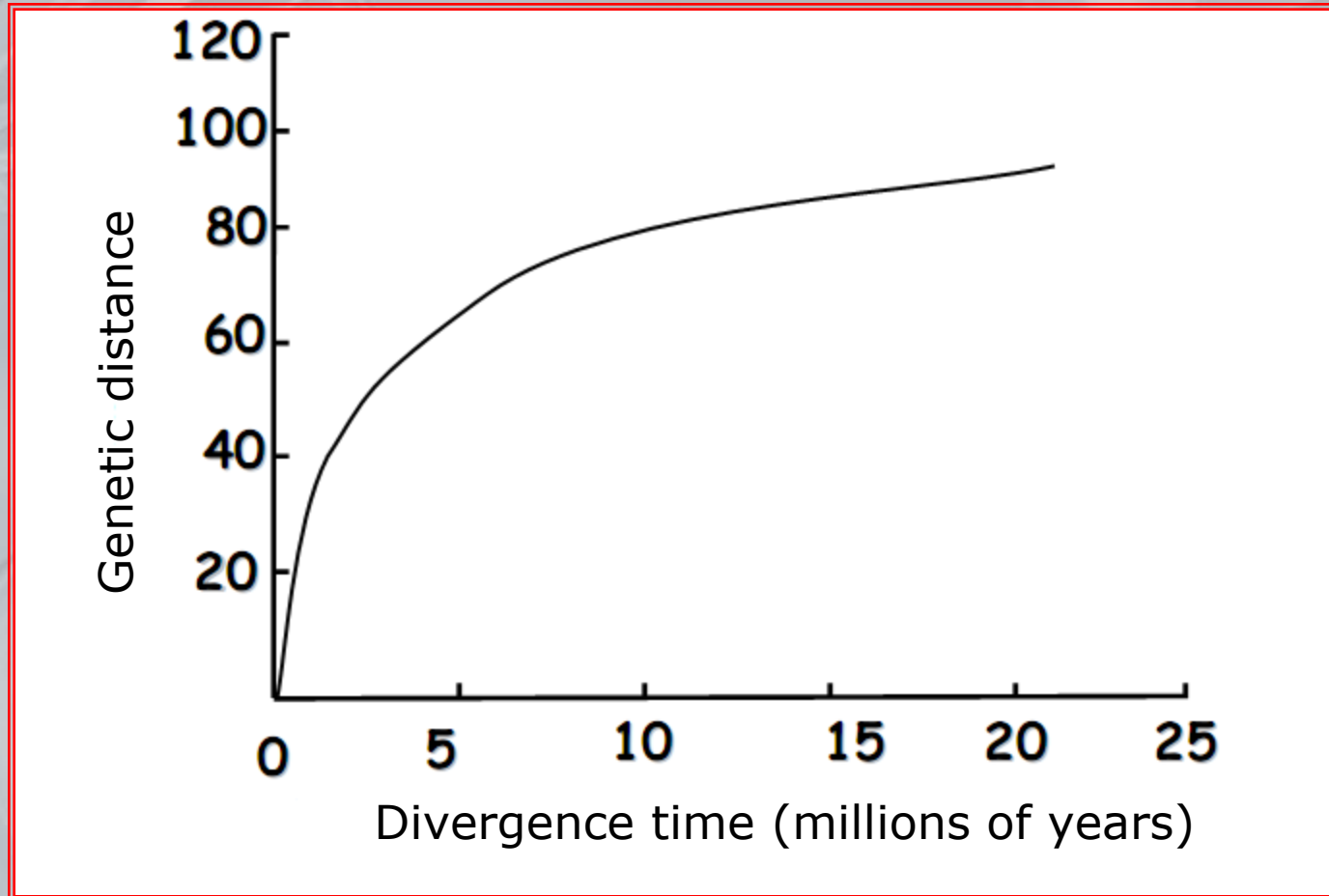
Trees of genes versus trees of species – 5

- ✦ Advantages in the use of trees of genes
 - Unambiguous description of the data
 - No interference with similarities due to non-genetic environmental effects (convergent evolution often implies similar phenotypes but different genotypes)
 - Divergence time (i.e. the length of the branches) easier to estimate
 - Rigorous statistical models
 - Possibility of analyzing also non-coding DNA sequences
 - All individuals have the DNA!

Trees of genes versus trees of species – 6

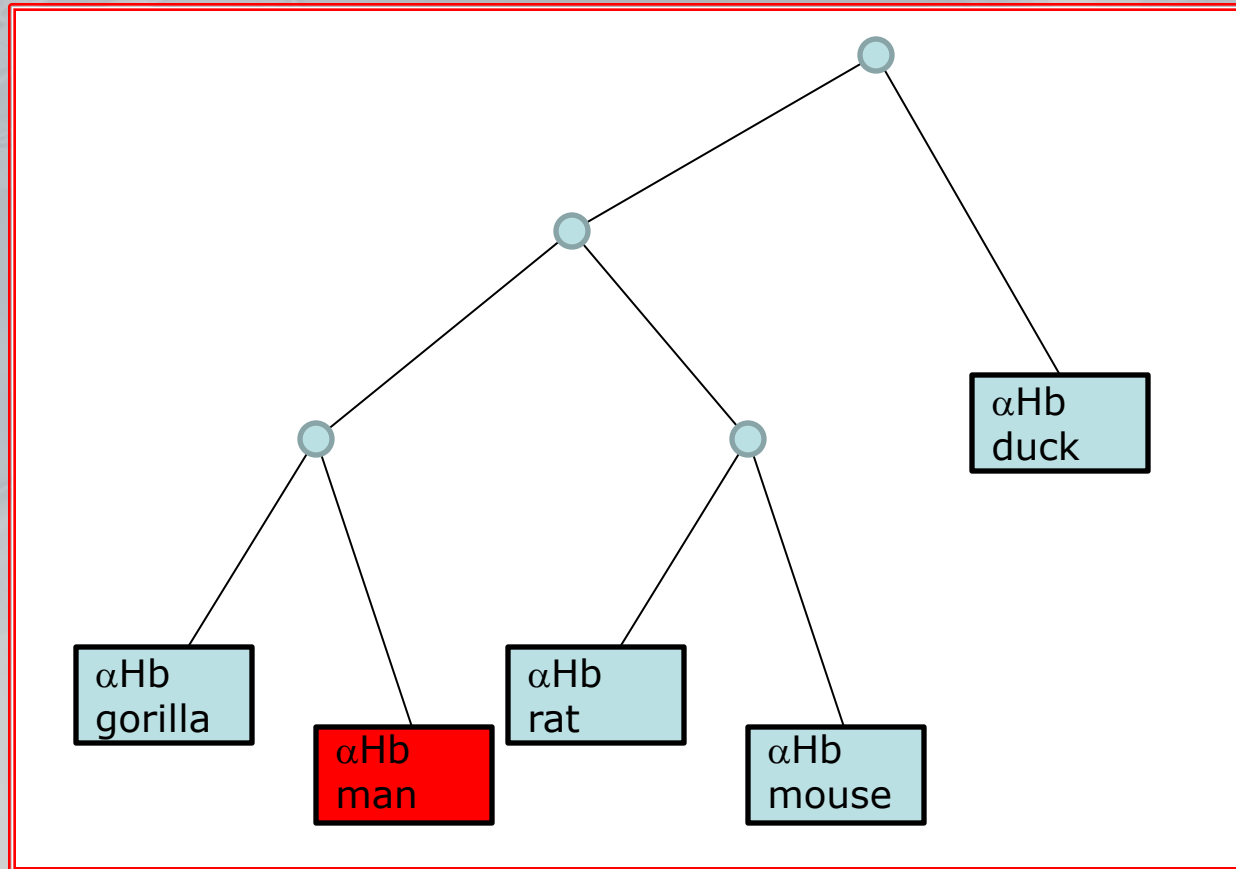
- ✦ Disadvantages of using trees of genes
 - Common, recurrent mutations may alter the relationship between genetic and temporal distances
 - Duplication and horizontal gene transfer can be identified, but they can create, anyway, some problems in the phylogenetic reconstruction
 - The homoplasy (which consists of a simple similarity with an ancestor who, despite having the same trait, has not hereditarily transmitted it to the subject under study) may be frequent
 - Homology (i.e., similarity due to inheritance from an ancestor who owns that particular character) and homoplasy cannot be easily distinguished through a detailed analysis, so as for the phenotypic traits

Trees of genes versus trees of species – 7



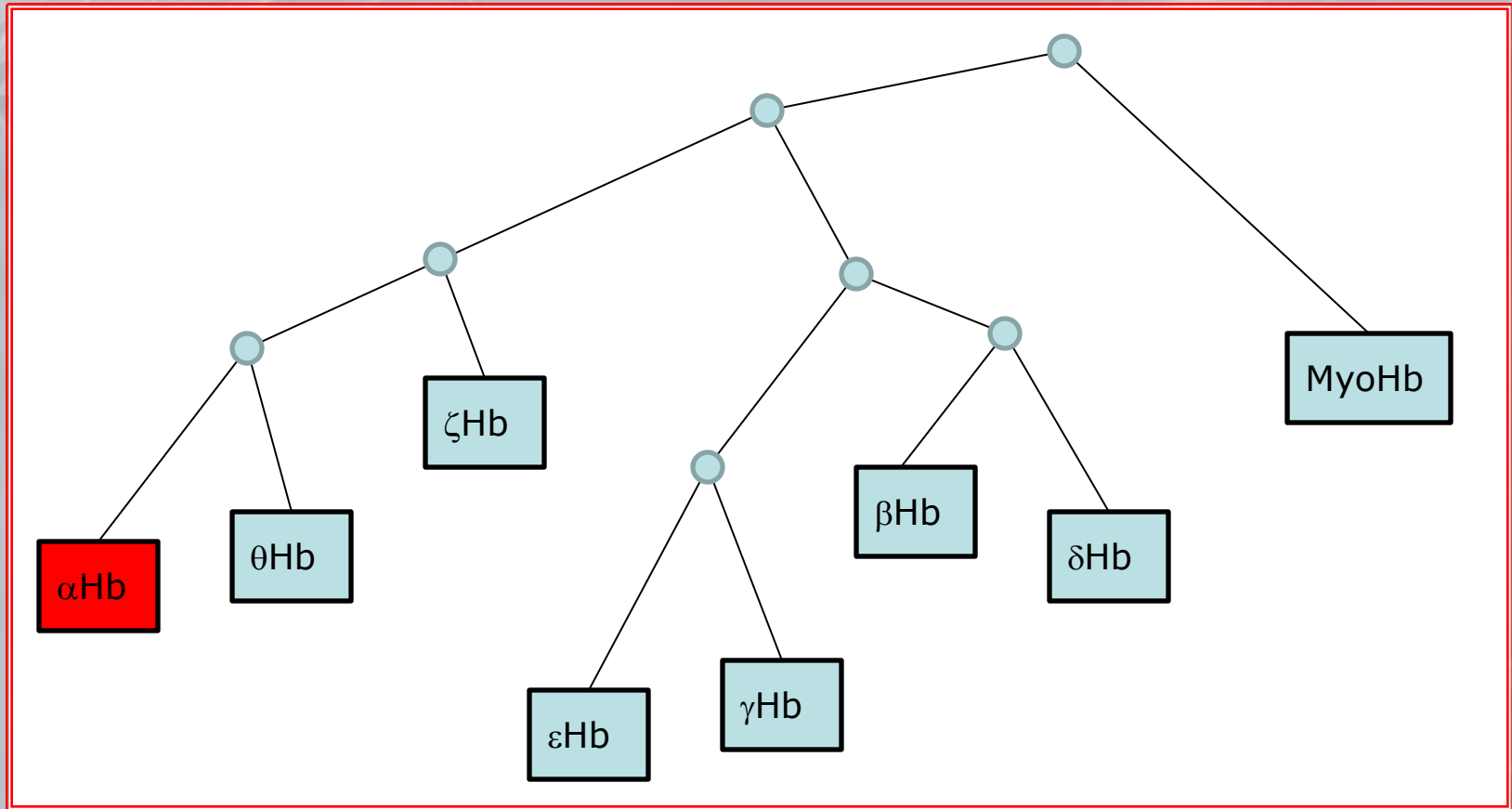
The relationship between genetic distance and divergence time is not linear, since the same locus may have undergone multiple substitutions during the evolution

Trees of genes: Orthologous genes



The phylogenetic tree for orthologous genes of α -globin in different species

Trees of genes: Paralogous genes



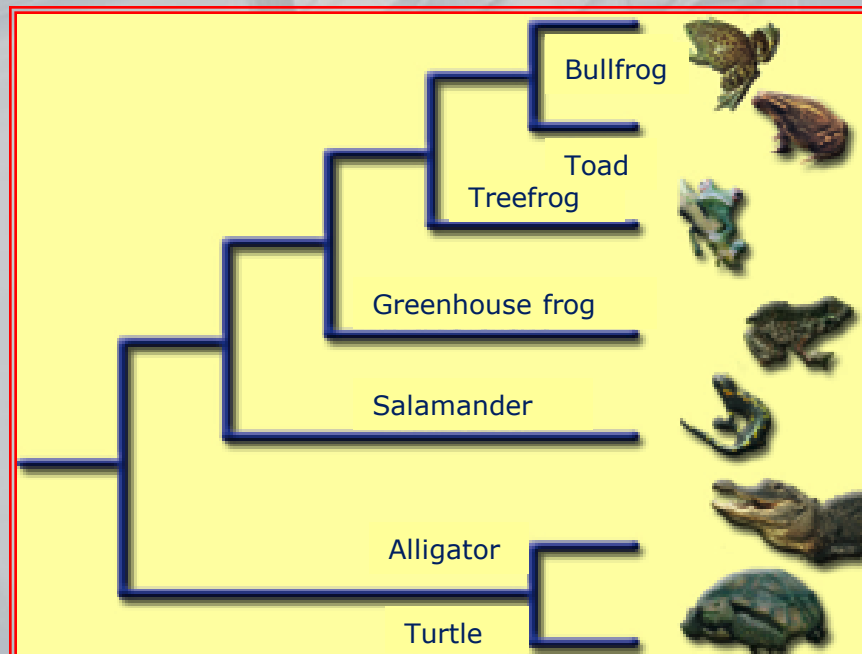
The phylogenetic tree for paralogous globin genes in man

Character and distance data – 1

- ✦ Molecular data, used to generate phylogenetic trees, belong to two categories
 - Characters (well-defined features which occur in a limited number of different instances)
 - Distances (measure of the difference between two sets of data)
- ✦ Both nucleotide and amino acid sequences are examples of data described by a finite alphabet (a set of discrete instances of characters)
- ✦ Other sets of character data are those that are encountered in the taxonomy based on anatomical or behavioral characteristics, such as the color of an organism or the amount of time it takes to react to a particular stimulus

Character and distance data – 2

Species	Tympanic membrane	Shell eggs	Aquatic life	Bellows mandible
Bullfrog	1	0	0	0
Toad	1	0	0	0
Treefrog	1	0	1/0	0
Greenhouse frog	1	0	1/0	0
Salamander	1	0	1	0
Alligator	0	1	1	1
Turtle	0	1	1	0



Character and distance data – 3

- DNA sequences are now so abundant that it is rare to have sets of data that originate from distance measures, such as those generated by DNA–DNA hybridization experiments between genomes of different organisms
- Nevertheless, character data can easily be converted into distance data, once established some appropriate criteria to determine the similarity between all the possible character states

Character and distance data – 4

- ✦ For example, a distance value D between two genes may be calculated as $D=n/l$, where n is the number of observed mismatches in an alignment, while l represents its length
- ✦ To enhance the distance:
 - Adjustments to take into account different frequencies for transition and transversion
 - Adjustments to account for multiple/local substitutions
 - Normalization to get “the number of changes per 100 nucleotides”

Character and distance data – 5

- ✦ The distance between proteins can be calculated in a similar way, by aligning the amino acid sequences
 - Loss of potentially useful information
 - Great difficulty in the comparison between protein sequences: not only is more likely that some amino acids are replaced with others depending on similar chemical activity of their functional groups, but also the number of substitutions at the DNA level can vary, in order to obtain an amino acid substitution

Character and distance data – 6

- ✦ Computational approaches, used for the construction of phylogenetic trees, generally neglect the importance of certain subtleties present in biological datasets
- ✦ **Phenetics**, the approach proposed by R. Sokal and P. Sneath in 1963, is an attempt to classify organisms based on overall similarities, usually related to morphological or other observable traits, regardless of their phylogeny or evolutionary relation
 - Phenetists do not give different weights to the various characters: to each of them a value is assigned (0 for the absence, 1 for the presence); then, the closer species are those that share a greater number of characters
 - The accuracy of the method improves as the number of selected characters increases
- ➡ Relationships between (measurable) sets of data are highlighted, without paying particular attention to the evolutionary paths followed to reach the current state

Character and distance data – 7

- ✦ Phenetics gives up on principle to the concept of species as a real entity in Nature
- ✦ Phenetists “divide living organisms based on sets of characters” (any quality detectable by observation), and – by claiming to be unable to discriminate between homologies and analogies – aseptically and similarly process all characters, using statistical algorithms
- ✦ The computer output is considered for what it is: an aseptic, ahistorical and artificial classification of life, which has the unique purpose to bring order in the living world, but does not pretend to infer anything
- ✦ Even the word “species” disappears from the pheneticistic jargon, being substituted by **Operational Taxonomic Unit** (OTU), a precisely operating concept, not defined as an ontological category

Character and distance data – 8

- ✦ **Cladists**, conversely, are more interested in evolutionary paths and patterns, preferring a “biological” approach for the construction of phylogenetic trees
 - The main objective of cladistics is in fact that of classifying living organisms, based on the phylogenetic hierarchy that results from the history of the life on the Earth; because this was unique, it provides absolute objectivity to this type of classification
 - Founder of the cladistic school was considered the German entomologist W. Hennig, even if he never spoke of cladistics, but of *phylogenetic systematics*
 - His idea was that to divide all the living beings into “clades”: since, in Nature, when a species is divided gives rise to two descendant species (sibling species), we can consider as a taxonomic group the set composed by the two new species and by their common ancestor
 - In this way, a natural classification arises, that can theoretically go up to the first living organism on the Earth

Character and distance data – 9

- ✦ Cladistics not only considers it possible to discriminate between homologies and analogies, but even further distinguishes two different types of homology, called “apomorphies” and “plesiomorphies”
- ✦ Apomorphies are “recent homologies”, those that clearly define a group, while plesiomorphies are still homologies, but so widely shared to be uninformative
- ✦ For example, the wings are apomorphies within vertebrates, in the sense that they define a subgroup that might be called “birds”, but become plesiomorphies when considering only birds, in the sense that it is totally unnecessary to refer to wings in order to categorize groups contained in the macrogroup of birds (since all the birds share this characteristic)

How to reconstruct the phylogeny?

- ✦ Distance-based methods
- ✦ Parsimony-based methods
- ✦ Likelihood-based methods

Likelihood-based approaches describe which is the probability that a certain hypothesis H , a phylogenetic tree, corresponds to a certain set of data D , a multiple alignment
High computational complexity

The principle of maximum parsimony – very important in the natural processes – searches for a tree that requires the smallest number of evolutionary changes to explain differences observed among the OTUs under study. Such a tree is called a maximum parsimony tree. Often more than one tree with the same minimum number of changes can be found, so that a unique tree cannot be inferred.

Distance-based methods

✦ Advantages

- Speed: suitable for analyzing large sets of data (because of their polynomial computational complexity)
- Based on the use of distance matrices (the only available data, for example, in the case of DNA hybridization, reactions to antibodies, etc.)
- Distance-based methods build **phenograms**

✦ Disadvantages

- Loss of information: starting from the distances, we cannot reconstruct the sequences!
- Problems with distances which are nonlinear in time

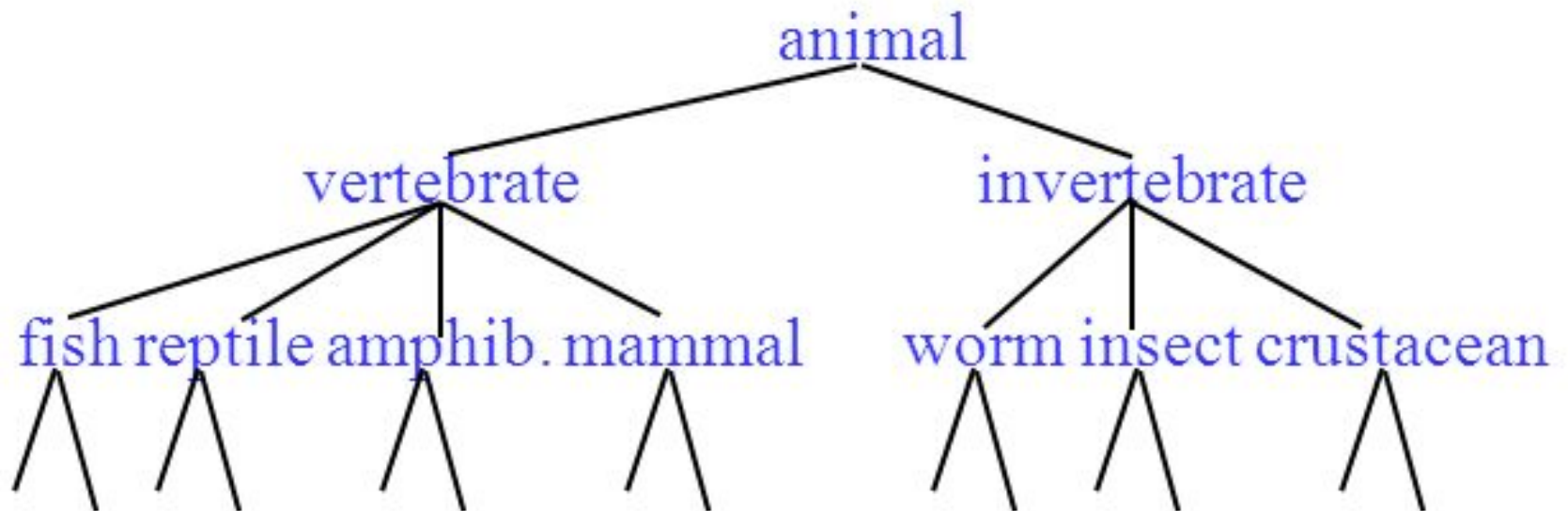
Hierarchical clustering – 1

- ✦ Grouping sequences means associating them together so that the sum of all the distances between the sequences in the same group is minimal
- ✦ The simplest clustering algorithms presuppose to know the number of (leaf) clusters a priori
- ✦ However, for building a tree, just dividing sequences into separate groups is not enough but, in turn, the groups must be grouped together to form larger entities, and so on, until there is a single group that includes all the sequences (defining the tree root)
- ✦ This type of clustering is called **hierarchical**: A single object does not belong to a single group, but to several groups that are contained within each other

Hierarchical clustering – 2

- ✦ Actually, in a phylogenetic tree the sequences are collected into groups in a hierarchical way
- ✦ In the so-called **agglomerative hierarchical clustering**, in particular, objects (single and/or clustered sequences) are grouped into pairs, starting from the closest objects, until there is only one group within which all the objects are contained
- ✦ In agglomerative hierarchical clustering, it will sometimes be necessary to calculate the distance between a sequence and a cluster already formed or between two clusters

Hierarchical clustering – 3



Hierarchical clustering – 4

- ✦ While it is clear how to estimate the distance between two sequences (using a distance matrix), the concept of distance in the case of compound entities needs to be defined
 - ▶ Minimum distance among all the sequences belonging to two different clusters
 - ▶ Maximum distance among all the sequences belonging to two different clusters
 - ▶ Average distance among all the sequences belonging to the two clusters
- ▶ Anyway... combine the two clusters with the smallest distance
- ✦ In the general case, the complexity of agglomerative clustering techniques is $\mathcal{O}(n^2 \log(n))$, with n representing the number of leaves – slow for grouping large data-sets

Distance matrix–based methods – 1

- ✦ Among all the possible trees, distinguishing which is the best one for describing the evolution of a group of genes or organisms is a difficult task
- ✦ Pairwise distance matrices — tabular representations of the differences between all the data to be analyzed — constitute the typical input to the algorithms for the calculation of phylogenetic trees
- ✦ **UPGMA** (*Unweighted–Pair–Group Method with Arithmetic mean*) is the oldest and also the simplest approach among distance matrix–based methods
 - Information on the genetic distance between all the considered taxa should be available, in order to construct the distance (lower triangular) matrix
 - UPGMA is a hierarchical (agglomerative) clustering method which uses average distances

Distance matrix-based methods – 2

- Let us assume that the distances between each pair of taxa in the set $\{A,B,C,D\}$ are collected in the following matrix:

Species	A	B	C
B	d_{AB}	–	–
C	d_{AC}	d_{BC}	–
D	d_{AD}	d_{BD}	d_{CD}

- d_{AB} represents the distance between A and B (the number of mismatched nucleotides, divided by the length of the aligned sequences, for instance)
- d_{AC} is the distance between A and C
- ...

Distance matrix–based methods – 3

- ✦ In the first phase of the UPGMA algorithm, the two species separated by the shortest distance are identified, placing them in the same composite group
 - Assuming that the smallest value within the matrix corresponds to d_{AB} , we first group together the two species A and B into (AB)
- ✦ After the first grouping, a new distance matrix is evaluated, in which the distances between the new group (AB) and the species C and D are calculated as the arithmetic mean of the original distances of the two species constituting the group

$$d_{(AB)C} = 1/2(d_{AC} + d_{BC})$$

$$d_{(AB)D} = 1/2(d_{AD} + d_{BD})$$

Distance matrix-based methods – 4

- ✦ Again, in the new matrix, the two species separated by the smallest distance will be identified, in order to group them in a new composite species
- ✦ The process is repeated until a single group is obtained, which includes all the species to be analyzed
- ✦ If, in order to represent the evolutionary distance between species, a scaled tree is used, from the branch points, two outgoing arcs of the same length will be obtained (each one having a length equal to a half of the distance between the grouped species – based on the molecular clock hypothesis)

Distance matrix-based methods – 5

✦ Example (to be continued)

Let us consider the following multiple alignment

A: GTGCTGCACG GCTGAGTATA GCATTTACCC TTCCATCTTC AGATCCTGAA
B: ACGCTGCACG GCTCAGTGCG GTGTTTACCC TCCCATCTTC AGATCCTGAA
C: GTGCTGCACG GCTCGGCGCA GCATTTACCC TCCCATCTTC AGATCCTATC
D: GTATCACACG ACTCAGCGCA GCATTTGCCC TCCCGTCTTC AGATCCTAAA
E: GTATCACATA GCTCAGCGCA GCATTTGCCC TCCCGTCTTC AGATCTAAAA

The pairwise comparison leads to the matrix

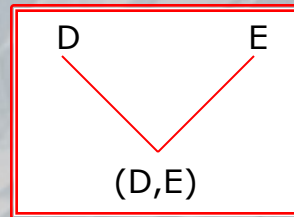
Species	A	B	C	D
B	9	–	–	–
C	8	11	–	–
D	12	15	10	–
E	15	18	13	5

- Given that all the sequences have the same length and contain no gaps, the distances are calculated as the number of mismatched nucleotides in each pairwise alignment

Distance matrix-based methods – 6

✦ Example (to be continued)

The shortest distance between two sequences for the considered multiple alignment corresponds to d_{DE} ; then the species D and E are grouped



while a new distance matrix will be calculated based on this new composite group (DE)

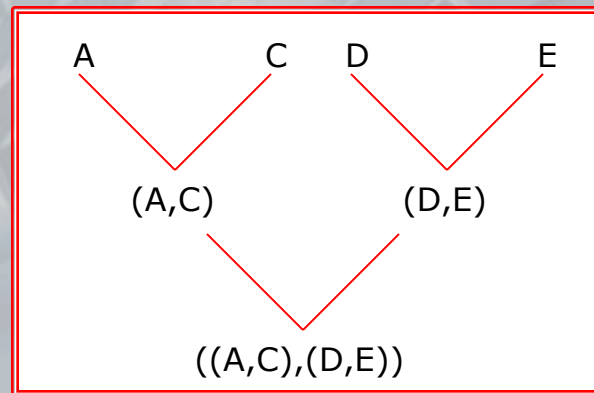
Species	A	B	C
B	9	–	–
C	8	11	–
DE	13,5	16,5	11,5

The distances between the remaining species and the new group will be determined by considering the average distance between its two components (D and E) and all the other species

Distance matrix-based methods – 7

✦ Example (to be continued)

In the new matrix, the shortest distance between two species is now related to A and C, that, therefore, form the new group (AC)



whereas, the distance matrix is recalculted as

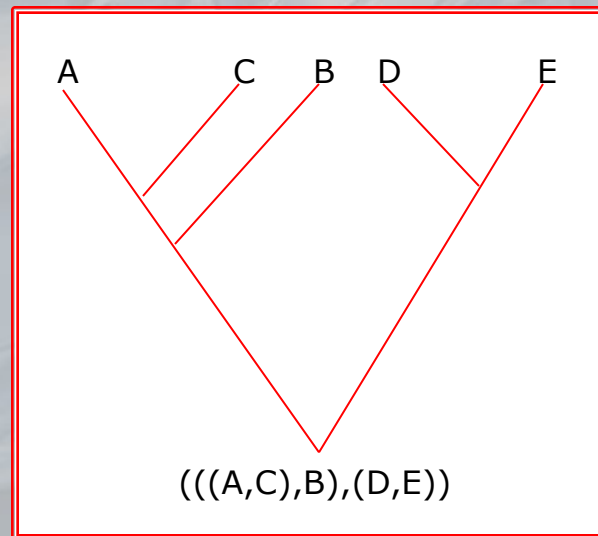
Species	B	AC
AC	10	–
DE	16,5	12,5

Distance matrix-based methods – 8

✦ Example

Finally, in this last matrix, the minor distance is that between (AC) and B ($d_{(AC)B}=10$), that are grouped together

Therefore, the complete phylogenetic tree is:



Distance matrix-based methods – 9

- ✦ The distance matrix evaluation, used by the UPGMA method, represents the computationally more expensive calculation in the process which leads to the construction of the phylogenetic tree
- ✦ While small data sets can be easily analyzed “by hand”, using UPGMA, the problem quickly becomes onerous (but still of polynomial complexity) for large datasets (both in the number and in the dimension of the sequences to be analyzed)
- ✦ A trivial implementation to construct the UPGMA tree has $\mathcal{O}(n^3)$ (actually $\mathcal{O}(n^2 \log(n))$) time complexity

Just one exercise...

- Let the following table represent evolutionary distances among four species

	Human	Chimp	Bonobo	Gorilla
Human	0	12	12	14
Chimp	12	0	4	14
Bonobo	12	4	0	14
Gorilla	14	14	14	0

- Reconstruct the rooted phylogenetic tree based on UPGMA

Solution

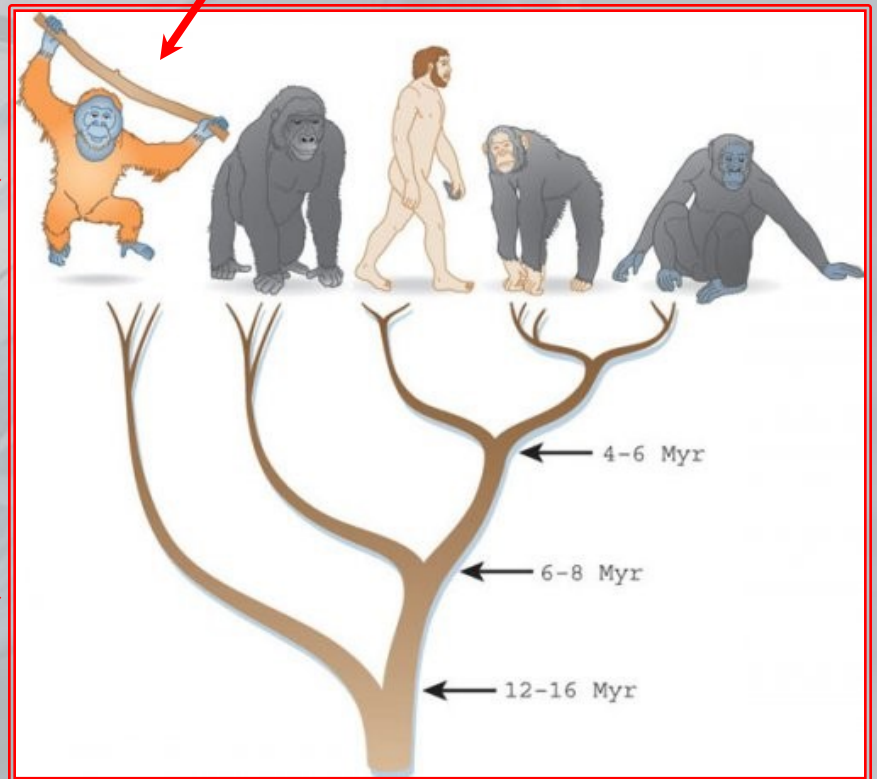
- ✦ We first consider the matrix

Species	H	C	B
C	12	–	–
B	12	4	–
G	14	14	14

and, after recalculation...

Species	H	CB
CB	12	–
G	14	14

Old World Monkeys
(Baboons, Macaques, etc.)



Arc length estimation – 1

- ✦ In addition to describing the evolutionary relationships among sequences, the phylogenetic tree topology can also provide information on their divergence degree
 - **Cladograms**, in which the arc length is proportional to the number of accumulated changes (or, using the molecular clock, to the speciation time)
 - ✗ The arc length is calculated based on the contents of the distance matrix
 - ✗ If we assume that the evolution rate is constant along all the lineages \Rightarrow the internal nodes are equidistant from each of the species to which they gave rise

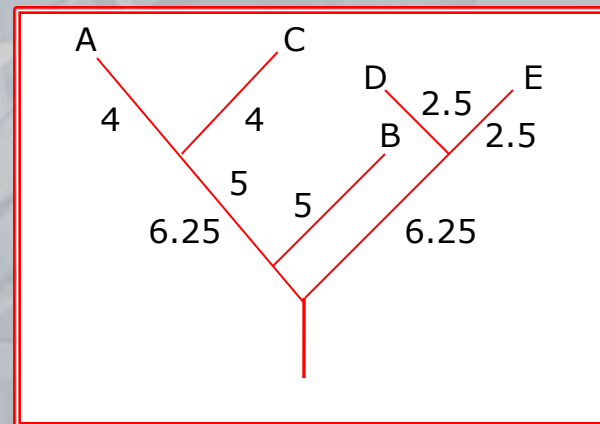
Arc length estimation – 2

✦ Example

Species	A	B	C	D
B	9	–	–	–
C	8	11	–	–
D	12	15	10	–
E	15	18	13	5

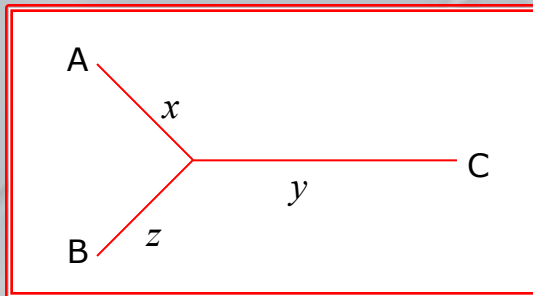
Species	A	B	C
B	9	–	–
C	8	11	–
DE	13,5	16,5	11,5

Species	B	AC
AC	10	–
DE	16,5	12,5



Arc length estimation – 3

- ✦ In scaled trees, the estimation of the arc length is difficult when the evolutionary speed cannot be assumed to be the same for all the lineages
- ✦ Let us consider the following unrooted tree:



$$d_{AC} = x + y$$

$$d_{AB} = x + z$$

$$d_{BC} = z + y$$

from which, with simple algebra, we can obtain:

$$x = (d_{AB} + d_{AC} - d_{BC})/2$$

$$y = (d_{AC} + d_{BC} - d_{AB})/2$$

$$z = (d_{AB} + d_{BC} - d_{AC})/2$$

Arc length estimation – 4

- ✦ The arc lengths of more complicated trees, which have more than one branch point, can be anyway estimated considering only three branches at a time
- ✦ The branches to be considered are:
 - the branches that connect the two closest phylogenetic species according to the distance matrix
 - the branch that connects the common ancestor of this two species to the common ancestor of all the other species
 - This procedure must be recursively applied until all the arc lengths are determined

Transformed distance method – 1

- ✦ The strength of the distance–matrix based approaches is that they work equally well with molecular or morphological data or, with a combination of both, having selected an appropriate metric
- ✦ Conversely, the weakness of UPGMA lies in the assumption of a constant rate of evolution along all lineages
 - Changes in substitution frequencies can cause the construction of topologically incorrect trees

Transformed distance method – 2

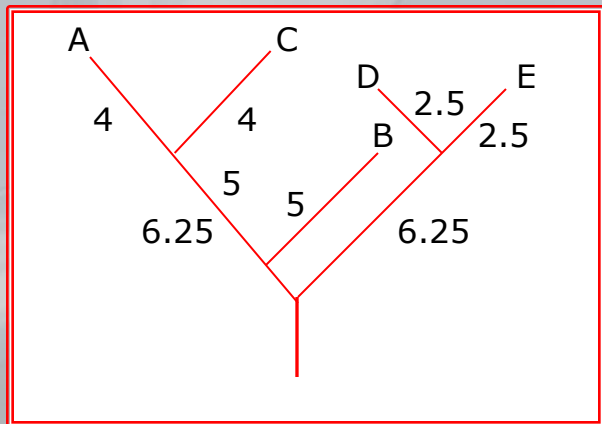
✦ **Example:** For the following distance matrices

Species	A	B	C	D
B	9	–	–	–
C	8	11	–	–
D	12	15	10	–
E	15	18	13	5

Species	A	B	C
B	9	–	–
C	8	11	–
DE	13,5	16,5	11,5

Species	B	AC
AC	10	–
DE	16,5	12,5

an indication that the rate of evolution is not constant is given by the lengths of the arcs in the cladogram, which are not additive



$d_{AE} = 4 + 6.25 + 6.25 + 2.5 = 19$
whereas, in the distance matrix,

$$d_{AE} = 15$$

Transformed distance method – 3

- ✦ Some alternative approaches to UPGMA based on distance matrices consider the possibility of different evolutionary rates in different lineages
- ✦ The **transformed distance method**, proposed by J. Farris in 1997, is based on the introduction of an **outer group**, a species that has undergone divergence from the common ancestor before all the other species represented in the matrix (also called **internal groups**)
- ✦ Hp.: This distance gives a good indicator of the relative location of sequences within a phylogenetic tree

Transformed distance method – 4

✦ **Example:** By considering the following sequences

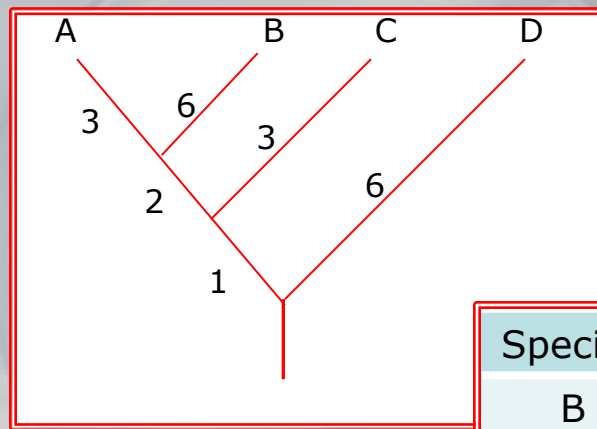
A: GTGCTGCACG GCTGAGTATA GCATTTACCC TTCCATCTTC AGATCCTGAA

B: ACGCTGCACG GCTCAGTGTG GTGTTTACCC TCCCATCTTC AGATCCTGAA

C: GTGCTGCACG GCTCGGCGCA GCATTTACCC TCCCATCTTC AGATCCTATC

D: GTATCACACG ACTCAGCGCA GCATTTGCCC TCCCGTCTTC AGATCCTAAA

we assume that the species D is an outer group compared to the species A, B and C, and that the true relationships among the species are represented by $((A,B),C),D)$ in the Newick format or by the phylogenetic tree...



The labels on the arcs correspond to the number of mutations, in the sequences, that have been accumulated along each lineage during each evolution stage

Species	A	B	C
B	9	–	–
C	8	11	–
D	12	15	10

Transformed distance method – 5

✦ Example (cont.)

In this situation, the external group D can thus be used as reference to transform the distances, by the following equation (Klotz *et al.*, 1979):

$$(d_{ij})' = (d_{ij} - d_{iD} - d_{jD})/2 + \overline{d_D}$$

where $(d_{ij})'$ is the transformend distance between the species i and j , and $\overline{d_D}$ is the average distance between the outer group and all the other internal groups (equal to $37/3$, in this case)

- The additive term that provides the average distance from the external group was introduced to ensure the positivity of the transformed distance (negative values do not make sense in an evolutionary perspective)

Transformed distance method – 6

✦ Example (cont.)

Consequently, we can calculate the transformed distance matrix for the species A, B and C

Specie	A	B
B	10/3	–
C	16/3	16/3

The classic approach UPGMA can then be used with the new matrix and produces the phylogenetic tree with the expected topology

Transformed distance method – 7

- ✦ The power of the described approach stems from a simple observation: the internal groups evolve separately only after their divergence and any difference in the number of accumulated substitutions must have occurred after speciation
 - ➡ External groups provide an objective reference system for comparing the substitution frequencies
- ✦ The transformed distance method creates trees with ultrametric distances (that is with all lineages that have diverged by equal amounts)
- ✦ Moreover, it should be noted that the transformed distance method only gives a (correct) tree topology and does not provide estimates of branch lengths (Nei, 1987)

Transformed distance method – 8

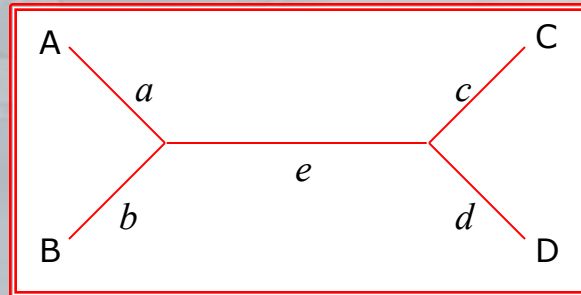
- ✦ The transformed distance method can also be applied when it is not possible to determine an external group
 - Even an internal group can act as a reference for the recalculation of the distances, but only outer groups allow correctly attaching a root to a phylogenetic tree
 - Solution with a two-stage approach:
 - Infer the root of the tree by the UPGMA method
 - After that, the taxa on one side of the root are used as references (outgroups) for making corrections for the unequal rates of evolution among the lineages on the other side of the root, and vice versa

Proximity relation methods – 1

- ✦ A (not too) different variant of the UPGMA method emphasizes the coupling of the species so as to construct trees with overall minimum arc lengths
- ✦ In an unrooted tree, the pairs of species that are separated by only one internal node are said to be **neighboring**
- ✦ From the topology of the tree, useful algebraic relations between neighbors can normally be obtained

Proximity relation methods – 2

✦ Example



- For a tree with additive arc lengths, it holds that:

$$d_{AC} + d_{BD} = d_{AD} + d_{BC} = a + b + c + d + 2e = d_{AB} + d_{CD} + 2e$$

where a , b , c and d are the lengths of the terminal branches, whereas e represents the length of the central branch

- The following conditions, known as the **four-points conditions**, hold

$$d_{AB} + d_{CD} < d_{AC} + d_{BD}$$

$$d_{AB} + d_{CD} < d_{AD} + d_{BC}$$

Proximity relation methods – 3

- ✦ We have to determine, among all the possible pair arrangements of the four species, those that satisfy the four-points conditions and then proceed to the grouping of the related elements
 - So far, it has been assumed that trees are additive: the method is not particularly sensitive to the deviation from this assumption that, if not checked, may anyway cause the construction of a topologically incorrect tree

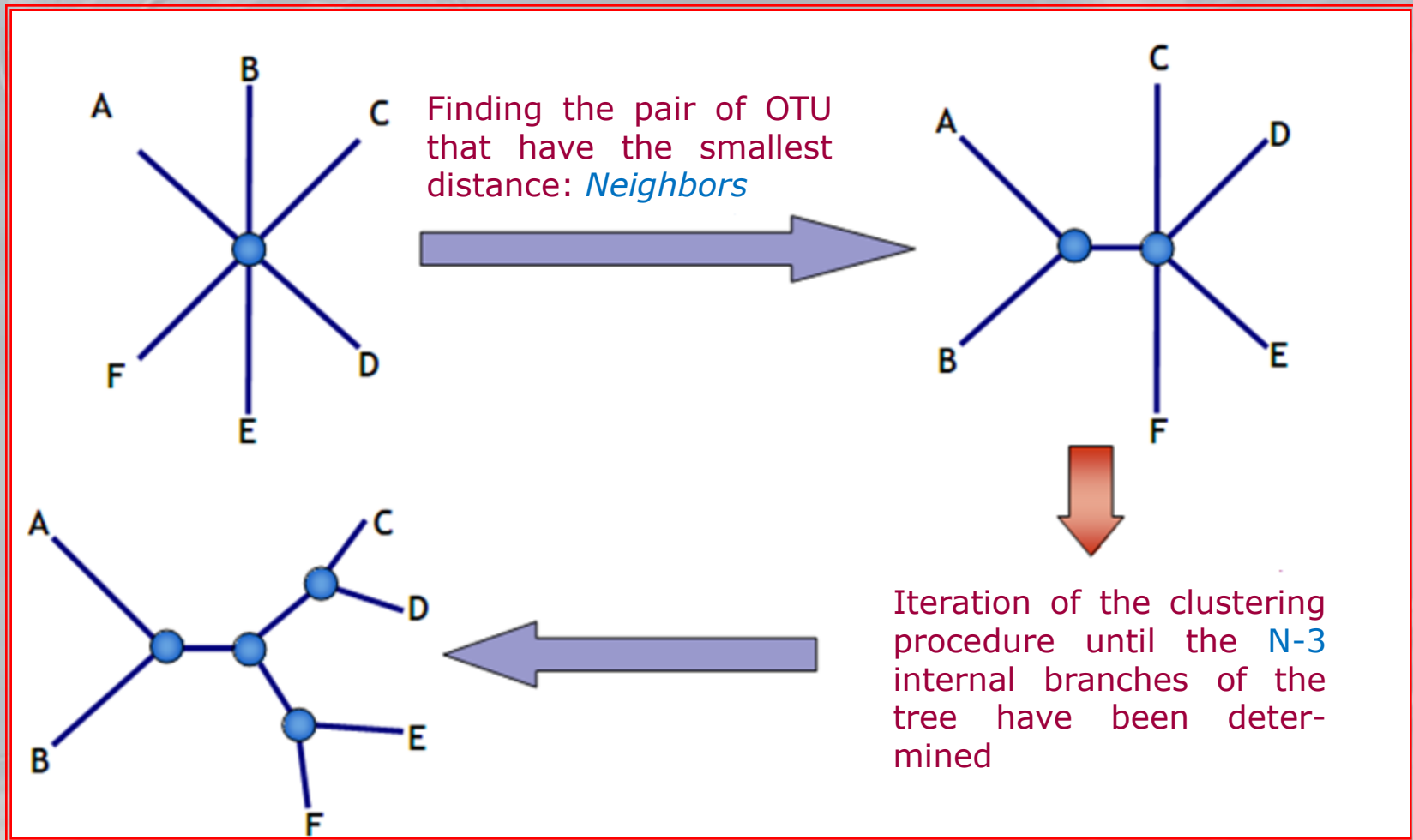
Proximity relation methods – 4

- ✦ In 1977, S. Sattah and A. Tversky suggested a way to apply the proximity approach to phylogenetic trees constituted by more than four species
 - 1) A distance matrix must be generated
 - 2) Based on its entries, we define for four species
 - ✗ $d_{AB} + d_{CD}, d_{AC} + d_{BD}, d_{AD} + d_{BC}$
 - 3) A score equal to 1 is assigned to the two neighboring couples that produce the minimum sum; on the contrary, 0 is assigned to the others
 - 4) The procedure is repeated with respect to all sets of four species that can be formed from the initial data
 - 5) At the end of the analysis, the pair of species with the highest score is grouped
 - 6) The distance matrix must be recalculated and the process is repeated from step 2) until there are only three species and the topology of the tree is uniquely determined
- ✦ Computationally burdensome for more than five or six species!

Neighbor-joining methods – 1

- ✦ There are many other possible approaches based on proximity, including different variants called **neighbor-joining** methods
 - We start with the creation of a star tree, where all the species, regardless of their number, descend from a single central node
 - The neighbors that minimize the total length of the branches of the tree are exhaustively searched
 - The main difference among the various neighbor-joining methods consists in the way in which the sum of the arc lengths is determined at each iteration
 - ➡ Neighbor-joining methods produce unrooted trees, having the additive properties

Neighbor-joining methods – 2



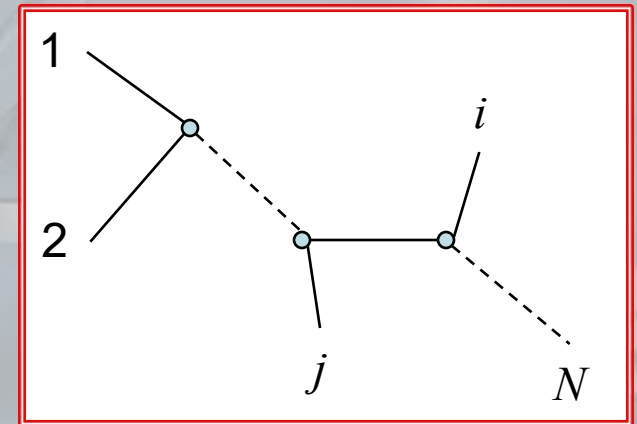
Neighbor-joining methods – 3

- ✦ N. Saitou and M. Nei (1987)

$$S_{12} = (1/(2(N-2))) \Sigma (d_{1k} + d_{2k}) + 1/2 d_{12} + (1/(N-2)) \Sigma d_{ij}$$
where each couple of species assumes the position 1 and 2 in the tree, N is the number of the species represented within the distance matrix, k is an outer group and d_{ij} is the distance between i and j

- ✦ J. Studier and K. Keppler (1988)

$$Q_{12} = (N-2)d_{12} - \Sigma d_{1i} - \Sigma d_{2i}$$



Neighbor-joining methods – 4

- ✦ **Algorithm:** Neighbor-joining takes as input a distance matrix specifying the distance between each pair of taxa
 1. Based on the current distance matrix calculate S or Q
 2. Find the pair of taxa for which it assumes its lowest value; add a new node to the tree, joining these taxa to the rest of the tree (and discard the original nodes – this pruning process converts the newly added common ancestor into a terminal node on a tree of reduced size)
 3. Calculate the distance from each taxon to the new node
 4. Start the algorithm again using the distances calculated in the previous step

Neighbor-joining methods – 5

- ✦ In each iteration of the procedure all the possible pairs of species are considered and the pair that produces a tree with the minimum value of the total length of the arcs (S o Q) is grouped, and then a new distance matrix is generated
- ✦ It has been proved that the two definitions for S and Q are theoretically equivalent, as well as the neighbor-joining and the proximity methods, since both depend on the four-points conditions and on the additivity assumption
 - ✦ They generate trees with very similar, if not identical, topologies

Multiple alignments – 1

- ✦ Sequence alignments are simpler for similar sequences, within which a few indel events have been occurred
- ✦ A multiple alignment of more than two sequences is a natural extension of pairwise alignments
 - The order in which the sequences are added to a multiple alignment can significantly change the final result
- ✦ Given that similar sequences can be aligned very easily and with a greater confidence, multiple alignments must consider the phylogenetic relations among the sequences

Multiple alignments – 2

- ✦ If the phylogenetic origin of the sequences is known, before the alignment is accomplished, the sequences can be added one at a time in this order
 - First, the sequences most closely related and then the sequences that are far from the evolutionary point of view
- ✦ However, multiple alignments are often used just to determine phylogenetic relationships among sequences
 - ▶ We need an integrated approach that simultaneously generates the alignment and establishes the phylogeny
 - ▶ It requires many cycles of alignment and phylogenetic analysis, and can be very costly

Multiple alignments – 3

✦ Algorithm

- 1) Generation of a pairwise distance matrix, based on all the possible pairwise alignments between the considered sequences
- 2) Use of a statistical approach, such as UPGMA, to construct an initial tree
- 3) Progressively realign the sequences in the order established by the deduced tree
- 4) Building of a new tree from the pairwise distances obtained by the new multiple alignment
- 5) Repeat the process if the new tree is not equal to the previous one

Concluding...

- ✦ Defining the true relationships among a set of homologous sequences is a very difficult task, without using some automatic techniques
 - ▶ The number of possible phylogenetic trees is very high even for a relatively small number of sequences
- ✦ A wide variety of approaches exists designed to infer phylogenetic relationships among genes or species, using the information encoded in the nucleotide or amino acid sequences
- ✦ Distance-based approaches:
 - Restrict the field to a few plausible phylogenies (trees)
 - Consider the overall similarity among the available sequences and progressively assemble them (starting from the closest ones)