Character–based phylogenetic methods

“Earth, which has seemed so large, must now be seen in its smallness. We live in a closed system, absolutely dependent on Earth and on each other for our lives and those of succeeding generations. The many things that divide us are therefore of infinitely less importance than the interdependence and danger that unite us.” (C. R. Darwin)

Table of contents

- Parsimony and phylogenetics
- Ancestral deduced sequences
- Quick search strategies
- Consensus trees
- Confidence of a tree
- Comparisons among phylogenetic methods
- Molecular phylogenies

Introduction

- Phylogenetic analysis has the aim of tracing the evolutionary relationships among different entities, called taxonomic units, mostly represented by sequences of nucleic acids, then reconstructing their evolutionary history ⇨ phylogenetic inference
- From the genetic point of view, evolution just consists in accumulating mutations: thus, it is possible to reconstruct evolutionary relationships among nucleic acids simply on the basis of the degree of similarity/ diversity of representing sequences
- The ultimate goal of phylogenetic analysis is that to construct a phylogenetic tree, able to describe the most probable evolutionary path followed by the species actually living on the Earth

Parsimony

- The concept of parsimony (from the Latin word ***parcere***, to save) is central for the character–based methods for phylogenetic reconstruction
- In the biological sense, the term is used to describe the process that leads to prefer a particular evolutionary path based on the lowest number of predicted mutational events
- The two premises that underlie the concept of biological parsimony can be summarized as:
  - mutations are extremely rare events
  - a model that postulates unlikely events, is probably incorrect
  - ➡ Those relationships that require the fewest number of mutations to explain the current status of the considered sequences are the most likely correct

Parsimony, why? – 1

- Philosophical principle enunciated in the fourteenth century by William of Ockham: among different explanations, the simplest is preferable; it looks needless to resort to many assumptions if the same event can be explained by few hypotheses

  Entia non sunt multiplicanda praeter necessitatem

  - God created all things, and God would not have created anything complex if he could reach the same goal in a simpler way
- Ockham's razor: It represents the basic principle of the modern scientific thought; in its most immediate form suggests the futility of formulating more theories of those that are strictly necessary to explain a given phenomenon

Parsimony, why? – 2

- Natural selection favors rapid adaptation, that is obtained through the least possible number of evolutionary steps
- Statistically speaking, evolutionary changes are rare, so it is unlikely that they occur many times
- Maximum parsimony is an optimality criterion under which the phylogenetic tree that minimizes the total number of character–state changes must be preferred

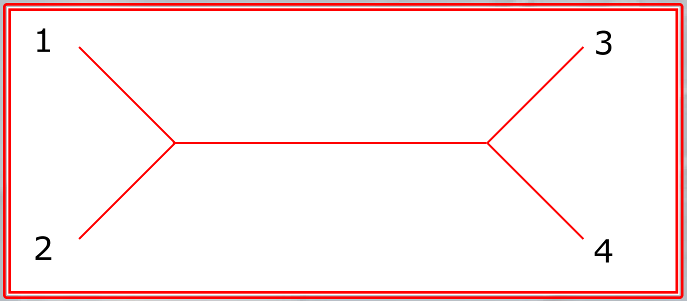
Informative and non-informative sites – 1

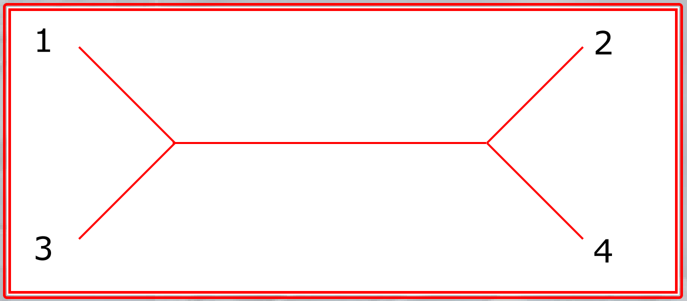
- There is an important distinction between informative and non–informative sites
- Which sites within a multiple alignment have an useful information content for a parsimonious approach?
- Example 1 (to be continued)

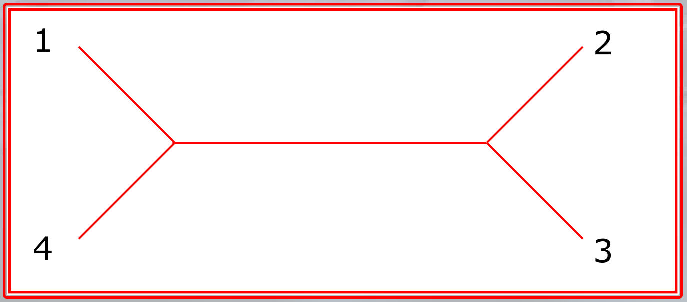
Sequences	a	b	c	d	e	f
1	G	G	G	G	G	G
2	G	G	G	A	G	T
3	G	G	A	T	A	G
4	G	A	T	C	A	T

Informative and non-informative sites – 2

- Example 1 (to be continued)
  - The relationship among four sequences may be described through three different unrooted trees ($N_U = (2s-5)!/[2^{s-3}(s-3)!]$); the informative sites are those that allow to distinguish one out of the three trees based on the number of postulated mutations

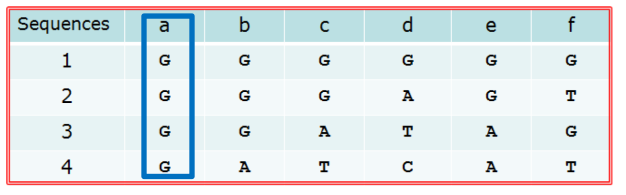




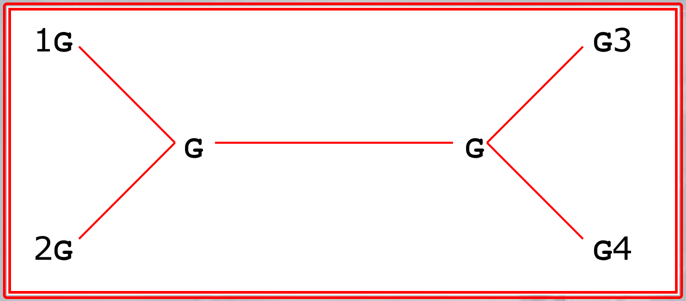


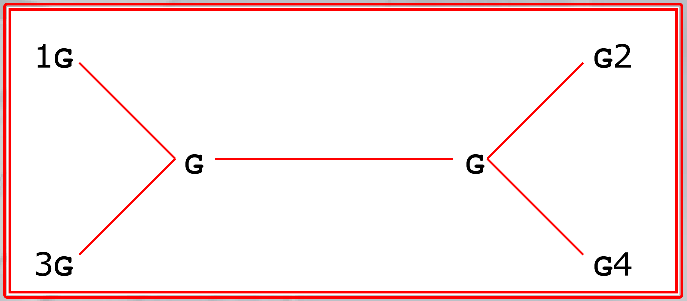
Informative and non-informative sites – 3

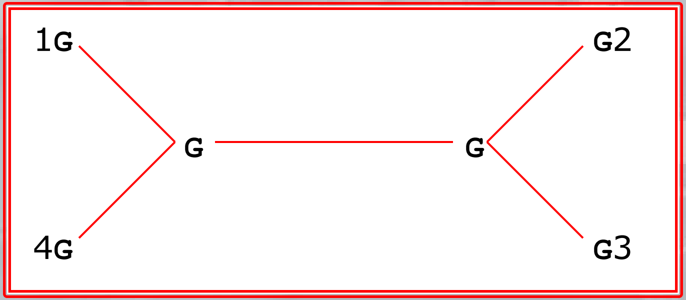
- Example 1 (to be continued)



Sequences	a	b	c	d	e	f
1	G	G	G	G	G	G
2	G	G	G	A	G	T
3	G	G	A	T	A	G
4	G	A	T	C	A	T







0 mutations

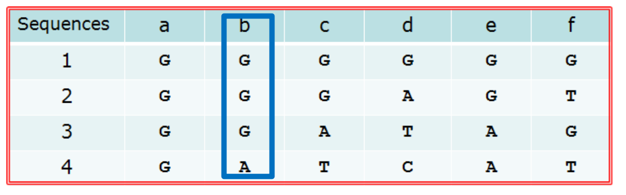
In the first position of the alignment, all four sequences share the same character (G) and the position is said to be invariant

Informative and non-informative sites – 4

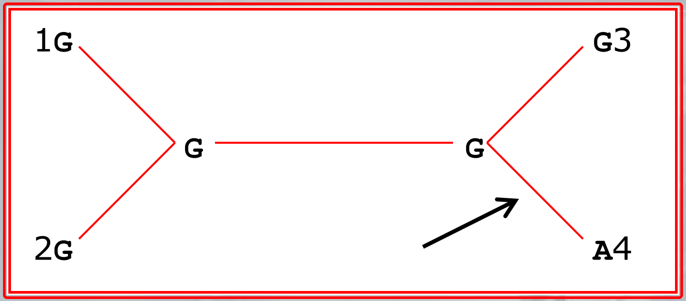
- Example 1 (to be continued)
  - The invariant sites are obviously non–informative sites, because each of the three possible trees that describe the relationship among the four sequences postulates exactly the same number of mutations (0)
  - Similarly, position b is non–informative from a parsimony point of view, since a mutation occurs in each of the three possible trees

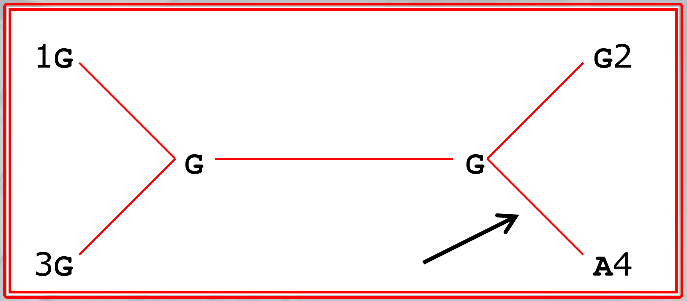
Informative and non-informative sites – 5

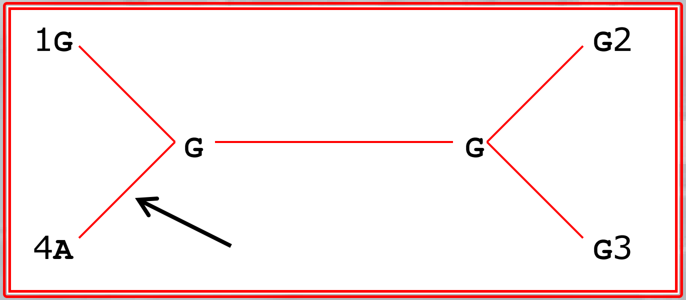
- Example 1 (to be continued)



Sequences	a	b	c	d	e	f
1	G	G	G	G	G	G
2	G	G	G	A	G	T
3	G	G	A	T	A	G
4	G	A	T	C	A	T



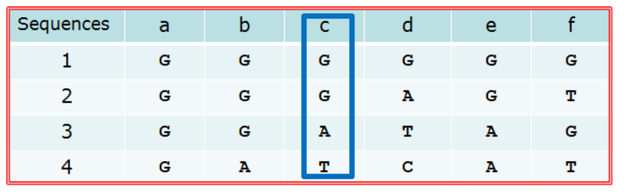




1 mutation

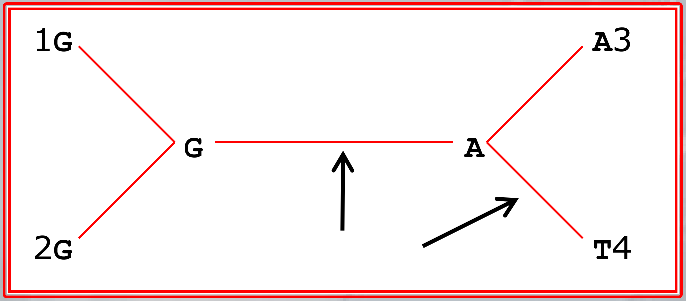
Informative and non-informative sites – 6

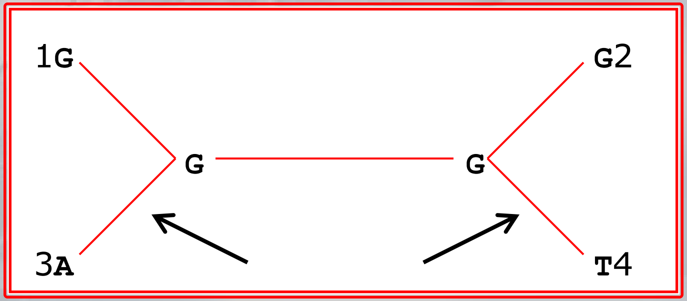
- Example 1 (to be continued)

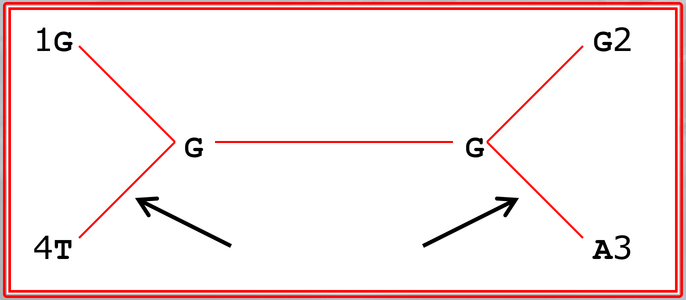


Sequences	a	b	c	d	e	f
1	G	G	G	G	G	G
2	G	G	G	A	G	T
3	G	G	A	T	A	G
4	G	A	T	C	A	T

Also position c is non–informative because all the trees require two mutations



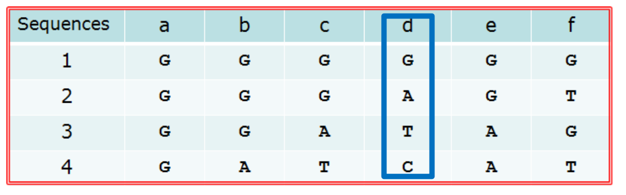




2 mutations

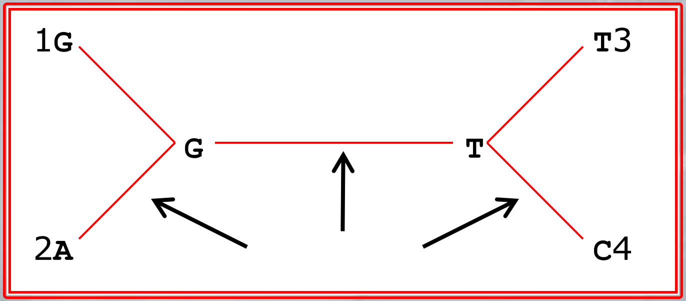
Informative and non-informative sites – 7

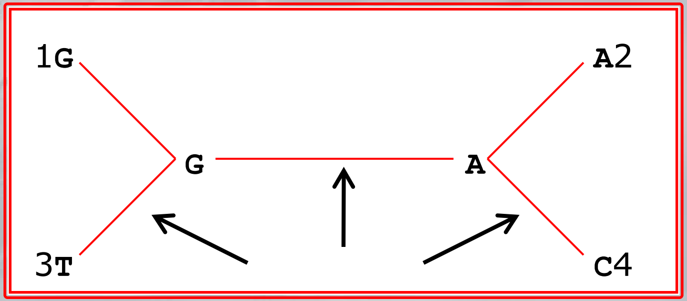
- Example 1 (to be continued)

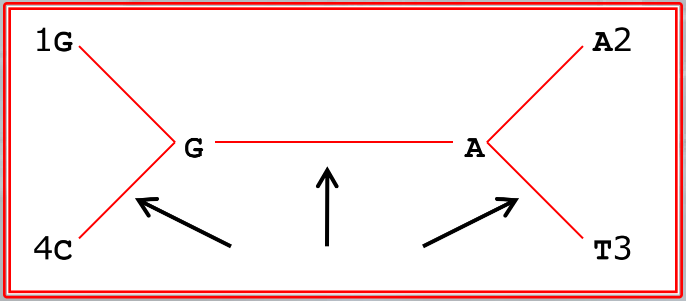


Sequences	a	b	c	d	e	f
1	G	G	G	G	G	G
2	G	G	G	A	G	T
3	G	G	A	T	A	G
4	G	A	T	C	A	T

…so as position d, in which all the trees postulate three mutations



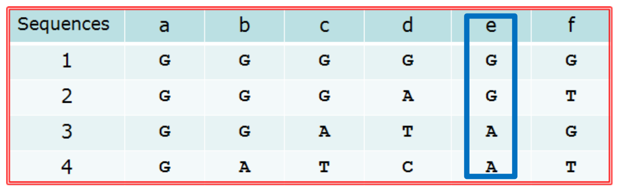




3 mutations

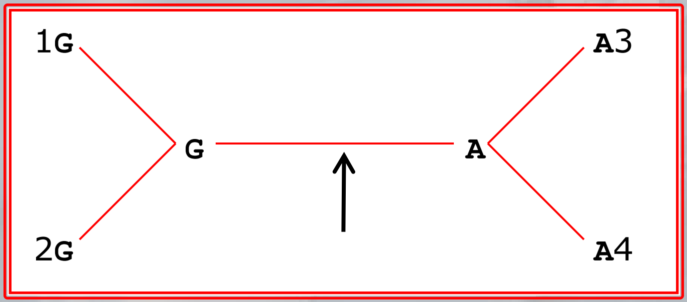
Informative and non-informative sites – 8

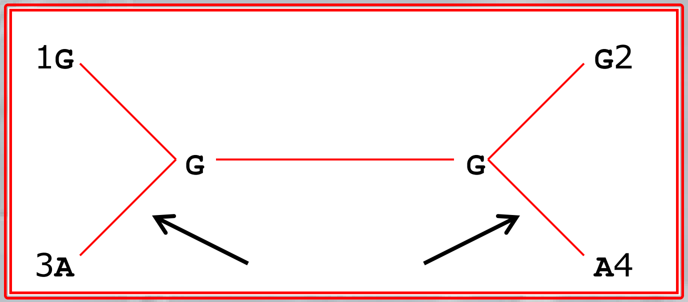
- Example 1 (to be continued)

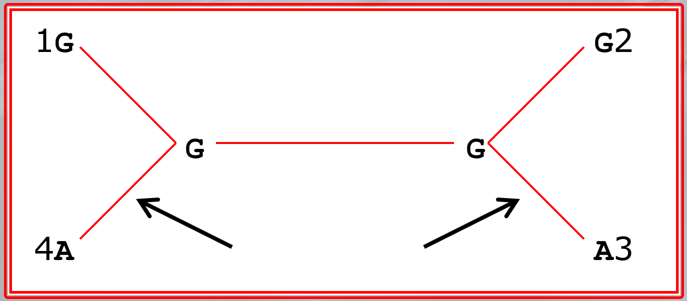


Sequences	a	b	c	d	e	f
1	G	G	G	G	G	G
2	G	G	G	A	G	T
3	G	G	A	T	A	G
4	G	A	T	C	A	T

In contrast, positions e and f are actually informative, because, in both cases, one of the trees postulates only one mutation, while the others require two mutations

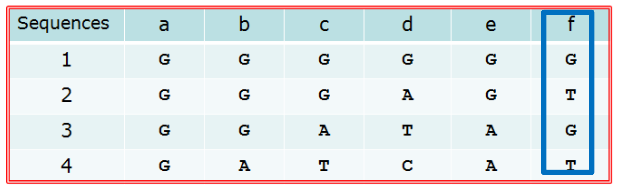




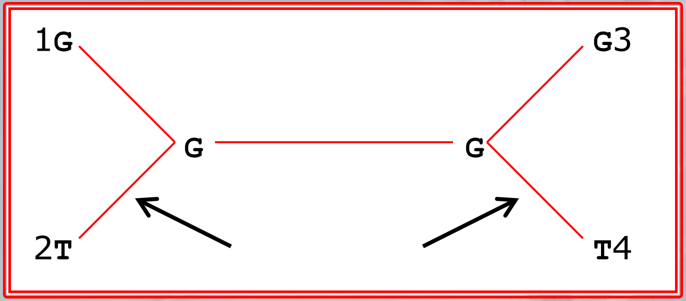


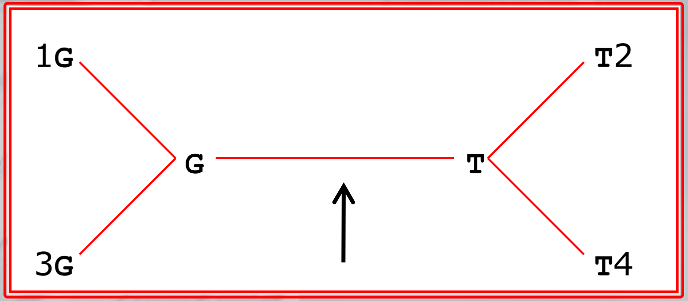
Informative and non-informative sites – 9

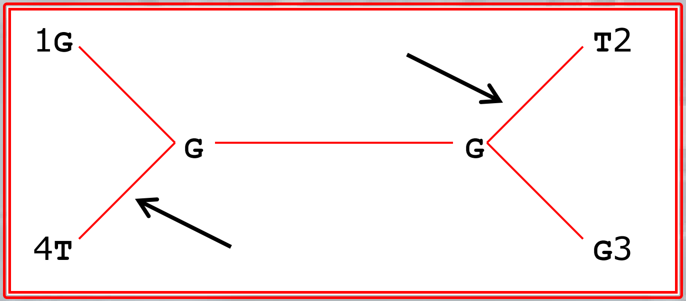
- Example 1



Sequences	a	b	c	d	e	f
1	G	G	G	G	G	G
2	G	G	G	A	G	T
3	G	G	A	T	A	G
4	G	A	T	C	A	T







Informative and non-informative sites – 10

- In general, in order for a position to be informative, regardless of how many sequences are aligned, it must contain at least two different nucleotides, each of which must be present at least twice
- The non–informative positions are simply discarded and not considered in the subsequent parsimony analysis
- Vice versa, non–informative positions do contribute to pairwise similarity scores used in distance–based approaches
  - Very different conclusions can be drawn depending on the chosen method (distance or character–based)

Unweighted parsimony – 1

- Once non–informative sites are identified and discarded, the parsimony approach can be implemented in its simplest form
  - For each informative site, we consider the three possible trees
  - For each tree, a score is maintained that keeps track of the minimum number of substitutions required for each position
  - After considering all the informative sites, the tree (or the trees) which postulates the fewest number of substitutions is, by definition, the most parsimonious
- Example 2: In an analysis involving only four sequences, each informative site favor only one of the three alternative trees, and the tree which is supported by the greatest number of informative sites is also the most parsimonious one

Unweighted parsimony – 2

- The evaluation of the alignments for five or more sequences is decidedly more complicated
  - The number of different unrooted trees grows exponentially with the number of sequences to be aligned
    - Even having identified a small number of informative sites, the approach “by hand” is inapplicable for more than ten sequences
  - The individual sites can support more than one alternative tree and the maximum parsimony tree does not necessarily coincide with that supported by the largest number of informative sites
  - Calculating the number of all postulated substitutions for each alternative tree is a hard problem just for only five sequences (15 trees)

Unweighted parsimony – 3

- Example 3 (to be continued)

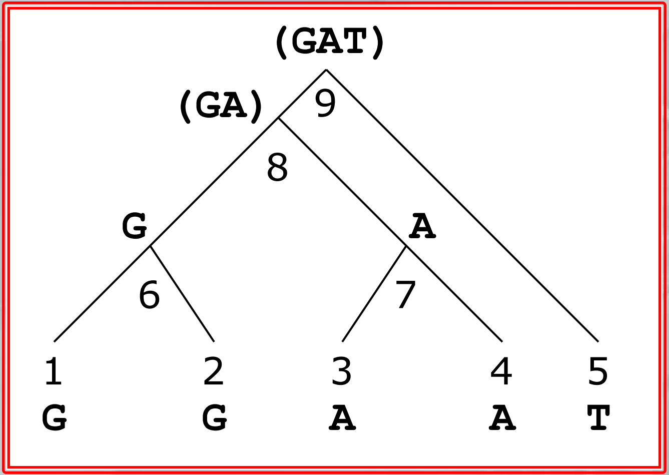
………G………

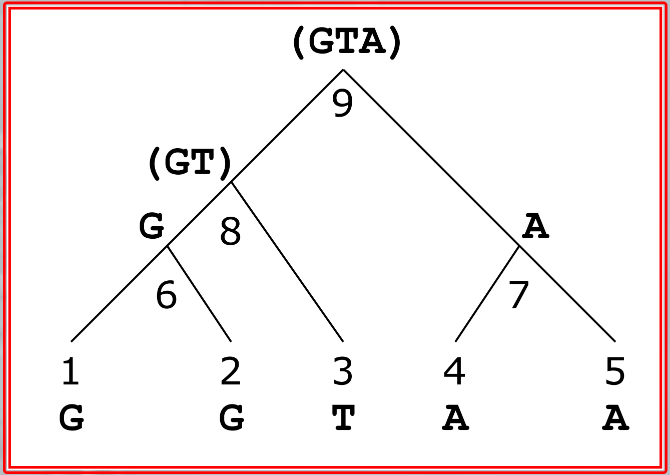
………G………

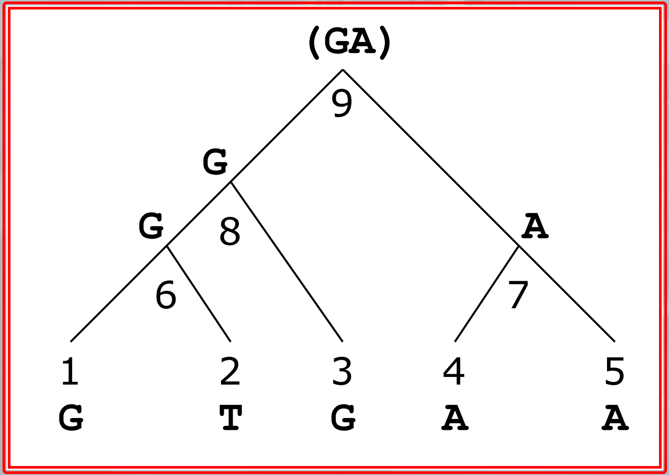
………A………

………A………

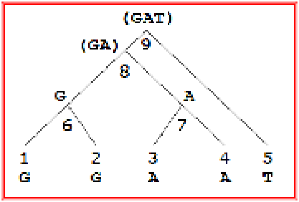
………T………

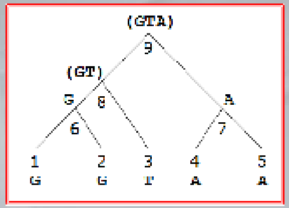






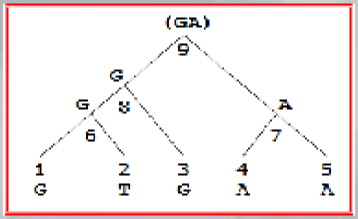
Unweighted parsimony – 4





- Example 3 (to be continued)
  - Determining the number of postulated substitutions for each tree requires to infer the most likely nucleotide in each of the four internal nodes from the nucleotides present in each of the five terminal nodes
    - The parsimony rule makes it easy to determine the nucleotide at position 6 (relative to the first two trees): the ancestral nucleotide must be a **G**, or a replacement should be happened along both the lineages leading to the terminal node 1 and 2
    - We can analogously justify the allocation of **A** in position 7
    - The nucleotide in the ancestral node 8, however, cannot be determined unambiguously, but based on the parsimony rule, it should be **A** or **G**, in the first tree, and **G** or **T**, in the second
    - At node 9, the triad **G**, **A**, **T** certainly contains the most parsimonious nucleotide

Unweighted parsimony – 5



- Example 3
  - Instead, for the last tree…
    - Nodes 1 and 2 suggest that the nucleotide in the ancestral node 6 is **G** or **T**
    - However, node 3 indicates **G** as the candidate nucleotide in node 8
      - By assigning **G** as the ancestral nucleotide to the nodes 6 and 8, for this portion of the tree, only one replacement must be postulated (along the lineage leading from node 6 to node 2)
    - All the other three alternatives (assigning a **T** to node 6, to node 8 or to both nodes 6 and 8) would require at least two substitutions

Unweighted parsimony – 6

- From a methodological point of view, the rule for assigning ancestral positions is the following:
  - The set of nucleotides that are the most probable candidates for an internal node is represented by the intersection of the two sets corresponding to its immediate descendant nodes, if the intersection is not empty
  - Otherwise, it is represented by the union of the sets corresponding to its descendant nodes
  - When a union is needed to form a set of nodes, a substitution has been occurred at a certain point of the evolutionary path that leads to that position
  - Thus, the number of unions represents also the minimum number of substitutions required to get the nucleotides at the terminal nodes, since they have shared a common ancestor

Unweighted parsimony – 7

- This method applies only to informative sites
- The minimum number of substitutions for a non–informative site is, instead, the number of different nucleotides present in the terminal nodes minus one
- Example 4: If the nucleotides present in a particular position in a five sequence alignment are **G**, **G**, **A**, **G**, **T**, then the minimum number of substitutions is 3–1=2, regardless of the tree topology
- The non–informative sites contribute with an equal number of replacements to all the alternative trees and are excluded from the parsimony analyses
- However, it is the total number of substitutions that defines the length of the tree and, to the length, also non–informative sites actually contribute

Weighted parsimony – 1

- Despite having established the general principle that “mutations are rare events”, inferring from this that all mutations are equivalent is an oversimplification (e.g., substitutions vs. indel events, indel length, transitions vs. transversions, etc.)
- If we could associate a value to the relative probability of different mutation events, these values would be translated into weights and used by parsimony algorithms
  - Difficulty in defining a single set of weights with universal validity or otherwise usable by many different sets of data, because...
    - some sequences (for example, non–coding sequences with tandem repeats) are more prone to indel events than others
    - the functional importance differs greatly from gene to gene and from species to species also for homologous genes
    - the predisposition to “soft” substitutions (e.g. **GC** with **AT**, or between codons that specify the same amino acid) usually varies from gene to gene and from species to species

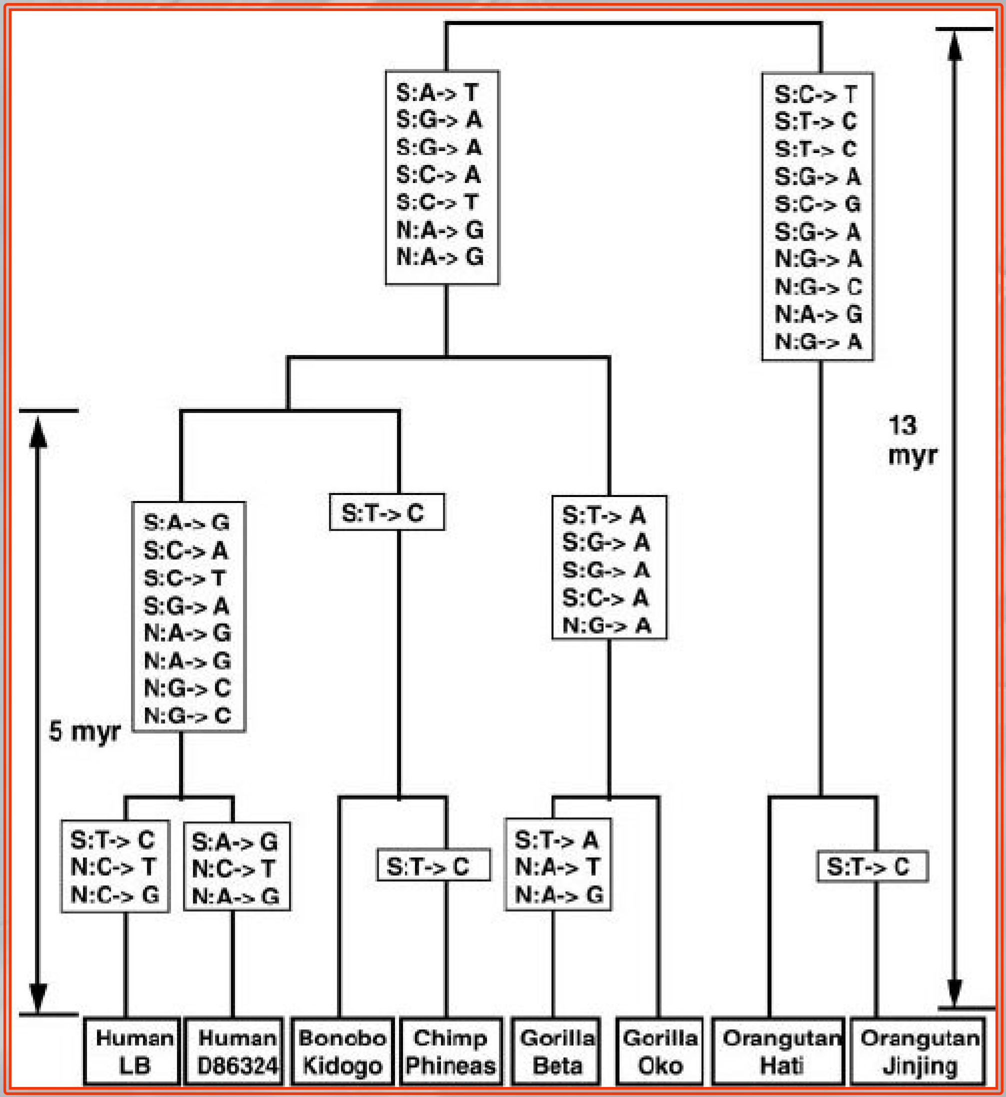
Weighted parsimony – 2

- The best choice for the weights is related to a particular set of empirical data
- Example 5: If, for a particular multiple alignment, comparisons between each single sequence and a consensus sequence indicate that the transitions are three times more common than transversions, then:
  - Values equal to 1 and 0.33 must be, respectively, associated to transversions and transitions
  - At the end of the analysis, the tree predicting the smallest number of the most common mutation events is the most parsimonious

Ancestral deduced sequences – 1

- A remarkable result of parsimony analysis is the deduction of ancestral sequences generated during the analysis itself
- In fact, if the process of deduction of a particular ancestral nucleotide may seem trivial, much less banal is its extension to a whole sequence representing a gene
- In particular, when the structure and the function of a set of homologous proteins are well known, the occurred amino acid substitutions may provide very interesting clues on the physiology of extremely ancient organisms and on the environment in which they lived
- Thanks to the deduced ancestors generated by parsimony analysis, the study of molecular evolution has no missing links and the intermediate states can be objectively inferred from the sequences of their living descendants

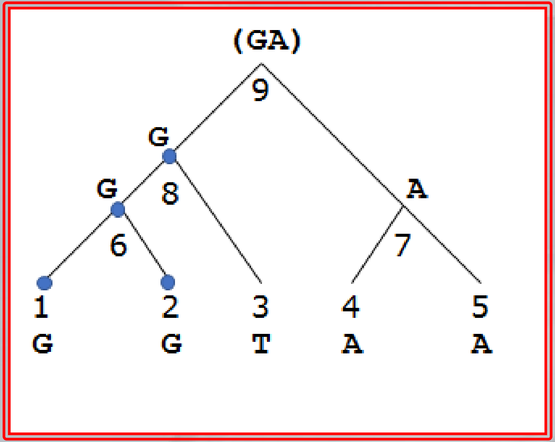
Ancestral deduced sequences – 2



Phylogenetic tree of hydroxylase sequences built by the maximum parsimony method.
Individual nucleotide substitutions on each branch (based on ancestral sequence deduced from rhesus monkey) are indicated with S or N, depending on whether the nucleotide substitution is a synonymous or nonsynonymous change, respectively (chronological order is not implied in the order of each listing). The rhesus monkey sequence (GenBank accession no. AB013814) was used as an outgroup to deduce an ancestral sequence of orangutan, gorillas, chimpanzees, and humans. A change that induced a new stop codon at the C terminus of the gorilla Beta sequence was observed and excluded from the analysis.

Ancestral deduced sequences – 3

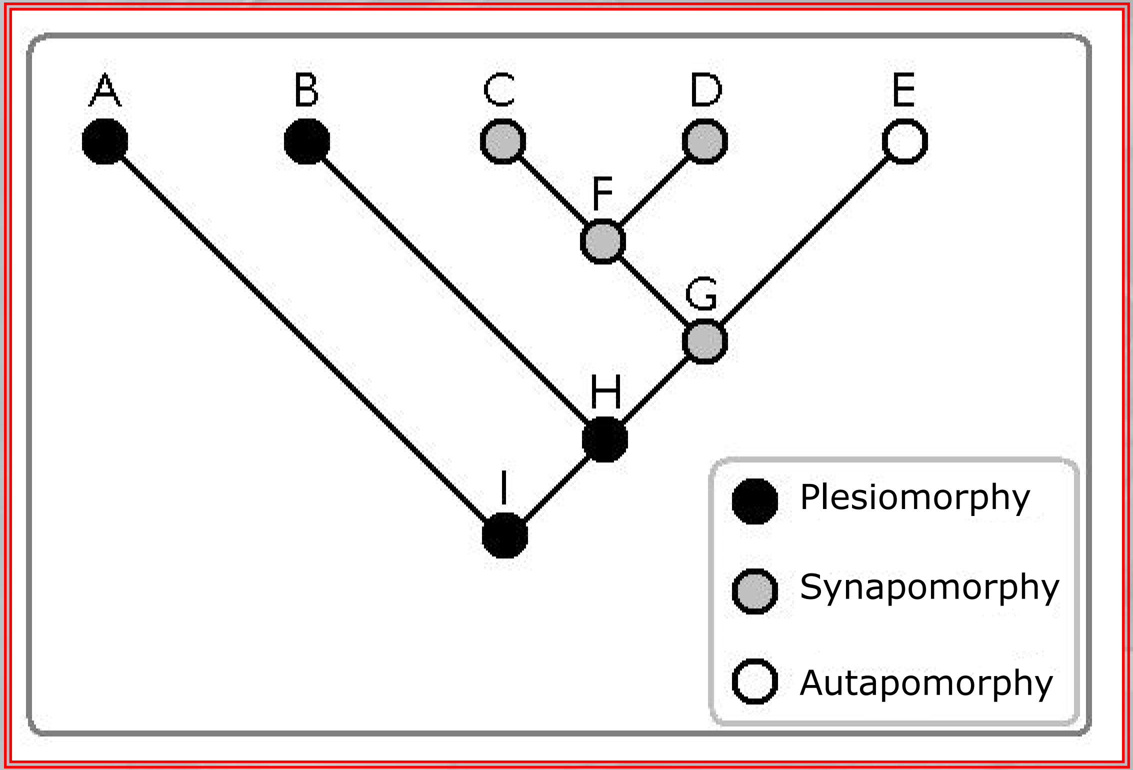
- The informative sites that support internal branches of the deduced tree are called synapomorphies



- The synapomorphy is a derived character, i.e., a new shared character, useful for reconstructing a phylogenetic tree
- Each hypothetical synapomorphy is subjected to a congruence test, that is, its pattern of distribution among various taxa is examined in comparison with other characters

- All the other informative sites are considered homoplasies (similar characters that appeared independently in different taxa rather than inherited from a common ancestor)

Ancestral deduced sequences – 4



Plesiomorphy – It describes the presence, in organisms belonging to different species, of an ancestral character that represents an innovative common evolution; for example, the spine is a plesiomorph character for the whole Vertebrata subphylum

Autapomorphy – It is a derived trait that is unique in each group; an autapomorph character is neither present in the closest relatives of the terminal group nor in the common ancestral progenitors

Quick search strategies

- The basic rules of parsimony remain the same both in the simplest case of an alignment involving only four sequences and for multiple alignments related to many sequences
- Anyway, using a standard parsimony approach, it quickly becomes impossible to perform even few alignments “by hand”, albeit containing a small number of informative sites
  - To analyze 10 sequences, more than 2 million trees must be considered and the exhaustive search becomes a prohibitive approach just for 12 sequences
  - Conversely, in real world applications, data to be processed are normally hundreds of times larger than that allowed by the above limitations
  - ➧ Efficient search algorithms

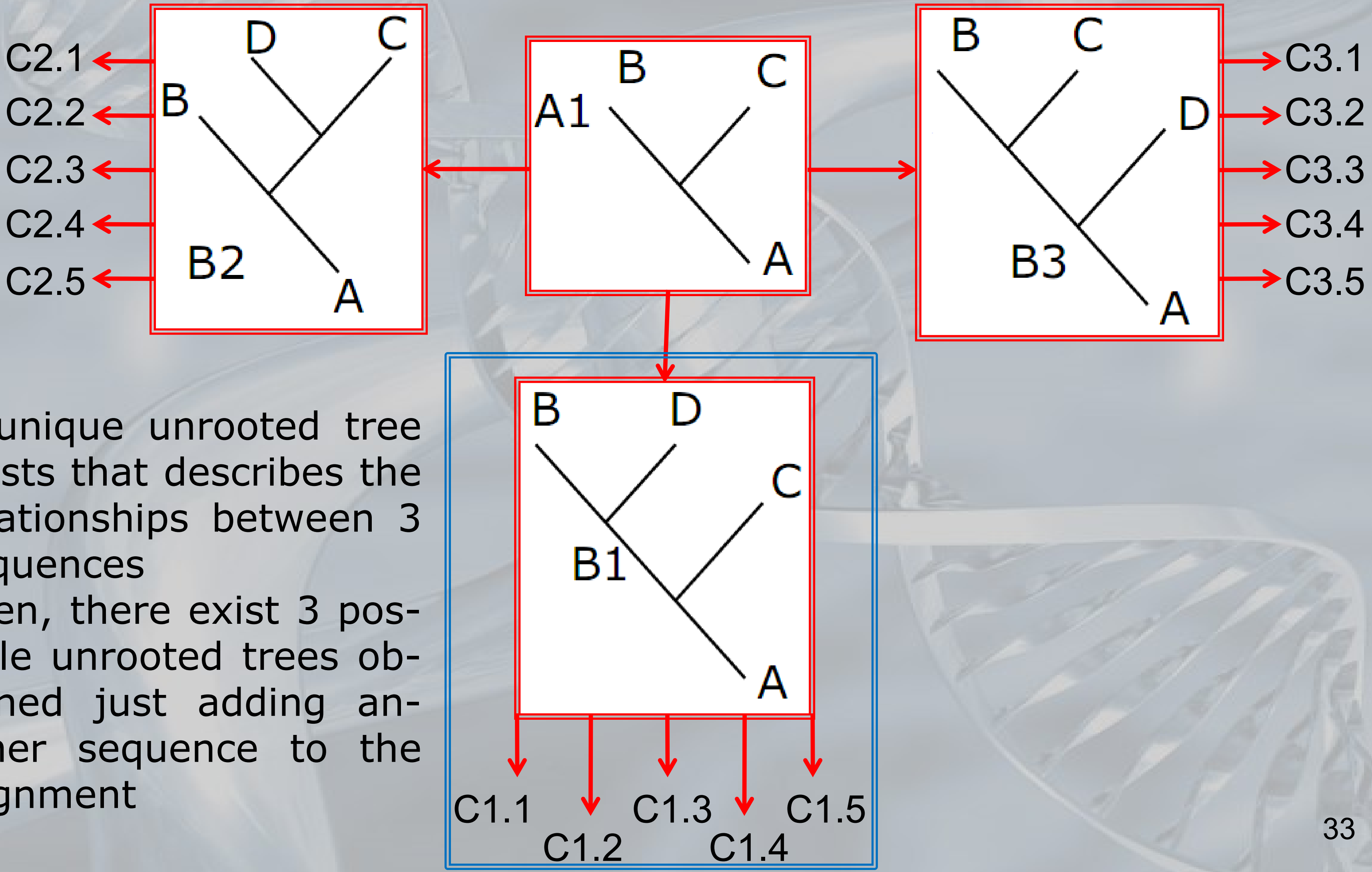
Branch and bound – 1

- The length of a tree, L, is defined as the sum of the minimum numbers of substitutions over all sites for the given topology
- Originally proposed by Hardy and Penny in 1982, the branch and bound method consists of two steps:
  1) Fixing an upper bound, L, for the most parsimonious tree length w.r.t. a certain set of data; L can be estimated…
     - …randomly choosing a tree that describes the relations among all the sequences to be analysed
     - …building a reasonable approximation of the most parsimonious tree (for example, by UPGMA)
  2) Construction of each tree, adding a branch at a time, to include all the sequences to be analysed, ending the procedure when the tree reaches the previously established length L

Branch and bound – 2

- What makes the method actually effective is the fact that each tree, consisting of a subset of the data, which requires more than L substitutions, must forcibly become longer with the addition of new sequences
  - It cannot be the most parsimonious tree
- If, during the analysis, we build trees with length smaller than L, L can be updated accordingly, making the method also more efficient

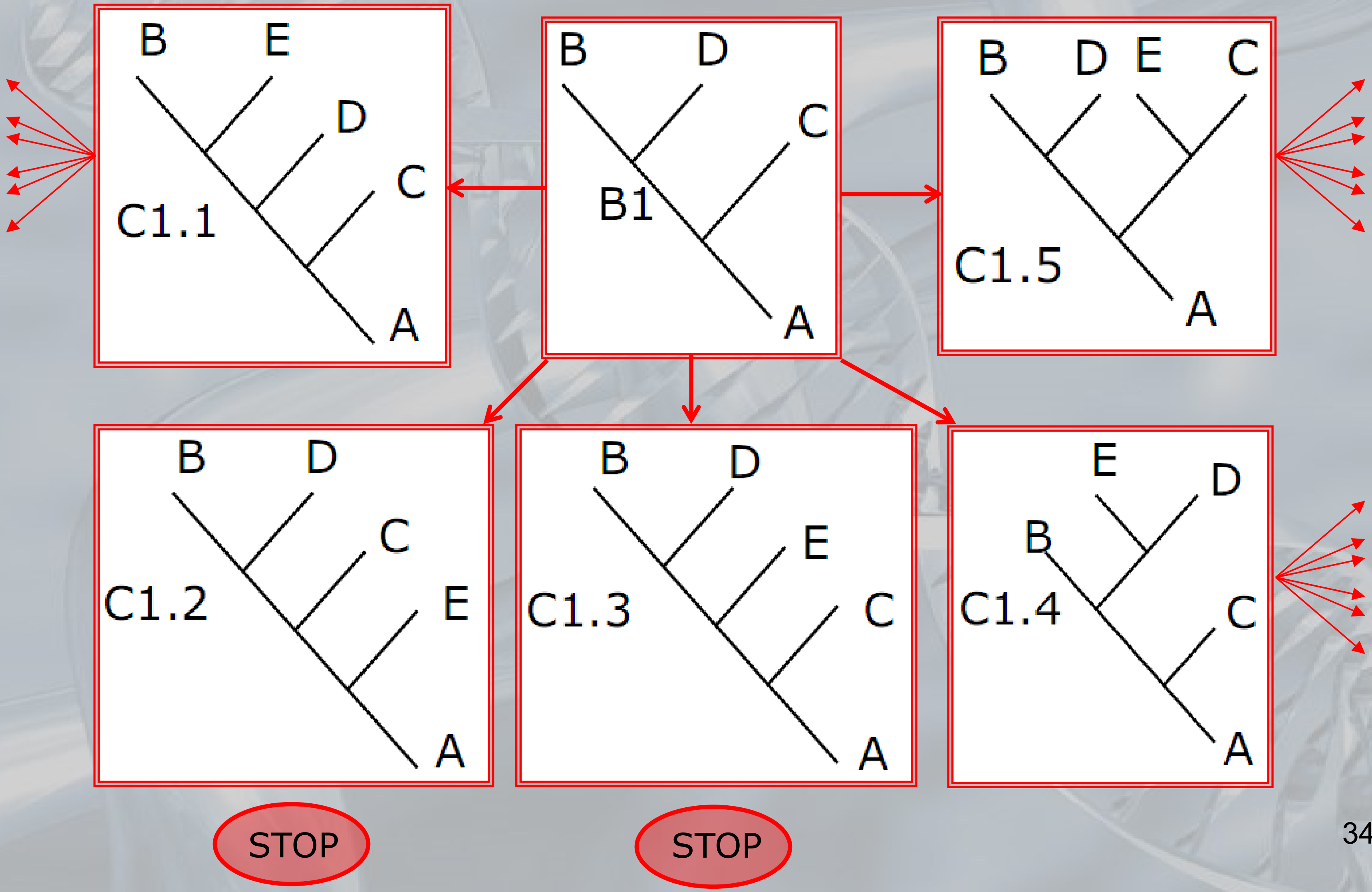
Branch and bound – 3



A unique unrooted tree exists that describes the relationships between 3 sequences

Then, there exist 3 possible unrooted trees obtained just adding another sequence to the alignment

Branch and bound – 4



Branch and bound – 5

- As the exhaustive search, the branch and bound method ensures that, at the end of the analysis, all the optimal trees – according to the maximum parsimony criterion – were found
- Branch and bound is several orders of magnitude faster than the exhaustive search
- However… it is useful for the alignment of at most twenty sequences, while it is computationally untenable for multiple alignments that involve the analysis of more than 10^{21} unrooted trees

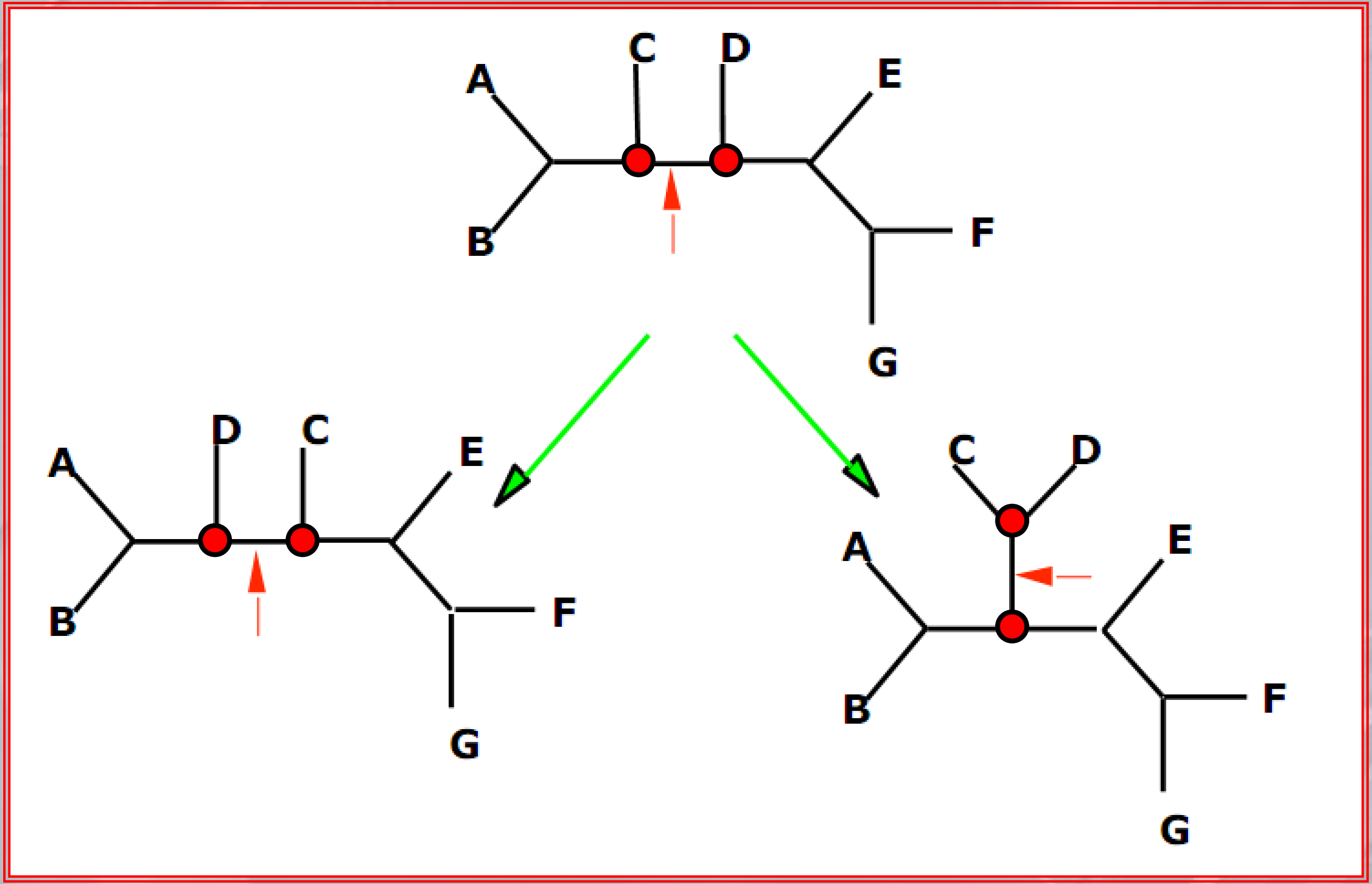
Heuristic search – 1

- The amount of sequence information is continuously increasing and it is quite common that multiple alignments involve more than twenty sequences
  - ➡ Necessity of using computationally less expensive algorithms that, anyway, cannot always guarantee the global optimum
- Assumptions underlying all heuristic methods:
  - The “alternative” trees are not independent each other
  - Since the most parsimonious trees should have very similar topologies to trees that are a little less thrifty, all heuristic searches begin with a tree building phase; such tree is used as a starting point for finding the shortest trees

Heuristic search – 2

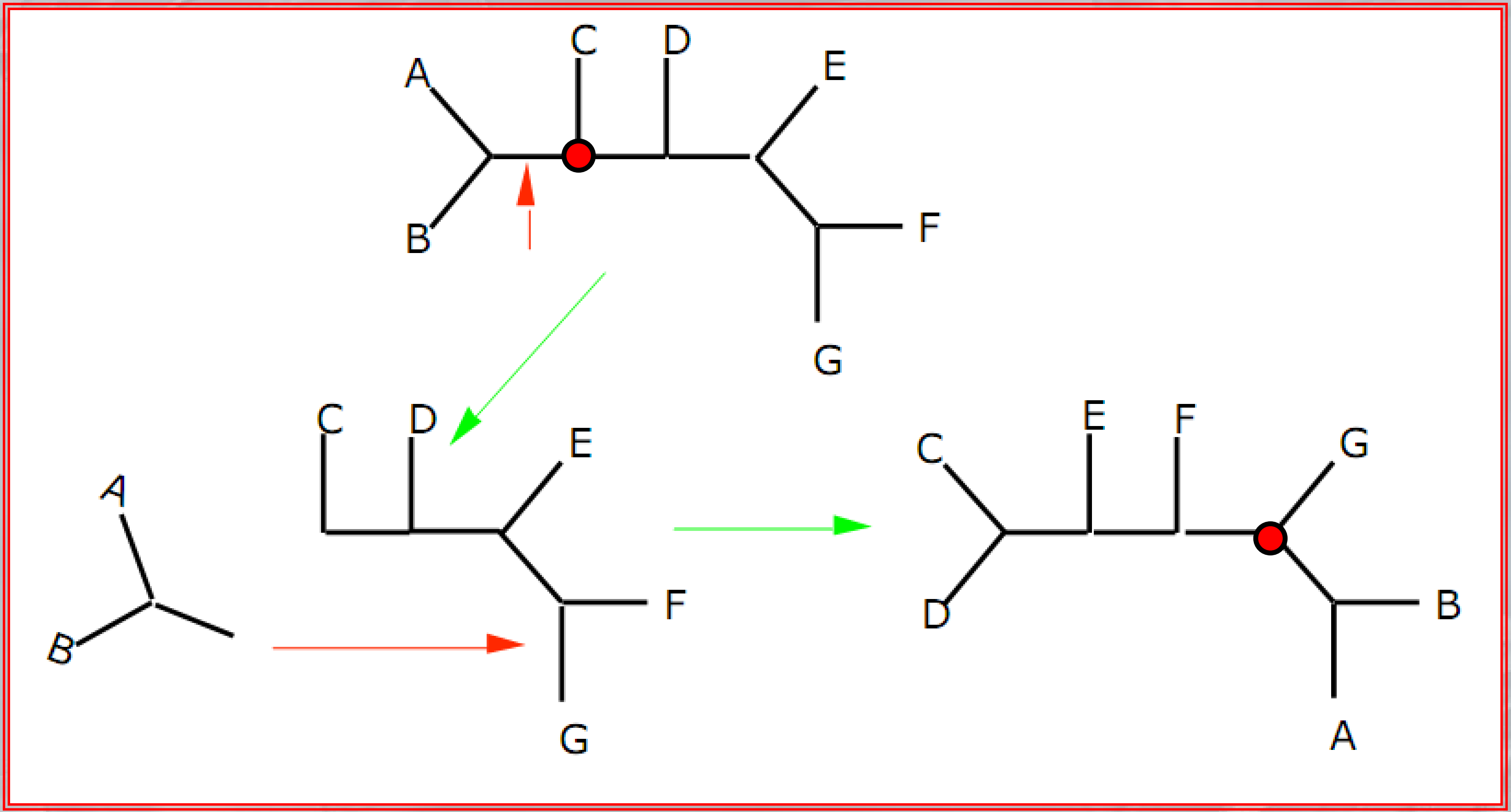
- Heuristic searches actually work well if the “initial” tree is a good approximation of the most parsimonious tree
- However, instead of building alternative trees branch by branch, the heuristic search generates complete trees, with topologies similar to that of the starting tree, performing exchanges in the tree branches and grafting them on other portions of the best tree found up to that point in the analysis
  - Nearest Neighbor Interchange
  - Subtree Pruning and Regrafting
  - Tree Bisection and Reconnection

Heuristic search – 3



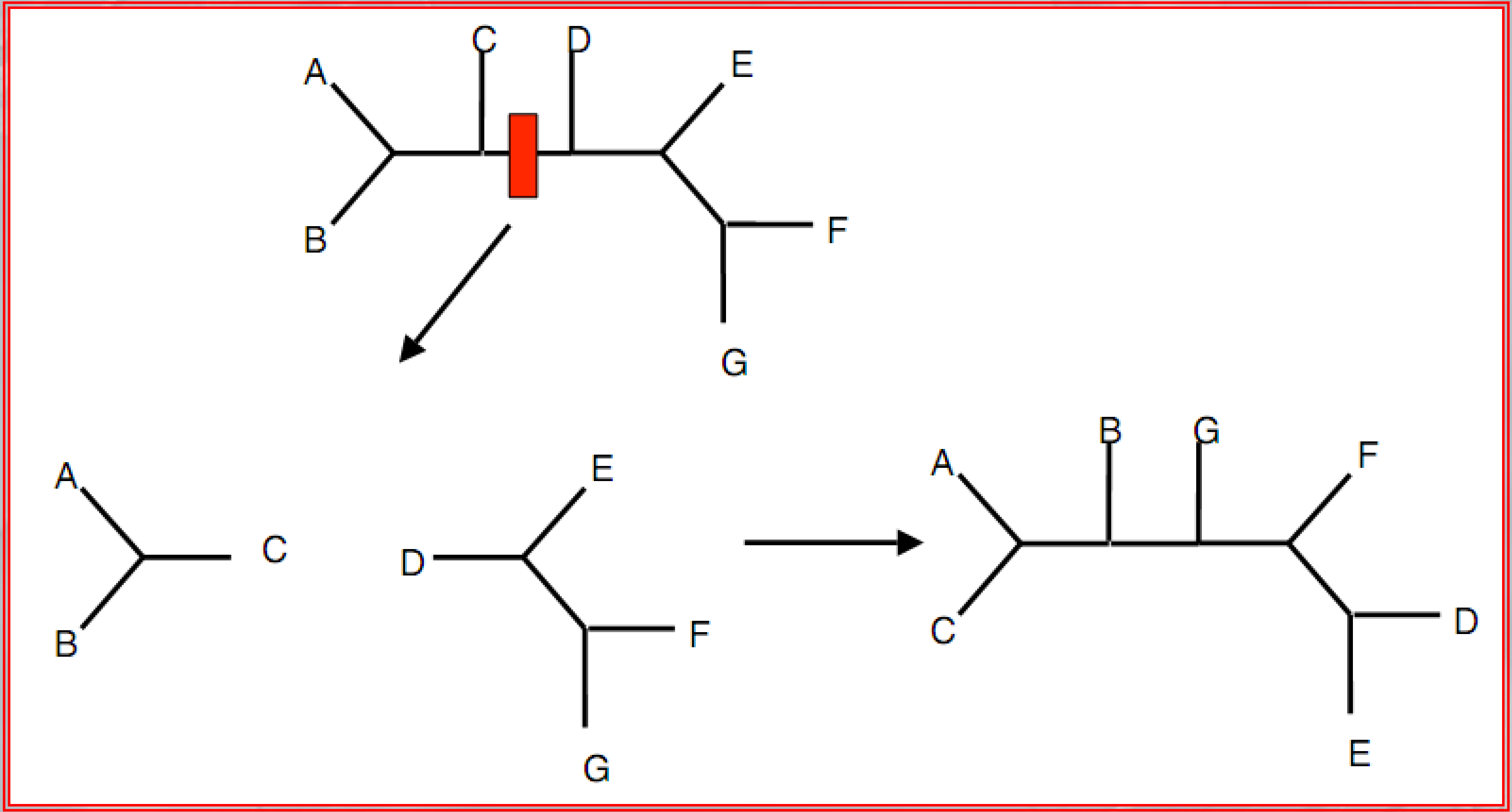
Nearest Neighbor Interchange

Heuristic search – 4



Subtree Pruning and Regrafting

Heuristic search – 5



Tree Bisection and Reconnection

Heuristic search – 6

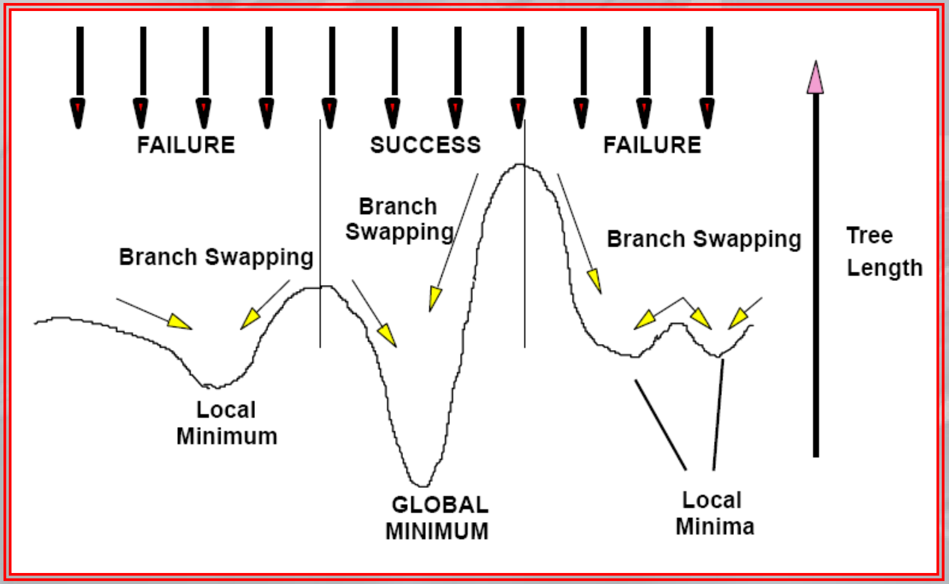
- In all the cases, a rearrangement is accepted if it produces a tree better (shorter) than the tree from which it is obtained
- The process is repeated until an exchange cycle fails to produce a tree that is equal to or shorter than the tree generated during the previous cycle of pruning and grafting

Heuristic search – 7

- The heuristic algorithms take into account the impossibility of examining all the enormous number of alternative unrooted trees obtained by complex multiple alignments, emphasizing the exchange of branches on trees more and more parsimonious
  - This process can give rise to the stall of the algorithm on topologies which do not necessarily exhibit the least number of substitutions
  - In other words, if the initial tree is far from the most parsimonious tree, it may not be possible to get to it without making an arrangement that, at first, increases the number of substitutions

Heuristic search – 8

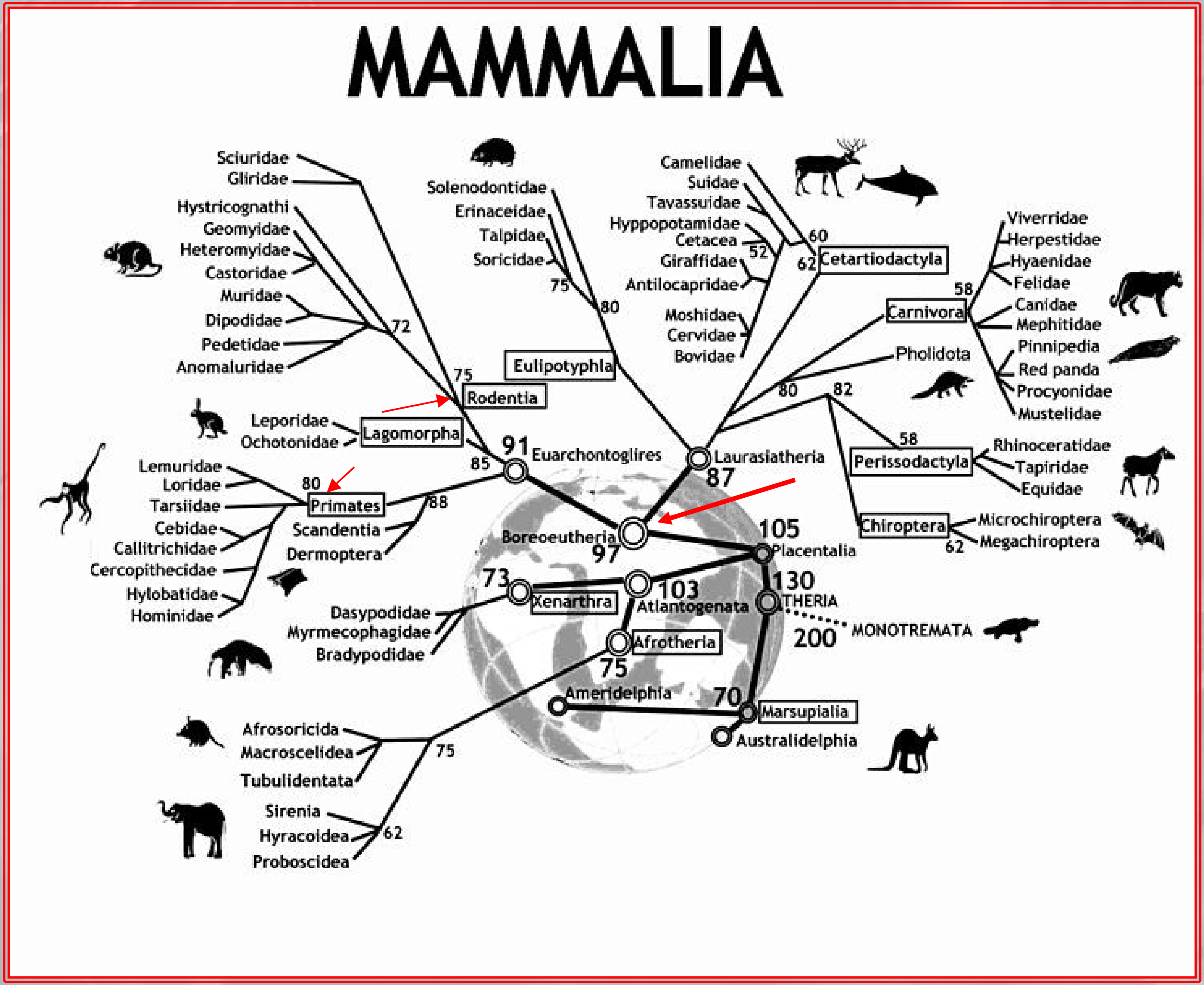
- Occasionally exploring ways to increase the length of the trees, in the hope of going beyond “local minima”, involves a very high computational cost
- Since it is the amount of alignments, and not their length, to create the largest computational problems, a plausible alternative is to split alignments, involving many sequences, into smaller groups



Heuristic search – 9

- Example
  Multiple alignments among a large number of homologous sequences of mammals can be realized by dividing/grouping:
  - The primates, to determine the relationships at the top of their tree trunk
  - The rodents, to determine the relationships at the top of their tree trunk
  - A subset of artiodactyls (cows), lagomorphs (rabbits), primates and rodents, to examine the oldest and the most recent divergence events
  - ➡ Opportunity to examine the most ancient divergences and the final positioning of the most detailed trunks of rodents and primates

Heuristic search – 10

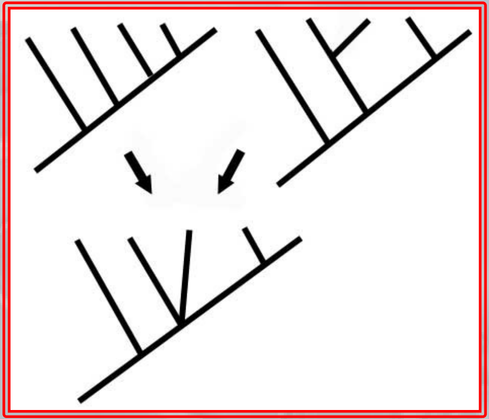


Heuristic search – 11

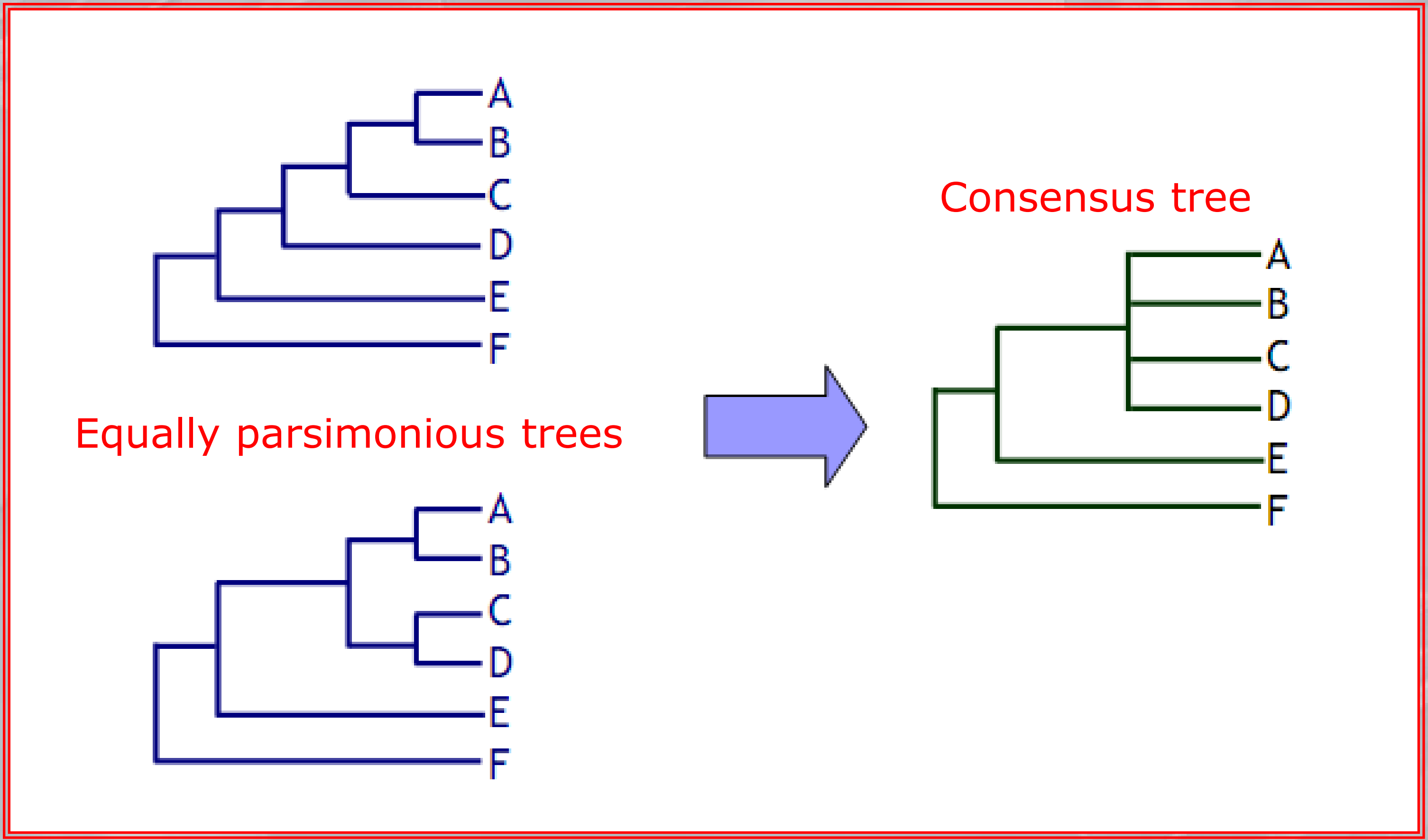
- When such a strategy is adopted, having an *a priori* knowledge of the general relations among the sequences (e.g., all the primates are related to each other more than they are to any other mammals) is crucial
- …but not essential, because a heuristic algorithm may also be required to consider separately each group of sequences that exceeds a particular threshold of pairwise similarity

Consensus trees – 1

- Parsimony approaches normally produce many equally parsimonious trees, too many to be used as a summary of the underlying phylogenetic information
- A consensus tree must be defined, that "summarizes" all the most parsimonious trees
  - The branch points where all the considered trees are in agreement are represented in the consensus tree as bifurcations
  - The points of disagreement are merged together in internal nodes that connect three or more descendant branches



Consensus trees – 2



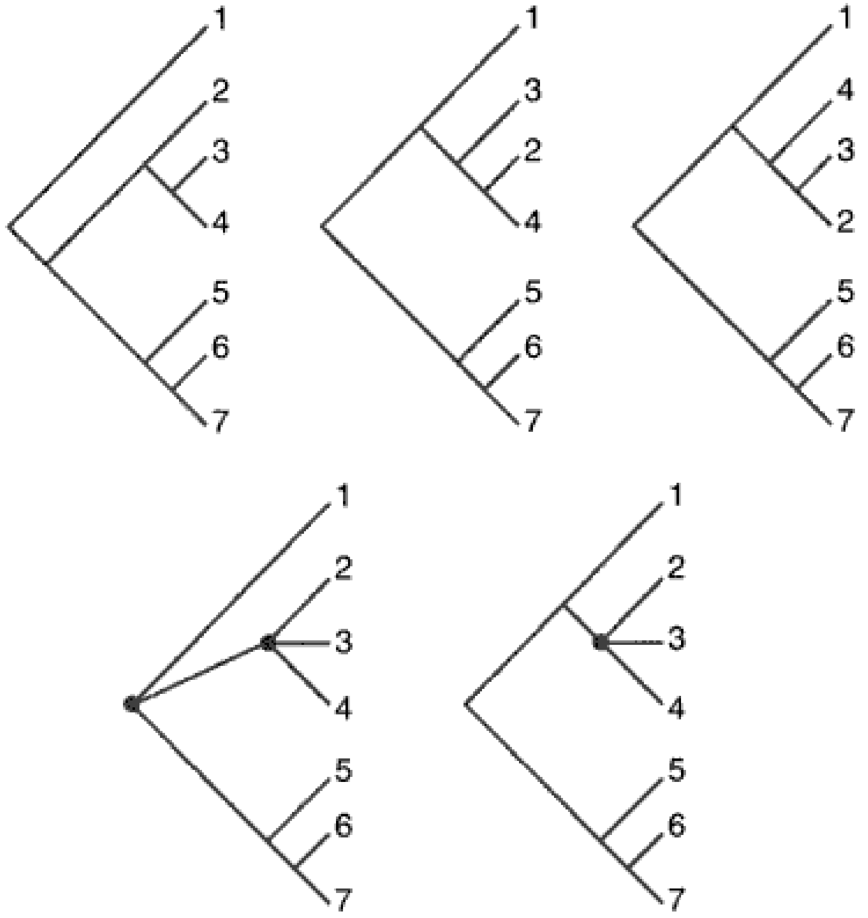
Equally parsimonious trees

Consensus tree

Consensus trees – 3

- In a strict consensus tree, all the disagreement points are treated in a uniform manner, even when a single tree is not consistent with hundreds of others, which agree with respect to a particular branch point
- Alternatively, using the “more than 50% consensus” rule, each internal node that is present in at least half of the trees is represented as a simple bifurcation, while the nodes on which less than half of the trees are in agreement are represented as multifurcations

Consensus trees – 4



Strict consensus

“More than 50% consensus” rule

Tree confidence

- All phylogenetic trees represent a hypothesis about the evolutionary history of the sequences that are collected in a dataset
- It is therefore appropriate to ask the following questions
  - How much confidence can be associated with a tree as a whole and with its constituent parts (subtrees/ arcs)?
  - ➔ Bootstrapping
  - Which is the probability that a certain tree is actually correct with respect to an alternative tree chosen *ad hoc* or at random?
  - ➔ Parametric tests

Bootstrapping – 1

- Different portions of inferred trees can be determined with varying confidence degrees
- The bootstrap test allows a rough quantification of such confidence levels
- Bootstrap
    - A subset of the original data is extracted (based on permutations) and a new tree is inferred from that subset
    - The process of creating new subsets is repeated in order to create hundreds/thousands of resampled datasets
    - The portions of the inferred trees that are mostly represented in the consensus tree are those particularly well supported by the original set of data

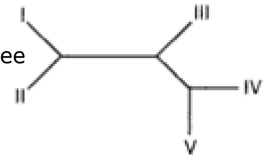
Bootstrapping – 2

Random permutation destroys any correlation among characters beyond that expected by chance alone

Position

Sequence	1	2	3	4	5	6	7	8	9	10
I	G	G	G	G	G	G	A	T	C	A
II	G	G	G	A	G	T	A	T	C	A
III	G	G	A	T	A	G	A	C	A	T
IV	G	A	T	C	A	T	G	T	A	T
V	G	T	T	C	A	T	A	T	C	T

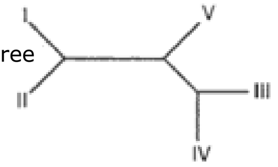
Inferred tree



Position

Sequence	1	1	3	5	5	5	6	9	9	9
I	G	G	G	G	G	G	G	C	C	C
II	G	G	G	G	G	G	T	C	C	C
III	G	G	A	A	A	A	G	A	A	A
IV	G	G	T	A	A	A	T	A	A	A
V	G	G	T	A	A	A	T	C	C	C

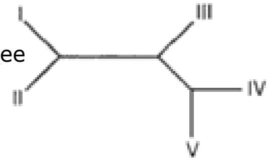
Bootstrap tree #1



Position

Sequence	1	2	2	4	5	5	5	7	7	10
I	G	G	G	G	G	G	G	A	A	A
II	G	G	G	A	G	G	G	A	A	A
III	G	G	G	T	A	A	A	A	A	T
IV	G	A	A	C	A	A	A	G	G	T
V	G	T	T	C	A	A	A	A	A	T

Bootstrap tree #2

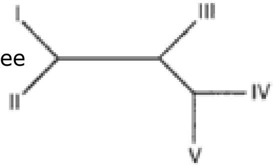


...

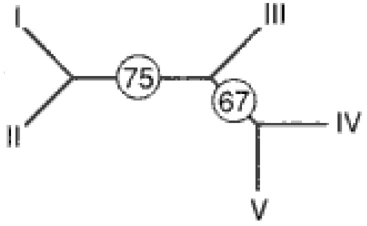
Position

Sequence	3	3	3	4	6	6	7	8	8	8
I	G	G	G	G	G	G	A	T	T	T
II	G	G	G	A	T	T	A	T	T	T
III	A	A	A	T	G	G	A	C	C	C
IV	T	T	T	C	T	T	G	T	T	T
V	T	T	T	C	T	T	A	T	T	T

Bootstrap tree #n



Bootstrap consensus tree



The numbers that count the fraction of bootstrap trees reproducing the same node are positioned close to the corresponding node in the consensus tree, to provide indications on the relative confidence of each part of the tree

Bootstrapping – 3

- The frequency with which different groups are found in the constructed consensus tree (called bootstrap proportions) is a measure of the statistical support for that group
- Values above 80% indicate a very strong support
- However, even values higher than 50% indicate that a group is frequently found in the pseudo–datasets
- A low statistical support does not necessarily imply a "wrong" clade

Bootstrapping – 4

- Despite the frequent use of bootstrap–like methods in the scientific literature, the bootstrap results should be treated with some caution
  - When they are based on “few” iterations, that is, cycles of resampling and tree generation, they are probably not very reliable, especially when a large number of sequences is involved
  - The confidence is normally underestimated at high levels and overestimated at low levels
  - *Fallacy of multiple tests*: simple fluctuations seem to have statistical significance
- Apart from the highlighted problems, in this way, trees that are more accurate representations of the “true” phylogenetic tree can be normally gained, with respect to the method of calculating the single most parsimonious tree

Parametric tests – 1

- Since the parsimony approaches often generate a lot of trees that have the same minimum number of substitutions, there are also many alternative trees that postulate a few more substitutions
- Even in this case, the principle underlying the concept of parsimony suggests that the tree which postulates the fewest number of substitutions most probably describes the true relationship among the sequences
- However, there does not exist a limit on the number of replacements postulated by the most parsimonious tree and, for datasets that involve many dissimilar sequences, many thousands of replacements can easily be estimated

Parametric tests – 2

- In such cases, it is reasonable to ask whether a tree, which is already so unlikely as to postulate 10000 substitutions, is significantly more likely than an alternative tree which postulates 10001 substitutions
- Or… how much greater is the probability of the most parsimonious tree with respect to a particular alternative tree previously proposed to describe the relationship among a given set of taxa?
- To this question it is possible to provide an answer, albeit partial, using a parametric test
- A parametric test is a statistical test that can be applied to normally distributed data

Parametric tests – 3

- This is done by performing a hypothesis testing on the value of a parameter, such as the standard deviation, the equality between two means, etc.
  - In a phylogenetic context, the parametric test most often used is due to H. Kishino and M. Hasegawa (1989)
  - It is assumed that the informative sites within an alignment are independent and equivalent; the difference of the minimum number of substitutions postulated by two trees, D, is used as a statistical test (calculating the variance V over the entire set of informative sites)
- Alternative parametric tests are available not only for parsimony analysis, but also for distance–based methods and maximum likelihood trees

Comparison among phylogenetic methods

- Neither phylogenetic reconstruction methods based on distance, nor those based on characters can guarantee to be able to describe the true tree that tracks the evolutionary history of a set of aligned sequences
- However…
  - Those datasets which allow a method to infer the correct phylogenetic relationship, generally, lead to good results with all the commonly used approaches
  - If many changes have been occurred within the data or if the substitution frequencies significantly vary from branch to branch, no method works in a truly reliable way
- However, if by processing a dataset with fundamentally different methods, we always obtain the same tree, that tree can be considered “reliable”

Molecular phylogenies

- Over the past forty years, numerous interesting examples of evolutionary relationships deciphered by sequence analysis have been accumulated
- These studies have had important implications in medicine, agriculture, conservation of the species
    - It is likely that a particular drug effective against a certain type of infection is also effective on infections caused by related organisms
    - Easy transfer of resistance factors to parasites among closely related plant species
    - Possibility of determining whether a given population of organisms is distinguished enough to be classified as a separate species, to eventually deserve it a special protection

The tree of life – 1

- One of the most striking cases in which the sequence analysis has provided new insights into the evolutionary relationships is related to the understanding of the fundamental classifications of life forms
- Originally, biologists divided all life forms into two main groups: plants and animals
- Nevertheless, with the subsequent discoveries of new organisms and with the study of their characteristics, this simple dichotomy became not convincing
- It was then later recognized that organisms could be divided into prokaryotes and eukaryotes, on the basis of their cellular structure

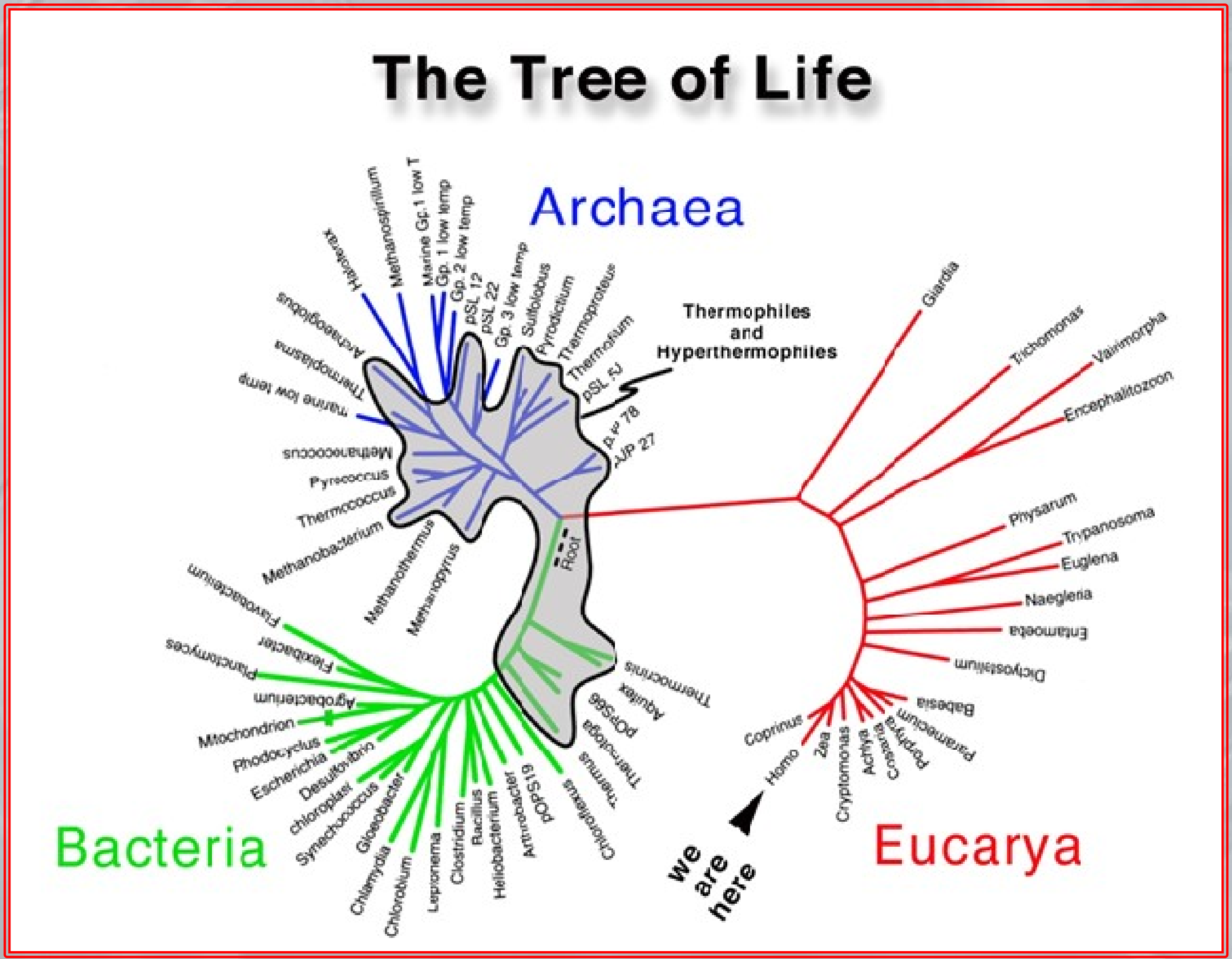
The tree of life – 2

- Most recently, several classifications have been accepted for living organisms, such as the five kingdoms proposed by Whittaker: prokaryotes, protists, plants, fungi and animals
- However, a negative test – i.e., the absence of internal membranes that distinguishes prokaryotes – has been universally recognized as inadequate to taxonomically group all the living organisms
- Since the late ‘70s, for the first time, RNA and DNA sequences were used to discover the basic lines of the evolutionary history of all the species

The tree of life – 3

- In a famous study, Carl Woese *et al.* built an evolutionary tree for all the forms of life based on the nucleotide sequences of the 16s rRNA (ribosomal RNA) gene, which is present in all the organisms
- The rRNA is the most conserved component of the cell
  - The genes coding for rRNA are sequenced to identify the taxonomic group of an organism, to recognize related groups and estimate the divergence rate among the various species
- The evolutionary tree reveals three main groups:
  - Bacteria – prokaryotes
  - Eucarya – eukaryotic organisms, such as plants, animals and fungi
  - Archaea – thermophilic bacteria and little known organisms, that can be studied only through their rRNA sequences

The tree of life – 4

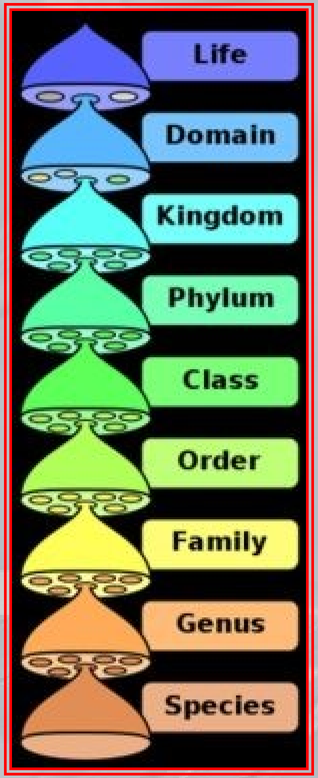


The tree of life – 5

- It was found that Bacteria and Archaea, although both prokaryotes as devoid of internal membranes, were so genetically different as Bacteria and Eucarya
- The deep evolutionary differences between Bacteria and Archaea were not obvious on the basis of the phenotype, whereas the fossil record was completely silent on this topic
- The differences became clear only after their nucleotide sequences were compared
    - Sequences of 5s rRNA and of some genes coding for fundamental proteins also support their membership to two different evolutionary groups

The origin of man – 1

Domain: Eukarya
Kingdom: Animalia
Subkingdom: Eumetazoa
Phylum: Chordata
Subphylum: Vertebrata
Class: Mammalia
Subclass: Eutheria
Order: Primates
Superfamily: Hominoidea
Family: Hominidae
Genus: Homo
Species: Homo sapiens
Subspecies: Homo sapiens sapiens



The origin of man – 2

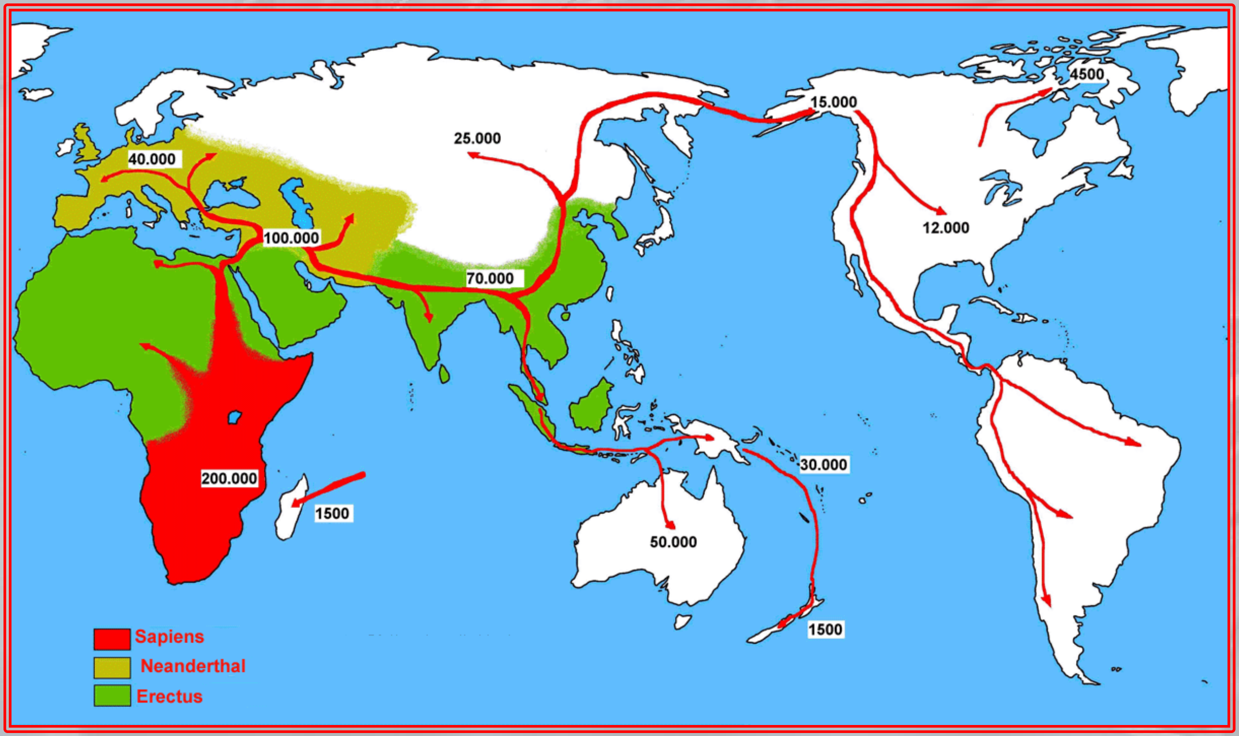
- In contrast to the large variability observed in size, in the body shape, in the facial features, w.r.t. the skin color, the muscle fibers, the bone density, etc., genetic differences between human populations are relatively small
- The analysis of mtDNA sequences (speciation event man–chimp dates back to 4–13 million years ago) reveals that the average difference between two human populations is approximately of 0.33%
- Other primates show much greater differences: the two orangutan subspecies differ for about 5%
  - Human groups are closely related even if they have some genetic differences

The origin of man – 3

- Surprisingly, the major differences are not found between populations located on different continents, but among the people living in Africa
- All other human populations show less significant differences than those detectable among the African people
  - Man originated and underwent the first evolutionary divergence in Africa
  - After the development of a number of genetically differentiated populations, a small group of humans could be migrated out of Africa and has originated all other human populations
    - *Out–of–Africa theory*: analysis of data coming both from the mitochondrial DNA and from the Y chromosome in the nucleus are consistent with this hypothesis

The origin of man – 4

- Further interpretations of the data suggest that all living humans share mitochondria that are derived from a “mitochondrial Eve” and that the Y chromosome of all men comes from a “Y Adam chromosome” of about 200,000 years ago



Just a curiosity… – 1

- Beleza *et al.*, *Molecular Biology and Evolution*, January 2013
  - Study of several genes that affect the skin color in order to understand when the divergence event has occurred
  - The results showed that the spread of an allele, shared by both Europeans and Asians, dated back to about 30,000 years ago, after the migration from Africa, that occurred about 100,000 years ago
  - Conversely, variants of other genes, typically related to European populations, would be much more recent, dating back to 11,000–19,000 years ago
  - But what have been the factors that influenced the selection of gene variants that code for a lighter color of the skin?

Just a curiosity… – 2

- The period between 11,000 and 19,000 years ago is at the peak of the last ice age and it is reasonable to believe that human beings, to protect themselves from the cold weather, covered themselves and lived in shelters, limiting their exposure to UV rays
- It is likely that these changes have encouraged the spread of alleles for clear skin, so as to ensure an adequate production of vitamin D, which is useful to fix calcium in the bones
- The selection of genes coding for a clearer complexion occurred, in European populations, relatively recently and the selective pressure has favored the cutaneous conditions for an adequate synopsis of vitamin D
- With a little sun exposure, a skin less rich in melanin is efficient at producing vitamin D, and reduce the risk of its lack and the related consequences

Concluding… – 1

- Character–based phylogenetic reconstruction methods mainly focus on the parsimony principle – substitutions are rare events and the phylogeny that invokes the fewest number of substitutions is the one that most likely reflects the true relationship between the considered sequences
- In addition to describe relationships among the sequences, parsimony approaches can provide potentially useful inferences about the sequence of long extinct ancestors of all the living organisms
- However, the parsimony analysis can be computationally heavy, especially if considering multiple alignments related to a great amount of sequences

Concluding… – 2

- The analysed data often lead to different trees that are equally parsimonious and, to summarize them, consensus trees can be used
- There are several methods to determine the robustness of parsimonious trees, including bootstrap and parametric tests, although we cannot guarantee that an inferred tree — both with character–based and distance–based approaches — represents the true evolutionary relationship among the considered sequences