Molecular Phylogenetic Trees

Lecture outline

- Introduction to molecular evolution
- Terminologies frequently used to characterize a phylogenetic tree
- Five stages of phylogenetic analysis

  (1) Choosing molecular markers;

  (2) Performing multiple sequence alignment;

  (3) Choosing substitution models;

  (4) Determining a tree building method;

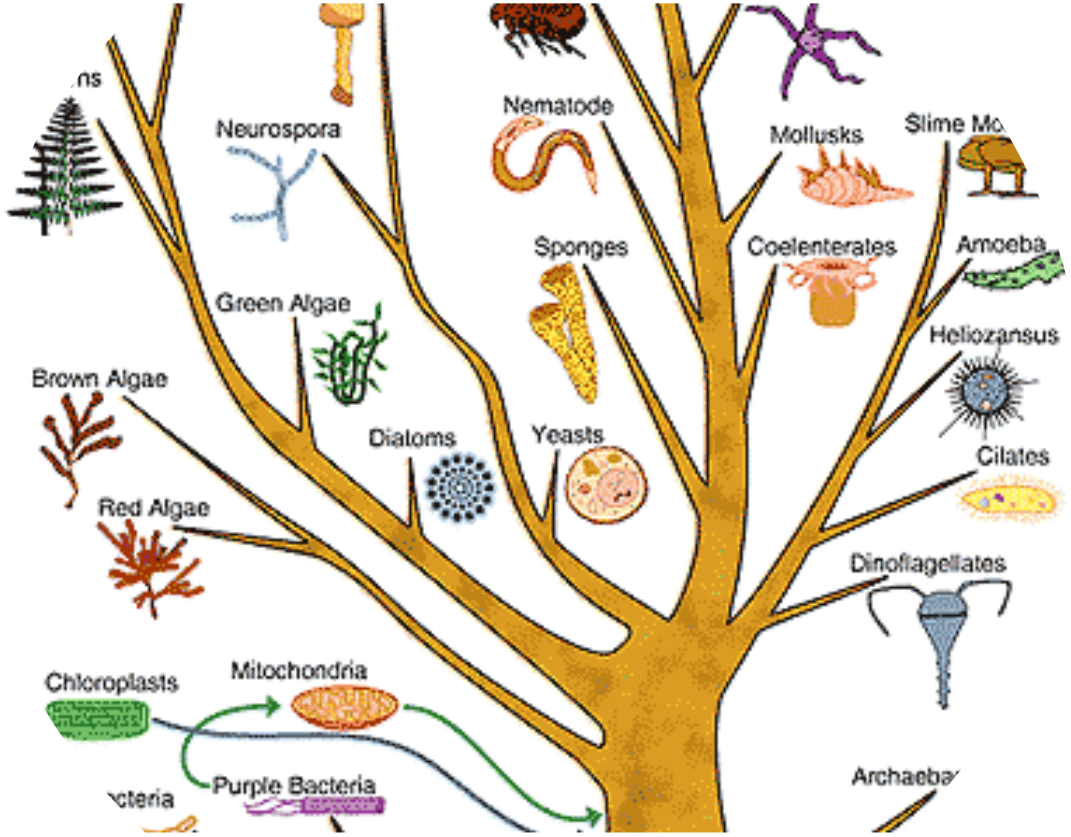
  (5) Assessing tree reliability.

What is biological evolution

- In the biological context, **evolution** refers to change through time as species become modified and diverge to produce multiple descendant species.
- **Evolution** is the historical occurrence of change, and **natural selection** is an important mechanism that can cause it.
- The underlying mechanism of evolution is genetic mutations occur spontaneously —— biological diversity within a population.

Phylogenetics

- **Phylogeny** is the evolutionary history and relationships between groups of organisms.
- **Phylogenetics** is the study of the evolutionary history of living organisms.
- **Phylogenetic trees** are tree-like diagrams to represent the evolutionary relationships between different organisms —— a visual representation.



Study phylogeny with fossil records

- Pros:
  - Fossil records contain morphological information about ancestors of current species and the timeline of divergence.
- Cons:
  - Available only for certain species, e.g. for microorganisms, fossils are essentially nonexistent
  - Existing fossil data can be fragmentary and their collection is often limited by abundance, habitat, geographic range, and other factors.
  - The descriptions of morphological traits are often ambiguous and subjective.
- Thus, using fossil records to determine phylogenetic relationships can often be biased.

Study phylogeny with molecular data

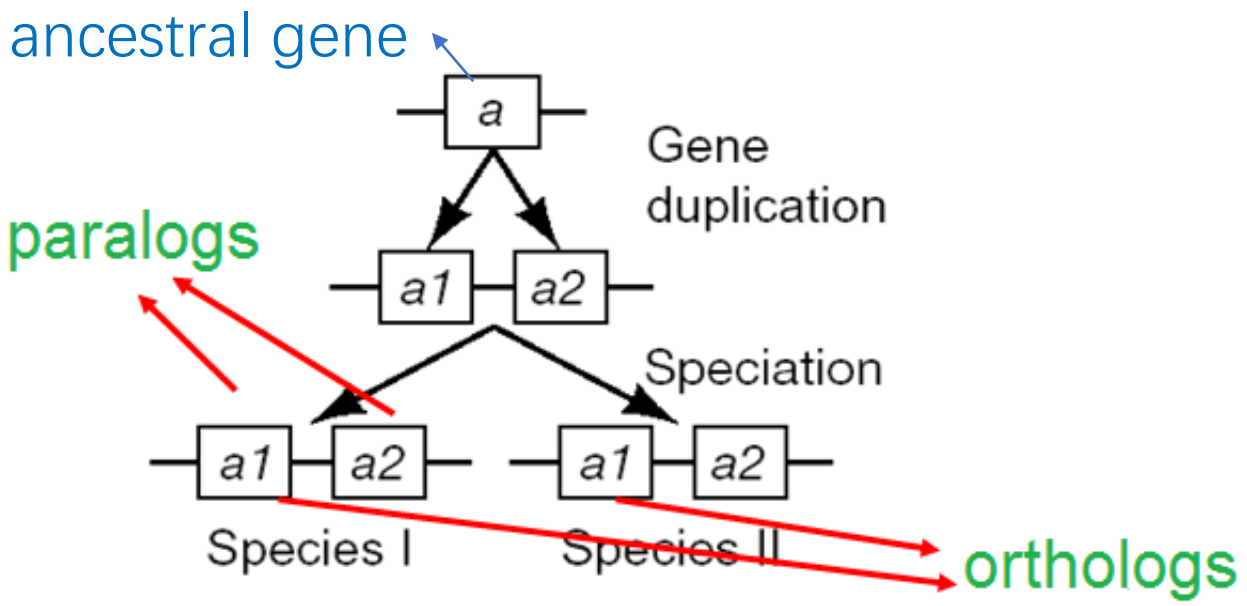
- Molecular data can also provide very useful evolutionary relationships of existing organisms because, as organisms evolve, the genetic materials accumulate mutations over time causing phenotypic changes.
- Genes —— molecular fossils
- Molecular evolution focuses on the molecular changes within macromolecules such as DNA, RNA and protein.
- Molecular evolution is the study of changes in genes and proteins by analyzing mutations at various positions in their sequences.
- Evolutionary relationships between the organism can often be inferred based on the sequence similarity of the molecules.

Study phylogeny with molecular data

- Pros:
  - Molecular data are more numerous than fossil records and are easier to obtain.
  - High-throughput sequencing generates tremendous amounts of molecular sequence data, which has led to the repaid development of molecular phylogenetics.
  - There is no sampling bias involved.
- Cons:
  - The species evolution is the combined result of evolution by multiple genes in a genome. The evolution of a particular gene does not necessarily correlate with the evolutionary path of the species.
  - A gene phylogeny (phylogeny inferred from a gene or protein sequence) only describes the evolution of that particular gene or encoded protein.

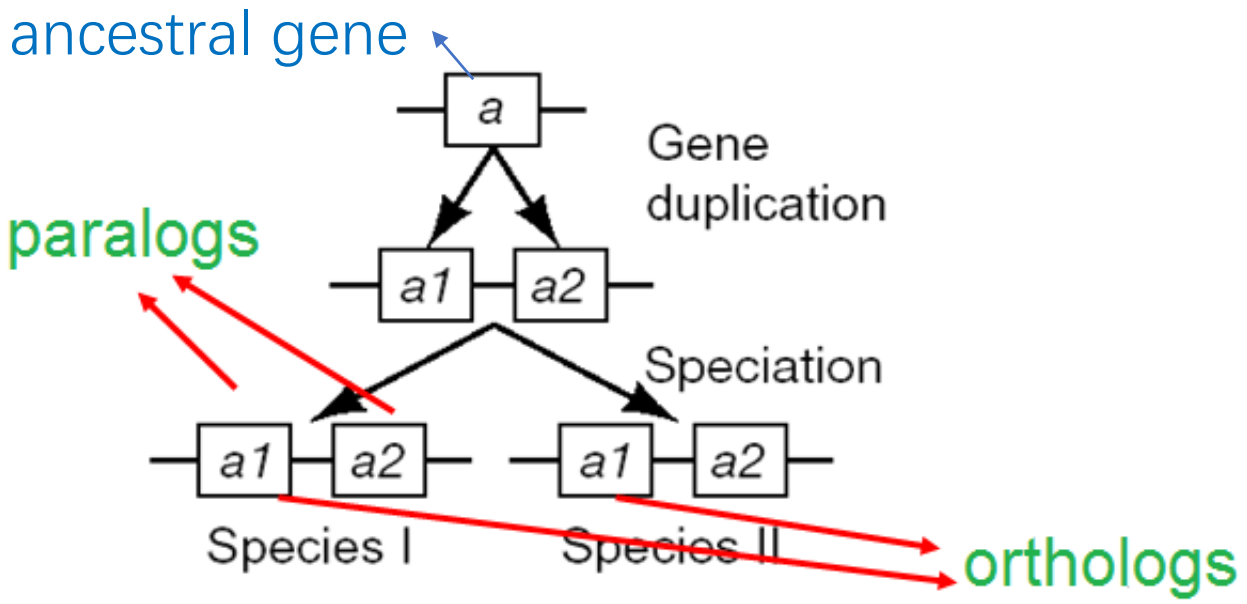
Biological definitions for related sequences

- **Homologues** are similar sequences in a single or two different organisms that have been **derived from a common ancestor sequence.**
- Homologues can be described as either **paralogues** or **orthologues**.

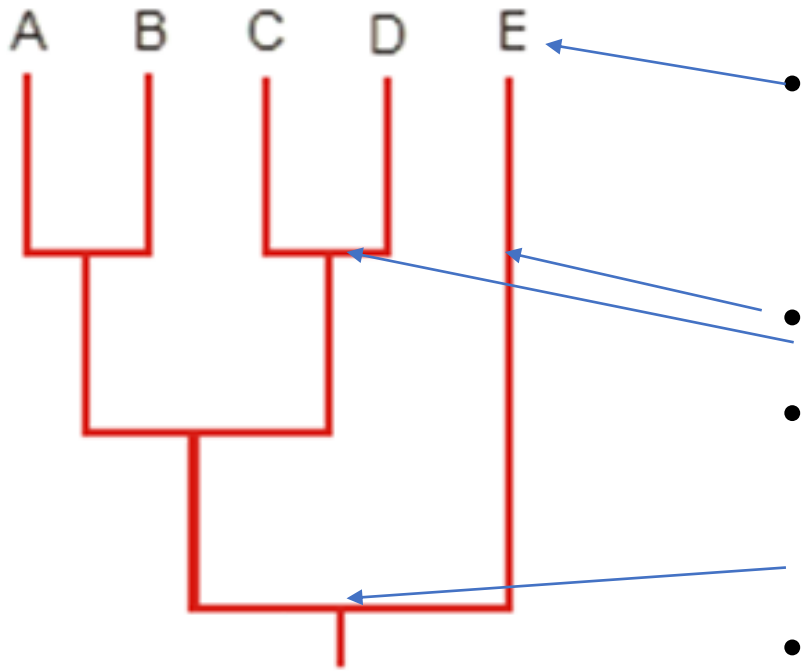


Biological definitions for related sequences

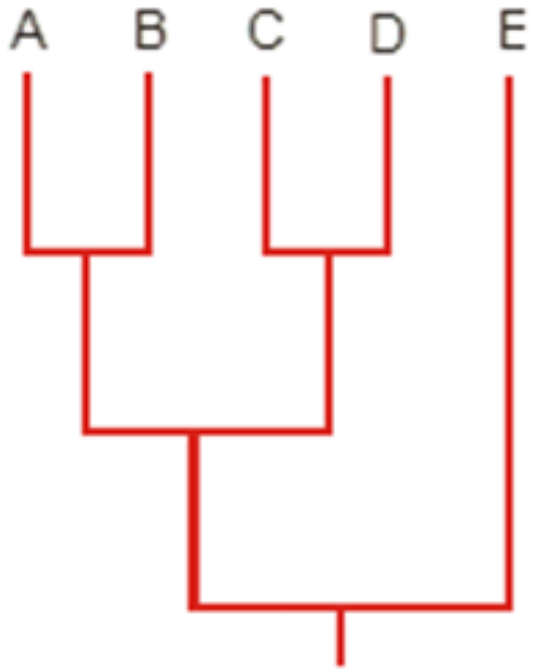
- **Paralogous genes** are homologous genes that have diverged **within one species**. A paralogous gene arises during **gene duplication** where one copy of the gene receives a mutation that gives rise to a new gene with a new, yet similar function.
- **Orthologous genes** are homologous genes that diverged after evolution giving rise to **different species**, an event known as **speciation**. The Orthologous genes generally maintain the same function to that of the ancestral gene.



Terminology



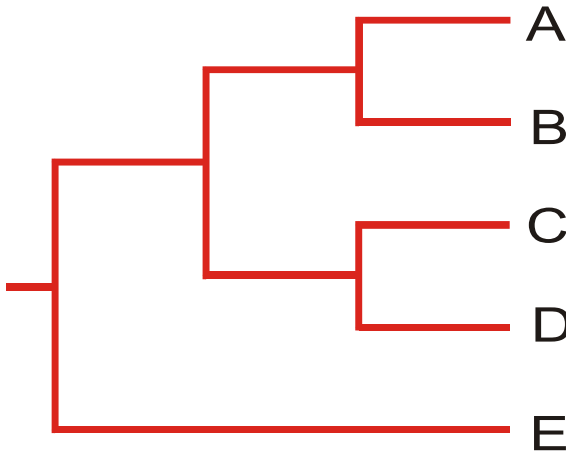
- **Terminal nodes** (leaves) = present-day species or sequences (for which we have data), known as **taxa**
- **Branches** (edges) show relationship
- **Internal nodes** = inferred ancestors, the connecting points where two adjacent branches join
- **Root node** = common ancestor of all members of the tree

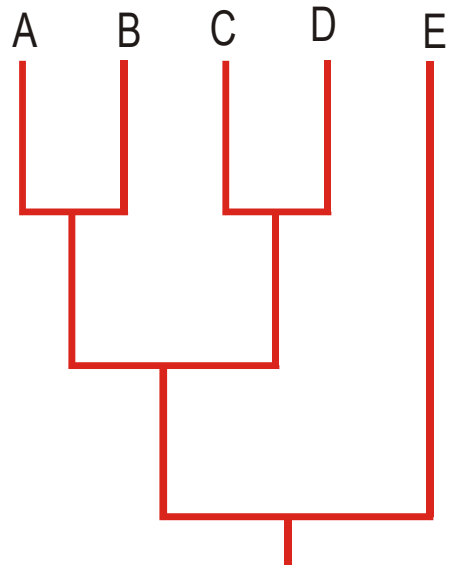


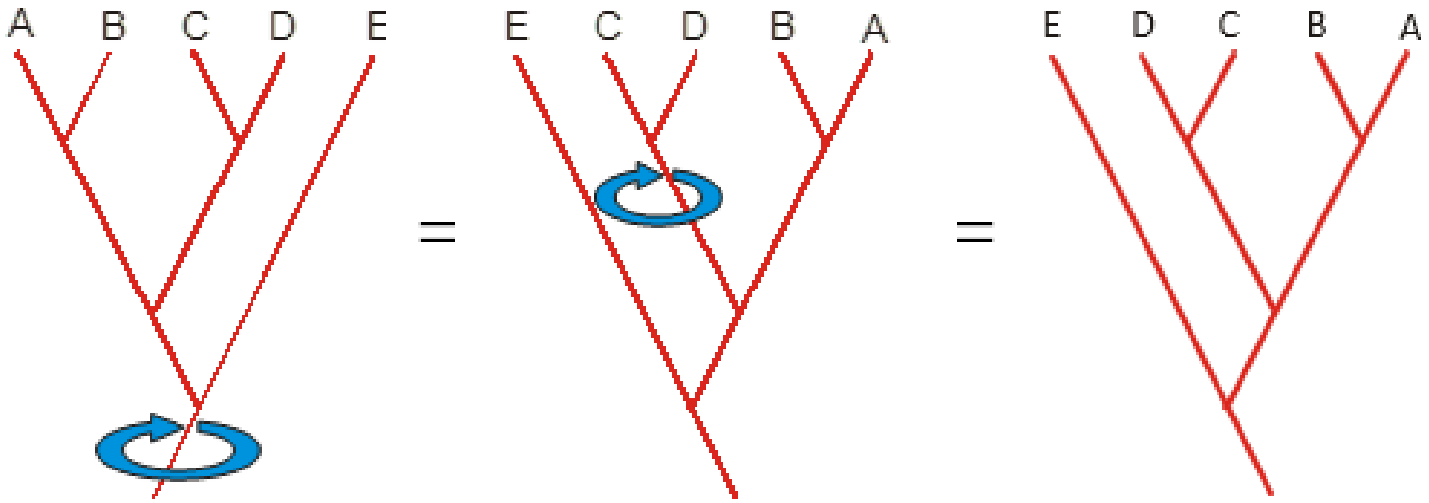
Terminology

- **Length of the branch** can reflect evolutionary distacne
- **Clade**: a group of two or more taxa that includes both their closest common ancestor and all its descendants
- **Tree topology**: the branching pattern in a tree
- Examples:
- Species tree – how similar are species
- Gene trees – how similar are genes

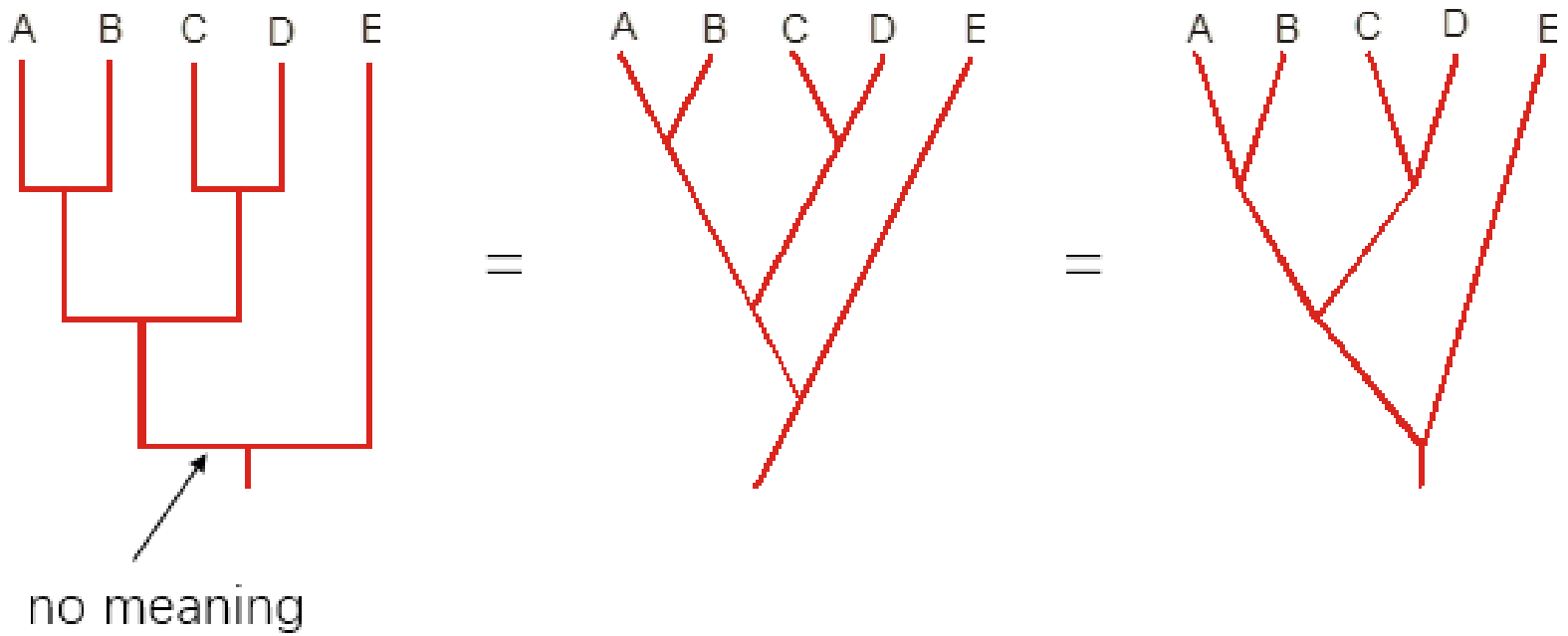
There are many ways of drawing a tree

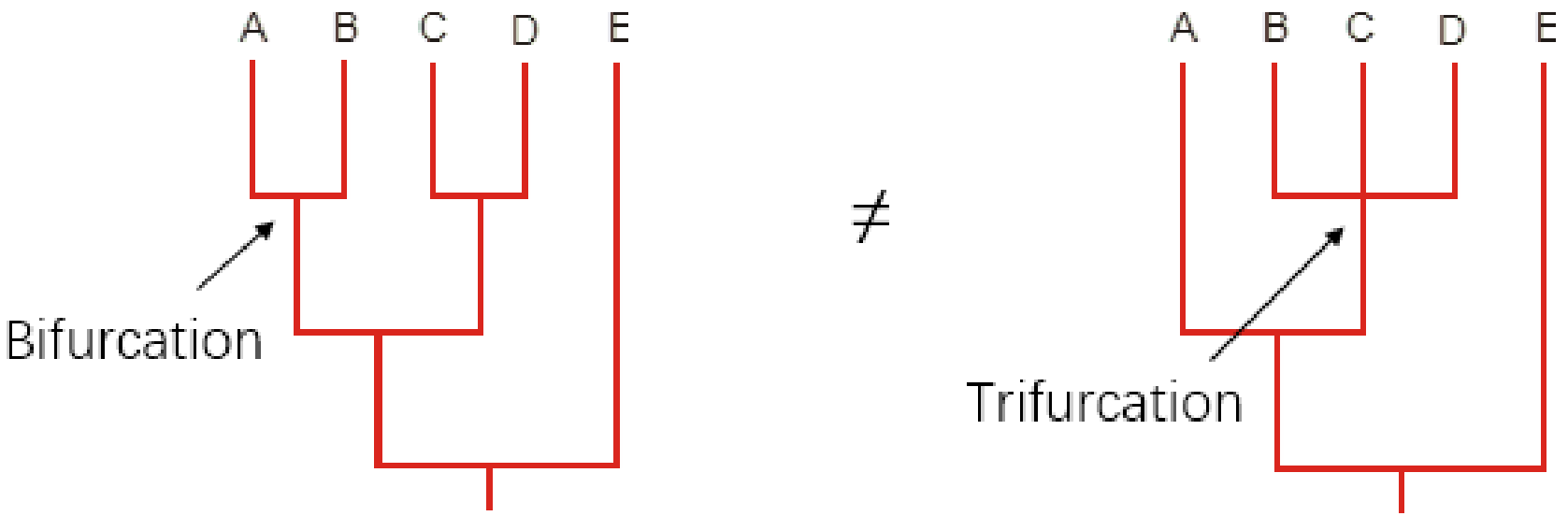






Branches of a tree can freely rotate without changing the relationships among the taxa.

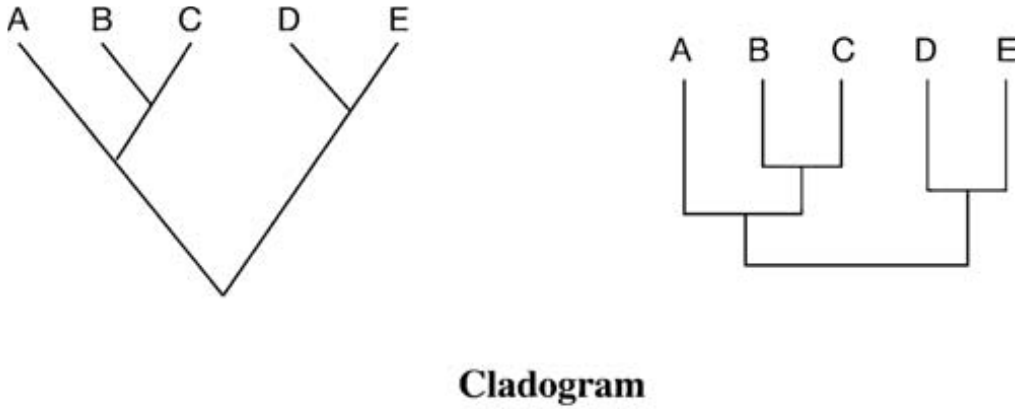

A B C D E
no meaning
=
A B C D E
=
A B C D E



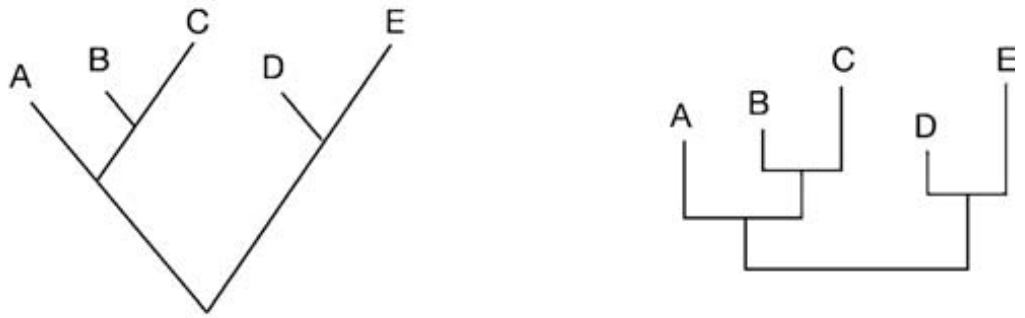
- Bifurcation: nodes divide and give rise to two descendants
- Multifurcation: nodes have more than two descendants. A multifurcation often represents an unresolved phylogeny in which the exact order of bifurcation can’t be determined precisely.

Scaled and unscaled trees

- Unscaled tree:
  - The taxa line up neatly in a row or column. The branch lengths have no phylogenetic meaning
  - Cladogram
- Scaled tree:
  - The branch lengths represent the amount of evolutionary divergence
  - Phylogram



Cladogram

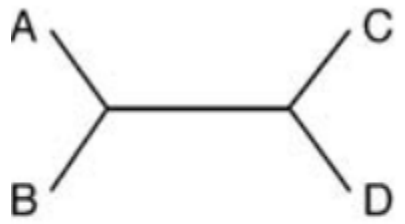


Phylogram

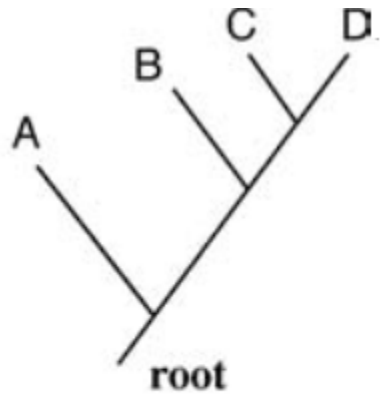
Unrooted and Rooted trees

Unrooted tree:

- does not assume knowledge of a common ancestor, but only show relative relationships among taxa;
- no indication of evolutionary direction.



Unrooted

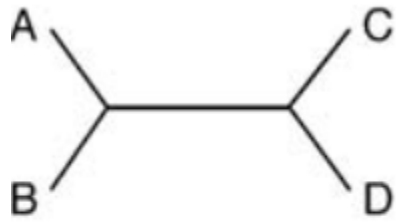


Rooted

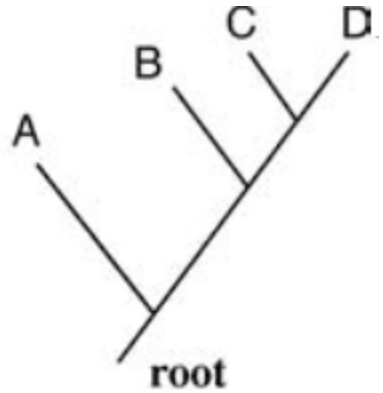
Unrooted and Rooted trees

Rooted tree:

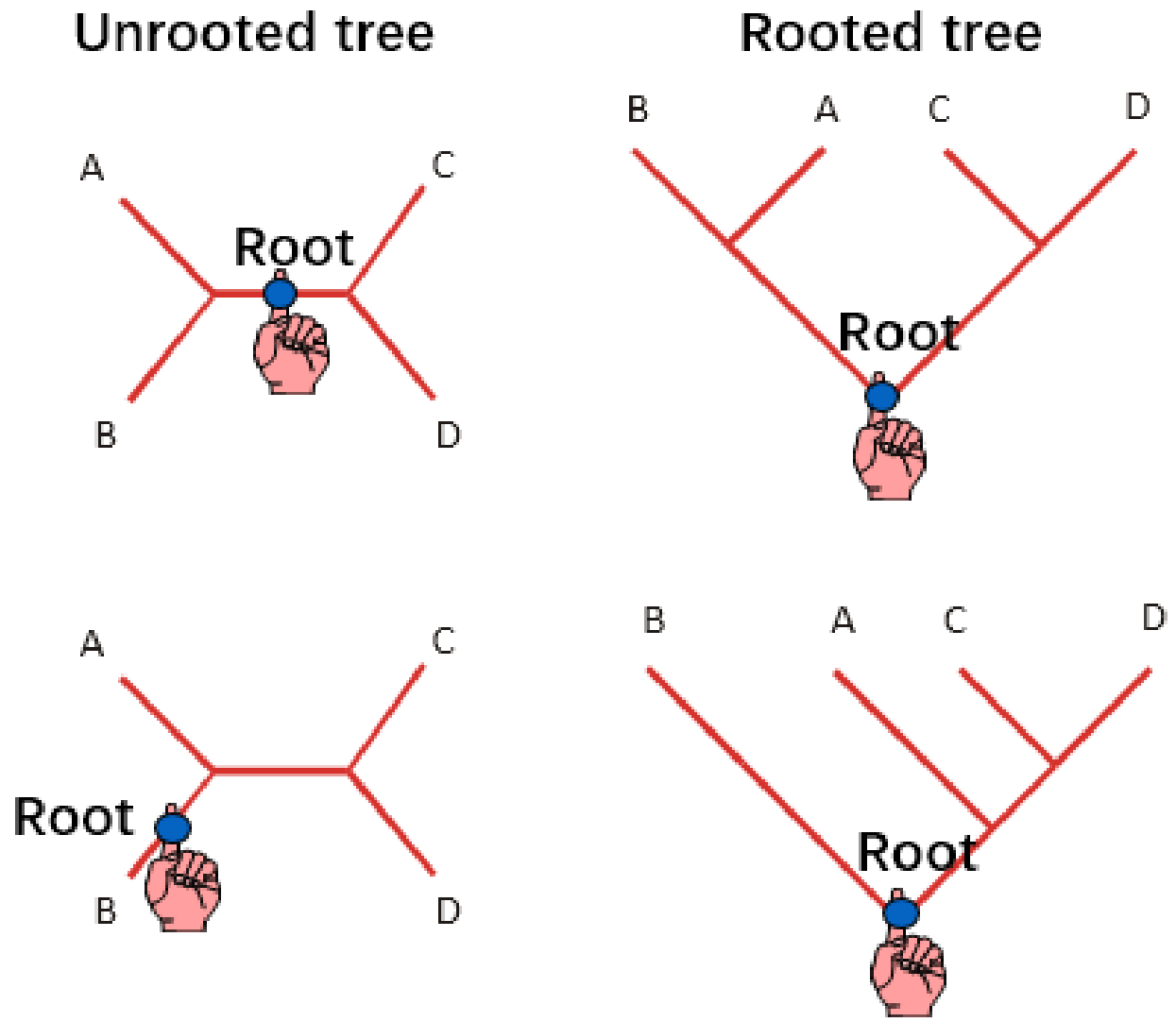
- all the species or sequences under study have a common ancestor or root node;
- have an evolutionary direction from root node leading to all other nodes .



Unrooted



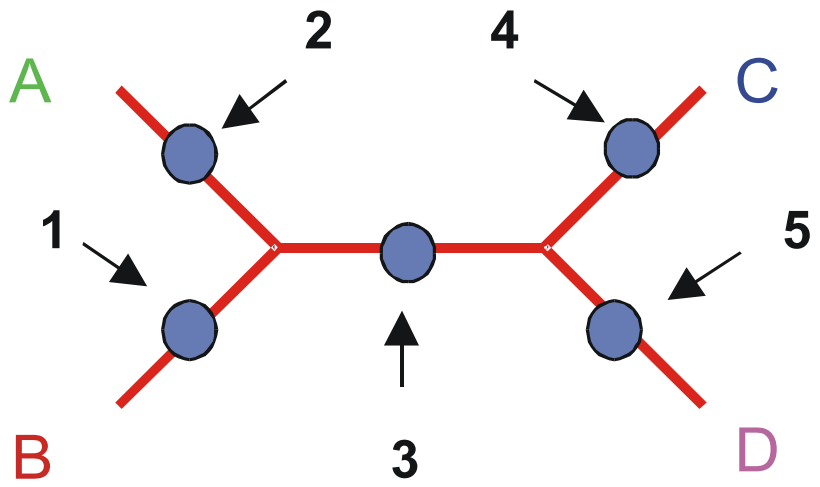
Rooted



A rooted tree is more informative than an unrooted one.

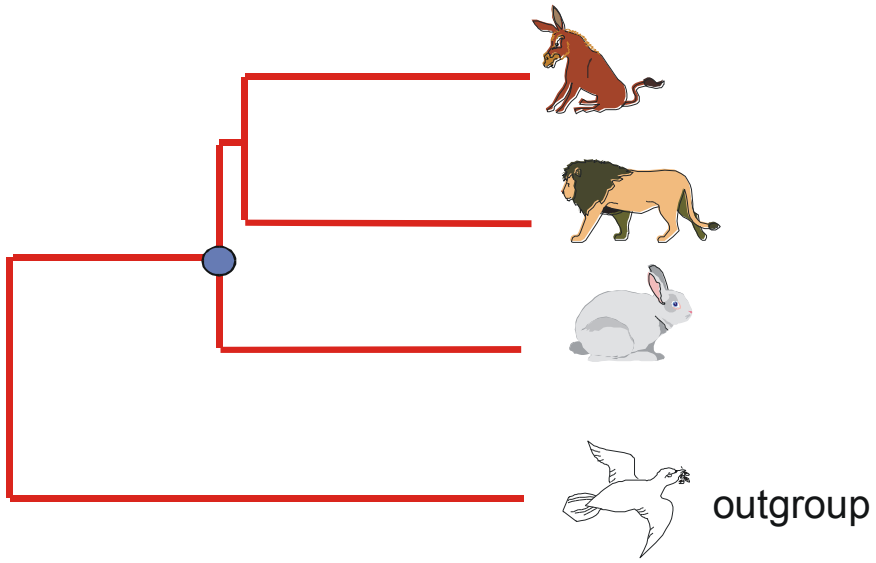
How to define the root of a tree

1. Use information from ancestors. In most cases not available as the common ancestor is already extinct.

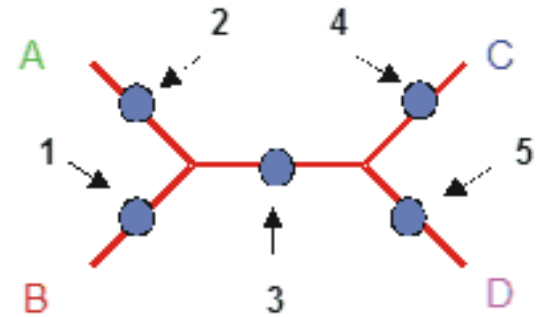


How to define the root of a tree

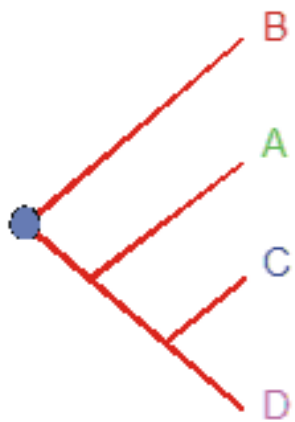
2. Use "**outgroup**", edge joining the outgroup to the rest of the tree is best candidate for root position
   - Outgroups:
     - are sequences that are homologous to the sequences under consideration (ingroup), but separated from those sequences at an early evolutionary time
     - are distinct from the ingroup sequences, but not too distant from the ingroup
     - are generally determined from independent sources of information
       e.g. a bird sequence can be used as a root for the phylogenetic analysis of mammals



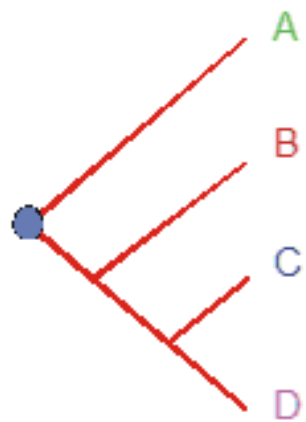
Unrooted tree



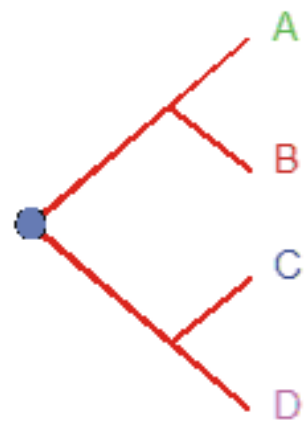
Rooted tree 1



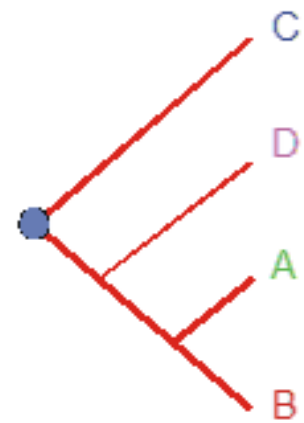
Rooted tree 2



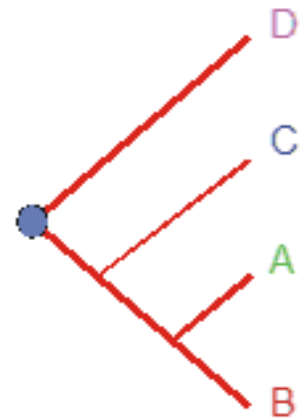
Rooted tree 3



Rooted tree 4



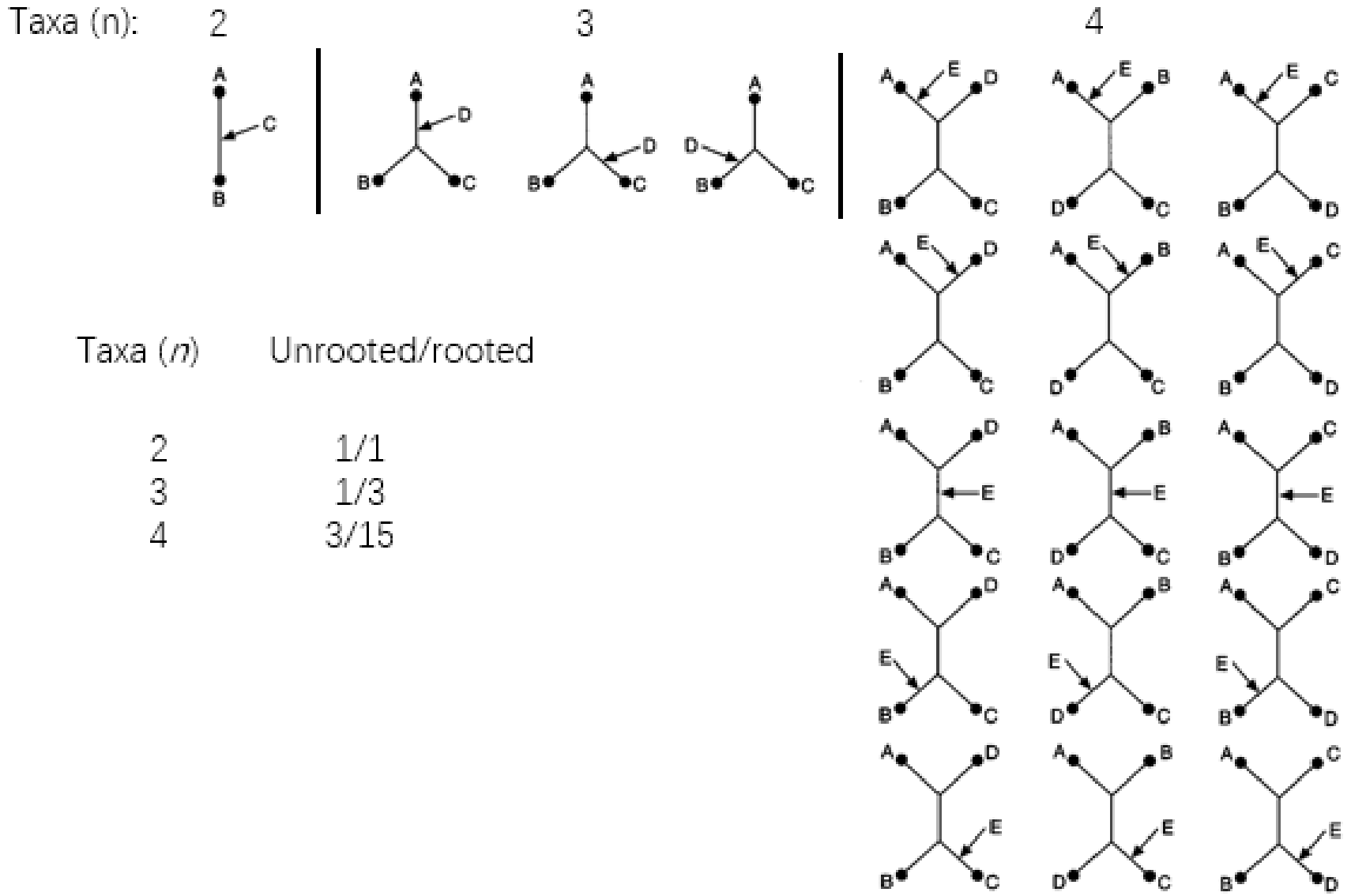
Rooted tree 5



These trees show five different evolutionary relationships among the taxa!

Why finding a true tree is difficult

- The main objective of molecular phylogenetics is to correctly reconstruct the evolutionary history based on the observed sequence divergence between organisms —— finding a correct tree topology with correct branch lengths.
- However, the search for a correct tree topology can sometimes be extremely difficult and computationally demanding.
- The reason is that the number of potential tree topologies can be enormously large even with a moderate number of taxa.
- The increase of possible tree topologies follows an exponential function.



Taxa (*n*)	Unrooted/rooted
2	1/1
3	1/3
4	3/15

There is only one of these 15 trees that accurately describes the evolutionary process by which these four sequences evolved.

Taxa (n)	rooted $(2n-3)!/(2n-2(n-2)!)$	unrooted $(2n-5)!/(2n-3(n-3)!)$
2	1	1
3	3	1
4	15	3
5	105	15
6	954	105
7	10,395	954
8	135,135	10,395
9	2,027,025	135,135
10	34,459,425	2,027,025

- If there are ten taxa, there can be over 2 million unrooted trees and over 30 million rooted ones.
- A true phylogenetic tree can be simplified by calculating the unrooted trees first. Once an optimal tree is found, rooting the tree to produce a rooted tree.

Molecular phylogenetic tree construction

(1) Choosing molecular markers;
(2) Performing multiple sequence alignment;
(3) Choosing substitution models;
(4) Determining a tree building method;
(5) Assessing tree reliability.

Choice of molecular markers

- For constructing molecular phylogenetic trees, one can use either nucleotide or protein sequence data.
- The decision to use nucleotide or protein sequences depends on the properties of the sequences and the purposes of the study.
- In many cases, protein sequences are preferable to nucleotide sequences because protein sequences are relatively more conserved.
- However, for studying closely related sequences, faster evolutionary rates at the DNA level become an advantage.

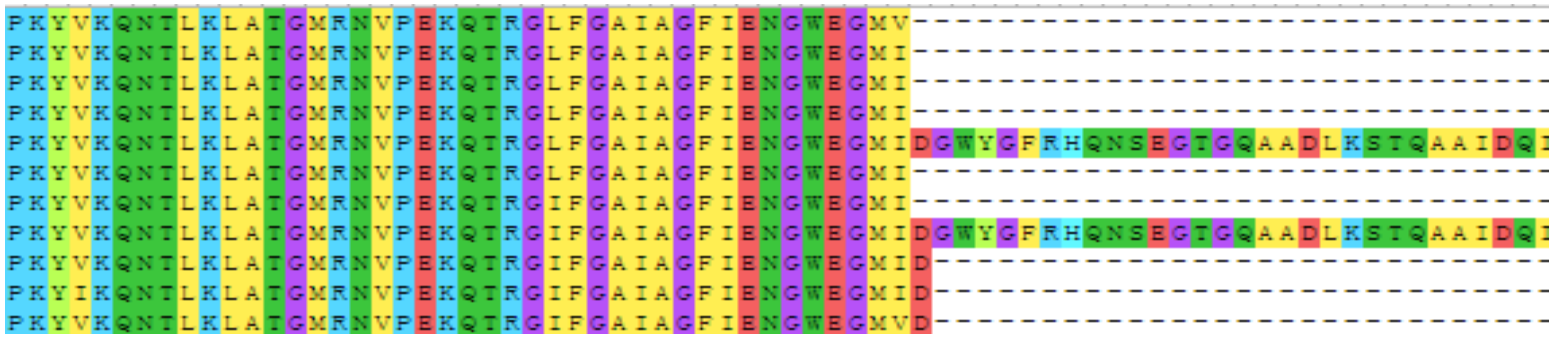
Multiple sequence alignment

- The second step in phylogenetic analysis is to construct sequence alignment——it establishes positional correspondence in evolution.
- Only the correct alignment produces correct phylogenetic inference.

① Confirm that all sequences are homologous

② Adjust gap creation and extension penalties as needed to optimize the alignment

③ Restrict phylogenetic analysis to regions of the multiple sequence alignment for which data are available for most taxa (delete columns having incomplete data)



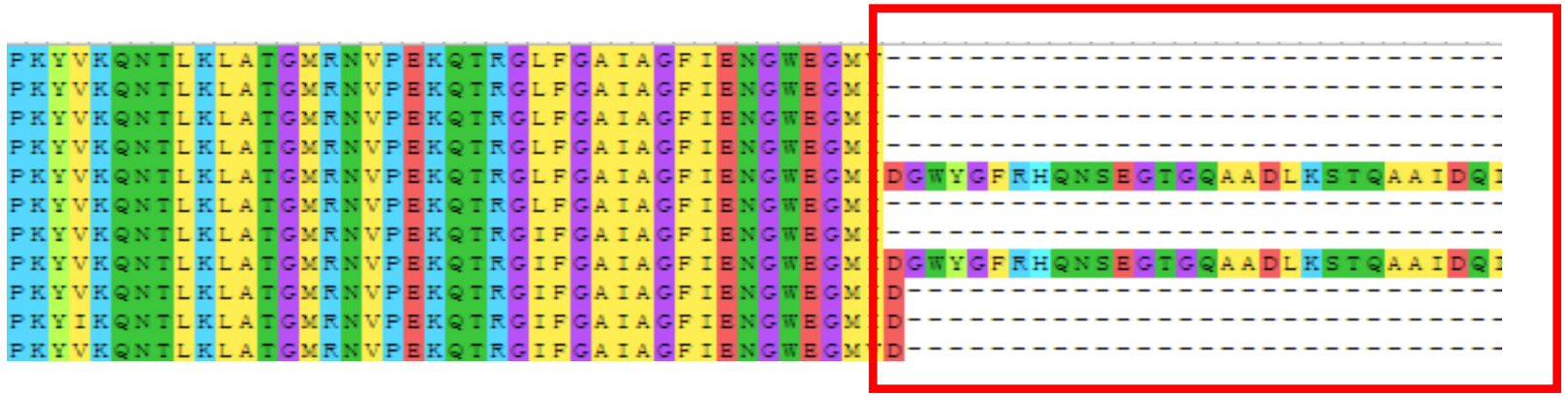
Multiple sequence alignment

- The second step in phylogenetic analysis is to construct sequence alignment——it establishes positional correspondence in evolution.
- Only the correct alignment produces correct phylogenetic inference.

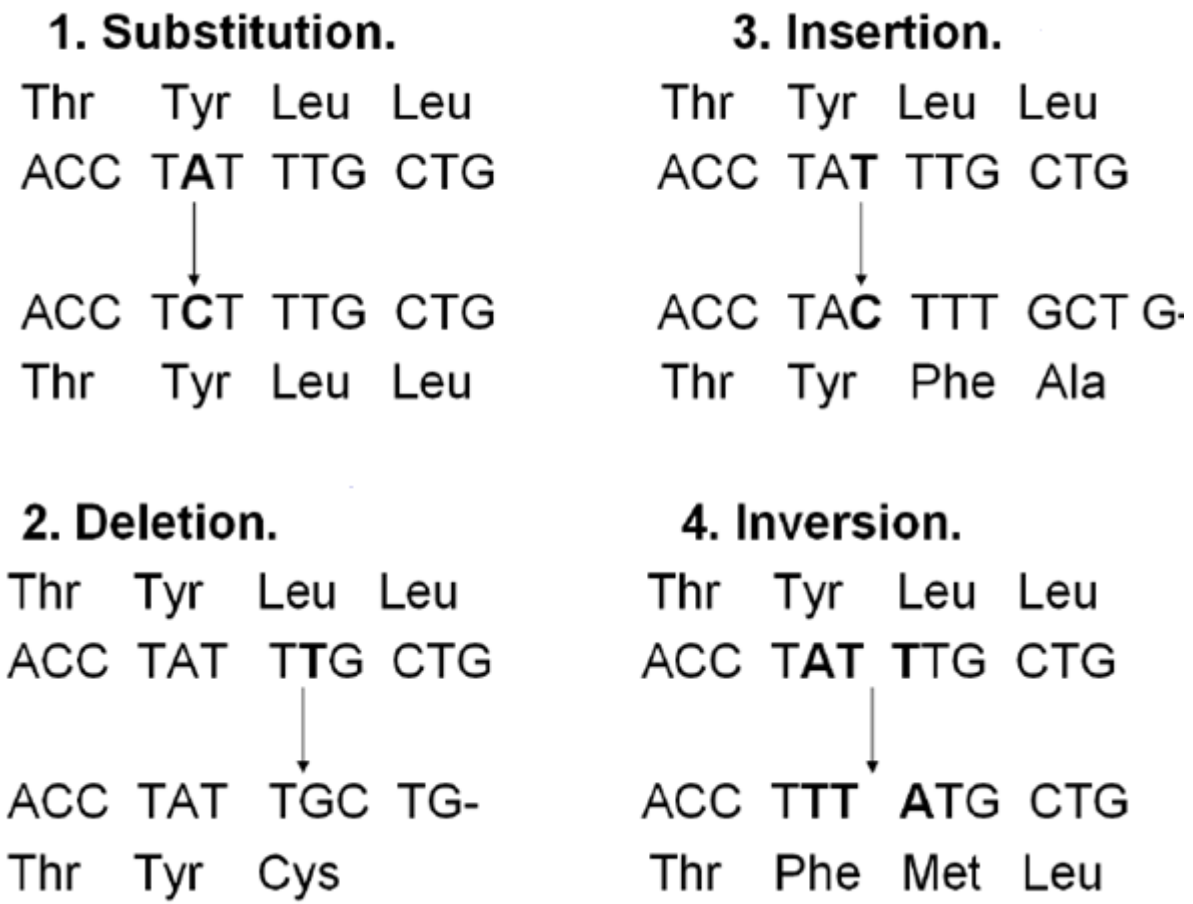
① Confirm that all sequences are homologous

② Adjust gap creation and extension penalties as needed to optimize the alignment

③ Restrict phylogenetic analysis to regions of the multiple sequence alignment for which data are available for all taxa (delete columns having incomplete data)



DNA mutation



Choosing substitution models

- The simplest approach to measuring distances between sequences is to align pairs of sequences, and then to count the number or percentage of differences.
- For an alignment of length N with m sites at which there are differences, the **observed distance** between the two sequences is called **p-distance**:

$$D = m/N$$

- For example, if an alignment of sequences A and B is 20 nucleotides long and 6 pairs are found to be different, the sequences differ by 30%, or have an observed distance 0.3.

Multiple Substitutions

- In reality, the observed number of substitutions often **underestimate** the true evolutionary events that actually occurred.
- When a mutation is observed as A replaced by C, the nucleotide A may have **actually undergone a number of intermediate steps** to become C, such as

$$A \rightarrow T \rightarrow G \rightarrow C.$$

- When the same nucleotide is observed, a **back mutation** could have occurred when a mutated nucleotide reverted back to the original nucleotide, such as

$$G \rightarrow C \rightarrow G.$$

- An identical nucleotide observed in the alignment could be due to **parallel mutations**, such as

$$G \rightarrow C$$

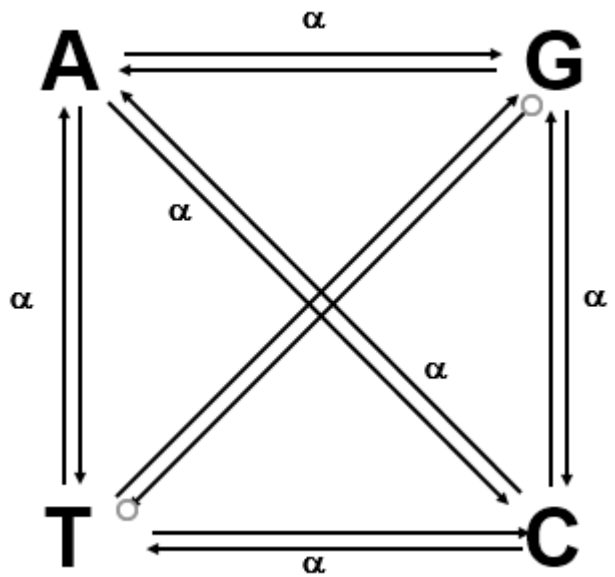
$$G \rightarrow C$$

Nucleotide substitution models

- Statistical models are needed to infer the true evolutionary distances between sequences.
  - Jukes-Cantor model
  - Kimura model

Jukes–Cantor model

- The model assumes that all nucleotides are substituted with equal probability a.



Jukes–Cantor model

- The evolutionary distance between sequences A and B is:

$$d_{\mathrm{AB}} = -(3/4)\,\ln[1 - (4/3)\,p_{\mathrm{AB}}]$$

where p_{AB} is the observed distance.

- For example, if an alignment of sequences A and B is 20 nucleotides long and 6 pairs are found to be different, the corrected evolutionary distance is:

$$d_{\mathrm{AB}} = -3/4\,\ln[1 - (4/3 \times 0.3)] = 0.38$$

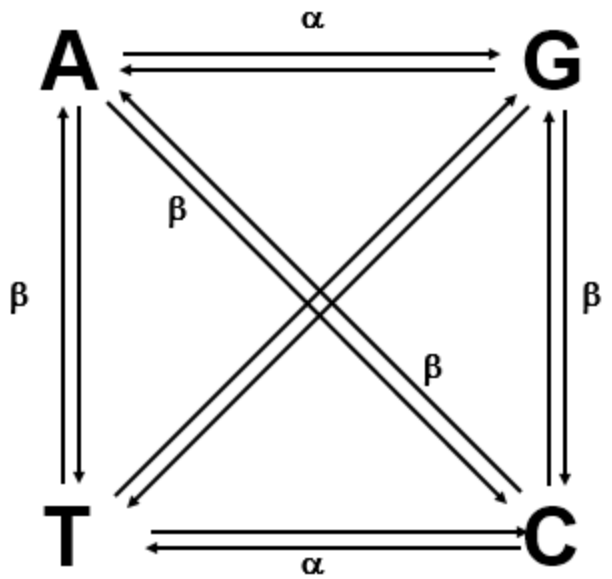
Jukes–Cantor model

- The Jukes–Cantor model can only handle reasonably closely related sequences.
- For distantly related sequences, the correction can become too large to be reliable.
- If two DNA sequences have 25% similarity, p_{AB} is 0.75. This leads the log value to be infinitely large.

$$d_{AB} = -(3/4)\ \ln[1 - (4/3)\, p_{AB}]$$

Kimura model

- This is a more sophisticated model in which mutation rates for transitions and transversions are assumed to be different, which is more realistic——transitions occur more frequently than transversions $\alpha > \beta$.



Kimura model

- The evolutionary distance between sequences A and B is:

$$d_{AB} = -(1/2)\ \ln(1 - 2\,p_{ti} - p_{tv}) - (1/4)\ \ln(1 - 2\,p_{tv})$$

- where p_{ti} is the observed frequency of transition, and p_{tv} the frequency of transversion.

- For example, sequences A and B differ by 30%. If 20% of changes are a result of transitions and 10% of changes are a result of transversions, the corrected evolutionary distance is:

$$d_{AB} = -1/2\ \ln(1 - 2 \times 0.2 - 0.1) - 1/4\ \ln(1 - 2 \times 0.1) = 0.40$$

Protein substitution models

- For protein sequences, the evolutionary distances from an alignment can be corrected using a PAM or other amino acid **substitution matrix** whose construction already takes into account the multiple substitutions.

Determining a tree building method

- Distance-based tree building methods
  - UPGMA
  - Neighbor-joining
- Character-based tree building methods
  - Maximum parsimony
  - Maximum likelihood

Distance-based tree building methods

- Assume that:
  - for any pair of sequences we have an estimation of evolutionary distance between them;
  - tree branches are additive —— the evolutionary distance between two taxa equals to the sum of all branch lengths connecting them.
- Goal: construct a tree which best approximates these pairwise evolutionary distances.

Distance-based tree building methods

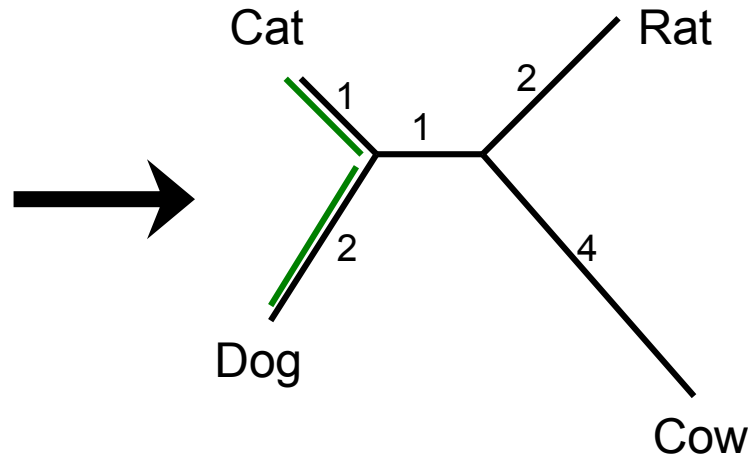
- Advantage
  - Make use of a number of substitution models to correct distances
  - Computationally fast
  - Single “best tree” found
- Disadvantage
  - The actual sequence information is lost when all the sequence variation is reduced to a single value

How to get distance matrix

1. Commonly obtained the observed distances from sequence alignments;
2. Correct the observed distances for multiple substitutions by using a substitution model, e.g. Jukes–Cantor model.

Perfectly “tree-like” distances

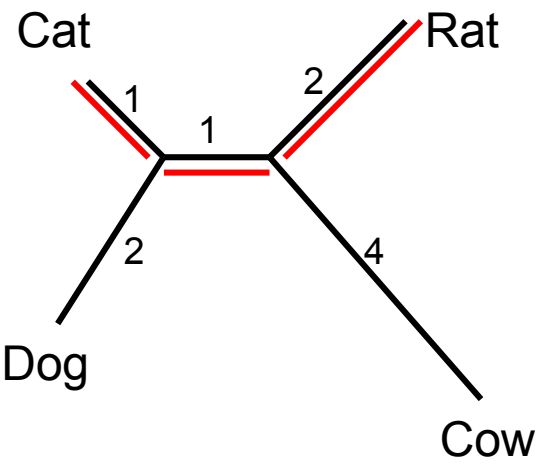
	Cat	Dog	Rat
Dog	3		
Rat	4	5	
Cow	6	7	6



- In a tree there is a unique path between any two nodes.
- **Assume that:** tree branches are additive —— the evolutionary distance between two taxa equals to the sum of all branch lengths connecting them.
  e.g. the distance between cat and dog is: 1 + 2 = 3

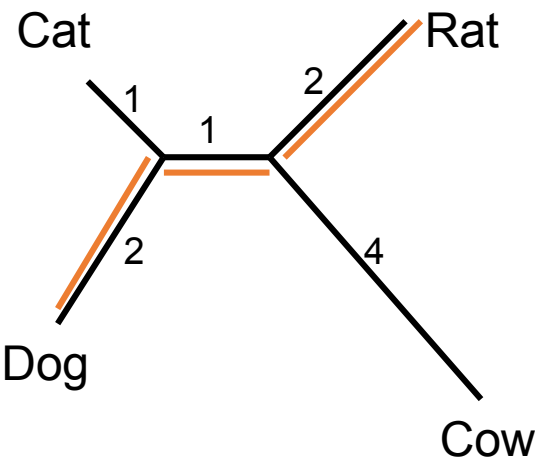
Perfectly “tree-like” distances

	Cat	Dog	Rat
Dog	3		
Rat	4	5	
Cow	6	7	6



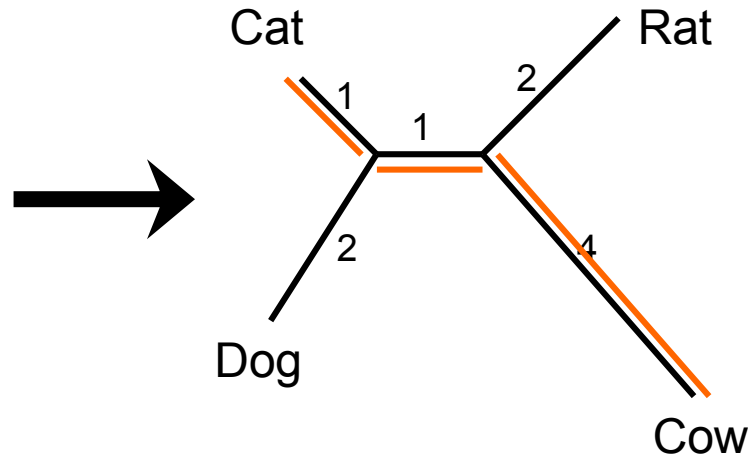
Perfectly “tree-like” distances

	Cat	Dog	Rat
Dog	3		
Rat	4	5	
Cow	6	7	6



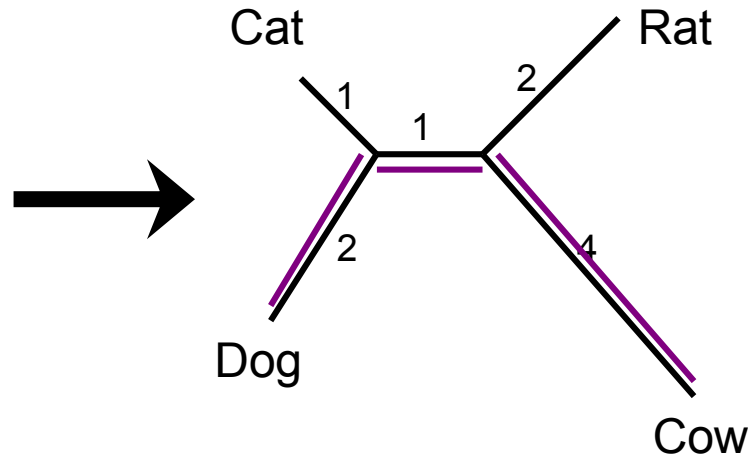
Perfectly “tree-like” distances

	Cat	Dog	Rat
Dog	3		
Rat	4	5	
Cow	6	7	6



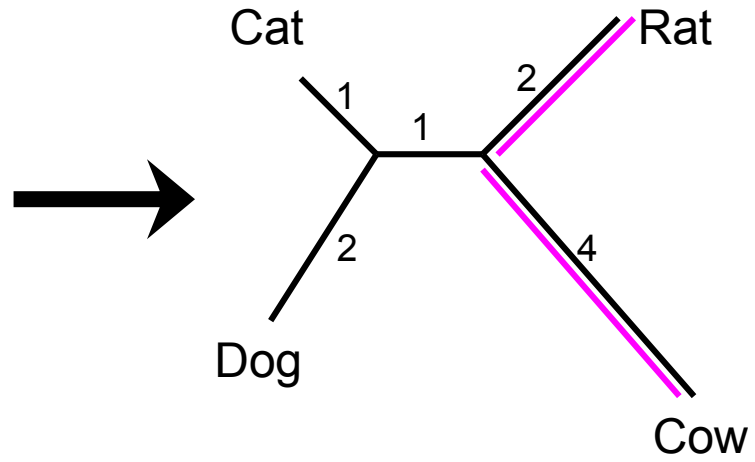
Perfectly “tree-like” distances

	Cat	Dog	Rat
Dog	3		
Rat	4	5	
Cow	6	7	6



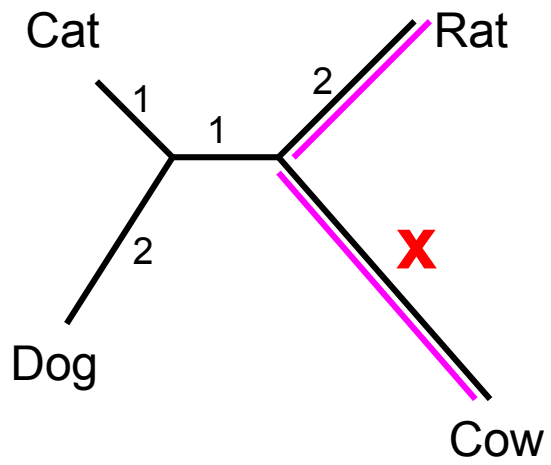
Perfectly “tree-like” distances

	Cat	Dog	Rat
Dog	3		
Rat	4	5	
Cow	6	7	6

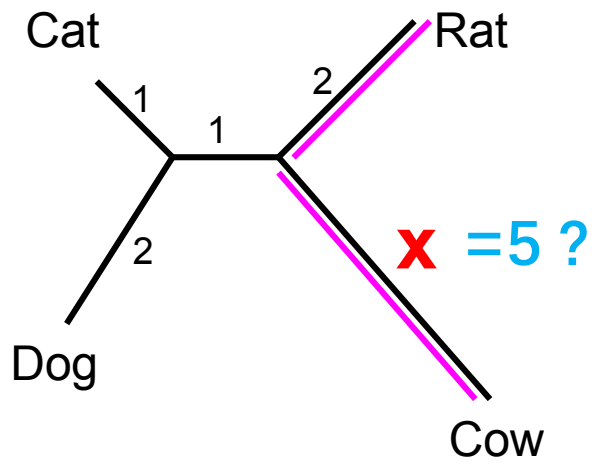


The tree exactly preserves the distance in the distance matrix.

	Cat	Dog	Rat
Dog	3		
Rat	4	5	
Cow	6	7	7

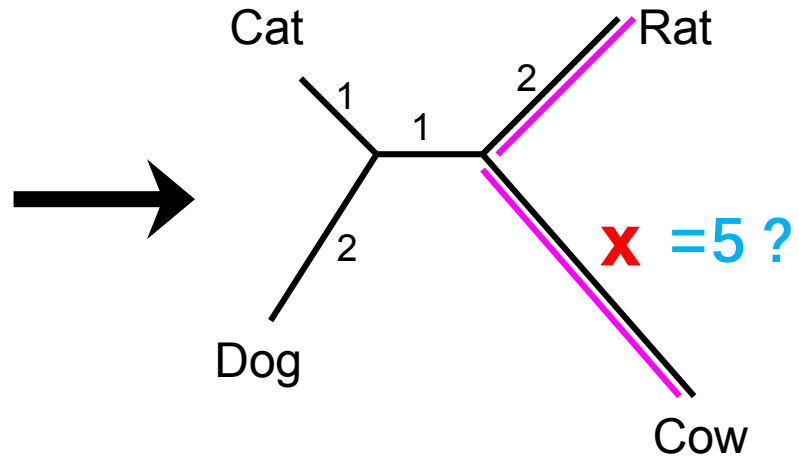


	Cat	Dog	Rat
Dog	3		
Rat	4	5	
Cow	6	7	7



Real matrices are almost never additive

	Cat	Dog	Rat
Dog	3		
Rat	4	5	
Cow	6	7	7



- There is no way to add x to the tree and preserve the distances in the matrix.
- However, we can construct a tree which best approximates these distances.

UPGMA trees

- Unweighted Pair Group Method with Arithmetic mean (UPGMA)
- Tree-building steps：

① compute the pairwise distances of all the sequences;

② find the two sequences with the smallest pairwise distance; join them into a new cluster;

③ create a reduced distance matrix by treating the new cluster as a single taxon;

④ repeat steps 2 and 3 until all sequences are placed on the tree.

UPGMA Example

Goal: reconstruct the phylogeny of four taxa A, B, C and D.

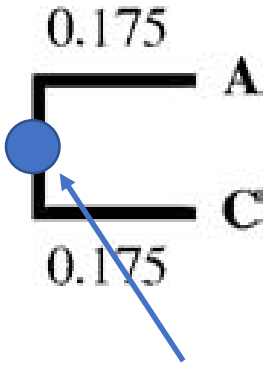
Step 1: compute the pairwise distances of all the sequences.

	A	B	C
B	0.40		
C	0.35	0.45	
D	0.60	0.70	0.55

UPGMA Example

Step 2: find the two sequences with the smallest pairwise distance. cluster and place a midpoint between them

	A	B	C
B	0.40		
C	0.35	0.45	
D	0.60	0.70	0.55



The branch length for A to the **midpoint** is AC/2 = 0.35/2 = 0.175.

UPGMA Example

Step 3: create a reduced distance matrix.

Because A and C are joined into a cluster, they are treated as one new composite taxon, which is used to create a reduced matrix.

We need to calculate the distance of the new taxon to every other taxa.

	A-C	B
B	$\frac{0.4 + 0.45}{2} = 0.425$	
D	$\frac{0.55 + 0.6}{2} = 0.575$	0.70

The distance of B to A–C is (AB + BC)/2; and that of D to A–C is (AD + CD)/2.

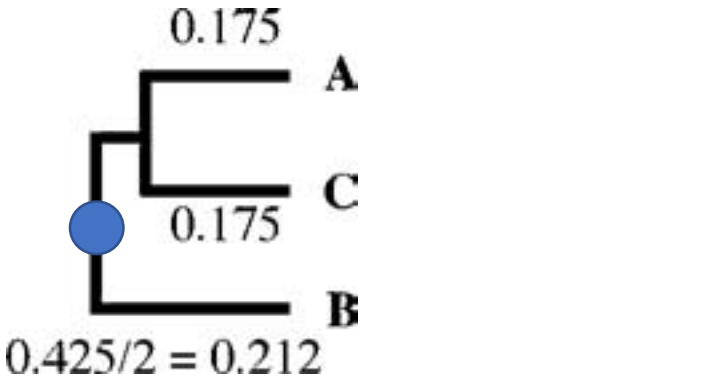
UPGMA Example

Step 4: repeat steps 2 and 3 until all sequences are placed on the tree.

Find the two sequences with the smallest distance in the reduced distance matrix.

Cluster and place a midpoint between them.

	A-C	B
B	$\frac{0.4 + 0.45}{2} = 0.425$	
D	$\frac{0.55 + 0.6}{2} = 0.575$	0.70



UPGMA Example

Step 4: repeat steps 2 and 3 until all sequences are placed on the tree.

B and A–C are grouped and treated as a single taxon, which allows the distance matrix to reduce further into only two taxa.

	B-A-C
D	$\frac{0.7 + 0.6 + 0.55}{3} = 0.617$

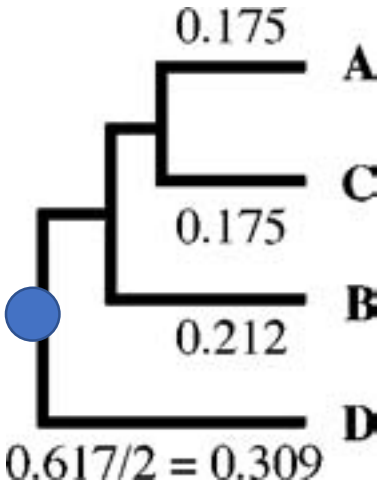
UPGMA Example

Step 4: repeat steps 2 and 3 until all sequences are placed on the tree.

Join D and B-A-C together and place a midpoint between them.

The iteration stops as all taxa are placed on the tree.

	B-A-C
D	$\frac{0.7 + 0.6 + 0.55}{3} = 0.617$

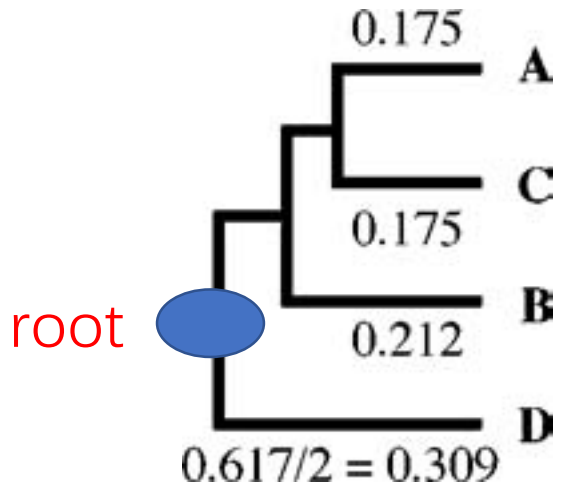


UPGMA Example

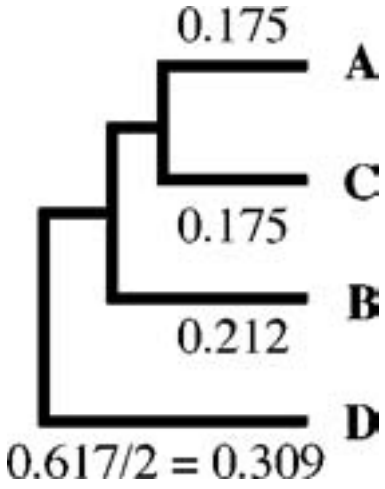
Step 4: repeat steps 2 and 3 until all sequences are placed on the tree.

	B-A-C
D	$\frac{0.7 + 0.6 + 0.55}{3} = 0.617$

The last sequence added is considered the outgroup producing a rooted tree.



UPGMA Example

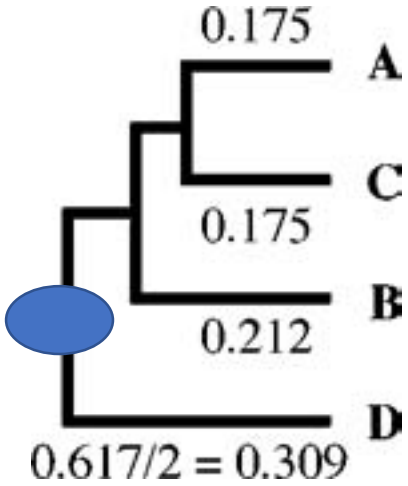


	A	B	C
B	0.42		
C	0.35	0.42	
D	0.62	0.62	0.62

	A	B	C
B	0.40		
C	0.35	0.45	
D	0.60	0.70	0.55

The **estimated distances** do not match the **actual evolutionary distances** shown, which illustrates the failure of UPGMA to precisely reflect the experimental observation.

UPGMA trees



- UPGMA is a simple approach for making trees.
- An UPGMA tree is always rooted.
- An assumption of the algorithm is that all taxa evolve at a constant rate (**molecular clock assumption**) and that they are equally distant from the root. If there are unequal substitution rates, the tree may be wrong.
- It is less accurate than the neighbor-joining approach.

Neighbor Joining trees

- The Neighbor Joining method is a method for re-constructing phylogenetic trees, and computing the lengths of the branches of this tree.
- Unlike UPGMA
  - doesn't make molecular clock assumption (i.e. does not assume all lineages evolve at the same rate)
  - produces unrooted trees
- Does assume the tree branches are additivity: distance between pair of leaves is sum of lengths of edges connecting them

Neighbor Joining trees

- Neighbor joining produce unrooted trees
  - Choose one distant sequence as an outgroup to find the root
- The output tree topology isn’t guaranteed to be correct. However, neighbor joining often constructs the correct tree topology anyway.
- It's computationally efficient which makes it practical for analyzing large data sets
- One of the most widely used methods for phylogenic tree construction

Character-based tree building methods

- Character-based methods are based directly on the sequence characters rather than on pairwise distances.
- They count mutational events accumulated on the sequences and may therefore avoid the loss of information when characters are converted to distances.

Maximum parsimony

- **Parsimony assumption (lowest cost)**: a tree with **minimal changes** is likely to be a good estimate of the true tree.
- Parsimony tree building works by searching for all possible tree topologies and reconstructing ancestral sequences that require the minimum number of changes to evolve to the current sequences.
- This method finds the **unrooted** tree that explains the considered sequences with the minimum number of substitutions.

Maximum parsimony

- Count the mutational events occurred in the multiple sequence alignment.
- To save computing time, only a small number of sites that have the richest phylogenetic information are used in tree determination. These sites are the so-called **informative sites**.
- Informative sites: sites that have at least two different kinds of characters, each occurring at least twice.

taxa \ sites	1	2	3	4	5	6	7	8
I	A	A	T	T	A	G	C	T
II	G	G	T	C	G	T	A	G
III	A	A	T	G	C	G	C	T
IV	A	G	T	A	A	G	C	A
V	A	C	T	T	C	G	C	G
VI	A	C	A	T	G	G	C	A

Maximum parsimony

- Tree-building steps：

① Make all possible trees

② Identify informative sites

③ Compute the minimum number of substitutions at each informative site for a given tree topology.

④ The total number of changes at all informative sites are summed up for each possible tree topology.

⑤ The tree that has the smallest number of changes is chosen as the best tree.

four homologous sequences

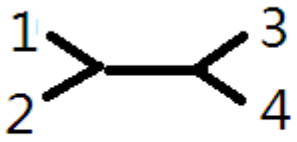
seq	p1	p2	p3	p4	p5	p6
1	G	G	G	G	G	G
2	G	G	G	A	G	T
3	G	G	A	T	A	G
4	G	A	T	C	A	T

For four taxa, how many possible unrooted trees?

four homologous sequences

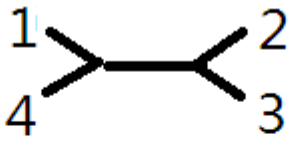
seq	p1	p2	p3	p4	p5	p6
1	G	G	G	G	G	G
2	G	G	G	A	G	T
3	G	G	A	T	A	G
4	G	A	T	C	A	T

Tree 1



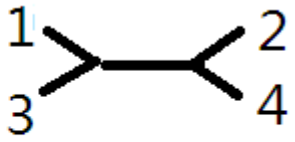
((1,2)(3,4))

Tree 2



((1,4)(2,3))

Tree 3



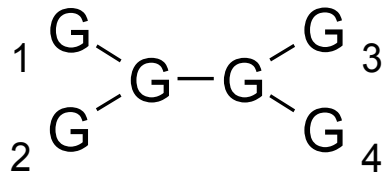
((1,3)(2,4))

For four taxa, there are three unrooted trees.

four homologous sequences

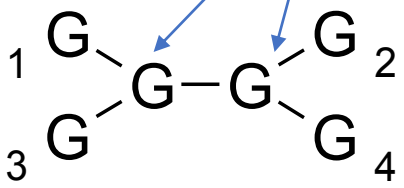
seq	p1	p2	p3	p4	p5	p6
1	G	G	G	G	G	G
2	G	G	G	A	G	T
3	G	G	A	T	A	G
4	G	A	T	C	A	T

Constant site

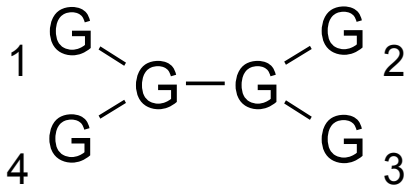


((1,2)(3,4))

ancestral character



((1,3)(2,4))



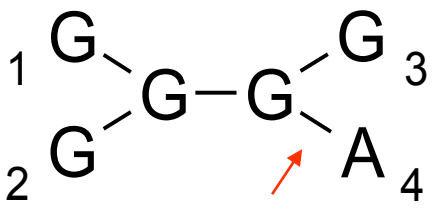
((1,4)(2,3))

Constant site has the same state in all taxa and is obviously useless in evaluating the various topologies.

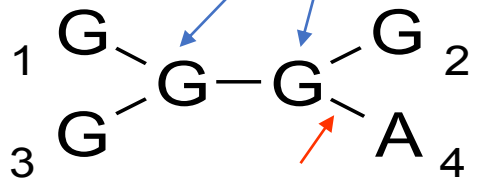
seq	p1	p2	p3	p4	p5	p6
1	G	G	G	G	G	G
2	G	G	G	A	G	T
3	G	G	A	T	A	G
4	G	A	T	C	A	T

Noninformative site

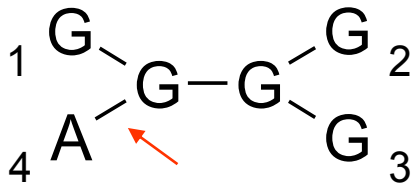
Assign **ancestral characters** that involve **minimum number of substitutions**



((1,2)(3,4))



((1,3)(2,4))

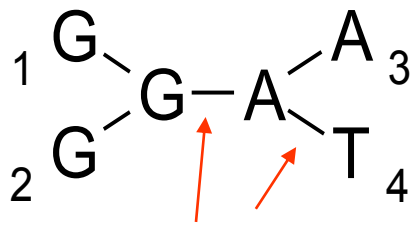


((1,4)(2,3))

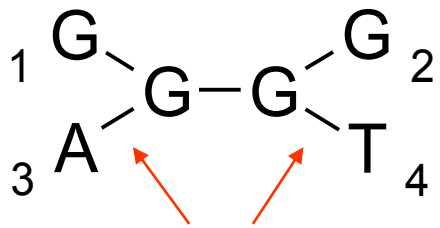
The sites that have changes occurring only once are noninformative sites because they can be explained by all tree topologies.

seq	p1	p2	p3	p4	p5	p6
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

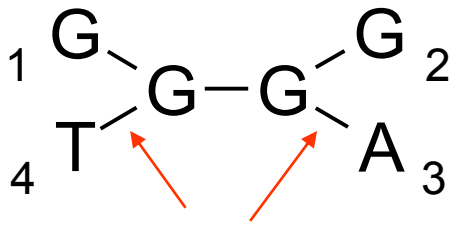
Noninformative site



((1,2)(3,4))



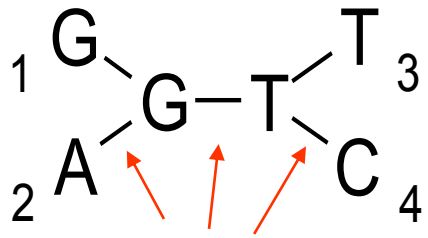
((1,3)(2,4))



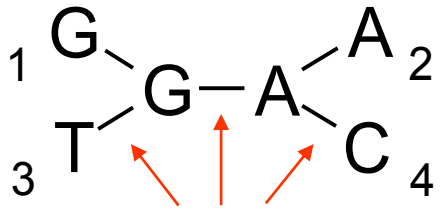
((1,4)(2,3))

seq	p1	p2	p3	p4	p5	p6
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

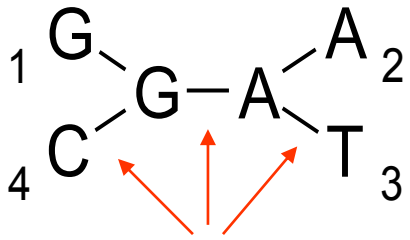
Noninformative site



((1,2)(3,4))



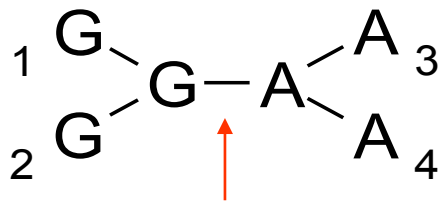
((1,3)(2,4))



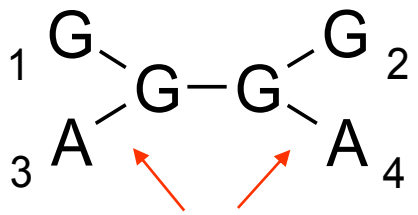
((1,4)(2,3))

seq	p1	p2	p3	p4	p5	p6
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

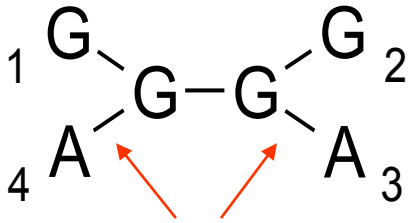
Informative site



((1,2)(3,4))



((1,3)(2,4))

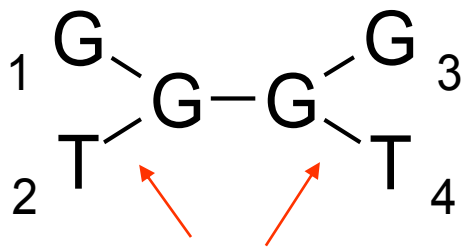


((1,4)(2,3))

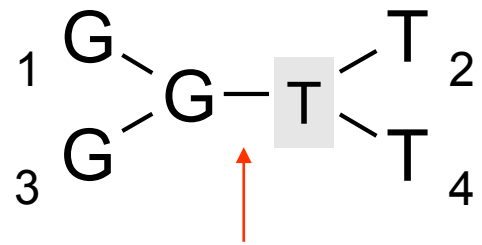
Informative site: a site that has at least two different kinds of characters, each occurring at least twice.

seq	p1	p2	p3	p4	p5	p6
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

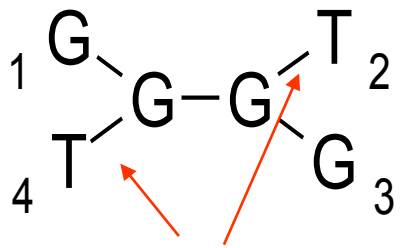
Informative site



((1,2)(3,4))



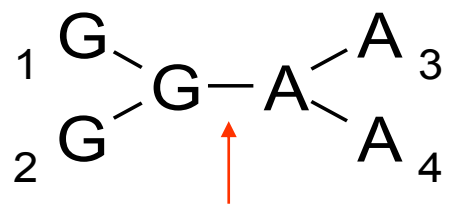
((1,3)(2,4))



((1,4)(2,3))

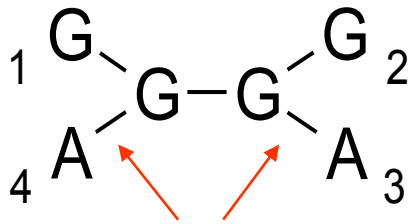
p5:

Tree 1



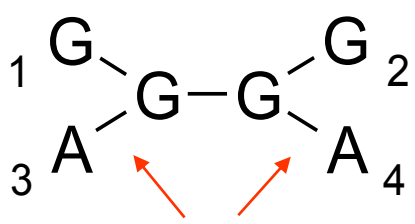
((1,2)(3,4))

Tree 2



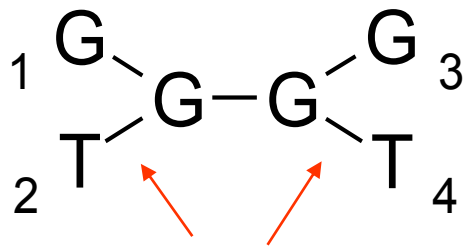
((1,4)(2,3))

Tree 3

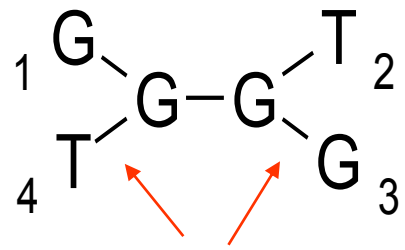


((1,3)(2,4))

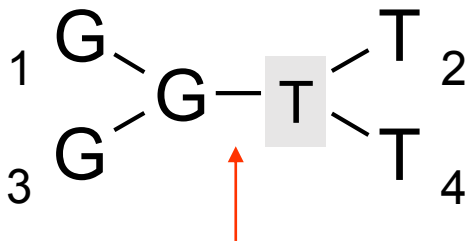
p6:



((1,2)(3,4))



((1,4)(2,3))

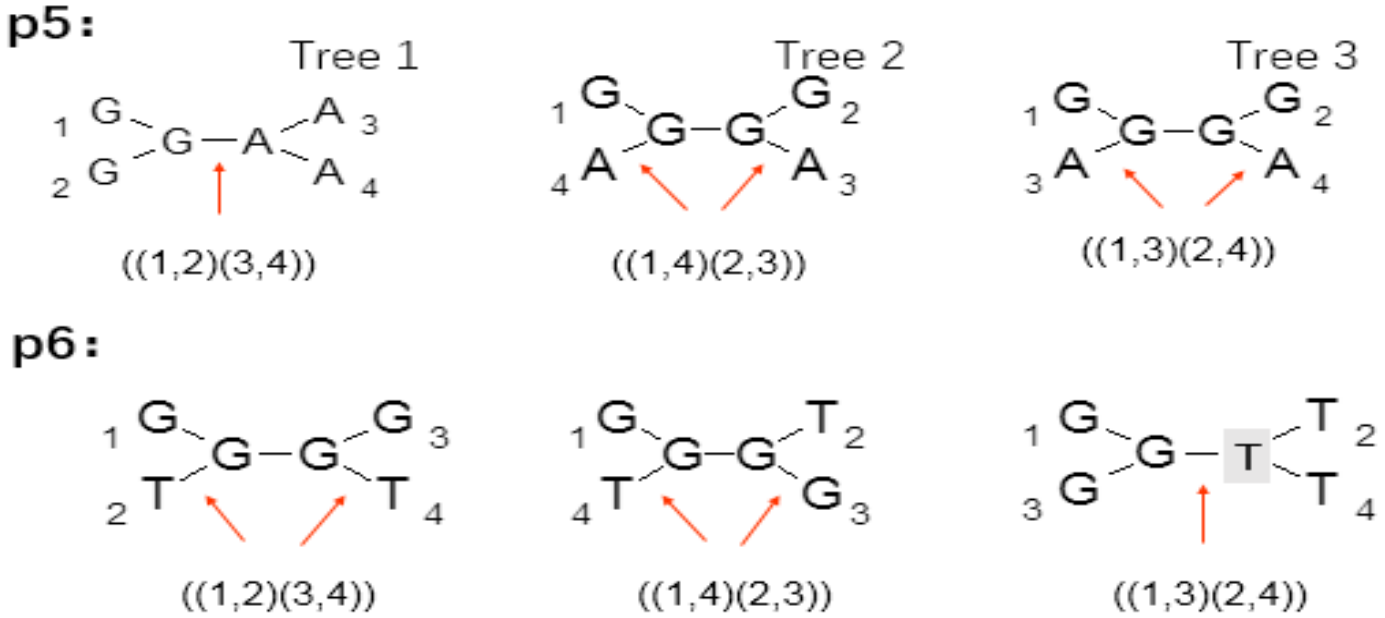


((1,3)(2,4))

For position 5, the minimum number of substitutions in Tree 1 is 1, in Trees 2 & 3 is 2.

For position 6, the minimum number of substitutions in Trees 1 & 2 is 2, in Tree 3 is 1.

Maximum parsimony example



- The total number of changes in Tree 1 is: 1+2 = 3
- The total number of changes in Tree 2 is: 2+2 = 4
- The total number of changes in Tree 3 is: 2+1 = 3

Trees 1 and 3 are the best trees (the most-parsimonious trees) having the smallest number of changes.

Maximum parsimony

➢Pros

- It tends to produce more accurate trees than the distance-based methods when sequence divergence is low because this is the circumstance when the parsimony assumption of rarity in evolutionary changes holds true.

➢Cons

- Often exists a number of equally most-parsimonious trees.
- MP does not employ substitution models to correct for multiple substitutions. This drawback can become prominent when dealing with divergent sequences.
- MP only considers informative sites, and ignores other sites. Consequently, certain phylogenetic signals may be lost.
- MP is also slow compared to the distance methods.

Maximum likelihood tree

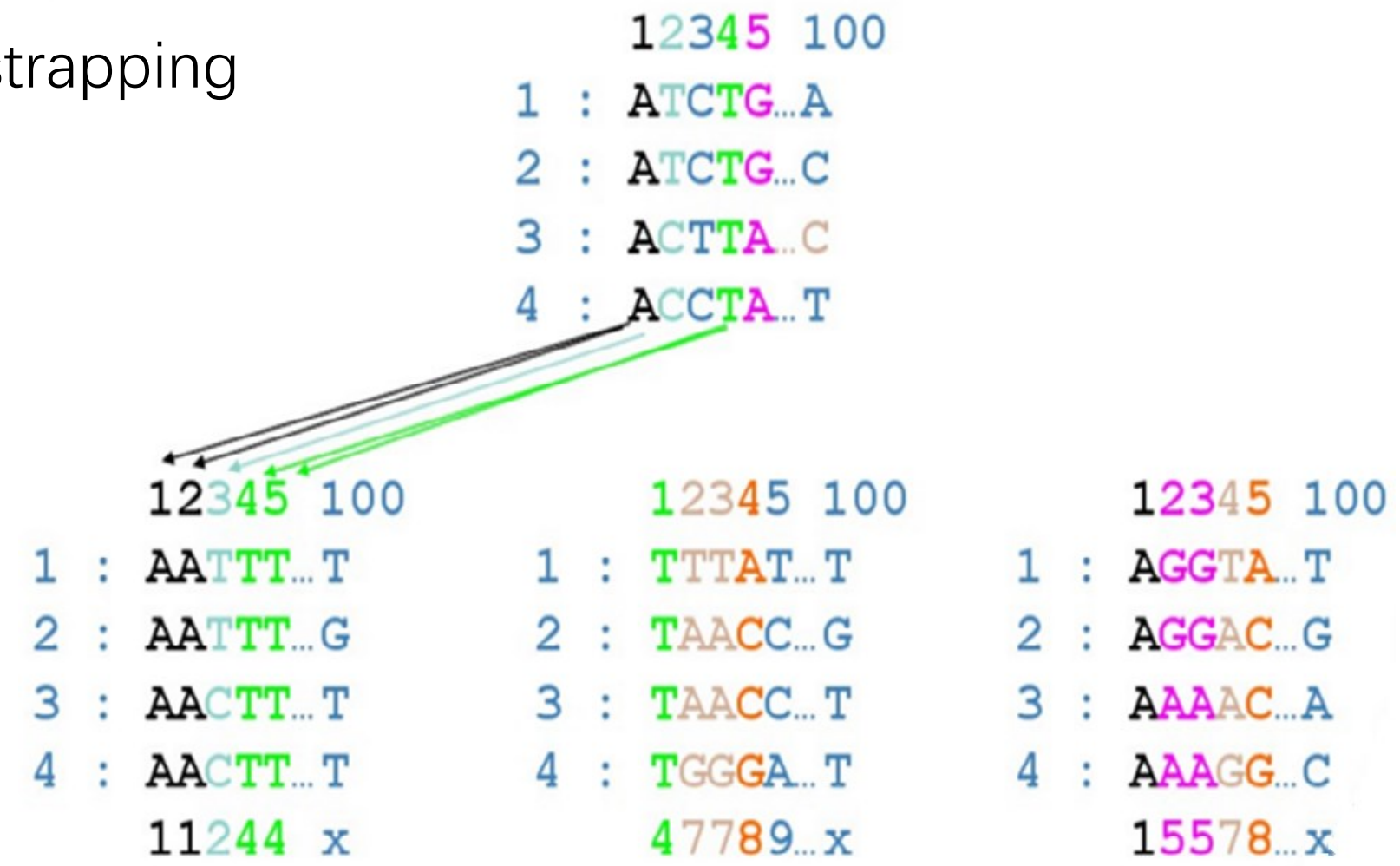
- Use probabilistic models to choose a best tree that has the highest probability or likelihood of reproducing the observed data.
- Maximum likelihood is computationally intensive as it evaluates:
  - all possible ancestral states
  - at all positions, not just informative sites
  - in all possible tree topologies
- It tends to produce more accurate trees than the Maximum parsimony.

Validation of the phylogenetic prediction

- After phylogenetic tree construction, the next step is to statistically evaluate the reliability of the inferred phylogeny —— how reliable the tree or a portion of the tree is.
- There are two ways to estimate the degree of confidence of a specific phylogenetic reconstruction. And it is recommended to use both:

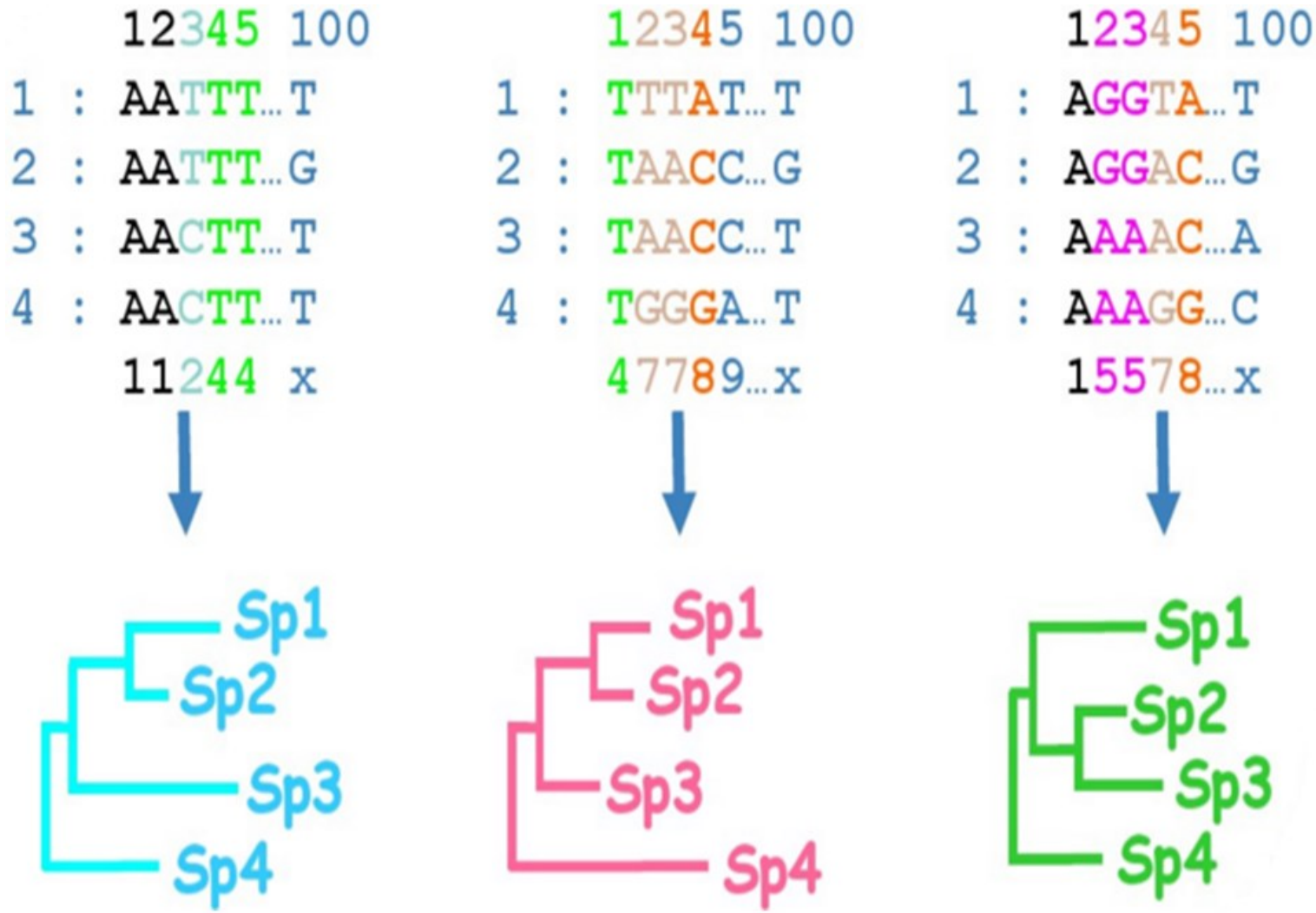
1. Comparison of topologies obtained with different methods for the construction of the tree, possibly one based on the distance and the other on the characters.
2. Statistical estimation of the reliability of the results obtained through random sampling the considering data (bootstrapping).

Bootstrapping



- Assume you have a multiple alignment of length 100.
- Repeat, say 3 times the following process:

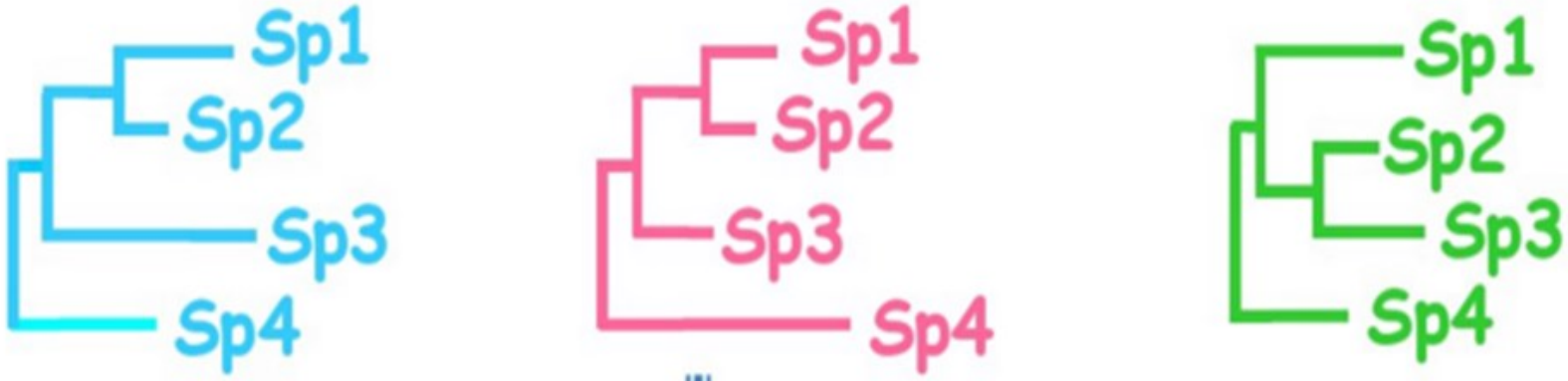
  -**Select randomly with replacement** 100 columns from the original alignment to produce a randomized alignment



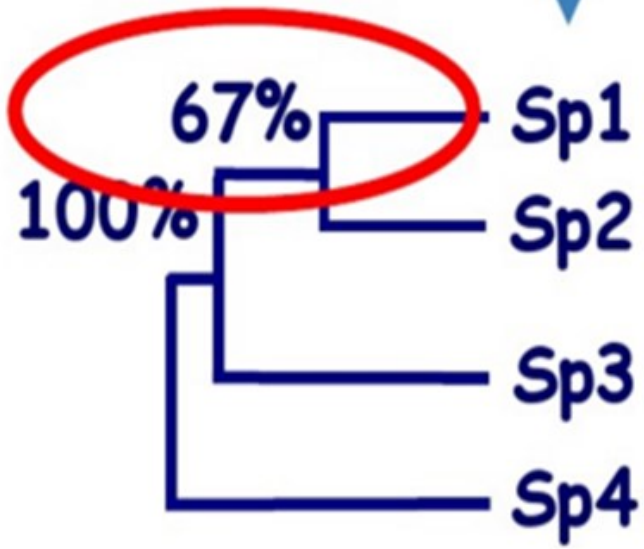
bootstrap trees

- Assume you have a multiple alignment of length 100.
- Repeat, say 3 times the following process:
  - **Select randomly with replacement** 100 columns of the alignment to produce a randomized alignment
  - The original tree-building method is applied to this randomized alignment

bootstrap trees



original tree



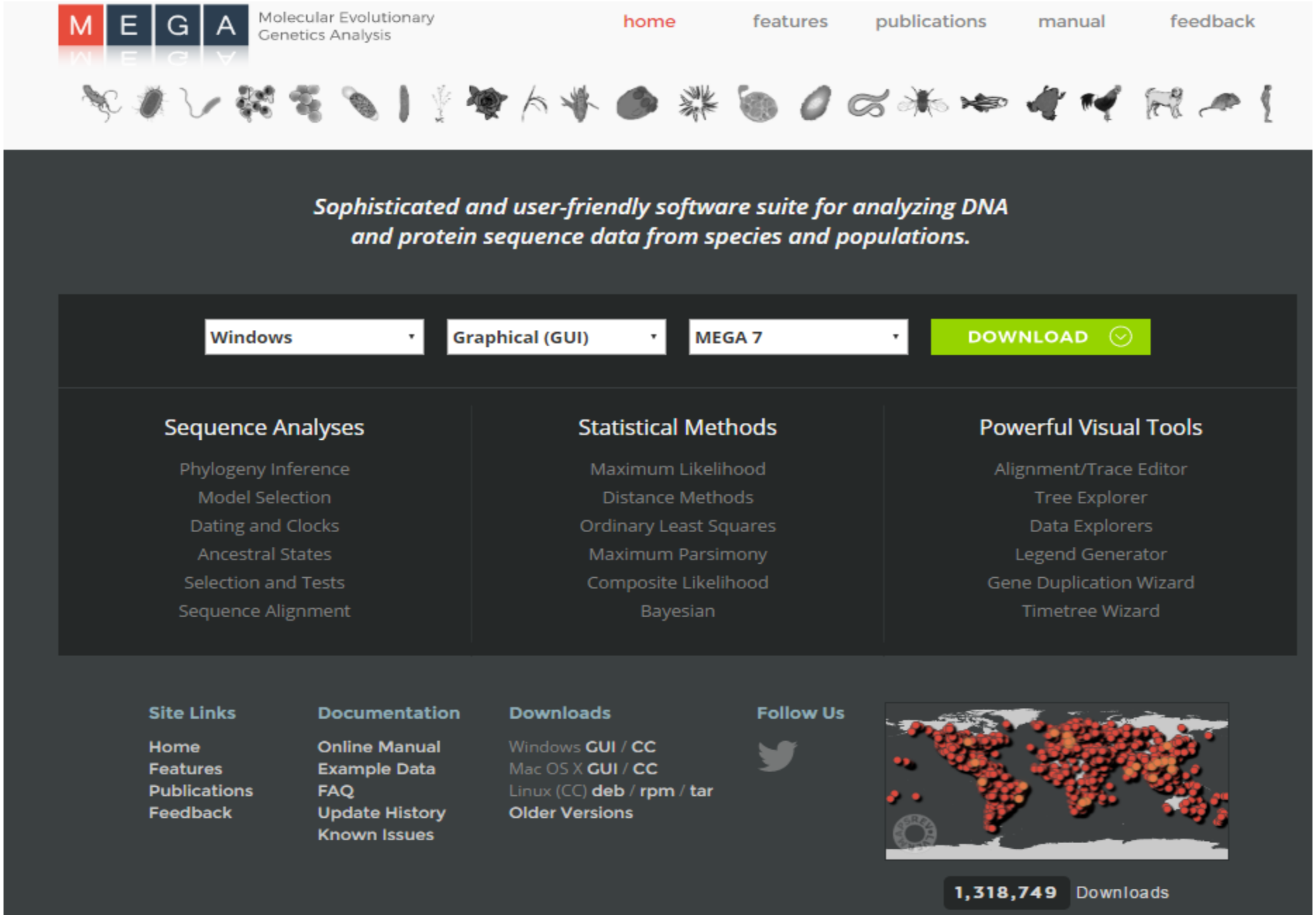
sp1 and sp2 occurred as a clade in 2 of the 3 bootstrap trees, giving 67% confidence.

- Assign **a confidence value** to each **clade** in the original tree.
- Calculate the proportions of bootstrap trees agreeing with the original tree —— bootstrap confidence values .
- “Agreeing” here refers to the topology of the tree and not to the length of its branches.

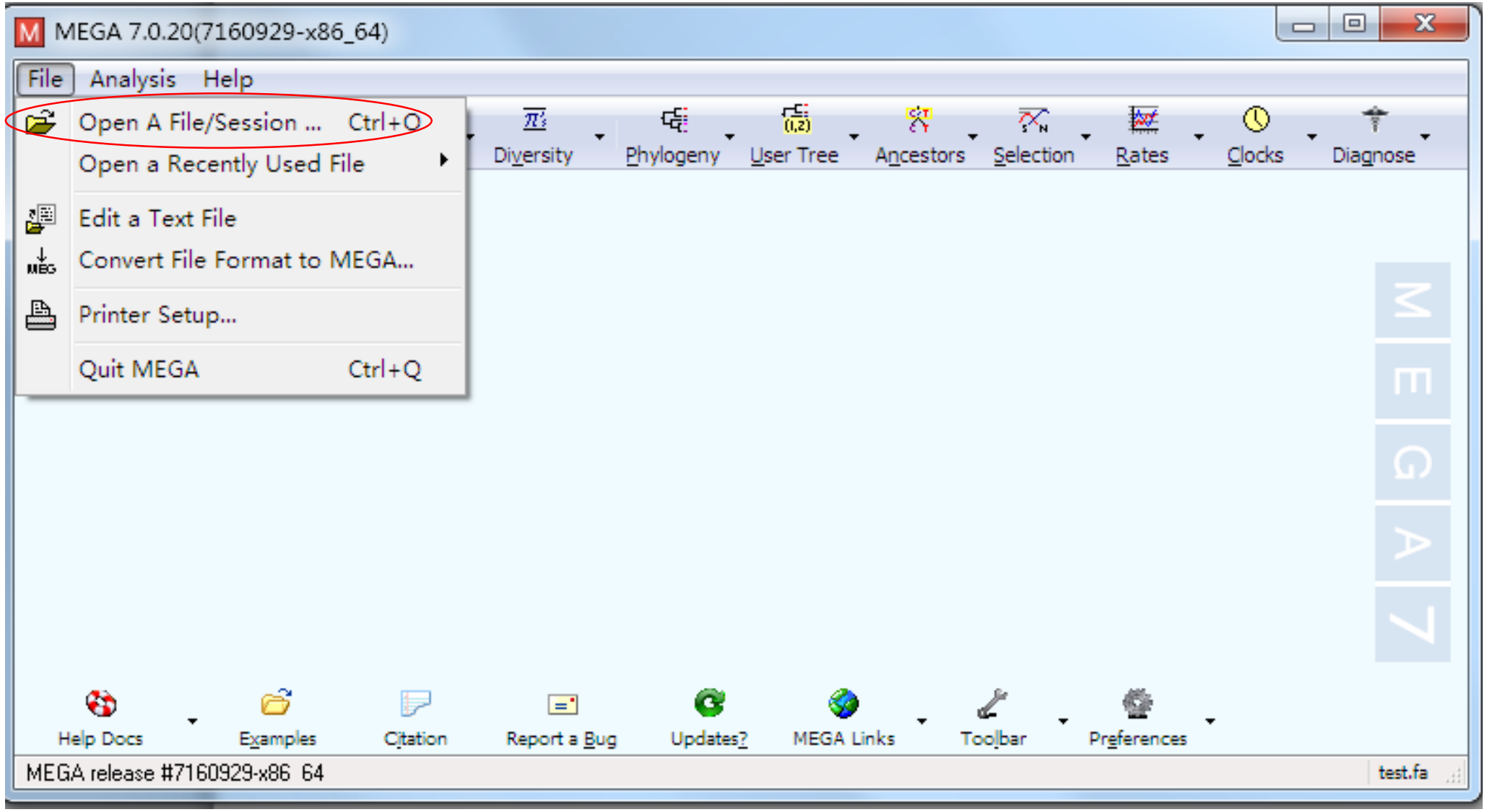
Bootstrapping

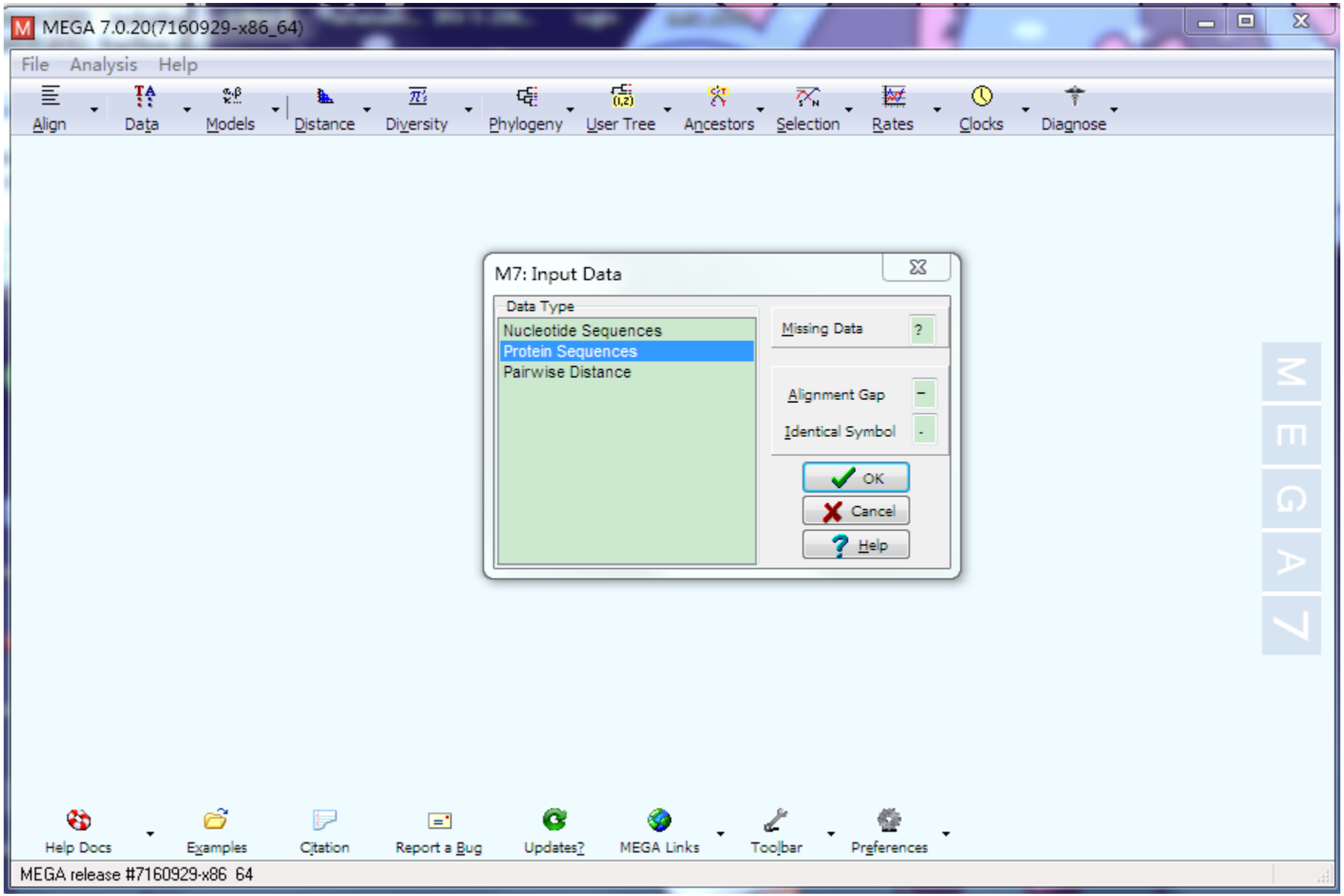
- The bootstrap method is a popular statistical technique used to assess the robustness of the inferred phylogenetic trees.
- The bootstrap test provides a measure for evaluating the confidence levels of the tree topology.
- It is generally recommended that a phylogenetic tree should be bootstrapped 500 to 1,000 times.
- Trusted clades: 70% or better bootstrap value.

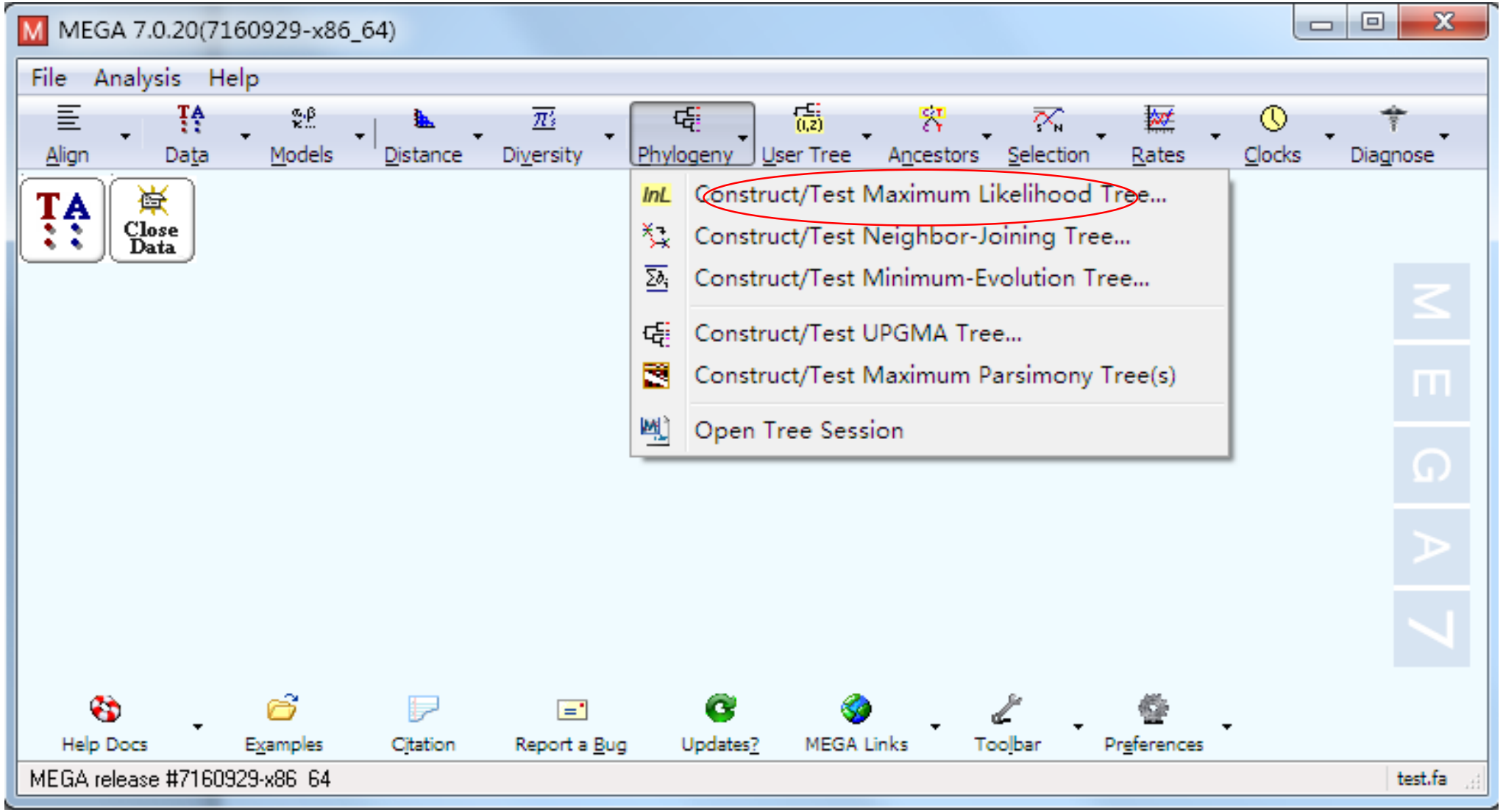
http://www.megasoftware.net/

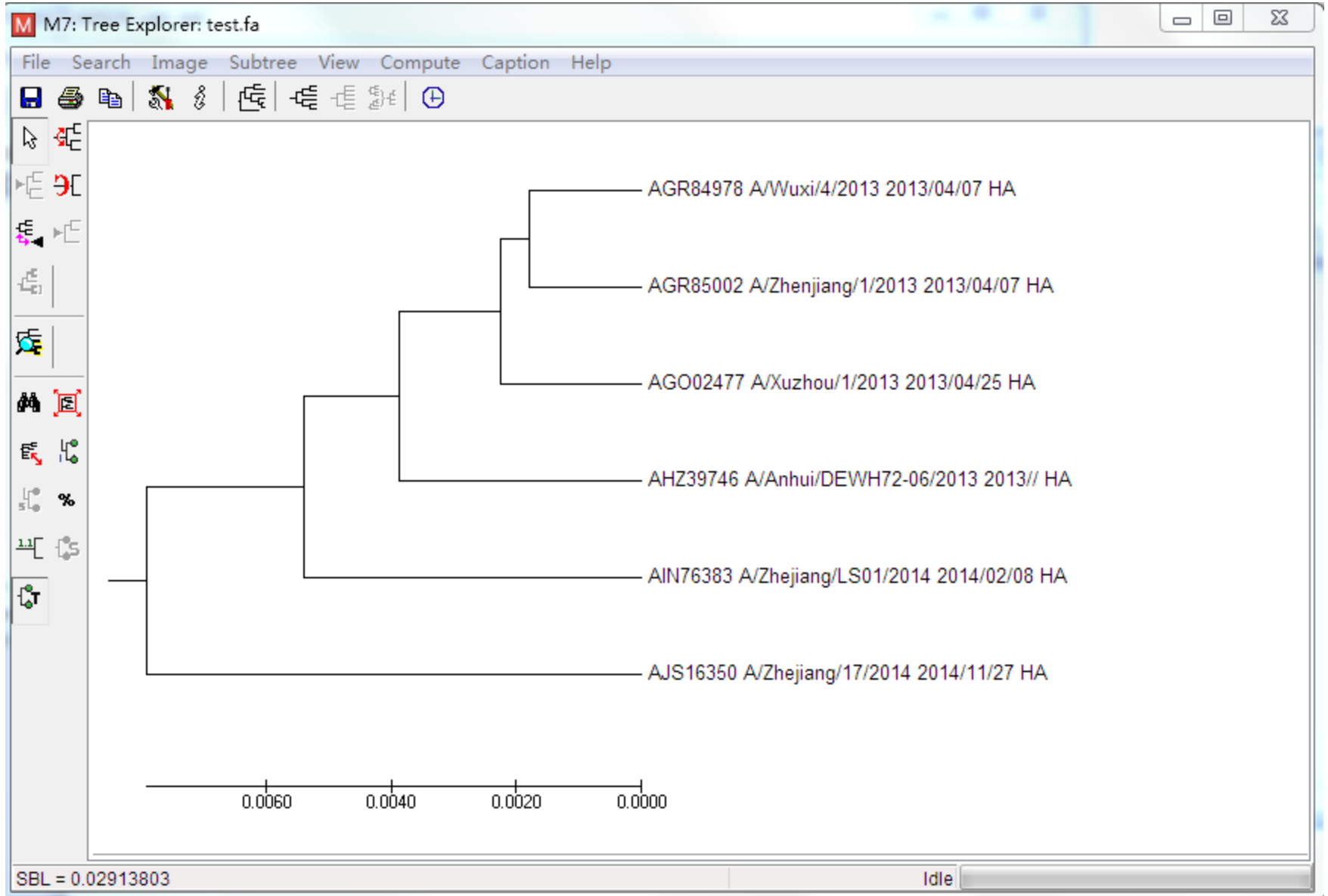


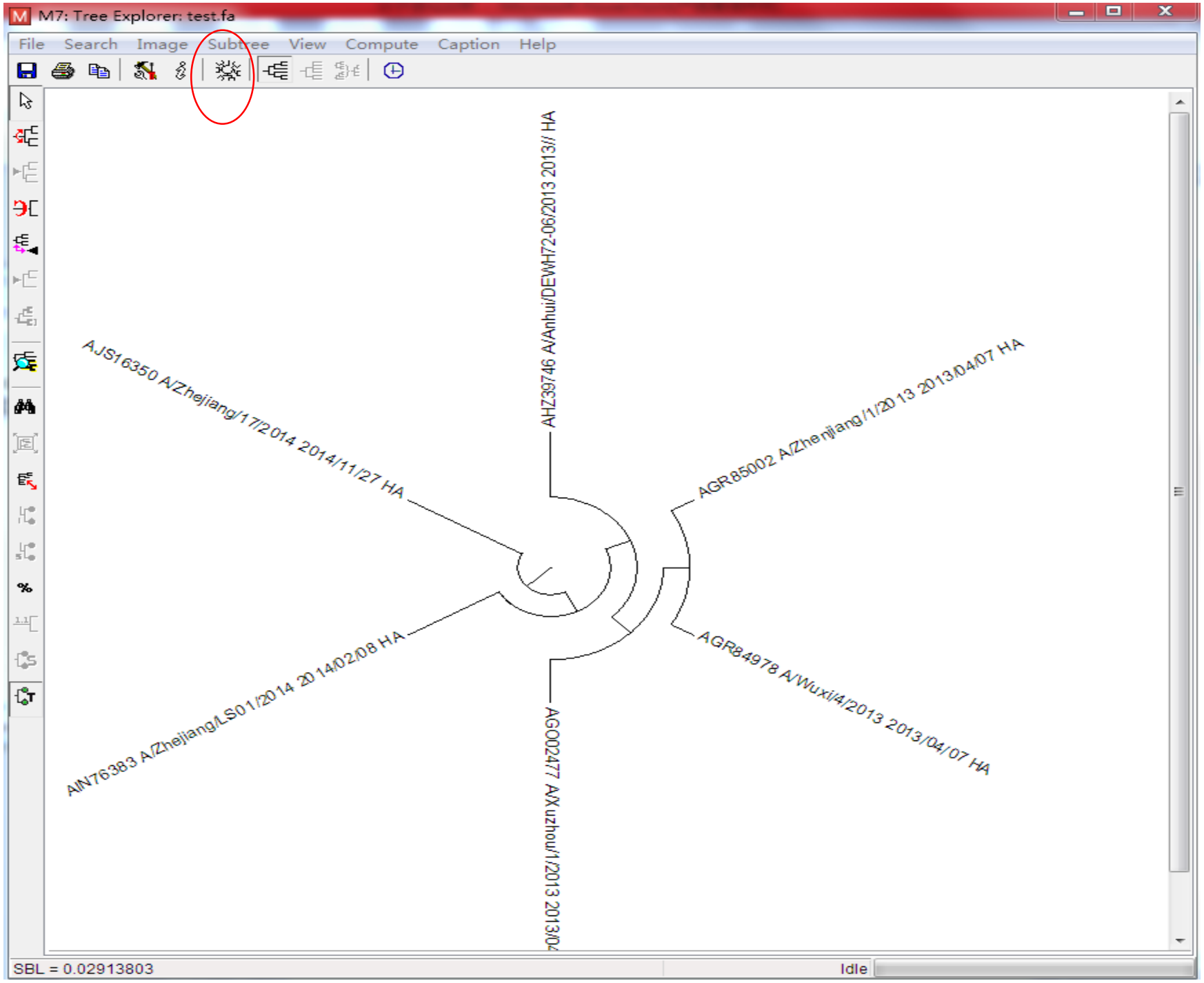
```
>AHZ39746 A/Anhui/DEWH72-06/2013 2013// HA
MNTQILVFALIAIIPTNADKICLGHHAVSNGTKVNTLTERGVEVVNATETVERTNIPRICSKGKRTVDLG
QCGLLGTITGPPQCDQFLEFSADLIIERREGSDVCYPGKFVNEEALRQILRESGGIDKEAMGFTYSGIRT
NGATSACRRSGSSFYAEMKWLLSNTDNAAFPQMTKSYKNTRKSPALIVWGIHHSVSTAEQTKLYGSGNKL
VTVGSSNYQQSFVPSPGERPQVNGLSGRIDFHWLMLNPNDTVTFSFNGAFIAPDRASFLRGKSMGIQSGV
QVDANCEGDCYHSGGTIISNLPFQNIDSRAVGKCPRYVKQRSLLLATGMKNVPEIPKGRGLFGAIAGFIE
NGWEGLIDGWYGFRHQNAQGEGTAADYKSTQSAIDQITGKLNRLIEKTNQQFELIDNEFNEVEKQIGNVI
NWTRDSITEVWSYNAELLVAMENQHTIDLADSEMDKLYERVKRQLRENAEEDGTGCFEIFHKCDDDCMAS
IRNNTYDHSKYREEAMQNRIQIDPVKLSSGYKDVILWFSFGASCFILLAIVMGLVFICVKNGNMRCTICI
>AGR84978 A/Wuxi/4/2013 2013/04/07 HA
MNTQILVFALIAIIPTNADKICLGHHAVSNGTKVNTLTERGVEVVNATETVERTNIPRICSKGKKTVDLG
QCGLLGTITGPPQCDQFLEFSADLIIERREGSDVCYPGKFVNEEALRQILRESGGIDKEAMGFTYSGIRT
NGATSACRRSGSSFYAEMKWLLSNTDNAAFPQMTKSYKNTRKSPALIVWGIHHSVSTAEQTKLYGSGNKL
VTVGSSNYQQSFVPSPGARPQVNGLSGRIDFHWLMLNPNDTVTFSFNGAFIAPDRASFLRGKSMGIQSGV
QVDANCEGDCYHSGGTIISNLPFQNIDSRAVGKCPRYVKQRSLLLATGMKNVPEIPKGRGLFGAIAGFIE
NGWEGLIDGWYGFRHQNAQGEGTAADYKSTQSAIDQITGKLNRLIEKTNQQFELIDNEFNEVEKQIGNVI
NWTRDSITEVWSYNAELLVAMENQHTIDLADSEMDKLYERVKRQLRENAEEDGTGCFEIFHKCDDDCMAS
IRNNTYDHSKYREEAMQNRIQIDPVKLSSGYKDVILWFSFGASCFILLAIVMGLVFICVKSRNMRCTICI
>AGO02477 A/Xuzhou/1/2013 2013/04/25 HA
MNTQILVFALIAIIPTNADKICLGHHAVSNGTKVNTLTERGVEVVNATETVERTNIPRICSKGKRTVDLG
QCGLLGTITGPPQCDQFLEFSADLIIERREGSDVCYPGKFVNEEALRQILRESGGIDKEAMGFTYSGIRT
NGATSACRRSGSSFYAEMKWLLSNTDNAAFPQMTKSYKNTRKSPALIVWGIHHSVSTAEQTKLYGSGSKL
VTVGSSNYQQSFVPSPGARPQVNGLSGRIDFHWLMLNPNDTVTFSFNGAFIAPDRASFLRGKSMGIQSGV
QVDANCEGDCYHSGGTIISNLPFQNIDSRAVGKCPRYVKQRSLLLATGMKNVPEIPKGRGLFGAIAGFIE
NGWEGLIDGWYGFRHQNAQGEGTAADYKSTQSAIDQITGKLNRLIEKTNQQFELIDNEFNEVEKQIGNVI
NWTRDSITEVWSYNAELLVAMENQHTIDLADSEMDKLYERVKRQLRENAEEDGTGCFEIFHKCDDDCMAS
IRNNTYDHSKYREEAMQNRIQIDPVKLSSGYKDVILWFSFGASCFILLAIVMGLVFICVKSRNMRCTICI
>AJS16350 A/Zhejiang/17/2014 2014/11/27 HA
MNTQILVFALIAIIPTNADKICLGHHAVSNGTKVNTLTEREVEVVNATETVERTNIPRICSKGKRTVDLG
QCGLLGTITGPPQCDQFLEFSADLIIERREGSDVCYPGKFVNEEALRQILRESGGIDKEAMGFTYNGIRT
NGVTSACRRSGSSFYAEMKWLLSNTDNAAFPQMTKSYKNTRKSPAIIVWGIHHSVSTAEQTKLYGSGNKL
VTVGSSNYQQSFVPSPGARPQVNGLSGRIDFHWLMLNPNDTVTFSFNGAFIAPDRASFLRGKSMGIQSGV
QVDANCEGDCYHSGGTIISNLPFQNIDSRAVGKCPRYVKQRSLLLATGMKNVPEIPKGRGLFGAIAGFIE
NGWEGLIDGWYGFRHQNAQGEGTAADYKSTQSAIDQITGKLNRLIAKTNQQFELIDNEFNEVEKQIGNVI
NWTRDSITEVWSYNAELLVAMENQHTIDLADSEMDKLYERVKRQLRENAEEDGTGCFEIFHKCDDDCMAS
IRNNTYDHRKYREEAMQNRIQIDPVKLSSGYKDVILWFSFGASCFILLAIVMGLVFICVKNGNMRCTICI
>AIN76383 A/Zhejiang/LS01/2014 2014/02/08 HA
MNTQILVFALIAIVPTNADKICLGHHAVSNGTKVNTLTERGVEVVNATETVERTNIPRICSKGKKTVDLG
QCGLLGTITGPPQCDQFLEFSADLIIERREGSDVCYPGKFVNEEALRQILRESGGIDKEAMGFTYSGIRT
NGTTSACRRSGSSFYAEMKWLLSNTDNAAFPQMTKSYKNTRKSPALIVWGIHHSVSTAEQTKLYGSGNKL
VTVGSSNYQQSFVPSPGARPQVNGLSGRIDFHWLMLNPNDTVTFSFNGAFIAPDRASFLRGKSMGIQSGV
QVDANCEGDCYHSGGTIISNLPFQNIDSRAVGKCPRYVKQKSLLLATGMKNVPEIPKGRGLFGAIAGFIE
NGWEGLIDGWYGFRHQNAQGEGTAADYKSTQSAIDQITGKLNRLIEKTNQQFELIDNEFNEVEKQIGNVI
NWTRDSITEVWSYNAELLVAMENQHTIDLADSEMDKLYERVKRQLRENAEEDGTGCFEIFHKCDDDCMAS
IRNNTYDHSKYREEAMQNRIQIDPVKLSSGYKDVILWFSFGASCFILLAIVMGLVFICVKNGNMRCTICI
>AGR85002 A/Zhenjiang/1/2013 2013/04/07 HA
MNTQILVFALIAIIPTNADKICLGHHAVSNGTKVNTLTERGVEVVNATETVERTNIPRICSKGKMTVDLG
QCGLLGTITGPPQCDQFLEFSADLIIERREGSDVCYPGKFVNEEALRQILRESGGIDKEAMGFTYSGIRT
NGATSACRRSGSSFYAEMKWLLSNTDNAAFPQMTKSYKNTRKSPALIVWGIHHSVSTAEQTKLYGSGNKL
VTVGSSNYQQSFVPSPGARPQVNGLSGRIDFHWLMLNPNDTVTFSFNGAFIAPDRASFLRGKSMGIQSGV
QVDANCEGDCYHSGGTIISNLPFQNIDSRAVGKCPRYVKQRSLLLATGMKNVPEIPKGRGLFGAIAGFIE
NGWEGLIDGWYGFRHQNAQGEGTAADYKSTQSAIDQITGKLNRLIEKTNQQFELIDNEFNEVEKQIGNVI
NWTRDSITEVWSYNAELLVAMENQHTIDLADSEMDKLYERVKRQLRENAEEDGTGCFEIFHKCDDDCMAS
IRNNTYDHSKYREEAMQNRIQIDPVKLSSGYKDVILWFSFGASCFILLAIVMGLVFICVKSRNKRCTICI
```

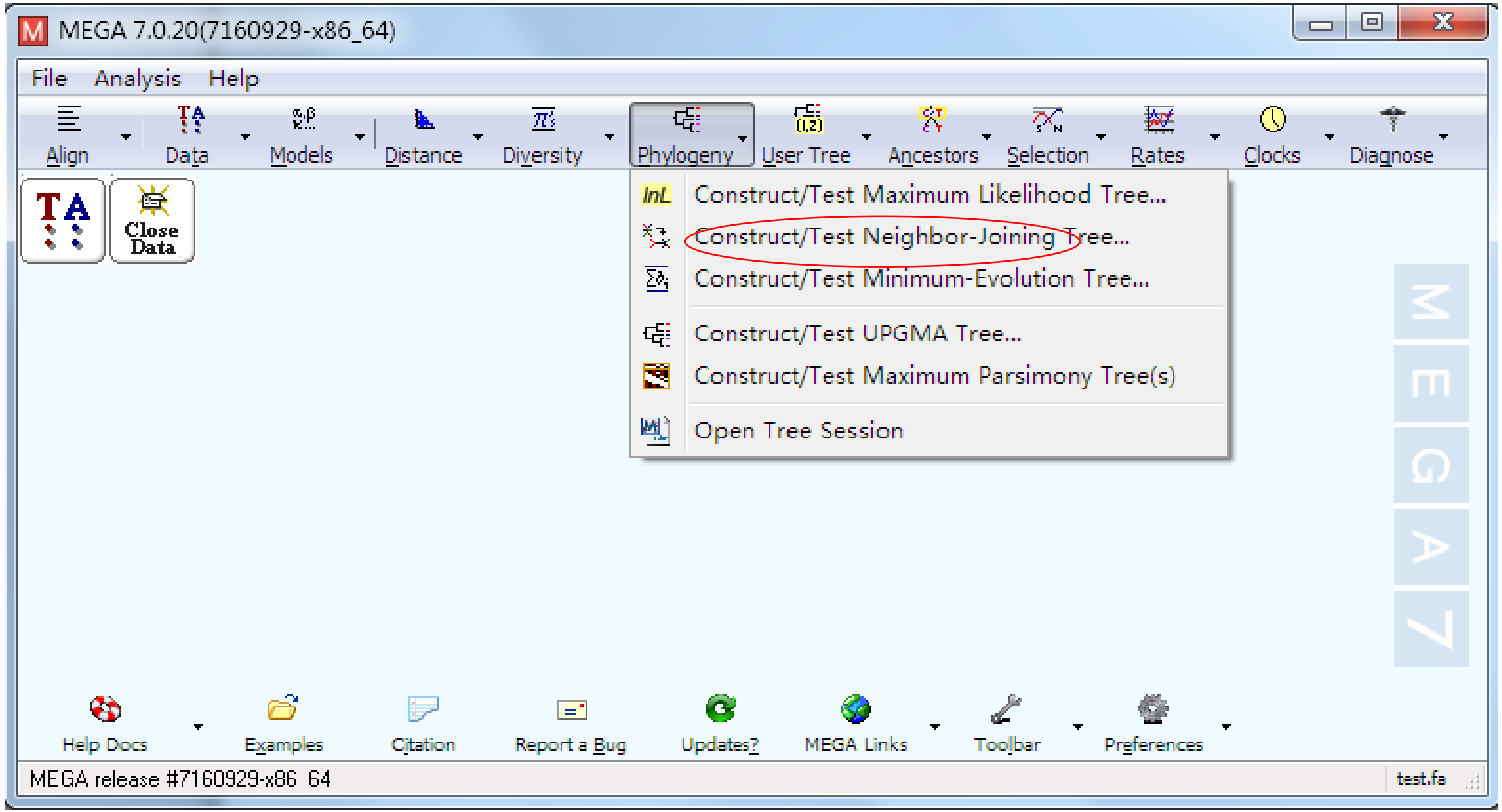

MEGA 7.0.20(7160929-x86_64)
File
Analysis
Help
Open A File/Session ... Ctrl+O
Open a Recently Used File
Edit a Text File
Convert File Format to MEGA...
Printer Setup...
Quit MEGA Ctrl+Q
Diversity
Phylogeny
User Tree
Ancestors
Selection
Rates
Clocks
Diagnose
MEGA7
Help Docs
Examples
Citation
Report a Bug
Updates?
MEGA Links
Toolbar
Preferences
MEGA release #7160929-x86_64
test.fa

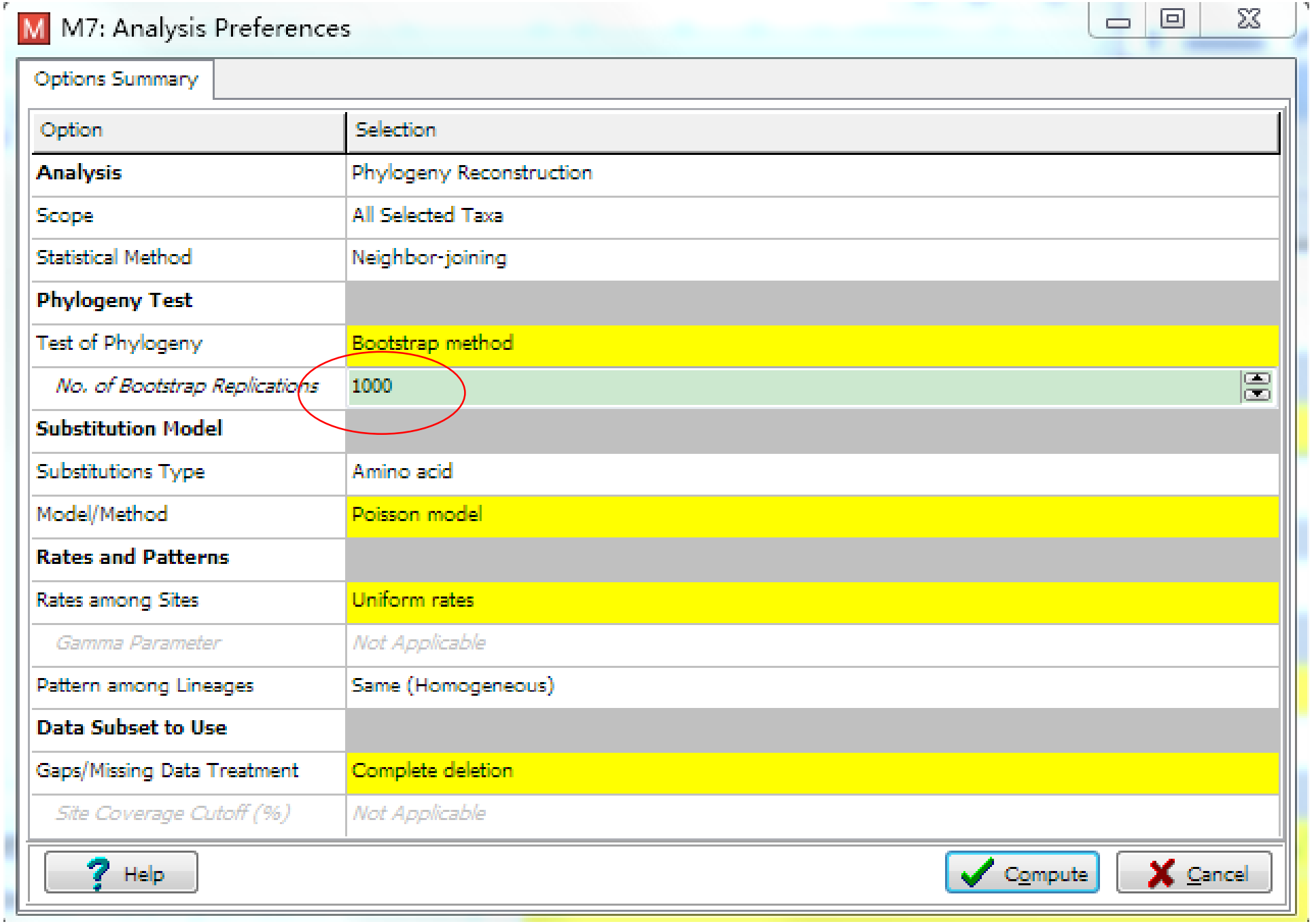

MEGA 7.0.20(7160929-x86_64)
File
Analysis
Help
Align
Data
Models
Distance
Diversity
Phylogeny
User Tree
Ancestors
Selection
Rates
Clocks
Diagnose
M7: Input Data
Data Type
Nucleotide Sequences
Protein Sequences
Pairwise Distance
Missing Data
?
Alignment Gap
-
Identical Symbol
.
OK
Cancel
Help
MEGA
7
Help Docs
Examples
Citation
Report a Bug
Updates?
MEGA Links
Toolbar
Preferences
MEGA release #7160929-x86_64

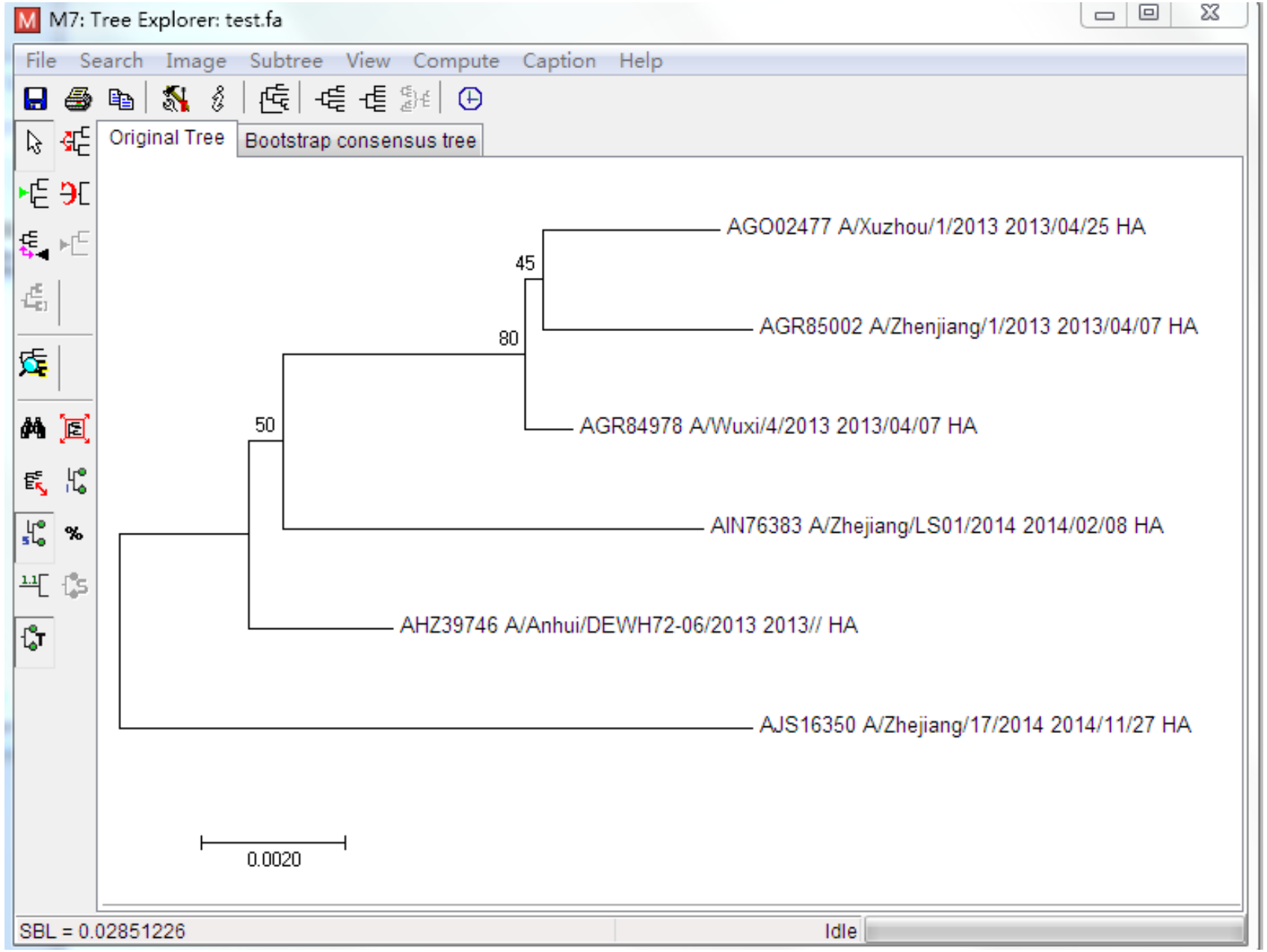

MEGA 7.0.20(7160929-x86_64)
File
Analysis
Help
Align
Data
Models
Distance
Diversity
Phylogeny
User Tree
Ancestors
Selection
Rates
Clocks
Diagnose
Close Data
InL
Construct/Test Maximum Likelihood Tree...
Construct/Test Neighbor-Joining Tree...
Construct/Test Minimum-Evolution Tree...
Construct/Test UPGMA Tree...
Construct/Test Maximum Parsimony Tree(s)
Open Tree Session
MEGA7
Help Docs
Examples
Citation
Report a Bug
Updates?
MEGA Links
Toolbar
Preferences
MEGA release #7160929-x86_64
test.fa

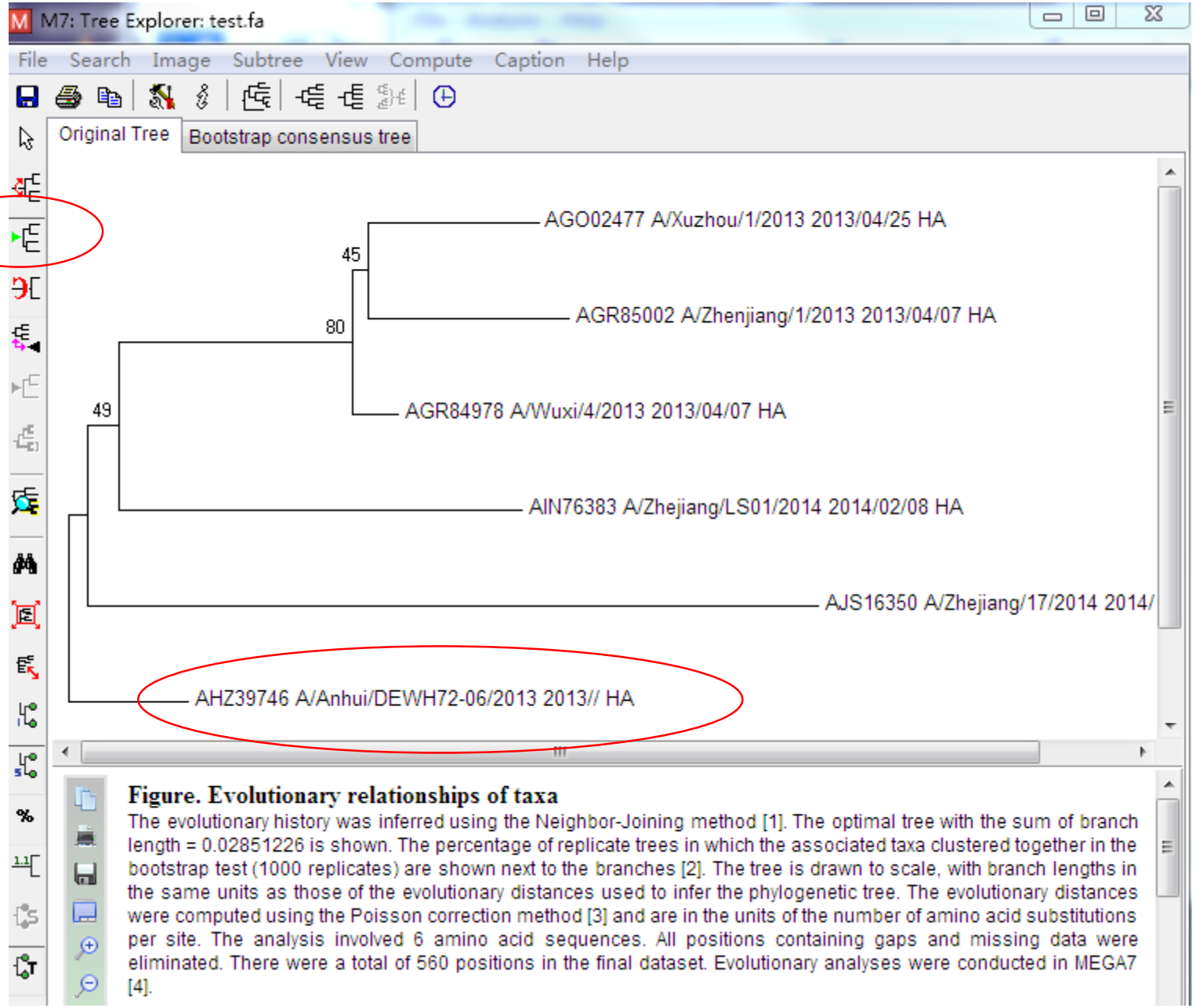

M7: Tree Explorer: test.fa
File
Search
Image
Subtree
View
Compute
Caption
Help
AGR84978 A/Wuxi/4/2013 2013/04/07 HA
AGR85002 A/Zhenjiang/1/2013 2013/04/07 HA
AGO02477 A/Xuzhou/1/2013 2013/04/25 HA
AHZ39746 A/Anhui/DEWH72-06/2013 2013// HA
AIN76383 A/Zhejiang/LS01/2014 2014/02/08 HA
AJS16350 A/Zhejiang/17/2014 2014/11/27 HA
0.0060
0.0040
0.0020
0.0000
SBL = 0.02913803
Idle

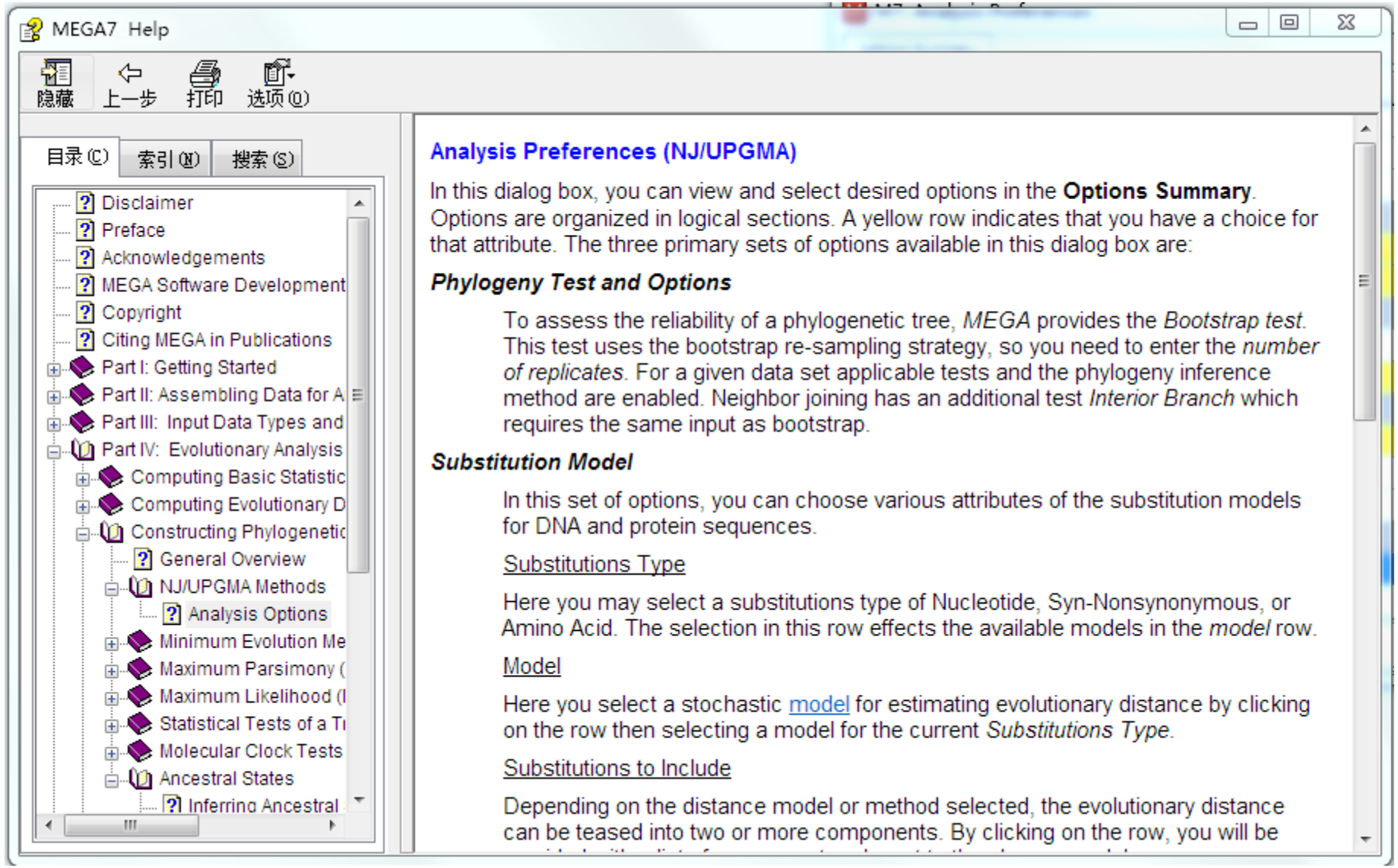

M7: Tree Explorer: test.fa
File
Search
Image
Subtree
View
Compute
Caption
Help
AJS16350 A/Zhejiang/17/2014 2014/11/27 HA
AHZ39746 A/Anhui/DEWH72-06/2013 2013// HA
AGR85002 A/Zhejiang/1/2013 2013/04/07 HA
AIN76383 A/Zhejiang/LS01/2014 2014/02/08 HA
AGR84978 A/Wuxi/4/2013 2013/04/07 HA
AGO02477 A/Xuzhou/1/2013 2013/0
SBL = 0.02913803
Idle


MEGA 7.0.20(7160929-x86_64)
File
Analysis
Help
Align
Data
Models
Distance
Diversity
Phylogeny
User Tree
Ancestors
Selection
Rates
Clocks
Diagnose
Close Data
Construct/Test Maximum Likelihood Tree...
Construct/Test Neighbor-Joining Tree...
Construct/Test Minimum-Evolution Tree...
Construct/Test UPGMA Tree...
Construct/Test Maximum Parsimony Tree(s)
Open Tree Session
MEGA7
Help Docs
Examples
Citation
Report a Bug
Updates?
MEGA Links
Toolbar
Preferences
MEGA release #7160929-x86_64
test.fa


M7: Analysis Preferences
Options Summary
Option
Selection
Analysis
Phylogeny Reconstruction
Scope
All Selected Taxa
Statistical Method
Neighbor-joining
Phylogeny Test
Test of Phylogeny
Bootstrap method
No. of Bootstrap Replications
1000
Substitution Model
Substitutions Type
Amino acid
Model/Method
Poisson model
Rates and Patterns
Rates among Sites
Uniform rates
Gamma Parameter
Not Applicable
Pattern among Lineages
Same (Homogeneous)
Data Subset to Use
Gaps/Missing Data Treatment
Complete deletion
Site Coverage Cutoff (%)
Not Applicable
Help
Compute
Cancel


M7: Tree Explorer: test.fa
File
Search
Image
Subtree
View
Compute
Caption
Help
Original Tree
Bootstrap consensus tree
AGO02477 A/Xuzhou/1/2013 2013/04/25 HA
AGR85002 A/Zhenjiang/1/2013 2013/04/07 HA
AGR84978 A/Wuxi/4/2013 2013/04/07 HA
AIN76383 A/Zhejiang/LS01/2014 2014/02/08 HA
AHZ39746 A/Anhui/DEWH72-06/2013 2013// HA
AJS16350 A/Zhejiang/17/2014 2014/11/27 HA
45
80
50
0.0020
SBL = 0.02851226
Idle


M7: Tree Explorer: test.fa
File Search Image Subtree View Compute Caption Help
Original Tree
Bootstrap consensus tree
AGO02477 A/Xuzhou/1/2013 2013/04/25 HA
AGR85002 A/Zhenjiang/1/2013 2013/04/07 HA
AGR84978 A/Wuxi/4/2013 2013/04/07 HA
AIN76383 A/Zhejiang/LS01/2014 2014/02/08 HA
AJS16350 A/Zhejiang/17/2014 2014/
AHZ39746 A/Anhui/DEWH72-06/2013 2013// HA
45
80
49
Figure. Evolutionary relationships of taxa
The evolutionary history was inferred using the Neighbor-Joining method [1]. The optimal tree with the sum of branch length = 0.02851226 is shown. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) are shown next to the branches [2]. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Poisson correction method [3] and are in the units of the number of amino acid substitutions per site. The analysis involved 6 amino acid sequences. All positions containing gaps and missing data were eliminated. There were a total of 560 positions in the final dataset. Evolutionary analyses were conducted in MEGA7 [4].


MEGA7 Help
隐藏
上一步
打印
选项(O)
目录(C)
索引(N)
搜索(S)
Disclaimer
Preface
Acknowledgements
MEGA Software Development
Copyright
Citing MEGA in Publications
Part I: Getting Started
Part II: Assembling Data for A
Part III: Input Data Types and
Part IV: Evolutionary Analysis
Computing Basic Statistic
Computing Evolutionary D
Constructing Phylogenetic
General Overview
NJ/UPGMA Methods
Analysis Options
Minimum Evolution Me
Maximum Parsimony (
Maximum Likelihood (I
Statistical Tests of a Tr
Molecular Clock Tests
Ancestral States
Inferring Ancestral
Analysis Preferences (NJ/UPGMA)
In this dialog box, you can view and select desired options in the Options Summary. Options are organized in logical sections. A yellow row indicates that you have a choice for that attribute. The three primary sets of options available in this dialog box are:
Phylogeny Test and Options
To assess the reliability of a phylogenetic tree, MEGA provides the Bootstrap test. This test uses the bootstrap re-sampling strategy, so you need to enter the number of replicates. For a given data set applicable tests and the phylogeny inference method are enabled. Neighbor joining has an additional test Interior Branch which requires the same input as bootstrap.
Substitution Model
In this set of options, you can choose various attributes of the substitution models for DNA and protein sequences.
Substitutions Type
Here you may select a substitutions type of Nucleotide, Syn-Nonsynonymous, or Amino Acid. The selection in this row effects the available models in the model row.
Model
Here you select a stochastic model for estimating evolutionary distance by clicking on the row then selecting a model for the current Substitutions Type.
Substitutions to Include
Depending on the distance model or method selected, the evolutionary distance can be teased into two or more components. By clicking on the row, you will be

Review

Essential Bioinformatics:

- Chapter ten
- Chapter eleven