Amira A. AL-Hosary
PhD of infectious diseases
Department of Animal Medicine
(Infectious Diseases)
Faculty of Veterinary Medicine
Assiut University-Egypt

PHYLOGENETIC ANALYSIS

Phylogenetic Basics:

- Biological sequence analysis is founded on solid evolutionary principles.
- Similarities and divergence among related biological sequences revealed by sequence alignment often have to be rationalized and visualized in the context of phylogenetic trees.
- Thus, molecular phylogenetic is a fundamental aspect of bioinformatics.

MOLECULAR EVOLUTION AND MOLECULAR PHYLOGENETICS

"What is evolution?"

- ✓ In the biological context, evolution can be defined as the development of a biological form from other preexisting forms or its origin to the current existing form through natural selections and modifications.
- ✓ The driving force behind evolution is natural selection in which "unfit" forms are eliminated through changes of environmental conditions or sexual selection so that only the fittest are selected.
- ✓ The underlying mechanism of evolution is genetic mutations that occur spontaneously.

Phylogenetic?

It is the study of the evolutionary history of living organisms using treelike diagrams to represent pedigrees of these organisms.

The tree branching patterns representing the evolutionary divergence are referred to as *phylogeny.* Phylogenetic can be studied in various ways.

It is often studied using fossil records, which contain

morphological information about ancestors of current species and the timeline of divergence.

However, fossil records have many limitations; they may be available only for certain species.

Molecular phylogenetic:

Defined as the study of evolutionary relationships of genes and other biological macromolecules by analyzing mutations at various positions in their sequences and developing hypotheses about the evolutionary relatedness of the biomolecules.

Based on the sequence similarity of the molecules, evolutionary relationships between the organisms can often be inferred.

Major Assumptions

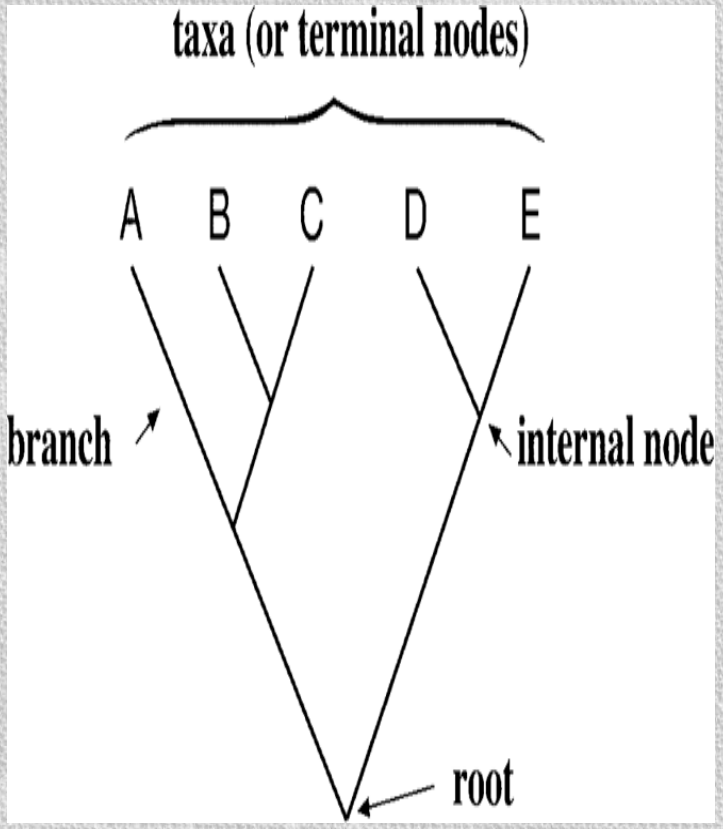
To use molecular data to reconstruct evolutionary history requires making a number of reasonable assumptions.

- The first is that the molecular sequences used in phylogenetic construction are homologous, meaning that they share a common origin and subsequently diverged through time.
- Phylogenetic divergence is assumed to be bifurcating, meaning that a parent branch splits into two daughter branches at any given point. Another assumption in phylogenetic is that each position in a sequence evolved independently.
- The variability among sequences is sufficiently informative for constructing unambiguous phylogenetic trees.

TERMINOLOGY

- phylogenetic tree
- The lines in the tree are called ***branches.***
- At the tips of the branches are present-day species or sequences known as ***taxa*** (the singular form is *taxon*).
- The connecting point where two adjacent branches join is called a ***node,*** which represents an inferred ancestor of extant taxa.
- The bifurcating point at the very bottom of the tree is the ***root node,*** which represents the common ancestor of all members of

the tree.



A typical bifurcating phylogenetic tree showing root, internal nodes, terminal nodes and branches.

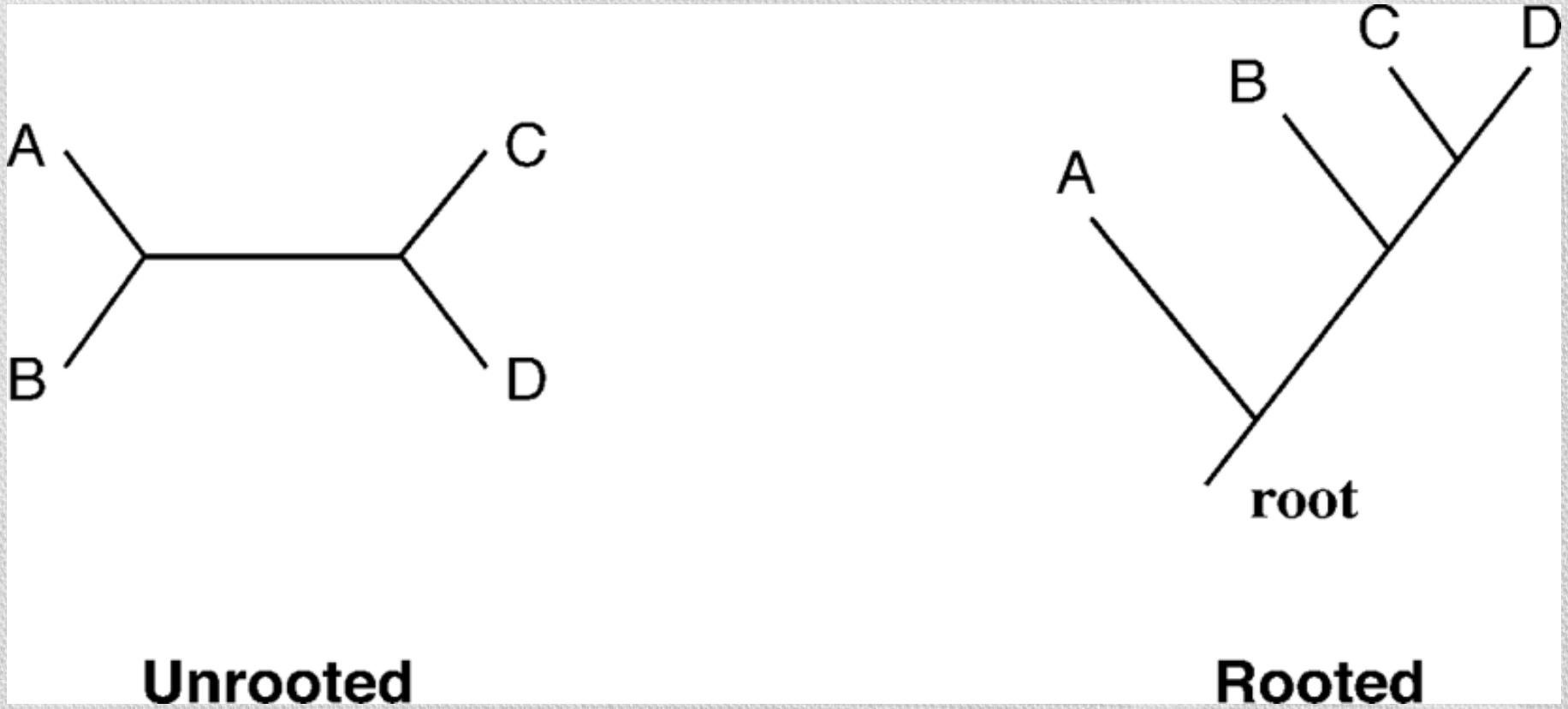
The rooted and un rooted tree:

A phylogenetic tree can be either rooted or un rooted.

An *un rooted phylogenetic tree:* does not assume knowledge of a common ancestor, but only positions the taxa to show their relative relationships. Because there is no indication of which node represents an ancestor, there is no direction of an evolutionary path in an un rooted tree.

In a *rooted tree,* all the sequences under study have a common ancestor or root node from which a unique evolutionary path leads to all other nodes. Obviously, a rooted tree is more informative than an un rooted one.

To convert an un rooted tree to a rooted tree, one needs to first determine where the root is.



An illustration of rooted versus un rooted trees. A phylogenetic tree without definition of a root is un rooted (*left*). The tree with a root is rooted (*right*).

Conversion of the un rooted tree to rooted:

Strictly speaking, the root of the tree is not known; the common ancestor is already extinct.

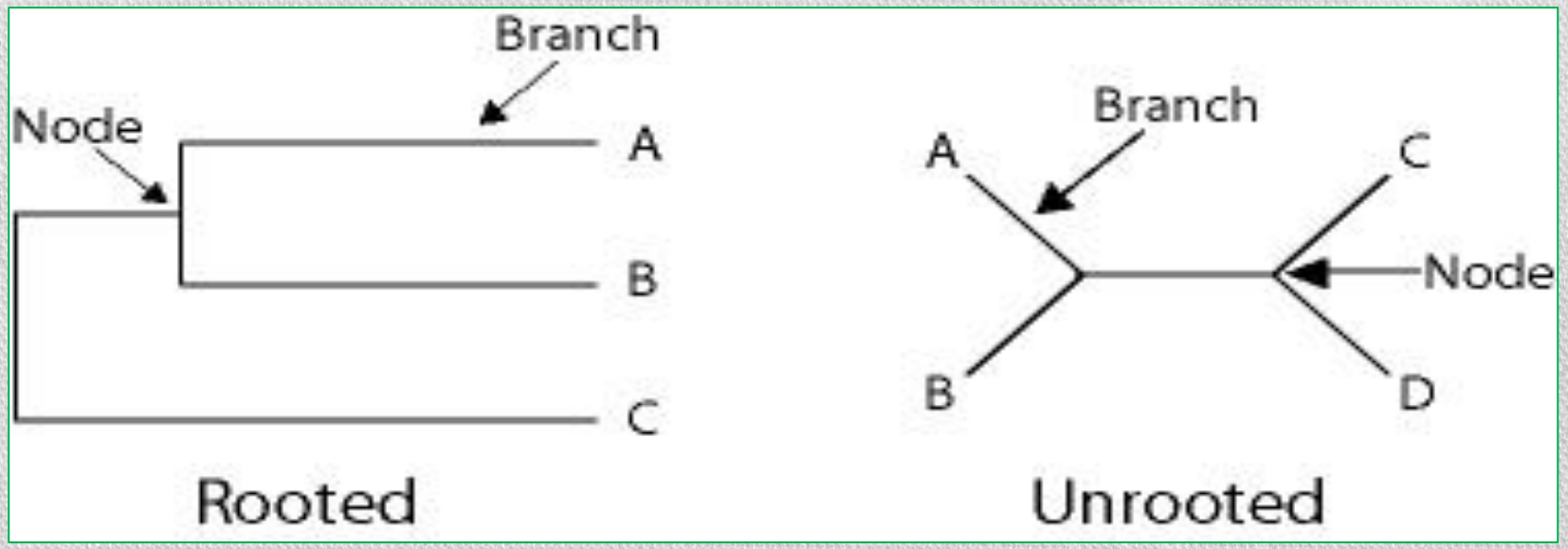
In practice, however, it is often desirable to define the root of a tree.

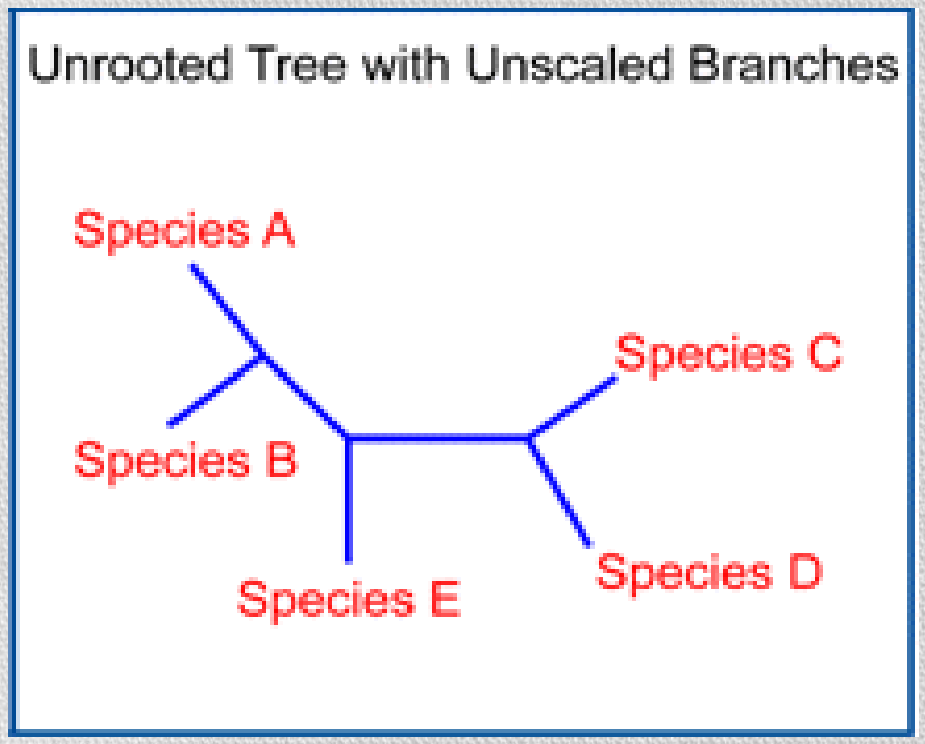
There are **two ways to define the root of a tree:**

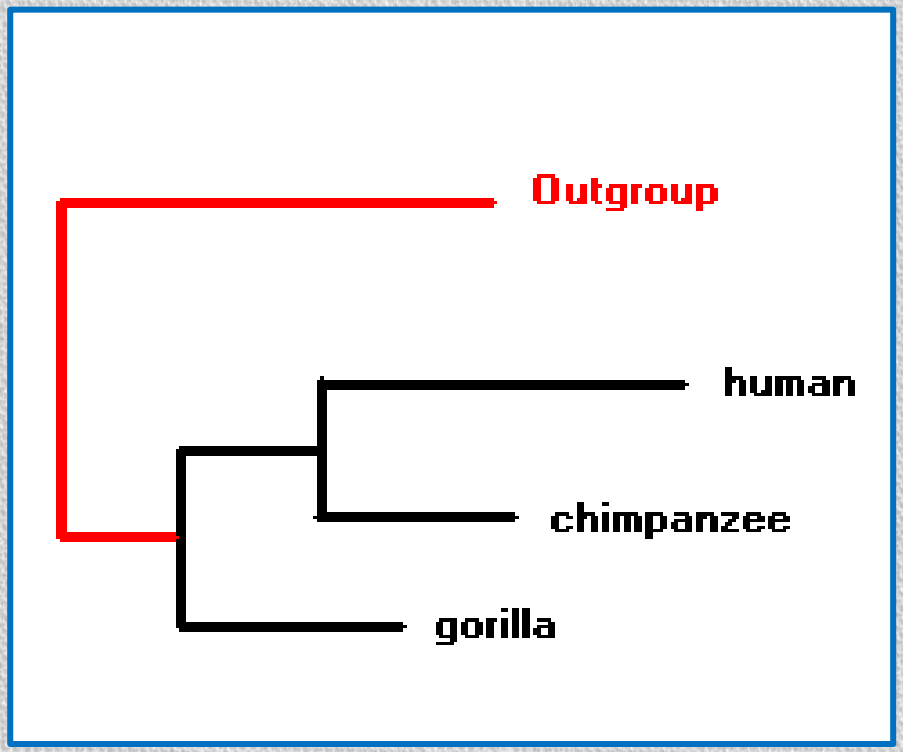
- One is to use an ***out group***, which is a sequence that is homologous to the sequences under consideration, but separated from those sequences at an early evolutionary time.

Out groups are generally determined from independent sources of information. *For example*, a bird sequence can be used as a

root for the phylogenetic analysis of mammals based on multiple lines of evidence that indicate that birds branched off prior to all mammalian taxa in the in group.


Branch
Node
A
B
C
Rooted
Branch
A
C
Node
B
D
Unrooted


Unrooted Tree with Unscaled Branches
Species A
Species C
Species B
Species E
Species D


Outgroup
human
chimpanzee
gorilla

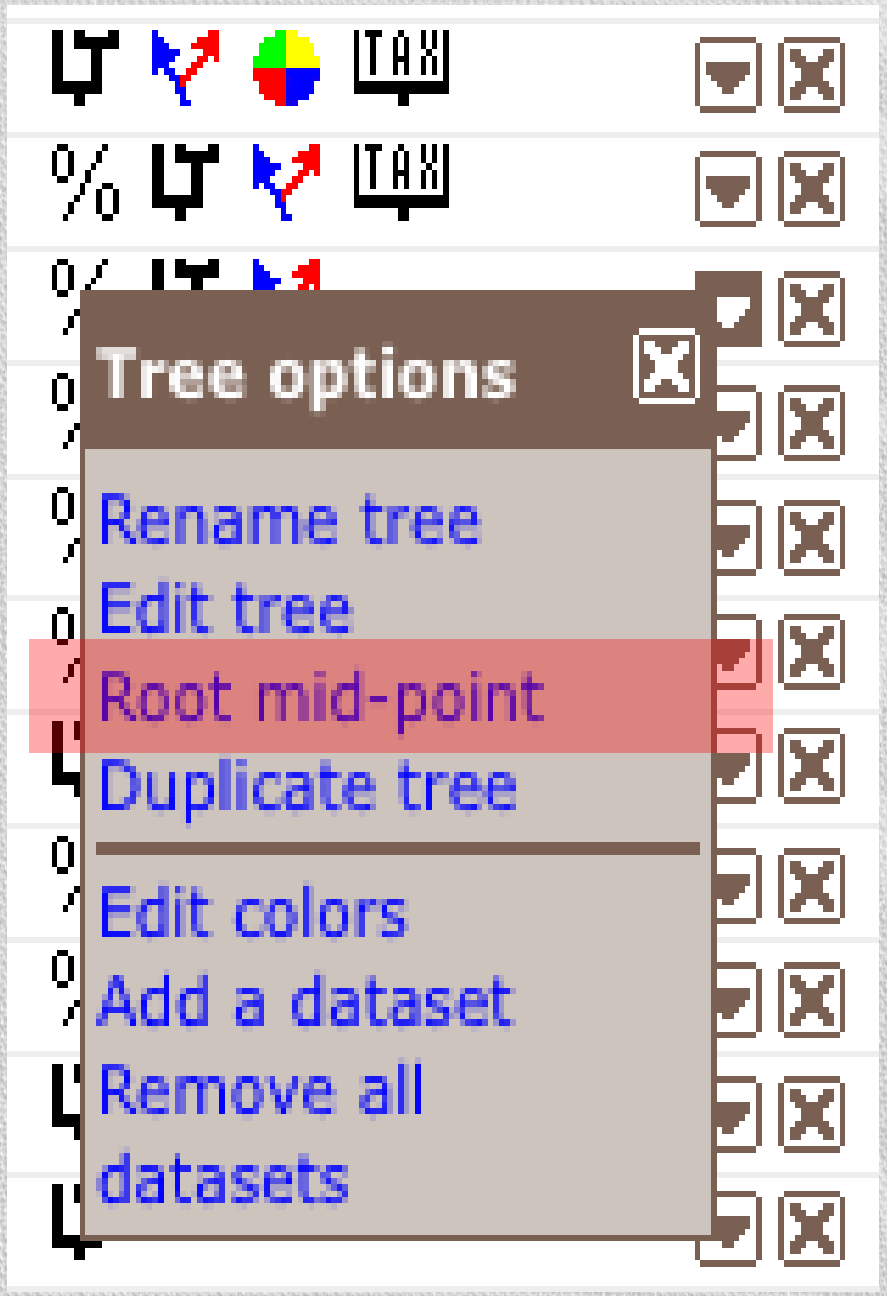
In the absence of a good out group, a tree can be rooted using

Midpoint rooting approach, in which the midpoint of the two most divergent groups judged by overall branch lengths is assigned as the root.

This type of rooting assumes that divergence from root to tips for both branches is equal and follows the "molecular clock" hypothesis.

Molecular clock is an assumption by which molecular sequences evolve at constant rates so that the amount of accumulated mutations is proportional to evolutionary time.

Based on this hypothesis, branch lengths on a tree can be used to estimate divergence time. This assumption of uniformity of evolutionary rates, however, rarely holds true in reality.


Tree options
Rename tree
Edit tree
Root mid-point
Duplicate tree
Edit colors
Add a dataset
Remove all
datasets

GENE PHYLOGENY VS SPECIES PHYLOGENY

Gene phylogeny (phylogeny inferred from a gene or protein sequence) only describes the evolution of that particular gene or encoded protein. This sequence may evolve more or less rapidly than other genes in the genome or may have a different evolutionary history from the rest of the genome.

The species evolution is the combined result of evolution by multiple genes in a genome.

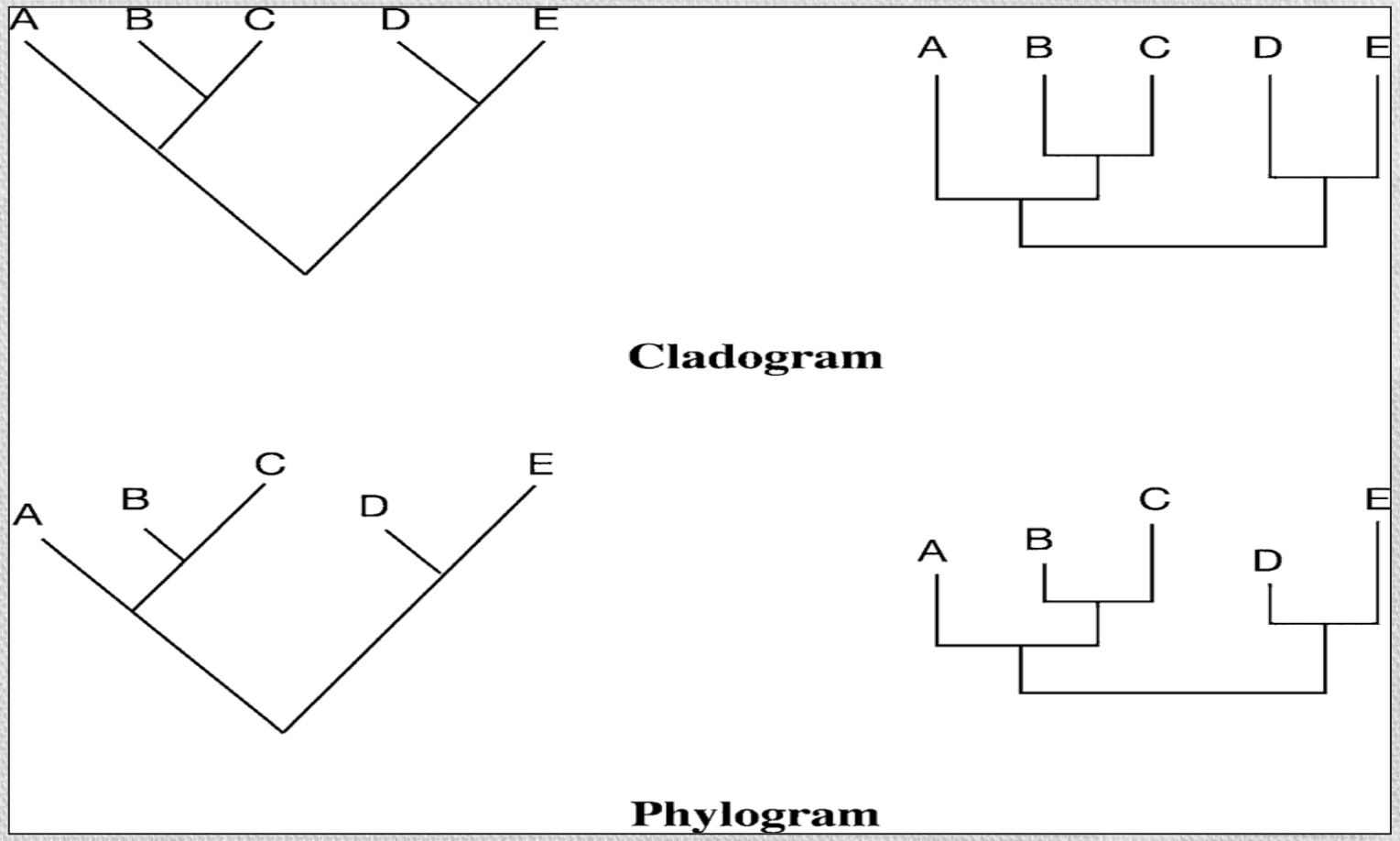
In a species tree, the branching point at an internal node represents the speciation event whereas, **in a gene tree**, the internal node indicates a gene duplication event.

Thus, to obtain a species phylogeny, phylogenetic trees from a variety of gene families need to be constructed to give an overall assessment of the species evolution.

FORMS OF TREE REPRESENTATION:

The topology of branches in a tree defines the relationships between the taxa. The trees can be drawn in different ways, such as a cladogram or a phylogram. In each of these tree representations, the branches of a tree can freely rotate without changing the relationships among the taxa.

In a *phylogram,* the branch lengths represent the amount of evolutionary divergence. Such trees are said to be scaled. The scaled trees have the advantage of showing both the evolutionary relationships and information about the relative divergence time of the branches.	**In a *cladogram,*** however, the external taxa line up neatly in a row or column. Their branch lengths are not proportional to the number of evolutionary changes and thus have no phylogenetic meaning. In such an un scaled tree, only the topology of the tree matters, which shows the relative ordering of the taxa.



Phylogenetic trees drawn as **cladograms** (*top*) and **phylograms** (*bottom*). **The branch lengths** are un scaled in the cladograms and scaled in the phylograms. **The trees** can be drawn as angled form (*left*) or squared form (*right*).

PROCEDURE:

Molecular phylogenetic tree construction can be divided into five steps:

(1) Choosing molecular markers.

(2) Performing multiple sequence alignment.

(3) Choosing a model of evolution.

(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 choice of molecular markers is an important matter because it can make a major difference in obtaining a correct tree.

The decision to use nucleotide or protein sequences depends on the properties of the sequences and the purposes of the study.

For example,

Studying very closely related organisms, nucleotide sequences, which evolve more rapidly than proteins.

Alignment

The second step in phylogenetic analysis is to construct sequence alignment.

This is probably the most critical step in the procedure because it establishes positional correspondence in evolution.

Only the correct alignment produces correct phylogenetic inference because aligned positions are assumed to be genealogically related.

Incorrect alignment leads to systematic errors in the final tree or even a completely wrong tree.

Choosing Substitution Models

The statistical models used to correct homoplasy are called *substitution models* or *evolutionary models.*

The most common substitution models or evolutionary models:

1- Jukes–Cantor Model.

2- Kimura Model.

Jukes–Cantor Model

The simplest nucleotide substitution model is the Jukes–Cantor model, which assumes that all nucleotides are substituted with equal probability. This model can only handle reasonably closely related sequences.

A formula for deriving evolutionary distances that include hidden changes is introduced by using a logarithmic function.

$$d\mathrm{AB} = -(3/4)\ \ln\left[1 - (4/3)\ p\mathrm{AB}\right]$$

where *d* AB is the evolutionary distance between sequences A and B and *p* AB is the observed sequence distance measured by the proportion of substitutions over the entire length of the alignment.

Kimura Model

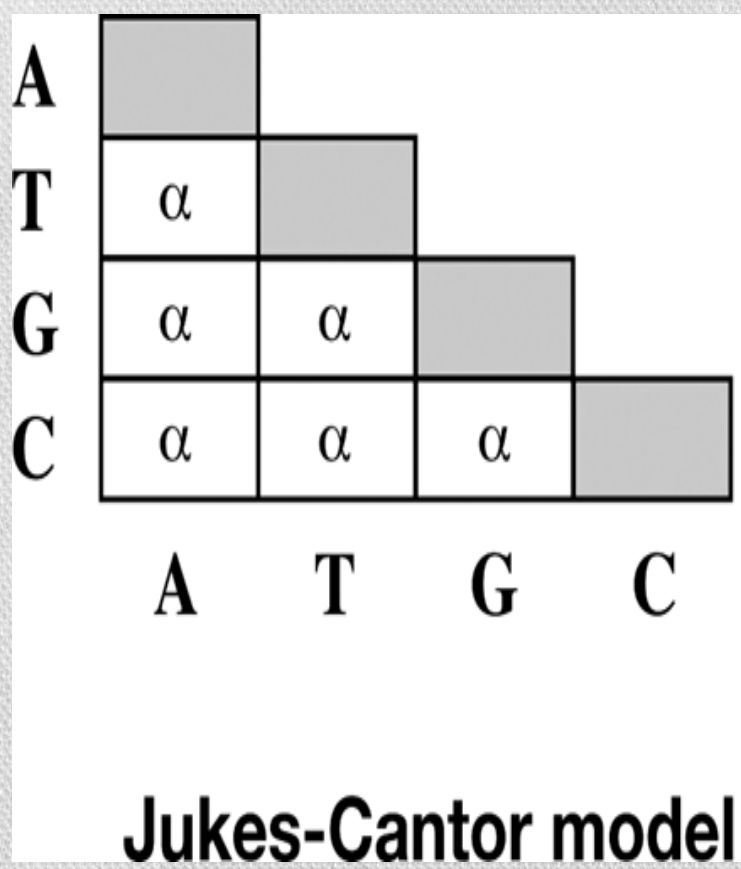
Another model to correct evolutionary distances is called the Kimura two-parameter model. This is a more sophisticated model in which mutation rates for transitions and transversion are assumed to be different, which is more realistic.

According to this model, transitions occur more frequently than trans versions, which, therefore, provides a more realistic estimate of evolutionary distances.

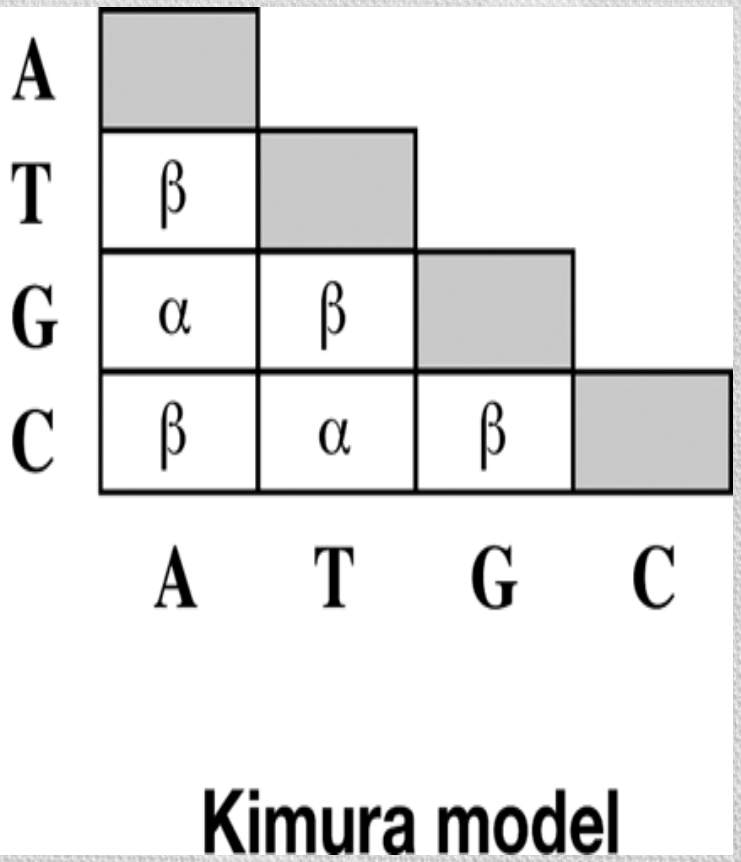
The Kimura model uses the following formula:

$$d\,\mathrm{AB} = -(1/2)\ln(1 - 2p\mathrm{ti} - p\,\mathrm{tv}) - (1/4)\ln(1 - 2p\mathrm{tv})$$

where *d* AB is the evolutionary distance between sequences A and B, *p*ti is the observed frequency for transition, and *p* tv the frequency of transversion.



Jukes-Cantor model



Kimura model

The Jukes–Cantor and Kimura models for DNA substitutions.

In the Jukes–Cantor model, all nucleotides have equal substitution rates (α).

In the Kimura model, there are unequal rates of transitions (α) and transversions (β).

Determining a tree building method.

Several methods are available for reconstructing phylogenetic trees.

Most of them use some criterion for evaluating the fit of a given data set to the topology and then search for the tree that gives the best score in terms of that criterion.

If the criterion used is realistic and the data are sufficient, the tree should represent the true phylogenetic relationship of the sequences.

There are two methods:-

1- DISTANCE-BASED METHODS

True evolutionary distances between sequences can be calculated from observed distances after correction using a variety of evolutionary models. The computed evolutionary distances can be used to construct a matrix of distances between all individual pairs of taxa

→ Neighbour-joining (NJ) methods.

2- CHARACTER-BASED METHODS.

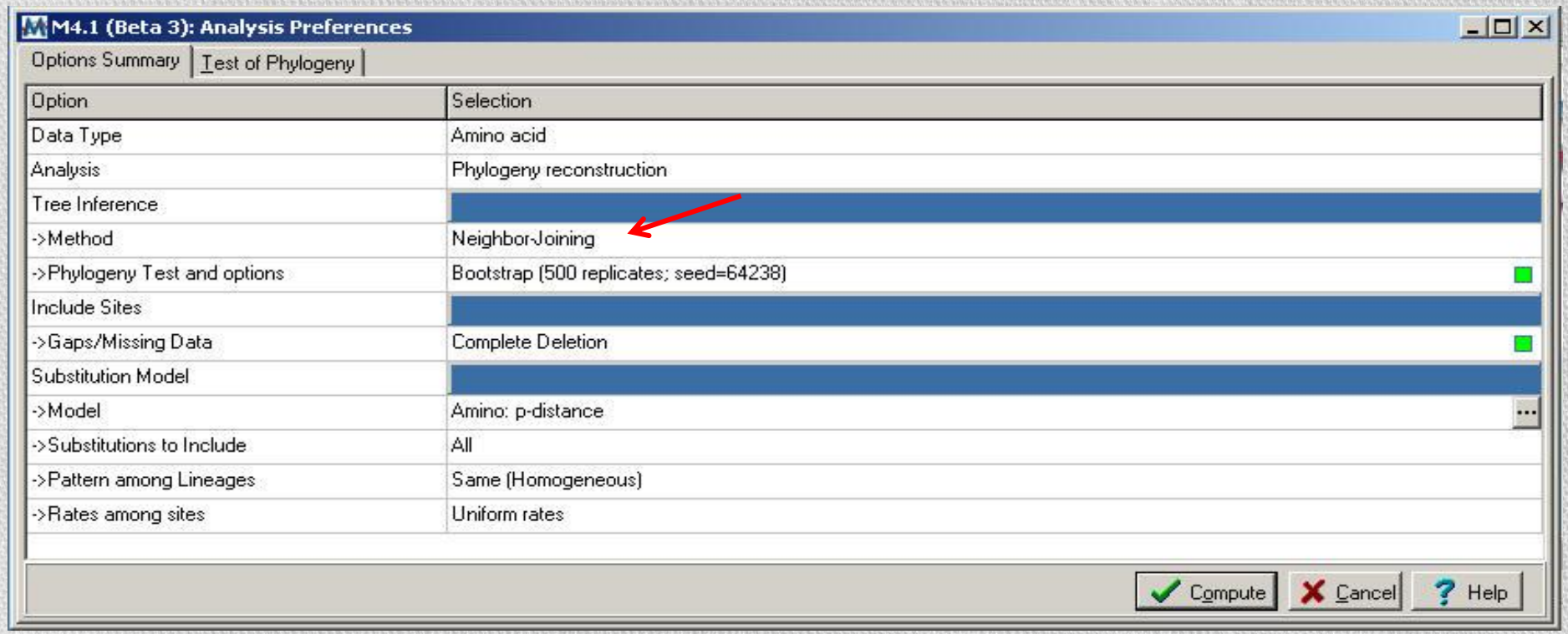
also called (*discrete methods*) based directly on the sequence characters rather than on pairwise distances. They count mutational events accumulated on the sequences and may therefore avoid the loss of information when characters are converted to distances.

Maximum parsimony (MP) methods

Maximum likelihood (ML) methods.

Neighbour-joining (NJ) methods.

The neighbor-joining (**NJ**) method has been frequently used in molecular phylogenetic, especially for large-scale data analyses. It has desirable statistical properties and is known to produce trees as accurate as more computationally intensive. This method works in a stepwise fashion by minimizing the sum of branch lengths at each step of sequence clustering.

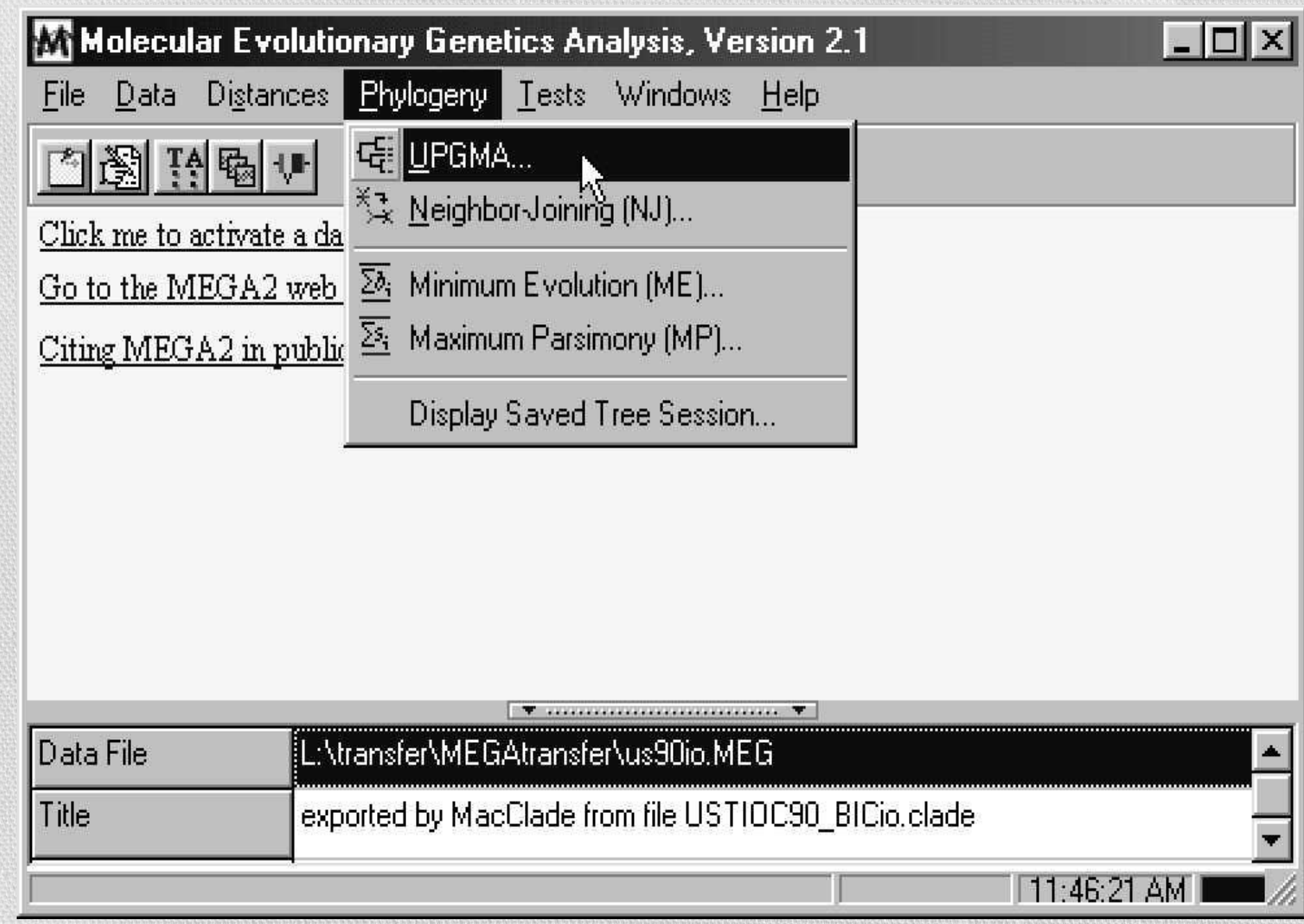


Maximum parsimony (MP) methods:

The topology requiring the smallest number of nucleotide changes to fit the observed sequence data is chosen to represent the true tree.

Maximum likelihood (ML) methods:

The topology with the greatest likelihood under a given probabilistic model of nucleotide substitutions is chosen.


Molecular Evolutionary Genetics Analysis, Version 2.1
File
Data
Distances
Phylogeny
Tests
Windows
Help
UPGMA...
Neighbor-Joining (NJ)...
Minimum Evolution (ME)...
Maximum Parsimony (MP)...
Display Saved Tree Session...
Click me to activate a da
Go to the MEGA2 web
Citing MEGA2 in public
Data File
L:\transfer\MEGAtransfer\us90io.MEG
Title
exported by MacClade from file USTIOC90_BICio.clade
11:46:21 AM

Assessing tree reliability.

After phylogenetic tree construction, the next step is to statistically evaluate the reliability of the inferred phylogeny.

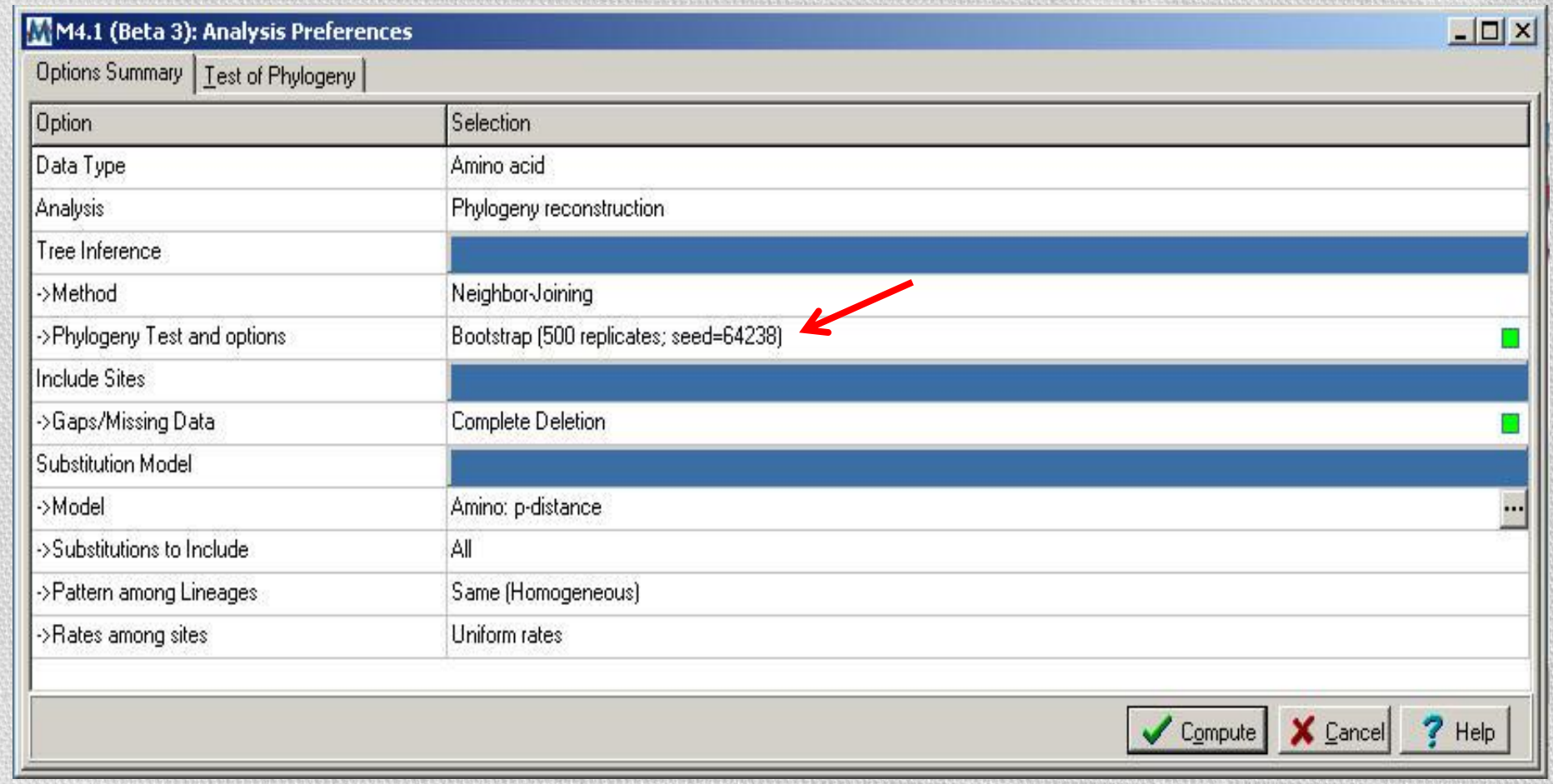
There are two questions that need to be addressed.

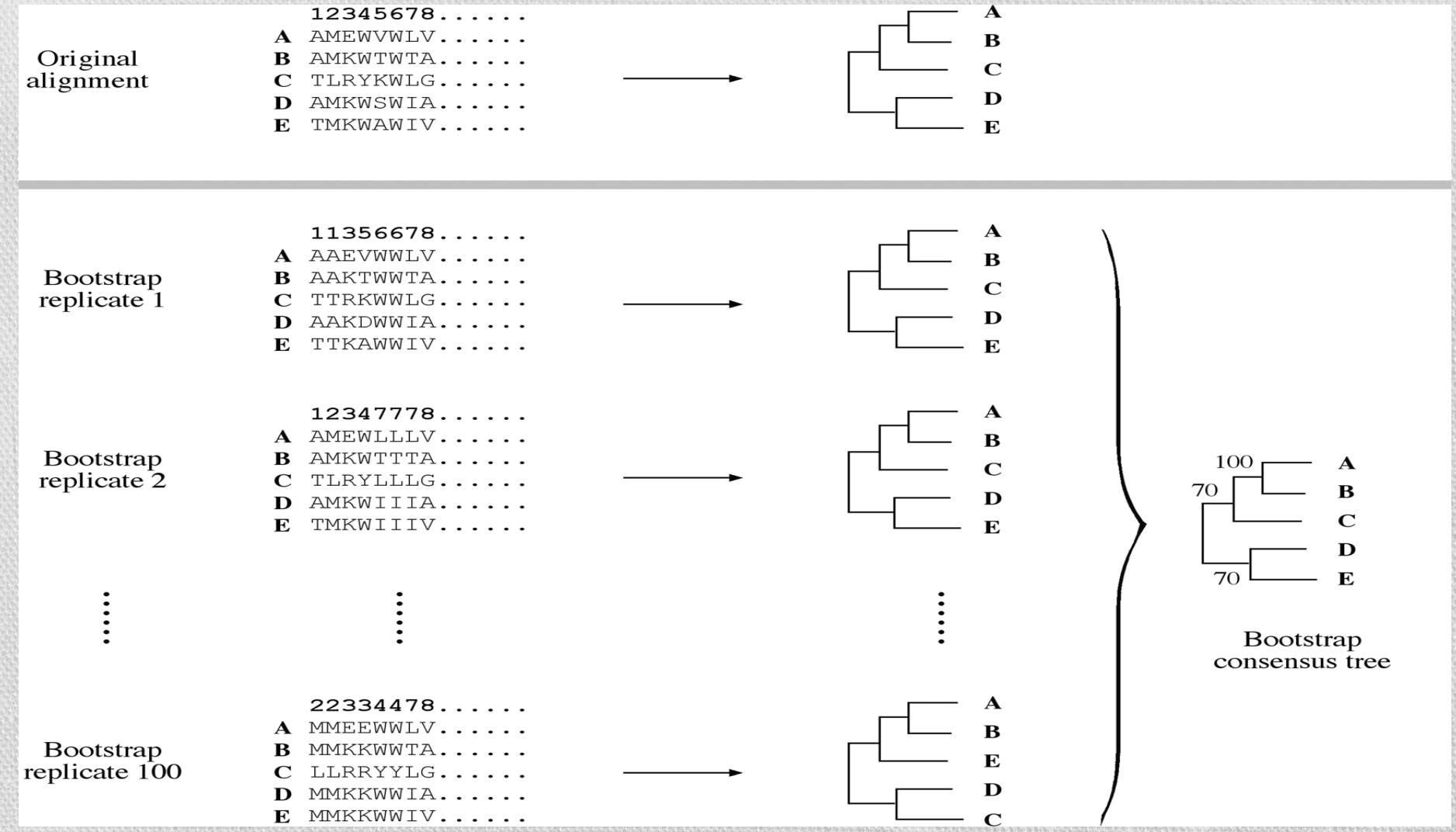
One is how reliable the tree or a portion of it.

Second is whether this tree is significantly better than another tree.

What Is Bootstrapping?

Bootstrapping **is a statistical technique that tests the sampling errors of a phylogenetic tree.**





Schematic representation of a bootstrap analysis showing the original alignment and modified replicates in which certain sites are randomly replaced with other existing sites. The resulting altered replicates are used to building trees for statistical analysis at each node.

The bootstrap test provides a measure for evaluating the confidence levels of the tree topology.

Analysis has shown that a bootstrap value of 70% approximately corresponds to 95% statistical confidence, although the issue is still a subject of debate.

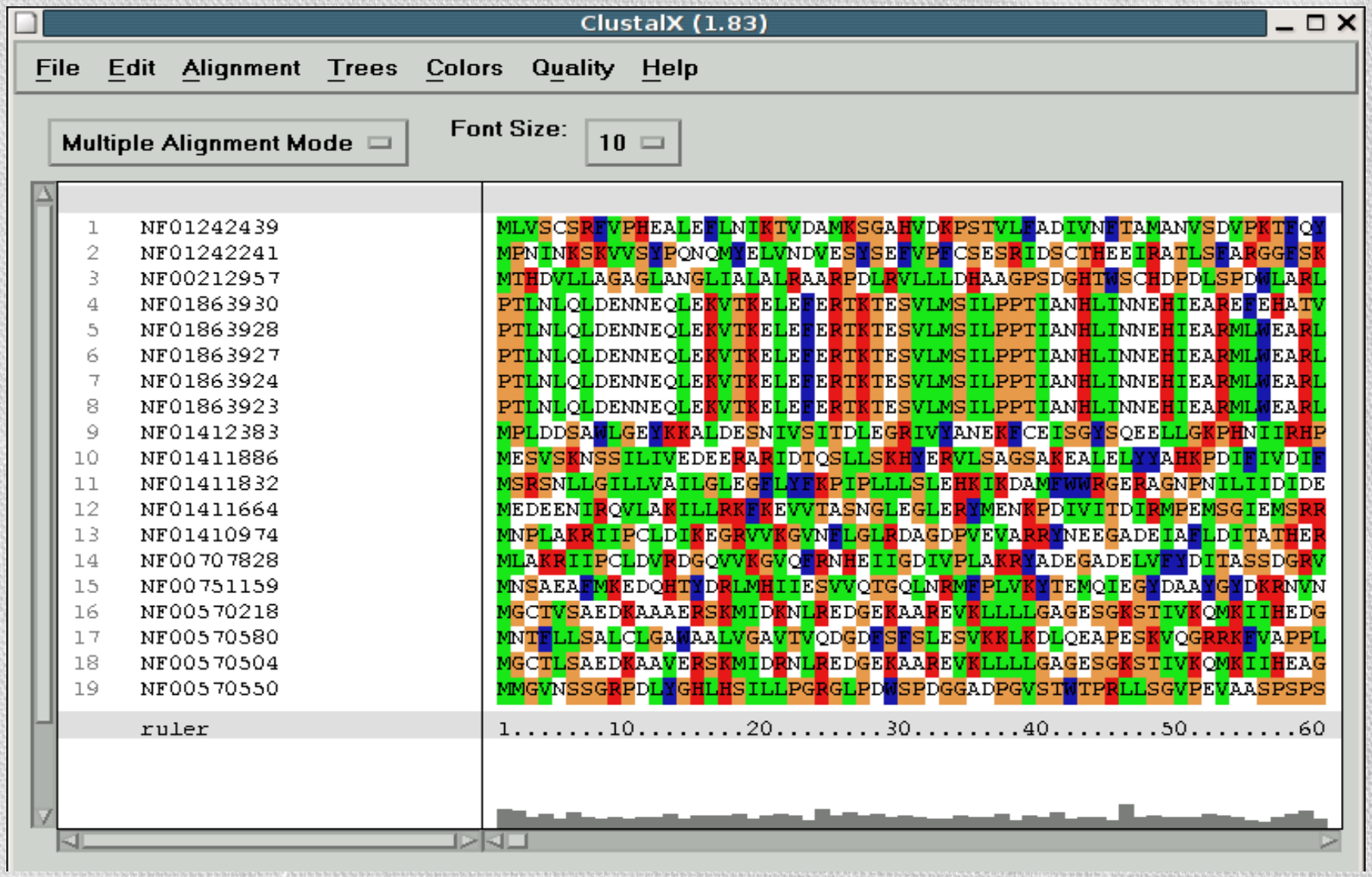
Bootstrapping does not assess the accuracy of a tree, but only indicates consistency and stability of individual clades of the tree.

This means that, because of systematic errors, wrong trees can still be obtained with high bootstrap values.

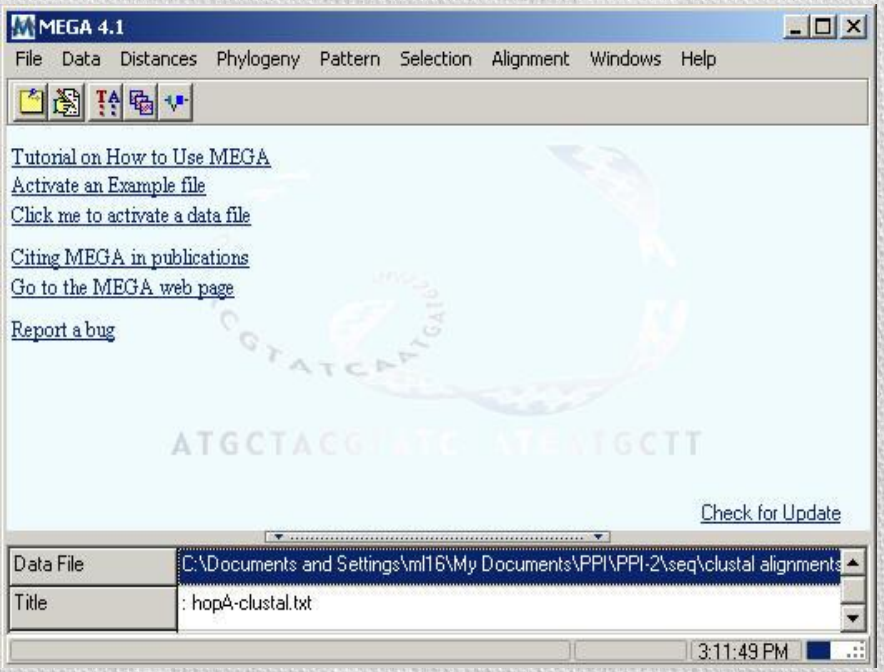
It is generally recommended that a phylogenetic tree should be bootstrapped 500 to 1,000 times

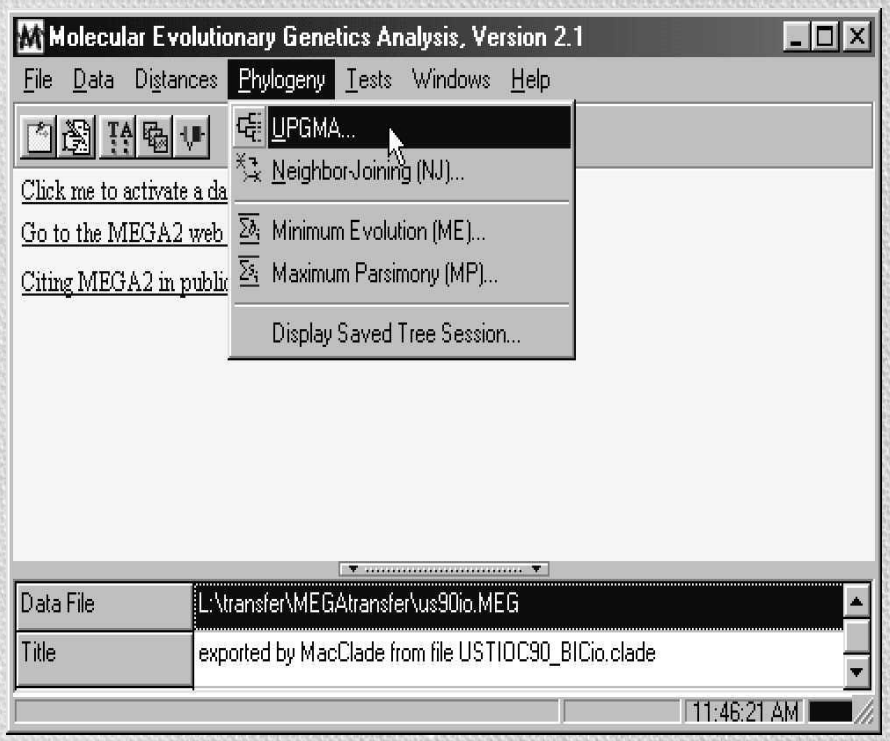
PRACTICAL

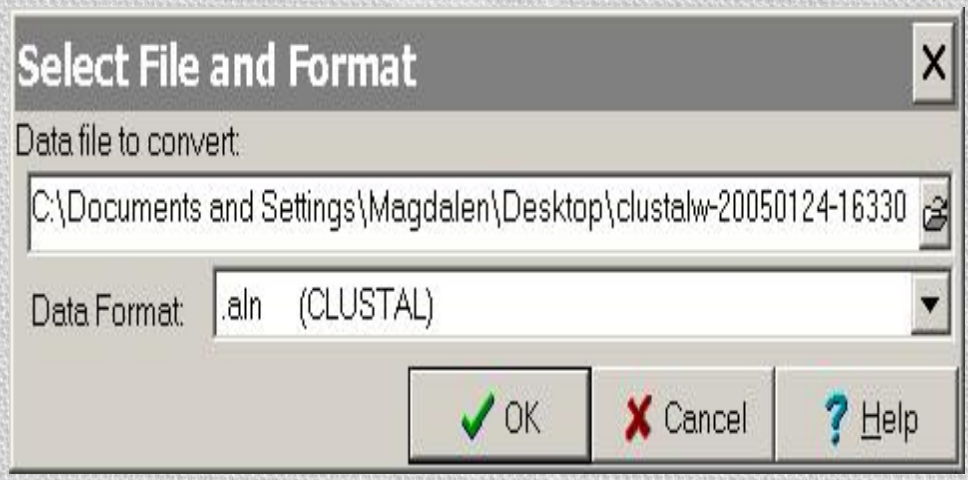
All Sequences alignment:

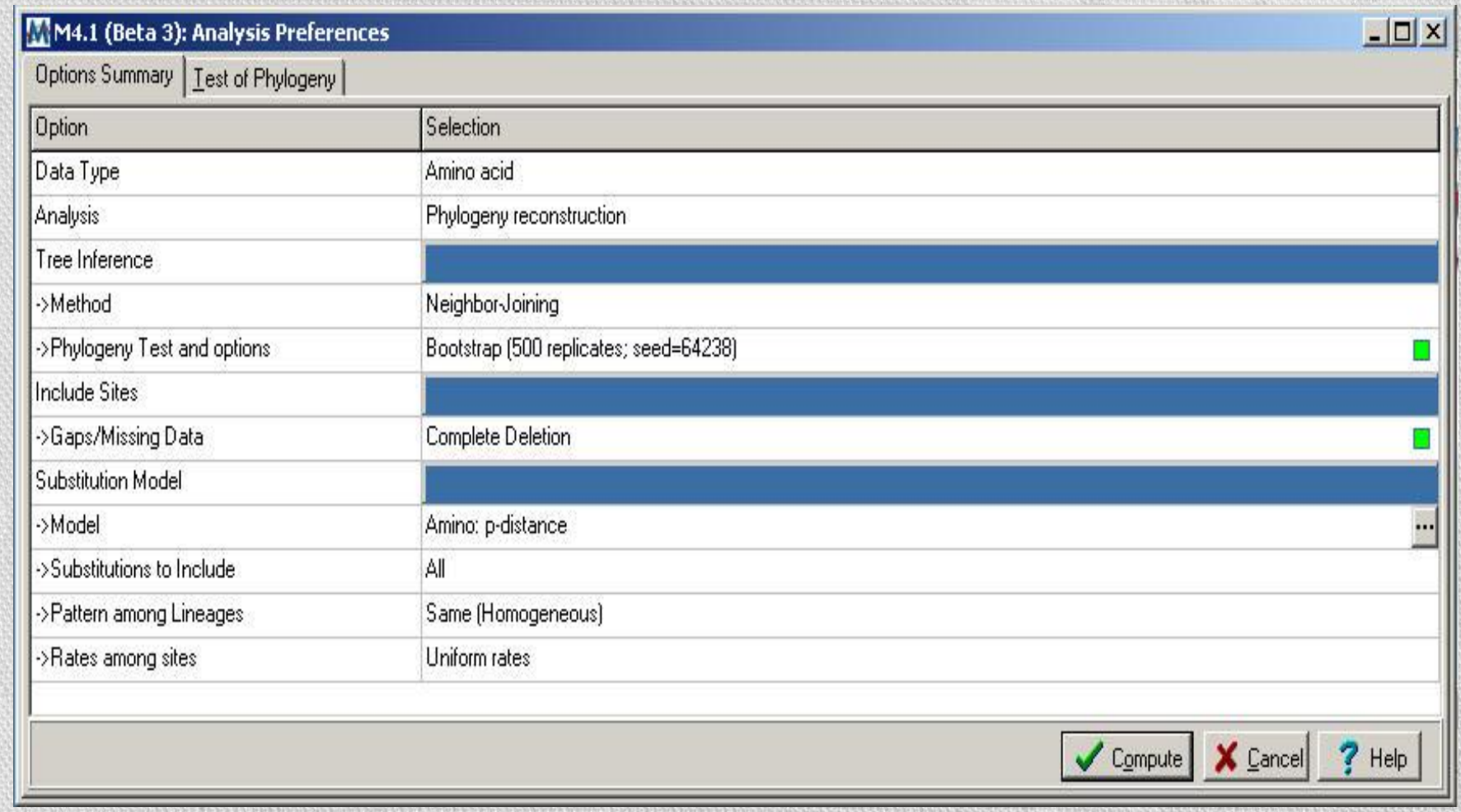


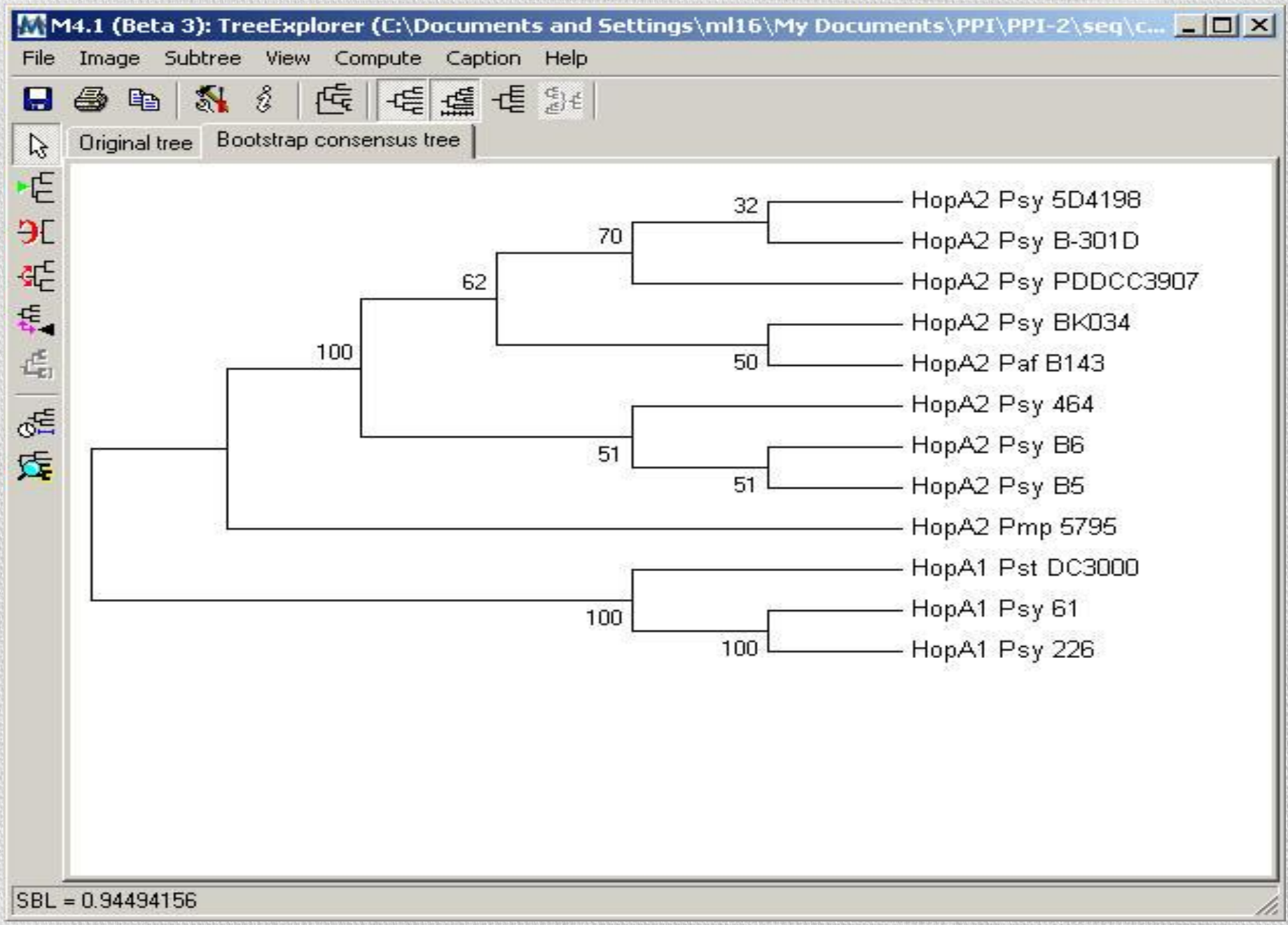

MEGA 4
©1993-2008
KOICHIRO TAMURA
JOEL DUDLEY
MASATOSHI NEI
SUDHIR KUMAR

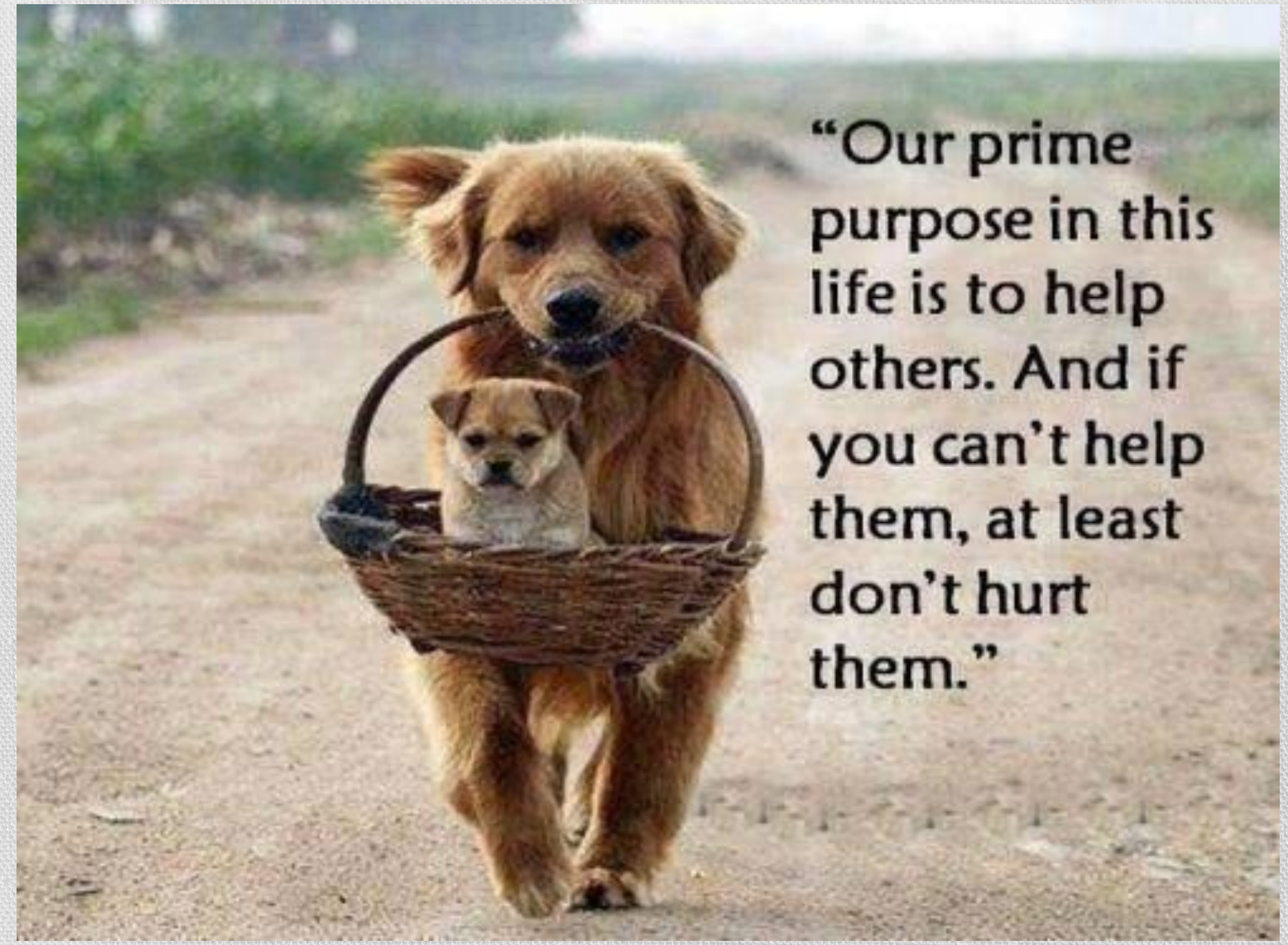

MEGA 4.1
File Data Distances Phylogeny Pattern Selection Alignment Windows Help
Tutorial on How to Use MEGA
Activate an Example file
Click me to activate a data file
Citing MEGA in publications
Go to the MEGA web page
Report a bug
Check for Update
Data File
C:\Documents and Settings\ml16\My Documents\PPI\PPI-2\seq\clustal alignments
Title
: hopA-clustal.txt
3:11:49 PM

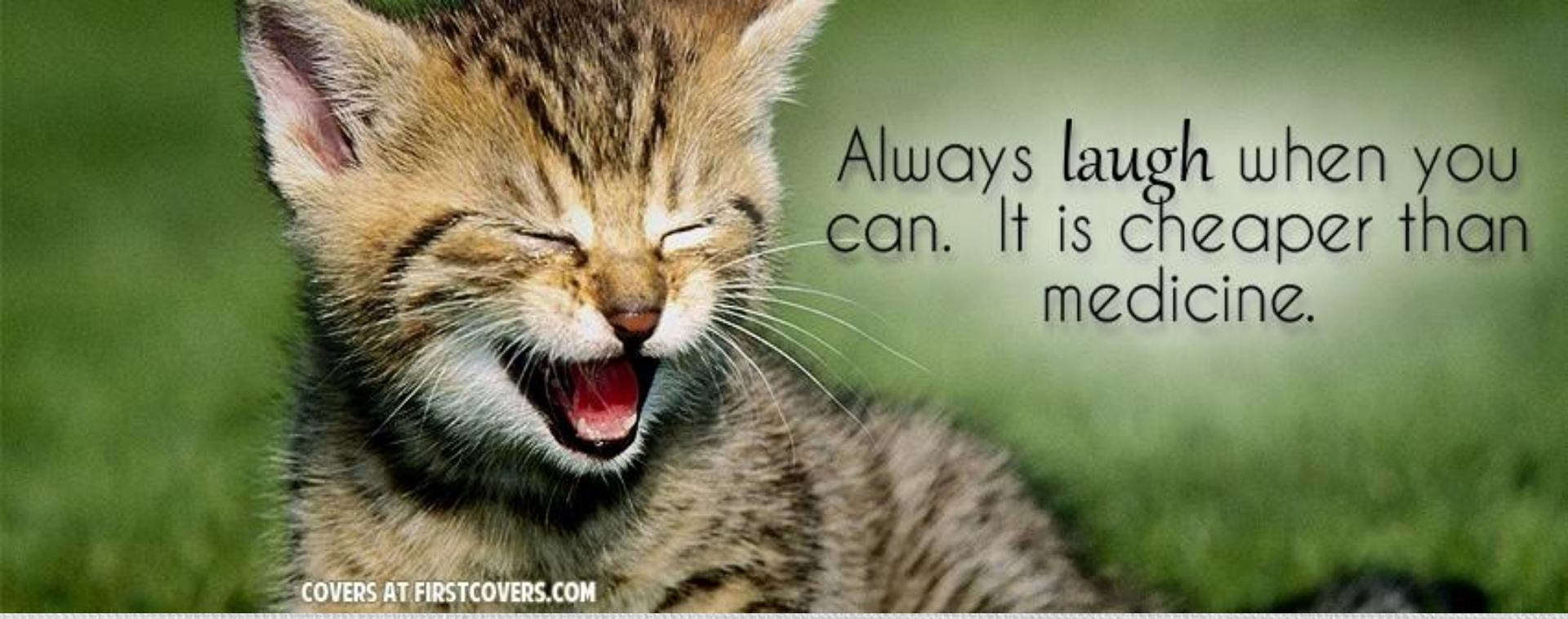

Molecular Evolutionary Genetics Analysis, Version 2.1
File Data Distances Phylogeny Tests Windows Help
UPGMA...
Neighbor-Joining (NJ)...
Minimum Evolution (ME)...
Maximum Parsimony (MP)...
Display Saved Tree Session...
Click me to activate a da
Go to the MEGA2 web
Citing MEGA2 in publi
Data File
L:\transfer\MEGAtransfer\us90io.MEG
Title
exported by MacClade from file USTIOC90_BICio.clade
11:46:21 AM


Select File and Format
Data file to convert:
C:\Documents and Settings\Magdalen\Desktop\clustalw-20050124-16330
Data Format:
.aln (CLUSTAL)
OK
Cancel
Help


M4.1 (Beta 3): Analysis Preferences
Options Summary
Test of Phylogeny
Option
Selection
Data Type
Amino acid
Analysis
Phylogeny reconstruction
Tree Inference
->Method
Neighbor-Joining
->Phylogeny Test and options
Bootstrap (500 replicates; seed=64238)
Include Sites
->Gaps/Missing Data
Complete Deletion
Substitution Model
->Model
Amino: p-distance
->Substitutions to Include
All
->Pattern among Lineages
Same (Homogeneous)
->Rates among sites
Uniform rates
Compute
Cancel
Help


M4.1 (Beta 3): TreeExplorer (C:\Documents and Settings\ml16\My Documents\PPI\PPI-2\seq\c...
File
Image
Subtree
View
Compute
Caption
Help
Original tree
Bootstrap consensus tree
32
70
62
100
50
51
51
100
100
HopA2 Psy 5D4198
HopA2 Psy B-301D
HopA2 Psy PDDCC3907
HopA2 Psy BK034
HopA2 Paf B143
HopA2 Psy 464
HopA2 Psy B6
HopA2 Psy B5
HopA2 Pmp 5795
HopA1 Pst DC3000
HopA1 Psy 61
HopA1 Psy 226
SBL = 0.94494156


"Our prime purpose in this life is to help others. And if you can't help them, at least don't hurt them."



Thanks a lot

with my Best Regards and My Best wishes

Amira A. AL-Hosary
E-mail: Amiraelhosary @yahoo.com
Mob. (002) 01004477501