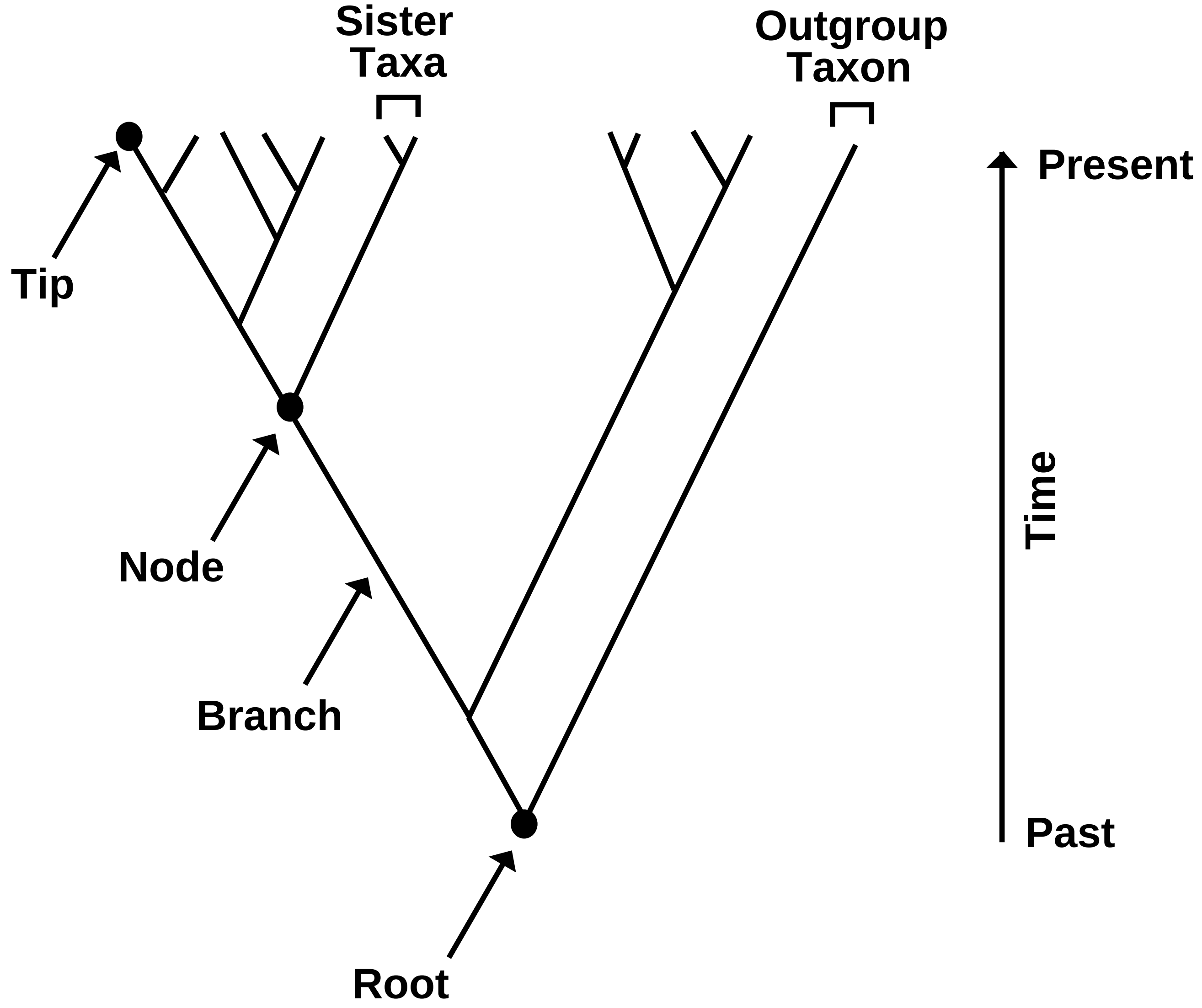
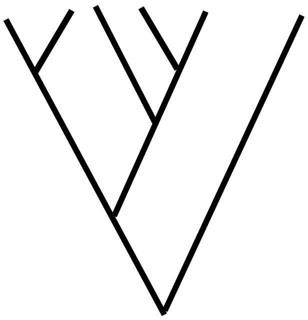
Phylogenetic Reconstruction

Phylogeny: The history of descent from a group of taxa such as species from their common ancestors, including the order of branching and sometimes absolute ages of divergence; also applied to the genealogy of genes derived from a common ancestral gene.

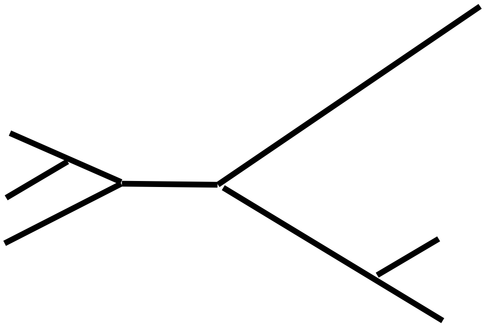
-- Futuyma 1998

PHYLOGENETIC TREE:





Rooted Tree



Unrooted Tree

"Rooted" trees make a statement about the passage of time.

Nodes near the bottom of a rooted tree represent older divergences between two lineages.

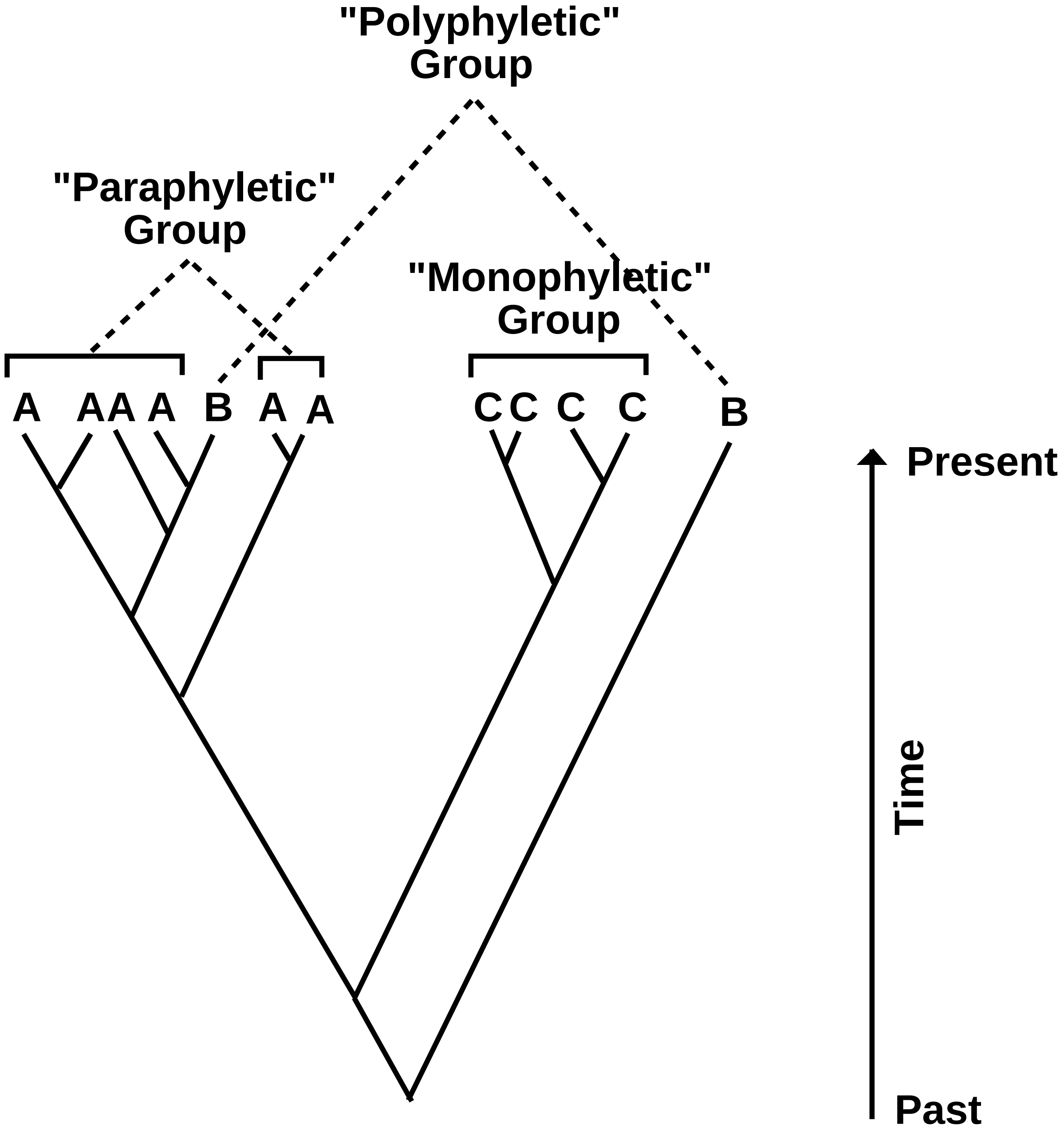
Nodes near the top of the tree represent recent divergences between two closely related lineages.

The root of a tree is often determined by an "outgroup".

An outgroup is *presumed* to be outside of the group of interest (i.e., it diverged prior to the taxa in a phylogenetic analysis).

An unrooted tree makes no claim about which of the divergences is oldest.

Phylogenetic trees sometimes do and sometimes do not correspond to the Linnean classification system.

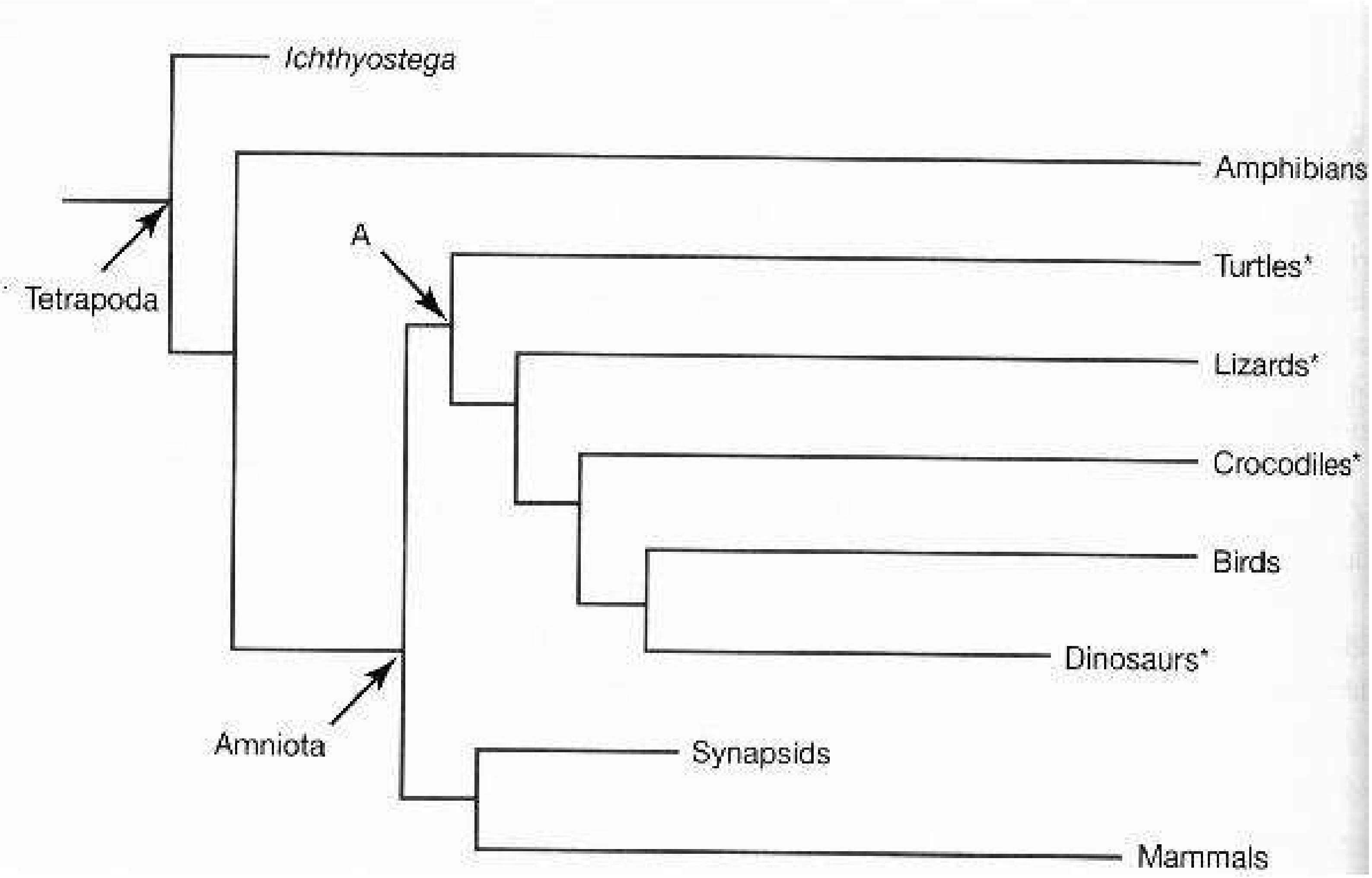


For instance, mammals make a good phylogenetic group (or clade), because all mammals are more closely related to each other than they are to any other taxon.

➡ Monophyletic group

Conversely, reptiles do not represent all the descendants of their common ancestor. Birds and mammals are also descended from the common ancestor of all reptiles (living and extinct). This means that some reptiles (e.g. crocodiles) may be more closely related to a non-reptile (e.g. birds) than they are to other reptiles.

➡ Paraphyletic group



Finally, some systematic groupings are completely artificial and based only on superficial resemblance and convergent evolution rather than true relatedness. For example, Linnaeus grouped together several unrelated worms into the artificial group "Vermes".

➡ Polyphyletic group

Should the classification scheme we use be based purely on monophyletic groups?

Choice of Characters

Phylogenetic trees may be based on many different forms of data: morphological, physiological, biochemical, molecular.

For any type of character, there are four attributes that are key to a successful phylogenetic analysis:

- **Numbers:** There should be a large number of characters.
- **Independence:** The characters should evolve independently of one another.
- **Homologous:** The characters must be derived from the same character in a common ancestor.
- **Low risk of convergence:** The characters should reflect common descent not "homoplasy".

Homoplasy: Similarity in the characters found in different species that is due to convergent evolution, parallelism, or reversal -- not common descent.

-- Freeman and Herron 1998

We are going to focus on reconstructing phylogenies from molecular data, specifically from DNA sequences.

Attributes of molecular data:

- **Numbers:** Large numbers of characters can be generated.
- **Independence:** Basepairs *largely* evolve independently of one another.
- **Homologous:** Sequences can be aligned using many different taxa to attempt to place basepairs in homologous positions.
- **Low risk of convergence:** No!!

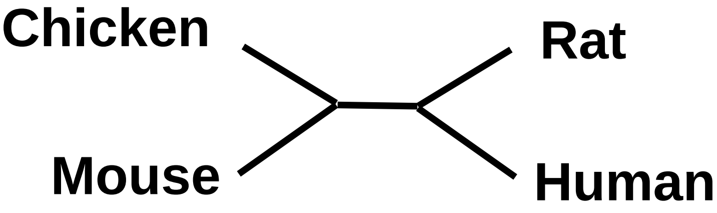
The main problem with using molecular data is that there is a high risk of homoplasy.

That is, if two sequences both have an adenine at a particular site, we do not know if this is because both descended from a common ancestor that had an adenine or because adenine happened to arise independently in both lineages.

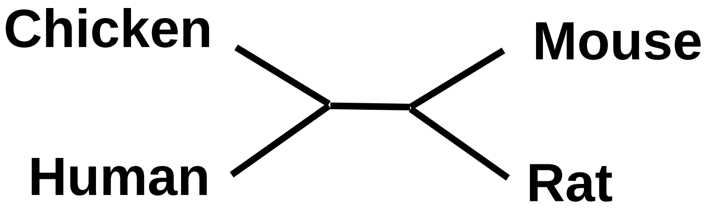
EXAMPLE OF HOMOPLASY

Sites 1220-1225 in cytochrome oxidase 1 (mtDNA)

Chicken:	**AATAGT**
Mouse:	**AATAGT**
Rat:	**TATGGT**
Human:	**TATGGT**



Tree based on these sites



True tree

➡ The risk of homoplasy is greatest if the DNA sequence evolves rapidly relative to the species divergences being examined.

Choice of Trees

There are several different *criteria* and *algorithms* used to reconstruct phylogenetic trees.

We'll focus on the conceptual criteria used in three different methods:

- Parsimony analysis
- Distance analysis
- Maximum Likelihood analysis

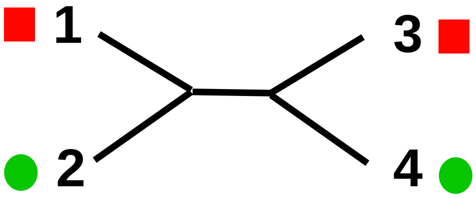
1. Parsimony analysis

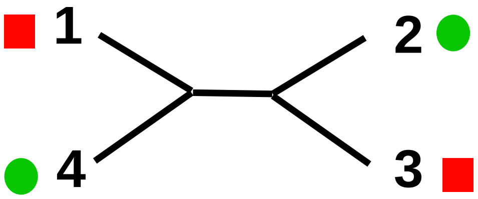
> "The principle of this method is to infer the amino acid or nucleotide sequences of the ancestral species and choose a tree that requires the *minimum number of mutational changes*".
>
> -- Nei (1987)

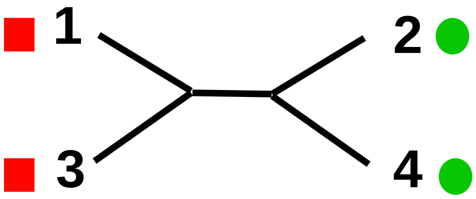
Parsimony's guiding principle is Occam's razor, the philosophical principle that it is preferable to choose the simplest of alternative explanations.

In practise, this means determining the tree (or trees) that require the *fewest number of mutations* in order to explain the data that you have.

Taxa:	1	2	3	4
Character:	🟥	🟢	🟥	🟢







With multiple characters, the minimum number of mutations on each possible tree has to be determined.

In Figure 17.13, Ridley provides an *algorithm* for determining some (but not all) possible ancestral sequences and for finding the smallest number of mutations required by a tree.

More sophisticated algorithms exist for searching all possible trees and all possible ancestral states.

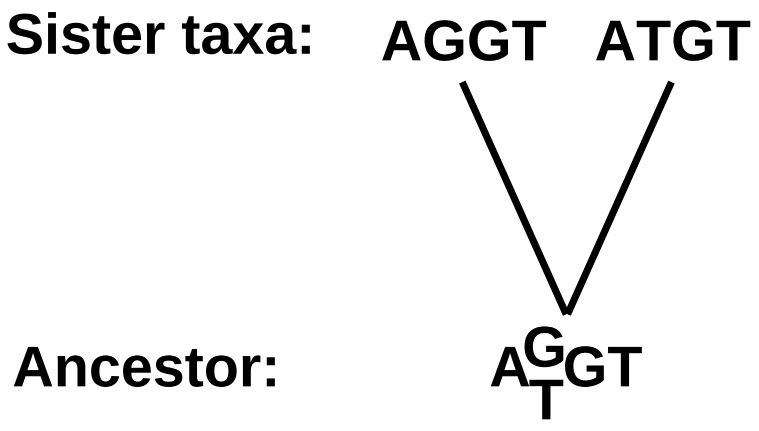
An algorithm for finding the minimum number of mutations on a tree

Part A: Determining ancestral states

(1) Pick a pair of sister taxa.

(2) Write an inferred sequence for the most recent common ancestor of these two taxa at the node connecting them. Site by site, determine:

- if the basepairs are the same in both sister taxa. Add this basepair to the ancestral sequence.
- if the basepairs differ between the sister taxa. Add both basepairs to the ancestral sequence (stacked on top of each other).



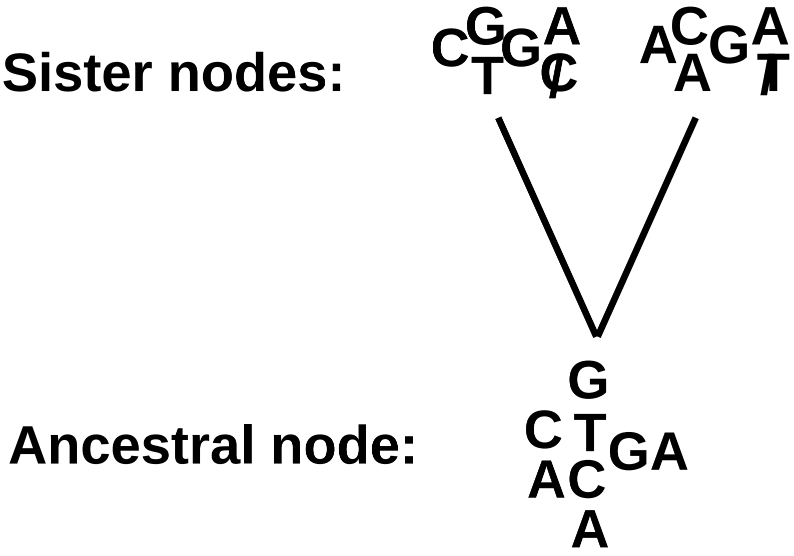
Sister taxa: **AGGT** **ATGT**

Ancestor: **A$\substack{G\\T}$GT**

(3) Now ignore the original sister taxa, and treat their ancestral node as a new taxon.

(4) Repeat steps (1)-(3), with the following additional instruction for step (2). When comparing sites that have stacks of possible basepairs:

- If the stacks have no basepairs in common, add all basepairs to the ancestral sequence (stacked on top of each other).
- If the stacks have a basepair in common, strike out all other members of the two stacks. Then move back up the tree to resolve the stacks in previous parts of the tree if necessary.



(5) Once all ancestral nodes have been determined, resolve any remaining stacks, being careful to choose the same basepair at a site on both sides of a branch whenever possible.

NOTE: There are often several different possible sets of ancestral states that would give the same minimum number of mutations.

Part B: Counting the minimum number of mutations

(1) Along each branch, make a mark for each difference between the two sequences at either end of the branch.

(2) Count the total number of marks on the tree.

➡ Minimum number of mutations required to explain this sequence data with this tree.

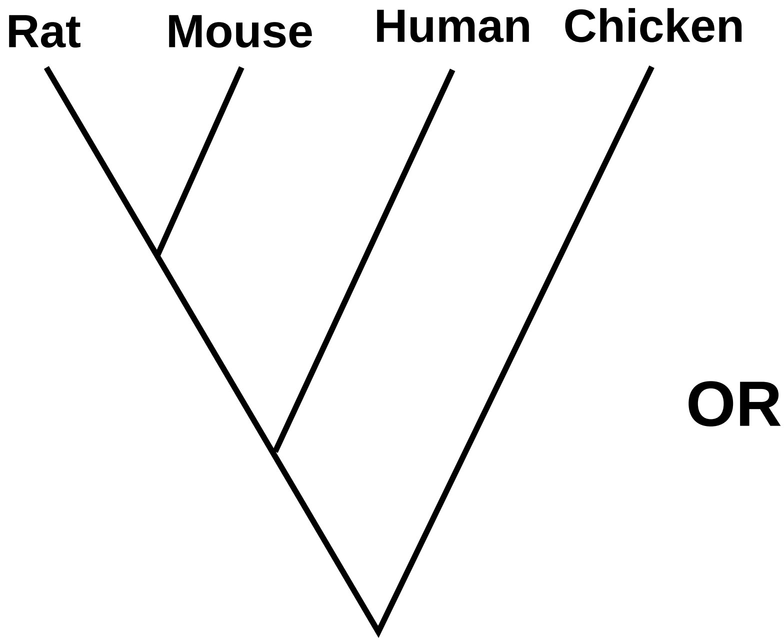
(Figure 17.13 describes this method Ridley.)

Example

For the following data set, which tree is most parsimonious?

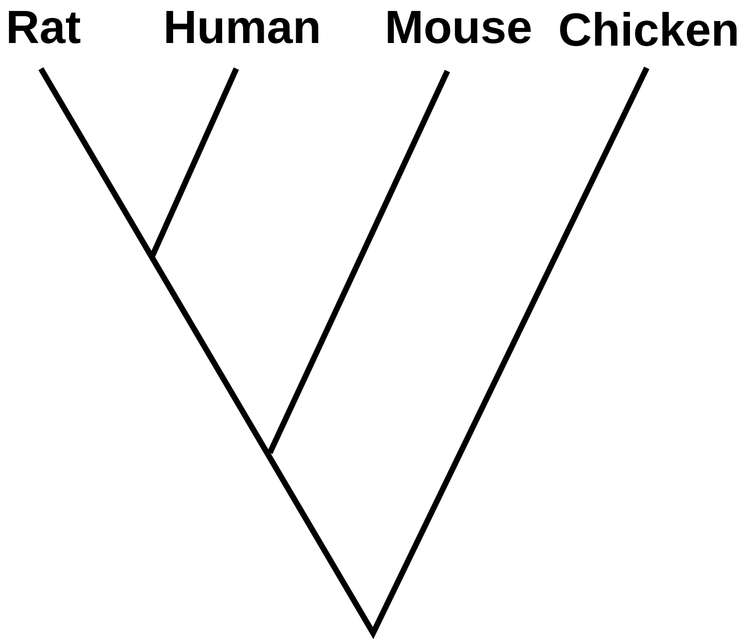
Sites 819-824 in cytochrome oxidase 1 (mtDNA)

Chicken:	**ACCCAT**
Mouse:	**ATGACA**
Rat:	**ATGACA**
Human:	**ACCAAA**



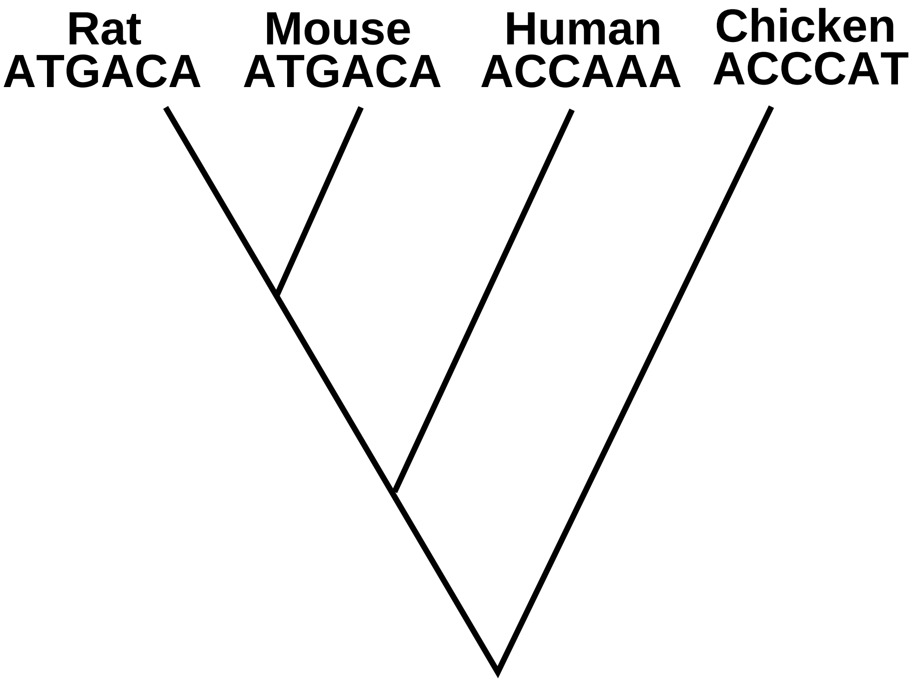
TREE A

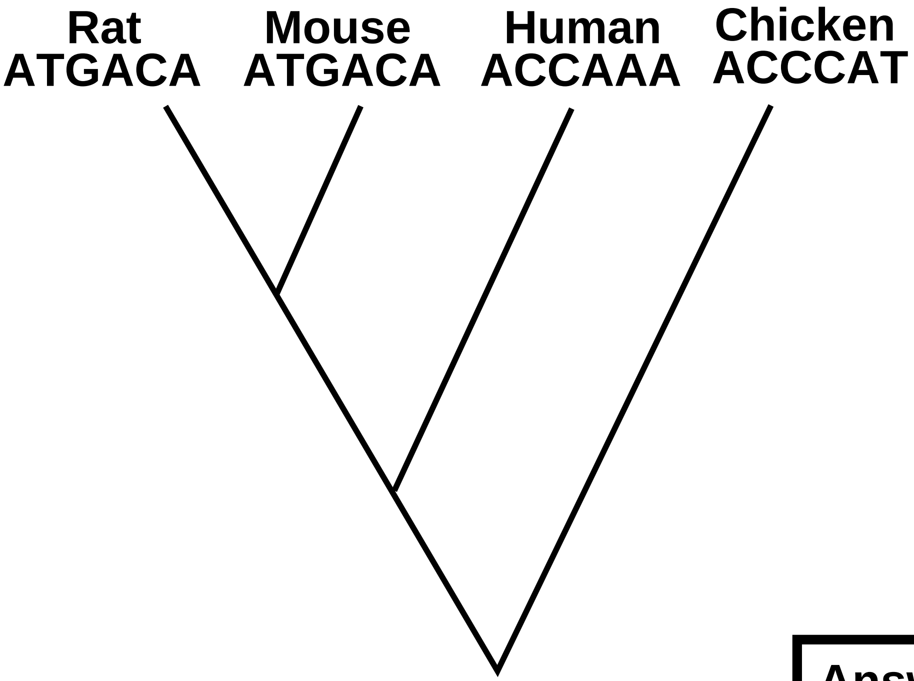
OR



TREE B

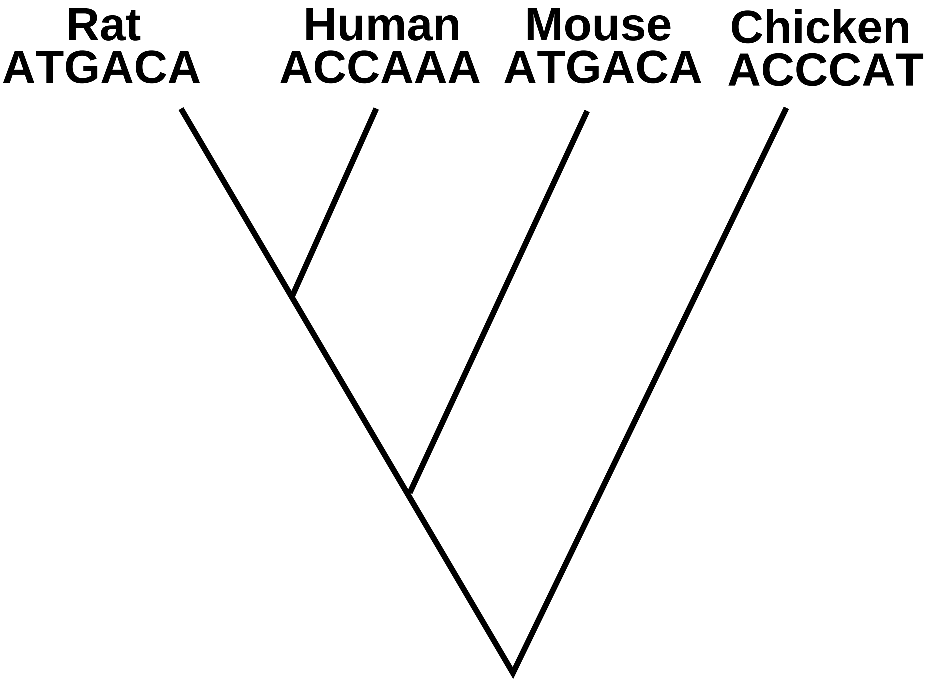
Worksheet for tree A

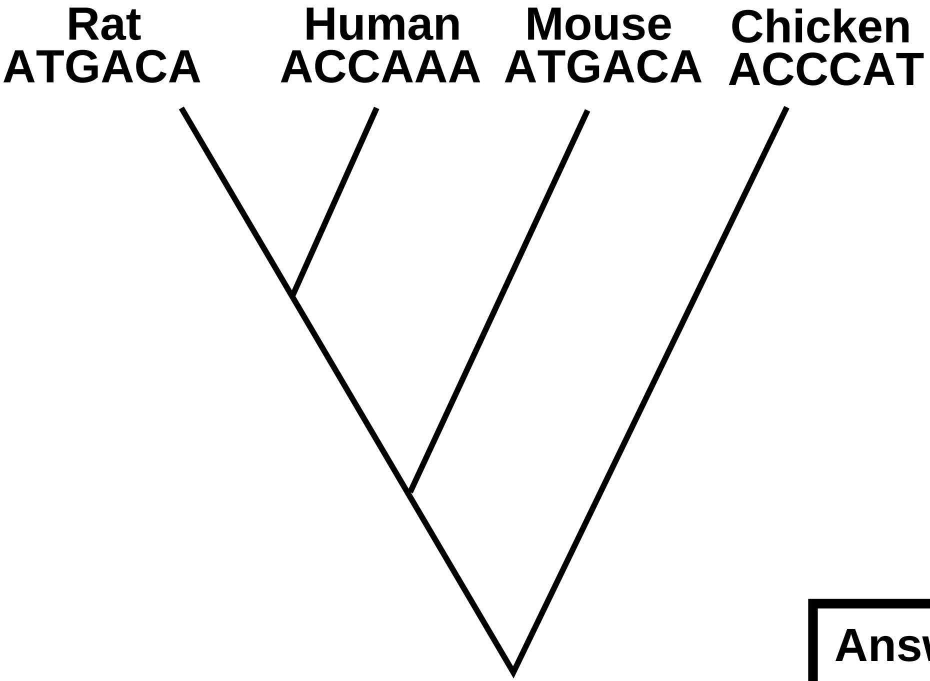




Answer:
Five mutations

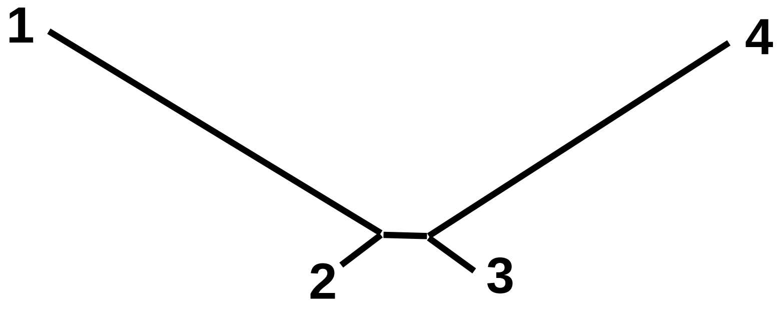
Worksheet for tree B





Answer:
Eight mutations

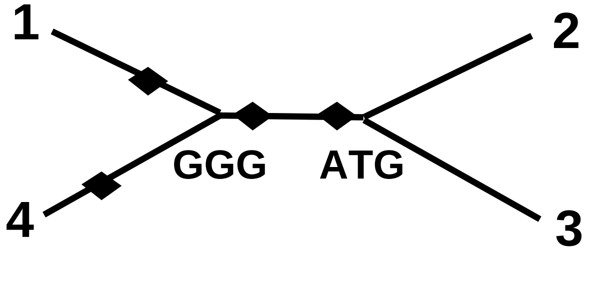
Interestingly, parsimony can fail as a method, because evolution may take more steps than absolutely necessary to get from the ancestral sequence to the current sequences.



True Tree

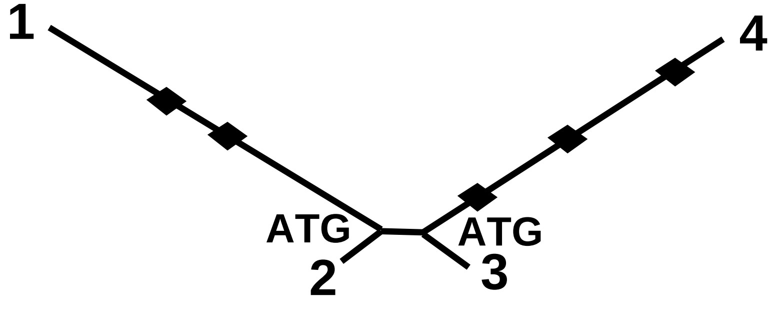
Mutations on both long branches may be common.
Mutations on internal branch may be rare.

Species 1: CGG
Species 2: ATG
Species 3: ATG
Species 4: GGT



Tree requires:
Four mutations

Most Parsimonious Tree



Tree requires:
Five mutations

True Tree

Advantages of Parsimony Analysis

- Conceptually easy to understand
- Straightforward to calculate the length of a tree
- Accurate if few evolutionary changes have occurred (homoplasy unlikely)

Disdvantages of Parsimony Analysis

- Underestimates the true amount of evolutionary change
- Can strongly favor the wrong tree ("positively misleading")

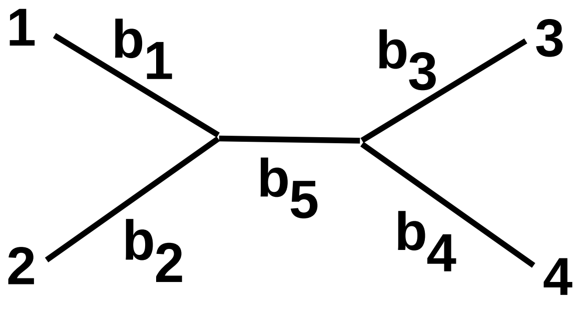
2. Distance analyses

Species comparisons are often presented as distances between each pair of species (e.g. the number of sequence differences).

Sometimes only distance data are available, such as the strength of DNA-DNA hybridization.

Distance methods choose a tree on the basis of how well it coincides with the observed distances between every pair of species.

For any particular tree, the expected distance between two taxa can be found by summing the branch lengths separating the two taxa:



For example, the expected distance between species 1 and species 3 is $d_{13} = b_1 + b_3 + b_5$.

Distance methods attempt to minimize the discrepancy between the observed distances (D_{ij}) and the expected distances (d_{ij}), e.g. by minimizing:

$$\sum w_{ij} (D_{ij} - d_{ij})^2,$$

where w_{ij} is a weighting term that can be used, for example, to diminish the importance of distantly related taxa.

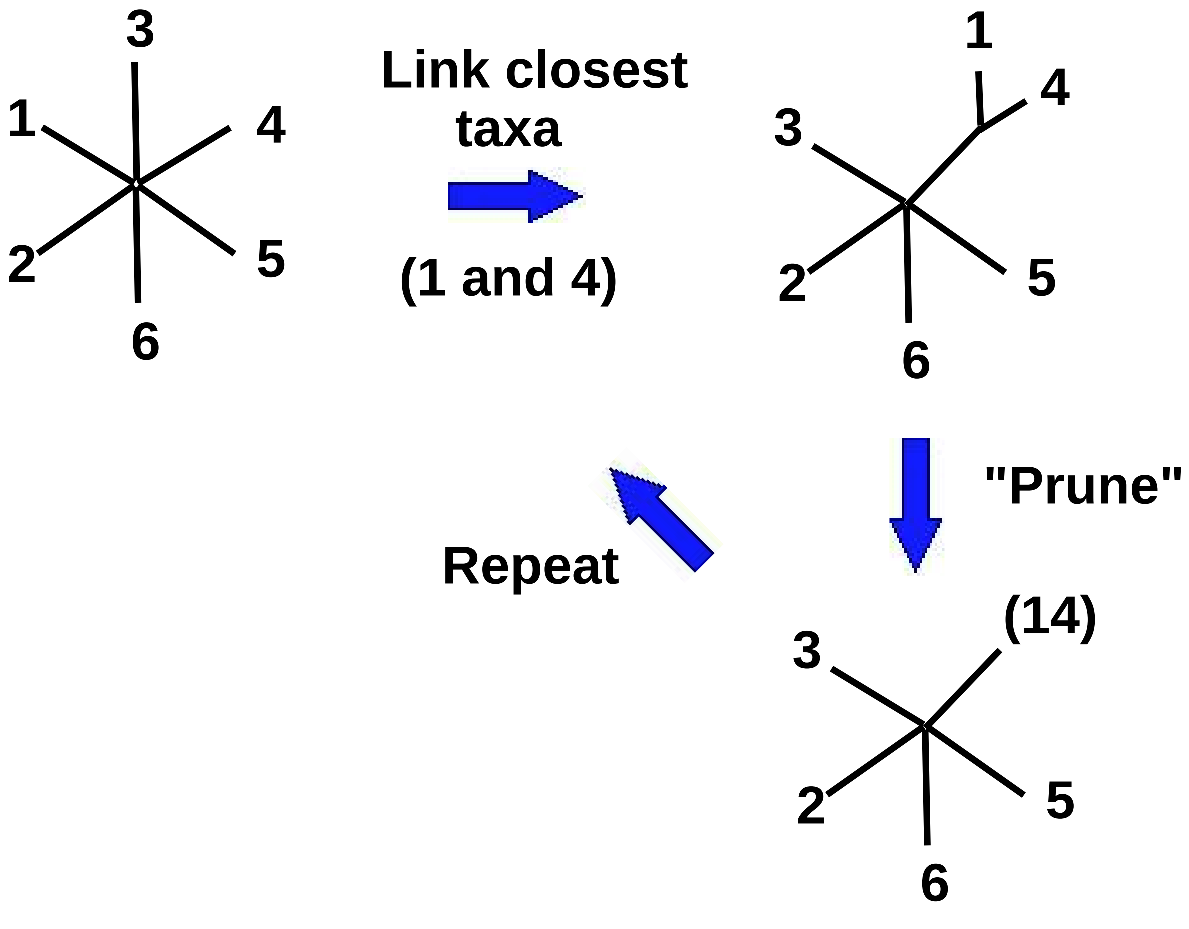
A particularly common distance method is neighbor-joining.

Neighbor-joining is an algorithm, meaning that one follows a recipe to get the tree rather than figuring out how to mimimize functions like the one above.

Neighbor-joining: Starting from a star-like tree, the two closest taxa are placed together as neighbors. [Aside: The distances are first corrected to take into account potential differences in rates between the taxa.]

These two taxa are then represented by their common ancestral node and removed from the analysis.

The procedure is repeated until the full tree is resolved.



Advantages of Distance Analysis

- The only method available for distance data
- Fast (especially neighbor-joining)
- Better able to handle large data sets

Disdvantages of Distance Analysis

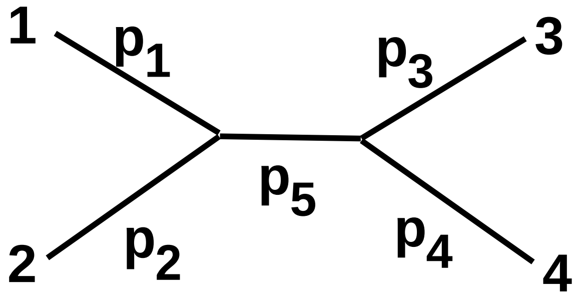
- Distances can hide convergent evolution (homoplasy)
- Distance methods can generate incorrect trees (when distances do not scale with time)

3. Maximum likelihood analysis

In a maximum likelihood analysis, a specific model is used to determine the probability that a given base substitution will occur along a given branch on a tree.

The maximum likelihood tree is the one that can generate the observed data with the highest probability.

For any one site, the likelihood of observing the data given a particular tree and a particular model of sequence evolution is calculated:



Likelihood of this tree:

The total probability of making each transition on the tree (= $p_1 p_2 p_3 p_4 p_5$), summed over all possible internal nodes.

The likelihood for the whole sequence is then calculated as the product of the likelihoods for each site.

Advantages of Maximum Likelihood Analysis

- Extremely flexible (any model can be used)
- Statistically justifiable
- Will always infer the right tree given enough data (if the model is correct)

Disdvantages of Maximum Likelihood Analysis

- Impossible to know if the model is correct
- Computer intensive
- Practically impossible with many taxa

Evaluating a Tree

Frequently, many trees are optimal or near optimal on the basis of a criterion. Generating a "best" tree does not say how much better it is than other trees.

One of the most common methods used to evaluate the support in the data for the phylogenetic relationships shown on a tree is the *bootstrap* resampling procedure.

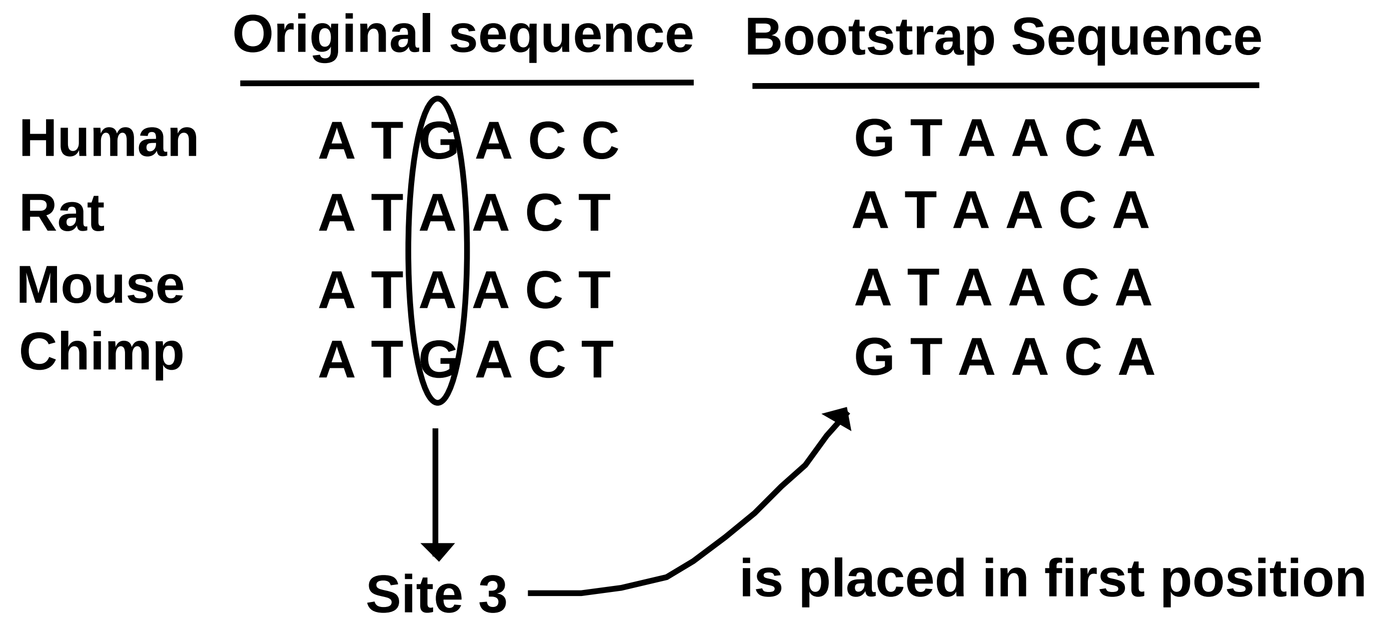
The bootstrap technique involves generating artificial sequences by randomly sampling sites from the original sequences with replacement.

This randomly generated data set has the same sequence length but a slightly different composition (i.e some sites will be oversampled and others not).

For example, consider a simple sequence with 6 sites.

Say that the first site chosen randomly is 3. For each species, site 3 is placed in the first position of the bootstrap sequence.

This is repeated until the bootstrap sequence is also 6 bp long.



	Original sequence	Bootstrap Sequence
Human	A T G A C C	G T A A C A
Rat	A T A A C T	A T A A C A
Mouse	A T A A C T	A T A A C A
Chimp	A T G A C T	G T A A C A

Site 3 **is placed in first position**

(Then the next five randomly chosen sites: 2, 1, 1, 5, 4, are placed in the next five positions.)

The "best" tree is then determined from the bootstrap sequences, using the same method as used with the original data set.

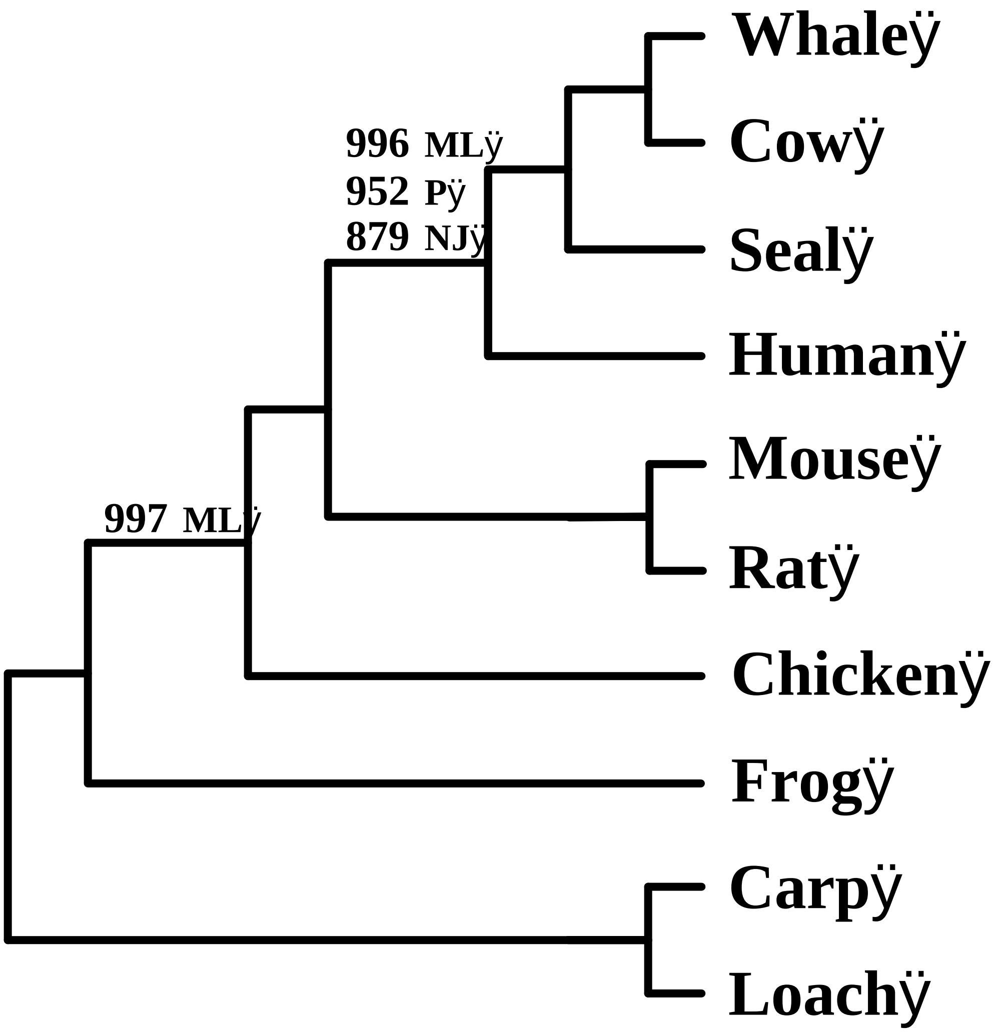
This whole process is repeated at least 100 times.

The number of times that a clade is seen among the bootstrap trees is reported.

The more often a clade is present among the bootstrap trees, the more strongly the data support that clade, because the result is insensitive to which basepairs happen to be sampled.

EXAMPLE:

Cummings et al (1995) used the entire mitochondrial genomes of ten vertebrates and obtained the following tree using parsimony (P), neighbor joining (NJ), and maximum likelihood (ML) methods:



All clades were supported in 1000/1000 bootstrap data sets, with the exception of the two clades shown, which still had strong support.

Conclusion

We have focused on the criteria used to build trees.

The criteria have been refined to take into account several other factors including:

- transition/transversion bias
- mutation rate heterogeneity across a sequence
- rate variation along a tree.

In practise, efficient algorithms have to be used in order to evaluate all the possible arrangements of taxa on trees.

Several computer programs are available that implement these phylogenetic methods (the commonly used, general-purpose programs are Phylip, PAUP, and MEGA).

For more information, check out this web-site of phylogenetic resources.

SOURCES:

- **Hillis, Moritz, and Mable (1996) Molecular Systematics. Sinauer Associates, MA.**

- **Felsenstein (1988) Phylogenies from molecular sequences: Inference and reliability. Annual Review of Genetics 22:521-565.**

- Futuyma (1998) Evolutionary Biology. Sinauer Associates, MA.

- Freeman and Herron (1998) Evolutionary Analysis. Prentise Hall.