Introduction to phylogenetics

by

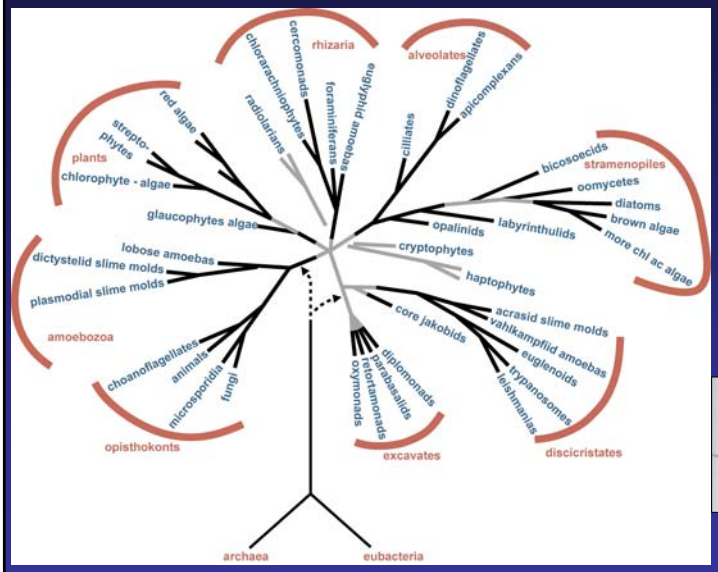
Peik Haugen

University of Tromsø, Norway
Department of Molecular Cell Biology
and
Department of Molecular Biotechnology

University of Iowa
Department of Biological Sciences and
Center for Comparative Genomics

Intron origin and evolution

Tree of life



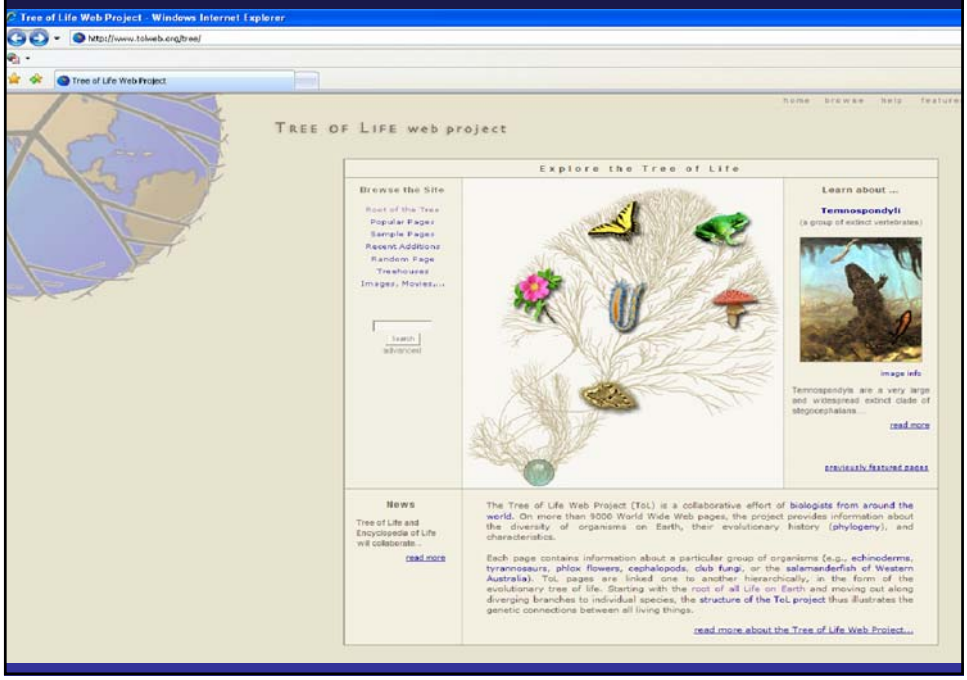
Eukaryotes
(1.8 billion years)

Classical tree



Prokaryotes (3.5 billion years)

Tree of life initiative



Phylogenetics : Reconstruction of the evolutionary relationships (that is, the phylogeny) of a group of taxa, such as species

From: Nature glossary

Molecular phylogenetics : Reconstruction of the evolutionary relationships of genes and taxa as inferred from molecular data and the methodology of tree reconstruction.

Mechanisms of evolution : The primary cause of evolution is the mutational change of genes

Examples of mutations:
- substitutions
- insertions/deletions
- recombination
- gene conversion
- inversions

1. Deleterious
2. Advantageous
3. Neutral (no effect on fitness)

Advantages to molecular phylogenies

Organisms with high degrees of molecular similarity are expected to be more closely related than those that are dissimilar.

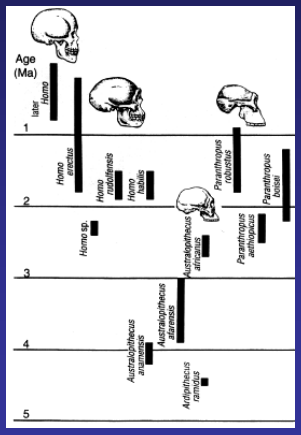
Before: rely on comparisons of phenotypes (how organisms look) to infer genotypes (the genes that gave rise to their physical appearance).

Now: use DNA and protein sequences many of which are essential to all living things.

Molecular approaches to generating phylogenies can usually be relied on to group organisms correctly.

Classical evolutionists

Use comparative morphology and comparative physiology

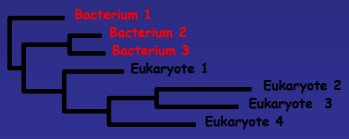


Todays evolutionists

Most use molecular data

```
GAA-GGA-AGC-TCC-TGG-TTA-CTC-CTG-GGA-TCC
GAG-GGT-TCC-AGC-TAT-CTA-TTA-ATT-GGT-AGC
GAC-GGC-AGT-GCA-TGG-TTG-CTT-TTG-GGC-AGT
GAT-GGG-TCA-GCT-TAC-CTC-CTG-GCC-GGG-TCA
```

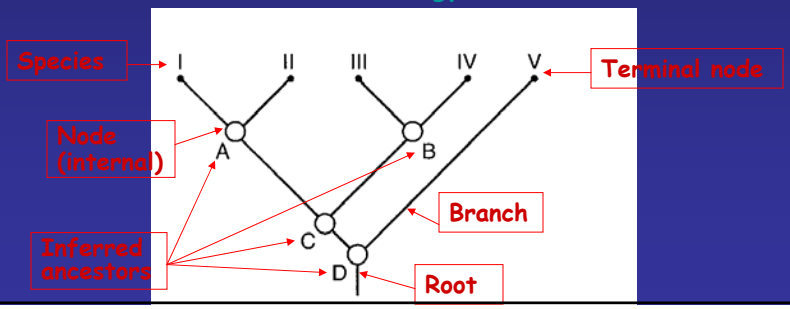
```
Glu-Gly-Ser-Ser-Trp-Leu-Leu-Leu-Gly-Ser
Glu-Gly-Ser-Ser-Tyr-Leu-Leu-Ile-Gly-Ser
Asp-Gly-Ser-Ala-Trp-Leu-Leu-Leu-Gly-Ser
Asp-Gly-Ser-Ala-Tyr-Leu-Leu-Ala-Gly-Ser
```



Phylogenetic Trees

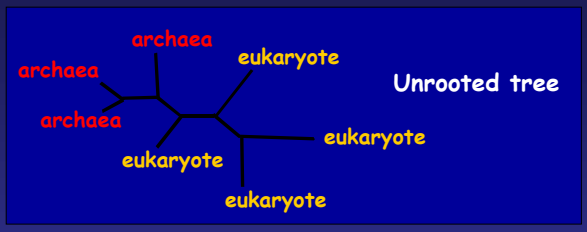
A graphical representation of the evolutionary relationship among three or more genes or organisms

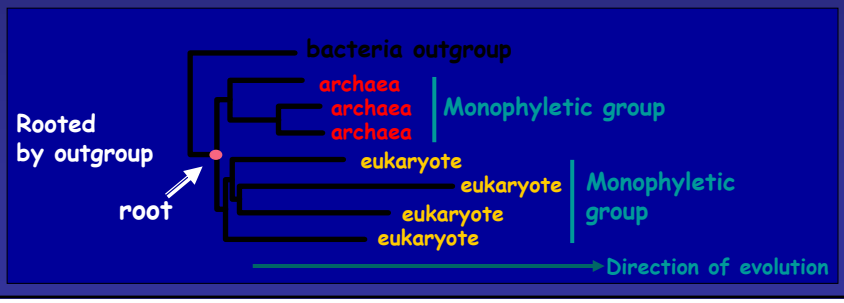
Terminology:



Phylogenetic Trees

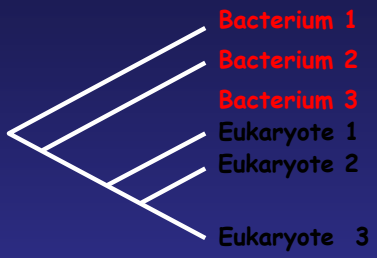
Terminology: Rooted and Unrooted Trees



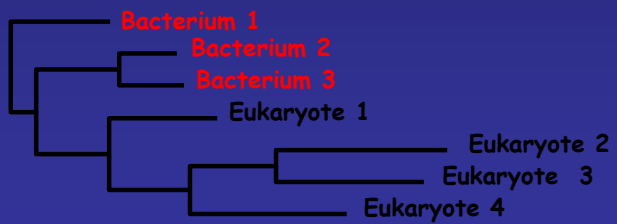


Cladograms and phylograms

Terminology:



Cladograms show branching order - branch lengths are meaningless (unscaled trees)



Phylograms show branch order and branch lengths (additive trees, scaled trees or metric trees)

Terminology:

Homology : Common ancestry of genes that is recent enough such that sequence information has been retained enough to be used in phylogenetic analyses

Orthologous genes: homologous genes that have started a separate evolution because of speciation.

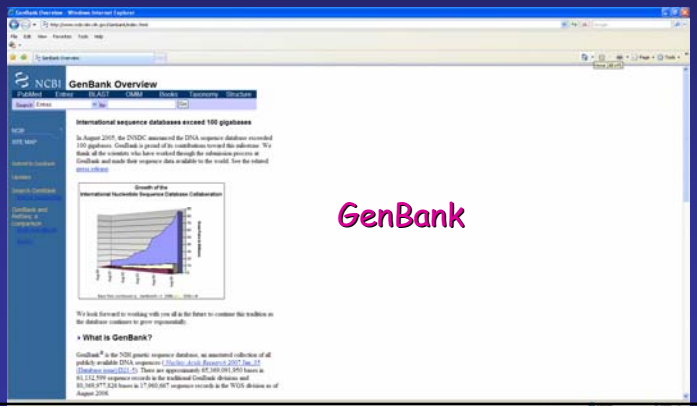
Paralogous genes: genes originating from a gene duplication event recent enough to reveal common ancestry

Tree-building

1. Sequence retrieval
2. Alignment
3. Phylogenetic analysis

Sequences : DNA, RNA, protein

DNA sequencing



GenBank

Alignments: can be easy or difficult

```
GCGGCCCA TCAGGTAGTT GGTGG
GCGGCCCA TCAGGTAGTT GGTGG
GCGTTCCA TCAGCTGGTT GGTGG
GCGTCCCA TCAGCTAGTT GGTGG
GCGGCGCA TTAGCTAGTT GGTGA
******** ********** *****
```

Easy

```
TTGACATG CCGGGG---A AACCG
TTGACATG CCGGTG--GT AAGCC
TTGACATG -CTAGG---A ACGCG
TTGACATG -CTAGGGAAC ACGCG
TTGACATC -CTCTG---A ACGCG
******** ?????????? *****
```

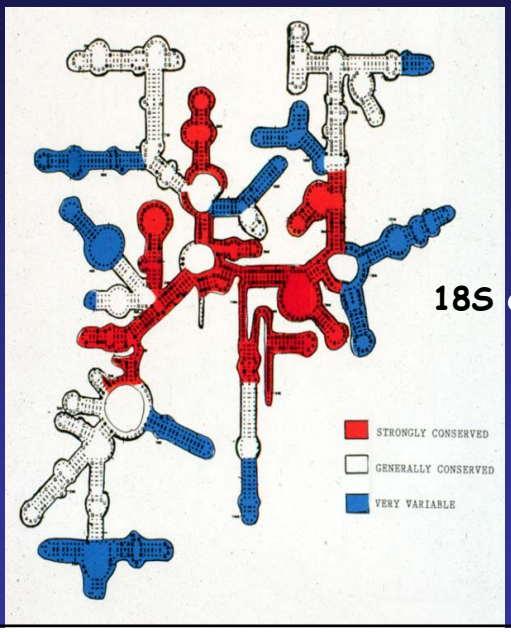
Difficult due to insertions or deletions (indels)

Alignment of 16S rRNA can be guided by secondary structure

```
•<--------------(--------------------HELIX 19--------------------)
•<---------------(22222222-000000-111111-00000-111111-0000-22222222
•Thermus ruber     UCCGAUGC-UAAAGA-CCGAAG-CUCAA-CUUCGG-GGGU-GCGUUGGA
•Th. thermophilus  UCCCAUGU-GAAAGA-CCACGG-CUCAA-CCGUGG-GGGA-GCGUGGGA
•E.coli            UCAGAUGU-GAAAUC-CCCGGG-CUCAA-CCUGGG-AACU-GCAUCUGA
•Ancyst.nidulans   UCUGUUGU-CAAAGC-GUGGGG-CUCAA-CCUCAU-ACAG-GCAAUGGA
•B.subtilis        UCUGAUGU-GAAAGC-CCCCGG-CUCAA-CCGGGG-AGGG-UCAUUGGA
•Chl.aurantiacus   UCGGCGCU-GAAAGC-GCCCCG-CUUAA-CGGGGC-GAGG-CGCGCCGA
•match             **       ***        * ** ** *                 **
```

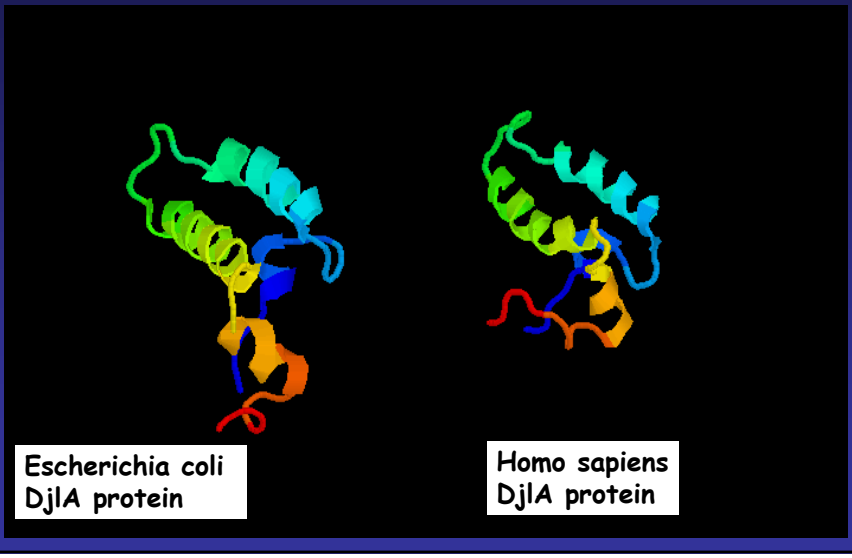
Alignment of 16S rRNA sequences from different bacteria

Small subunit ribosomal RNA



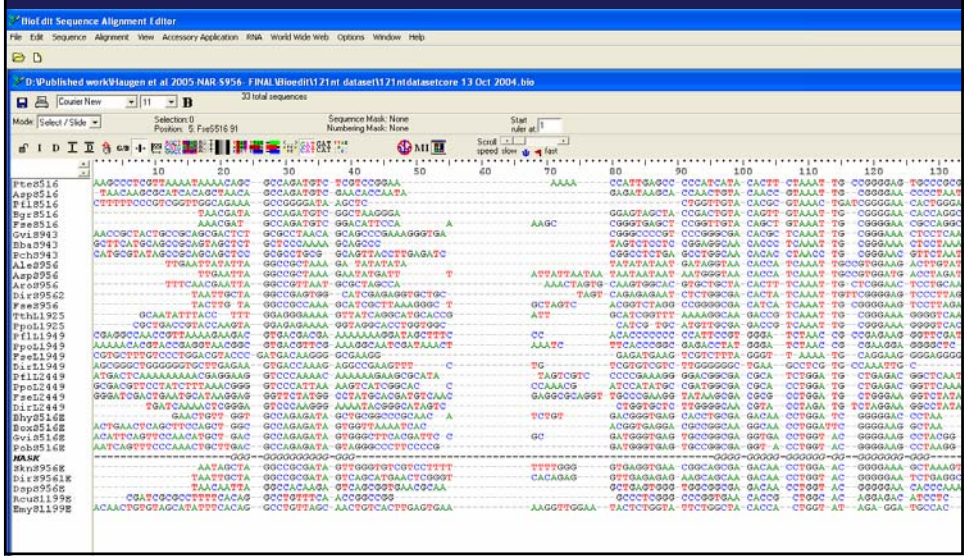
18S or 16S rRNA

Protein alignment may be guided by tertiary structure interactions



Alignment tool

BioEdit is the present-day standard tool to administrate/adjust/export alignments



Tree-building methods

Not straighforward or easy. Your end-result will likely be a hypothesis rather than the final answer.

Distance methods:

- UPGMA
- Neighbor-Joining
- Minimum evolution

Charcter-based methods:

- Maximum Parsimony
- Maximum Likelihood
- Bayesian inference of phylogeny

Distance methods

Two basic steps:

- **Construction of distance matrix** (based on an evolutionary model).
- **Construction of phylogenetic tree**.

Distance methods

Distance matrix:

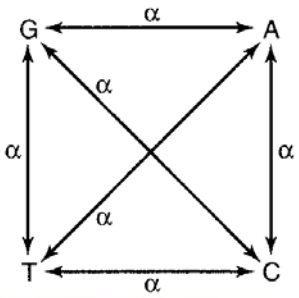
→ **Input: sequence alignment**

- **p-distance** (Uncorrected)
- **Corrected models** (e.g. Jukes and Cantor, Kimura 2, Tamura-Nei, HKY, GTR, etc)

Distance methods

Jukes-Cantor Model

- Assumption: Each nucleotide is just as likely to change into any other nucleotide. α = nucleotide substitution rate

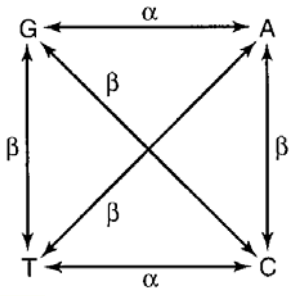


(21)

Distance methods

Kimura's Two-Parameter Model

- Take into account the different rates of transistions and transversions.



Models with even more parameters

- Since each of the four nucleotides can change into any of the other three, 12 different types of substitutions are possible:

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

- Such complex models invoke more assumptions than the one- and two-parameter models
- Sampling errors due to the comparatively small numbers of instances of each kind of change
- **Consequently**, simulation studies suggests that one- and two-parameter models often give more realistic results than more complex alternatives. (30)

Distance methods

Tree-building:

→ **Input: distance matrix**

★ **UPGMA** (Unweighted Pair-Group Method with Arithmic Mean). Clustering method.The program first find the pair of taxa with smallest distance between them and defines the branching between them as half of that distance in placing a node at the midpoint of the branch.

★ **Minimal Evolution (ME)**. The sum (S) of all branch length estimates, i.e. $S = \sum_{i}^{T} \hat{b}_i$ is computed for all or all plausible topologies, and the topology that has the smallest S value is chosen as the best tree.

★ **Neigbor Joining (NJ)**.Based on ME principle, but does not examine all possible topologies

UPGMA (Unweighted-pair-group method with arithmetic mean)(clustering method)

The oldest distance method and also the simplest of all methods for tree reconstruction.

Pairwise distance matrix:

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

1. UPGMA begins by clustering the two species with smallest distance e.g., d_{AB}
2. A and B are grouped (AB), and the new distance matrix is composed where the distances between (AB) and species C and D are calculated.
3. The species separated with the smallest distances in the new matrix are clustered to make another new composite species and the process is repeated.

Scaled branch length: branch points are positioned at a distance halfway between each species being grouped (i.e., at $d_{AB}/2$ for the first clustering)

A practical example using UPGMA:

Alignment

```
          10         20         30         40         50
A:  GTGCTGCACGG  CTCAGTATA  GCATTTACCC  TTCCATCTTC  AGATCCTGAA
B:  ACGCTGCACGG  CTCAGTGCG  GTGCTTACCC  TCCCATCTTC  AGATCCTGAA
C:  GTGCTGCACGG  CTCGGCGCA  GCATTTACCC  TCCCATCTTC  AGATCCTATC
D:  GTATCACACGA  CTCAGCGCA  GCATTTGCCC  TCCCGTCTTC  AGATCCTAAA
E:  GTATCACATAG  CTCAGCGCA  GCATTTGCCC  TCCCGTCTTC  AGATCTAAAA
```

FIGURE 4.5 *A five-way alignment of homologous DNA sequences.*

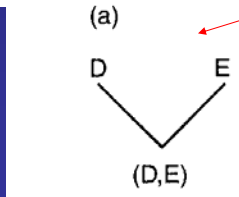
Distance matrix 1

TABLE 4.2 A pairwise distance matrix that summarizes the number of nonmatching nucleotides between all possible pairs of sequences shown in Figure 4.5.

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

Smallest distance

Tree construction 1

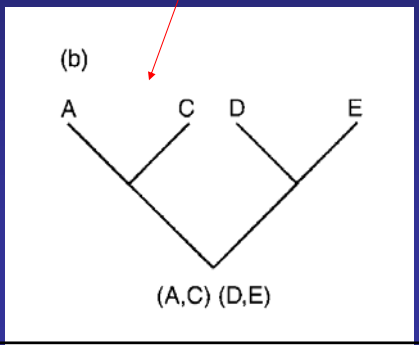


Distance matrix 2

TABLE 4.3 The distance matrix that results when species D and E of Table 4.2 are combined and considered as a single group.

Species	A	B	C
B	9	—	—
C	8	11	—
DE	13.5	16.5	11.5

Tree construction 2

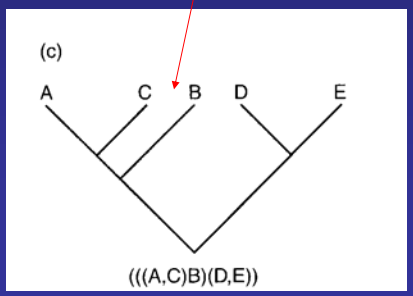


Distance matrix 3.

TABLE 4.4 The distance matrix that results when species A and C of Table 4.3 are combined and considered as a single group.

Species	B	AC
AC	10	—
DE	16.5	12.5

Tree construction 3



Distance methods

Tree-building:

→ **Input: distance matrix**

- **UPGMA** (Unweighted Pair-Group Method with Arithmic Mean). Clustering method.The program first find the pair of taxa with smallest distance between them and defines the branching between them as half of that distance in placing a node at the midpoint of the branch.
- **Minimal Evolution (ME).** The sum (S) of all branch length estimates, i.e. $S = \sum_{i}^{T} \hat{b}_i$ is computed for all or all plausible topologies, and the topology that has the smallest S value is chosen as the best tree.
- **Neigbor Joining (NJ).** Based on ME principle, but does not examine all possible topologies

Character-based methods

- **Parsimony** (based on an evolutionary model).
- **Maximum likelihood**

  Bayesian inference of phylogeny

Parsimony

Biological: Used to describe the preference to one evolutionary pathway over another on the basis of which pathway requires the smallest number of events.

Two premises that underlie biological parsimony:

1. Mutations are exceedingly rare events
2. The more unlikely events a model invokes, the less likely the model is to be correct

Maximum parsimony chooses the tree (trees) that requires the smallest number of evolutionary changes (e.g., nucleotide substitutions) to explain the observed differences among the species under study.

Informative and Uninformative Sites

What sites within a multiple sequence alignment might have useful information content for parsimony?

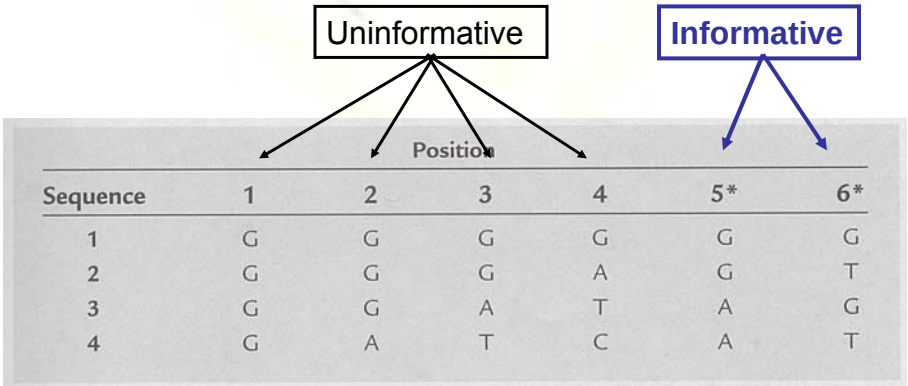
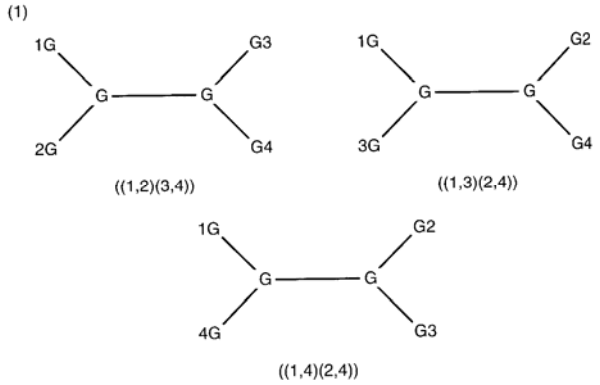
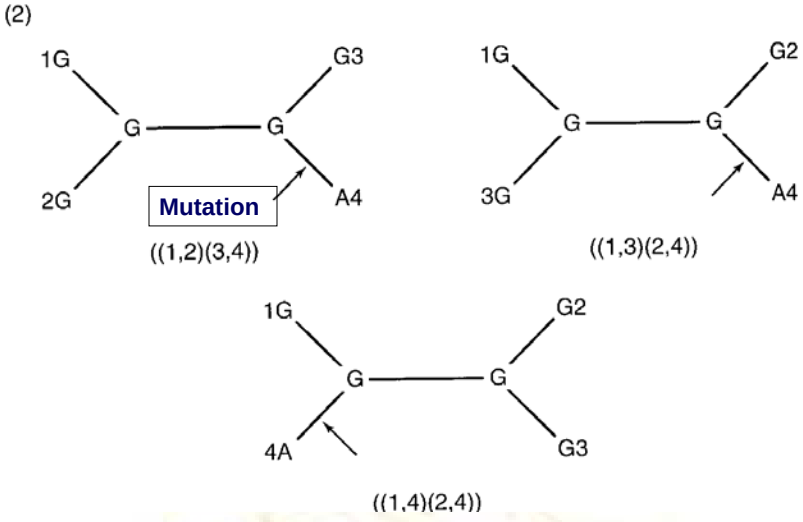


FIGURE 5.1 *An alignment of four homologous sequences, each of which is six nucleotides long. Asterisks (*) indicate the two columns within the alignment that correspond to informative sites. The four remaining sites are uninformative.*

FIGURE 5.2 *Three possible unrooted trees that describe the possible relationships between the four sequences at all 6 positions shown in Figure 5.1. Each position is numbered as in Figure 5.1. The Newick format description of each of the trees is also shown. The two internal nodes of each tree contain a possible common ancestor's sequence, and lineages where mutations must have occurred are highlighted with arrows.*

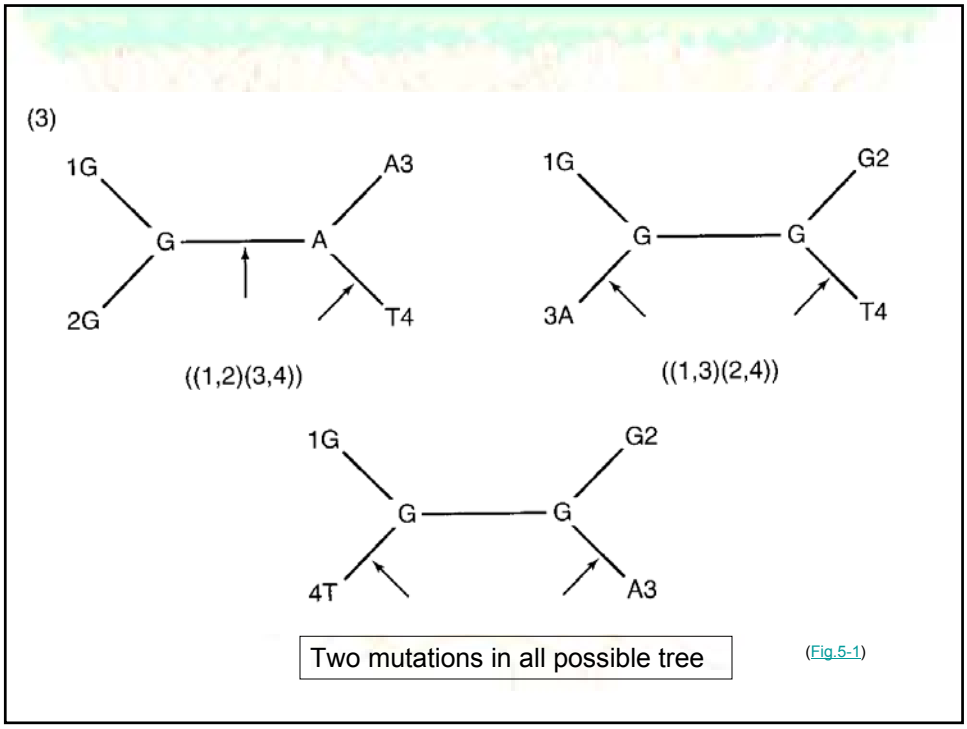


Invariant



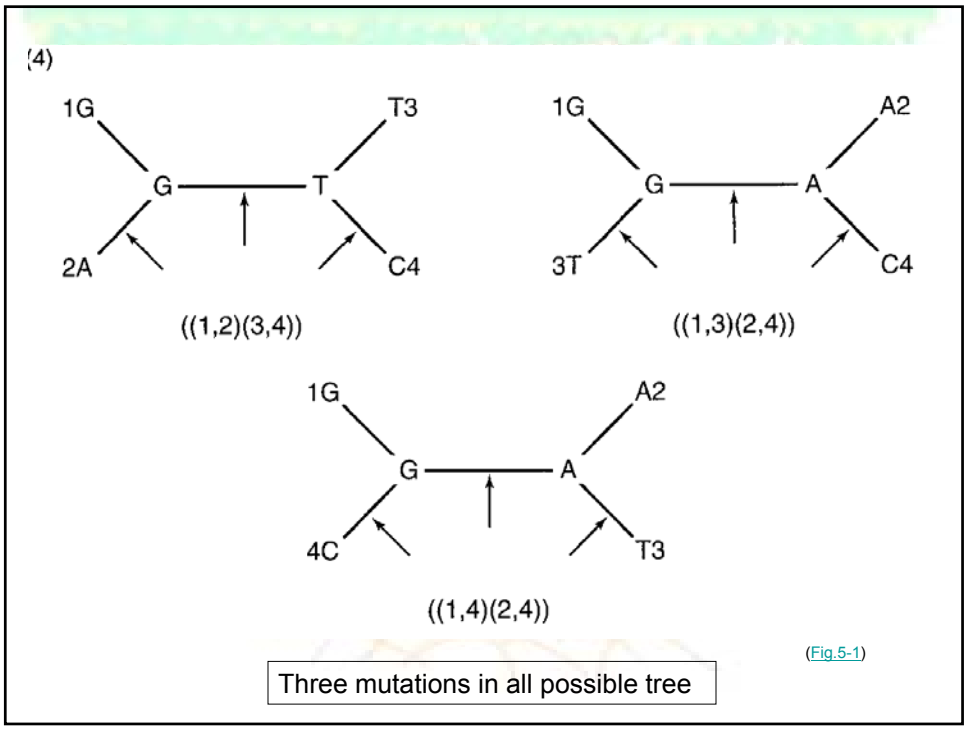
One mutation in all possible tree

(Fig.5-1)



Two mutations in all possible tree

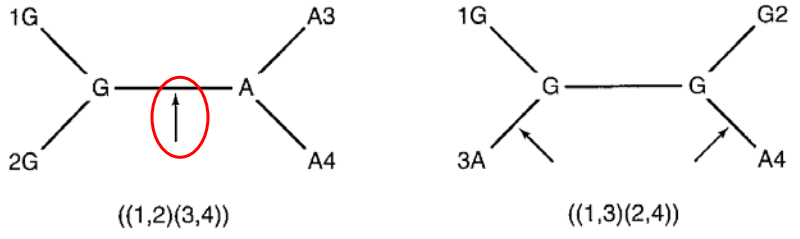
(Fig.5-1)



(Fig.5-1)

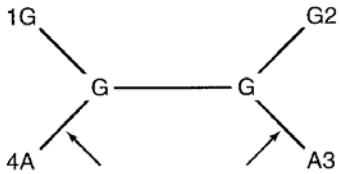
Three mutations in all possible tree

(5)



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

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

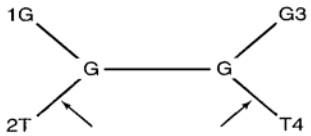


((1,4)(2,4))

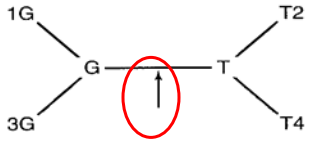
One mutation in one tree and two in two trees

(Fig.5-1)

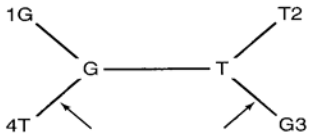
(6)



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



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



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

In general:

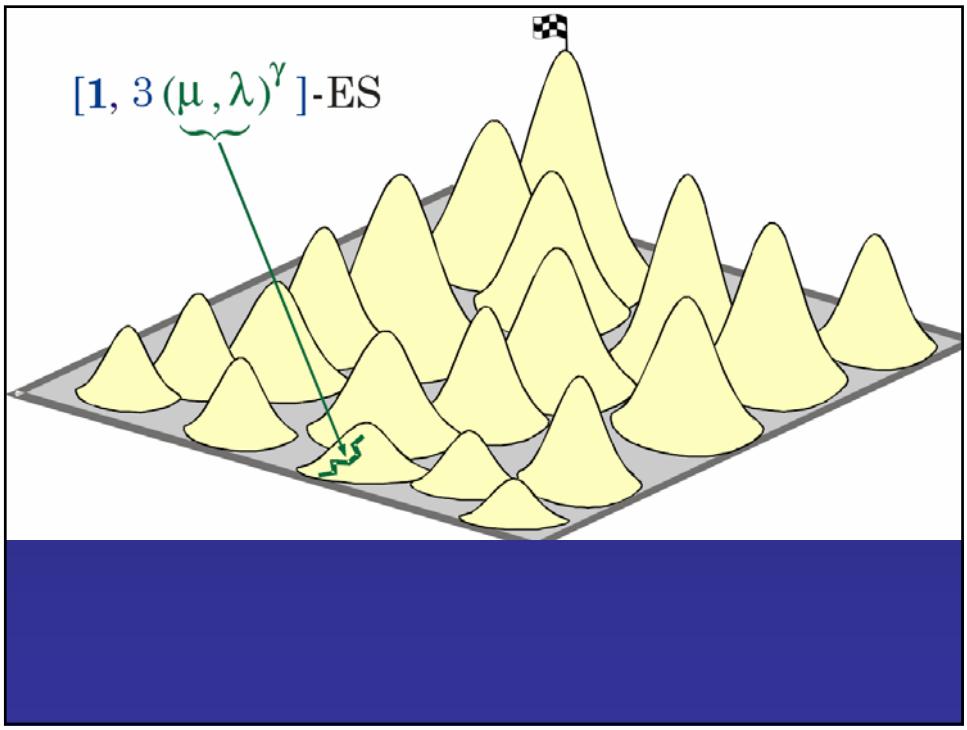
For a position to be informative regardless of how many sequences are aligned, it has to have at least two different nucleotides, and each of these nucleotides has to be present at least twice.

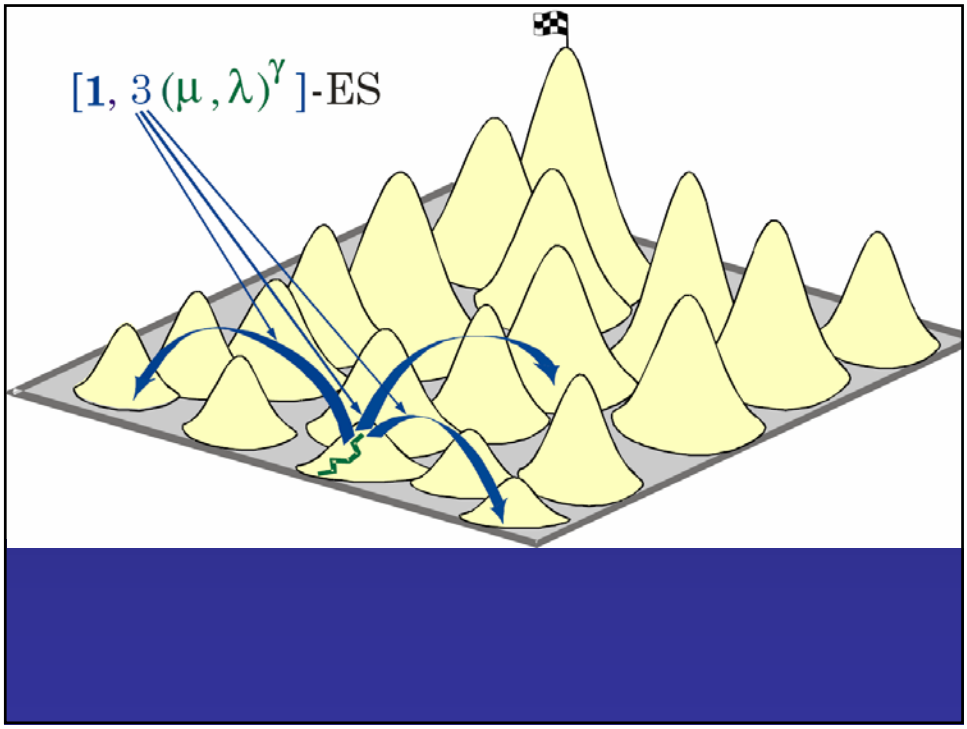
Maximum likelihood (ML)

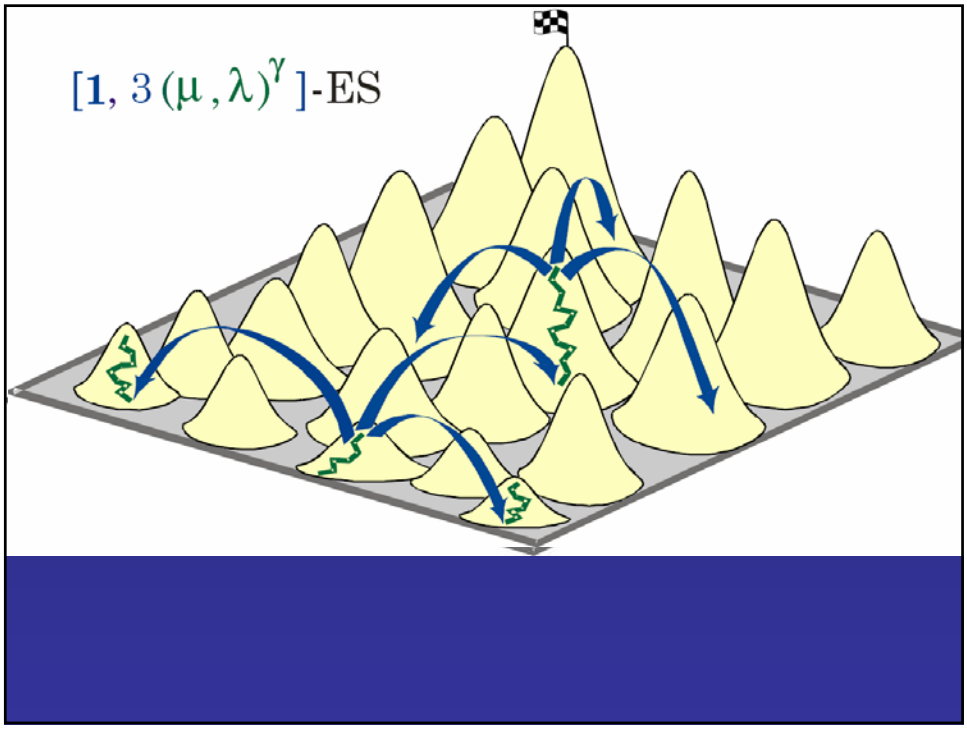
ML attempts to answer the question: What is the probability that I would observe the data (sequence alignment), given a particular model of evolution (a tree and a process).

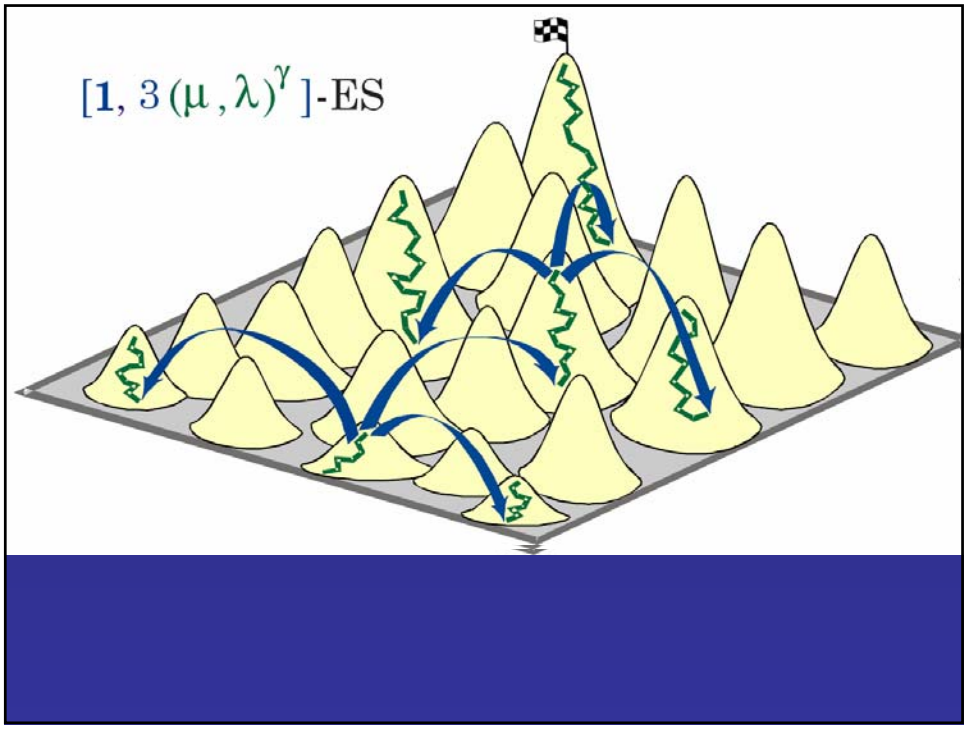
Bayesian method

- Recent variant of Maximum Likelihood
- Instead of seeking **the tree** that maximizes the likelihood of observing the data, it seeks **those trees** with greatest likelihood's given the data.
- Instead of producing a single tree, Bayesian analyses produces a set of trees of roughly equal likelihoods




[1, 3 (μ , λ)^γ]-ES


[1, 3 (μ , λ)^γ]-ES



Tree Confidence

- All phylogenetic trees represent hypothesis regarding the evolutionary history of the sequences that make up a date set.
- Reasonable to ask: "*how well the hypothesis describes the data*"

How much confidence can be attached to the overall tree and its component parts (branches)?

Bootstrap and Bayesian analyses

Bootstrapping

Bootstrapping is a modern statistical technique that uses computer intensive random resampling of data to determine sampling error or confidence intervals for some estimated parameter



Bootstrapping

- The most commonly used method to test the reliability of an inferred tree.
- A method in which one takes a subsample of the sites in an alignment and creates trees based on these subsamples.
  - This process is repeated multiple times (usually > 1000) and the results are compiled to a consensus tree
  - Numbers that correspond to the fraction of bootstrapped trees yielding the same grouping are often placed next to the corresponding nodes in phylogenetic trees.

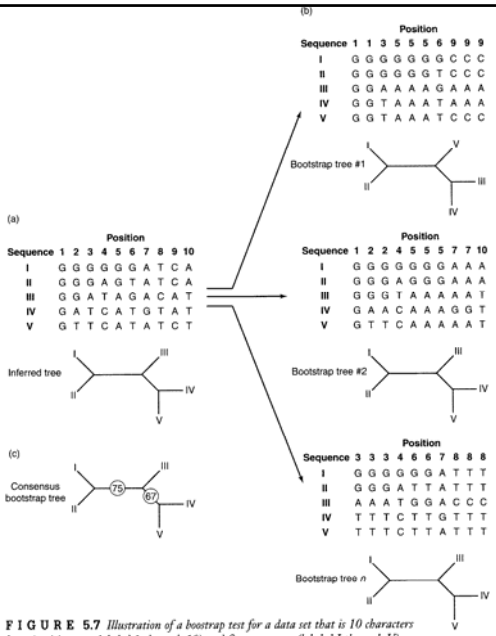


FIGURE 5.7 *Illustration of a boostrap test for a data set that is 10 characters long (positions are labeled 1 through 10) and five sequences (labeled I through V) deep. (a) The original data set and its most parsimonious tree is also shown. A random number generator chooses 10 times from the original 10 positions and creates the three resampled data sets with their corresponding trees (b). A consensus tree (c) for the three resampled data sets is shown with circled values indicating the fraction of bootstrapped trees that support the clustering of sequences I and II and sequences IV and V.*

Tree-building programs

PAUP* (http://paup.csit.fsu.edu/index.html).

Maximum likelihood (ML), Parsimony (MP), Distance methods; nucleotides

Puzzle (http://www.tree-puzzle.de/). **ML analysis for nucleotide, amino acid, and two-state data.**

Phylip (http://evolution.genetics.washington.edu/phylip.html).

ML, MP, NJ, UPGMA and more; nucleotides and proteins.

MEGA (http://www.megasoftware.net/). **MP, ME, NJ, UPGMA; nucleotides and proteins**