

WG1-WG4: School on bioinformatical analyses of phytoplasma sequences

Phylogenetic tree construction

Bojan Duduk



Institute for Pesticides and Environmental Protection, Belgrade

Homology :

the starting point of molecular phylogeny



Phylogenetic Tree: A branching diagram or “tree” showing the evolutionary relationships among various species, based upon similarities and differences in their genetic characteristics

- **Sequence comparison**

- **Bioinformatics tools** like ClustalW, JalView, and BLAST

- **Reference Sequence:** A sequence that has been chosen for the purpose of comparison. In genetic testing, a reference sequence is a known and well studied DNA or protein sequence. The reference sequences are chosen because they are of high quality and are thought to represent the sequence from the original organism

- **Query Sequence:** When performing genetic research, your “query sequence” is the sequence you are analyzing or trying to match

- **Mutation:** A change in a DNA or protein sequence

Sequence comparison

- the number of changes between different sequences is used to understand the evolutionary relatedness of the organisms
 - ▣ When sequences from two species are very similar, they are thought to be closely related
 - ▣ when sequences from two species are more dissimilar, the species are thought to be more distantly related
- DNA sequences that are more similar to one another are believed to share a more recent common ancestor than DNA sequences that are more different from one another

Pairs of Sequences are Compared to Each Other

A: ATGGTGCCG
B: ATGCTGCCG

B : ATGCTGCCG
C : ATGGACACG

B : ATGGTGCCG
D: ATGGTGAAG

A : ATGGTGCCG
D: ATGCAGCCG

D : ATGCAGCCG
C: ATGGACACG

A: ATGGTGCCG
C: ATGGACACG

Number of Nucleotide Differences:

	A	B	C	D
A	0	1	2	3
B	1	0	2	4
C	2	2	0	3
D	3	4	3	0

Pairwise Comparison: The process of comparing two DNA or protein sequences to one another to look for similarities and differences between the two sequences

Comparing DNA Sequences

Example: Genetic Testing using BLAST

Reference Sequence		<u>A T A G C T G</u>
Query Sequence(s):	1	<u> A </u>
	2	<u> A C </u>
	3	<u> </u>

Look for **mutations** or changes relative to **Reference Sequence**

Example: Multiple Sequence Alignments Using ClustalW

Sequence 1	A T G G T G C
Sequence 2	A T G C T G C
Sequence 3	A T G G A C A
Sequence 4	A T G C A G C

Look for **changes** relative to **each other**

The amount of changes among the sequences reflects the evolutionary relatedness of the organisms

Multiple Sequence Alignment



- The process of comparing more than two DNA or protein sequences to one another by aligning the sequences and looking for similarities and differences
 - Predicting protein structure, function
 - Primer design
- The information obtained from multiple sequence alignments can be used to construct phylogenetic trees

Multiple sequence alignment (MSA)

- For the construction of reliable phylogenetic trees the quality of a multiple alignment is of the utmost importance
- There are many programs available for the multiple alignment
 - ▣ A good program in the public domain is: ClustalW
 - ▣ A similar program is Pileup of the GCG package
- They quickly align sequence pairs and roughly determine the degrees of identity between each pair
- Then the sequences are aligned more precisely in a progressive way starting with the two closest sequences

Most programs work better when the sequences have similar length

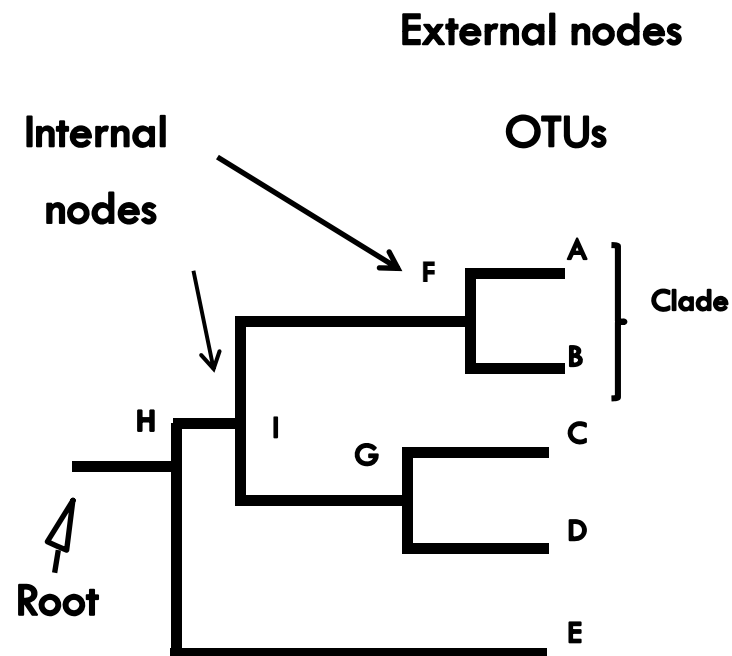
Phylogenetic tree and MSA



- Phylogenetic trees are a graphical representation of the evolutionary relatedness among the species in the tree
- Multiple sequence alignment (MSA) is closely related to constructing of a phylogenetic tree
- Every position in MSA is a character

Phylogenetic Trees Reflect Evolution

Phylogenetics: The study of evolutionary relationships among organisms



Distances are reflected in branch lengths

Remarks



In general, the output tree of a phylogenetic analysis is an estimate of the *character's* phylogeny (i.e. a gene tree) and not the phylogeny of the taxa (i.e. species tree) from which these characters were sampled, though ideally, both should be very close

They do not necessarily accurately represent the species evolutionary history

the analysis can be confounded by horizontal gene transfer, hybridization between species, convergent evolution, and conserved sequences

- Noncoding regions are more variable than coding regions
- Some positions in the protein coding genes are more variable than the others
- Some genes evolve faster than the other
- Same genes in the different organisms evolve faster than in other

Steps of making a phylogenetic tree



1. Find and download the sequences to be included in the tree
 - ▣ NCBI
2. Align the acquired sequences, check and trim the alignment
 - ▣ Clustal
 - ▣ MEGA 5
3. Construct the phylogenetic tree
 - ▣ MEGA 5

Program packages

There are more than 190 different packages related to phylogenetic analyses

- **GCG (Genetics Computer Group)** package:
PAUP (*Phylogenetic Analysis Using Parsimony*)
- PHYLIP (*PHYLogeny Inference Package*) open source

➤ **MEGA 5** MOLECULAR EVOLUTIONARY GENETICS ANALYSIS

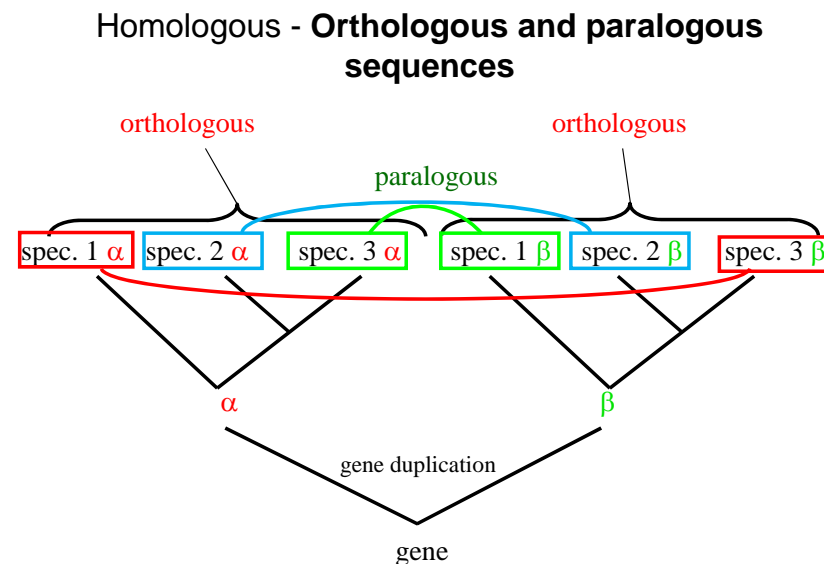


1. Find and download the sequences to be included in the tree

13

➤ Orthologous and paralogous genes

- ❑ Two genes are orthologous if they diverged after a speciation event
- ❑ Two genes are paralogous if they diverged after a duplication event
- ❑ It is likely that two orthologs have similar function, these functions are not necessarily "identical"
- ❑ Paralogous usually have different function



- Homologous sequences as result of horizontal transfer between 2 species, and not common ancestor
- Homologous sequences as result of convergence

1. Find and download the sequences to be included in the tree

14

- Sequence databases
 - NCBI: <http://www.ncbi.nlm.nih.gov/>
 - EMBL: <http://www.ebi.ac.uk/>
 - DDBJ: <http://www.ddbj.nig.ac.jp/>

- NCBI Home
- Site Map (A-Z)
- All Resources
- Chemicals & Bioassays
- Data & Software
- DNA & RNA
- Domains & Structures
- Genes & Expression
- Genetics & Medicine
- Genomes & Maps
- Homology
- Literature
- Proteins
- Sequence Analysis
- Taxonomy
- Training & Tutorials
- Variation

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

[About the NCBI](#) | [Mission](#) | [Organization](#) | [Research](#) | [RSS Feeds](#)

Get Started

- [Tools](#): Analyze data using NCBI software
- [Downloads](#): Get NCBI data or software
- [How-To's](#): Learn how to accomplish specific tasks at NCBI
- [Submissions](#): Submit data to GenBank or other NCBI databases

Human Microbiome Project

NIH Roadmap Initiative designed to characterize the community of microorganisms living on and in the human body.



1 2 3 4 5

Popular Resources

- [BLAST](#)
- [Bookshelf](#)
- [Gene](#)
- [Genome](#)
- [Nucleotide](#)
- [OMIM](#)
- [Protein](#)
- [PubChem](#)
- [PubMed](#)
- [PubMed Central](#)
- [SNP](#)

NCBI News

[Preliminary genomic assemblies from two isolates from the European E. coli outbreak now available](#)

07 Jun 2011

[Preliminary genomic assemblies of two isolates are in the](#)


























[New version of Cn3D \(v.4.3\) now available](#)

07 Jun 2011

[A new version of this popular 3D molecular visualization program](#)

Search across databases [Help](#)

- Result counts displayed in gray indicate one or more terms not found

<input type="text" value="none"/>  PubMed: biomedical literature citations and abstracts	<input type="text" value="none"/>  Books: online books
<input type="text" value="none"/>  PubMed Central: free, full text journal articles	<input type="text" value="none"/>  OMIM: online Mendelian Inheritance in Man
<input type="text" value="none"/>  Site Search: NCBI web and FTP sites	
<input type="text" value="34"/>  Nucleotide: Core subset of nucleotide sequence records	<input type="text" value="none"/>  dbGaP: genotype and phenotype
<input type="text" value="none"/>  EST: Expressed Sequence Tag records	<input type="text" value="none"/>  UniGene: gene-oriented clusters of transcript sequences
<input type="text" value="none"/>  GSS: Genome Survey Sequence records	<input type="text" value="none"/>  CDD: conserved protein domain database
<input type="text" value="none"/>  Protein: sequence database	<input type="text" value="none"/>  UniSTS: markers and mapping data
<input type="text" value="none"/>  Genome: whole genome sequences	<input type="text" value="1"/>  PopSet: population study data sets
<input type="text" value="none"/>  Structure: three-dimensional macromolecular structures	<input type="text" value="none"/>  GEO Profiles: expression and molecular abundance profiles
<input type="text" value="none"/>  Taxonomy: organisms in GenBank	<input type="text" value="none"/>  GEO DataSets: experimental sets of GEO data
<input type="text" value="none"/>  SNP: single nucleotide polymorphism	<input type="text" value="none"/>  Epigenomics: Epigenetic maps and data sets
<input type="text" value="none"/>  dbVar: Genomic structural variation	<input type="text" value="none"/>  Cancer Chromosomes: cytogenetic databases
<input type="text" value="none"/>  Gene: gene-centered information	<input type="text" value="none"/>  PubChem BioAssay: bioactivity screens of chemical substances

Firefox 34 selected items - Nucleotide result

http://www.ncbi.nlm.nih.gov/nucleotide/M30790,U15442,AY197655,AB052876,AY197648,AY390261,AF092209,AF515637,AJ542541,AJ542544,AJ542543,X9

NCBI Resources How To My NCBI Sign In

Nucleotide

Alphabet of Life

Search: [Limits](#) [Advanced search](#) [Help](#)

[Display Settings:](#) Summary, 20 per page, Sorted by Default order [Send to:](#)

Results: 1 to 20 of 34 << First < Prev Page **1** of 2 Next > Last >>

[Oenothera phytoplasma 86-7 16S ribosomal RNA, complete sequence](#)

1. 1,535 bp linear rRNA
Accession: M30790.1 GI: 175280
[GenBank](#) [FASTA](#) [Graphics](#) [Related Sequences](#)

[Candidatus Phytoplasma aurantifolia 16S ribosomal RNA, complete sequence; tRNA-Ile gene, complete sequence; and 23S ribosomal RNA, partial sequence](#)

2. 1,788 bp linear rRNA
Accession: U15442.1 GI: 567051
[GenBank](#) [FASTA](#) [Graphics](#) [Related Sequences](#)

[Elm yellows phytoplasma strain EY1 16S ribosomal RNA gene, partial sequence](#)

3. 1,527 bp linear DNA
Accession: AY197655.1 GI: 31074421
[GenBank](#) [FASTA](#) [Graphics](#) [PopSet](#) [Related Sequences](#)

[Phytoplasma sp. JWB-G1 gene for 16S rRNA, partial sequence](#)

4. 1,367 bp linear DNA
Accession: AB052876.1 GI: 27530586
[GenBank](#) [FASTA](#) [Graphics](#) [Related Sequences](#)

[Rubus stunt phytoplasma strain RUS 16S ribosomal RNA gene, partial sequence](#)

5. 1,529 bp linear DNA
Accession: AY197648.1 GI: 31074414
[GenBank](#) [FASTA](#) [Graphics](#) [PopSet](#) [Related Sequences](#)

[Candidatus Phytoplasma trifolii 16S ribosomal DNA gene, partial sequence; 16S-23S ribosomal DNA intergenic spacer and tRNA-Ile gene, complete](#)

Filter your results:

All (34)

[Bacteria \(33\)](#)

[INSDC \(GenBank\) \(33\)](#)

mRNA (0)

RefSeq (0)

[Manage Filters](#)

Top Organisms [Tree]

- Phytoplasma sp. (2)
- Mycoplasma sp. (1)
- Acholeplasma laidlawii (1)
- Oenothera phytoplasma 86-7 (1)
- Candidatus Phytoplasma pini (1)
- All other taxa (27)
- More...

Find related data

Database:

Recent activity

[Turn Off](#) [Clear](#)

Windows taskbar: Start, litvanjaskola, Phylogenetic tree constru..., 34 selected items - N..., SR, 11:59

Display Settings: Summary, 20 per page, Sorted by Default order

Results: 1 to 20 of 34

[Oenothera phytoplasma 86-7 16S ribosomal RNA, complete sequence](#)

1. 1,535 bp linear rRNA
Accession: M30790.1 GI: 175280
[GenBank](#) [FASTA](#) [Graphics](#) [Related Sequences](#)

[Candidatus Phytoplasma aurantifolia 16S ribosomal RNA, complete sequence; tRNA-Ile gene, complete sequence; and 2 sequence](#)

2. 1,788 bp linear rRNA
Accession: U15442.1 GI: 567051
[GenBank](#) [FASTA](#) [Graphics](#) [Related Sequences](#)

[Elm yellows phytoplasma strain EY1 16S ribosomal RNA gene, partial sequence](#)

3. 1,527 bp linear DNA
Accession: AY197655.1 GI: 31074421
[GenBank](#) [FASTA](#) [Graphics](#) [PopSet](#) [Related Sequences](#)

[Phytoplasma sp. JWB-G1 gene for 16S rRNA, partial sequence](#)

4. 1,367 bp linear DNA
Accession: AB052876.1 GI: 27530586
[GenBank](#) [FASTA](#) [Graphics](#) [Related Sequences](#)

[Rubus stunt phytoplasma strain RUS 16S ribosomal RNA gene, partial sequence](#)

5. 1,529 bp linear DNA
Accession: AY197648.1 GI: 31074414
[GenBank](#) [FASTA](#) [Graphics](#) [PopSet](#) [Related Sequences](#)

[Candidatus Phytoplasma trifolii 16S ribosomal DNA gene, partial sequence; 16S-23S ribosomal DNA intergenic spacer and tRNA-Ile gene, complete](#)

Send to: Filter your results:

Choose Destination
 File Clipboard
 Collections
Download 34 items.
Format
FASTA
Create File

Top Organisms [Tree]
Phytoplasma sp. (2)
Mycoplasma sp. (1)
Acholeplasma laidlawii (1)
Oenothera phytoplasma 86-7 (1)
Candidatus Phytoplasma pini (1)
All other taxa (27)
More...

Find related data
Database: Select
Find items

Recent activity
Turn Off Clear

Multiple sequence alignment



2. Align the acquired sequences, check and trim the alignment

Programs that performs multiple sequence alignments.

- ▣ Muscle
- ▣ ClustalW: performs very well in practice.
 - MEGA 5

Multiple sequence alignment



A multiple sequence alignment (MSA) is obtained by inserting gaps ('-') into the original sequences such that all resulting sequences have equal length and no column consists of gaps only

The most commonly used approach to MSA is probably progressive alignment (ClustalW)

One of the first progressive alignment algorithms was published 1987 by Feng and Doolittle

CLUSTAL is one of the most popular programs for computing an MSA. It is based on the Feng-Doolittle method

Feng, D-F & Doolittle, RF. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J. Mol. Evol.* 25:351-360, 1987

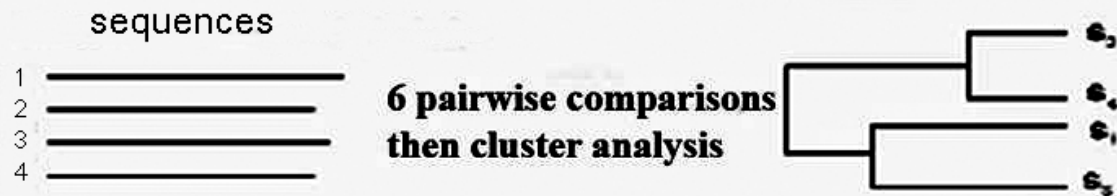
Multiple sequence alignment



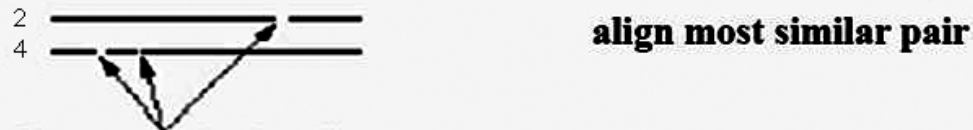
- The algorithm starts by computing a rough distance matrix between each pair of sequences based on pairwise sequence alignment scores
- Next, the algorithm uses the neighbor-joining method with midpoint rooting to create a guide tree, which is used to generate a global alignment. The guide tree serves as a rough template for clades that tend to share insertion and deletion features
- The most similar sequences, that is, those with the best alignment score are aligned first. Then progressively more distant groups of sequences are aligned until a global alignment is obtained
 1. Fast
 2. This generally provides a close-to-optimal result, especially when the data set contains sequences with varied degrees of divergence

Progressive alignment

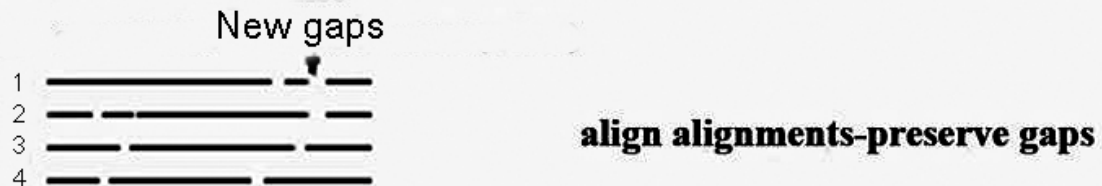
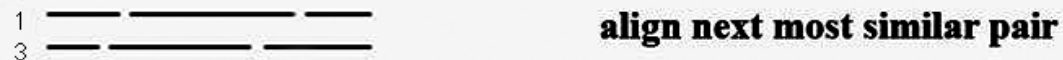
Pairwise Alignment



Multiple alignment according to the tree



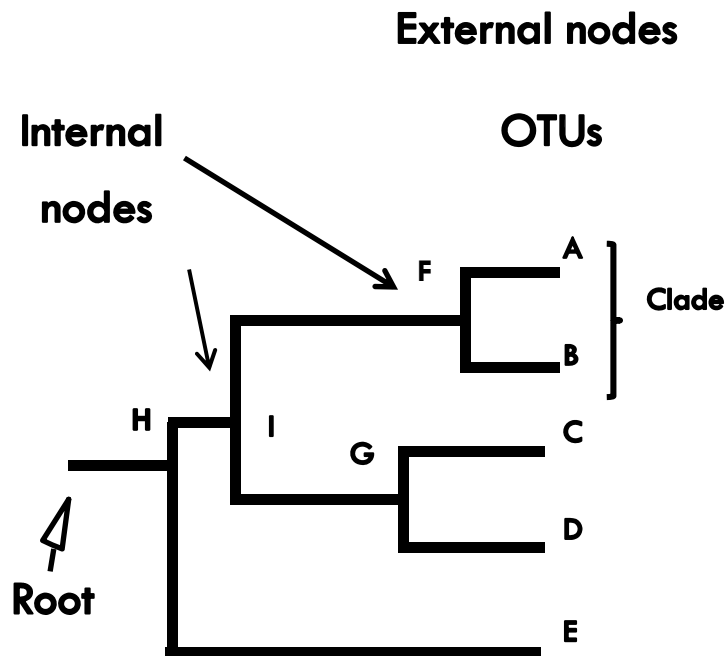
Gaps to optimize alignment



Align the sequences



Some useful information about phylogenetic trees



A-E are external nodes OTUs are operational taxonomic units
They can be: species

F-J are internal (ancestral) nodes
They are existing or extinct

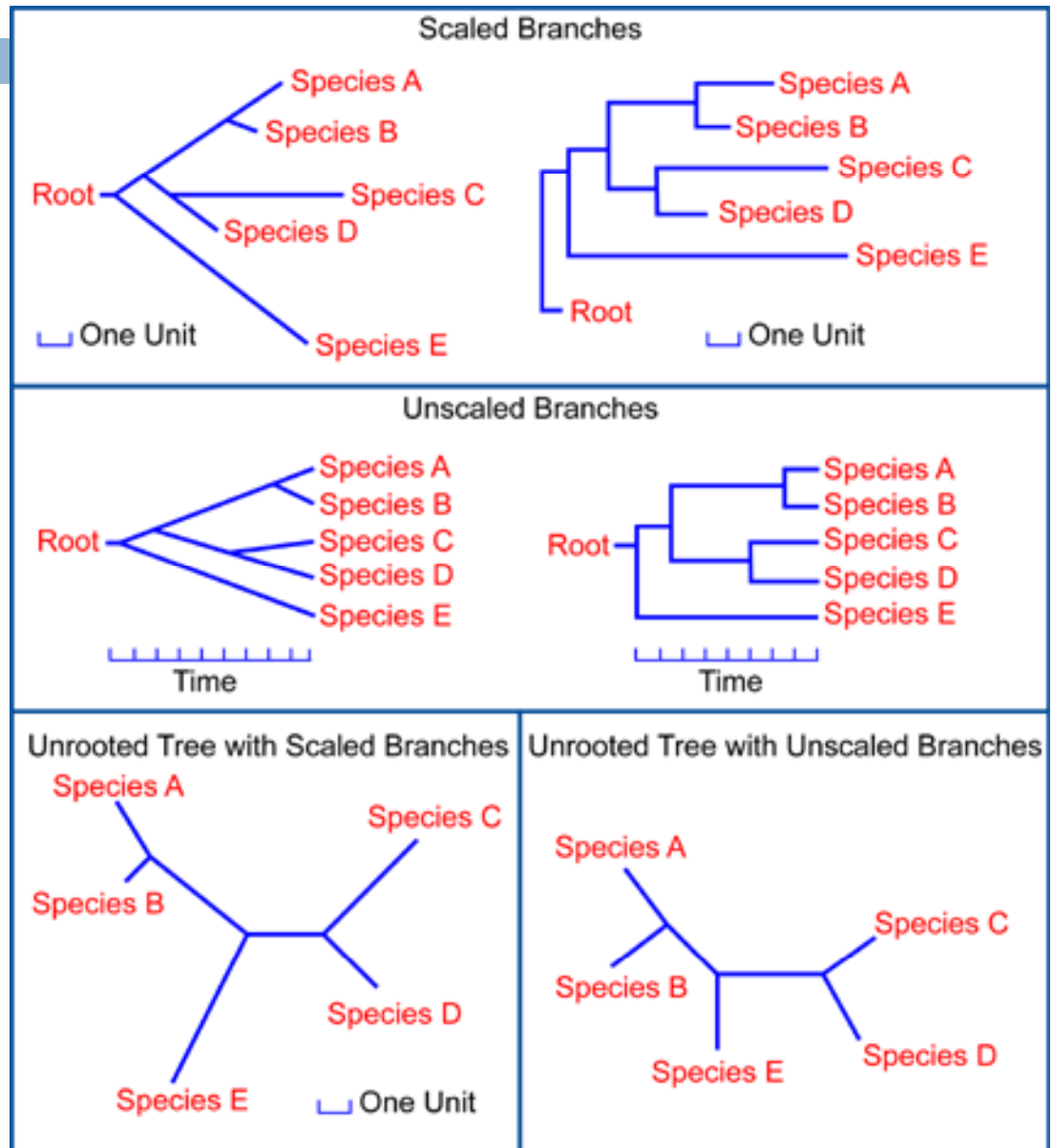
Root: a common ancestor for all sequences

Clade: a group consisting of an organism and all its descendants. In the terms of biological systematics, a clade is a single branch on the tree

Branch length: represents number of changes

Topology: order of the nodes on the tree

A phylogenetic tree can be



rooted path from root to a node represents an evolutionary path - the root represents the common ancestor

unrooted specifies relationships among things, but not evolutionary paths

How to root an unrooted tree?



- To root a tree one should add an outgroup to the dataset. An outgroup is an operational taxonomic unit (OUT) that branched off before all other taxa
- Do not choose an outgroup that is very distantly related to your taxa. This may result in serious topological errors
- Do not choose either an outgroup that is too closely related to the taxa in question. In this case it may not be a true outgroup
- The use of more than one outgroup generally improves the estimate of tree topology
- In the absence of a good outgroup the root may be positioned by assuming approximately equal evolutionary rates over all the branches. In this way the root is put at the midpoint of the longest pathway between two OTUs

Bootstrapping

statistical method for obtaining an estimate of errors

- Bootstrapping is a way of testing the reliability of a phylogenetic tree
- The pseudo-replicate datasets are generated by randomly sampling the original character matrix to create new matrices of the same size as the original
- The frequency with which a given branch is found is recorded as the bootstrap proportion, and it can be used as a measure of the reliability
- is used to examine how often a particular cluster in a tree appears when nucleotides or aminoacids are re-sampled

Phylogenetic tree building methods



- **Molecular phylogenetic tree building methods**
- Are mathematical and/or statistical methods for inferring the divergence order of taxa, as well as the lengths of the branches that connect them
- There are many phylogenetic methods available today, each having strengths and weaknesses
- None of the methods is reliable when OTUs with highly unequal evolutionary separation are included in the data set
- Most can be classified as follows:

Phylogenetic tree building methods

Distance-based methods— Methods for making phylogenetic trees with DNA or protein sequences that involves calculating the percent difference between each pair of sequences, and using these percent differences to construct the phylogenetic tree

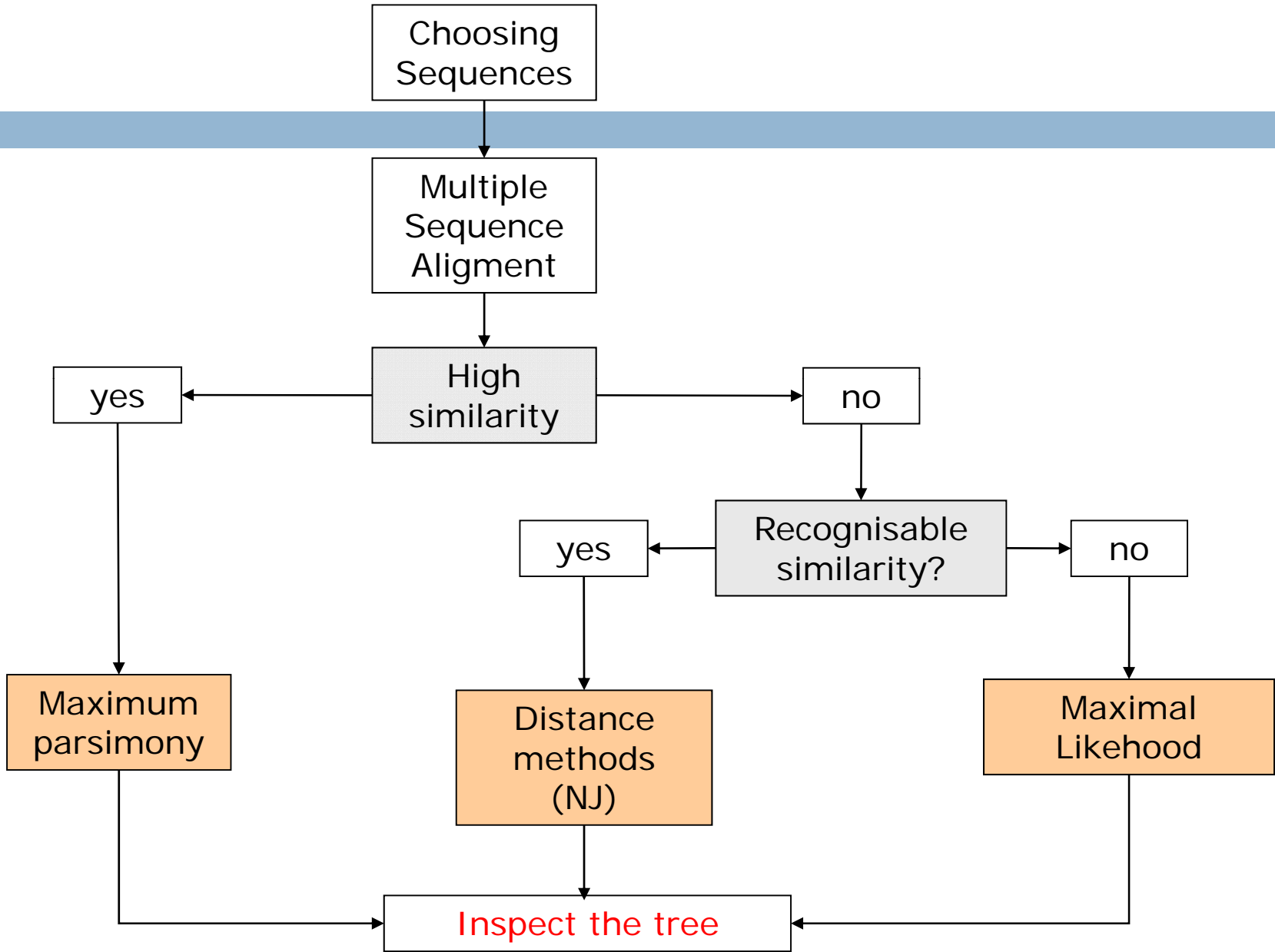
- ▣ Neighbor Joining

Character-based methods - are said to be more powerful than distance methods because they use the raw data

- ▣ Parsimony — searches in all possible phylogenetic trees that needs the minimum number of substitutions of nucleic acids or amino acids (mutations), so the best tree is the one that have the minimum number of mutations
- ▣ Maximum likelihood — the best estimate of a parameter is that giving the highest probability that the observed set of measurements will be obtained

Phylogenetic tree building methods

- Distance methods or distance based trees are easy to set up, and you can apply them in most situations, but they aren't necessarily the most accurate. The main disadvantage of distance-matrix methods is their inability to efficiently use information about local high-variation regions
- Distance approaches (UPGMA, **Neighbor Joining** etc.) do not use the original (sequence) data, but calculate the percent difference between each pair of sequences, and are using these percent differences to construct the phylogenetic tree. Some information is said to be lost
- Character-state approaches (maximum parsimony, maximum likelihood) are said to be more powerful than distance methods because they use the raw data
- **Maximum parsimony** uses only the relevant sites. So when the number of informative sites is not large, this method is often less efficient than distance methods (Saitou and Nei, 1986). Maximum parsimony is notorious for its sensitivity to codon bias and unequal rates of evolution
- **Likelihood methods** are the most accurate and the best, because it uses all data, but the problem is that they run very slow because of their long algorithms



Distance-based methods

Transform the sequence data into pairwise distances (dissimilarities), and then use the matrix during tree building

	A	B	C	D	E
Species A	----	0.20	0.50	0.45	0.40
Species B	0.23	----	0.40	0.55	0.50
Species C	0.87	0.59	----	0.15	0.40
Species D	0.73	1.12	0.17	----	0.25
Species E	0.59	0.89	0.61	0.31	----

← Example 1:
Uncorrected
“p” distance
(=observed percent
sequence difference)



Example 2: Kimura 2-parameter distance
(estimate of the true number of substitutions between taxa)

Maximum Composite Likelihood-increases the accuracy of calculating the pairwise distances

Distance-based methods



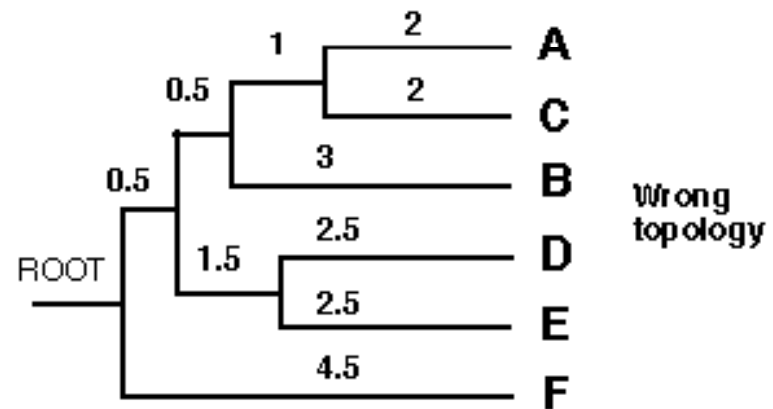
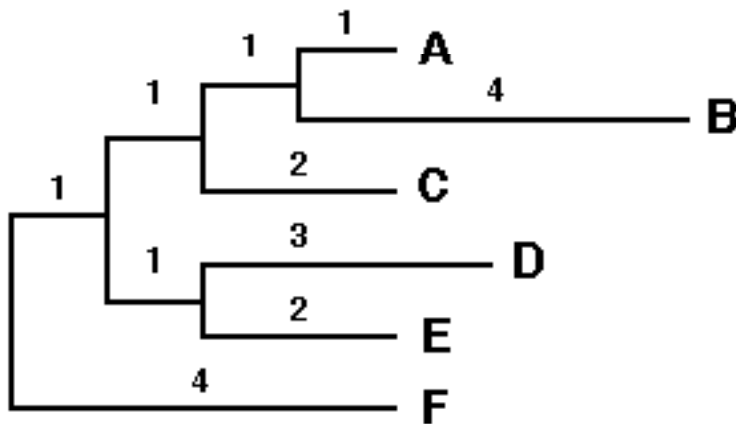
UPGMA (**U**nweighted **P**air **G**roup **M**ethod with **A**rithmetic **M**ean) assumes a constant rate of evolution, and is not a well-regarded method for inferring relationships unless this assumption has been tested and justified for the data set being used. \Rightarrow construct a rooted tree

NJ (***N**eighbor **J**oining*) - unlike UPGMA does not assume a constant rate of evolution across lineages \Rightarrow different branch lengths, unrooted tree

Unequal rates of mutation lead to wrong trees

UPGMA

- The UPGMA clustering method is very sensitive to unequal evolutionary rates
- UPGMA tree construction based on the data of the left tree would result in the erroneous tree at the right



Neighbor Joining (NJ) (Saitou and Nei, 1987)



- The principle of this method is to find pairs of operational taxonomic units (OTUs) that minimize the total branch length at each stage of clustering of OTUs starting with a starlike tree
- The branch lengths as well as the topology of a parsimonious tree can quickly be obtained by using this method

Neighbor Joining (NJ)

The algorithm Step 1

- The raw data of the tree are represented by the following distance matrix

	A	B	C	D	E
B	5				
C	4	7			
D	7	10	7		
E	6	9	6	5	
F	8	11	8	9	8

- We have in total 6 OTUs ($N=6$)

Neighbor Joining (NJ)

The algorithm Step 2



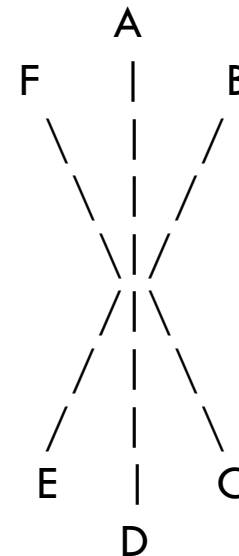
- We calculate the net divergence $r(i)$ for each OTU from all other OTUs
- $r(A) = 5+4+7+6+8=30$
- $r(B) = 42$
- $r(C) = 32$
- $r(D) = 38$
- $r(E) = 34$
- $r(F) = 44$

Neighbor Joining (NJ)

The algorithm Step 3

- Now we calculate a new distance matrix using for each pair of OUTs the formula
- $M(ij) = d(ij) - [r(i) + r(j)] / (N - 2)$
- $M(AB) = d(AB) - [(r(A) + r(B))] / (N - 2) = -13$

	A	B	C	D	E
B	-13				
C	-11.5	-11.5			
D	-10	-10	-10.5		
E	-10	-10	-10.5	-13	
F	-10.5	-10.5	-11	-11.5	-11.5

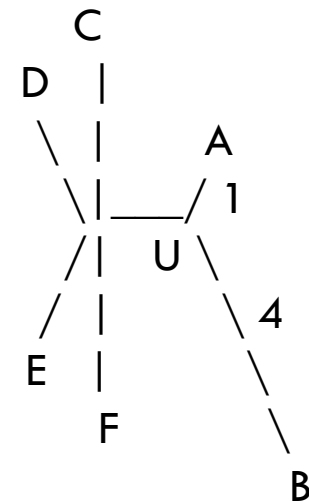


Neighbor Joining (NJ)

The algorithm Step 4

- Now we choose as neighbors those two OTUs for which M_{ij} is the smallest. These are A and B and D and E. Let's take A and B as neighbors and we form a new node called U. Now we calculate the branch length from the internal node U to the external OTUs A and B.
- $S(AU) = d(AB) / 2 + [r(A) - r(B)] / 2(N - 2) = 1$
- $S(BU) = d(AB) - S(AU) = 4$

- The resulting tree will be the following



Neighbor Joining (NJ)

The algorithm Step 5

- Now we define new distances from U to each other terminal node:

- $d(CU) = d(AC) + d(BC) - d(AB) / 2 = 3$

- $d(DU) = d(AD) + d(BD) - d(AB) / 2 = 6$

- $d(EU) = d(AE) + d(BE) - d(AB) / 2 = 5$

- $d(FU) = d(AF) + d(BF) - d(AB) / 2 = 7$

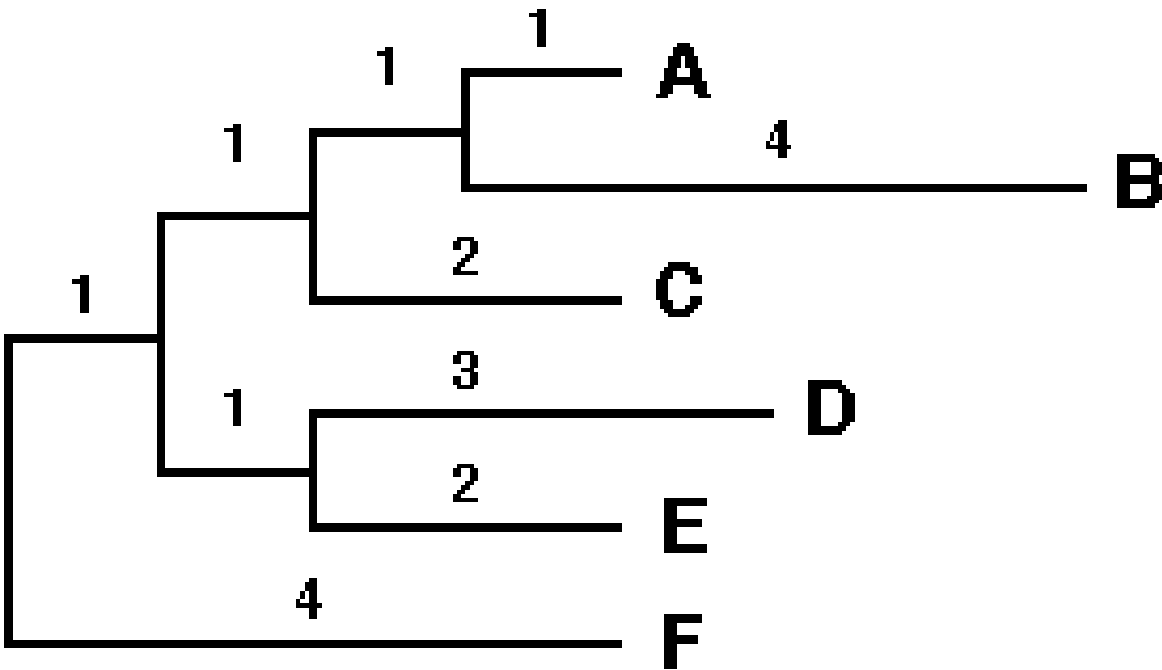
- and we create a new matrix

- $N = N - 1 = 5$

-

The entire procedure is repeated starting at step 1

	U(AB)	C	D	E
C	3			
D	6	7		
E	5	6	5	
F	7	8	9	8



Neighbor Joining (NJ)

- **Advantages and disadvantages of the neighbor-joining method**
- **Advantages**
 - is fast and thus suited for large datasets and for bootstrap analysis
 - permits lineages with largely different branch lengths
 - permits correction for multiple substitutions (from Jukes-Cantor model)
 - gives only one possible tree
- **Disadvantages**
 - sequence information is reduced
 - strongly dependent on the model of evolution used
 - gives only one possible tree

NJ tree



Character-based methods



- **Maximum parsimony (MP)** is a method of identifying the potential phylogenetic tree that requires the smallest total number of evolutionary events to explain the observed sequence data
- **Maximum likelihood method (ML)** - Inferring the most likely evolutionary tree for a group of sequences by considering the probability of all possible mutational paths between them

Maximum Parsimony analysis



- Parsimony implies that simpler hypotheses are preferable to more complicated ones
- Maximum parsimony is a **character-based method** that infers a phylogenetic tree by minimizing the total number of evolutionary steps required to explain a given set of data, or in other words by minimizing the total tree length

Maximum Parsimony Methods



- Use sequence information rather than distance information
- Calculate for all possible trees and find the tree that represents the minimum number of substitutions at each informative site

Maximum parsimony-minimum change



- The tree that requires the smallest number of changes to explain the data is the most likely tree (the most parsimonious tree)
- MP method does not use specific models to estimate the trees
- By changing the topology or OTUs the parsimony score is changed
- The MP method produces many equally parsimonious trees

Informative sites

1 2 3 4 5 6 7 8 9

Sequence

1	A	A	G	A	G	T	G	C	A
2	A	G	C	C	G	T	G	C	G
3	A	G	A	T	A	T	C	C	A
4	A	G	A	G	A	T	C	C	G
				*	*	*			

Maximum Parsimony analysis

- The number of **rooted** trees (N_r) for n OTUs is given by:

$$N_r = (2n - 3)! / (2^{n-2} (n - 2)!)$$

- The number of **unrooted** trees (N_u) for n OTUs is given by:

$$N_u = (2n - 5)! / (2^{n-3} (n - 3)!)$$

Number of OTUs	unrooted trees	rooted trees
2	1	1
3	1	3
4	3	15
5	15	105
6	105	945
7	954	10,395
8	10,395	135,135
9	135,135	34,459,425
10	34,459,425	2.13E15
15	2.13E15	8.E21

This rapid increase in number of trees to be analysed may make it impossible to apply the method to very large datasets. In that case the parsimony method may become very time consuming, even on very fast computers

Parsimony



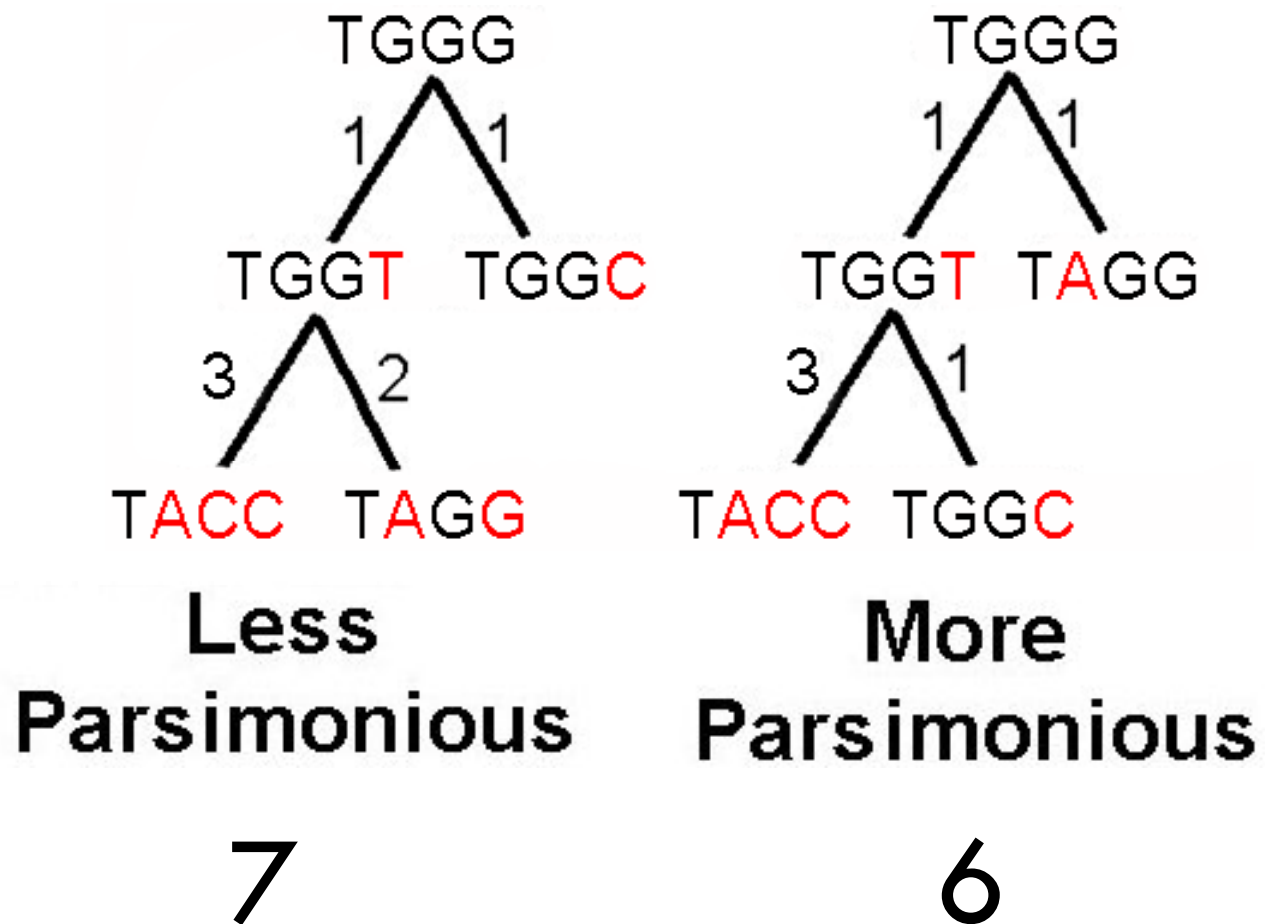
- Two problems
 - The Small Parsimony Problem
 - to compute the parsimony score for a given tree
 - The large parsimony problem
 - How to find the best tree ?

The Small Parsimony Problem



- The Fitch algorithm
- In 1971, Walter Fitch published a dynamic programming algorithm that solves the small parsimony problem efficiently

Parsimony



Large parsimony problem

- Number of trees to be searched is HUDGE:
 - ▣ $(2n - 3)!!$ Number of possible rooted trees
 - ▣ $(2n - 5)!!$ Number of possible unrooted trees
- Exhaustive enumeration of all possible tree topologies will only work for small number of sequences ($n \leq 10$)

# seq.	# unrooted trees	# rooted trees
10	2,027,025	34,459,425

- ▣ Thus, we need more efficient strategies that either solve the problem exactly, such as the “branch and bound” technique, or return good approximations, such as “heuristic searches”

How to find the best tree ?



- **Maximum parsimony** searches for the optimal (minimal) tree. In this process more than one minimal trees may be found. In order to guarantee to find the best possible tree an exhaustive evaluation of all possible tree topologies has to be carried out. However, this becomes impossible when there are more than 12 OTUs in a dataset
- **Branch and Bound**: is a variation on maximum parsimony that guarantees to find the minimal tree without having to evaluate all possible trees. This way a larger number of taxa can be evaluated but the method is still limited
- **Heuristic searches** is a method with step-wise addition and rearrangement (branch swapping) of OTUs. Here it is not guaranteed to find the best tree



- Tree Searching Methods

- Exhaustive search (exact)

- Branch and bound search (exact)

- Heuristic search methods (approximate)

- Stepwise addition

- Star decomposition

- Branch swapping

- Close-Neighbor-Interchange (CNI)

Branch and bound search

- Application of branch-and-bound to evolutionary trees was first suggested by Mike Hendy and Dave and Penny (1982)
- While this algorithm is guaranteed to find all the MP trees, a branch-and-bound search often is too time consuming for more than 15 sequences
- In practice, using branch and bound one can obtain exact solutions for data sets of twenty or more sequences, depending on the sequence length and the “messiness” of the data

Hendy, M. D. and Penny, D. (1982). Branch and bound algorithms to determine minimal evolutionary trees. *Mathematical Biosciences* 59: 277-290

Close-Neighbor-Interchange (CNI)

- This algorithm reduces the time spent searching by first producing a temporary tree, and then examining all of the topologies that are different from this temporary tree by a topological distance of $dT = 2$ and 4. If this is repeated many times, and all the topologies previously examined are avoided, it can usually obtain the tree being sought
- For the MP method, the CNI search can start with a tree generated by the random addition of sequences. This process can be repeated multiple times to find the MP tree

Nei M & Kumar S (2000) Molecular Evolution and Phylogenetics. Oxford University Press, New York.

Maximum parsimony can be inconsistent



- Under certain conditions ***long branch attraction*** can occur
 - where there are long branches (a high level of substitutions) for two characters, but short branches for another two. And all diverged from a common ancestor

Some final notes on maximum parsimony



- MP positive points:
 - is based on shared and derived characters
 - does not reduce sequence information to a single number
 - evaluates different trees

- MP negative points:
 - is slow in comparison with distance methods
 - does not use all the sequence information (only informative sites are used)
 - does not correct for multiple mutations (does not imply a model of evolution)
 - does not provide information on the branch lengths
 - the most parsimonious tree is not always the correct one;
 - similarity between sequences on long branches may be explained by independent substitutions to the same nucleotide and not by their closer relationship

MP tree



Maximum likelihood



- The method of maximum likelihood is a contribution of RA Fisher, who first investigated its properties in 1922
- **Principle:** evaluate all possible trees (topology and branch lengths) and substitution model parameters (TS/TV, base freq, rate heterogeneity etc.). Choose the one that maximizes the likelihood of your data (the alignment)

Maximum likelihood



- Pick an Evolutionary Model
- For each position, Generate all possible tree structures
- Based on the Evolutionary Model, calculate Likelihood of these Trees and Sum them to get the Column Likelihood for each OTU cluster
- Calculate Tree Likelihood by multiplying the likelihood for each position
- Choose Tree with Greatest Likelihood

Maximum likelihood



- Similar to maximum parsimony, an optimal MLE tree is determined by a search in tree space
- The method searches for the tree with the highest probability or likelihood
- The likelihood of observing a given set of data is maximized for each topology, and the topology that gives the highest maximum likelihood is chosen as the final tree
- The parameters to be considered are not the topologies but the branch lengths for each topology, and the likelihood is maximized to estimate branch lengths rather than the topology

Likelihood for the full tree

The likelihood for the full tree is the product of the likelihood at each site


Since the individual likelihoods are extremely small numbers it is convenient to sum the log likelihoods at each site and report the likelihood of the entire tree as the log likelihood

$$\ln L = \ln L(1) + \ln L(2) \dots + \ln L(N) = \sum_{j=1}^N \ln L(j)$$

The maximum likelihood tree



- This procedure is repeated for all possible topologies, and the topology that shows the highest likelihood is chosen as the final tree
- According to this method, the nucleotides or amino acids of all sequences at each site are considered separately (as independent), and the log-likelihood of having these bases are computed for a given topology by using a particular probability model
- The method requires that evolution at different sites and along different lineages must be statistically independent
- Maximum likelihood is thus well suited to the analysis of distantly related sequences, but because it formally requires search of all possible combinations of tree topology and branch length, it is computationally expensive

- 
- Parsimony picks the most probable path, likelihood method sums over all paths
 - Parsimony ignores evolution time t

Advantages and disadvantages of the maximum likelihood method



- There are some supposed advantages of maximum likelihood methods over other methods.
 - It is the estimation method least affected by sampling error
 - with very short sequences it tends to outperform alternative methods such as parsimony or distance methods.
 - evaluates different tree topologies
 - uses all the sequence information

- There are also some supposed disadvantages
 - maximum likelihood is very CPU intensive and thus extremely slow
 - result is dependent on the model of evolution used

Bayesian Inference of Tree

- Mr Bayes (<http://mrbayes.csit.fsu.edu/>)
- Character based
- **Posterior probability** The posterior probability distribution of trees is impossible to calculate analytically; instead, MrBayes uses a simulation technique called Markov chain Monte Carlo (or MCMC) to approximate the posterior probabilities of trees
- Begins with a tree (randomly chosen)
- Evaluate the tree
- Change the tree and evaluate it (better-accept)
- Calculate the consensus of the recorded trees (with posterior probabilities)

Conclusions



- Neighbor-joining is good when evolutionary rates vary. Proven to construct the correct tree
- Parsimony is good for closely related sequences
- Likelihood method is the most general of all
- Using several phylogenetic methods is instructive
- If more characters are used to construct the phylogenetic tree it is better



Thank you for your attention