

## **Molecular phylogenetic analysis using ribosomal RNA (rRNA)**

*mostly taken from Prof. James W. Brown, NC State University*

Molecular phylogenetic analysis is the use of macromolecular sequences to reconstruct the evolutionary relationships between organisms. The extent of difference between homologous DNA, RNA, or protein sequences in different organisms is used as a measure of how much these organisms have diverged from one another evolutionarily.

### **Steps of a molecular phylogenetic analysis:**

1. Decide what organisms/sequence/gene/region to examine
2. Determine the sequence(s) experimentally
3. Identify homologous residues
4. Phylogenetic analysis: compare sequences to determine relationships.

### **Deciding on a sequence for analysis --**

Some sequences are better than others - the most important factors to consider in a sequence for phylogenetic analysis are:

1. Clock-like behavior, i.e. sequence divergence in the gene between two organisms should be proportional to how long ago they diverged. Clock-like behavior depends mostly on functional constancy of the sequence - function change leads to large, directed sequence change. The sequence should be long enough to provide information.
2. Phylogenetic range. The sequence must be present in all of the organisms to be analysed and must have retained its structure & function in these organisms. The gene must have enough variation in sequence to give information but must be similar enough so that homologous residues can be recognized and lined up (alignment). Non-functional sequences (e.g. introns) usually change too fast for analysis except of the very closest of relatives.
3. No horizontal transfer. This means that the gene must be acquired only by inheritance, not by transfer from another organism. An example of frequently horizontally transferred genes are those encoding antibiotic resistance. You can still generate a tree with these sequences, but they do not reflect the genealogy of the organisms as a whole, only of those genes.
4. Must have a large existing data set with which to compare your sequences. In other words, other sequences of that gene should be in GenBank!

## 16S rRNA ("the molecule" of modern genomics)

The gene encoding the RNA in the small subunit of the ribosome (ssu rRNA) is often the best choice because:

- \* It is present in all cells
- \* It has exactly the same function in all cells
- \* It is conserved enough in sequence & structure of be readily & accurately aligned.
- \* It contains both rapidly & slowly evolving regions - the fast regions are useful for determining closely related species, whereas the slow regions are useful for determining distant relationships
- \* Horizontal transfer of rRNA genes is absent or rare.
- \* There is a large database (>100,000) of aligned sequences available

### 16S rRNA Facts

The 16S rRNA gene (or rDNA gene) codes for one part of the ribosomal RNA in a cell.

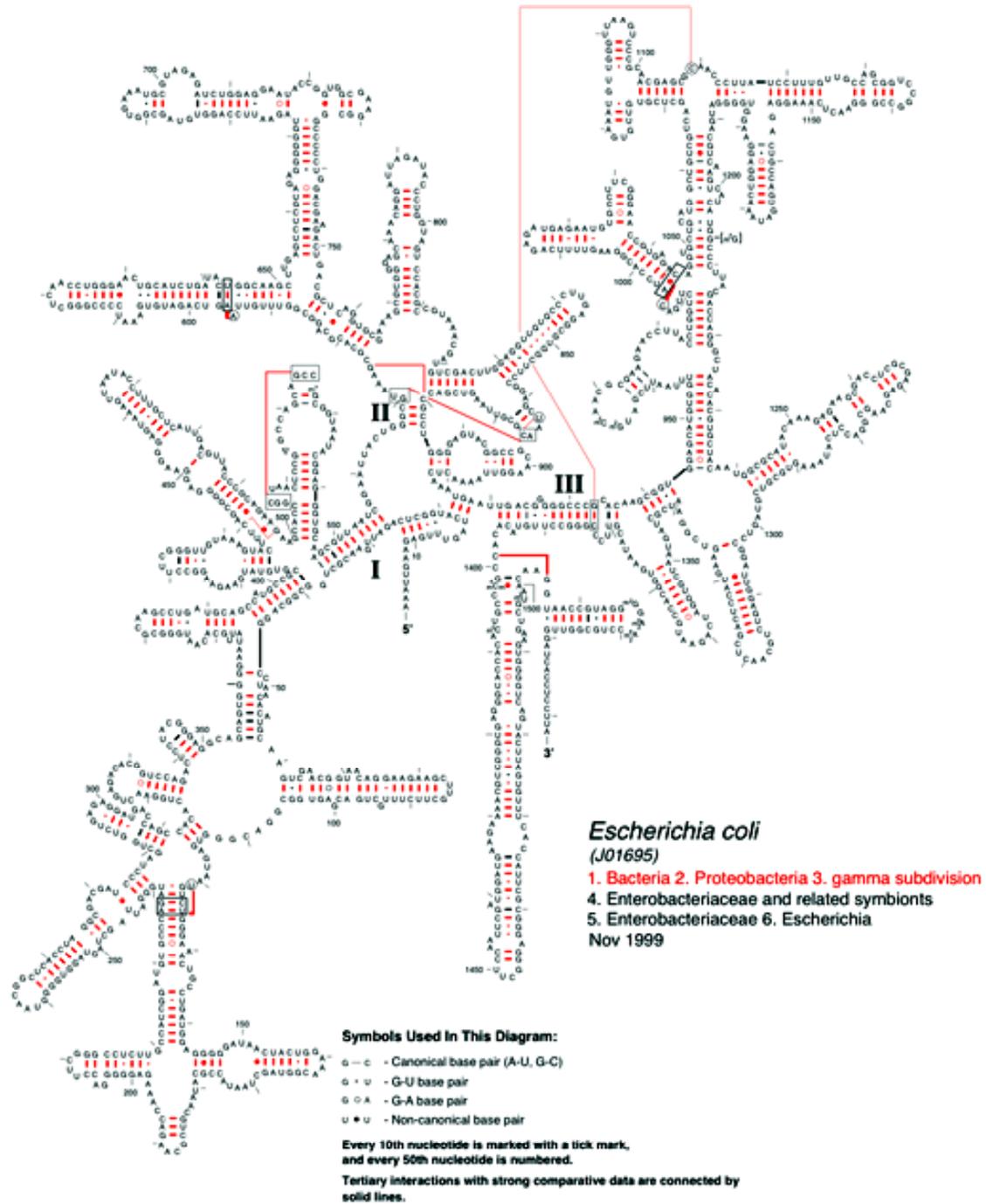
Every cell has a 16S rRNA gene; there are 1-10 copies in Bacteria, depending on the species, up to hundreds of copies in animal cells.

It is ~1550 bases in length.

Ribosomal RNA (rRNA) is an integral part of the *ribosome*, which is responsible for making cellular proteins.

The 16S rRNA molecule folds itself into a shape that must fit like a puzzle piece with other molecules (2 other RNAs & many proteins). For this to work. some structural elements must be maintained. Therefore, some parts of the sequence may vary and others may not.

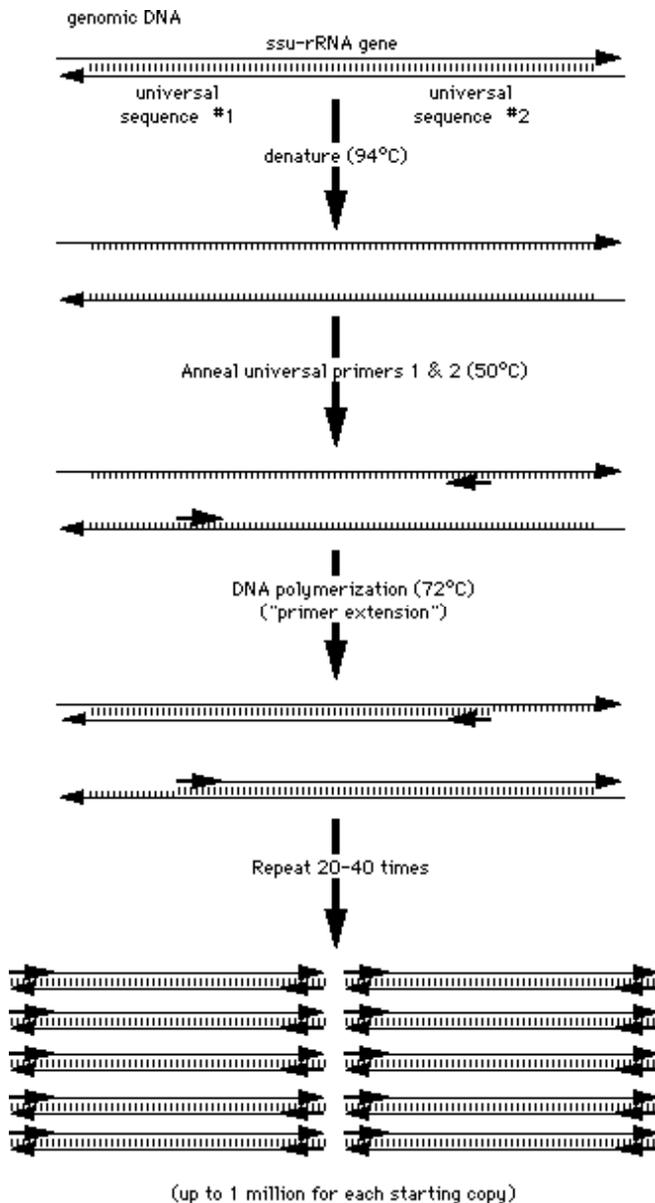
## Secondary Structure: small subunit ribosomal RNA



## Obtaining the sequence experimentally

The commonly used method these days, and the one we'll use in lab, is Polymerase Chain Reaction (PCR). PCR amplifies genes logarithmically - a single molecule of a gene, imbedded in the rest of the genomic DNA, is specifically amplified to up to a million molecules in 30 cycles! In a PCR reaction, 3 steps (denaturation, primer annealing, and DNA polymerization) are cycled over-&-over, each time doubling the amount of the specific DNA fragment.

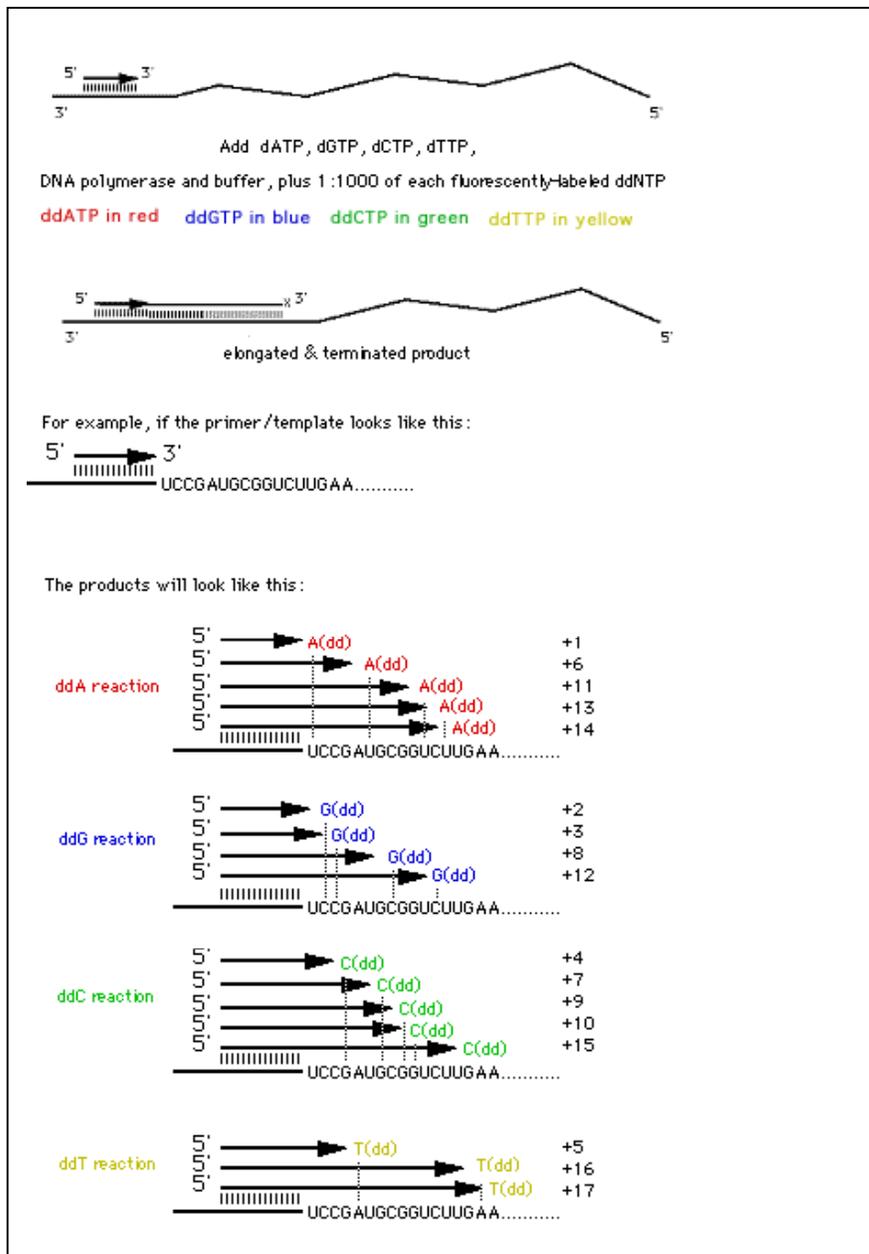
### PCR STEPS:



### SEQUENCING STEPS:

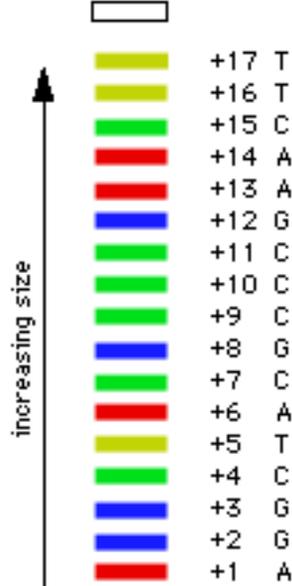
The PCR product is then sequenced - often using the same oligonucleotide primers that were used in the PCR reaction. Sequencing involves denaturing the DNA, annealing an oligonucleotide primer, and extending from this primer with DNA polymerase in the presence of dNTPs and small amounts of 'chain terminator' dideoxynucleotides (analogs of dNTPs that DNA polymerase cannot continue extending from):

1. Denature the DNA (separate the strands) with heat or high pH.
2. Anneal an oligonucleotide primer complementary to the DNA:

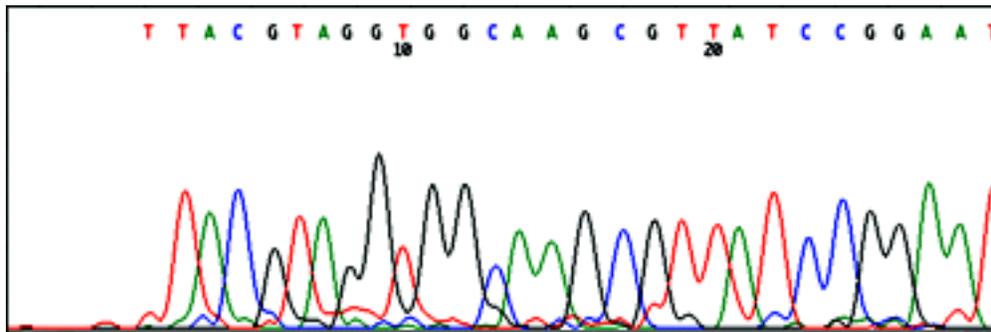


3. Add all the 4 dNTPs and a small amount of each of the 4 dideoxydNTP (ddNTP), each with a different fluorescent 'tag', and DNA polymerase.

4. Run sample on a high-resolution gel (one that can separate DNAs that differ by only a single base):



A fluorometer at the bottom of the gel detects the termination dyes as they run past in each lane of the gel (there are usually about 50 lanes per gel!). The connected computer collects this data and 'reads' the sequence from the pattern of peaks. The output from the computer looks like this:



(notice that the colors used here don't match the example)

Each reaction gives 300-700 bases of sequence, so it is usually necessary to use several primers spaced along the length of the molecule to get the complete sequence on an rRNA gene.

## Identifying homologous residues - i.e aligning the sequences

This is the most important part of the analysis!

example:

```

position: 1 2 3 4 5 6 7 8 9 10
seq A  A A A C U U G U U U
seq B  A C A C U U G U G U
seq C  A G A U U U - U C U

```

An alignment is a 2-dimensional matrix of multiple sequences. Each sequence is in a line (row) of the matrix. Each position (column) in an alignment contains homologous residues of each sequence. Gaps (shown as dashes) are added where needed to maintain the alignment. These gaps represent bases absent in that sequence that are present in some other sequence in the alignment.

Sequences must be fairly similar in sequence and length to be readily alignable 'by eye', or by computer alignment programs (e.g. Clustal). Thank goodness, most of the length of ssu-rRNAs are highly conserved and can (with experience) be manually aligned without much trouble.

Some of the tricks to aligning sequences are:

1. Sequences are often aligned sequentially - start by aligning the two most similar sequences, then add sequences to the alignment one at a time after this, starting with the sequences most similar to those already aligned and finishing with the most distantly related sequences. Likewise if you're adding a single sequence to an existing alignment, start by identifying the most similar sequence in the alignment & use that sequence as a guide.
2. Alternatively, you can identify conserved blocks of sequence in all of the sequences, and align these. You have now broken the alignment problem into smaller, easier chunks. Add gaps as need to align the space between pre-aligned chunks according to the criteria below.
3. Start out by finding patches of very similar sequences and align these, then work out in both directions from these, adding gaps sparingly when needed. Everything after this is about rearranging (and potentially adding or removing) these gaps.
4. Where there are sequence differences, slide the gaps around to keep purines (G, A) aligned with purines & pyrimidines (C, U) aligned with pyrimidines.
5. Try also to keep differences together in variable sequence positions, and align gaps together in columns wherever possible. A single gap of two positions is a lot better than two separate gaps of one position each!
6. Try to keep what look like conserved positions (columns) conserved, and all things being equal put differences into positions already known to be variable.

In more extreme cases, you can use the secondary structure of the RNA to identify homologous parts of the RNA, rather than relying only on sequence similarity.

This works because, in general, it doesn't matter (so much) to the RNA what the bases in the helices are, what matters is that they can form the correct secondary structure. As a result, the secondary structure of an RNA is much more conserved than it's sequence, because co-evolution of bases that form base-pairs maintains the secondary structure as the sequence changes. Variation in the length of the RNA is usually in hairpin lengthening or shortening. So it's usually possible to keep track of homologous parts of an RNA structure even if the sequences are completely different!

**See exercise on sequence alignment!**

**Questions for thought:**

- \* What are some DNA sequences that would not be good for phylogenetic analysis? Why?
- \* What are some other sequences that would be good for phylogenetic analysis, and in what situations would these be good?
  
- \* How did people get large amounts of a specific DNA for sequencing before PCR was invented?
  
- \* In an episode of the X-files, Agent Scully sequences some alien DNA and finds 'missing bands' in the sequences that she interprets to correspond to bases that are unique to aliens (not found in earthling DNA). Why is this not technically feasible?
  
- \* Given the variation of sequences in the context of the same secondary structure, can you imagine how scientists solve these secondary structures by comparative sequence analysis?
  
- \* Mutations occur one-at-a-time. Can you imagine, then, how the basepairs in a helix could change without disrupting the structure of the RNA? Does this explain (at least in part) why basepair changes that keep the purines (G or A) & pyrimidines (C or U) in the same positions are more common than those that switch them?
  
- \* Although RNA tertiary structures are rarely known, there are hundreds of protein 3D structures, determined by X-ray crystallography. Can you imagine a way to use these structures, in ways analogous to the use of RNA secondary structures are used, to align protein sequences more reliably?

## Tree construction

Phylogenetic trees represent the evolutionary relationships of sequences or species. There are several methods for building trees, including DISTANCE and PARSIMONY methods.

All methods of tree building from molecular data start with a sequence alignment. Here is an example alignment of 5 sequences with 25 positions in the alignment:

```
Seq A  AGAUUCGUCUGUAGGUUUCCACCAA
Seq B  ACAUUCGUGUAUAGGUUUCCACUAA
Seq C  ACAUUCGUGUAGAGGUUUCCACUAA
Seq D  AAGUUCGCUUGGAGGUUUCCACGAA
Seq E  AUCGUGAGAUC CAGGUAUCCACAAU
```

DISTANCE: The first step toward building a DISTANCE tree is to generate a distance matrix: Just count the fraction of identical bases in every pair of sequences in the alignment.

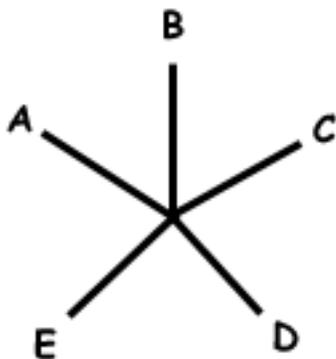
```
Seq A  AGAUUCGUCUGUAGGUUUCCACCAA
      |X|||||X|X|||||||X|| 3/25 = 0.16
Seq B  ACAUUCGUGUAUAGGUUUCCACUAA
```

	A	B	C	D	E
A	----	----	----	----	----
B	0.16	----	----	----	----
C	0.20	0.04	----	----	----
D	0.24	0.28	0.24	----	----
E	0.48	0.48	0.48	0.48	----

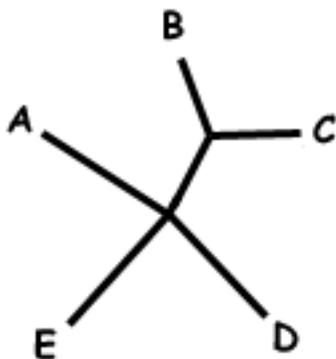
In this example, sequences A and B differ by 0.16 (= 16%), A and C are separated by 0.20, B and C differ by 0.04, etc, etc.

These distances can be used to make a tree by joining the most similar sequences, one after another, in pairs. This is called the “neighbor-joining method”. The structure of the tree is determined first, then the branch lengths are fit to this skeleton. The tree starts out with a single internal node at a branch out to each sequence.

First, the closest relatives ("neighbors") are "joined" onto a single branch, and then the process is repeated using the average distances from the original neighbors. Using our distance matrix, the tree starts out like this (remember that we're sorting out the structure of the tree, not the branch lengths, yet).



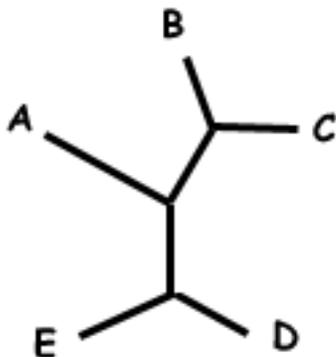
The closest neighbors in the distance matrix are B and C (0.04), so these branches are joined:



The distances from all other sequences to B and C are then averaged to reduce the distance matrix:

	A	B/C	D	E
A	-	-	-	-
B/C	0.18	-	-	-
D	0.29	0.26	-	-
E	0.48	0.48	0.48	-

Now the closest neighbors are A and B/C, so join them:

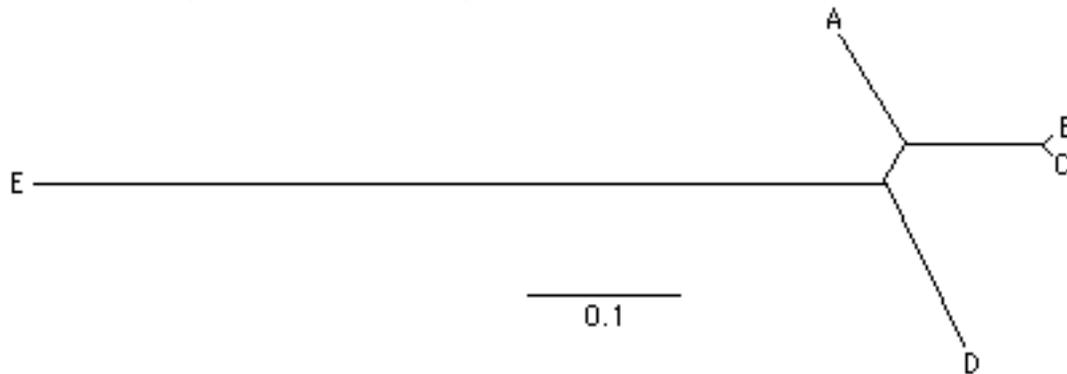


That's it! The tree is completely resolved when there are only 3 lines coming from each node.

If there were more sequences, you'd re-reduce the matrix as before (A with B and C to make A/B/C), & repeat the process over-and-over until all of the nodes were resolved. But usually this is done with a computer program. Imagine how hard this would be if you had 30 sequences, each 1000 bases or much longer. Neighbor-joining is really fast in terms of computer time, and so can be used on trees with much larger numbers of sequences than other methods.

In the Ribosomal Database Project, there are over 100,000 16S rRNA sequences, each about 1500 bases long. These have been used to produce a universal Tree of Life.

Once the structure (topology) of the tree is determined, it is an easy computational task to calculate the lengths of all the branches that fit best to the original distance matrix by minimizing the total least-squares deviation. When you do this, the tree looks like this:



This tells you that seq E is more distantly related relative to the others and that B and C are very closely related.

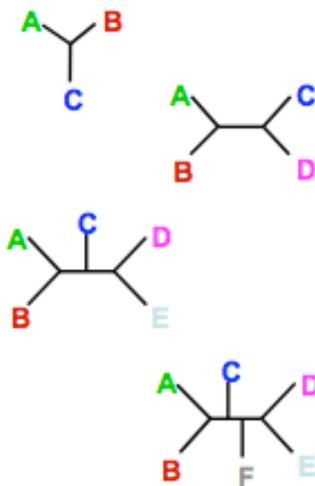
Notice that the distance between any two sequences is (approximately) equal to the sum of the length of the line segments joining those two sequences - in other words, the tree is additive.

#### PARSIMONY:

Another tree-making method, called Parsimony or Minimum Evolution, finds the tree that requires the fewest number of total changes on the whole tree. In other words, the method calculates the “Length” of each tree, as the total changes on all the branches and then compares this value, selecting the tree that has the lowest value.

### A) HOW MANY PHYLOGENETIC TREES ARE THERE?

The number of unrooted trees increases in a greater than exponential manner with number of OTUs (taxa)

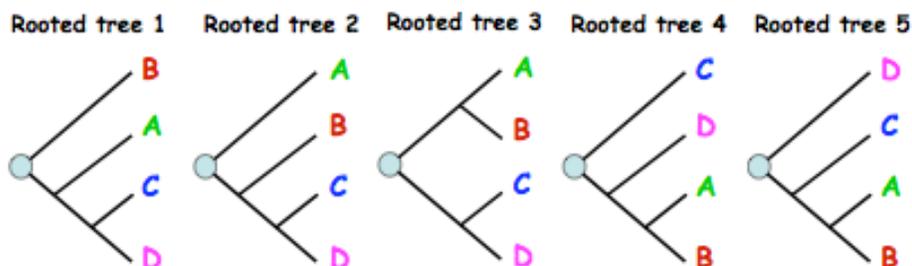
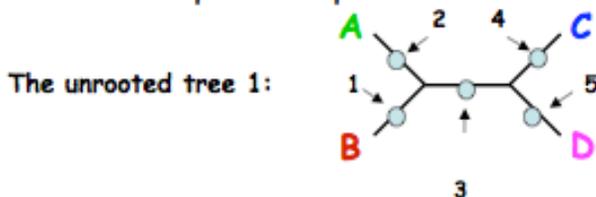


# Taxa (N)	# Unrooted trees
3	1
4	3
5	15
6	105
7	945
8	10,935
9	135,135
10	2,027,025
.	.
.	.
.	.
30	$\approx 3.58 \times 10^{36}$

$$Nu = (2N - 5)! / 2^{N-3} (N-3)! \\ = \# \text{ unrooted trees for } n \text{ taxa}$$

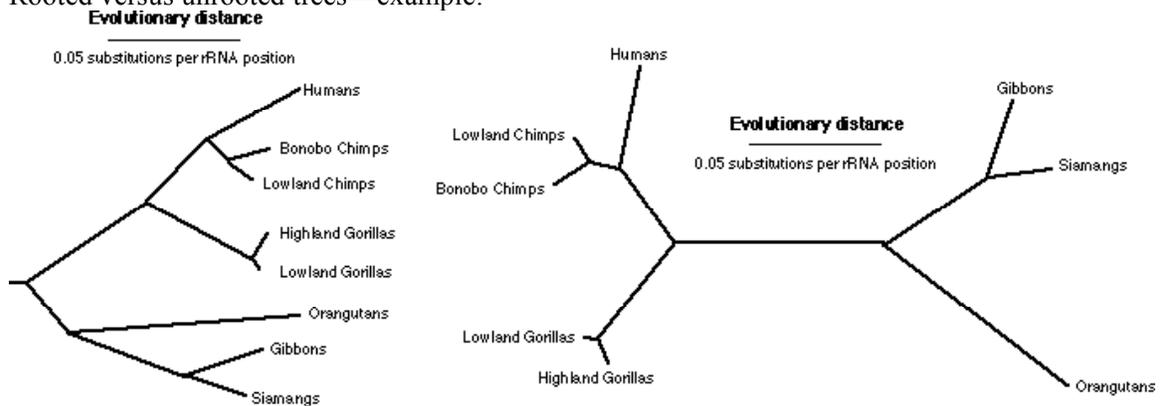
### A) HOW MANY PHYLOGENETIC TREES ARE THERE?

An unrooted, four-taxon tree theoretically can be rooted in five different places to produce five different rooted trees



Why might it be difficult to use PARSIMONY for finding the shortest tree if you have 100 species (sequences)? (How many trees must be evaluated?)

Rooted versus unrooted trees—example:



### Rooting the tree with an outgroup

An outgroup is a sequence that is known to be outside of the group you're interested in treeing. By including an outgroup can you locate the root of a tree. For example, if you were building trees from mammalian sequences, you might include the sequence from a reptile as an outgroup.

More on how to read a phylogenetic tree

A phylogenetic tree is a representation of the evolutionary/geneological relationships between a collection of organisms (or molecular sequences). There are many different ways to draw these trees, but they share a common set of features (please note that the trees that follow are generalized approximations):

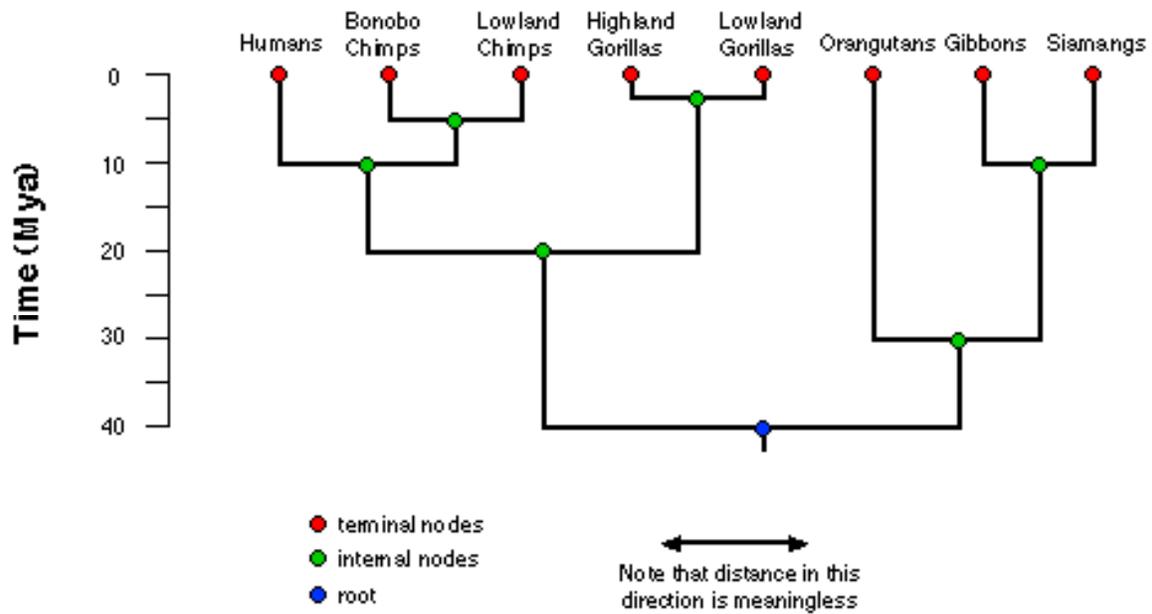
- \* **Scale.** This typically is either time or evolutionary divergence. Trees with a time scale are based on some form of physical data, such as a fossil record, that provide dating information. If the scale is time, all modern organisms should obviously be shown at the same part of the scale. More often, the scale is evolutionary divergence, some measure of change in the organisms (or molecules). Because the extent of divergence is usually different in various parts of the tree, this is usually depicted by varying the lengths of the branches and providing a scale bar.

- \* **Terminal nodes.** These are the ends of the branches of the evolutionary tree - typically the modern organisms (or molecules) that are being compared, but in some cases the ends of evolutionary branches that became extinct.

- \* **Internal nodes.** These represent the last common ancestors of all of the organisms (or molecules) bound by this node.

- \* **Root.** This is the 'base' of the tree - the last common ancestor of all of the organisms (or molecules) in the tree. It is not always possible to identify the root of a tree - typically, this requires either physical data (e.g. a fossil record) or data about organisms outside of the part of the tree shown.

- \* **Branches.** These are the connections between nodes in the tree. These represent to evolutionary pathway between common ancestors (internal nodes) and modern organisms (terminal nodes). The length of these branches is defined by the scale - each branch represents a certain amount of historical time, if time is the scale used in the tree, or a certain amount of evolutionary change, if evolutionary divergence is the scale used in the tree.



Questions for thought:

\* What organism would you choose for an outgroup for an rRNA tree of mammals? As outgroup for an rRNA tree of Bacteria? What about for a "universal tree" containing sequences or all kinds of organisms?

\* What would a tree of some animals look like if constructed from globin genes where some of the sequences were alpha globins & others were beta globins?

\* What would a tree (no pun intended) of plants look like if some of the sequences (rRNAs) were accidentally taken from the chloroplast instead of the nucleus. What if all of the sequences were from the chloroplasts?