We need an optimality criterion to choose a best estimate (tree)

Parsimony: begins with the assumption that the simplest hypothesis that explains the data is probably the correct one. Assume that change is rare, and select the tree that requires the least amount of change along its branches to produce the data.

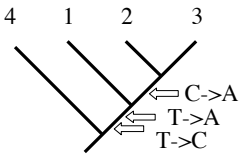
(In this example, we use simple morphological characters, but this method is also used with molecular sequence data.)

Other optimality criteria used to choose a best estimate (tree)

Distance: Based on the assumption that closely related organisms are going to be more similar. Construct a distance matrix, and select the tree that *minimizes the differences* (distances) between taxa.

1 ATGCGTCTC 2 ATGAGTCTC 3 ATGAGTCTC 4 TTGCGTCCC

	1	2	3	4
1	-	1	1	2
2	1	-	0	3
3	1	0	-	3
4	2	3	3	-



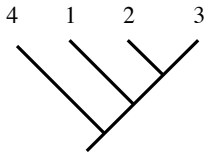
Tree rooted arbitrarily

Other optimality criteria used to choose a best estimate (tree)

Maximum Likelihood (for DNA sequence data): Start with a model of nucleotide evolution, then begin examining possible trees. Ask: what is the likelihood that a given tree would have produced the actual observed sequence data under the model of evolution? The most optimal tree is the one with the highest likelihood score.

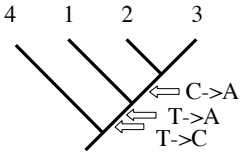
1 ATGGGTCTC 2 ATGAGTCTC 3 ATGAGTCTC 4 ATGCGTCTC

	A	T	G	C
A	-	.17	.20	.12
T	.17	-	.14	.20
G	.20	.14	-	.17
C	.12	.20	.17	-



Tree rooted arbitrarily

Note that in this simple example: all three optimality criteria (parsimony, distance, and maximum likelihood) would have given us the same answer. This increases our confidence in the results.



Tree rooted arbitrarily

In more complex analyses, there is usually *conflict* (disagreement) between trees derived from different optimality criteria (or even different assumptions within the same criterion). An important part of phylogenetic analysis is sorting through this conflict to arrive at the best phylogenetic estimate

Ideally, under any optimality criterion (parsimony, distance, or maximum likelihood) we would like to examine every possible tree and give it an optimality score before selecting the best one.

However, this quickly becomes impossible, even with a computer.

No. of taxa	No. of possible trees
4	3
5	15
6	105
7	945
10	2×10^6
11	34×10^6
50	3×10^{74}

Therefore, scientists use algorithms that explore the *tree space* without examining every possible tree. These methods are not guaranteed to find the best phylogenetic estimate(s) for the data, but they often do.

Non-exhaustive ways to explore tree space:

Neighbor-joining: use distance information to assemble a tree additively, one taxon at a time. This method does not evaluate every possible tree.

Heuristic: use random starting trees and "swap" branches around, looking for more optimal alternatives. Replicate many times.

The key point is: since we cannot evaluate every possible tree, we do everything we can to increase our confidence that we have found the best "island" in treespace (the most optimal set of trees under our optimality criterion). This is why we replicate 1000, 10,000, or even a million times or more.

Why is this all so complicated? What is the TRUE TREE?

A true tree does exist -- it is the evolutionary history of the organisms or genes in question.

But since we don't have a time machine, all we can do is attempt to *reconstruct* that history, which requires us to make assumptions, choose optimality criteria, and model evolution

Consider that a gene may contain both conserved areas that evolve slowly, and variable areas that evolve more rapidly. Thus, no model of molecular evolution could ever accurately describe what has happened to the whole gene sequence.

Robustness:
How strongly is a phylogenetic hypothesis supported by the data?

Bootstrap replicates generate new data sets by randomly sampling from the actual data, with replacement. These new data sets should contain phylogenetic signal similar to that in the original data. A high percentage of replicates (75%+) that support a grouping of interest indicates that the actual data support that grouping well.

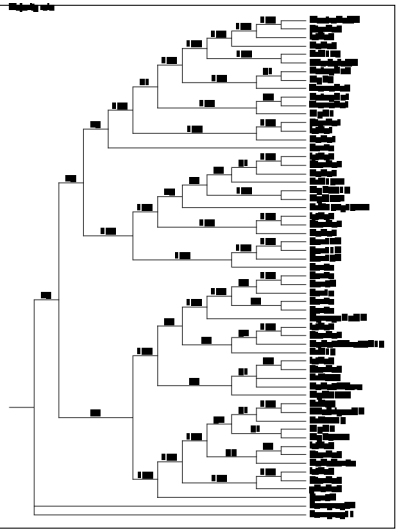
Robustness:
How strongly is a phylogenetic hypothesis supported by the data?

Bootstrap replicates generate new data sets by randomly sampling from the actual data, with replacement. These new data sets should contain phylogenetic signal similar to that in the original data. A high percentage of replicates (75%+) that support a grouping of interest indicates that the actual data support that grouping well.

Bayesian methods examine a large sample of possible trees with the best likelihoods, and ask what percentage of those trees retain a grouping of interest. This percentage is the posterior probability. Generally we are interested in p.p.'s of 95% and up.

REMEMBER: Analyses and results are only as good as the data!

For example, if these numbers were bootstrap values, I'd be in good shape with my tree, relative to my data. However, these numbers are Bayesian posterior probabilities, and many deep nodes have low support.



WHY ANALYZE ONE TYPE OF DATA, AND NOT ANOTHER?

- Some genes are very **conserved,** and will be useful for examining **ancient** divergences, or splits. **Highly conserved genes evolve slowly.**

WHY ANALYZE ONE TYPE OF DATA, AND NOT ANOTHER?

- Some genes are very **conserved,** and will be useful for examining **ancient** divergences, or splits. **Highly conserved genes evolve slowly.**

- However, a gene may be **so** conserved that it will be invariant (identical) among the descendants of more recent evolutionary splits. In such cases, pick a gene that is **less conserved** and has more **variation,** i.e., pick a gene that evolves **more rapidly.**

WHY ANALYZE ONE TYPE OF DATA, AND NOT ANOTHER?

- Some genes are very **conserved,** and will be useful for examining **ancient** divergences, or splits. **Highly conserved genes evolve slowly.**
- However, a gene may be **so** conserved that it will be invariant (identical) among the descendants of more recent evolutionary splits. In such cases, pick a gene that is **less conserved** and has more **variation,** i.e., pick a gene that evolves **more rapidly.**
- **BUT**, if the gene you pick is **too** variable, the sequence data will also be too variable to analyze. It may even approach a random distribution!

WHY ANALYZE ONE TYPE OF DATA, AND NOT ANOTHER?

- Some genes are very **conserved,** and will be useful for examining **ancient** divergences, or splits. **Highly conserved genes evolve slowly.**
- However, a gene may be **so** conserved that it will be invariant (identical) among the descendants of more recent evolutionary splits. In such cases, pick a gene that is **less conserved** and has more **variation,** i.e., pick a gene that evolves **more rapidly.**
- **BUT**, if the gene you pick is **too** variable, the sequence data will also be too variable to analyze. It may even approach a random distribution!
- **Therefore**, what is really needed is a gene which evolves at a rate that provides a good **balance** between conservation and variation. Or better yet, resolve splits of different ages by sequencing more than one gene

How did we estimate the phylogeny of the **Tree of Life**, when organisms are so different? There are not likely to be many sequence homologies between bacteria, archaea, and eukaroytes.

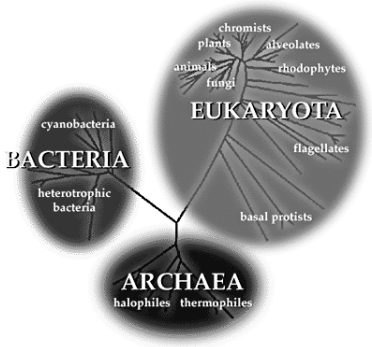
Solution:

Sequence information that is so ancient and so fundamental to living things that all organisms must have it.

RNA

Or more specifically,

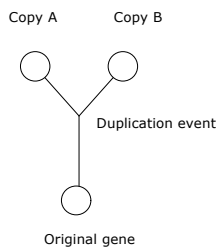
Ribosomal RNA



Gene duplication

- Physical duplication of a stretch of DNA, producing two (initially) identical sequences in the genome
- Can occur at a range of scales from a few bases to the entire genome
- A range of different mechanisms

Gene duplication

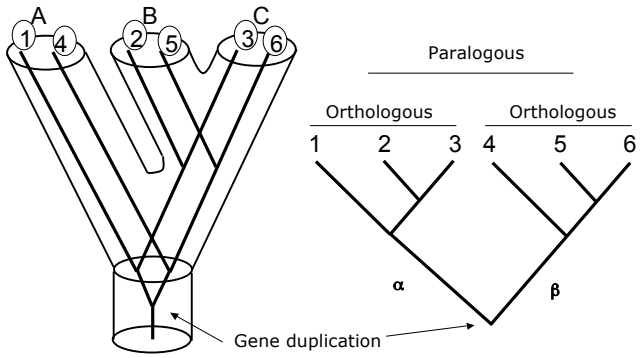


- Two copies of original gene
- Copy B may be "lost" (e.g., lose function due to mutation)
- Copy B may evolve new function (A retains original function)
- Copy B may persist relatively unchanged (provides redundancy)
- Copy A & B may divide the function of the original

So why are they interesting?

- New gene functions
- Gene duplications structure genomes
- Important for molecular phylogenetics

Orthology and Paralogy



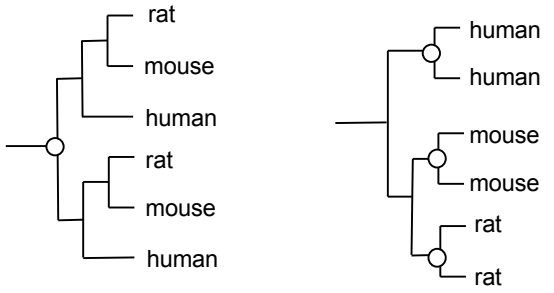
Why orthology matters

- Inference of function is best made between orthologous sequences (paralogues may have different function)
- Inference of species relationships should be based on orthologous genes

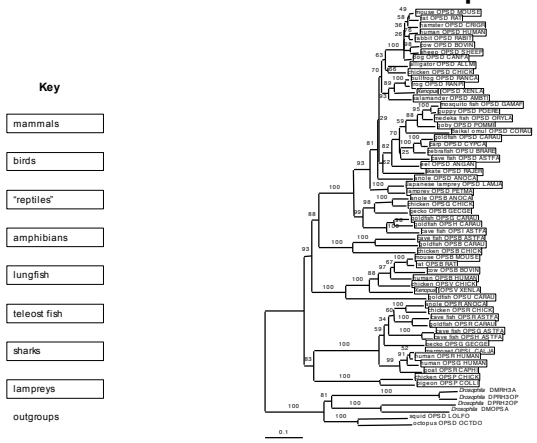
Recognising Orthology and Paralogy

- Sequence Similarity
  e.g. BLAST search

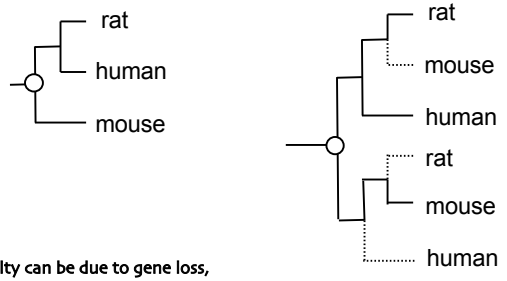
Easy..



More difficult - Rhodopsin



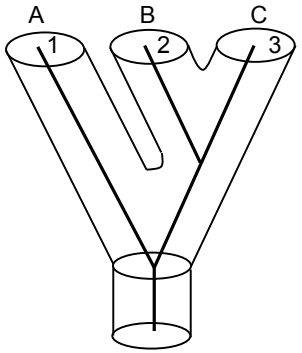
Impossible?



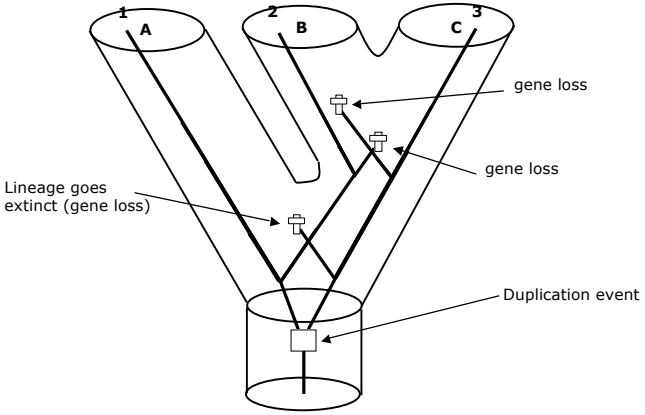
Difficulty can be due to gene loss, gene deletion or failure to sample

- Using DNA sequences to infer something about SPECIES relationships makes a fundamental assumption..

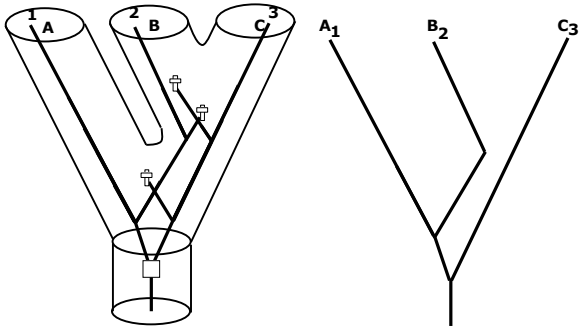
Assumption:
Gene tree = species tree



Duplication and loss



= incongruent gene and species trees



Is paralogy common?

Rates of Gene Duplication are high..
Drosophila maybe 10^{-4} or 10^{-6} per gene per generation
0.001 - 0.03 /gene/myr for a range of eukaryotes

Gene families are very common.

Up to 75% of genes in vertebrates are non-unique genes (I.e., are part of some gene family)

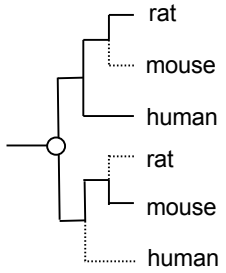
Why orthology matters

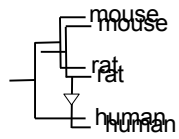
- Inference of species relationships should be based on orthologous genes
- But we don't (for sure) know they're orthologous until we know the relationships

What to do?

- Use putatively non-duplicating genes (mitochondria, rRNA)
- Sometimes we can spot paralogues (look for variation in introns, regulatory regions etc.)
- Do a series of different analyses, using different genes each time.

Lateral (Horizontal) Gene Transfer can look exactly like duplication-and-loss





One duplication and 3 losses
OR
1 LGT event