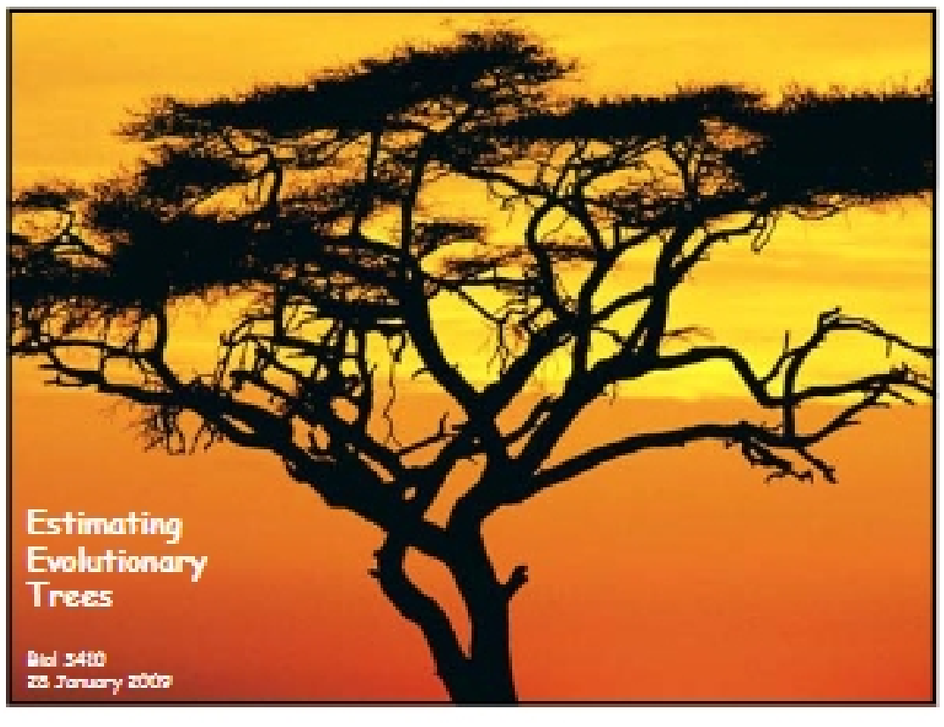
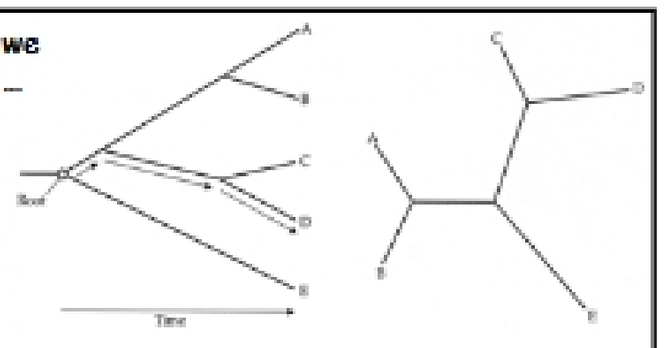
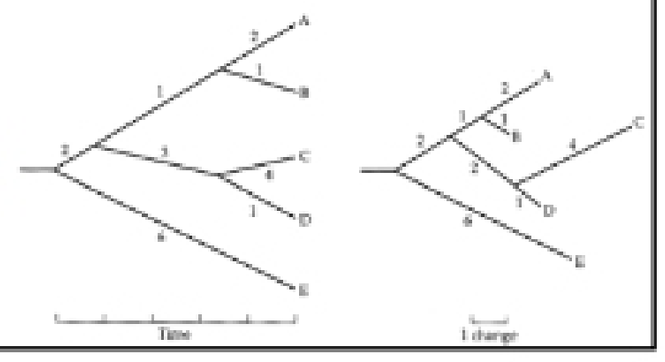


About the ways we represent trees …

They can be drawn *rooted* (with time flowing down from the root), or *unrooted* (with the direction of time ambiguous).



They can be drawn with branch lengths proportional to estimated *time*, or to estimated *numbers of evolutionary changes* (*e.g.*, nucleotide substitutions).

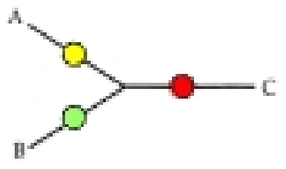


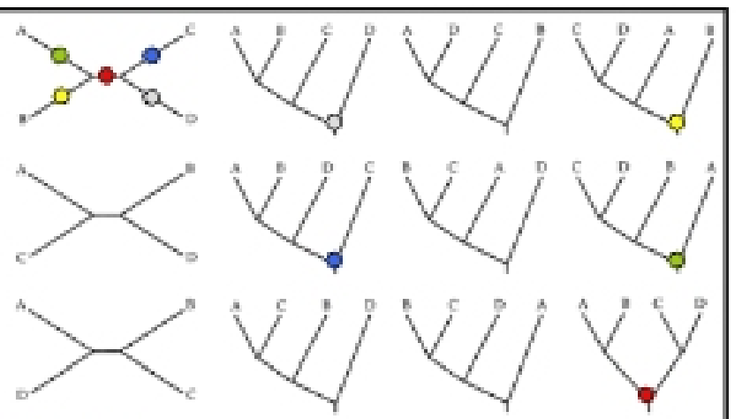
Rooting an unrooted tree

A different rooted tree results from placing the root on each of the tree's branches.

An unrooted tree has 2n-3 branches, so there are 2n-3 possible rootings.

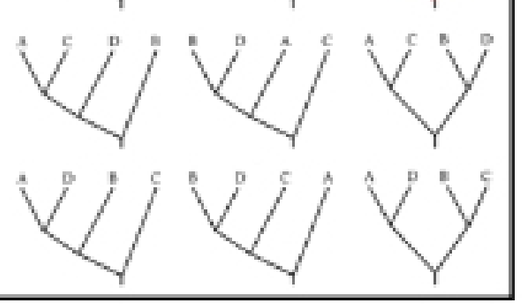
For three tips (Operational Taxonomic Units, or OTUs), there is only one unrooted tree, and hence a total of 2*3 - 3 = 3 rooted trees.



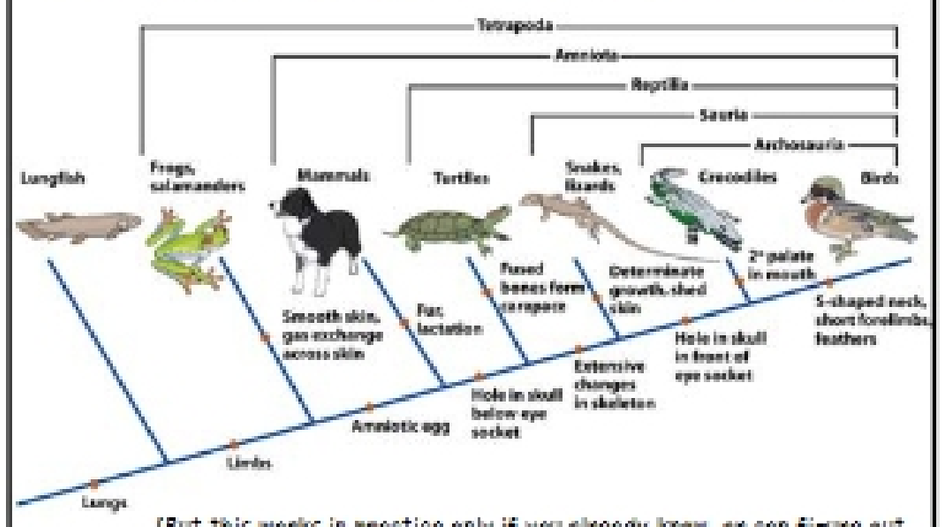


With 4 OTUs there are 3 different unrooted trees, each giving rise to 5 different rooted trees, for a total of 15.

(Problem: You do the others.)



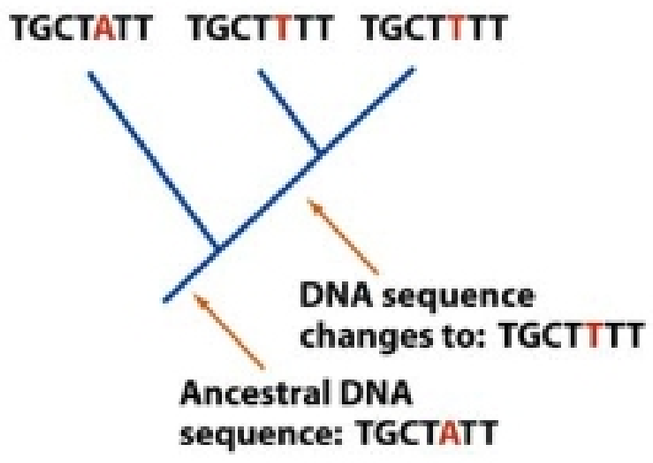
Synapomorphies ("states together toward the tips of the tree") or *shared derived characters* can identify monophyletic groups



(But this works in practice only if you already know, or can figure out, which similarities are *synapomorphies* and which are *homoplasies*).

Mutations can create synapomorphies

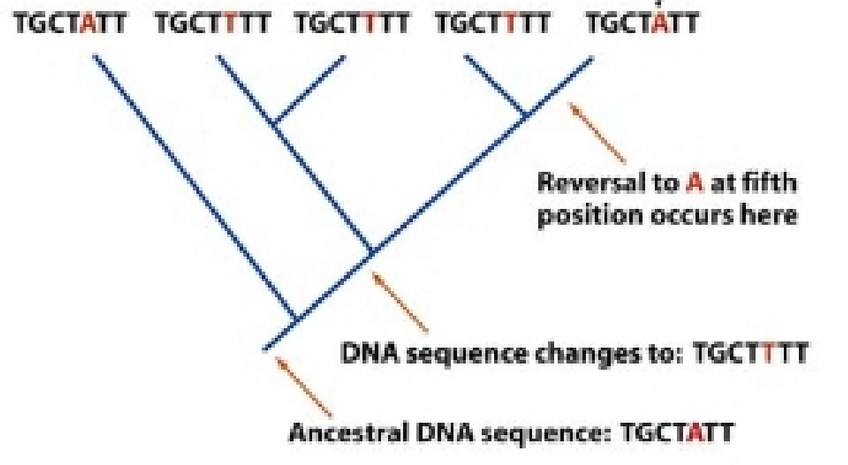
DNA sequences in descendants:



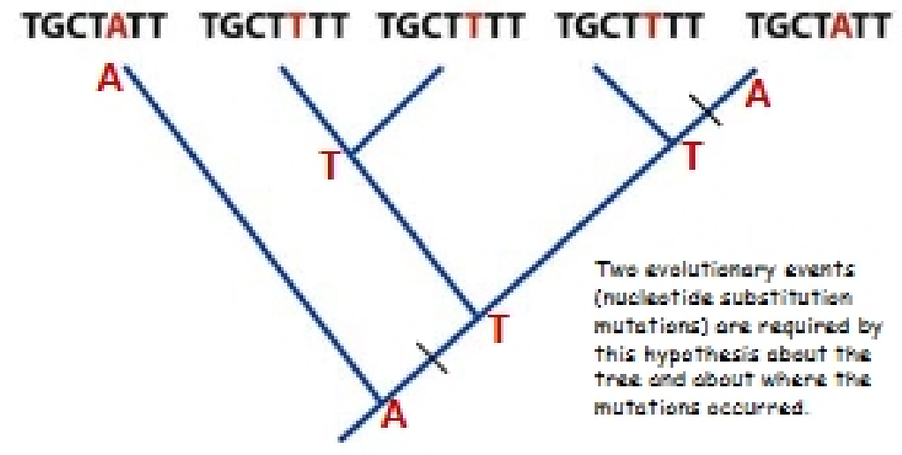
Reversals ("back-mutations") can remove synapomorphies

"Homoplasy": Similarity not caused by descent from a shared ancestor

DNA sequences in descendants:

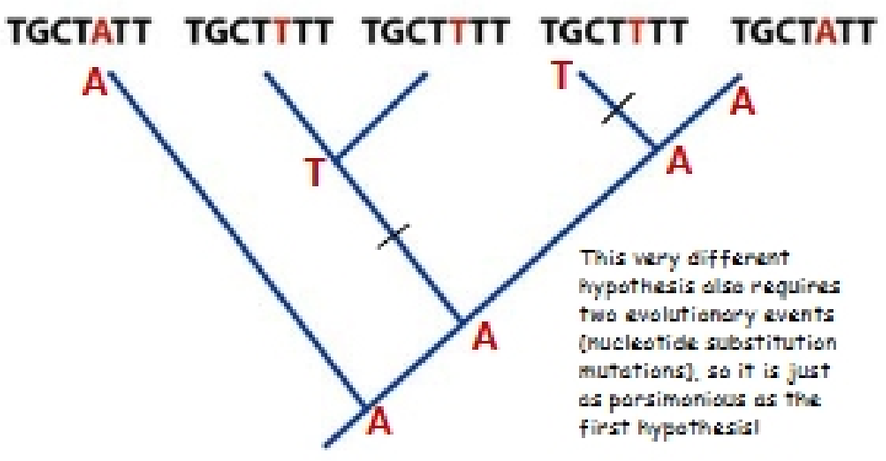


The principle of *parsimony* compares alternative hypotheses about the history of character-state change



Two evolutionary events (nucleotide substitution mutations) are required by this hypothesis about the tree and about where the mutations occurred.

An alternative hypothesis (or reconstruction, or inference)

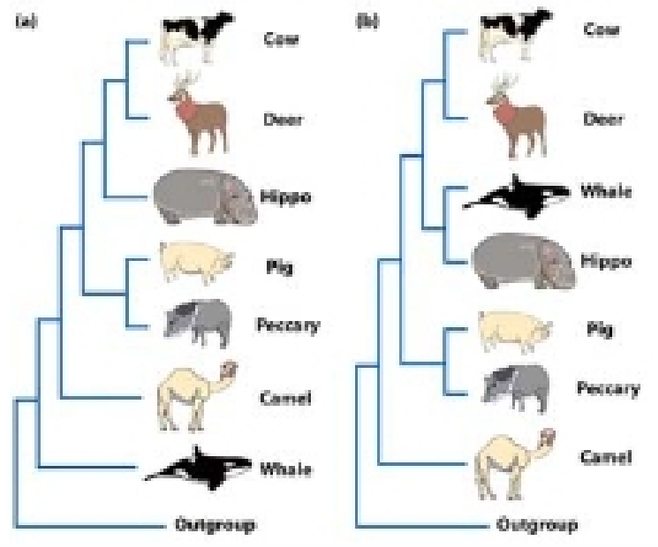


This very different hypothesis also requires two evolutionary events (nucleotide substitution mutations), so it is just as parsimonious as the first hypothesis!

The *cladistic* method (grouping by shared derived characters) runs into trouble when we can't be sure which is the derived state.

One solution: use characters where you know which states are derived!

Question: Are whales *outside* or *inside* the Artiodactyla?



SINEs (Short Interspersed Nuclear Elements) transpose with "borrowed" reverse transcriptase

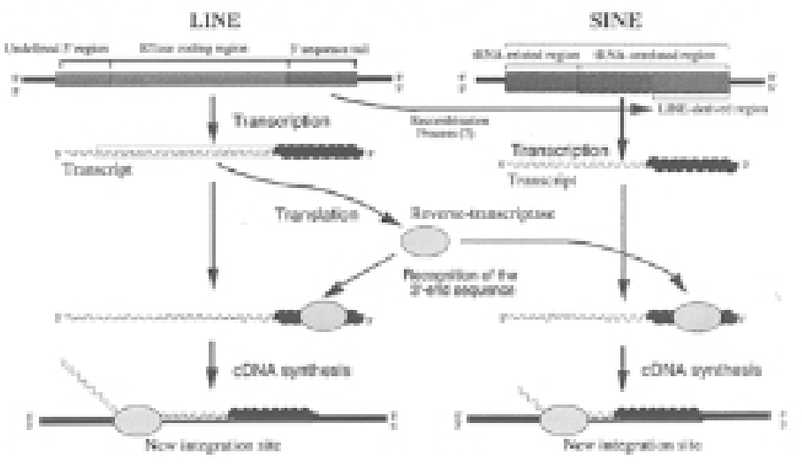


Figure 1. General model for a possible RNA-derived SINE amplification process. Corresponding LINE and SINE components are color-coded and share a common region (green) due to a recombination process. Reverse transcriptase (yellow) is generated by a LINE and the corresponding SINE transcript can be recognized by its LINE-derived tail region (dots; black 3′ end). The SINE transcript is then reverse transcribed into cDNA and integrates into the host genome (red site) by the target-DNA-primed mechanism adopted by LINEs.[REVIEWS]

SINE insertions: unique, irreversible, and easily assayed

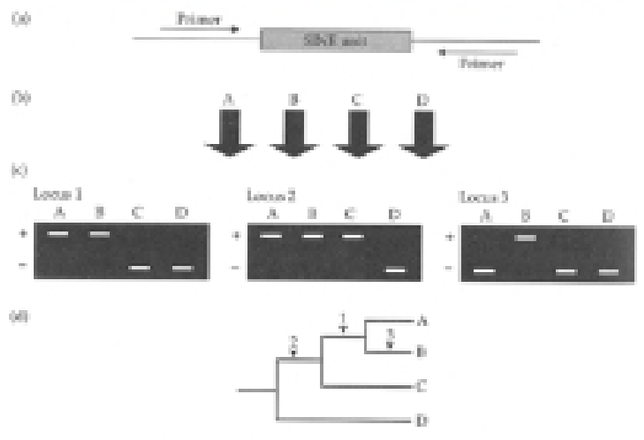
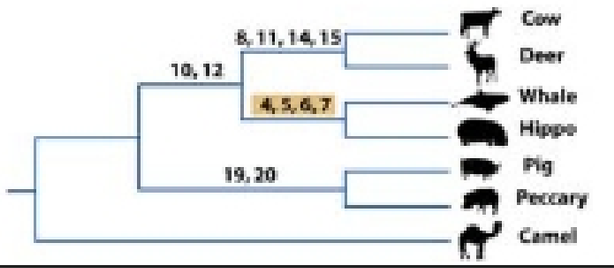
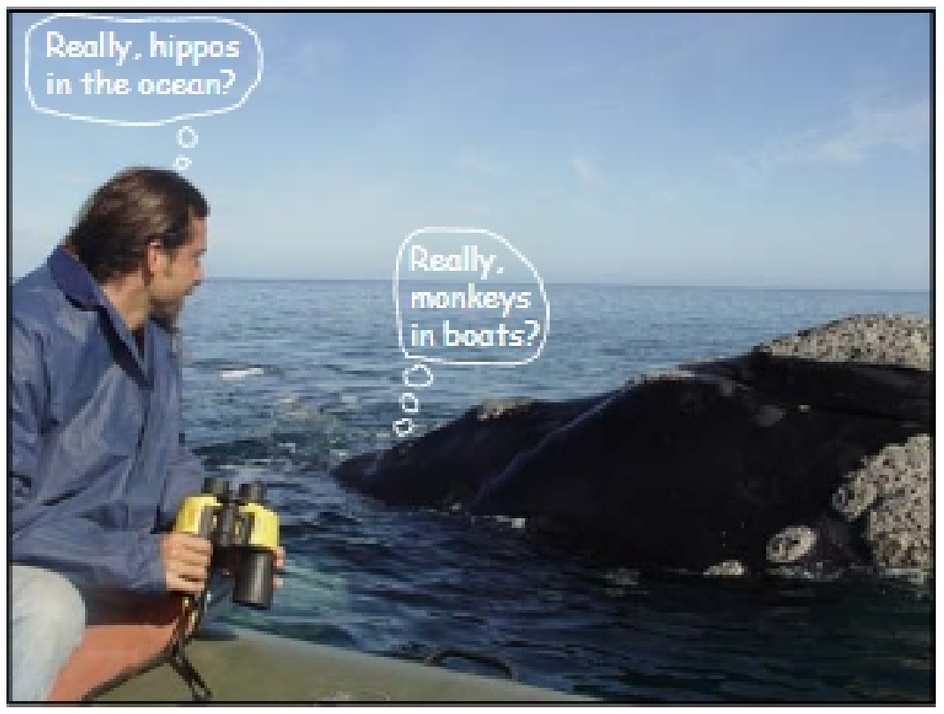


FIGURE 8.32 Inference of phylogeny from insertion patterns of SINEs. (a) The primers identify the genomic location (locus) of a SINE unit. (b) PCR is used to amplify the homologous loci from the genomic DNA of several species under study (A, B, C, and D). (c) The PCR products are subjected to electrophoresis and separation by length. A long PCR product indicates the presence (+) of a SINE unit; a short PCR product indicates absence (–). (d) Because SINE insertion is essentially an irreversible character state, the presence of a SINE at a certain locus may be treated as a synapomorphy defining monophyletic clades (arrowheads 1 and 2) or as an autapomorphy for a single taxon (arrowhead 3). Courtesy of Professor Norihiro Okada.

And the SINEs settle it; whales are hippos!

Locus	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Cow	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	0	0
Deer	0	0	0	0	0	0	0	1	?	1	1	1	1	1	1	?	1	1	0	0
Whale	1	1	1	1	1	1	1	0	?	1	0	1	1	0	0	0	?	1	0	0
Hippo	0	?	0	1	1	1	1	0	1	1	0	1	1	0	0	0	?	1	0	0
Pig	0	0	0	?	0	0	0	0	?	0	0	0	?	?	0	0	0	1	1	1
Peccary	?	?	1	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	1	1
Camel	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0





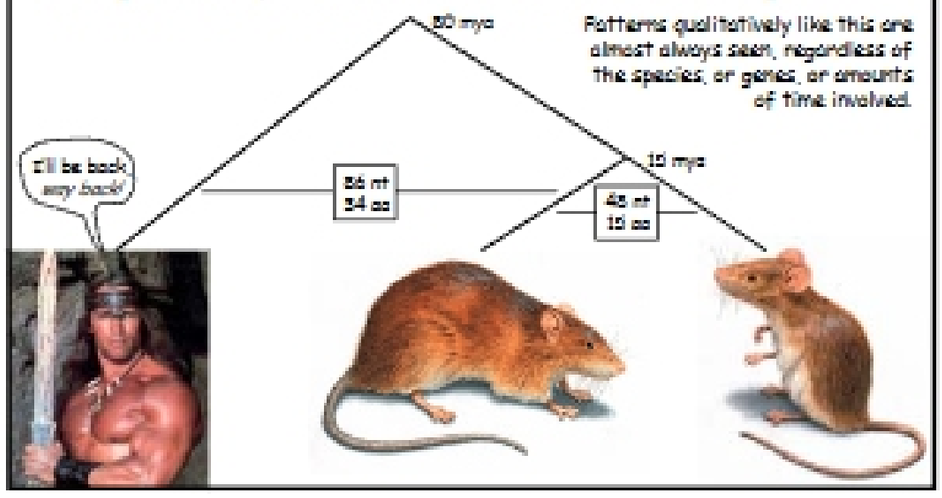
A simpler approach: use sequence differences as a molecular "clock"

Rat and mouse last had a common ancestor around 15 million years ago (mya).

Their I7 odorant receptor genes differ at 48/981 nucleotide positions, and the I7 proteins encoded by those genes differ at 15/327 amino-acid positions.

Humans and rodents last had a common ancestor around 80 mya.

Their I7 genes differ by around 86 nucleotides and 34 amino acids, on average.

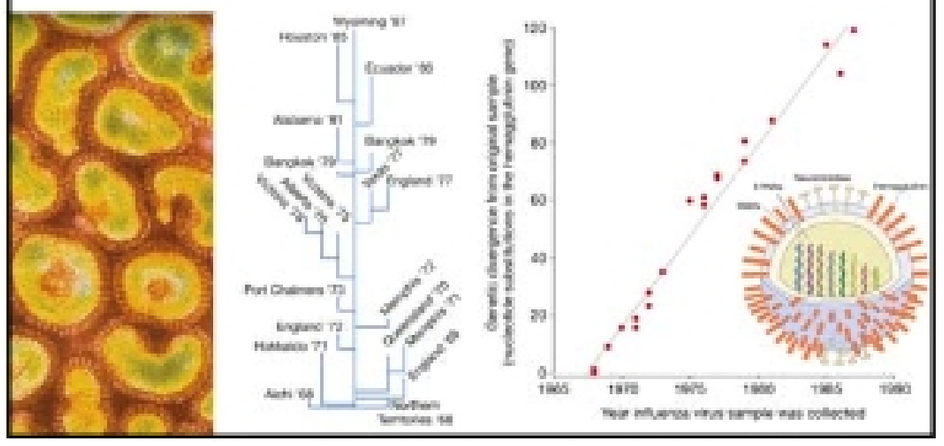


Patterns qualitatively like this are almost always seen, regardless of the species, or genes, or amounts of time involved.

Molecular clocks are never perfect, but they are often remarkably consistent, and they can be calibrated with fossils (or history!)

The flu-virus clock has been calibrated directly, by analyzing viruses sampled at many times during the last several decades.

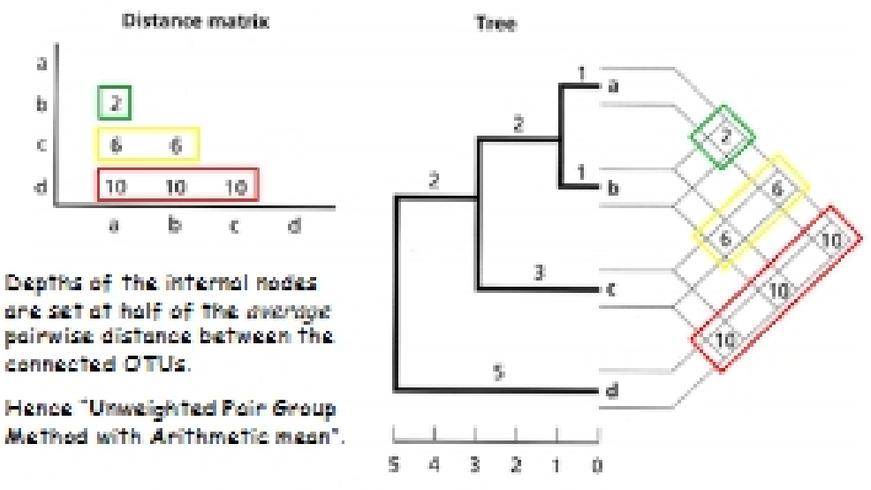
These data for the virus's hemagglutinin gene show a steady accumulation of nucleotide substitutions over a period of more than 20 years.



Trees are unambiguous when distances are proportional to time!

And that's what we really want to know: when did the lineages diverge?

The simplest tree-inference algorithm is UPGMA (sequential clustering)

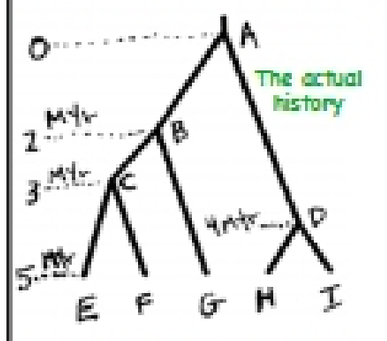


Depths of the internal nodes are set at half of the *average* pairwise distance between the connected OTUs.

Hence "Unweighted Pair Group Method with Arithmetic mean".

From D.M. Hillis and B.K. Holmes, *Molecular Evolution: A Phylogenetic Approach* (Blackwell, 1990)

A made-up example where we know exactly what happened

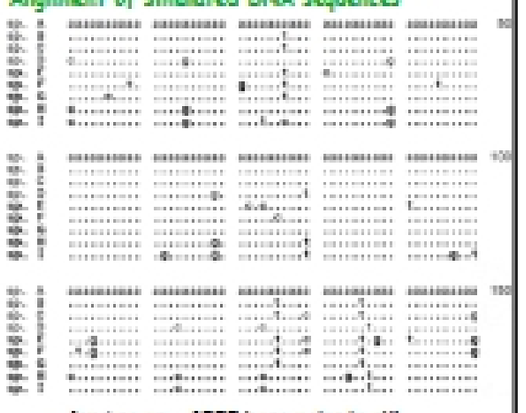


Jon's computer "evolved" the ancestral sequence (1000 a's) down this tree, at a rate of 0.01 substitutions per site per million years, or $\mu = 1 \times 10^{-8}$ subs/site/yr.

Matrix of pairwise differences

	E	F	G	H	I
Sp. E		39	60	103	104
Sp. F	39		52	98	99
Sp. G	60	52		96	96
Sp. H	103	98	96		15
Sp. I	104	99	96	15	

Alignment of simulated DNA sequences



(and so on ... 1000 base pairs in all)