

Some phylogenetic analysis (“tree-thinking”) problems involving whales

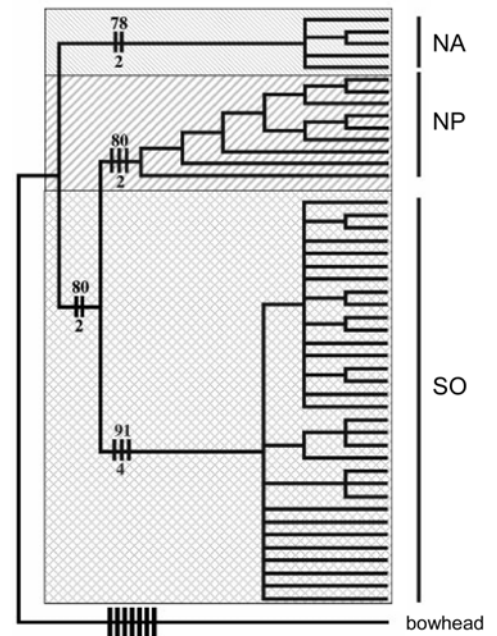
The relationships of North Pacific (NP), North Atlantic (NA), and Southern Ocean (SO) right whales have long been controversial. They all look very similar to each other, but they appear not to cross the equator. Some systematists believe they should be considered a single species; some believe they should be considered three species; and some distinguish just two (the southern right whale *Eubalaena australis* and the northern right whale *E. glacialis*, defined to include both the NP and NA populations).

1. The matrix below gives pairwise nucleotide differences for the combined sequences of two mitochondrial genes (*COI* and *cytb*, totaling around 2600 base pairs) recently sequenced for NP, NA and SO right whales and the bowhead whale, which is their closest living relative. Using the UPGMA method (successive clustering, with distance averaging), infer the phylogeny. Be sure to label the tips of your tree, and also indicate the depths of the internal nodes (in *approximate* numbers of substitutions from the present).

diffs	Bow.	NP	NA	SO
Bowhead	--	110	111	103
NP	110	--	29	29
NA	111	29	--	26
SO	103	29	26	--

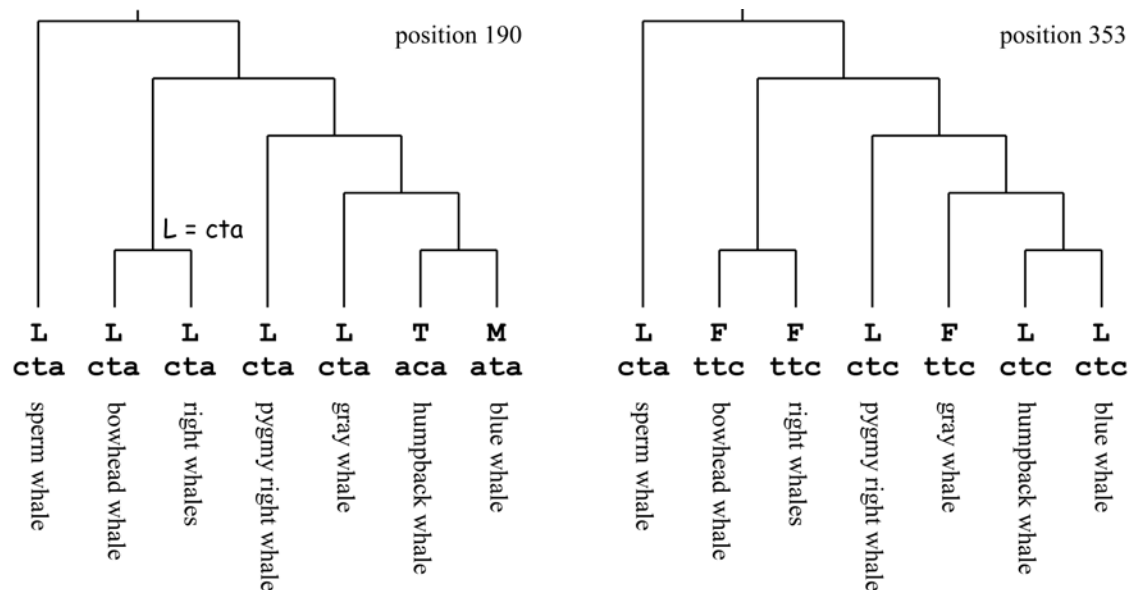
2. Fossil evidence suggests that the last common ancestor of the bowhead and the right whales lived about 16 million ( $1.6 \times 10^7$ ) years ago (MYA). About when did the last common ancestor of the three right whales live?

3. The tree on the right is a simplified “gene tree” for *population* samples of the mitochondrial control-region sequences of all three right-whale populations plus one bowhead, from a paper that challenges the two-species classification. The ticks on interior branches represent “diagnostic” mutations that occur in one or two of the populations but not in the others. In what ways does this tree agree with your analysis of *COI* and *cytb*, and in what ways does it disagree? Given all the evidence, do you think it is most reasonable to consider the world’s right whales as one, two, or three species? As we will discuss at some length later in the course, the criterion for species status is “lack of significant genetic exchange” with other species, for an “evolutionarily significant” amount of time. The arguments among experts arise mainly from differing opinions about what levels of gene flow and amounts of time are “significant”. You are free to argue any position as long as you refer to *relevant facts* and state them correctly!



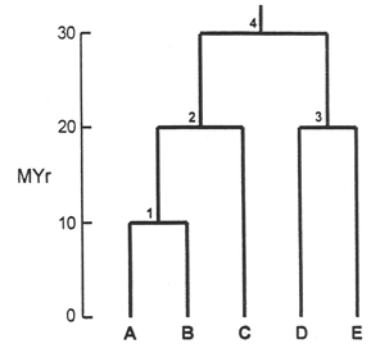
4. There are no amino-acid differences between the *COI* and *cytb* proteins of right whales, but there are many differences between right whales, the bowhead, and more distantly related baleen whales. Two variable amino-acid positions in the *cytb* gene and protein are shown below, at the tips of a phylogenetic tree relating six baleen whales and the sperm whale. Using the principle of parsimony (under which the *simplest hypothesis is best*), reconstruct the evolution of these two positions both at the *nucleotide* level and at the *amino-acid* level. Show both the inferred *codon* and the inferred *amino acid* at each internal node of the tree, including the root.

Some nodes may be ambiguous; if so, indicate the alternative possibilities. (One node on the tree for position 190 is done for you, to show what you need to do for the other nodes.) There is one place where your reconstruction at the DNA level resolves what would otherwise be an ambiguity at the protein level; please identify that place!



## Problems involving a gene family

5. Odorant receptors form the largest known family of proteins and genes, with over 1000 members in rodents and several to many hundreds in most other vertebrates. Assume that this hypothetical tree represents the *actual* phylogenetic history of five odorant-receptor genes that can be found in some modern rodent species. And assume that the *mutation rate* for nucleotide substitutions is  $5 \times 10^{-9}$  substitutions/site/year (which is near the mean for mammals).



(a) Fill in the matrix of expected pairwise synonymous distances ( $K_s$  = synonymous substitutions per synonymous site) for the five sequences. We'll discuss the distinction between synonymous and nonsynonymous nucleotide substitutions later in the course. For now, all you need to know is that synonymous mutations do not change the amino-acid sequence of a protein, so for all practical purposes they are not subject to selection and should therefore accumulate at approximately the rate of mutation, along any given line of descent. Remember that a distance matrix is *symmetrical* about its main diagonal; in other words,  $K_s(A-B) = K_s(B-A)$ , so you only need to fill in the 10 cells above or below the diagonal.

(b) The five sequences (A-E) occur in two species: A, B, and D are from Species X, while C and E are from Species Y. The four internal nodes in the tree therefore represent two kinds of events: one or more of the nodes correspond(s) to the *speciation* event that separated the populations ancestral to species X and Y, and one or more of the nodes correspond to *gene duplication* events that gave rise to different (paralogous) members of the gene family. Which of the four nodes are of each kind?

	A	B	C	D	E
A	0				
B		0			
C			0		
D				0	
E					0

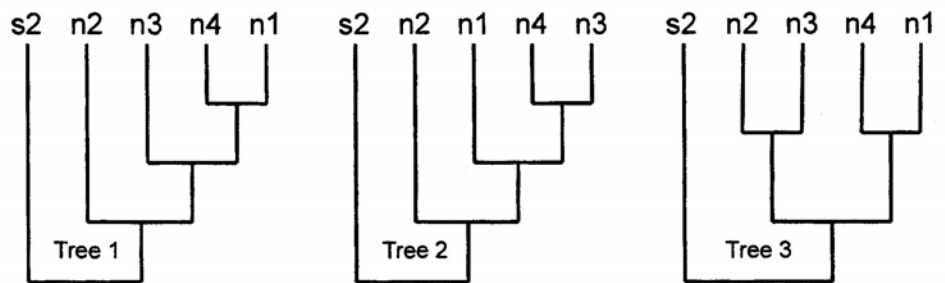
## More fun with Neandertals!

6. Several *Homo sapiens neanderthalensis* HV1 sequences have been determined in the last few years. For detail and discussion, you might want to look at Schmitz *et al.* (*PNAS* 99:13342-13347, 1 October 2002). You can download the PDF from the journal's web site ([www.pnas.org](http://www.pnas.org)) from any computer on campus, or go through Marriott Library if you're off campus. Two new sequences are more similar to the first one (nea1) from the "type" specimen discovered in 1856, than they are to the second one (nea2) from Mezmaiskaya cave in the Caucasus. Here are the variable positions from an alignment of four Neandertal sequences and a modern sequence (sap2) from the lecture slides, and also the matrix of pairwise differences among these five sequences.

### The variable positions

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
nea1	g	t	c	t	t	g	c	g	t	a	c	c	c	t	t	a	a	g	t	t	c	g	t	c	c
nea2	a	c	t	t	t	g	t	a	t	c	c	c	c	t	t	a	a	a	t	t	c	g	t	t	c
nea3	g	t	t	t	t	g	c	g	t	c	c	c	c	t	t	a	a	a	t	t	c	g	t	c	c
nea4	g	t	t	t	t	g	c	g	t	c	c	c	c	t	t	a	a	g	t	t	c	g	t	c	c
sap2	a	t	t	a	c	a	t	g	c	a	a	t	t	c	c	g	c	a	c	c	t	a	c	c	t

	n1	n2	n3	n4	s2
nea1	--	8	3	2	21
nea2	8	--	5	6	21
nea3	3	5	--	1	20
nea4	2	6	1	--	21
sap2	21	21	20	21	--



(a) Using the UPGMA method, infer a phylogeny for these five HV1 sequences.

(b) Only *four* of the 25 variable nucleotide positions are "informative" for the *parsimony* method, in the sense that they will favor some trees over others. Which are they? Hint: there must be at least two of one nucleotide and two of another.

(c) Using *just these four positions*, evaluate the three alternative trees, above. How many substitutions does each tree require, at a minimum? Hint: one requires 5 substitutions ("steps" in the jargon of parsimony), one requires 6 steps, and one requires 7 steps. Does the *most parsimonious* tree have the same topology as the UPGMA tree?

(d) Where are your reconstructions unambiguous, and where are there alternative, equally parsimonious ways to reconstruct the changes?

(e) Put *all* the substitutions (including the "uninformative" ones) on the most parsimonious tree.