* **Readings:**

Chapter 6 in Page and Holmes (1998). Molecular Evolution: A Phylogenetic Approach

- especially pages 172-193 for midterm
- scan the sections on quartets

Optional: Chapter 11 in Felsenstein (2004)

Optional: Strimmer and Haeseler (2009): “Genetic Distances and Nucleotide Substitution Models”

* outline with 5 references due by end of day
* midterm next Monday

Inferring a phylogeny is an estimation procedure
It can be done as a one- or two-step process

(1) chose an *algorithm* that uses data to generate a tree

(2) use an *optimality criterion* to chose among best trees, e.g.,

parsimony: shortest

likelihood and Bayesian methods: most likely, etc.

minimum evolution: shortest (distance) tree

* *algorithmic* procedures combine these steps into a single process
  e.g., UPGMA
* *optimality criterion*-based methods consider suboptimal trees or alternative solutions
  e.g., parsimony, Bayesian inference, minimum evolution

Distance Methods

* distance methods are algorithmic methods that replace character data with pairwise distances i.e., they use a (taxon x taxon matrix)

* pairwise distances are then used to infer (1) topology and (2) estimate branch lengths

* if observed distances reflect all evolutionary changes and rates of evolution are constant across lineages no problem…distance methods do fine, they have the advantages of being simple and fast

Stumbling blocks for distance methods

1) multiple hits: more than a single mutation/nucleotide substitution at a single site along a nucleotide sequence
   - distance methods underestimate numbers of changes
   - all methods challenged by multiple hits, but distance methods are especially so because of lost information that occurs when converting from taxon x character matrix to taxon-x-taxon distance matrix.

Hidden multiple hits

	actual changes	observ. changes
Multiple hit: Back mutation		
ACACCTCGTA → ACTCCTCGTA → ACACCTCGTA	**2**	**0**
Multiple hit: Unseen change		
ACACCTCGTA → ACGCCTCGTA → ACCCCTCGTA	**2**	**1**
Convergence		
Taxon A ACCCCTCGTA → ACGCCTCGTA	**2**	**0**
Taxon B ACACCTCGTA → ACGCCTCGTA		

Stumbling blocks for distance methods

1) multiple hits: more than a single mutation/nucleotide substitution at a single site along a nucleotide sequence
   - distance methods underestimate numbers of changes
   - all methods challenged by multiple hits, but distance methods are especially so because of lost information that occurs when converting from taxon x character matrix to taxon-x-taxon distance matrix.
2) unequal rates across lineages: often slowly evolving lineages may appear similar relative to derived taxa
   - particularly a problem for UPGMA (ultrametric method)
3) can stumble on small (e.g., morphological) data sets
   - perform better with more taxa and/or more characters

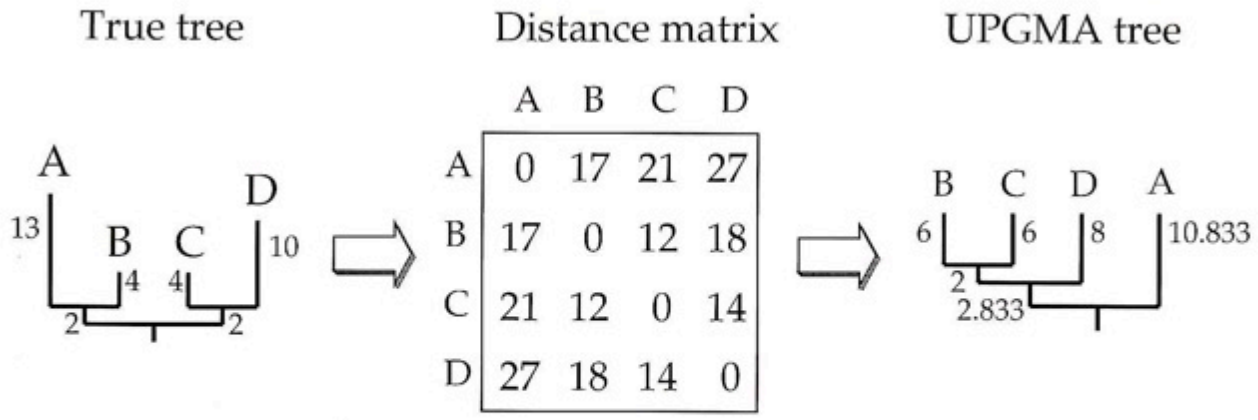


Figure 11.7: A four-species, nonclocklike tree and the expected data matrix it yields, when distances are the sums of branch lengths. The tree estimated by applying the UPGMA method to this distance matrix is shown—it does not have the correct tree topology. In both trees the branch lengths are proportional to the vertical length of the branches.

from Felsenstein 2004. Inferring Phylogenies

UPGMA: unweighted pair group method using arithmetic averages

* build distance matrix
  - values in matrix are simple pairwise distances
* connect closest two taxa, join to tree at midpoint
* calculate new distance matrix using the mean value for group as new distance
* repeat iteratively until all taxa are attached

Immunolgical distances for pinnapeds (log transformed)

	dog	bear	raccoon	weasel	**seal**	**sea lion**	cat	monkey
dog	0	32	48	51	**50**	**48**	98	148
bear	32	0	26	34	**29**	**33**	84	136
raccoon	48	26	0	42	**44**	**44**	92	152
weasel	51	34	42	0	**44**	**38**	86	142
* **seal**	**50**	**29**	**44**	**44**	**0**	**24**	**89**	**142**
* **sea lion**	**48**	**33**	**44**	**38**	**24**	**0**	**90**	**142**
cat	98	84	92	86	**89**	**90**	0	148
monkey	148	136	152	142	**142**	**142**	148	0

from Felsenstein. 2004. Inferring Phylogenies.

		dog	* **bear**	* **raccoon**	weasel	SS	cat	monkey
	dog	0	**32**	**48**	51	49	98	148
*	**bear**	**32**	**0**	**26**	**34**	**31**	**84**	**136**
*	**raccoon**	**48**	**26**	**0**	**42**	**44**	**92**	**152**
	weasel	51	**34**	**42**	0	41	86	142
	SS	49	**31**	**44**	41	0	89.5	142
	cat	98	**84**	**92**	86	89.5	0	148
	monkey	148	**136**	**152**	142	142	148	0

		dog	* **BR**	weasel	* **SS**	cat	monkey
	dog	0	**40**	51	**49**	98	148
*	**BR**	**40**	**0**	**38**	**37.5**	**88**	**144**
	weasel	51	**38**	0	**41**	86	142
*	**SS**	**49**	**37.5**	**41**	**0**	**89.5**	**142**
	cat	98	**88**	86	**89.5**	0	148
	monkey	148	**144**	142	**142**	148	0

from Felsenstein. 2004. Inferring Phylogenies.

		dog	* **BRSS**	* **weasel**	cat	monkey
	dog	0	**44.5**	**51**	98	148
*	**BRSS**	**44.5**	**0**	**39.5**	**88.75**	**143**
*	**weasel**	**51**	**39.5**	**0**	**86**	**142**
	cat	98	**88.75**	**86**	0	148
	monkey	148	**143**	**142**	148	0

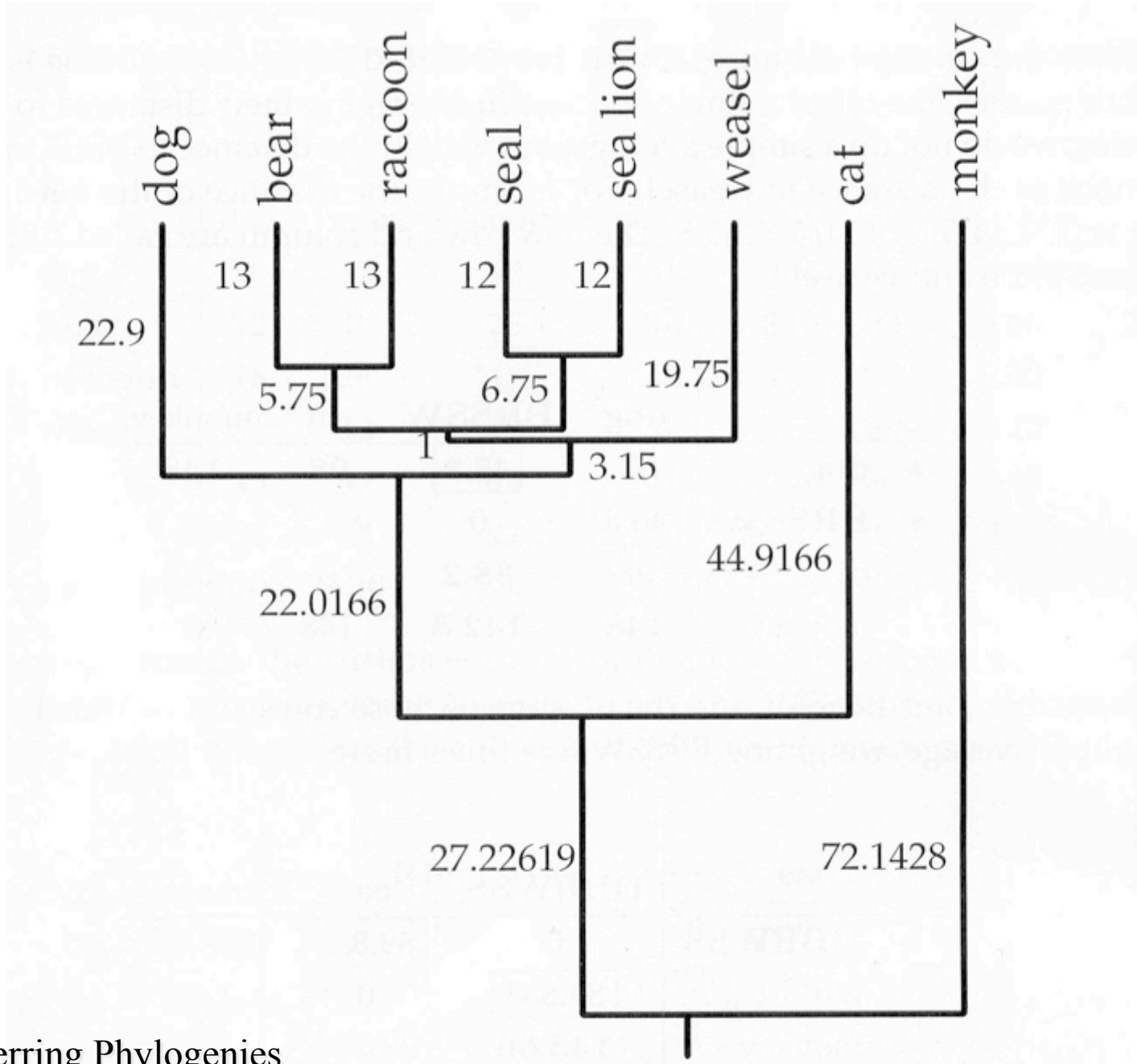
		* **dog**	* **BRSSW**	cat	monkey
*	**dog**	**0**	**45.8**	**98**	**148**
*	**BRSSW**	**45.8**	**0**	**88.2**	**142.8**
	cat	**98**	**88.2**	0	148
	monkey	**148**	**142.8**	148	0

from Felsenstein. 2004. Inferring Phylogenies.

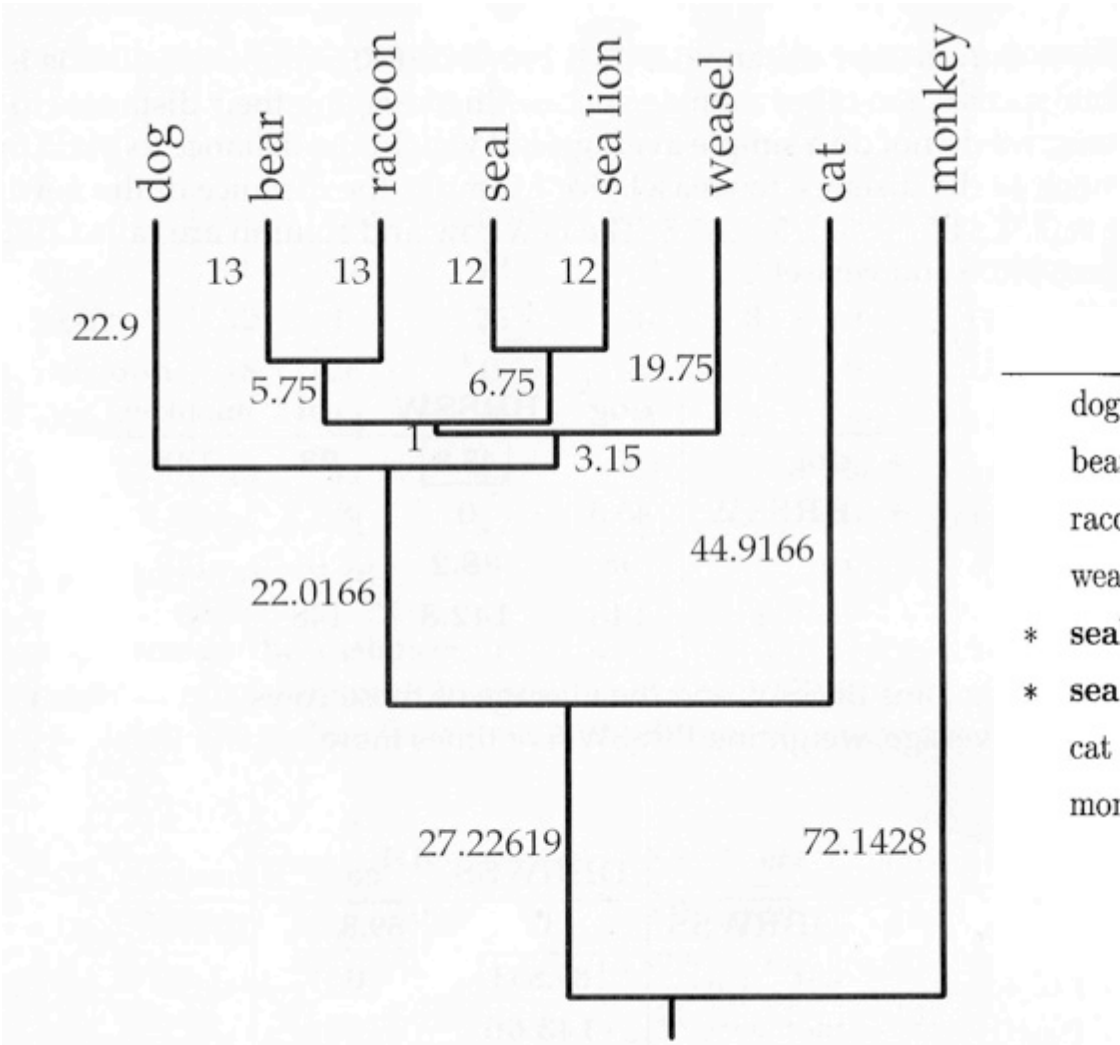
		* DBRWSS	* cat	monkey
*	**DBRWSS**	**0**	**89.833**	**143.66**
*	**cat**	**89.833**	**0**	**148**
	monkey	**143.66**	**148**	0

	DBRWSSC	monkey
DBRWSSC	0	144.2857
monkey	144.2857	0

from Felsenstein. 2004. Inferring Phylogenies.



from Felsenstein. 2004. Inferring Phylogenies



		dog	bear	raccoon	weasel	**seal**	**sea lion**	cat	monkey
	dog	0	32	48	51	**50**	**48**	98	148
	bear	32	0	26	34	**29**	**33**	84	136
	raccoon	48	26	0	42	**44**	**44**	92	152
	weasel	51	34	42	0	**44**	**38**	86	142
*	**seal**	**50**	**29**	**44**	**44**	**0**	**24**	**89**	**142**
*	**sea lion**	**48**	**33**	**44**	**38**	**24**	**0**	**90**	**142**
	cat	98	84	92	86	**89**	**90**	0	148
	monkey	148	136	152	142	**142**	**142**	148	0

from Felsenstein. 2004. Inferring Phylogenies.

* UPGMA proved to be a poor performer with morphological data
* UPGMA sometimes still used for
  a. gene frequency data, e.g., Nei's genetic distance
  b. DNA-DNA hybridization (raw data is a distance)
  c. immunological distances (raw data is a distance)

Its downfall is the assumption of ultrametricity
- in real life, distances from root to tip are rarely the same across a tree
- extreme example: alpha globin gene in primates (Shaw et al. 989):
  baboon and rhesus - differ by 9 nucleotide sites
  baboon and human - differ by 11 sites
  human and rhesus - differ by only 5 sites
- fast rate of alpha globin evolution in baboon throws off UPGMA

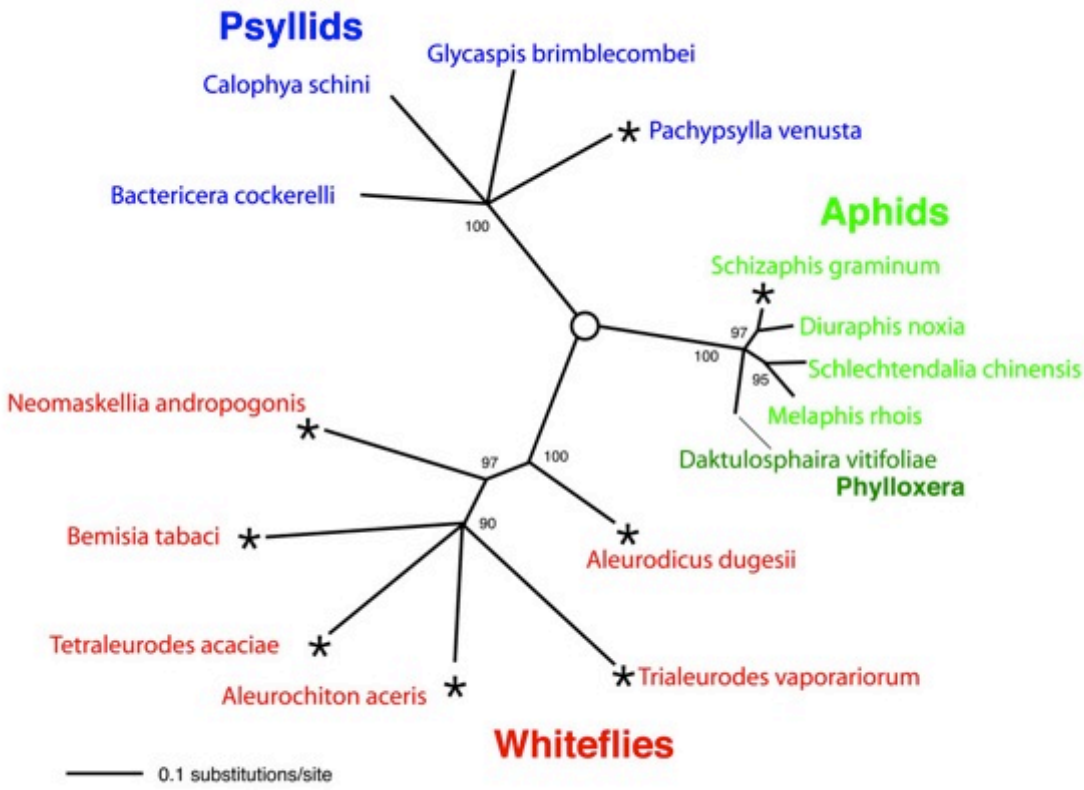
Additive Distance Methods

* several rate-insensitive distance methods have been developed for phylogenetic inference
* where branch lengths on trees are estimates of actual amount of evolutionary change
* different rates across lineages may obtain (ultrametricity not enforced)
* sum of the branch lengths between taxa estimates the (*corrected*) observed distance between taxa (and hence are ***additive***)
* the problem (as it is for all distance methods) is that of multiple hits or unobserved nucleotide changes (DNA substitutions)

Additive Distance Methods

* three common methods:
  - goodness of fit methods, such as least squares
  - minimum evolution
  - neighbor joining

* all produce unrooted trees (that can be rooted afterwards by pulling down node between ingroup and outgroup)



Goodness of fit measures: Least Squares

$$F_{\alpha} = \mathbf{\Sigma}_{1 \leq i < j <= n} |d_{ij} - p_{ij}|^{\alpha}$$

Where:

n = number of taxa

d_{ij} = observed distances

p_{ij} = path distances on tree

α = usually 1 or 2

if $\alpha=1$ then minimizes absolute differences between estimated lengths and path lengths

if $\alpha=2$ then = least squares

minimizes squares of differences between estimated lengths and path lengths

most common method

Optimality Criterion: chose tree that minimizes Σ or residual sum of squares

Table 6.1 Kimura 2-parameter distances between hominoid sequences (above diagonal) and tree distances obtained by least squares (below the diagonal) for the tree shown in Fig. 6.7. Tree distances larger than the observed distances are shown in bold, tree distances smaller than the observed are shown in italics.

	Human	Chimp	Gorilla	Orang-utan	Gibbon
Human	–	0.09190	0.1083	0.1790	0.2057
Chimp	0.0919	–	0.1134	0.1940	0.2168
Gorilla	*0.1068*	**0.1151**	–	0.1882	0.2170
Orang-utan	**0.1816**	*0.1898*	**0.1893**	–	0.2172
Gibbon	**0.2078**	*0.2160*	*0.2155*	0.2172	–

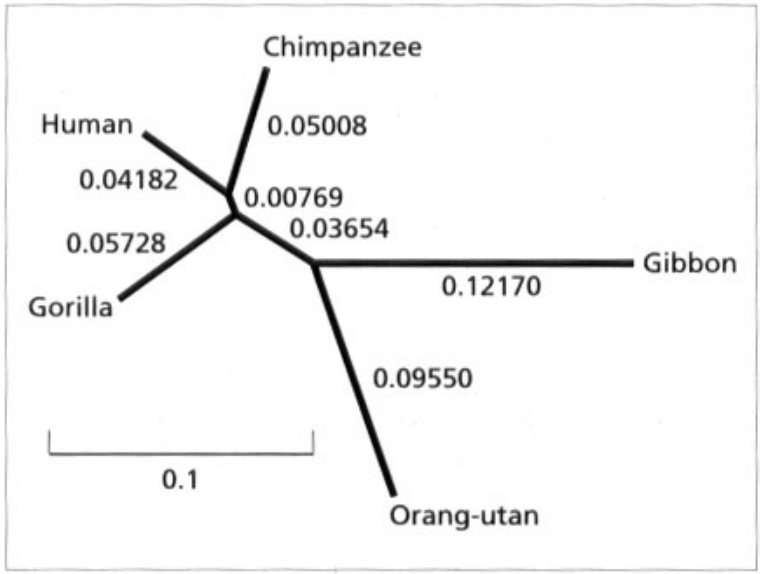


Fig. 6.7 Additive tree for hominoid mtDNA sequences showing branch lengths computed using least squares. The pairwise tree distances for this tree are given in the lower left triangle of Table 6.1.

from Page and Holmes. 1998. Molecular Evolution: A Phylogenetic Approach.

Minimum Evolution [Optional]

$$L = \sum_{i=1}^{2n-3} e_i$$

n = number of taxa

e_i = each path length (branch)

optimality criterion = chose sum of the paths lengths that minimizes the sum of the squared deviation between the estimated and fitted branch lengths

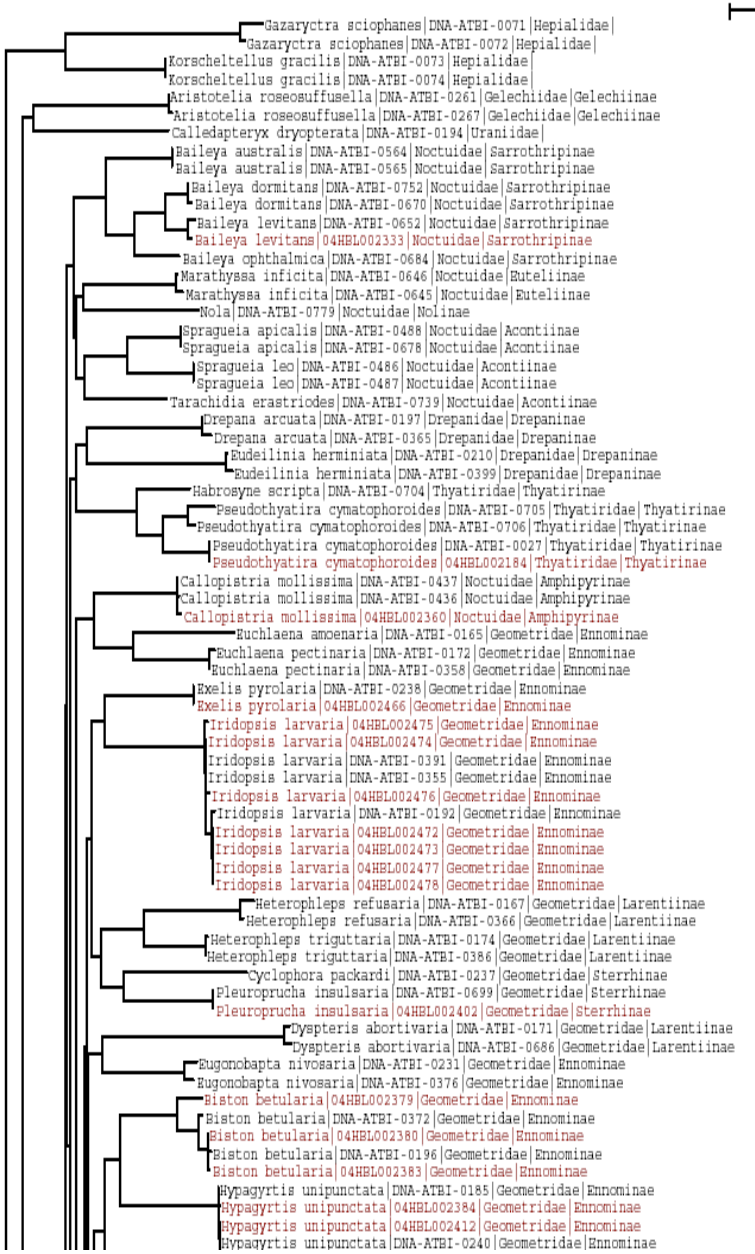
The shortest tree….

Neighbor-joining (Saitou and Nei 1987)

* a heuristic for approximating minimum evolution tree (shortest set of paths connecting all taxa)
* two steps

1) construct topology with algorithm
2) adjust branch lengths (by least squares)

* works with taxa/nodes
* calculates a taxon's/node's distance relative to all other taxa/nodes in analysis
* rather rate insensitive (meaning accommodates unequal branches)
* much better performer than UPGMA



Neighbor-joining (Saitou and Nei 1987)

The Method

Cycle 1

1) begin with a star tree and build distance matrix using (corrected) distances
2) construct a second matrix with each taxon's net distance from all other taxa (a rate-adjusting measure) that helps correct the unequal rates problem
3) connect two taxa that yield shortest tree length (most similar neighbors)
4) calculate (additive) branch lengths to these two neighbors and then prune the neighbors from the tree and use their new common node

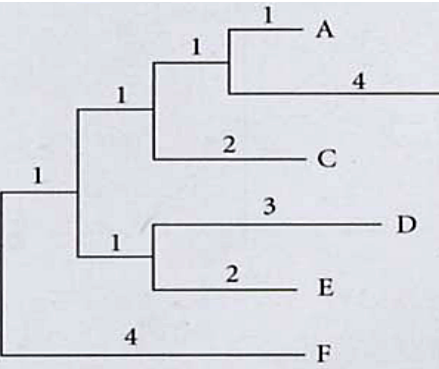
Cycle Two

5) using new common node, recalculate distances to remaining taxa
6) generate a second matrix with each taxon's net distance new (rate-adjusted) distances
7) connect two neighbors that yield shortest tree length (most similar neighbors)
8) calculate (additive) branch lengths to the two closest neighbors and then prune neighbors from the tree and use their new common node

Cycle Three+

9) repeat 5-8 until all taxa are connected

Neighbor-joining (Saitou and Nei 1987)

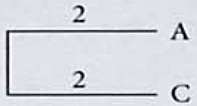


Hypothetical tree topology: since the divergence of sequences A and B, B has accumulated 4 times as many mutations as sequence A.

Assume the following matrix of pairwise evolutionary distances:

	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

Clustering methods (discussed in Box 5.1) would erroneously group sequences A and C because they assume clock-like behavior. Although sequences A and C look more similar, sequences A and B are more closely related.



1. Compute the net divergence r for every end node (N = 6)

$r_A = 5+4+7+6+8 = 30$ $\quad r_D = 38$
$r_B = 5+7+10+9+11 = 42$ $\quad r_E = 34$
$r_C = 32$ $\quad r_F = 44$

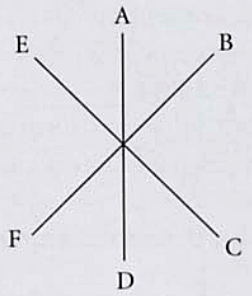
2. Create a rate-corrected distance matrix; the elements are defined by $M_i = d_{ij} - (r_i + r_j)/(N-2)$

$M_{AB} = d_{AB} - (r_A + r_B)/(N-2) = 5 - (30+42)/4 = -13$

N= taxa in analysis

$M_{AC} = \ldots$

	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



3. Define a new node that groups OTUs i and j for which M_i is minimal. For example, sequences A and B are neighbors and form a new node U (but, alternatively, OTUs D and E could have been joined; see below).

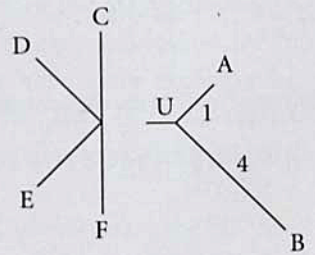
4. Compute the branch lengths from node U to A and B:

$S_{AU} = d_{AB}/2 + (r_A - r_B)/2(N-2) = 1$
$S_{BU} = d_{AB} - S_{AU} = 4$

5. Compute new distances from node U to every other terminal node:

$d_{CU} = (d_{AC} + d_{BC} - d_{AB})/2 = 3$
$d_{DU} = (d_{AD} + d_{BD} - d_{AB})/2 = 6$
$d_{EU} = (d_{AE} + d_{BE} - d_{AB})/2 = 5$
$d_{FU} = (d_{AF} + d_{BF} - d_{AB})/2 = 7$

	U	C	D	E
C	3			
D	6	7		
E	5	6	5	
F	7	8	9	8



6. N = N − 1; repeat Steps 1 through 5.

Van de Peer, Y. 2003. Phylogeny Inference Based on Distance Methods. *In* The Phylogenetic Handbook, M. Salemi and A.-M. Vandamme (eds).

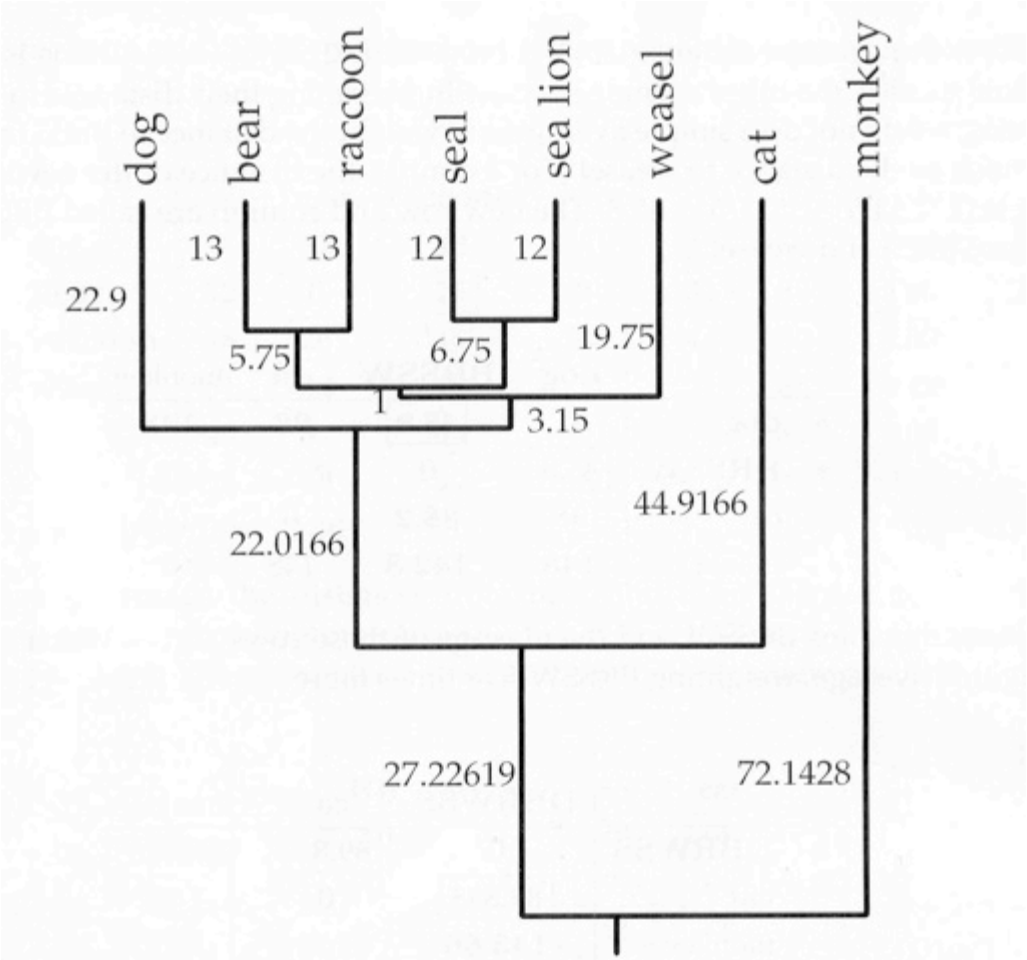


Figure 11.6: UPGMA tree

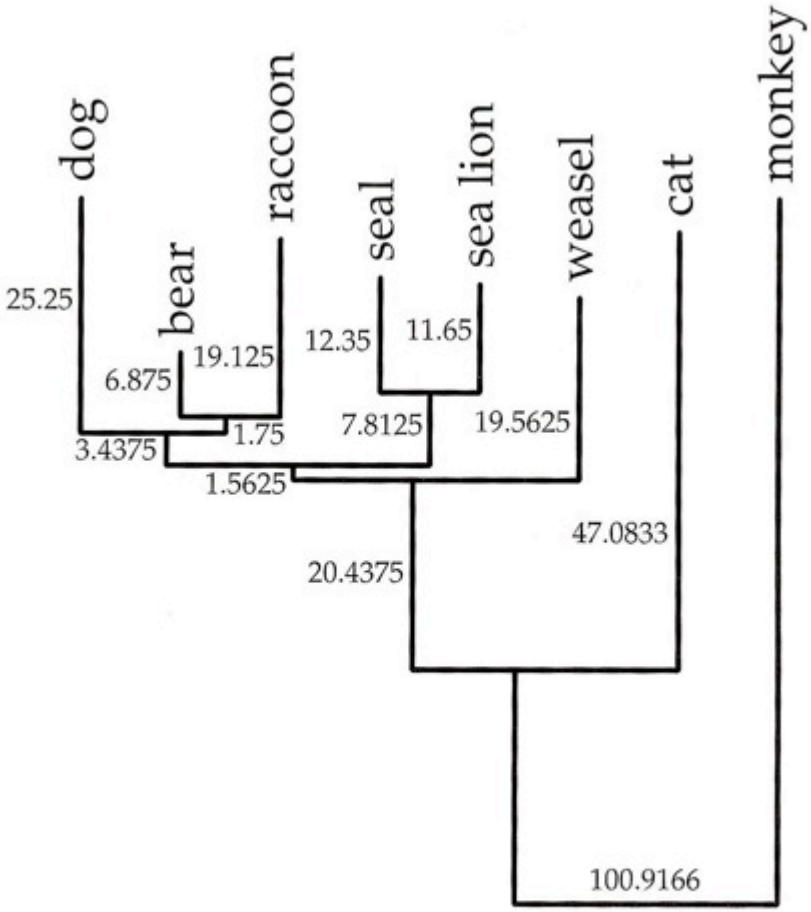


Figure 11.8: The neighbor-joining tree for the data set of Sarich (1969) rooted at the midpoint of the longest path between species. It may be compared with Figure 11.6.

Felsenstein. 2004. Inferring Phylogenies.

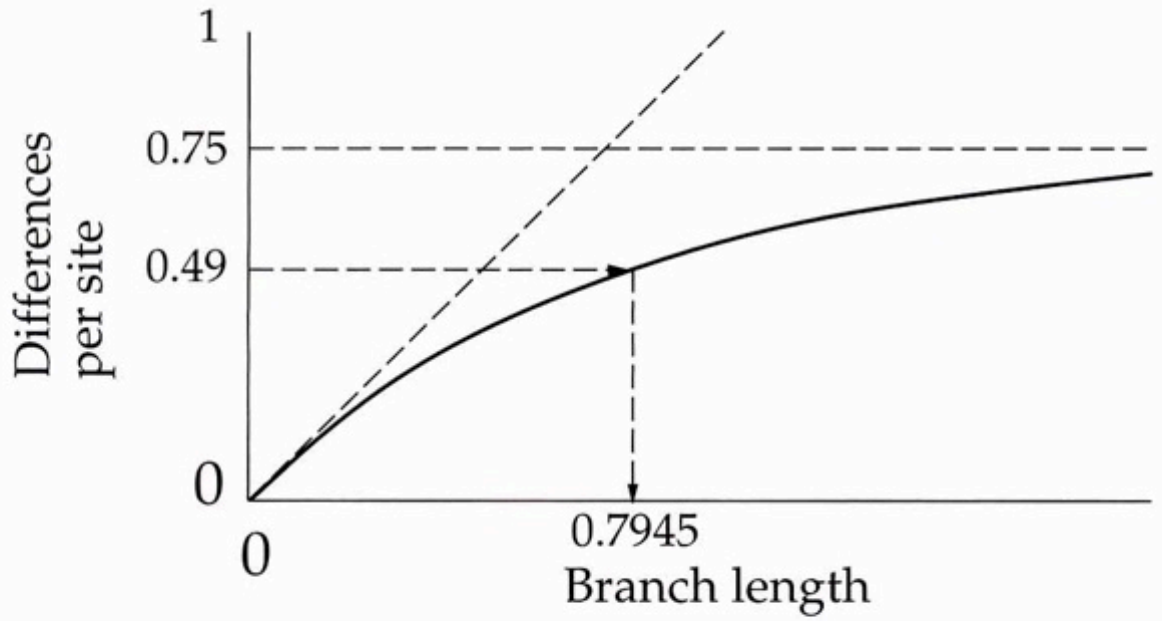
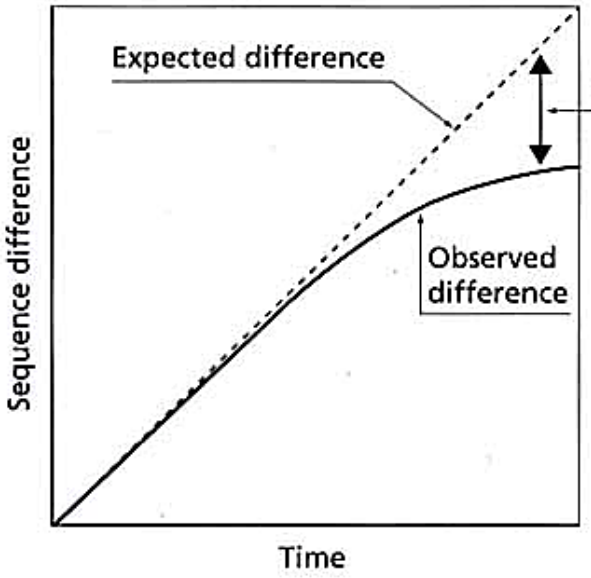


Figure 11.4: The expected difference per site between two sequences in the Jukes-Cantor model, as a function of branch length (the product of rate of change and time). The process of inferring the branch length from the fraction of sites that differ between two sequences is also shown.

from Felsenstein 2004. Inferring Phylogenetics.



shape of line varies with gene

with goal of generating branch lengths that more reflect number of expected substitutions

Fig. 5.12 The need to correct observed sequence differences. The extent of observed differences between two sequences is not linear with time (as we would expect if the rate of molecular evolution is approximately constant) but curvilinear due to multiple hits. The goal of distance correction methods is to recover the amount of evolutionary change that the multiple hits have overprinted and to 'correct' the distances for unobserved hits. In effect, the methods seek to 'straighten out' the line representing observed differences.

from Page and Holmes. 1998. Molecular Evolution: A Phylogenetic Approach.

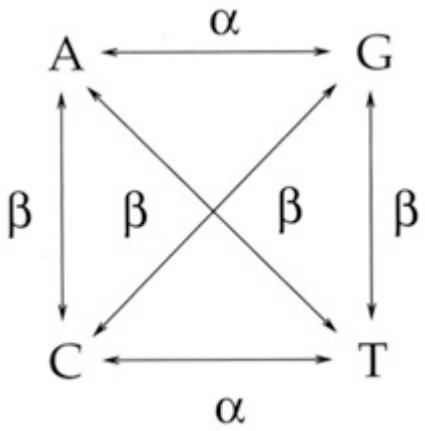
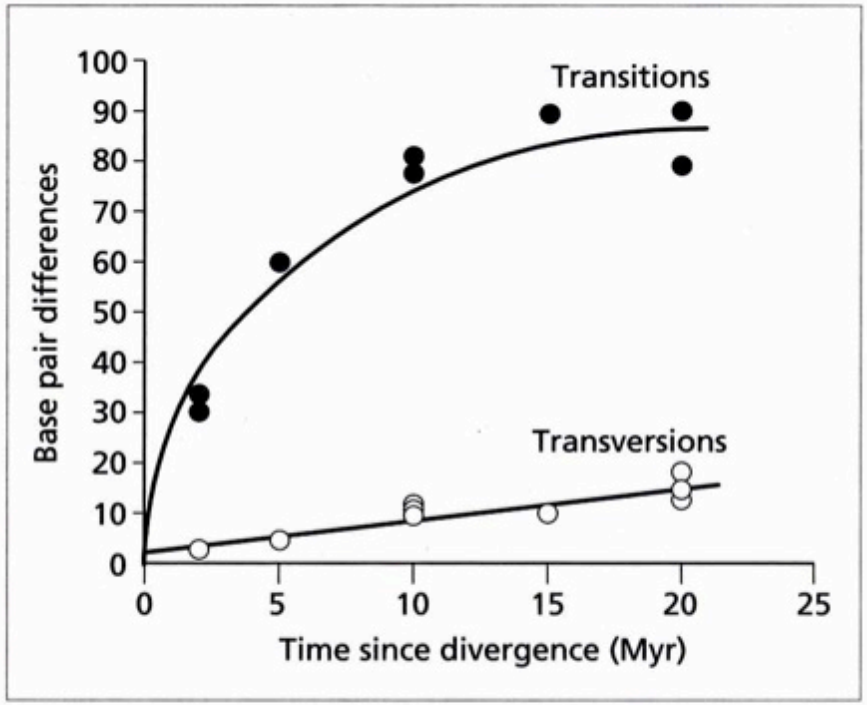
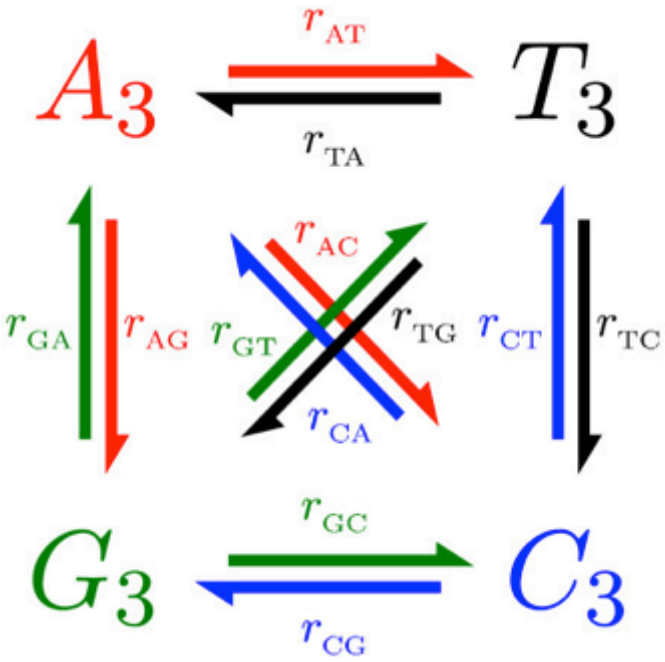


Fig. 5.13 The number of transitions and transversions between the same bovid mammal sequences used in Fig. 5.11. Transitions accumulate much more rapidly than transversions and become saturated, whereas transversions accumulate more slowly and show no evidence of saturation.



from Page and Holmes. 1998. Molecular Evolution: A Phylogenetic Approach.

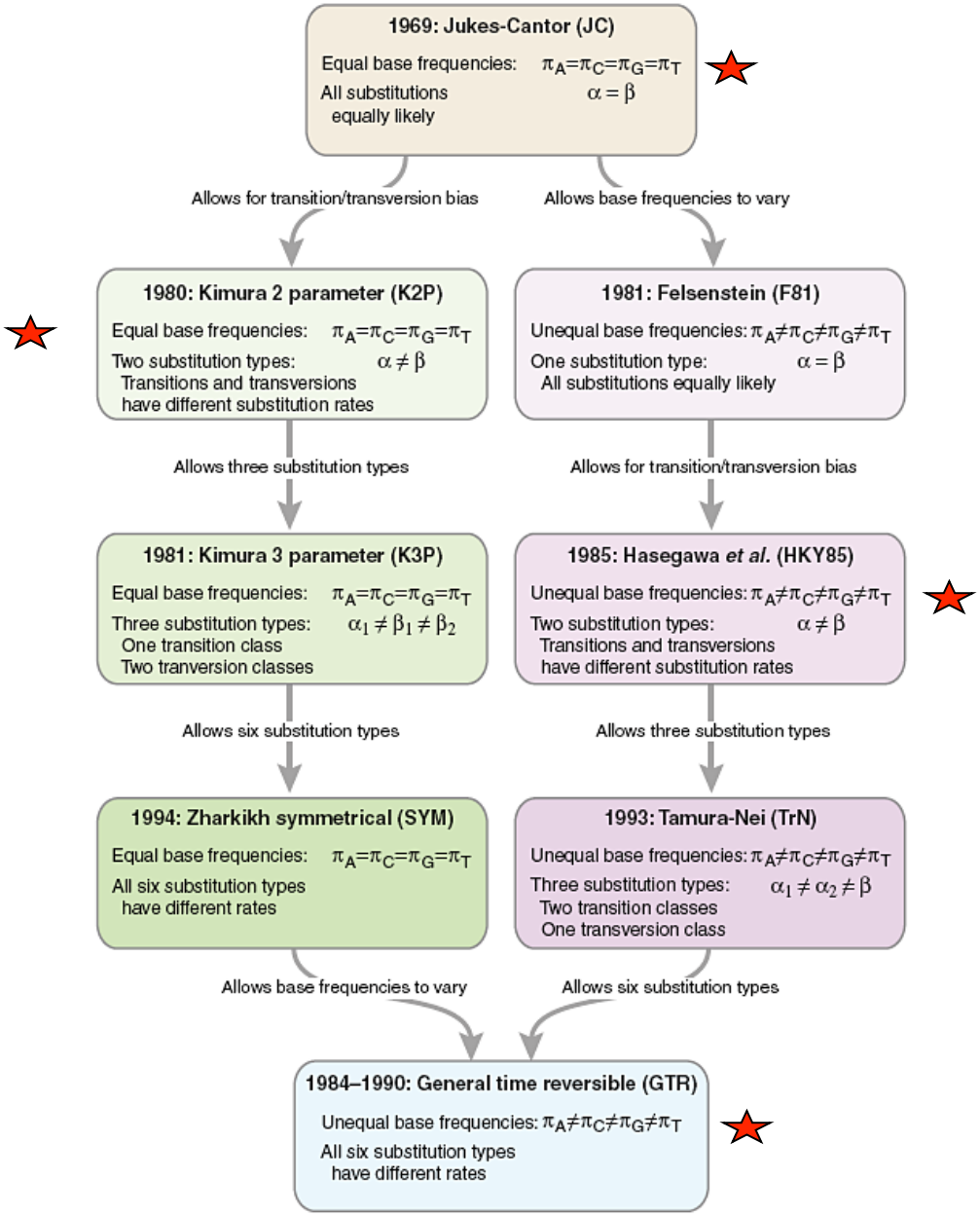


$$Q = \begin{array}{c} \\ \text{From base:} \end{array} \begin{array}{c} \text{To base:} \\ \begin{array}{cccc} \text{T} & \text{C} & \text{A} & \text{G} \end{array} \\ \begin{array}{c} \text{T} \\ \text{C} \\ \text{A} \\ \text{G} \end{array} \begin{pmatrix} - & r_{TC} & r_{TA} & r_{TG} \\ r_{CT} & - & r_{CA} & r_{CG} \\ r_{AT} & r_{AC} & - & r_{AG} \\ r_{GT} & r_{GC} & r_{GA} & - \end{pmatrix} \end{array}$$

The nucleotide substitution rate matrix summarizes the instantaneous rate of change from each of the four nucleotides to each of the other four nucleotides (from: www.biomedcentral.com/content/figures/1471-21...)

Nucleotide Substitution Models

* JC: equal rates of substitution among all four bases
* F81: unequal base frequencies are taken into account
* K2P: different rates for transitions and transversions
* HKY85: different rates for transitions and transversions + unequal bases freq.
* GTR (general time reversible): unequal base freq. + six different substitution rates; most general = least simple
* Log-det: allows base freq. proportions to change across different portions of the tree



Simon et al. 2006. Annual Rev. Ecol. Evol. Syst. 2006. 37:545–79

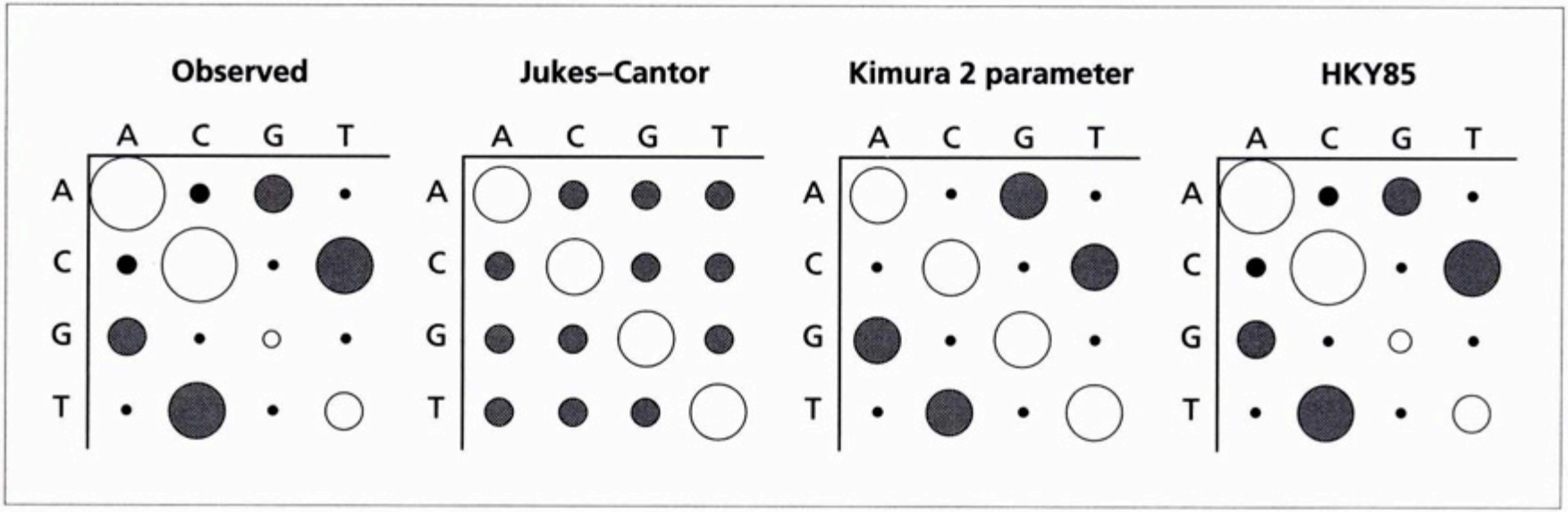


Fig. 5.15 Observed and expected numbers of nucleotide pairs between human and chimpanzee mtDNA sequences for three different models. As the models add parameters they more closely approximate the observed pattern. Data from Tamura (1994).

from Page and Holmes. 1998. Molecular Evolution: A Phylogenetic Approach.

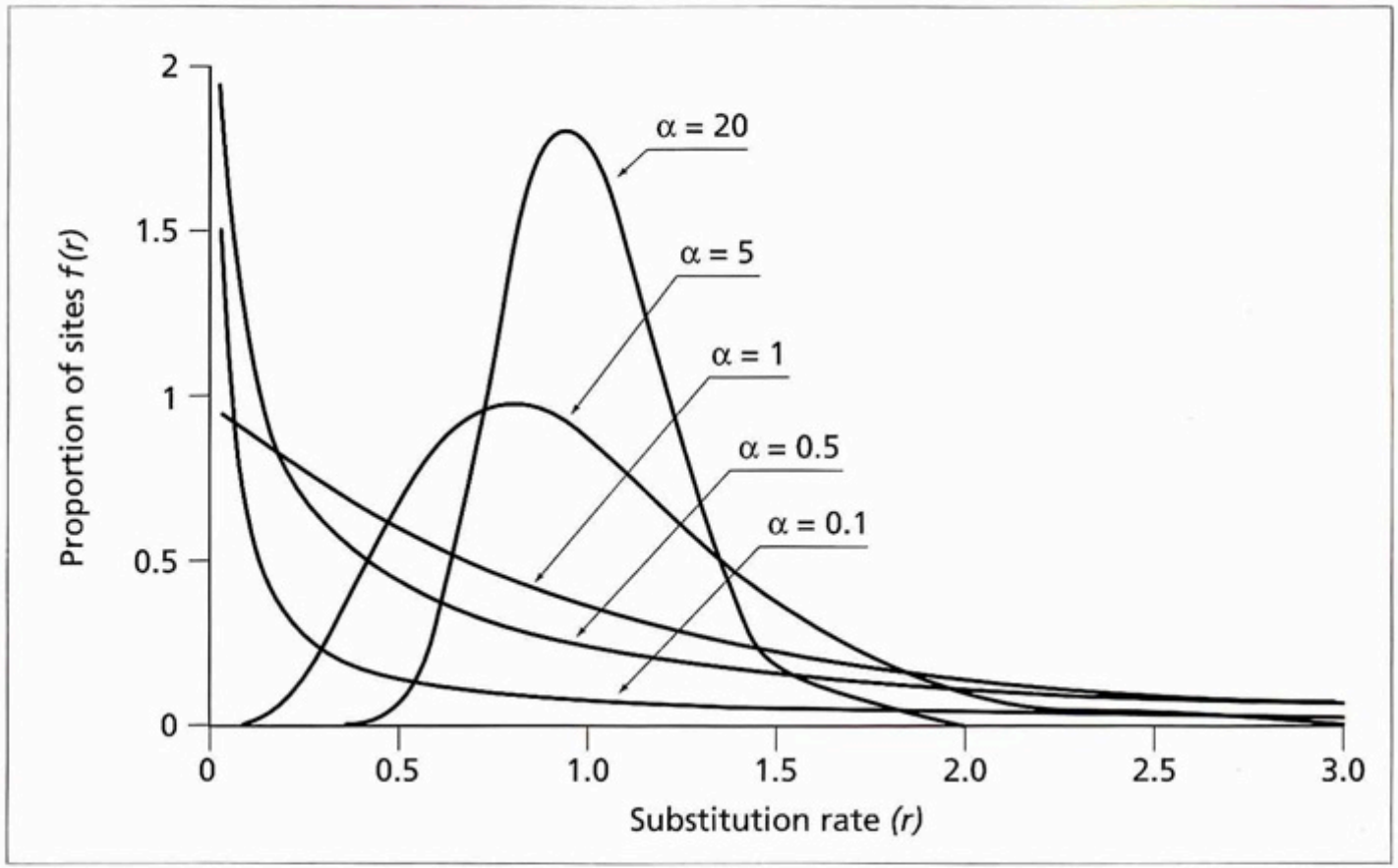


Fig. 5.20 The distribution of relative substitution rate *r* corresponding to different values of the gamma shape parameter α. Low α corresponds to large rate variation. As α gets larger the range of variation diminishes, until as α approaches ∞ all sites have the same substitution rate. After Yang (1996: Fig. 1).

from Page and Holmes. 1998. Molecular Evolution: A Phylogenetic Approach.

Type of sequences	α
Nuclear genes	
Albumin genes	1.05
Insulin genes	0.40
c-myc genes	0.47
Prolactin genes	1.37
16S-like rRNAs, stem region	0.29
16S-like rRNAs, loop region	0.58
$\psi\eta$-globin pseudogenes	0.66
Viral genes	
Hepatitis B virus genomes	0.26
Mitochondrial genes	
12S rRNAs	0.16
Position 1 of four genes	0.18
Position 2 of four genes	0.08
Position 3 of four genes	1.58
D-loop region	0.17
Cytochrome *b*	0.44

Advantages & disadvantages of distance methods

Advantages:

A. Computationally fast (especially if they don't have to consider alternative trees, e.g., in neighbor joining)

* speed determined principally by number of taxa

- e.g. number of nucleotides has little impact on neighbor joining

B. Use when a phylogeny but not necessary – just want a phylogenetic address, e.g., DNA barcoding

C. The only choice for some types of data which are inherently "phenetic"

DNA-DNA hybridization and immunological distances

Advantages & disadvantages of distance methods

Disadvantages:

A. Character data is lost
- subtleties of nature and distribution of character changes lost

B. Can't see where changes occur
- character-based methods plot changes on branches

C. Branch lengths have unclear meaning (in ultrametric trees)

D. Can't combine with character data

E. Generally do not evaluate suboptimal trees
- good to know nature of tree topologies that are almost as good

F. Distances can be asymmetrical (e.g., immunological distances)

G. Corrected distances are just guesses, apt to break down with deeper branches