Lecture 4. Distance methods. Models of DNA change.

Joe Felsenstein

Department of Genome Sciences and Department of Biology

Approximate variances for distances

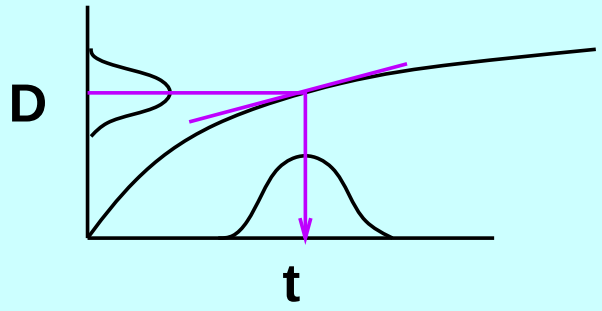
under the Jukes-Cantor model

Distance as a function of fraction of nucleotide differences is

$$\hat{t} = -\frac{3}{4}\ln\left(1 - \frac{4}{3}D\right)$$

The "delta method" approximates the variance of one as a function of the variance of the other:

$$\mathrm{Var}(\hat{t}) \simeq \left(\frac{\partial \hat{t}}{\partial D}\right)^2 \mathrm{Var}(D)$$



Approximate variances, continued

The variance of fraction of nucleotide difference with n sites is the binomial variance

$$\mathrm{Var}(\mathsf{D}) \;=\; \mathsf{D}\,(1-\mathsf{D})/\mathsf{n}$$

and since

$$\frac{\partial \hat{\mathsf{t}}}{\partial \mathsf{D}} \;=\; \frac{1}{1-\frac{4}{3}\mathsf{D}}$$

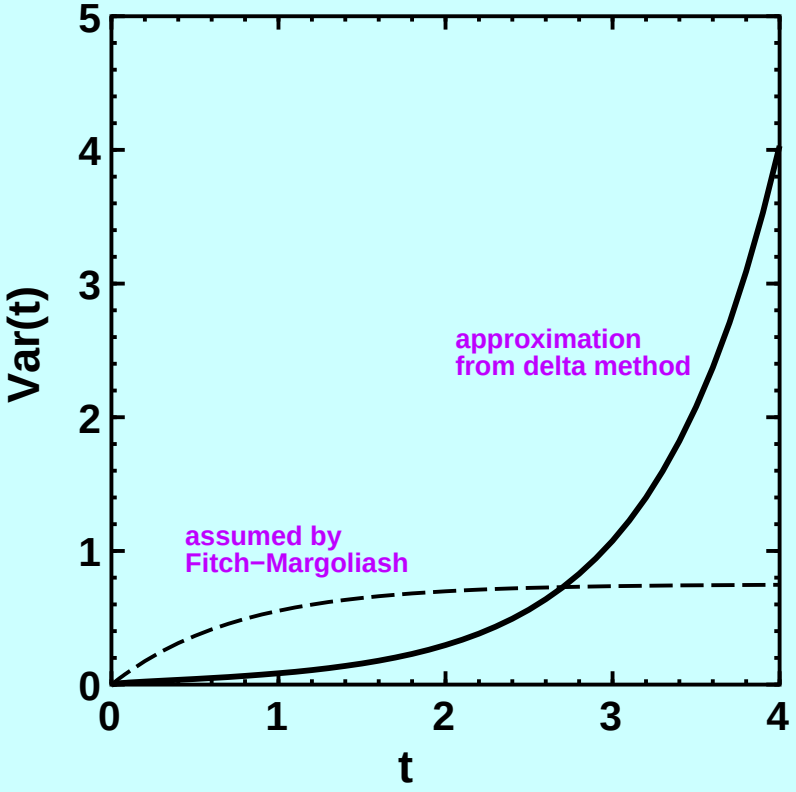
we get

$$\mathrm{Var}(\hat{\mathsf{t}}) \;\simeq\; \frac{\mathsf{D}(1-\mathsf{D})/\mathsf{n}}{\left(1-\frac{4}{3}\mathsf{D}\right)^2}$$

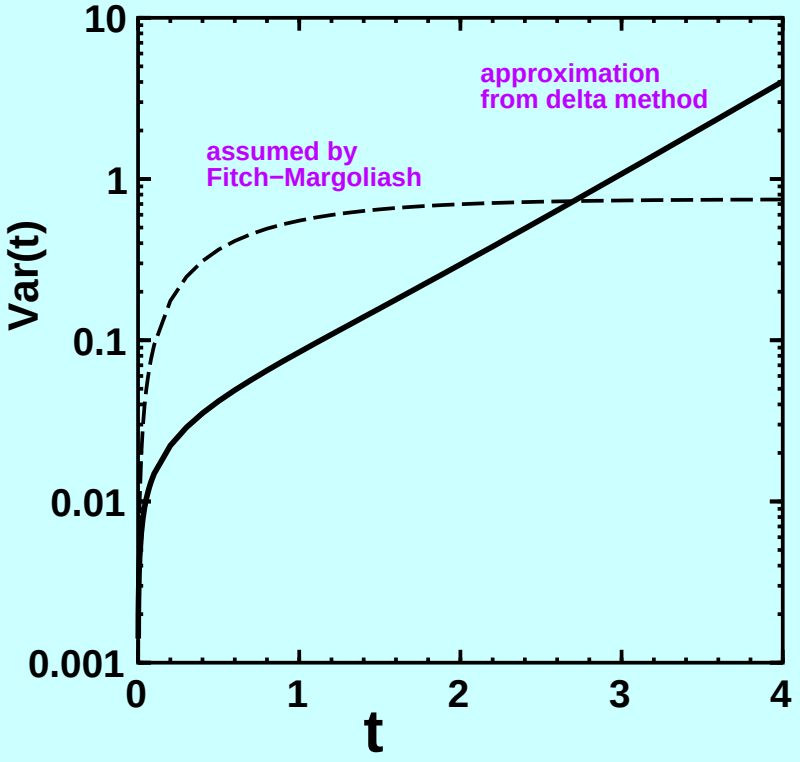
Variance of distance

as it increases with distance (given the JC model)

Variance on regular scale



Variance on log scale



The UPGMA algorithm

1. Choose the smallest of the D_{ij}
2. make a new "tip" (ij)
3. Have i and j connected to this new tip, by a node whose "time" ago in branch length units is $D_{ij}/2$.
4. Have the weight of the new tip be $w_{(ij)} = w_i + w_j$
5. For each other tip, aside from i and j, compute

$$D_{(ij),k} = D_{k,(ij)} = \frac{w_i D_{ik} + w_j D_{jk}}{w_i + w_j}$$

6. Delete the rows and columns of the D matrix for i and j.
7. If only one row left, stop, else return to step 1.

This can be done in $O(n^2)$ time if you save minimum elements of each row.

Sarich's (1969) immunological distances

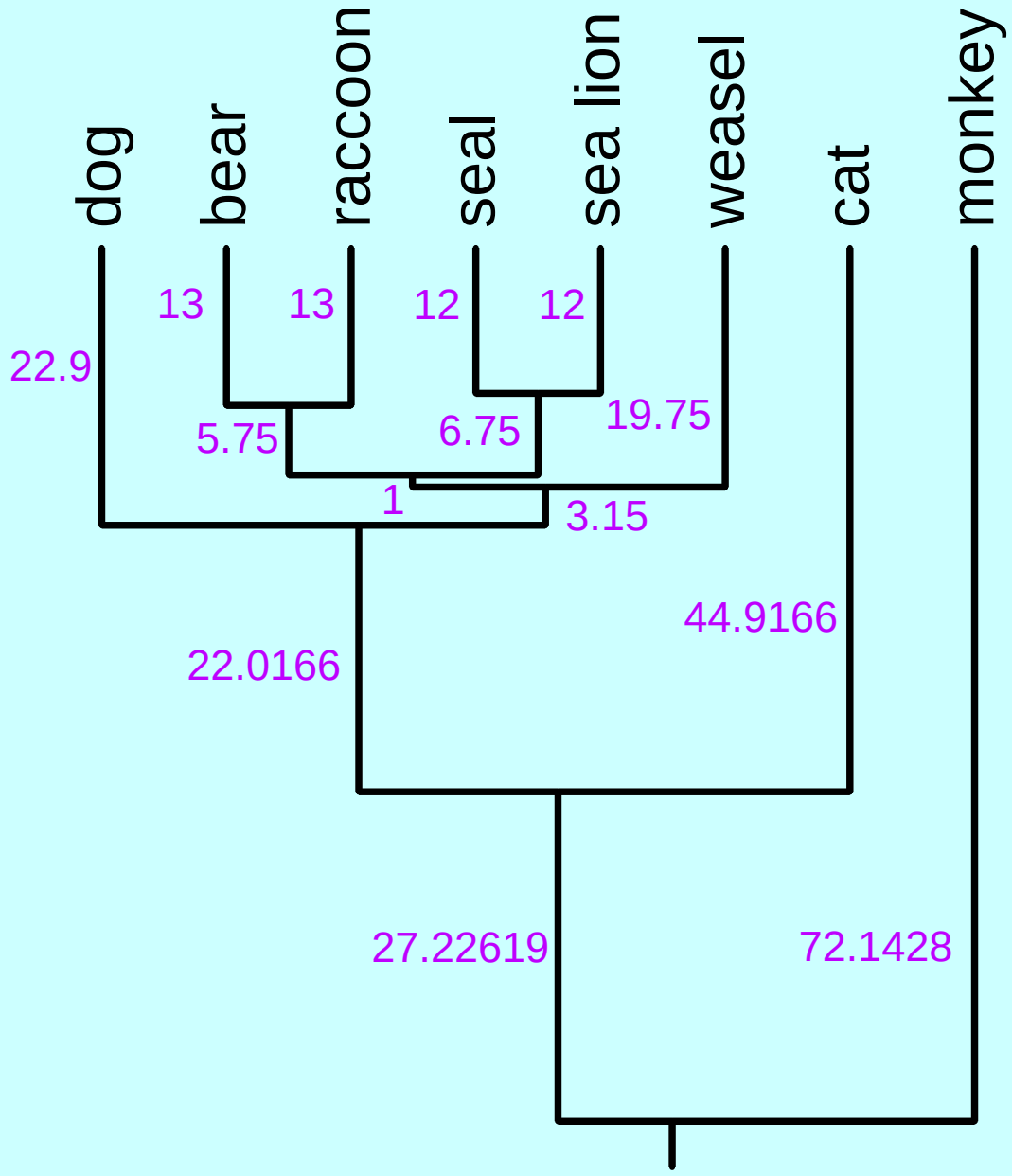
	dog	bear	raccoon	weasel	seal	sea lion	cat	monkey
dog	0	32	48	51	50	48	98	148
bear	32	0	26	34	29	33	84	136
raccoon	48	26	0	42	44	44	92	152
weasel	51	34	42	0	44	38	86	142
seal	50	29	44	44	0	24	89	142
sea lion	48	33	44	38	24	0	90	142
cat	98	84	92	86	89	90	0	148
monkey	148	136	152	142	142	142	148	0

Sarich's (1969) immunological distances

with the columns and rows corresponding to the smallest distance highlighted and a box shown for the smallest distance.

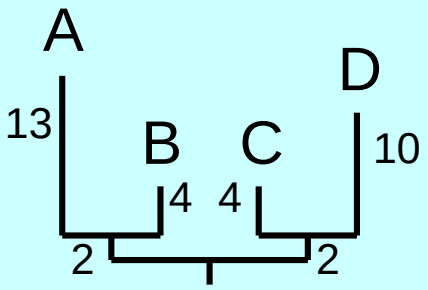
	dog	bear	raccoon	weasel	**seal**	**sea lion**	cat	monkey
dog	0	32	48	51	**50**	**48**	98	148
bear	32	0	26	34	**29**	**33**	84	136
raccoon	48	26	0	42	**44**	**44**	92	152
weasel	51	34	42	0	**44**	**38**	86	142
seal	**50**	**29**	**44**	**44**	**0**	**24**	**89**	**142**
sea lion	**48**	**33**	**44**	**38**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

UPGMA tree for Sarich (1969) data



UPGMA misleads on a nonclocklike tree

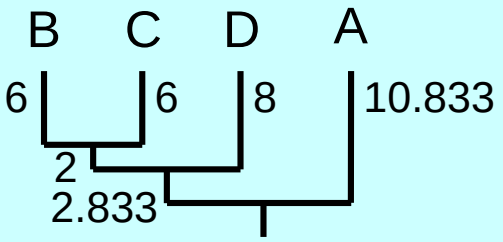
True tree



Distance matrix

	A	B	C	D
A	0	17	21	27
B	17	0	12	18
C	21	12	0	14
D	27	18	14	0

UPGMA tree



An unclocklike tree (left), the distances from it (center) and the UPGMA tree from those distances (right)

The distortion of the tree is due to “short-branch attraction” in which B and C, close to each other in the true tree, cluster first.

Neighbor-joining algorithm

1. For each tip, compute $u_i = \sum_{j \neq i}^{n} D_{ij}/(n-2)$
2. Choose the i and j for which $D_{ij} - u_i - u_j$ is smallest.
3. Join items i and j. Compute the branch length from i to the new node (v_i) and from j to the new node (v_j) as

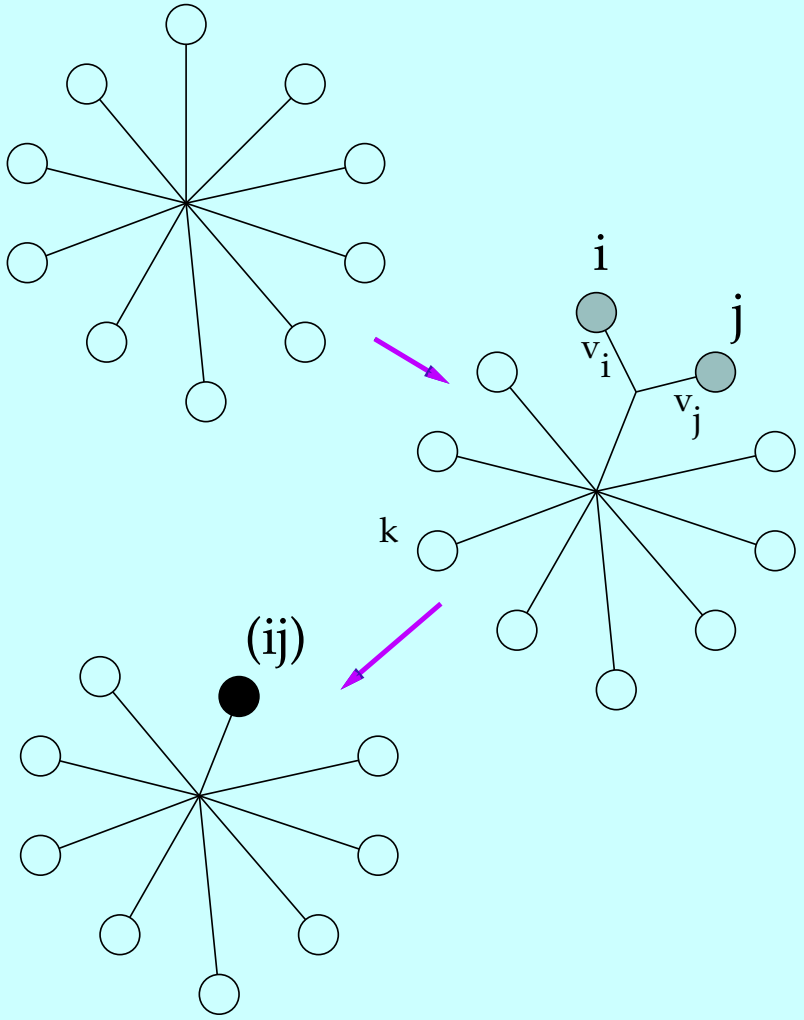
$$\begin{aligned} v_i &= \tfrac{1}{2} D_{ij} + \tfrac{1}{2}(u_i - u_j) \\ v_j &= \tfrac{1}{2} D_{ij} + \tfrac{1}{2}(u_j - u_i) \end{aligned}$$

4. compute the distance between the new node (ij) and each other tip as

$$D_{(ij),k} = (D_{ik} + D_{jk} - D_{ij}) \Big/ 2$$

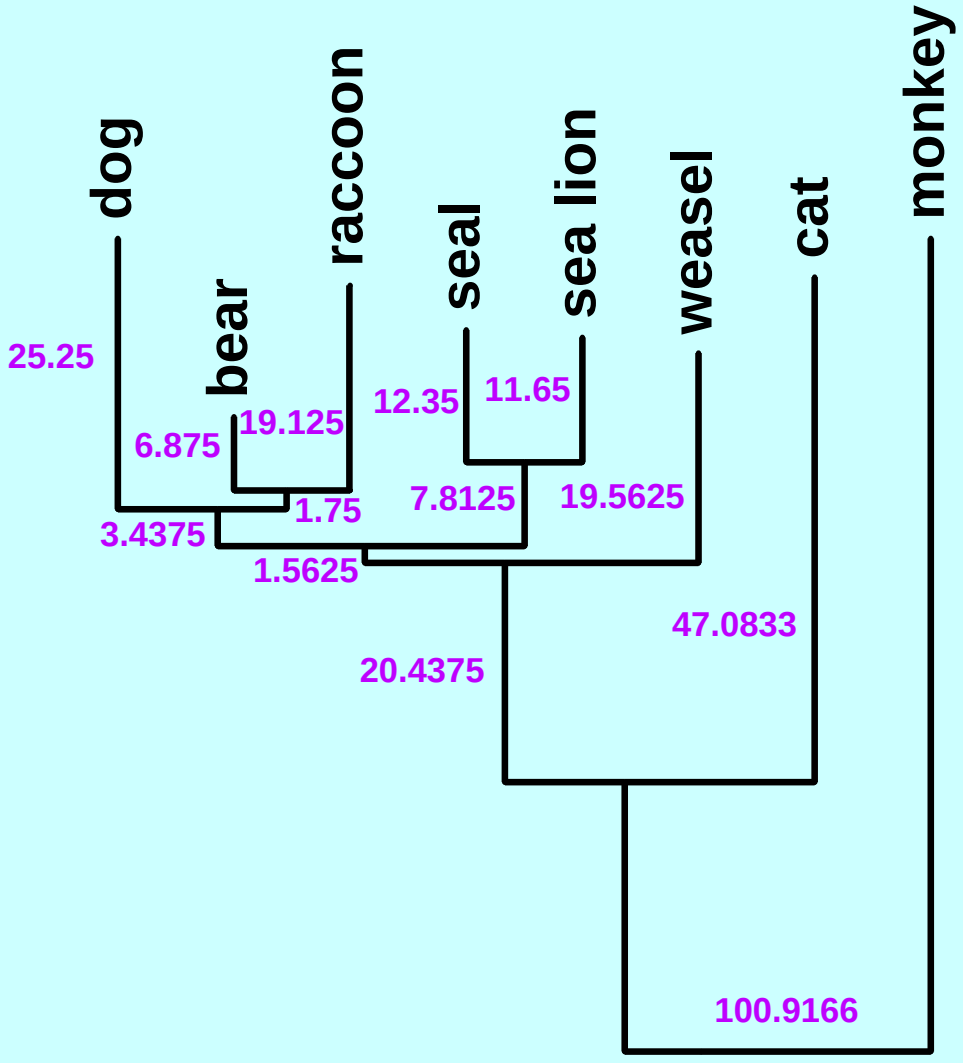
5. delete tips i and j from the tables and replace them by the new node, (ij), which is now treated as a tip.
6. If more than two nodes remain, go back to step 1. Otherwise connect the two remaining nodes by a branch of length D_{ij}.

Star decomposition search



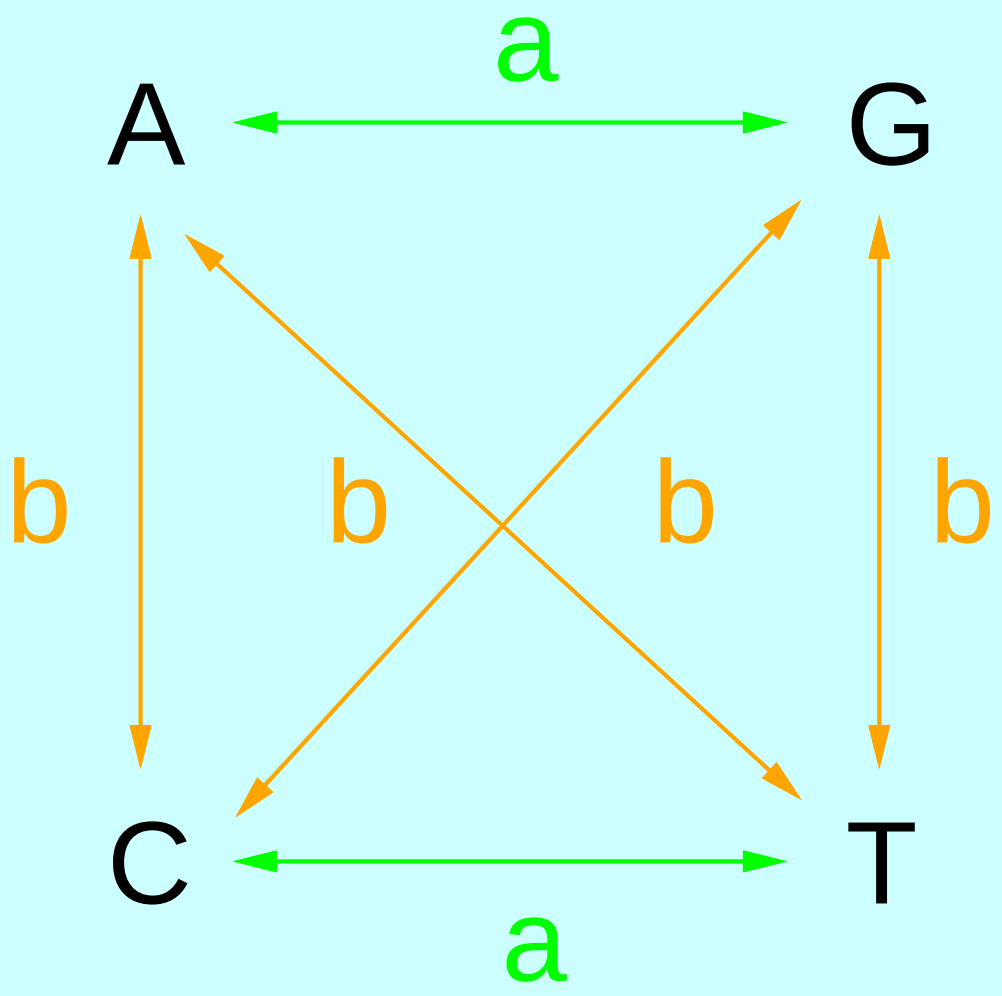
“Star decomposition” tree search method used in Neighbor-Joining method

NJ tree for Sarich's (1969) data



Neighbor-joining tree for the Sarich (1969) imunological distance data

Kimura's (1980) K2P model of DNA change,



which allows for different rates of transitions and transversions,

Motoo Kimura

Motoo Kimura, with family in Mishima, Japan about 1968

Transition probabilities for the K2P model

with two kinds of events:

I. At rate α, if the site has a purine (A or G), choose one of the two purines at random and change to it. If the site has a pyrimidine (C or T), choose one of the pyrimidines at random and change to it.

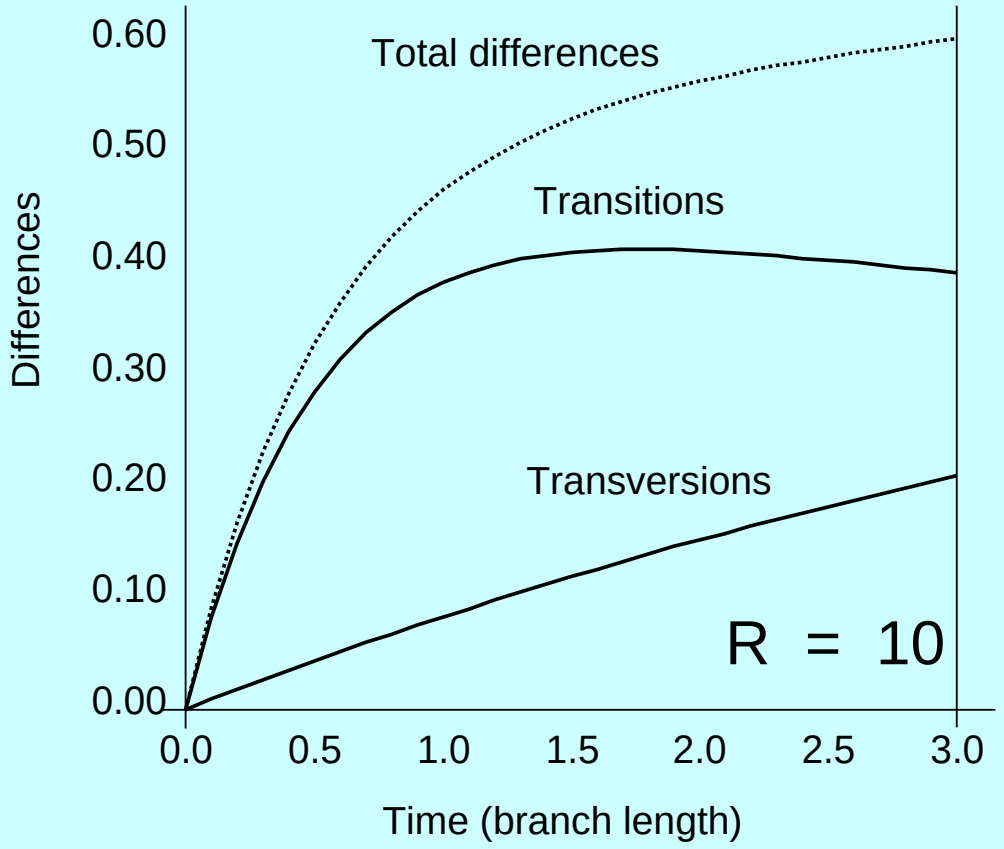
II. At rate β, choose one of the 4 bases at random and change to it.

By proper choice of α and β one can achieve the overall rate of change and T_s/T_n ratio R you want. For rate of change 1, the transition probabilities (*warning: terminological tangle*) and the transversion probabilities are:

$$\begin{aligned} \text{Prob}\,(\text{transition}|t) &= \tfrac{1}{4} - \tfrac{1}{2}\exp\left(-\tfrac{2R+1}{R+1}t\right) + \tfrac{1}{4}\exp\left(-\tfrac{2}{R+1}t\right) \\ \text{Prob}\,(\text{transversion}|t) &= \tfrac{1}{2} - \tfrac{1}{2}\exp\left(-\tfrac{2}{R+1}t\right). \end{aligned}$$

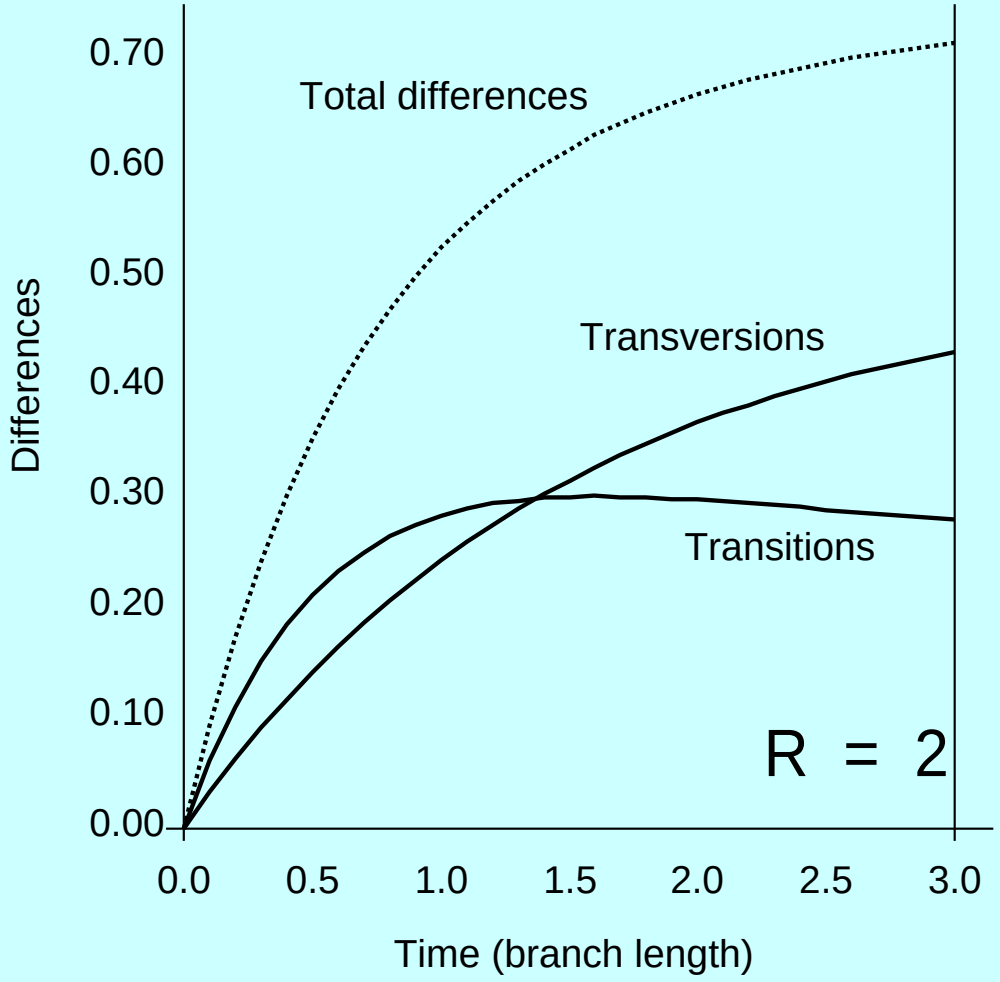
(the transversion probability is the sum of the probabilities of both kinds of transversions).

Transitions, transversions expected



in different amounts of branch length under the K2P model, for $T_s/T_n = 10$

Transitions, transversions expected



in different amounts of branch length under the K2P model, for $T_s/T_n = 2$

Other commonly used models include:

Two models that specify the equilibrium base frequencies (you provide the frequencies $\pi_A, \pi_C, \pi_G, \pi_T$ and they are set up to have an equilibrium which achieves them), and also let you control the transition/transversion ratio:

The **Hasegawa-Kishino-Yano (1985) model**:

to : from :	A	G	C	T
A	$-$	$\alpha\pi_G + \beta\pi_G$	$\alpha\pi_C$	$\alpha\pi_T$
G	$\alpha\pi_A + \beta\pi_A$	$-$	$\alpha\pi_C$	$\alpha\pi_T$
C	$\alpha\pi_A$	$\alpha\pi_G$	$-$	$\alpha\pi_T + \beta\pi_T$
T	$\alpha\pi_A$	$\alpha\pi_G$	$\alpha\pi_C + \beta\pi_C$	$-$

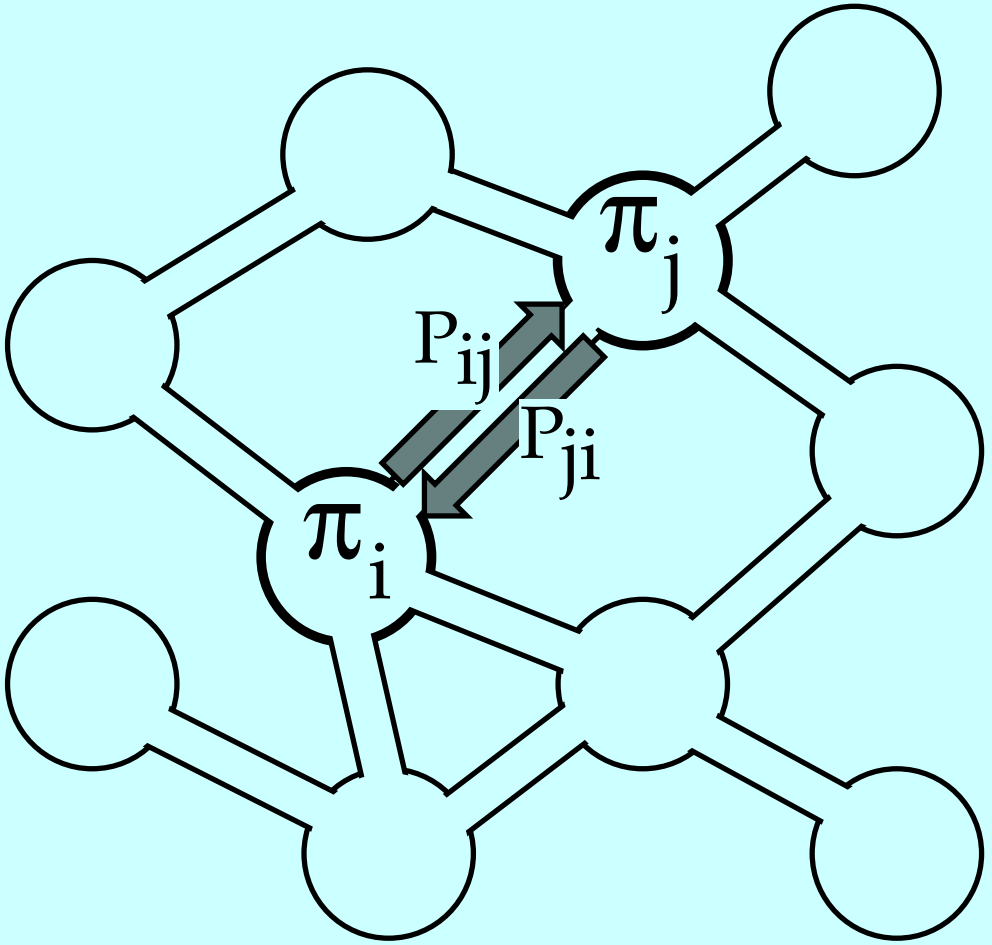
My F84 model

to : from :	A	G	C	T
A	$-$	$\alpha\pi_G + \beta\frac{\pi_G}{\pi_R}$	$\alpha\pi_C$	$\alpha\pi_T$
G	$\alpha\pi_A + \beta\frac{\pi_A}{\pi_R}$	$-$	$\alpha\pi_C$	$\alpha\pi_T$
C	$\alpha\pi_A$	$\alpha\pi_G$	$-$	$\alpha\pi_T + \frac{\beta\pi_T}{\pi_Y}$
T	$\alpha\pi_A$	$\alpha\pi_G$	$\alpha\pi_C + \beta\frac{\pi_C}{\pi_Y}$	$-$

where $\pi_R = \pi_A + \pi_G$ and $\pi_Y = \pi_C + \pi_T$ (The equilibrium frequencies of purines and pyrimidines)

Both of these models have formulas for the transition probabilities, and both are subcases of a slightly more general class of models, the **Tamura-Nei model** (1993).

Reversibility



The General Time-Reversible model (GTR)

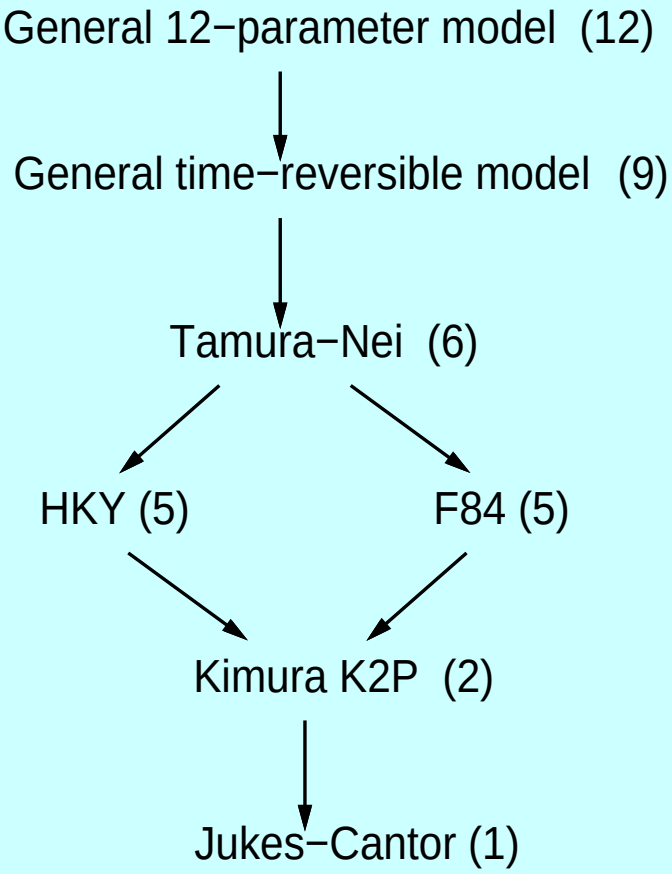
It maintains "detailed balance" so that the probability of starting at (say) A and ending at (say) T in evolution is the same as the probability of starting at T and ending at A:

to : / from :	A	G	C	T
A	$-$	$\alpha\pi_G$	$\beta\pi_C$	$\gamma\pi_T$
G	$\alpha\pi_A$	$-$	$\delta\pi_C$	$\epsilon\pi_T$
C	$\beta\pi_A$	$\delta\pi_G$	$-$	$\upsilon\pi_T$
T	$\gamma\pi_A$	$\epsilon\pi_G$	$\upsilon\pi_C$	$-$

And there is of course the **general 12-parameter model** which has arbitrary rates for each of the 12 possible changes (from each of the 4 nucleotides to each of the 3 others). (Neither of these has formulas for the transition probabilities, but those can be done numerically.)

Relation between models

There are many other models, but these are the most widely-used ones. Here is a general scheme of which models are subcases of which other ones:



The number next to each model is the number of parameters, one of which is (in effect) the branch length.

References, page 1

Rohlf, F. J. 1962. A numerical taxonomic study of the genus *Aedes* (Diptera: Culicidae) with emphasis on the congruence of larval and adult classifications. Ph.D. thesis, Department of Entomology, University of Kansas. [UPGMA – one of two introductions of it]

Sneath, P. H. A. 1962. The construction of taxonomic groups, pp. 289-332 in *Microbial Classification*, eds. G. C. Ainsworth and P. H. A. Sneath. Cambridge University Press, Cambridge. [UPGMA – one of two introductions of it]

Saitou, N., and M. Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* **4:** 406-425. [Neighbor-joining]

Rzhetsky, A. and M. Nei. 1992. A simple method of estimating and testing minimum-evolution trees. *Molecular Biology and Evolution* **9:** 945-967. [The ME method fits branch lengths by least squares, then chooses among topologies using the total branch length rather than the sum of squares]

Farris, J. S. 1981. Distance data in phylogenetic analysis. pp. 3-23 in *Advances in Cladistics. Proceedings of the first meeting of the Willi Hennig Society.*, ed. V. A. Funk and D. R. Brooks. New York Botanical Garden, Bronx. [Criticism of distance methods]

References, page 2

Felsenstein, J. 1984. Distance methods for inferring phylogenies: a justification. *Evolution* **38:** 16-24. [Argument for statistical interpretation of distance methods]

Farris, J. S. 1985. Distance data revisited. *Cladistics* **1:** 67 -85. [Reply to my 1984 paper]

Felsenstein, J. 1986. Distance methods: reply to Farris. *Cladistics* **2:** 130-143. [reply to Farris 1985]

Farris, J. S. 1986. Distances and statistics. *Cladistics* **2:** 1 44-157. [debate was cut off after this]

Bryant, D., and P. Waddell. 1998. Rapid evaluation of least-squares and minimum-evolution criteria on phylogenetic trees. *Molecular Biology and Evolution* **15:** 1346-1359. [quicker least squares distance trees]

Bruno, W. J., N. D. Socci, and A. L. Halpern. 2000. Weighted neighbor joining: a likelihood-based approach to distance-based phylogeny reconstruction. *Molecular Biology and Evolution* **17:** 189-197. [A weighted version of NJ which de-weights large distances appropriately]

References, page 3

Gascuel O. 1997. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Molecular Biology and Evolution* **14:** 685-695. [Like the Bruno et al. method but a compromise which is not as thoroughly corrected statistically, but much faster]

Kimura, M. 1980. A simple model for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution* **16:** 111-120. [Kimura's 2-parameter model]

Hasegawa, M., H. Kishino, and T. Yano. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution* **22:** 160-174. [HKY model]

Tamura, K. and M. Nei. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular Biology and Evolution* **10:** 512-526. [Tamura-Nei model]

Lanave, C., G. Preparata, C. Saccone, and G. Serio. 1984. A new method for calculating evolutionary substitution rates. *Journal of Molecular Evolution* **20:** 86-93. [General reversible model]

Barry, D., and J. A. Hartigan. 1987. Statistical analysis of hominoid molecular evolution. *Statistical Science* **2:** 191-210. [Early use of full 12-parameter model]

References, page 4

Lockhart, P. J., M. A. Steel, M. D. Hendy, and D. Penny. 1994. Recovering evolutionary trees under a more realistic model of sequence evolution. *Molecular Biology and Evolution* **11:** 605-612. [The LogDet distance for correcting for changing base composition]

Felsenstein, J. 2004. *Inferring Phylogenies*. Sinauer Associates, Sunderland, Massachusetts. [material is in chapters 11, 13]

Yang, Z. 2006. *Computational Molecular Evolution*. Oxford University Press, Oxford. [material is in pages 89-93, and chapters 1, 2]

Semple, C. and M. Steel. 2003. *Phylogenetics*. Oxford University Press, Oxford. [Material is in pages 145-160 and chapter 8]