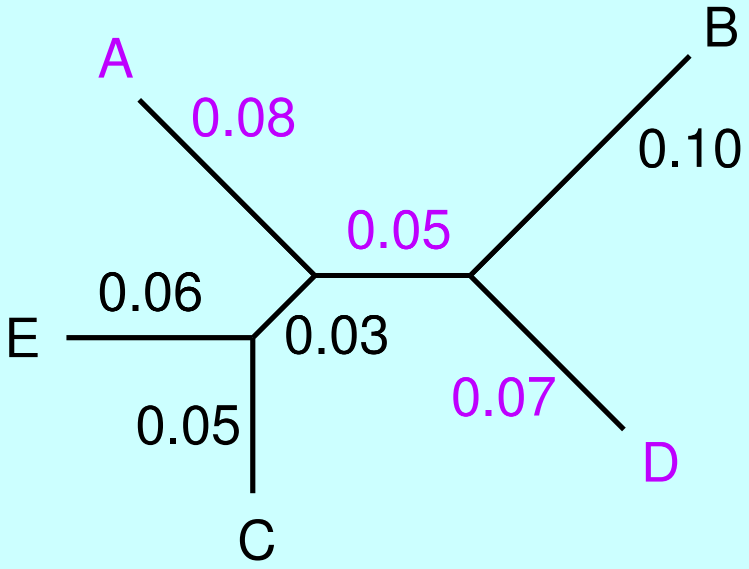
Week 5: Distance methods, DNA and protein models

Genome 570

February, 2016

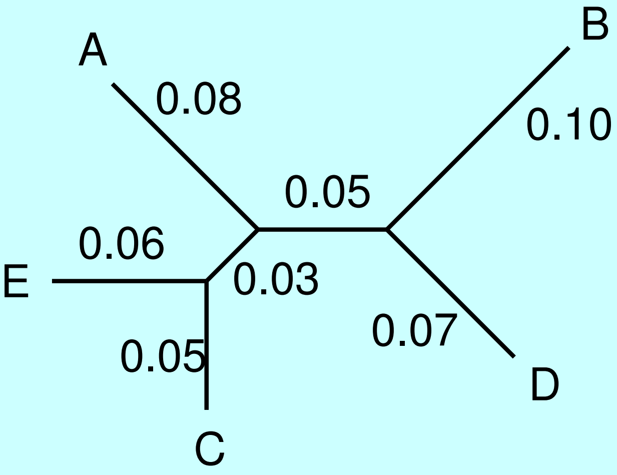
A tree and the expected distances it predicts

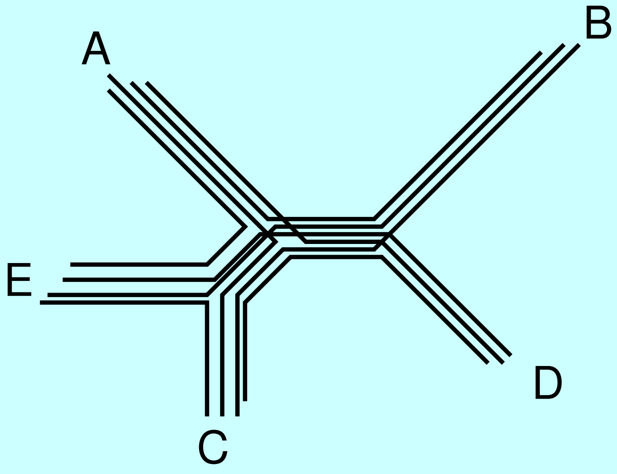


	A	B	C	D	E
A	0	0.23	0.16	0.20	0.17
B	0.23	0	0.23	0.17	0.24
C	0.16	0.23	0	0.15	0.11
D	0.20	0.17	0.15	0	0.21
E	0.17	0.24	0.11	0.21	0

The predicted distances are the sums of branch lengths between those two species.

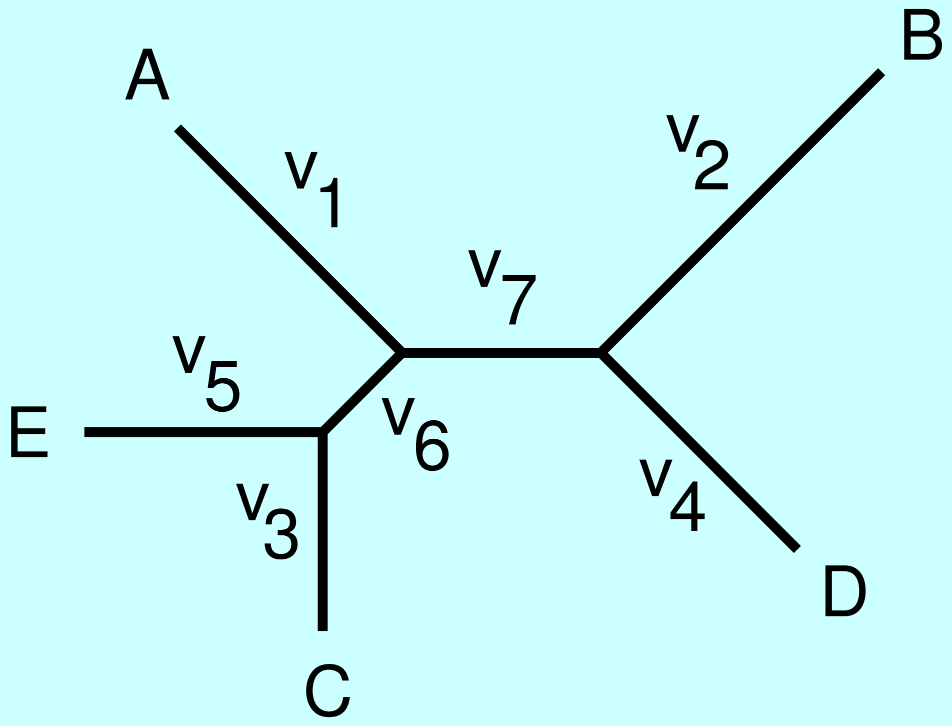
A tree and a set of two-species trees





The two-species trees correspond to the pairwise distances observed between the pairs of species. The tree also predicts two-species trees, because clipping off all other species you are left simply with the path between that pair of species.

A tree with branch lengths



Distance matrix methods always infer trees that have branch lengths, and they assume models of change of the characters (sequences).

Distance matrix methods

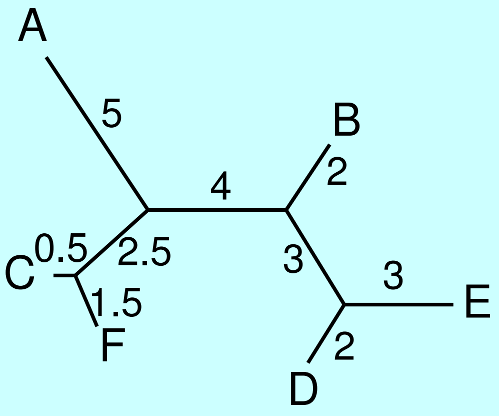
observed distances calculated from the sequence data

	A	B	C	D	E	F
A	0	10	9	12	16	9
B	10	0	10	6	9	9
C	9	10	0	10	15	2
D	12	6	10	0	6	13
E	16	9	15	6	0	15
F	9	9	2	13	15	0

Find the tree which comes closest to predicting the observed pairwise distances

compare

Each possible tree (with branch lengths) predicts pairwise distances



	A	B	C	D	E	F
A	0	11	8	14	15	9
B	11	0	9	7	8	10
C	8	9	0	13	14	2
D	14	7	13	0	5	13
E	15	8	14	5	0	14
F	9	10	2	13	14	0

The math of least squares trees

$$Q = \sum_{i=1}^{n} \sum_{j \neq i} w_{ij} (D_{ij} - d_{ij})^2$$

$$d_{ij} = \sum_k x_{ij,k} v_k$$

$$Q = \sum_{i=1}^{n} \sum_{j \neq i} w_{ij} \left(D_{ij} - \sum_k x_{ij,k} v_k \right)^2 .$$

$$\frac{dQ}{dv_k} = -2 \sum_{i=1}^{n} \sum_{j \neq i} w_{ij} \, x_{ij,k} \left(D_{ij} - \sum_k x_{ij,k} v_k \right) = 0.$$

Equations to solve to infer branch lengths by least squares on a given tree topology, which specifies the $x_{ij,k}$.

Solving for least squares branch lengths

$$\begin{aligned}
\mathsf{D}_{AB} + \mathsf{D}_{AC} + \mathsf{D}_{AD} + \mathsf{D}_{AE} &= 4v_1 + v_2 + v_3 + v_4 + v_5 + 2v_6 + 2v_7 \\
\mathsf{D}_{AB} + \mathsf{D}_{BC} + \mathsf{D}_{BD} + \mathsf{D}_{BE} &= v_1 + 4v_2 + v_3 + v_4 + v_5 + 2v_6 + 3v_7 \\
\mathsf{D}_{AC} + \mathsf{D}_{BC} + \mathsf{D}_{CD} + \mathsf{D}_{CE} &= v_1 + v_2 + 4v_3 + v_4 + v_5 + 3v_6 + 2v_7 \\
\mathsf{D}_{AD} + \mathsf{D}_{BD} + \mathsf{D}_{CD} + \mathsf{D}_{DE} &= v_1 + v_2 + v_3 + 5v_4 + v_5 + 2v_6 + 3v_7 \\
\mathsf{D}_{AE} + \mathsf{D}_{BE} + \mathsf{D}_{CE} + \mathsf{D}_{DE} &= v_1 + v_2 + v_3 + v_4 + 4v_5 + 3v_6 + 2v_7 \\
\mathsf{D}_{AC} + \mathsf{D}_{AE} + \mathsf{D}_{BC} & \\
+\mathsf{D}_{BE} + \mathsf{D}_{CD} + \mathsf{D}_{DE} &= 2v_1 + 2v_2 + 3v_3 + 2v_4 + v_5 + 6v_6 + 4v_7 \\
\mathsf{D}_{AB} + \mathsf{D}_{AD} + \mathsf{D}_{BC} & \\
+\mathsf{D}_{BE} + \mathsf{D}_{CD} + \mathsf{D}_{DE} &= 2v_1 + 3v_2 + 3v_4 + 2v_5 + 4v_6 + 6v_7
\end{aligned}$$

A vector of all distances, stacked

$$\mathsf{d} \;=\; \begin{bmatrix} \mathsf{D}_{\mathsf{AB}} \\ \mathsf{D}_{\mathsf{AC}} \\ \mathsf{D}_{\mathsf{AD}} \\ \mathsf{D}_{\mathsf{AE}} \\ \mathsf{D}_{\mathsf{BC}} \\ \mathsf{D}_{\mathsf{BD}} \\ \mathsf{D}_{\mathsf{BE}} \\ \mathsf{D}_{\mathsf{CD}} \\ \mathsf{D}_{\mathsf{CE}} \\ \mathsf{D}_{\mathsf{DE}} \end{bmatrix}$$

They are in lexicographic (dictionary) order.

The design matrix and least squares equations

$$\mathsf{X} \;=\; \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix}$$

$$\mathsf{X}^{\mathsf{T}}\mathbf{D} \;=\; \left(\mathsf{X}^{\mathsf{T}}\mathsf{X}\right)\,\mathsf{v}.$$

So solution of equations is

$$\mathsf{v} \;=\; (\mathsf{X}^{\mathsf{T}}\mathsf{X})^{-1}\,\mathsf{X}^{\mathsf{T}}\mathbf{D}$$

A diagonal matrix of weights

$$\mathsf{W} = \begin{bmatrix} \mathsf{w_{AB}} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \mathsf{w_{AC}} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \mathsf{w_{AD}} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \mathsf{w_{AE}} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \mathsf{w_{BC}} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \mathsf{w_{BD}} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \mathsf{w_{BE}} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mathsf{w_{CD}} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mathsf{w_{CE}} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mathsf{w_{DE}} \end{bmatrix},$$

Weighted least squares equations

$$\mathsf{X}^{\mathsf{T}}\mathsf{W}\mathsf{D} \;=\; \left(\mathsf{X}^{\mathsf{T}}\mathsf{W}\mathsf{X}\right)\mathsf{v},$$

$$\mathsf{v} \;=\; \left(\mathsf{X}^{\mathsf{T}}\mathsf{W}\mathsf{X}\right)^{-1}\mathsf{X}^{\mathsf{T}}\mathsf{W}\mathsf{D}.$$

These matrix equations solve for weighted least squares estimates of the branch lengths on a given tree. The tree topology is specified by the design matrix X and the weights are the elements of the diagonal matrix W.

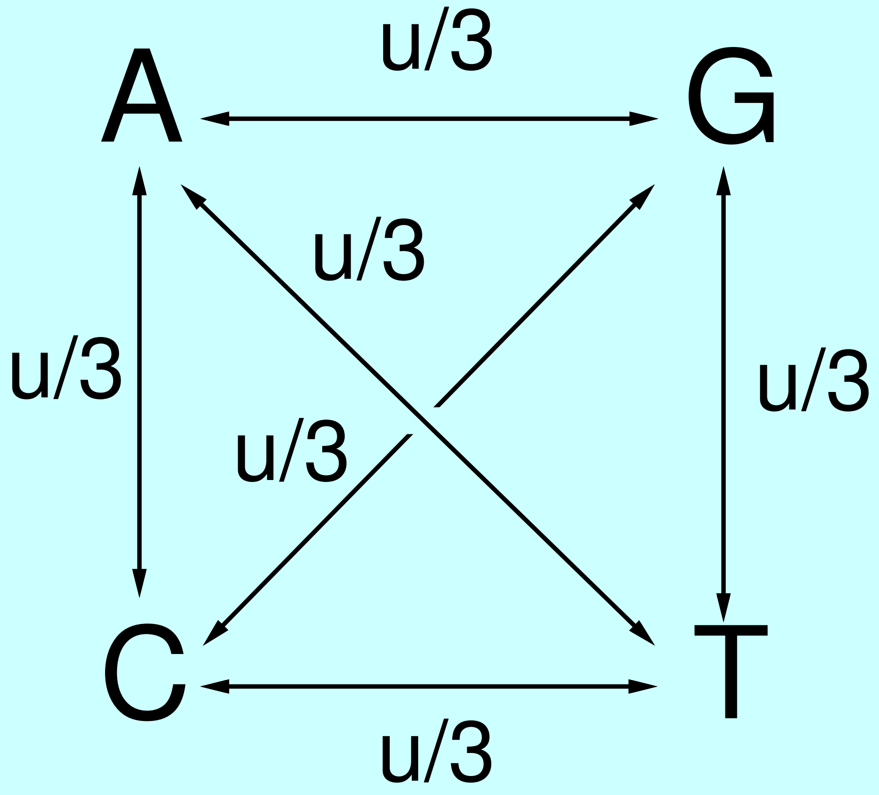
A statistical justification for least squares

$$\mathsf{SSQ} = \sum_{\mathsf{i}=1}^{\mathsf{n}} \sum_{\mathsf{j} \neq \mathsf{i}} \frac{\left(\mathsf{D}_{\mathsf{ij}} - \mathsf{E}\left(\mathsf{D}_{\mathsf{ij}}\right)\right)^2}{\mathrm{Var}\left(\mathsf{D}_{\mathsf{ij}}\right)}.$$

This least squares method

- ... is what we would get by standard statistical least squares approaches *if* the distances were normally distributed, independently varying, and had expectation and variance as shown
- ... but they actually aren't independent in almost all cases (such as molecular sequences), but ...
- ... it can be shown that the estimate of the tree will be a consistent estimate in the case of non-independence, just not as efficient

The Jukes-Cantor model



The simplest and most symmetrical of models of DNA evolution. In a small interval of time `dt` the probability of change at a site is `u dt`, and it is equally likely to go to each of the other three bases.

A simple derivation of the probabilities of net change

- Imagine a (fictional) kind of event that could change the base to one of the four possible bases (including the same base) with equal probability (instead of changing to one of the other three).

A simple derivation of the probabilities of net change

- Imagine a (fictional) kind of event that could change the base to one of the four possible bases (including the same base) with equal probability (instead of changing to one of the other three).
- If you set the probability of change in that model to $\frac{4}{3}u\,dt$, it would be indistinguishable from the actual Jukes-Cantor model.

A simple derivation of the probabilities of net change

- Imagine a (fictional) kind of event that could change the base to one of the four possible bases (including the same base) with equal probability (instead of changing to one of the other three).
- If you set the probability of change in that model to $\frac{4}{3}\mathtt{u}\,\mathtt{dt}$, it would be indistinguishable from the actual Jukes-Cantor model.
- If any nonzero number of these fictional events occur on a branch, the probability of ending up with (say) base C is $\frac{1}{4}$.

A simple derivation of the probabilities of net change

- Imagine a (fictional) kind of event that could change the base to one of the four possible bases (including the same base) with equal probability (instead of changing to one of the other three).
- If you set the probability of change in that model to $\frac{4}{3}\mathrm{u}\,\mathrm{dt}$, it would be indistinguishable from the actual Jukes-Cantor model.
- If any nonzero number of these fictional events occur on a branch, the probability of ending up with (say) base C is $\frac{1}{4}$.
- The number of these events that occur in a branch has Poisson distribution with expected number $\frac{4}{3}\mathrm{u}\,\mathrm{t}$, so the probability of no event is $\mathrm{e}^{-\frac{4}{3}\mathrm{ut}}$

The Jukes-Cantor model

Probability of no event: $\mathsf{e}^{-\frac{4}{3}\mathsf{ut}}$

Probability of some event: $1 - \mathsf{e}^{-\frac{4}{3}\mathsf{ut}}$

Probability of changing to C given start at A, have rate u, time t:

$$\mathrm{Prob}\,(\mathsf{C}|\mathsf{A},\mathsf{u},\mathsf{t}) \;=\; \frac{1}{4}\left(1 - \mathsf{e}^{-\frac{4}{3}\mathsf{ut}}\right)$$

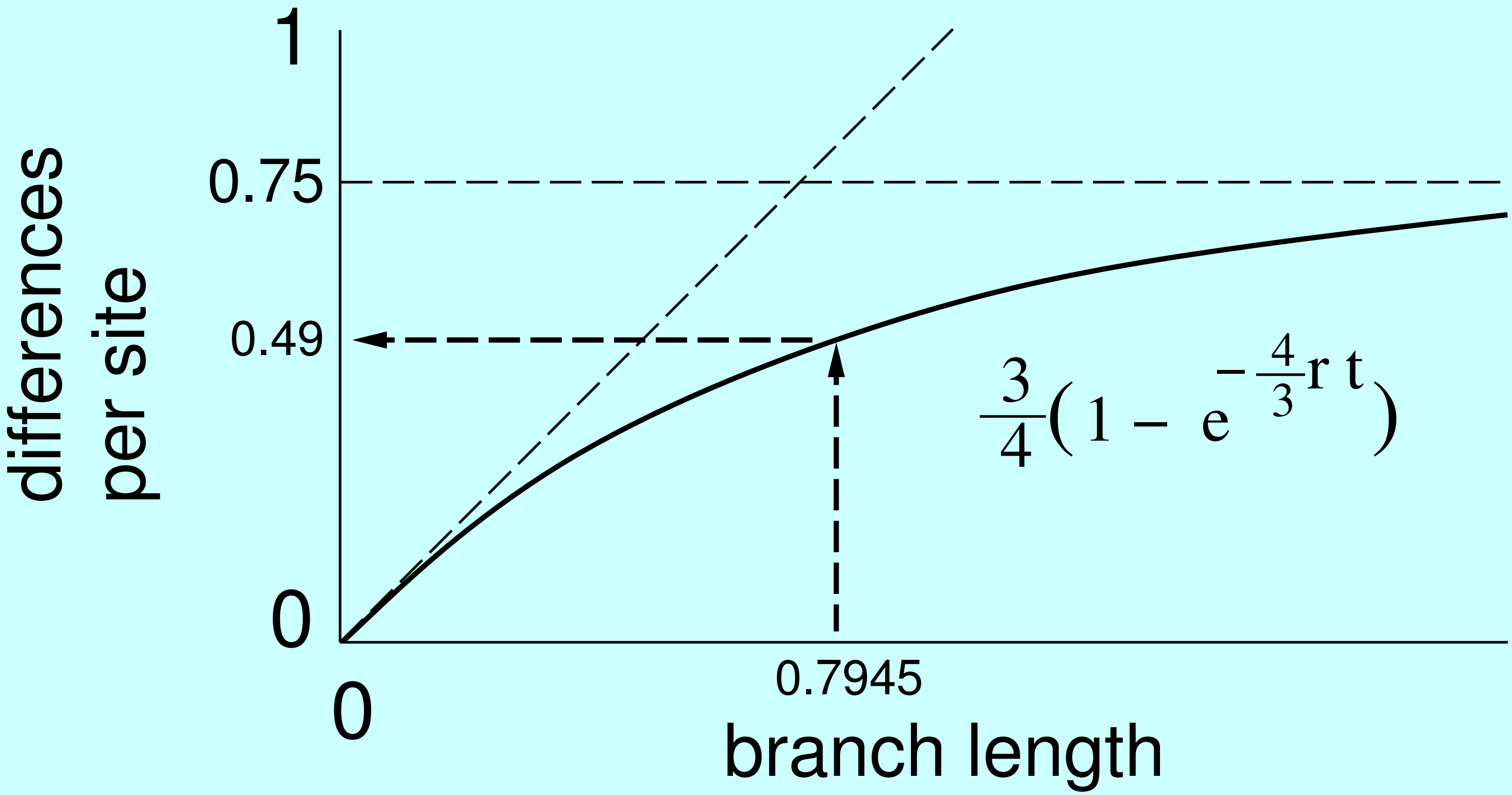
fraction of sites different:

$$\mathsf{f_D} \;=\; \frac{3}{4}\left(1 - \mathsf{e}^{-\frac{4}{3}\mathsf{ut}}\right).$$

Solving, the distance as function of the fraction of sites different is

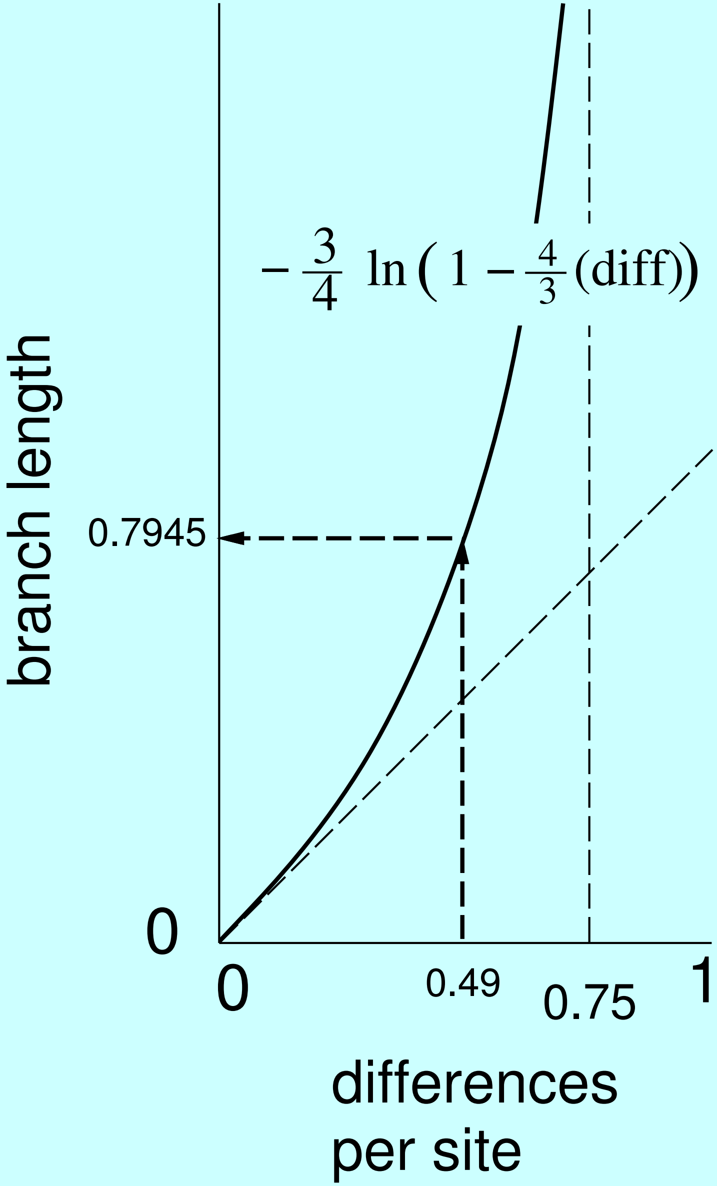
$$\mathsf{D} \;=\; \widehat{\mathsf{ut}} \;=\; -\frac{3}{4}\ln\left(1 - \frac{4}{3}\mathsf{f_D}\right)$$

Fraction of sites different versus branch length

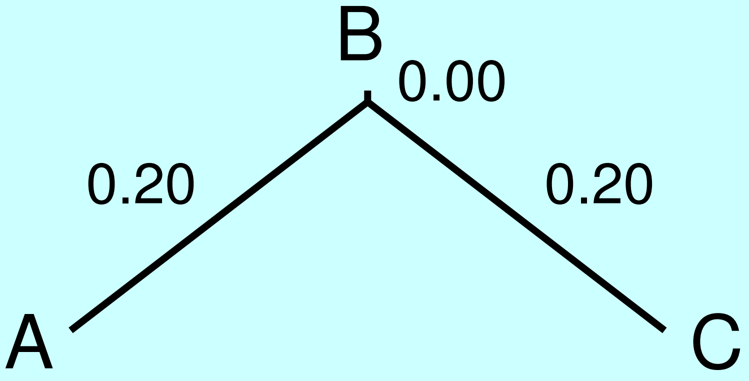


... as predicted by the Jukes-Cantor model.

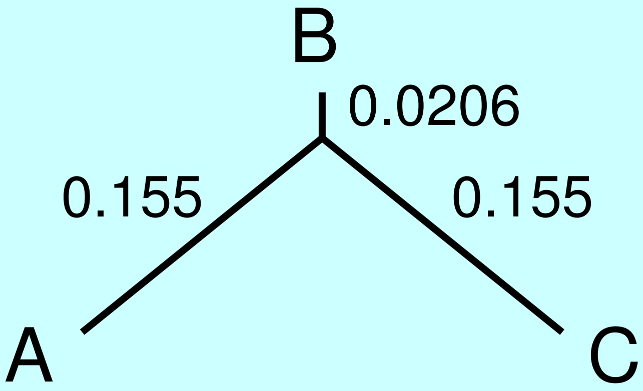
Branch length versus fraction of sites different



If you don’t correct for multiple changes



The true (unrooted) tree



What we estimate

This happening because the long path between A and C is shortened more, proportionally, than the shorter paths between A and B and between C and B. The only way to do this is to put in a spurious branch leading to B.

In effect, there is a “war” between the long paths and the short paths.

With properly corrected distances, each path approaches its true length when we look at a very large number of sites.

Least squares methods

These are just differently weighted least squares methods, as mentioned previously.

Fitch and Margoliash (1967) used a weight of $1/\mathrm{D}_{\mathrm{ij}}^2$, so the quantity to be minimized is

$$\mathrm{Q} \;=\; \sum_{\mathrm{ij}} \frac{(\mathrm{D}_{\mathrm{ij}} - \mathrm{d}_{\mathrm{ij}})^2}{\mathrm{D}_{\mathrm{ij}}^2}$$

Cavalli-Sforza and Edwards (1967) used the unweighted least squares method:

$$\mathrm{Q} \;=\; \sum_{\mathrm{ij}} (\mathrm{D}_{\mathrm{ij}} - \mathrm{d}_{\mathrm{ij}})^2$$

These amount to different assumptions about the how the size of a distance will affect its variance.

The Minimum Evolution method

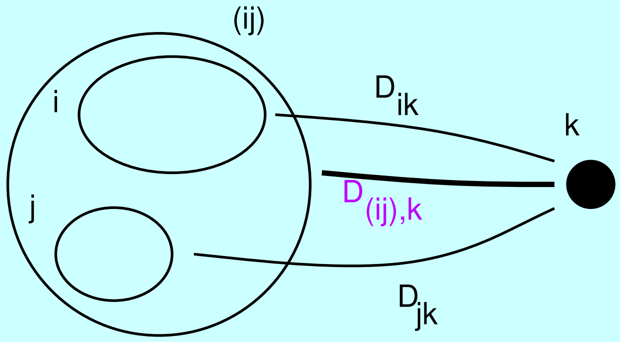
Kidd and Sgaramella-Zonta (1971) and (independently) Rzhetsky and Nei (1992ff.) came up with this method:

- Search through tree space as usual
- For each tree estimate branch lengths by least squares, not allowing negative branch lengths
- Then actually evaluate the tree *not* by the sum of squares, but by the total sum of branch lengths.

This does fairly well, in spite of the mixture of two optimization criteria.

Note that it is *not* directly related to parsimony, in spite of its name.

The UPGMA algorithm



The UPGMA algorithm

1. Assign each species weight $w_i = 1$
2. Pick the values of i and j that are for the smallest of the D_{ij}
3. Make a new group (ij)
4. Assign time depth $D_{ij} / 2$ to the connection between i and j.
5. Its weight is $w_i + w_j$
6. The distance between (ij) and another (say k) is computed as

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

7. Delete i and j from the table, put in a row and column for (ij).
8. If there is only one group left, stop. Otherwise go to step 2.

Note that the weighting in effect weights each of the original tips equally, so that a group's distance to an outside species or another group is the average of all the distances between species in one and species in the other. This is a natural "unweighted" choice.

Sarich 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

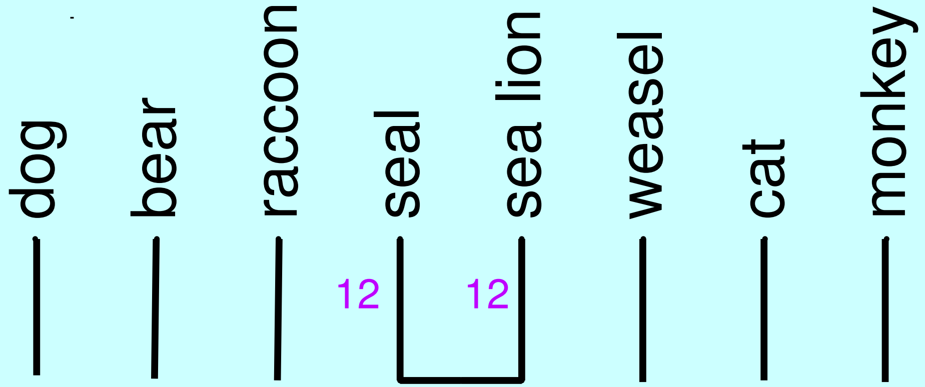
Find smallest element, its rows, columns

						*	*		
		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**	$\boxed{\mathbf{24}}$	**89**	**142**
*	**sea lion**	**48**	**33**	**44**	**38**	$\boxed{\mathbf{24}}$	**0**	**90**	**142**
	cat	98	84	92	86	**89**	**90**	0	148
	monkey	148	136	152	142	**142**	**142**	148	0

We do the following averaging

		dog	bear	raccoon	weasel	* **seal**	* **sea lion**	cat	monkey
	dog	0	32	48	51	$(\mathbf{50}+\mathbf{48})/\mathbf{2}$		98	148
	bear	32	0	26	34	$(\mathbf{29}+\mathbf{33})/\mathbf{2}$		84	136
	raccoon	48	26	0	42	$(\mathbf{44}+\mathbf{44})/\mathbf{2}$		92	152
	weasel	51	34	42	0	$(\mathbf{44}+\mathbf{38})/\mathbf{2}$		86	142
*	**seal**	**49**	**31**	**44**	**41**	**0**	$\boxed{\mathbf{24}}$	**89.5**	**142**
*	**sea lion**					$\boxed{\mathbf{24}}$	**0**		
	cat	98	84	92	86	$(\mathbf{89}+\mathbf{90})/\mathbf{2}$		0	148
	monkey	148	136	152	142	$(\mathbf{142}+\mathbf{142})/\mathbf{2}$		148	0

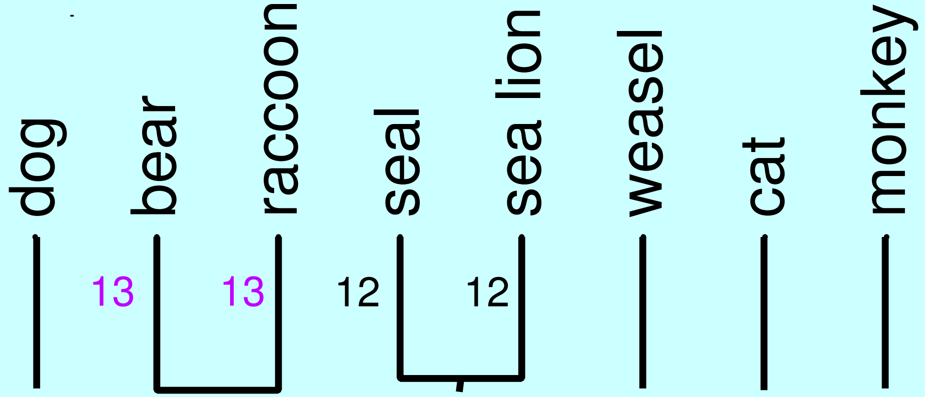
Clustering seal and sea lion



After clustering seal and sea lion

			*	*				
		dog	**bear**	**raccoon**	weasel	SS	cat	monkey
	dog	0	**32**	**48**	51	49	98	148
*	**bear**	**32**	**0**	$\boxed{\textbf{26}}$	**34**	**31**	**84**	**136**
*	**raccoon**	**48**	$\boxed{\textbf{26}}$	**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

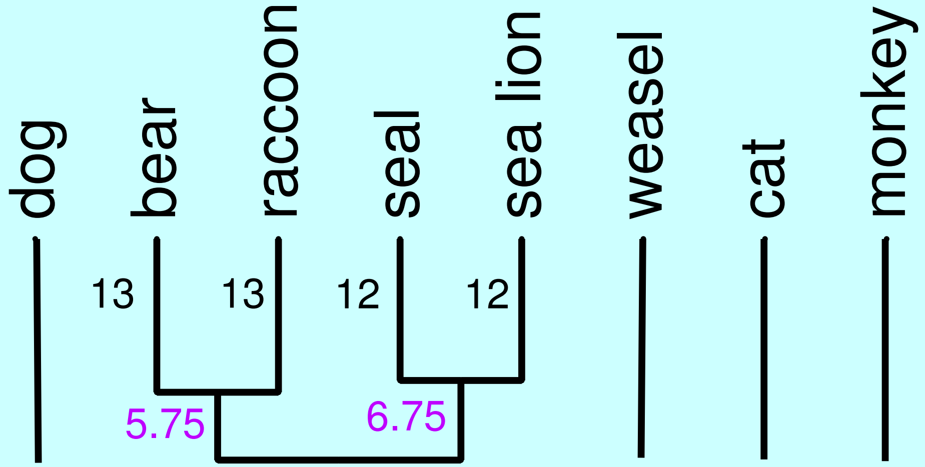
Clustering bear and racoon



After clustering bear and raccoon

			*		*		
		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

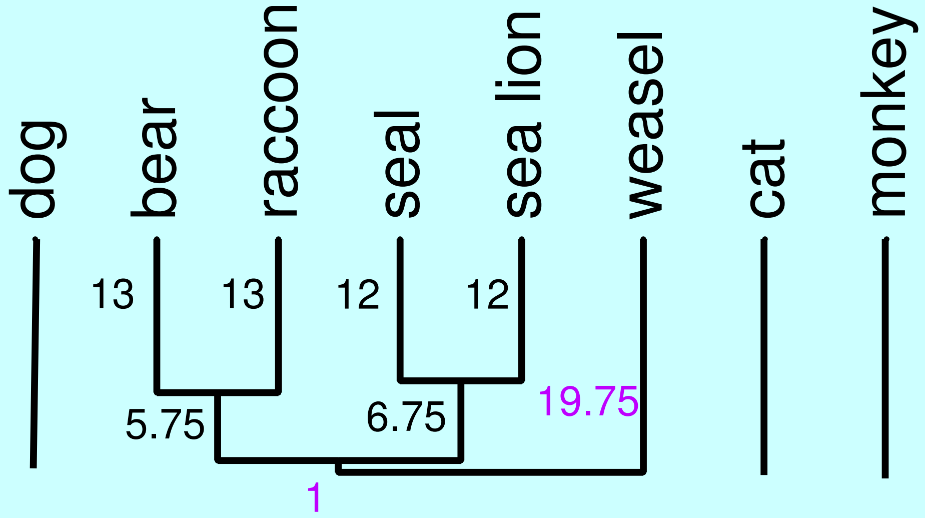
Clustering bear-raccoon with seal-sealion



After clustering those two clusters

			*	*		
		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

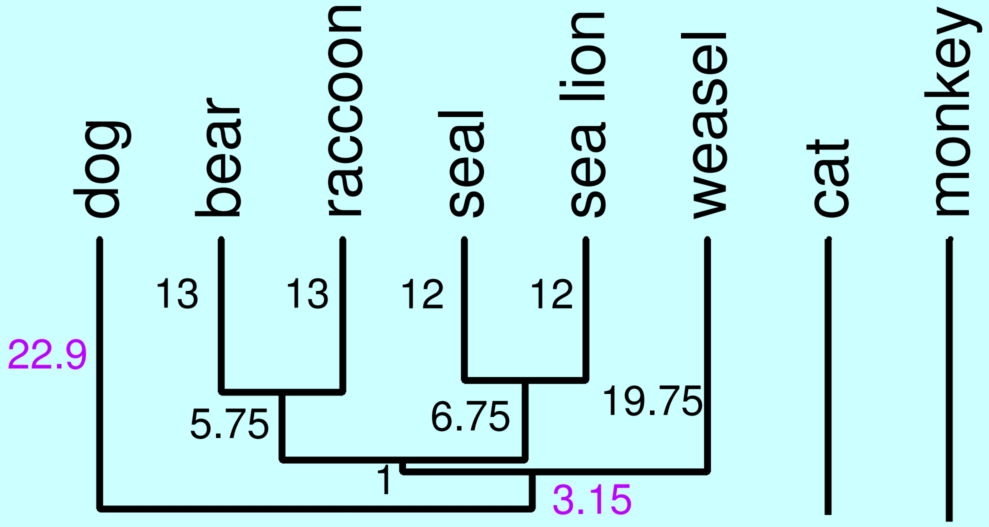
Clustering weasel with BRSS



After adding weasel to that cluster

		*	*		
		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

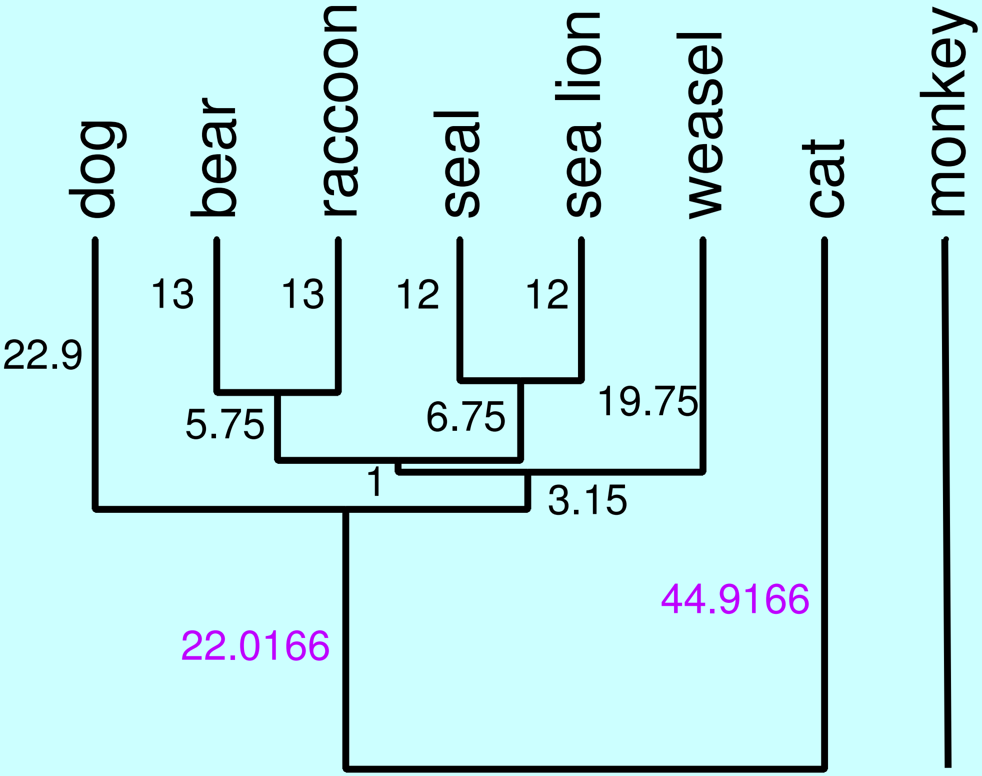
Clustering the dog with BRSSW



After adding dog to it

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

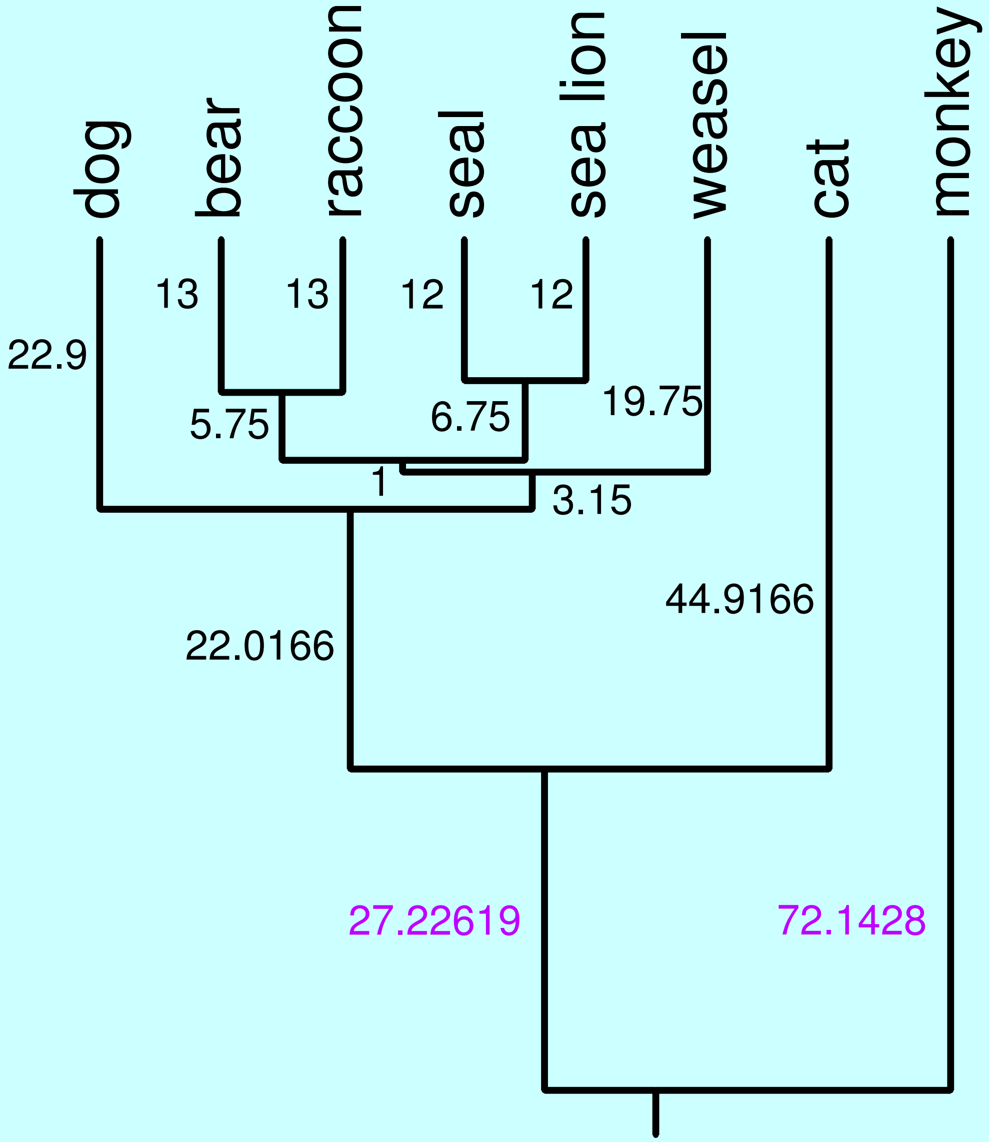
Clustering all the carnivores and pinnipeds



Finally, just monkey remaining

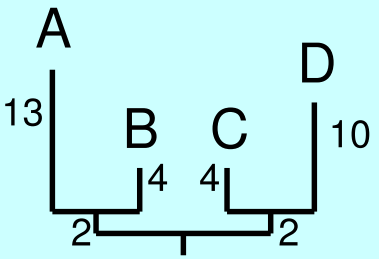
	DBRWSSC	monkey
DBRWSSC	0	144.2857
monkey	144.2857	0

The UPGMA tree



UPGMA can mislead

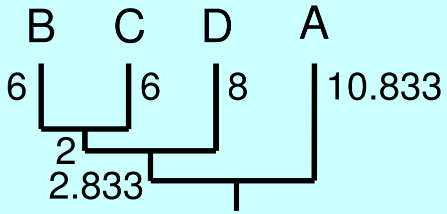
True tree



True 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



Neighbor-Joining

- For each tip, compute $u_i = \sum_{j:j \neq i}^{n} D_{ij}/(n-2)$. Note that the denominator is (deliberately) not the number of items summed.
- Choose the i and j for which $D_{ij} - u_i - u_j$ is smallest.
- 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

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

(continued on next slide)

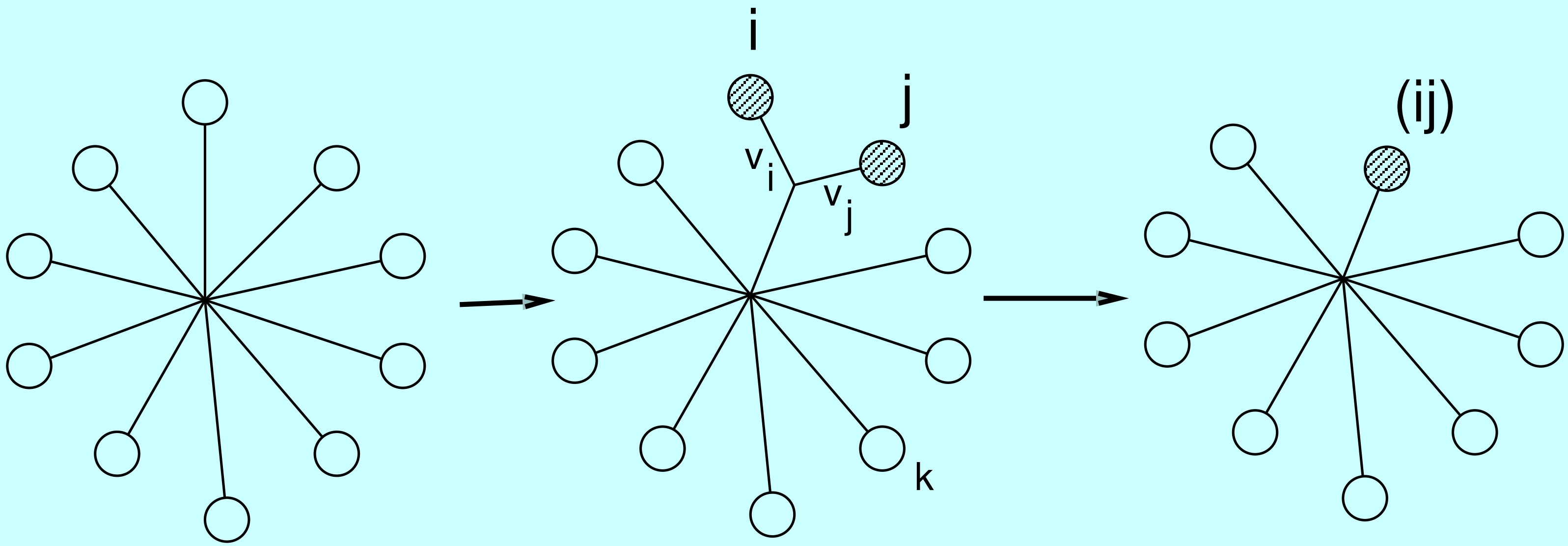
continued ...

- Compute the distance between the new node (ij) and each of the remaining tips as

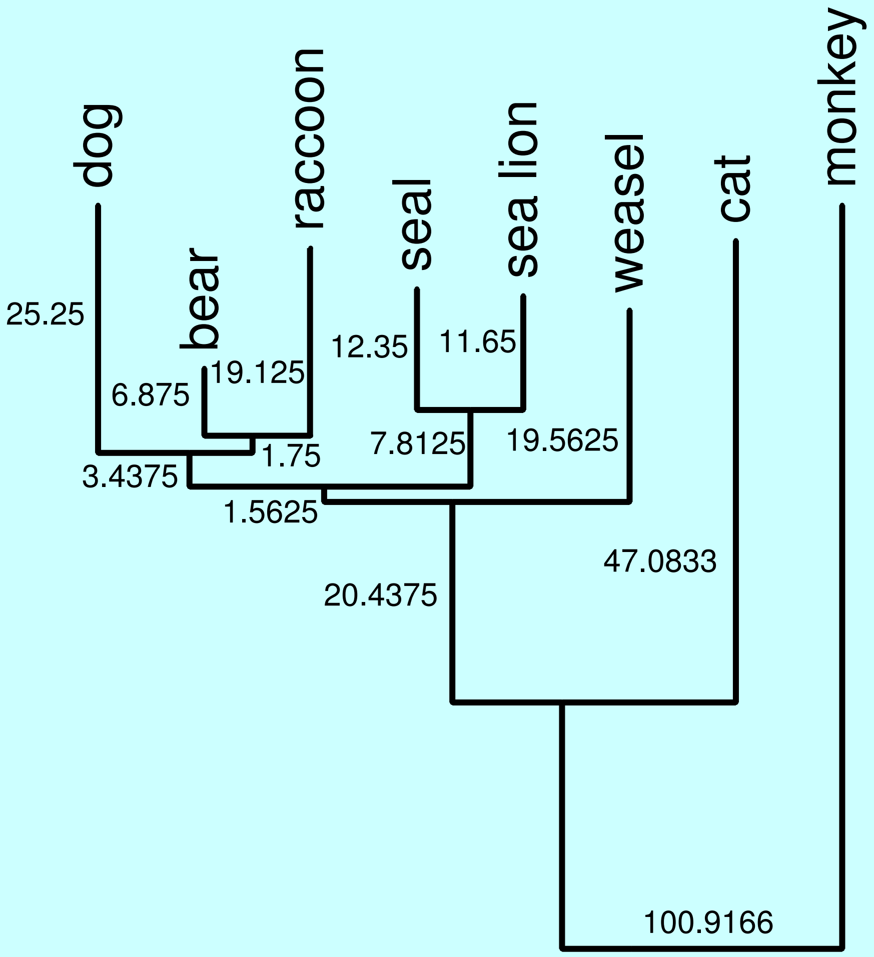
$$\mathrm{D}_{(\mathrm{ij}),\mathrm{k}} \; = \; \left(\mathrm{D}_{\mathrm{ik}} + \mathrm{D}_{\mathrm{jk}} - \mathrm{D}_{\mathrm{ij}}\right) \Big/ 2$$

- Delete tips i and j from the tables and replace them by the new node, (ij) , which is now treated as a tip.
- If more than two nodes remain, go back to step 1. Otherwise, connect the two remaining nodes (say, ℓ and m) by a branch of length $\mathrm{D}_{\ell\mathrm{m}}$.

The NJ star decomposition

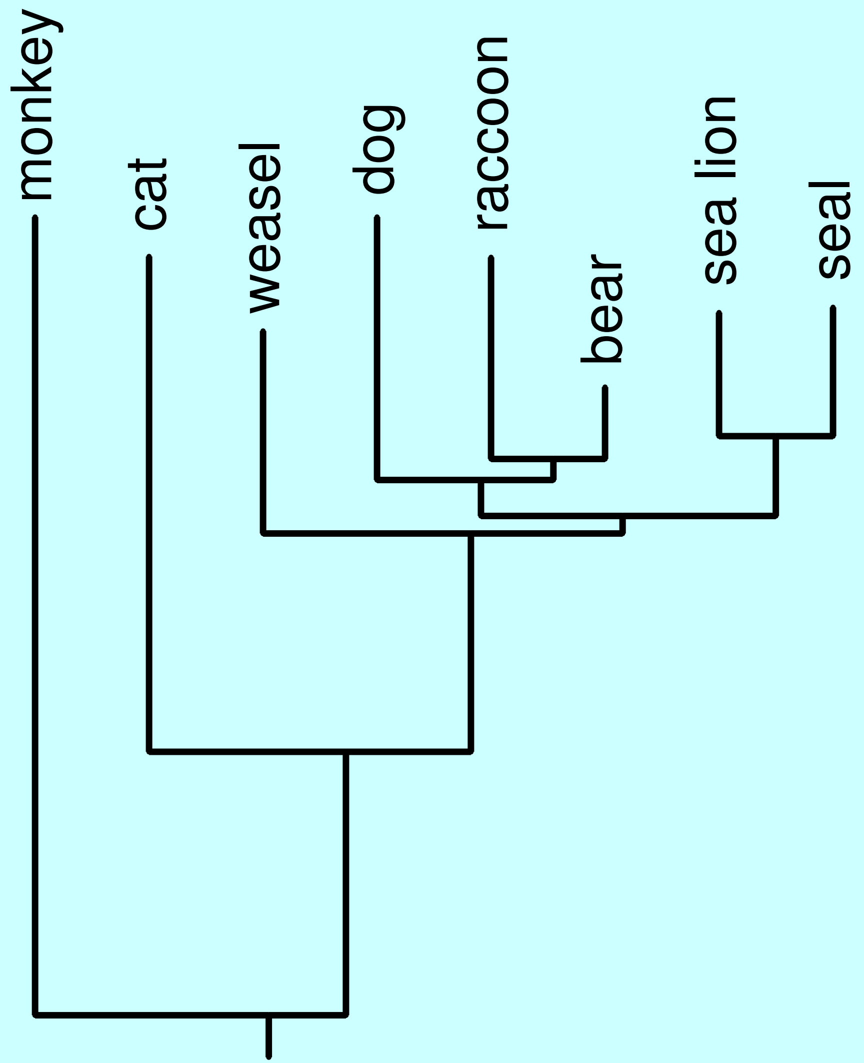


The NJ tree on the Sarich data



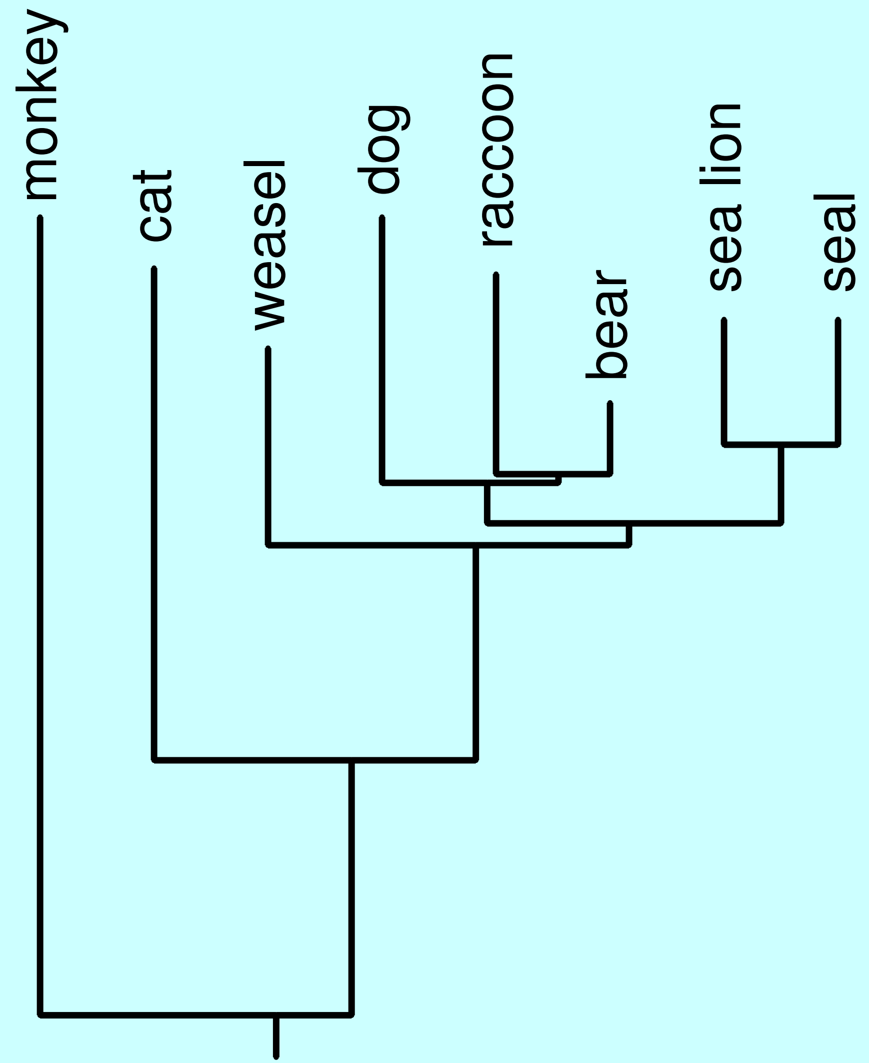
Same as UPGMA? No. Not clocklike.

Unweighted least squares on the Sarich data



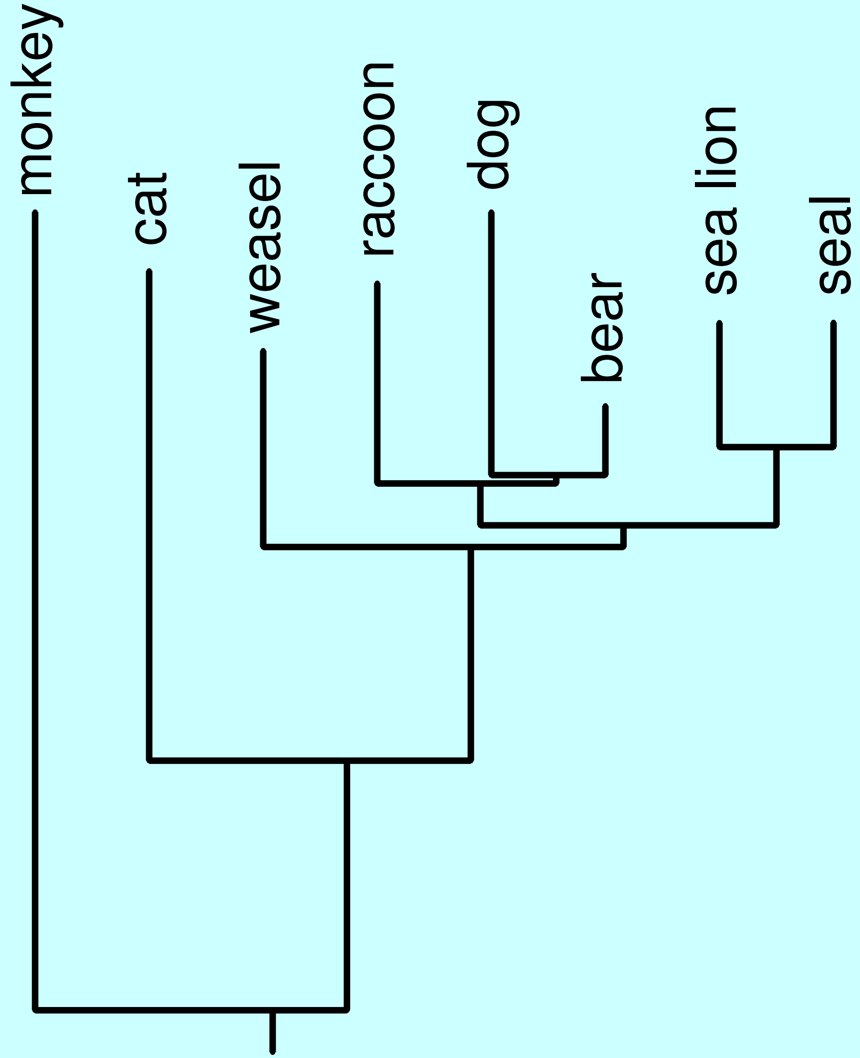
Same as Neighbor-Joining tree? Close but not the same.

Fitch-Margoliash tree on the Sarich (1969) data



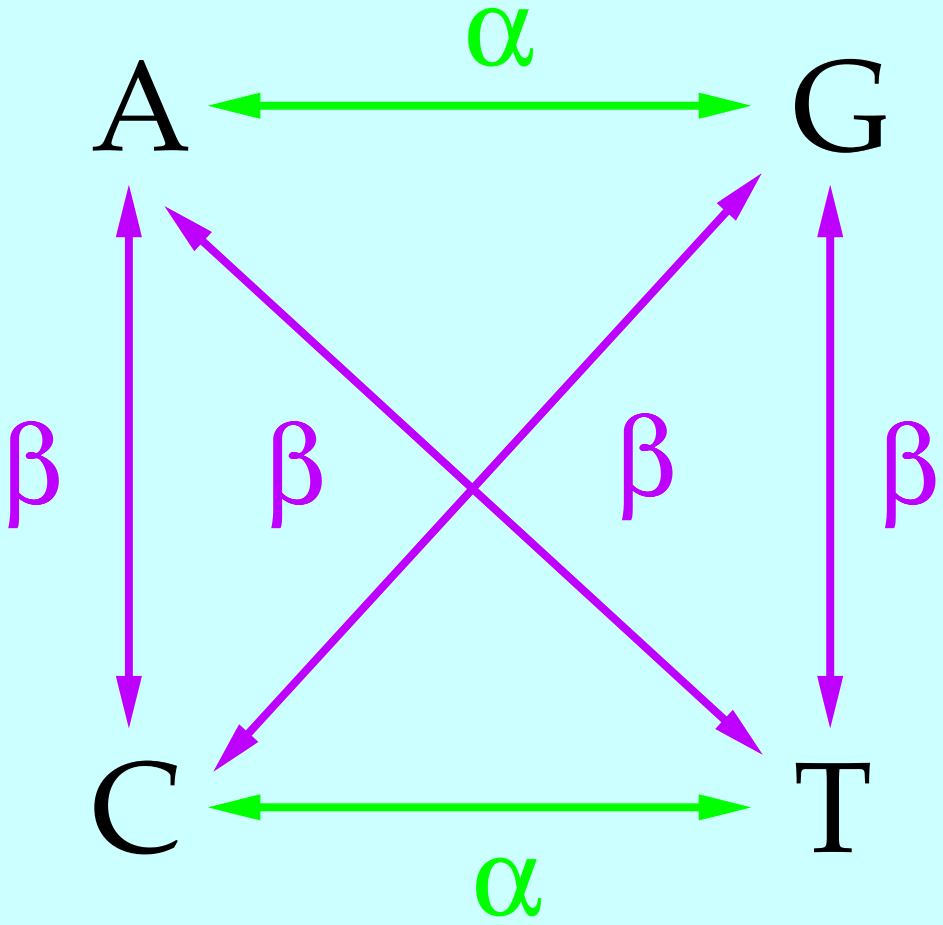
Same as the unweighted least squares tree? No.

Minimum evolution tree on the Sarich (1969) data



Same as either least squares tree? As NJ tree? No.

Kimura 2-parameter model



The simplest, most symmetrical model that has different rates for transitions than for transversions.

Parameters of K2P in terms of Ts/Tn ratio

If we let R be the expected ratio of transition changes to transversions, it turns out that

$$\alpha = \frac{\mathsf{R}}{\mathsf{R}+1}$$

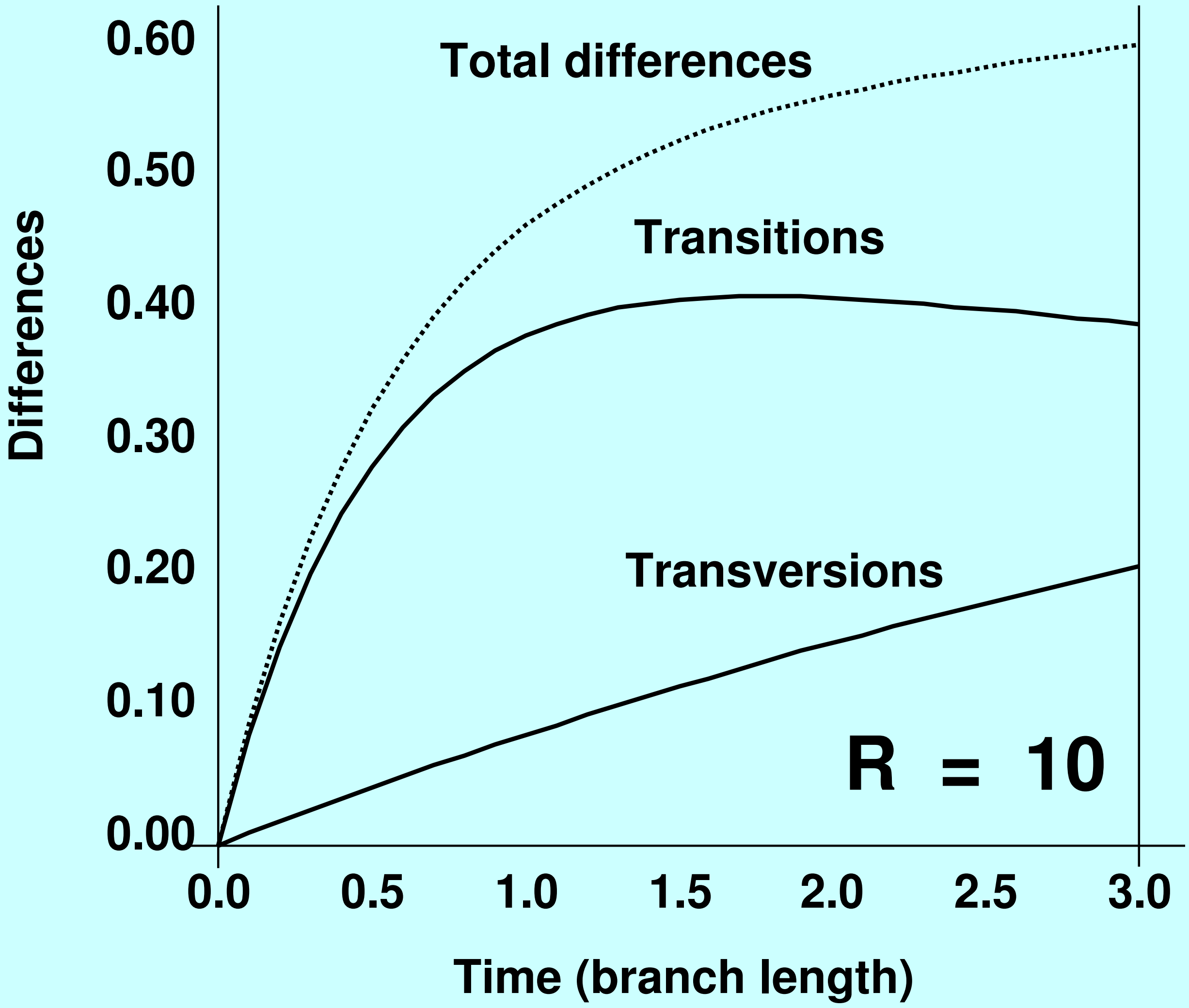
$$\beta = \left(\frac{1}{2}\right)\frac{1}{\mathsf{R}+1}$$

Transition [sic] probabilities for K2P

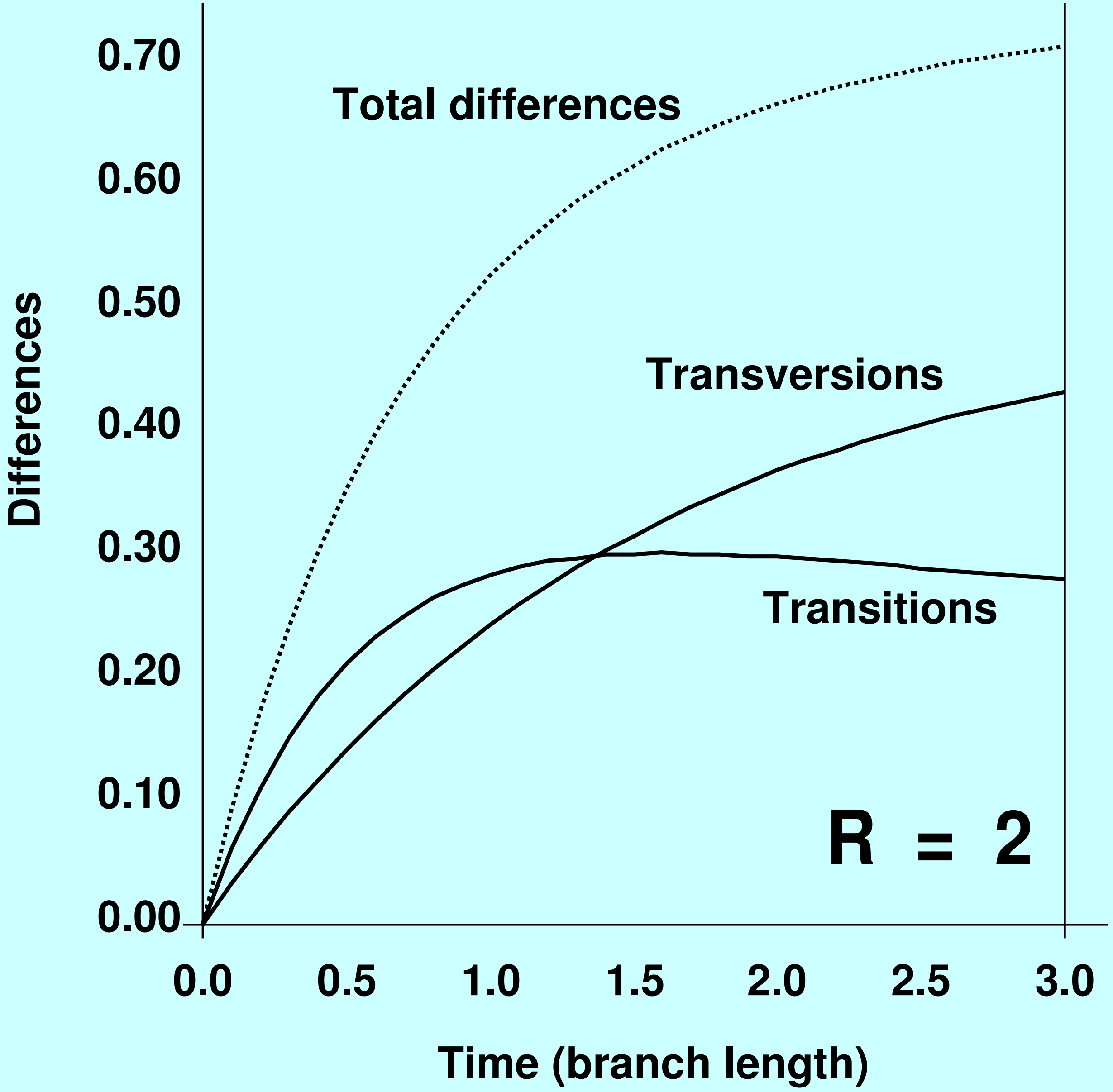
Computing the net probabilities of change (“transition” in the stochastic process sense rather than as molecular biologists use it), it turns out that

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

Transition and transversion when $R = 10$



Transition and transversion when $R = 2$



ML estimates for the K2P model

The sufficient statistics for comparison of DNA sequences under the K2P model are simply the observed fractions P and Q of transitions and transversions. Then the maximum likelihood estimates of the branch length t and of R are obtained by finding the values that predict exactly those fractions. That done by solving the equations given three slides ago.

$$\begin{aligned} \widehat{t} &= -\tfrac{1}{4}\ln\left[(1-2Q)(1-2P-Q)^2\right] \\ \widehat{R} &= \tfrac{-\ln(1-2P-Q)}{-\ln(1-2Q)} - \tfrac{1}{2} \end{aligned}$$

Likelihood for two species under the K2P model

$$\begin{aligned} L &= \mathrm{Prob}\,(\mathrm{data}\,|\,t, R) \\ &= \left(\tfrac{1}{4}\right)^{n} \ (1 - P - Q)^{n-n_1-n_2}\, P^{n_1} \ \left(\tfrac{1}{2}Q\right)^{n_2} \end{aligned}$$

where n_1 is the number of sites differing by transitions, and n_2 is the number of sites differing by transversions. P and Q are the expected fractions of transition and transversion differences, as given by the expressions four screens above.

This is a function of the two parameters t and R . The values that maximize it were given above, if we estimate both of them. But if R is given rather than estimated, there is no closed-form equation solving for t, it has to be inferred numerically by finding the value of t that maximizes L.

The Tamura/Nei model, F84, and HKY

To :	A	G	C	T
From :				
A	$-$	$\alpha_{\mathsf{R}}\pi_{\mathsf{G}}/\pi_{\mathsf{R}} + \beta\pi_{\mathsf{G}}$	$\beta\pi_{\mathsf{C}}$	$\beta\pi_{\mathsf{T}}$
G	$\alpha_{\mathsf{R}}\pi_{\mathsf{A}}/\pi_{\mathsf{R}} + \beta\pi_{\mathsf{A}}$	$-$	$\beta\pi_{\mathsf{C}}$	$\beta\pi_{\mathsf{T}}$
C	$\beta\pi_{\mathsf{A}}$	$\beta\pi_{\mathsf{G}}$	$-$	$\alpha_{\mathsf{Y}}\pi_{\mathsf{T}}/\pi_{\mathsf{Y}} + \beta\pi_{\mathsf{T}}$
T	$\beta\pi_{\mathsf{A}}$	$\beta\pi_{\mathsf{G}}$	$\alpha_{\mathsf{Y}}\pi_{\mathsf{C}}/\pi_{\mathsf{Y}} + \beta\pi_{\mathsf{C}}$	$-$

For the F84 model, $\alpha_{\mathsf{R}} \;=\; \alpha_{\mathsf{Y}}$

For the HKY model, $\alpha_{\mathsf{R}}/\alpha_{\mathsf{Y}} \;=\; \pi_{\mathsf{R}}/\pi_{\mathsf{Y}}$

Transition/transversion ratio for the Tamura-Nei model

$$\begin{aligned} \mathsf{T_s} &= 2\,\alpha_{\mathsf{R}}\,\pi_{\mathsf{A}}\,\pi_{\mathsf{G}}\,/\,\pi_{\mathsf{R}} \;+\; 2\,\alpha_{\mathsf{Y}}\,\pi_{\mathsf{C}}\,\pi_{\mathsf{T}}\,/\,\pi_{\mathsf{Y}} \\ &\quad + \beta\,\left(\,\pi_{\mathsf{A}}\pi_{\mathsf{G}} \;+\; 2\pi_{\mathsf{C}}\,\pi_{\mathsf{T}}\right) \\ \mathsf{T_v} &= 2\,\beta\,\pi_{\mathsf{R}}\,\pi_{\mathsf{Y}} \end{aligned}$$

To get $\mathsf{T_s}/\mathsf{T_v} = \mathsf{R}$ and $\mathsf{T_s} + \mathsf{T_v} = 1$,

$$\beta \;=\; \frac{1}{2\pi_{\mathsf{R}}\pi_{\mathsf{Y}}(1+\mathsf{R})}$$

$$\alpha_{\mathsf{Y}} \;=\; \frac{\pi_{\mathsf{R}}\pi_{\mathsf{Y}}\mathsf{R} - \pi_{\mathsf{A}}\pi_{\mathsf{G}} - \pi_{\mathsf{C}}\pi_{\mathsf{T}}}{(1+\mathsf{R})\left(\pi_{\mathsf{Y}}\pi_{\mathsf{A}}\pi_{\mathsf{G}}\rho + \pi_{\mathsf{R}}\pi_{\mathsf{C}}\pi_{\mathsf{T}}\right)}$$

$$\alpha_{\mathsf{R}} \;=\; \rho\,\alpha_{\mathsf{Y}}$$

Using fictional events to mimic the Tamura-Nei model

We imagine two types of events:

- Type I:
    - If the existing base is a purine, draw a replacement from a purine pool with bases in relative proportions $\pi_A : \pi_G$. This event has rate α_R.
    - If the existing base is a pyrimidine, draw a replacement from a pyrimidine pool with bases in relative proportions $\pi_C : \pi_T$. This event has rate α_Y.
- Type II: No matter what the existing base is, replace it by a base drawn from a pool at the overall equilibrium frequencies: $\pi_A : \pi_C : \pi_G : \pi_T$. This event has rate β.

Transition [sic] probabilities with the Tamura-Nei model

If the branch starts with a purine:

No events	$\exp(-(\alpha_{\mathsf{R}} + \beta)\mathsf{t})$
Some type I, no type II	$\exp(-\beta\,\mathsf{t})\,(1\ -\ \exp(-\alpha_{\mathsf{R}}\,\mathsf{t}))$
Some type II	$1\ -\ \exp(-\beta\,\mathsf{t})$

If the branch starts with a pyrimidine:

No events	$\exp(-(\alpha_{\mathsf{Y}}\ +\ \beta)\,\mathsf{t})$
Some type I, no type II	$\exp(-\beta\,\mathsf{t})\,(1\ -\ \exp(-\alpha_{\mathsf{Y}}\,\mathsf{t}))$
Some type II	$1\ -\ \exp(-\beta\,\mathsf{t})$

A transition probability

So if we want to compute the probability of getting a G given that a branch starts with an A, we add up

- The probability of no events, times 0 (as you can't get a G from an A with no events)
- The probability of “some type I, no type II” times π_G/π_Y (as the last type I event puts in a G with probability equal to the fraction of G's out of all purines).
- The probability of “some type II” times π_G (as if there is any type II event, we thereafter have a probability of G equal to its overall expected frequency, and further type I events don't change that).

A transition probability

So that, for example

$$
\begin{aligned}
\mathrm{Prob}\,(\mathsf{G}|\mathsf{A},\ \mathsf{t}) \ = \\
& \exp(-\beta\,\mathsf{t})\ (\mathsf{1}\ -\ \exp(-\alpha_{\mathsf{R}}\,\mathsf{t}))\ \frac{\pi_{\mathsf{G}}}{\pi_{\mathsf{R}}} \\
& +\ (\mathsf{1}\ -\ \exp(-\beta\,\mathsf{t}))\ \pi_{\mathsf{G}}
\end{aligned}
$$

A more compact expression

More generally, we can use the Kronecker delta notation δ_{ij} and the "Watson-Kronecker" notation $\varepsilon_{\mathsf{ij}}$ to write

$$
\begin{aligned}
\mathrm{Prob}\,(\mathsf{j}\,|\,\mathsf{i},\mathsf{t}) \;=&\\
&\exp(-(\alpha_{\mathsf{i}}+\beta)\mathsf{t})\;\delta_{\mathsf{ij}}\\
&+\exp(-\beta\mathsf{t})\,(1-\exp(-\alpha_{\mathsf{i}}\mathsf{t}))\left(\frac{\pi_{\mathsf{j}}\varepsilon_{\mathsf{ij}}}{\sum_{\mathsf{k}}\varepsilon_{\mathsf{jk}}\pi_{\mathsf{k}}}\right)\\
&+(1-\exp(-\beta\mathsf{t}))\,\pi_{\mathsf{j}}
\end{aligned}
$$

where δ_{ij} is 1 if the two bases i and j are different (0 otherwise), and ε_{ij} is 1 if, of the two bases i and j, one is a purine and one is a pyrimidine (0 otherwise).

Reversibility and the GTR model

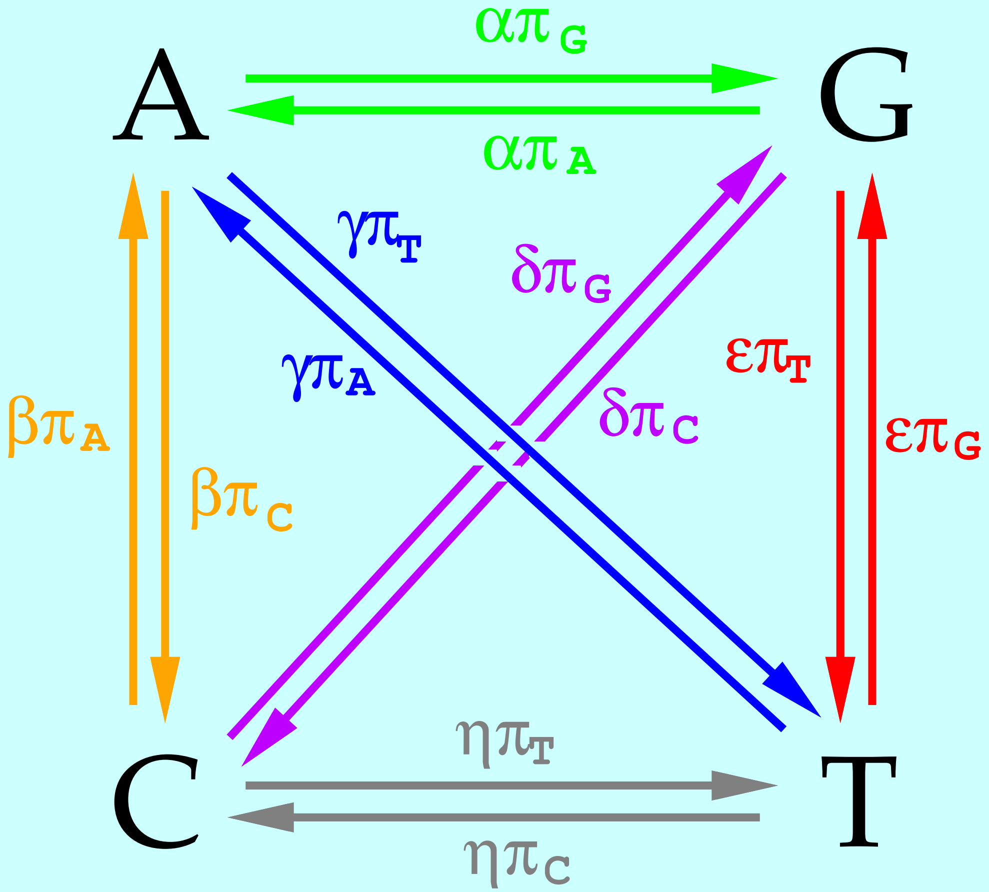
The condition of *reversibility* in a stochastic process is written in terms of the equilibrium frequences of the states (π_{i} and the transition probabilities as

$$\pi_{\mathsf{i}} \;\; \mathrm{Prob}\,(\mathsf{j}|\mathsf{i},\mathsf{t}) \;=\; \pi_{\mathsf{j}} \;\; \mathrm{Prob}\,(\mathsf{i}|\mathsf{j},\mathsf{t})$$

The general time-reversible model:

To :	A	G	C	T
From :				
A	$-$	$\pi_{\mathsf{G}}\ \alpha$	$\pi_{\mathsf{C}}\ \beta$	$\pi_{\mathsf{T}}\ \gamma$
G	$\pi_{\mathsf{A}}\ \alpha$	$-$	$\pi_{\mathsf{C}}\ \delta$	$\pi_{\mathsf{T}}\ \varepsilon$
C	$\pi_{\mathsf{A}}\ \beta$	$\pi_{\mathsf{G}}\ \delta$	$-$	$\pi_{\mathsf{T}}\ \eta$
T	$\pi_{\mathsf{A}}\ \gamma$	$\pi_{\mathsf{G}}\ \varepsilon$	$\pi_{\mathsf{C}}\ \eta$	$-$

The GTR model



It is the most general model possible that still has the changes satisfy the condition of reversibility – so that one cannot tell which direction evolution has gone by examining the sequences before and after.

Standardizing the rates

$$2\pi_{\mathsf{A}}\,\pi_{\mathsf{G}}\,\alpha \;+\; 2\,\pi_{\mathsf{A}}\,\pi_{\mathsf{C}}\,\beta \;+\; 2\,\pi_{\mathsf{A}}\,\pi_{\mathsf{T}}\,\gamma$$

$$+\,2\,\pi_{\mathsf{G}}\,\pi_{\mathsf{C}}\,\delta \;+\; 2\,\pi_{\mathsf{G}}\,\pi_{\mathsf{T}}\,\varepsilon \;+\; 2\,\pi_{\mathsf{C}}\,\pi_{\mathsf{T}}\,\eta \;=\; 1$$

This is done so that the expected rate of change is 1 per unit time.

General Time Reversible models – inference

A data example (simulated under a K2P model, true distance 0.2
transition/transversion ratio = 2

	A	G	C	T	total
A	93	13	3	3	112
G	10	105	3	4	122
C	6	4	113	18	141
T	7	4	21	93	125
total	116	126	140	118	500

The K2P model is a special case of the GTR model (as are all the other models that have been mentioned here). So if all goes well we should infer parameters that come close to specifying a K2P model.

Averaging across the diagonal ...

	A	G	C	T	total
A	93	11.5	4.5	5	114
G	11.5	105	3.5	4	124
C	4.5	3.5	113	19.5	140.5
T	5	4	19.5	93	121.5
total	114	124	140.5	121.5	500

Dividing each column by its sum

(column, because P_{ij} is to be the probability of change from j to i)

$$\hat{\mathbf{P}} = \begin{bmatrix} 0.815789 & 0.0927419 & 0.0320285 & 0.0411523 \\ 0.100877 & 0.846774 & 0.024911 & 0.0329218 \\ 0.0394737 & 0.0282258 & 0.80427 & 0.160494 \\ 0.0438596 & 0.0322581 & 0.13879 & 0.765432 \end{bmatrix}$$

Rate matrix from the matrix logarithm

If the rate matrix is $\mathbf{A}$,

$$\mathbf{P} \;=\; \mathsf{e}^{\mathbf{A}\,\mathsf{t}}$$

so that

$$\widehat{\mathbf{A}\mathsf{t}} \;=\; \log\left(\hat{\mathbf{P}}\right)$$

$$= \left[\begin{array}{llll} -0.212413 & 0.110794 & 0.034160 & 0.046726 \\ 0.120512 & -0.174005 & 0.025043 & 0.035554 \\ 0.0421002 & 0.028375 & -0.236980 & 0.205579 \\ 0.0498001 & 0.034837 & 0.177778 & -0.287859 \end{array}\right].$$

Standardizing the rates

If we denote by $\hat{\mathrm{D}}$ the diagonal matrix of observed base frequencies, and we require that the rate of (potentially-observable) substitution is 1:

$$-\mathrm{trace}(\hat{\mathbf{A}}\hat{\mathbf{D}}) = 1$$

We get:

$$\hat{\mathrm{t}} = -\mathrm{trace}(\widehat{\mathbf{A}\mathrm{t}}\hat{\mathbf{D}}) = -\mathrm{trace}\left(\log(\hat{\mathbf{P}})\hat{\mathbf{D}}\right)$$

and that also gives us an estimate of the rate matrix:

$$\hat{\mathbf{A}} = \log\left(\hat{\mathbf{P}}\right) / -\mathrm{trace}\left(\log(\hat{\mathbf{P}})\hat{\mathbf{D}}\right)$$

The rate estimates

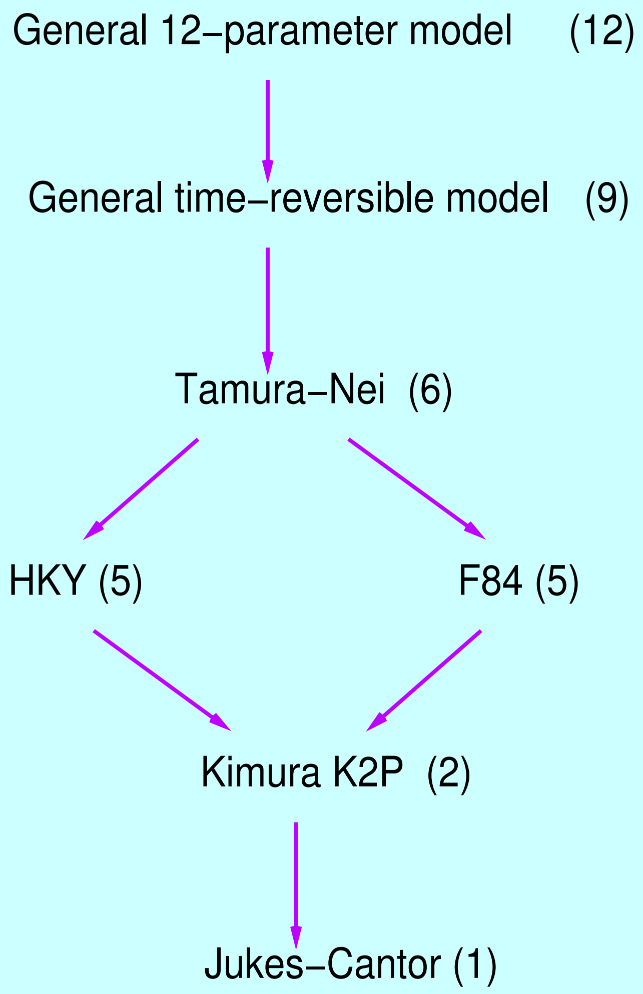
$$\hat{\mathbf{A}} = \begin{bmatrix} -0.931124 & 0.485671 & 0.149741 & 0.204826 \\ 0.528274 & -0.762764 & 0.109776 & 0.155852 \\ 0.184549 & 0.124383 & -1.038820 & 0.901168 \\ 0.218302 & 0.152710 & 0.779302 & -1.261850 \end{bmatrix}.$$

moderately close to the actual K2P rate matrix used in the simulation which was:

$$\mathbf{A} = \begin{bmatrix} -1 & 2/3 & 1/6 & 1/6 \\ 2/3 & -1 & 1/6 & 1/6 \\ 1/6 & 1/6 & -1 & 2/3 \\ 1/6 & 1/6 & 2/3 & -1 \end{bmatrix}$$

(but if any of the eigenvalues of $\log(\hat{\mathbf{P}})$ are negative, this doesn't work and the divergence time is estimated to be infinite).

The lattice of these models



General 12−parameter model (12)

General time−reversible model (9)

Tamura−Nei (6)

HKY (5)

F84 (5)

Kimura K2P (2)

Jukes−Cantor (1)

There are a great many other models, but these are the most commonly used. They allow for unequal expected frequencies of bases, and for inequalities of transitions and transversions.