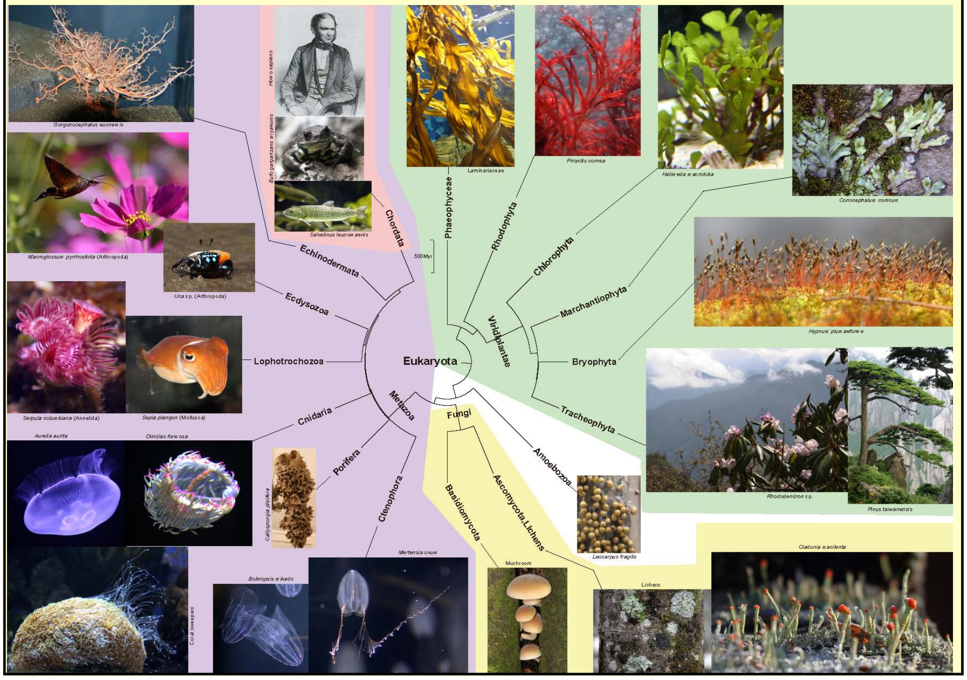
Eukaryotes Tree of Life

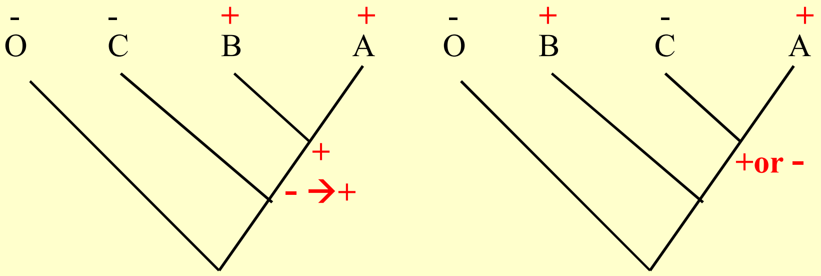


Nothing in biology makes sense without evolution

Theodosuis Dobzhansky (1973)

Nothing in evolution makes sense without phylogeny

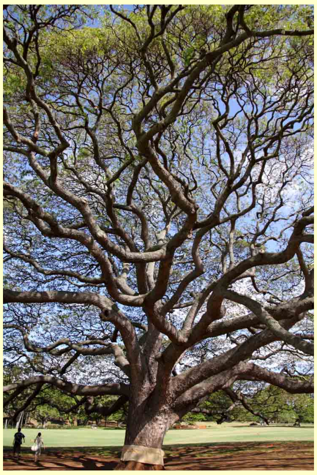
John Avise (2006)

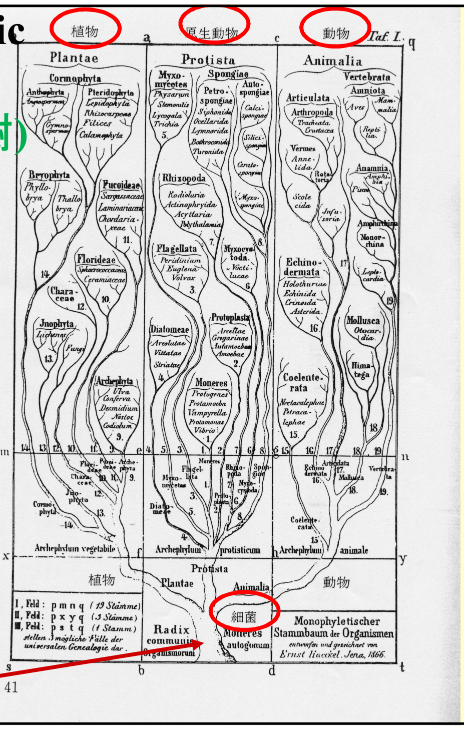


The interpretation of how a particular character evolved depends on which tree is correct.

Haeckel's Phylogenetic Tree (1866)

Tree of Life (生命の樹)



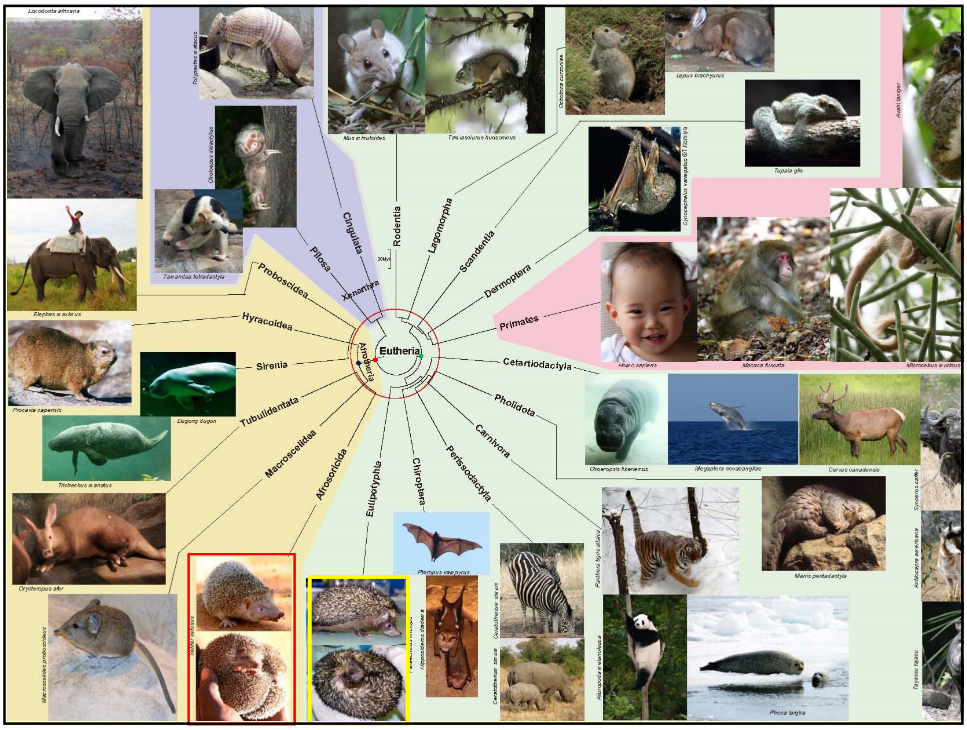


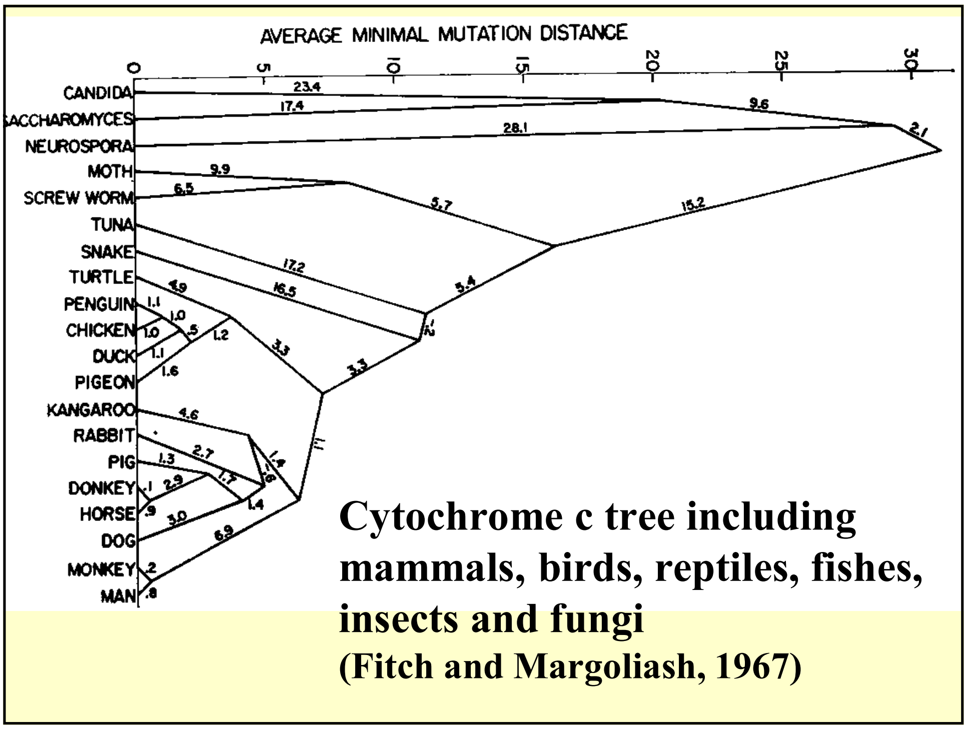
Universal Common Ancestor(UCA)

食虫目(Insectivora)?

Hedgehog
インドハリネズミ
Paraechinus microps

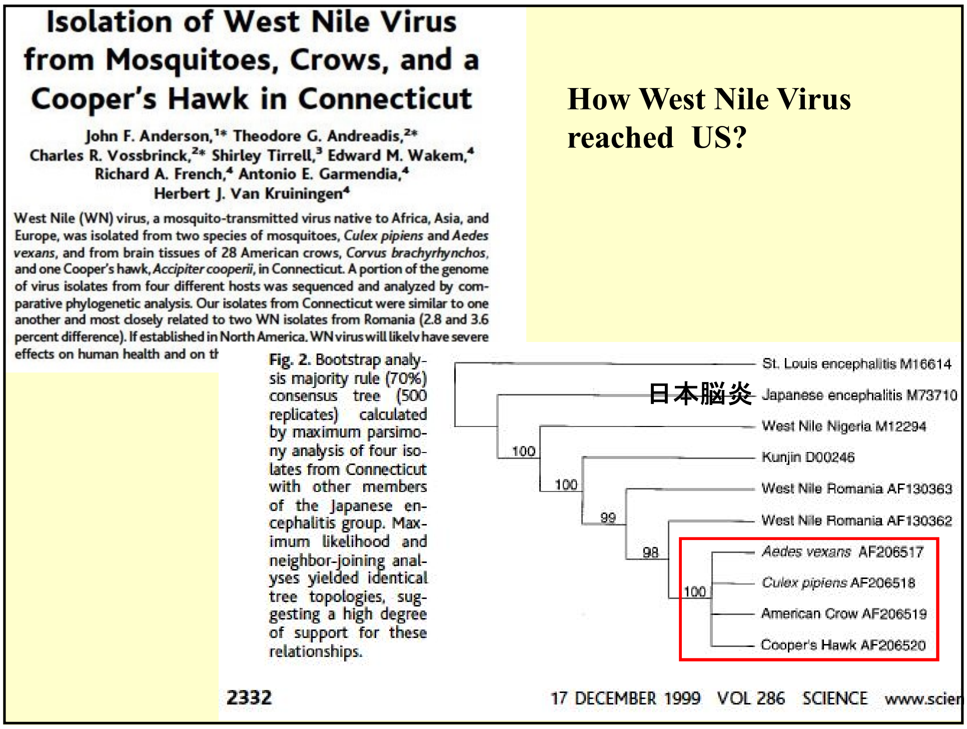
Tenrec
ハリテンレック
Setifer setosus





Cytochrome c tree including mammals, birds, reptiles, fishes, insects and fungi

(Fitch and Margoliash, 1967)



Isolation of West Nile Virus from Mosquitoes, Crows, and a Cooper's Hawk in Connecticut

John F. Anderson,[1]* Theodore G. Andreadis,[2]* Charles R. Vossbrinck,[2]* Shirley Tirrell,[3] Edward M. Wakem,[4] Richard A. French,[4] Antonio E. Garmendia,[4] Herbert J. Van Kruiningen[4]

West Nile (WN) virus, a mosquito-transmitted virus native to Africa, Asia, and Europe, was isolated from two species of mosquitoes, *Culex pipiens* and *Aedes vexans*, and from brain tissues of 28 American crows, *Corvus brachyrhynchos*, and one Cooper's hawk, *Accipiter cooperii*, in Connecticut. A portion of the genome of virus isolates from four different hosts was sequenced and analyzed by comparative phylogenetic analysis. Our isolates from Connecticut were similar to one another and most closely related to two WN isolates from Romania (2.8 and 3.6 percent difference). If established in North America, WN virus will likely have severe effects on human health and on th

How West Nile Virus reached US?

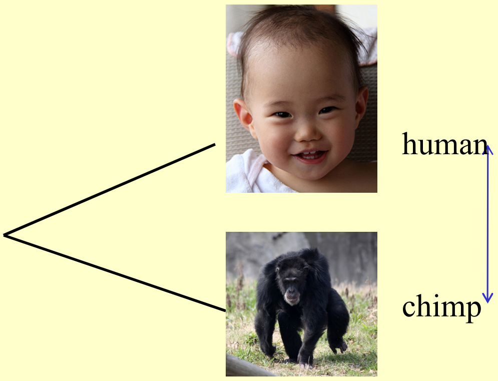
Fig. 2. Bootstrap analysis majority rule (70%) consensus tree (500 replicates) calculated by maximum parsimony analysis of four isolates from Connecticut with other members of the Japanese encephalitis group. Maximum likelihood and neighbor-joining analyses yielded identical tree topologies, suggesting a high degree of support for these relationships.

2332 17 DECEMBER 1999 VOL 286 SCIENCE www.scien

Neutral Theory of Molecular Evolution Motoo Kimura (1968)

Evolution in the molecular level is driven mostly by neutral substitutions, which are not necessarily advantageous.

Dr. Motoo Kimura (1990)
at Cold Spring Harbor



Mutations which occur in the individual level are not sufficient to produce the difference of DNA between the two species. The mutation must be fixed in the population so as to produce the difference of the species.

A mutation in the individual level must be distinguished from a substitution in the population level.

Rate of molecular evolution:

v (substitution/site/year)

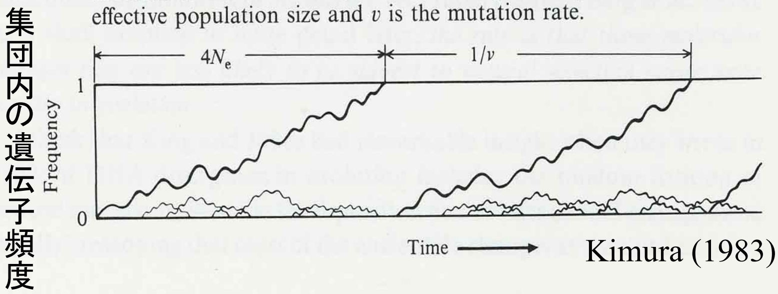
Population size: N

Mutation rate: μ

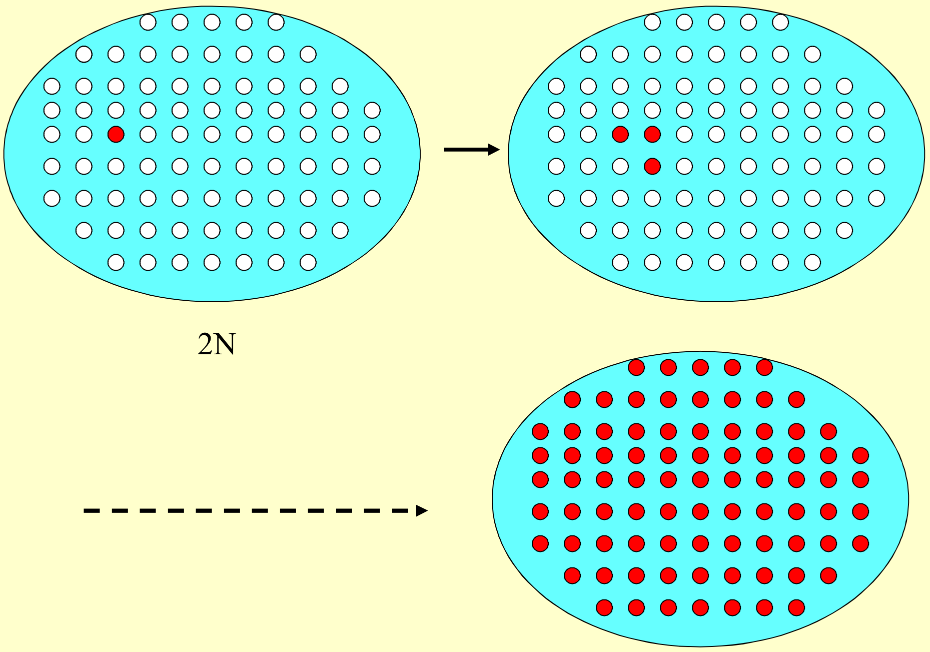
Fixation probability of mutant gene: u

$v = 2N\mu u$

Fig. 3.1. Behavior of mutant genes following their appearance in a finite population. Courses of change in the frequencies of mutants destined to fixation are depicted by thick paths. N_e stands for the effective population size and v is the mutation rate.



Kimura (1983)

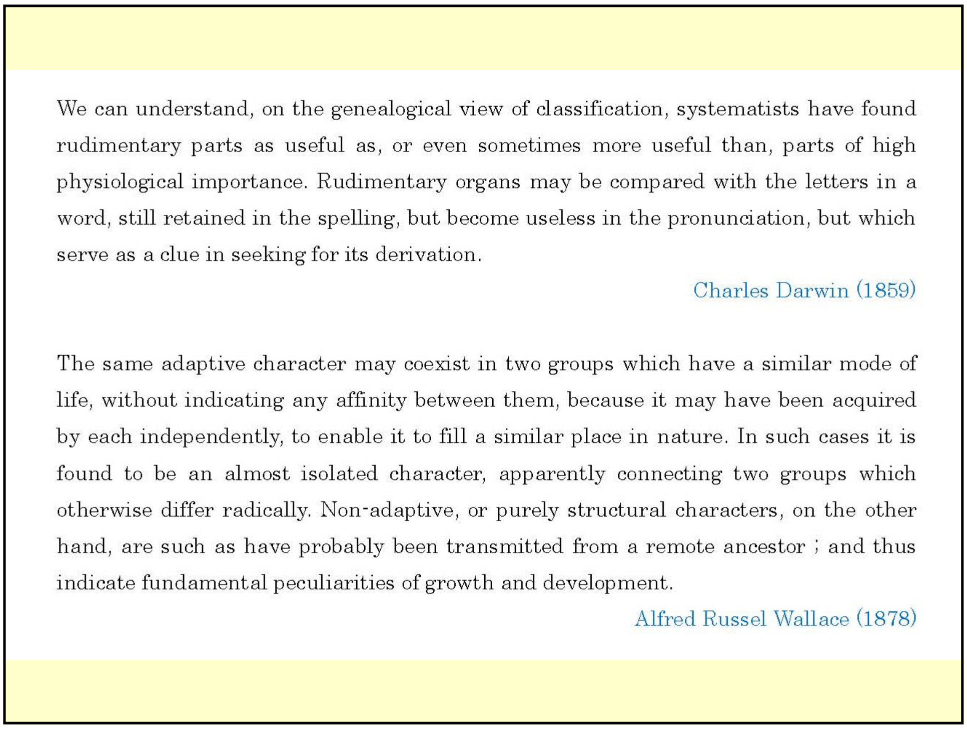


Fixation of a mutant gene in the population

Molecular evolutionary rate of neutral mutation

Rate of molecular evolution:
v (substitution/site/year)
Population size: N
Mutation rate: μ
Fixation probability of mutant gene: u
$v = 2N\mu u$

In the neutral case: $u = 1/(2N)$
→ $v = 2N\mu/(2N) = \mu$

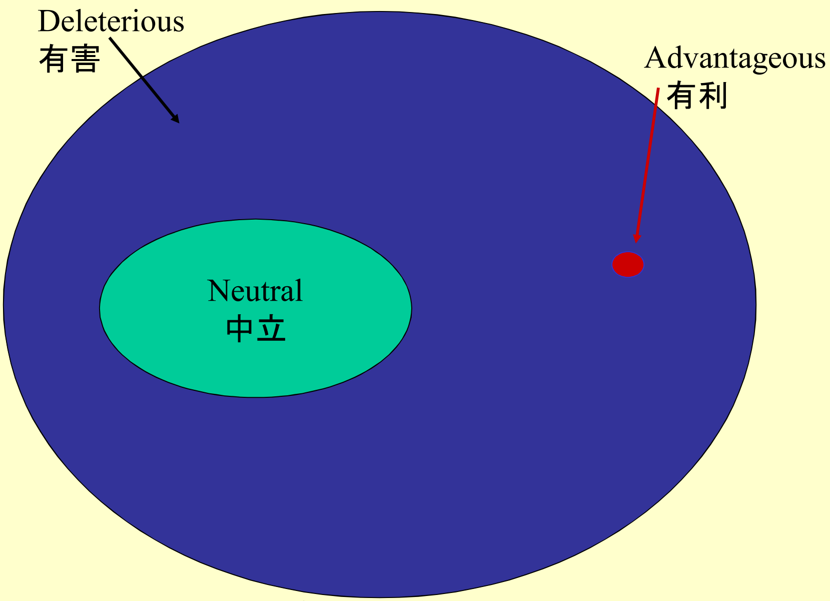


We can understand, on the genealogical view of classification, systematists have found rudimentary parts as useful as, or even sometimes more useful than, parts of high physiological importance. Rudimentary organs may be compared with the letters in a word, still retained in the spelling, but become useless in the pronunciation, but which serve as a clue in seeking for its derivation.

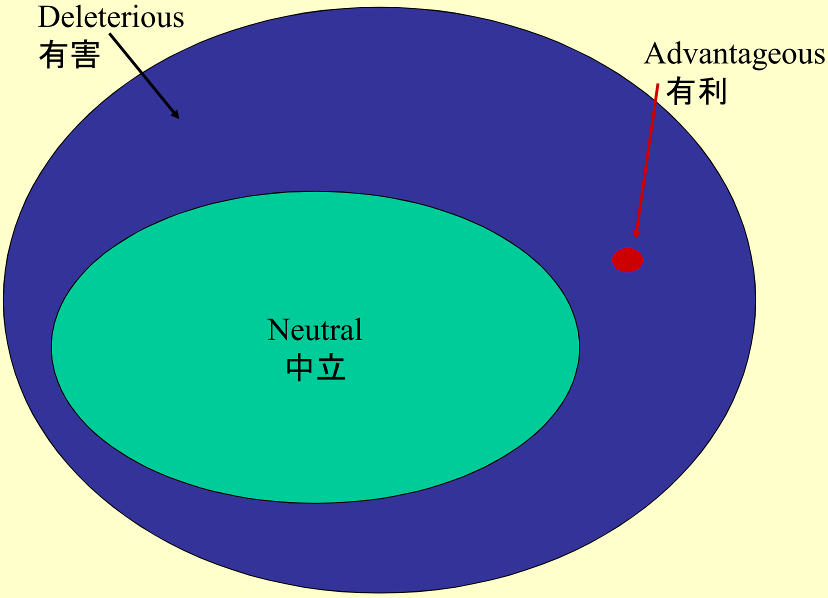
Charles Darwin (1859)

The same adaptive character may coexist in two groups which have a similar mode of life, without indicating any affinity between them, because it may have been acquired by each independently, to enable it to fill a similar place in nature. In such cases it is found to be an almost isolated character, apparently connecting two groups which otherwise differ radically. Non-adaptive, or purely structural characters, on the other hand, are such as have probably been transmitted from a remote ancestor ; and thus indicate fundamental peculiarities of growth and development.

Alfred Russel Wallace (1878)



v = μ (Neutral mutation rate)



Deleterious
有害

Advantageous
有利

Neutral
中立

Structure of Cytochrome *c* and the Rates of Molecular Evolution 37

Dickerson (1971)

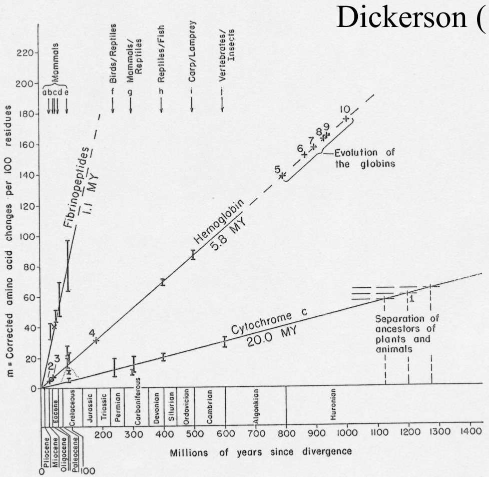
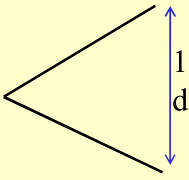


Fig. 3. The rates of macromolecular evolution in the fibrinopeptides, hemoglobin, and cytochrome *c*. The amino acid differences between divergent lines of evolution, corrected for multiple changes at the same locus, are from Table 2. The dating of branch points in evolution is from Table 3. Mean errors in amino acid differences are indicated by vertical bars. Comparisons for which no adequate time coordinate is available are indicated by numbered crosses. Point *1* represents a date of 1200 ± 75 MY (million years) for the separation of plants and animals, based on a linear extrapolation of the cytochrome curve. Points *2–10* refer to events in the development of the globin family, as detailed in Table 2. The δ/β separation is at point *3*, γ/β is at *4*, and α/β is at 500 MY (carp/lamprey). The *Unit Evolutionary Period* in MY is given by each curve. An earlier version of this figure has appeared in Ref. [5]

UEP (Unit of Evolutionary Period): Time required to produce 1% difference



1%
difference

Time = 1UEP

Dickerson (1971)

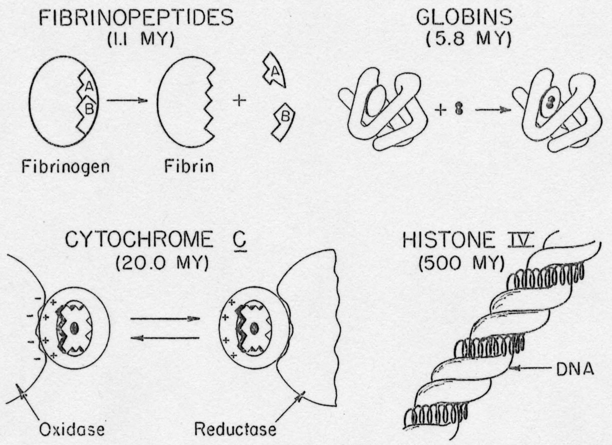
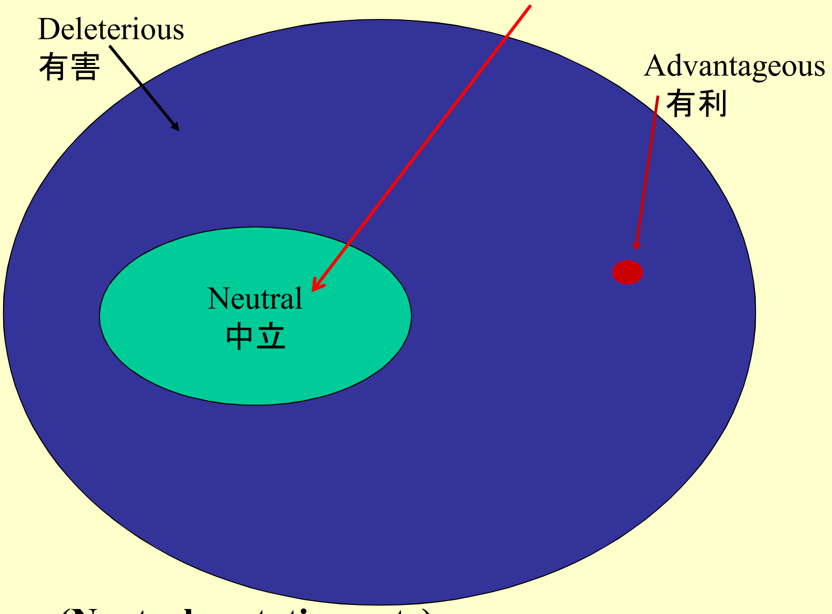


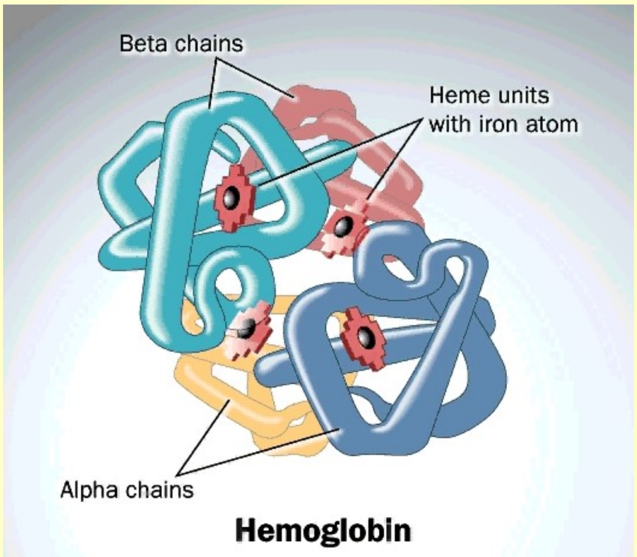
Fig. 4. The more complex the interactions of a protein with other molecules or macromolecules, the longer will be its *Unit Evolutionary Period*. The discarded fibrinopeptides have a UEP of slightly over 1 MY; the Histone IV bound to DNA within the nucleus has a UEP 500 times as long. The UEP for cytochrome *c* is longer than that for the globins primarily because cytochrome *c* interacts with other macromolecular complexes, whereas hemoglobin binds to O_2 and CO_2 in solution

Selective constraint vs Neutral area



v = μ (Neutral mutation rate)

Hemoglobin

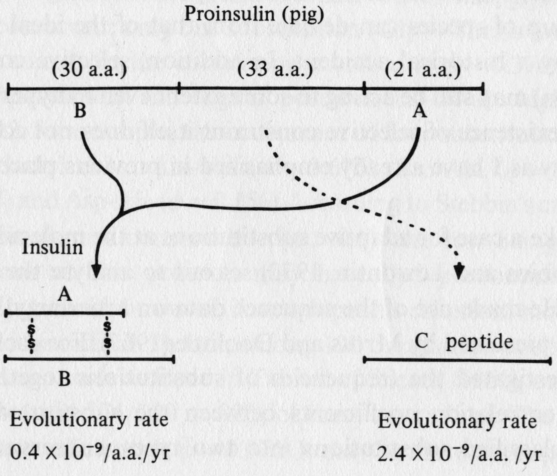


Molecular evolutionary rate of hemoglobin: Surface area vs. Heme pocket

Region	Hemoglobinα	Hemoglobinβ
Surface	1.35 (10^{-9}/year/site)	2.73 (10^{-9}/year/site)
Heme pocket	0.165	0.236

After Kimura and Ohta (1973)

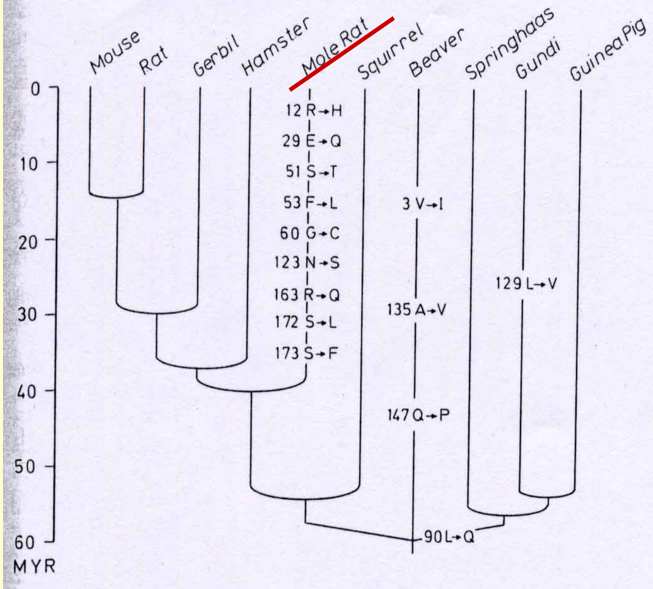
Fig. 7.3. Comparison between the evolutionary rate of insulin (A + B peptides) and that of the middle segment (C peptide) of proinsulin.



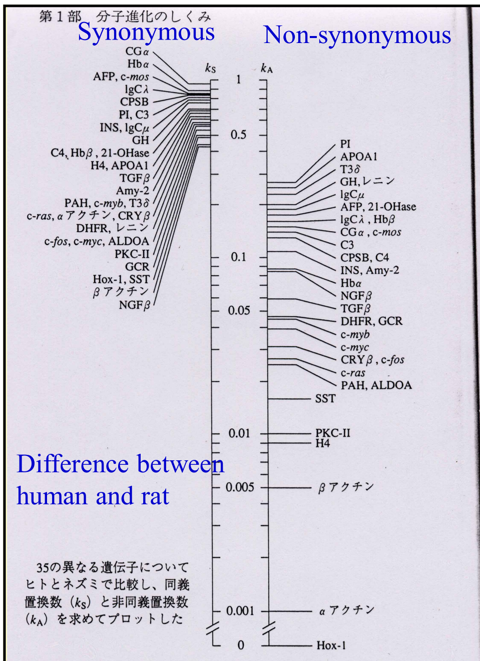
Kimura (1983)

αA-crystallin tree

Evolution: Hendriks *et al.*



第１部　分子進化のしくみ



35の異なる遺伝子についてヒトとネズミで比較し、同義置換数（k_S）と非同義置換数（k_A）を求めてプロットした

図６－３　同義置換速度（左側）と非同義置換速度（右側）の比較

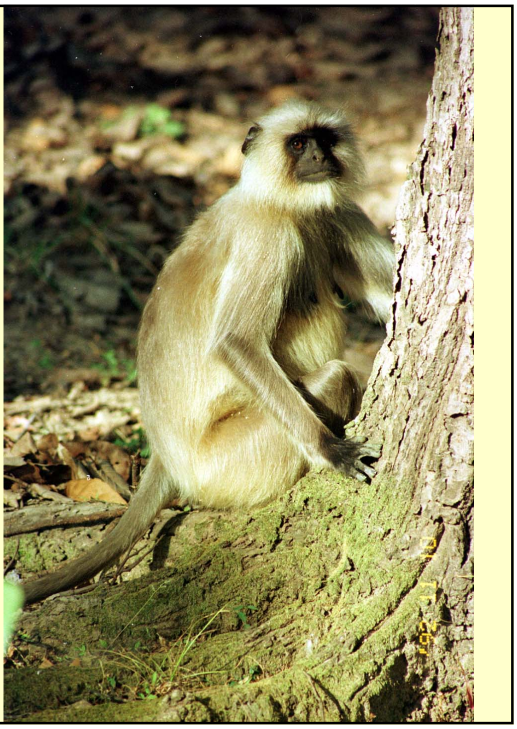
宮田隆「分子進化学への招待」(講談社)

93

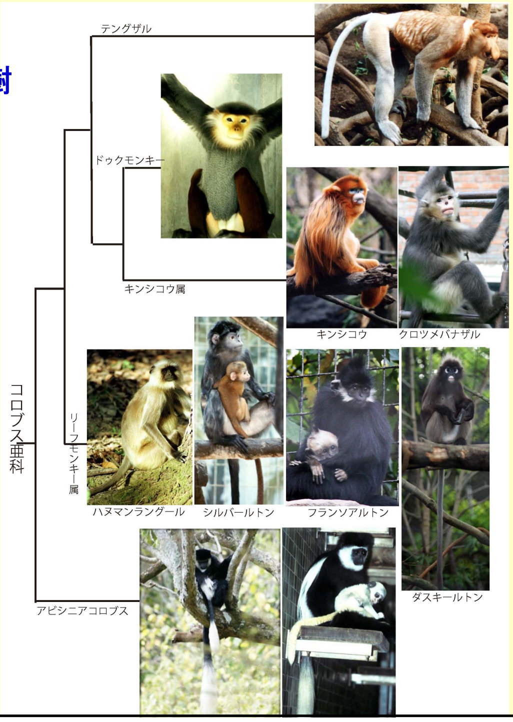
Non-synonymous differences vary very much among different genes because of difference of constraints. On the other hand, synonymous differences do not differ very much, and higher than non-synonymous differences.

Ruminants 反芻類

Hanuman langur (Colobus; Leaf-eating monkey)



コロブス亜科の系統樹
（疣猴亚科）
Colobinae:
Leaf-eating monkeys



Convergent evolution of lysozyme

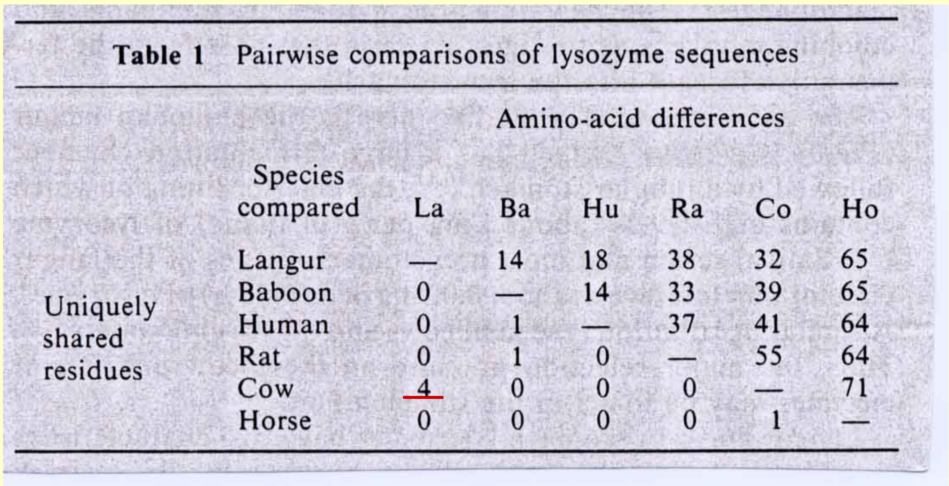


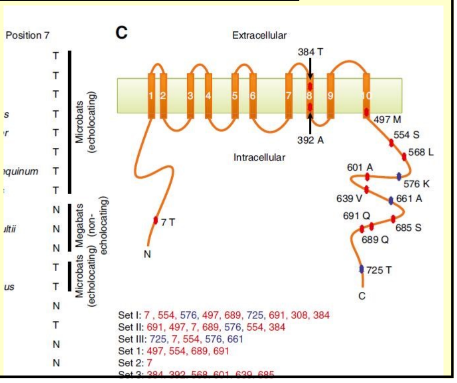
Table 1 Pairwise comparisons of lysozyme sequences

	Species compared	Amino-acid differences La	 Ba	 Hu	 Ra	 Co	 Ho
	Langur	—	14	18	38	32	65
Uniquely shared residues	Baboon	0	—	14	33	39	65
	Human	0	1	—	37	41	64
	Rat	0	1	0	—	55	64
	Cow	4	0	0	0	—	71
	Horse	0	0	0	0	1	—

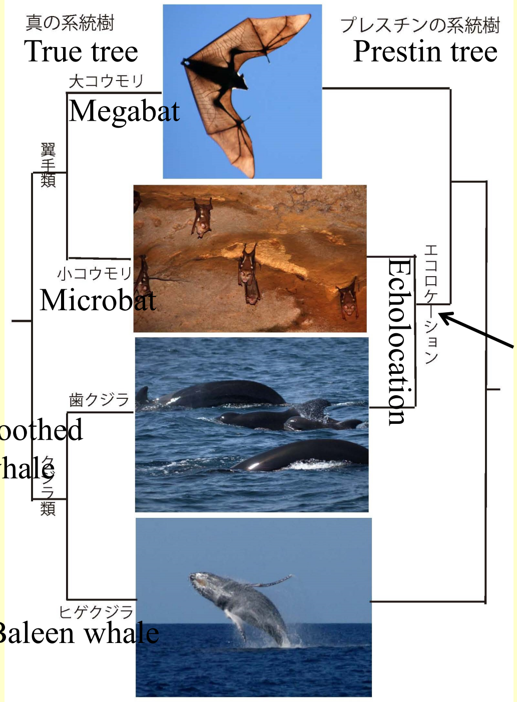
Stewart et al. (1987) Nature 330:401--404

Echolocation of toothed whales and microbats

http://blogs.discovermagazine.com/notrocketscience/2010/01/25/echolocation-in-bats-and-whales-based-on-same-changes-to-same-gene/



Prestin Li et al. (2010)



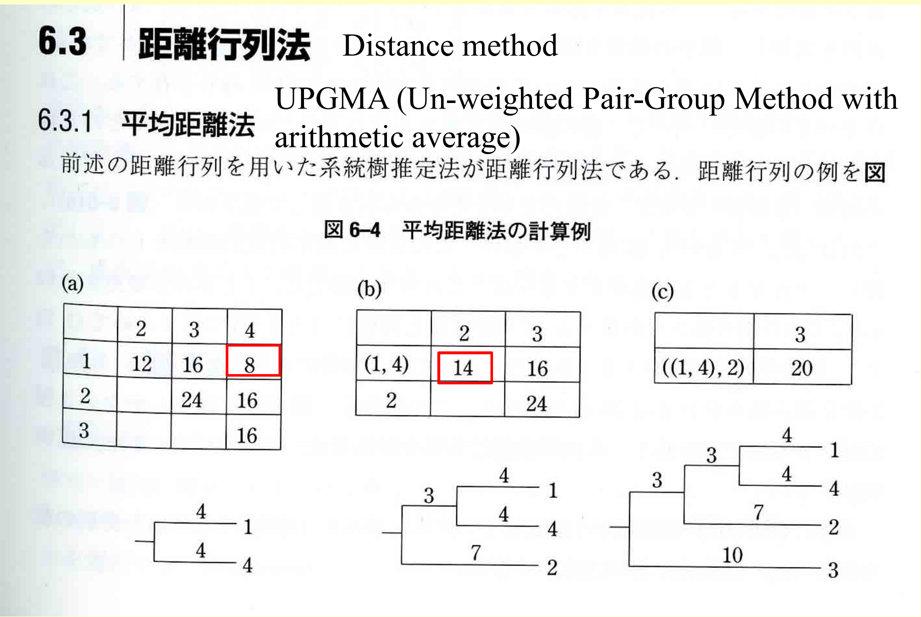
Convergent evolution for echolocation

How to reconstruct a molecular phylogenetic tree?

Comparison of DNA or protein sequences from various organisms.

1. human
2. chimpanzee
3. gorilla
4. orangutan

```
1 CTAGGCTATATACAACTACGCAAAGGCCCCAACGTTGTAGGCCCCTAC
2 CTAGGCTACATACAACTACGCAAAGGTCCCAACATTGTAGGTCCTTAC
3 TTAGGCTATATACAACTACGTAAAGGCCCCAACGTCGTAGGCCCCTAC
4 CTAGGCTATACACAACTACGCAAGGGACCTAACATCGTAGGCCCCTGC
```



6.3 距離行列法 Distance method

6.3.1 平均距離法 UPGMA (Un-weighted Pair-Group Method with arithmetic average)

前述の距離行列を用いた系統樹推定法が距離行列法である．距離行列の例を図

図 6–4　平均距離法の計算例

(a)

	2	3	4
1	12	16	8
2		24	16
3			16

(b)

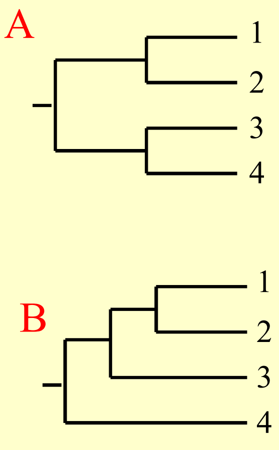
	2	3
(1, 4)	14	16
2		24

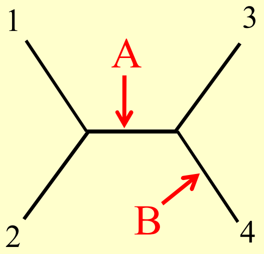
(c)

	3
((1, 4), 2)	20

4
1
4
4
3
4
1
4
4
7
2
3
3
4
1
4
4
7
2
10
3

宮田隆「新しい分子進化学入門」（講談社、2010）

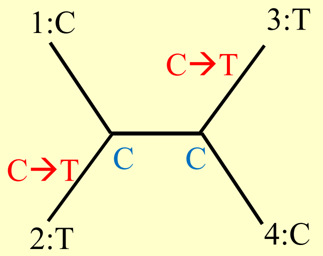


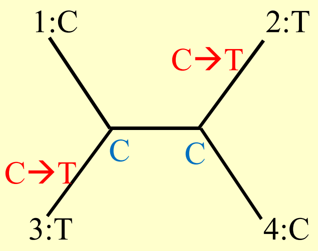


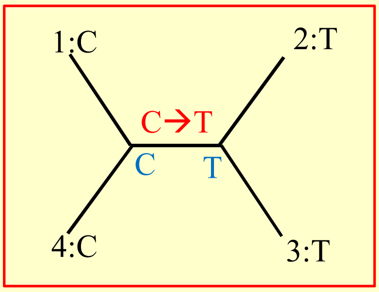
Unrooted tree with 4 OTUs:

When rate constancy is not assumed, the root cannot be determined. In order to root the tree, an outgroup, known to be outside of the ingroup species, is necessary.

最節約法(Maximum Parsimony Method)

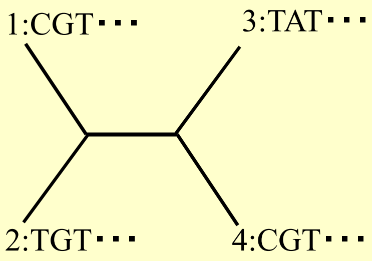


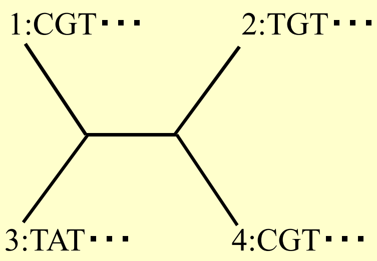


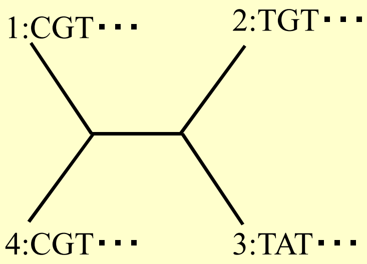


The parsimony method chooses a tree with the minimum number of substitutions.

最節約法(Maximum Parsimony Method)







Choose the tree with the minimum number of substitutions in total of the sequence.

Number of OTUs	Possible number of trees
3	1
4	3
5	3x5=15
6	3x5x7=105
7	3x5x7x9=945
8	3x5x7x9x11=10,395
9	3x5x7x9x11x13=135,135
10	3x5x7x9x11x13x15=2,027,025
22	$3\text{x}10^{23}$
50	$3\text{x}10^{74}$
100	$2\text{x}10^{182}$

近隣結合法 Neighbor-joining method (NJ法)

Saitou and Nei (1987)
Mol. Biol. Evol. 4, 406

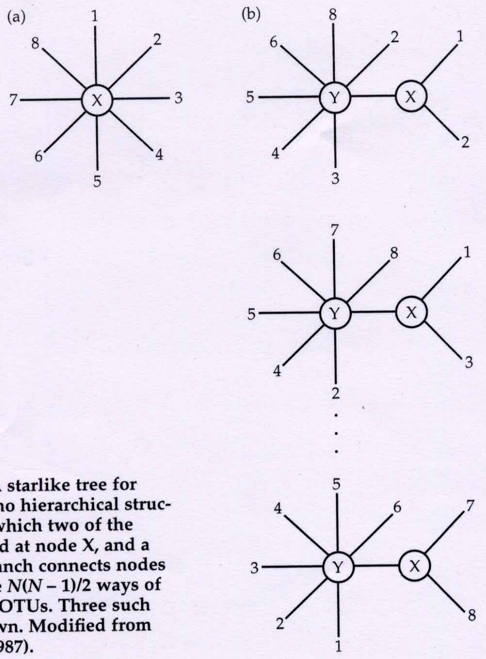
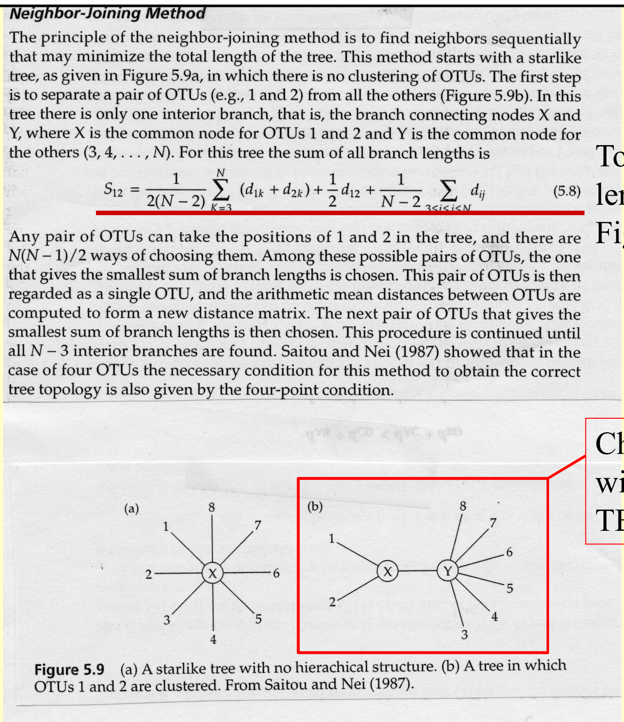


FIGURE 5.13 (a) A starlike tree for eight OTUs with no hierarchical structure. (b) Trees in which two of the OTUs are clustered at node X, and a single internal branch connects nodes X and Y. There are $N(N-1)/2$ ways of choosing pairs of OTUs. Three such examples are shown. Modified from Saitou and Nei (1987).



Neighbor-Joining Method

The principle of the neighbor-joining method is to find neighbors sequentially that may minimize the total length of the tree. This method starts with a starlike tree, as given in Figure 5.9a, in which there is no clustering of OTUs. The first step is to separate a pair of OTUs (e.g., 1 and 2) from all the others (Figure 5.9b). In this tree there is only one interior branch, that is, the branch connecting nodes X and Y, where X is the common node for OTUs 1 and 2 and Y is the common node for the others (3, 4, . . . , *N*). For this tree the sum of all branch lengths is

$$S_{12} = \frac{1}{2(N-2)} \sum_{K=3}^{N} (d_{1k} + d_{2k}) + \frac{1}{2} d_{12} + \frac{1}{N-2} \sum_{3 \le i < j \le N} d_{ij} \qquad (5.8)$$

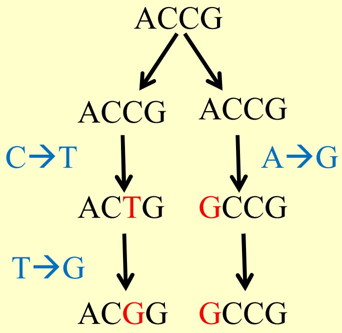
Any pair of OTUs can take the positions of 1 and 2 in the tree, and there are $N(N-1)/2$ ways of choosing them. Among these possible pairs of OTUs, the one that gives the smallest sum of branch lengths is chosen. This pair of OTUs is then regarded as a single OTU, and the arithmetic mean distances between OTUs are computed to form a new distance matrix. The next pair of OTUs that gives the smallest sum of branch lengths is then chosen. This procedure is continued until all $N-3$ interior branches are found. Saitou and Nei (1987) showed that in the case of four OTUs the necessary condition for this method to obtain the correct tree topology is also given by the four-point condition.

Figure 5.9 (a) A starlike tree with no hierachical structure. (b) A tree in which OTUs 1 and 2 are clustered. From Saitou and Nei (1987).

Total branch lengths (TBL) for Fig.5.9(b)

Choose the tree with minimum TBL

Multiple substitutions



While substitutions occurred 3 times, only 2 sites differ because of multiple substitutions in a site. Maximum parsimony does not take account of this. Although the distance methods such as NJ can take account of this to some extent, it is difficult to evaluate the effect of multiple substitutions between distantly related sequences pair-wisely without taking account of ancestral sequences.

Joe Felsenstein (1981)
Evolutionary trees from DNA sequences:
a maximum likelihood approach.
J. Mol. Evol. 17:368-376.

A statistical method for phylogenetic inference based on an explicit model for substitutions during evolution

$$L = P(\text{data}|\text{model})$$

model: substitution model + tree topology

Joe Felsenstein in 1998 at ISM

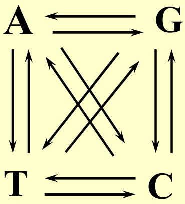
Maximum Likelihood Method（最尤法、似然法）

Likelihood $L = P(\text{data}|\text{model})$

Likelihood is the probability of realizing the data under the given evolutionary model.

Model: substitution model+ tree topology

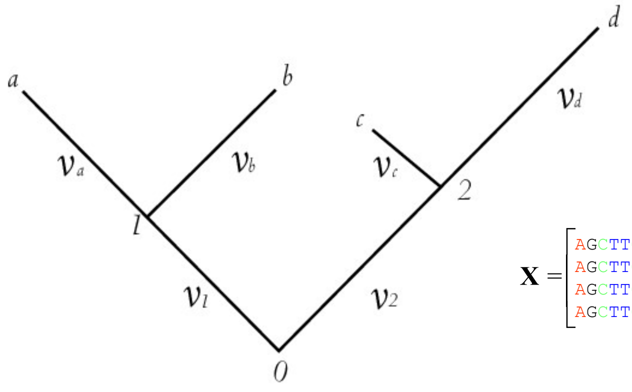
1. Human
2. Chimp
3. Gorilla
4. Orang



Model for nucleotide substitutions

```
1  CTAGGCTATATACAACTACGCAAAGGCCCCAACGTTGTAGGCCCCTAC
2  CTAGGCTACATACAACTACGCAAAGGTCCCAACATTGTAGGTCCTTAC
3  TTAGGCTATATACAACTACGTAAAGGCCCCAACGTCGTAGGCCCCTAC
4  CTAGGCTATACACAACTACGCAAGGGACCTAACATCGTAGGCCCCTGC
```

Likelihood function



$$\mathbf{X} = \begin{bmatrix} \texttt{AGCTTCACC GGCGCAGTCA TTCTCATAAT} \\ \texttt{AGCTTCACC GGCGCAATTA TCCTCATAAT} \\ \texttt{AGCTTCACC GGCGCAGTTG TTCTTATAAT} \\ \texttt{AGCTTCACC GGCGCAACCA CCCTCATGAT} \end{bmatrix}$$

$$L = \sum_{s_0}\sum_{s_1}\sum_{s_2} \pi_{s_0} P_{s_0 s_1}(v_1) P_{s_1 s_a}(v_a) P_{s_1 s_b}(v_b) P_{s_0 s_2}(v_2) P_{s_2 s_c}(v_c) P_{s_2 s_d}(v_d)$$

P: transition probability, π: base composition

Markov model of nucleotide (or amino acid) substitution

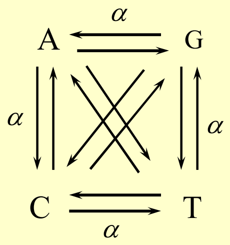
Transition probability matrix P(*t*) during time *t*

$$\mathbf{P}\,(t) = e^{\mathbf{Q}t}$$

where Q is instantaneous rate matrix during infinitesimal time interval *dt*

$$\mathbf{P}\,(dt) = \mathbf{1} + \mathbf{Q}\,dt$$

Poisson model
(Jukes and Cantor model)

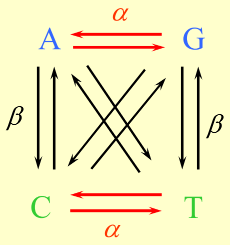


$$\mathbf{Q} = \begin{array}{c|cccc} \nearrow & T & C & A & G \\ \hline T & -3\alpha & \alpha & \alpha & \alpha \\ C & \alpha & -3\alpha & \alpha & \alpha \\ A & \alpha & \alpha & -3\alpha & \alpha \\ G & \alpha & \alpha & \alpha & -3\alpha \end{array}$$

Fig. 1 - Jukes-Cantor model.

$$\mathbf{P}(dt) = \begin{array}{c|cccc} & T & C & A & G \\ \hline T & 1\text{-}3\alpha dt & \alpha dt & \alpha dt & \alpha dt \\ C & \alpha dt & 1\text{-}3\alpha dt & \alpha dt & \alpha dt \\ A & \alpha dt & \alpha dt & 1\text{-}3\alpha dt & \alpha dt \\ G & \alpha dt & \alpha dt & \alpha dt & 1\text{-}3\alpha dt \end{array}$$

Kimura 2-parameter model
(Kimura, 1980)



$$\mathbf{Q} = \begin{array}{c|cccc} \nearrow & T & C & A & G \\ \hline T & -(\alpha+2\beta) & \alpha & \beta & \beta \\ C & \alpha & -(\alpha+2\beta) & \beta & \beta \\ A & \beta & \beta & -(\alpha+2\beta) & \alpha \\ G & \beta & \beta & \alpha & -(\alpha+2\beta) \end{array}$$

$$\mathbf{P}(dt) = \begin{array}{c|cccc} & T & C & A & G \\ \hline T & 1-(\alpha+2\beta)dt & \alpha dt & \beta dt & \beta dt \\ C & \alpha dt & 1-(\alpha+2\beta)dt & \beta dt & \beta dt \\ A & \beta dt & \beta dt & 1-(\alpha+2\beta)dt & \alpha dt \\ G & \beta dt & \beta dt & \alpha dt & 1-(\alpha+2\beta)dt \end{array}$$

— transversion

— transition

purines

pyrimidines

Motoo Kimura
(1924—1994)

Hasegawa, Kishino and Yano (1985) model (HKY model)

$$Q=\begin{bmatrix} -(\alpha\pi_C+\beta\pi_R) & \alpha\pi_C & \beta\pi_A & \beta\pi_G \\ \alpha\pi_T & -(\alpha\pi_T+\beta\pi_R) & \beta\pi_A & \beta\pi_G \\ \beta\pi_T & \beta\pi_C & -(\alpha\pi_G+\beta\pi_Y) & \alpha\pi_G \\ \beta\pi_T & \beta\pi_C & \alpha\pi_A & -(\alpha\pi_A+\beta\pi_Y) \end{bmatrix}$$

$$P(dt)=\begin{bmatrix} 1-(\alpha\pi_C+\beta\pi_R)dt & \alpha\pi_C dt & \beta\pi_A dt & \beta\pi_G dt \\ \alpha\pi_T dt & 1-(\alpha\pi_T+\beta\pi_R)dt & \beta\pi_A dt & \beta\pi_G dt \\ \beta\pi_T dt & \beta\pi_C dt & 1-(\alpha\pi_G+\beta\pi_Y)dt & \alpha\pi_G dt \\ \beta\pi_T dt & \beta\pi_C dt & \alpha\pi_A dt & 1-(\alpha\pi_A+\beta\pi_Y)dt \end{bmatrix}$$

where $\pi_R=\pi_A+\pi_G$, $\pi_Y=\pi_T+\pi_C$

Highly biased nucleotide frequencies of mammalian mtDNA

3rd codon positions: $\pi_T=0.169$, π_C=0.429, $\pi_A=0.364$, π_G=0.038

General time-reversible model (GTR model)

$$Q=\begin{bmatrix} -\mu(a\pi_C+b\pi_A+c\pi_G) & \mu a\pi_C & \mu b\pi_A & \mu c\pi_G \\ \mu a\pi_T & -\mu(a\pi_T+d\pi_A+e\pi_G) & \mu d\pi_A & \mu e\pi_G \\ \mu b\pi_T & \mu d\pi_C & -\mu(b\pi_T+d\pi_c+f\pi_G) & \mu f\pi_G \\ \mu c\pi_T & \mu e\pi_C & \mu f\pi_A & -\mu(c\pi_T+e\pi_c+f\pi_A) \end{bmatrix}$$

$$P(dt)=\begin{bmatrix} 1-\mu(a\pi_C+b\pi_A+c\pi_G)dt & \mu a\pi_C dt & \mu b\pi_A dt & \mu c\pi_G dt \\ \mu a\pi_T dt & 1-\mu(a\pi_T+d\pi_A+e\pi_G)dt & \mu d\pi_A dt & \mu e\pi_G dt \\ \mu b\pi_T dt & \mu d\pi_C dt & 1-\mu(b\pi_T+d\pi_C+f\pi_G)dt & \mu f\pi_G dt \\ \mu c\pi_T dt & \mu e\pi_C dt & \mu f\pi_A dt & 1-\mu(c\pi_T+e\pi_C+f\pi_A)dt \end{bmatrix}$$

The most general model with time-reversibility

$$\pi_i Q_{ij} = \pi_j Q_{ji}$$

Heterogeneity among sites

- **Partition among different categories of sites**
- **Taking account of invariable sites → Later improved with the discrete G-distribution model by Ziheng Yang**

Neglect of these factors gives gross underestimation of the number of nucleotide substitutions, and accordingly an older estimation of the date when calibration is taken at a deeper node.

Amino acid substitution model (Empirical matrix)

- Dayhoff (1972) model
- JTT model (Jones, Taylor and Thornton, 1992)
- mtREV model (Adachi and Hasegawa, 1996)
- cpREV model (Adachi, Waddell, Martin, and Hasegawa, 2000)
- WAG model (Whelan and Goldman, 2001)

Amino acid substitution of proteins encoded by nuclear genome for the time period of 1 substitution per 100 amino acids

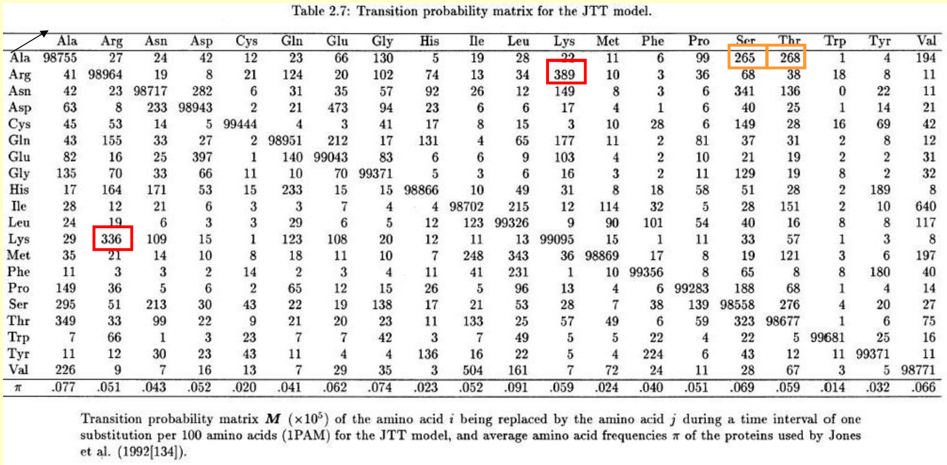


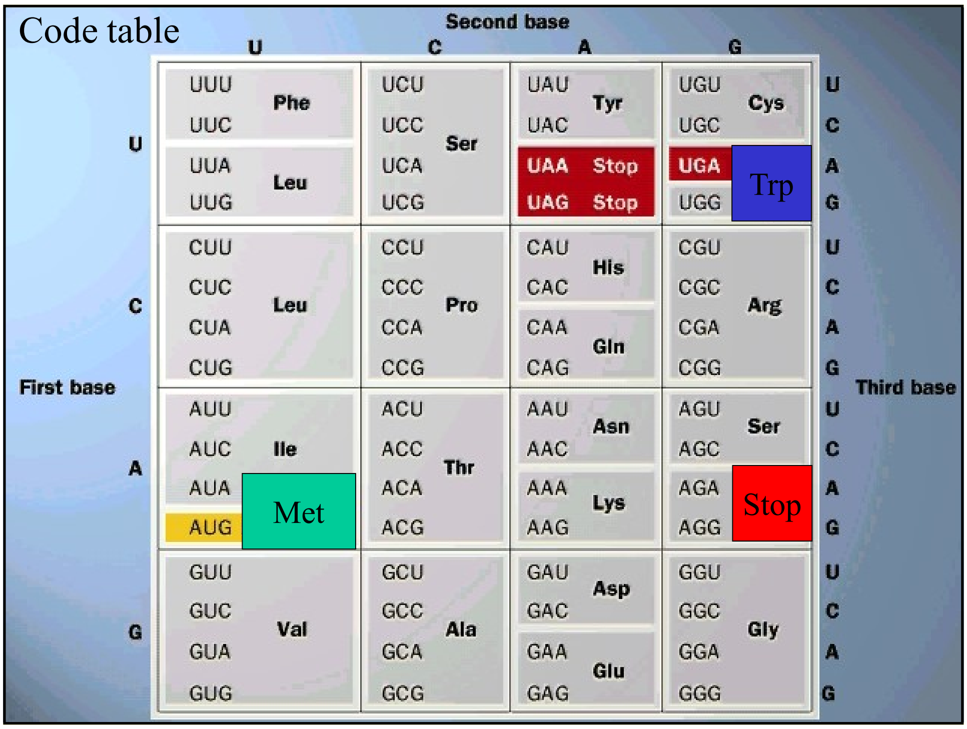
Table 2.7: Transition probability matrix for the JTT model.

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	98755	27	24	42	12	23	66	130	5	19	28	22	11	6	99	265	268	1	4	194
Arg	41	98964	19	8	21	124	20	102	74	13	34	389	10	3	36	68	38	18	8	11
Asn	42	23	98717	282	6	31	35	57	92	26	12	149	8	3	6	341	136	0	22	11
Asp	63	8	233	98943	2	21	473	94	23	6	6	17	4	1	6	40	25	1	14	21
Cys	45	53	14	5	99444	4	3	41	17	8	15	3	10	28	6	149	28	16	69	42
Gln	43	155	33	27	2	98951	212	17	131	4	65	177	11	2	81	37	31	2	8	12
Glu	82	16	25	397	1	140	99043	83	6	6	9	103	4	2	10	21	19	2	2	31
Gly	135	70	33	66	11	10	70	99371	5	3	6	16	3	2	11	129	19	8	2	32
His	17	164	171	53	15	233	15	15	98866	10	49	31	8	18	58	51	28	2	189	8
Ile	28	12	21	6	3	3	7	4	4	98702	215	12	114	32	5	28	151	2	10	640
Leu	24	19	6	3	3	29	6	5	12	123	99326	9	90	101	54	40	16	8	8	117
Lys	29	336	109	15	1	123	108	20	12	11	13	99095	15	1	11	33	57	1	3	8
Met	35	21	14	10	8	18	11	10	7	248	343	36	98869	17	8	19	121	3	6	197
Phe	11	3	3	2	14	2	3	4	11	41	231	1	10	99356	8	65	8	8	180	40
Pro	149	36	5	6	2	65	12	15	26	5	96	13	4	6	99283	188	68	1	4	14
Ser	295	51	213	30	43	22	19	138	17	21	53	28	7	38	139	98558	276	4	20	27
Thr	349	33	99	22	9	21	20	23	11	133	25	57	49	6	59	323	98677	1	6	75
Trp	7	66	1	3	23	7	7	42	3	7	49	5	5	22	4	22	5	99681	25	16
Tyr	11	12	30	23	43	11	4	4	136	16	22	5	4	224	6	43	12	11	99371	11
Val	226	9	7	16	13	7	29	35	3	504	161	7	72	24	11	28	67	3	5	98771
π	.077	.051	.043	.052	.020	.041	.062	.074	.023	.052	.091	.059	.024	.040	.051	.069	.059	.014	.032	.066

Transition probability matrix $\boldsymbol{M}$ ($\times 10^5$) of the amino acid i being replaced by the amino acid j during a time interval of one substitution per 100 amino acids (1PAM) for the JTT model, and average amino acid frequencies π of the proteins used by Jones et al. (1992[134]).

The substitutions with red squares occur frequently because of similar physico-chemical properties.

Code table



Second base

First base	U		C		A		G		Third base
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
	UUC		UCC		UAC		UGC		C
	UUA	Leu	UCA		UAA	Stop	UGA	Trp	A
	UUG		UCG		UAG	Stop	UGG		G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Gln	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U
	AUC		ACC		AAC		AGC		C
	AUA	Met	ACA		AAA	Lys	AGA	Stop	A
	AUG		ACG		AAG		AGG		G
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	Glu	GGA		A
	GUG		GCG		GAG		GGG		G

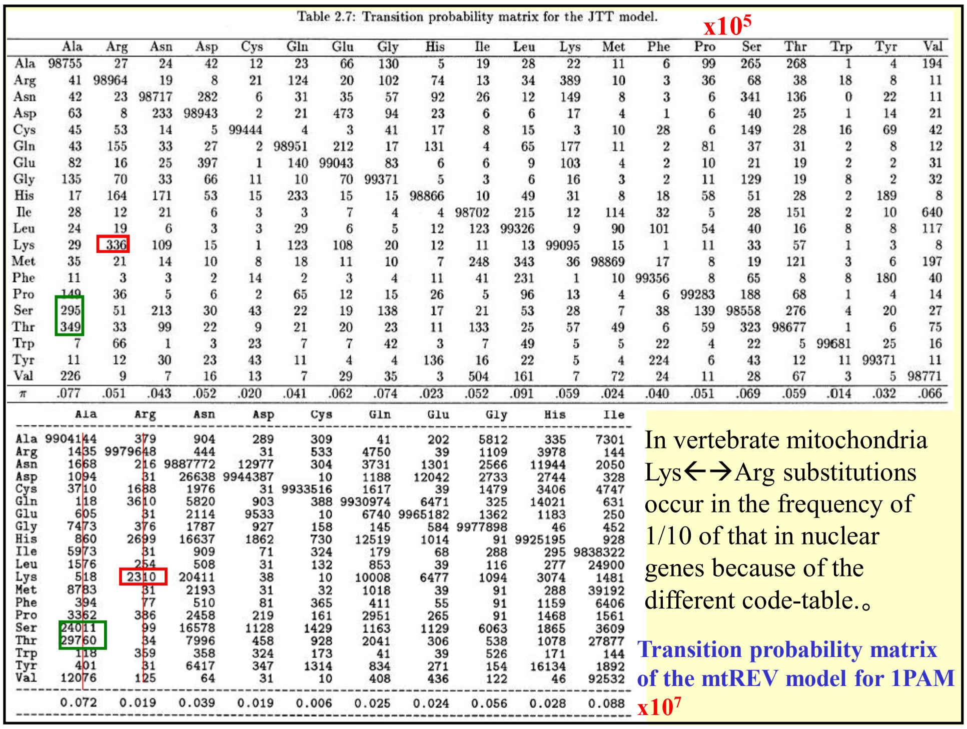


Table 2.7: Transition probability matrix for the JTT model.

x10^5

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	98755	27	24	42	12	23	66	130	5	19	28	22	11	6	99	265	268	1	4	194
Arg	41	98964	19	8	21	124	20	102	74	13	34	389	10	3	36	68	38	18	8	11
Asn	42	23	98717	282	6	31	35	57	92	26	12	149	8	3	6	341	136	0	22	11
Asp	63	8	233	98943	2	21	473	94	23	6	6	17	4	1	6	40	25	1	14	21
Cys	45	53	14	5	99444	4	3	41	17	8	15	3	10	28	6	149	28	16	69	42
Gln	43	155	33	27	2	98951	212	17	131	4	65	177	11	2	81	37	31	2	8	12
Glu	82	16	25	397	1	140	99043	83	6	6	9	103	4	2	10	21	19	2	2	31
Gly	135	70	33	66	11	10	70	99371	5	3	6	16	3	2	11	129	19	8	2	32
His	17	164	171	53	15	233	15	15	98866	10	49	31	8	18	58	51	28	2	189	8
Ile	28	12	21	6	3	3	7	4	4	98702	215	12	114	32	5	28	151	2	10	640
Leu	24	19	6	3	3	29	6	5	12	123	99326	9	90	101	54	40	16	8	8	117
Lys	29	336	109	15	1	123	108	20	12	11	13	99095	15	1	11	33	57	1	3	8
Met	35	21	14	10	8	18	11	10	7	248	343	36	98869	17	8	19	121	3	6	197
Phe	11	3	3	2	14	2	3	4	11	41	231	1	10	99356	8	65	8	8	180	40
Pro	149	36	5	6	2	65	12	15	26	5	96	13	4	6	99283	188	68	1	4	14
Ser	295	51	213	30	43	22	19	138	17	21	53	28	7	38	139	98558	276	4	20	27
Thr	349	33	99	22	9	21	20	23	11	133	25	57	49	6	59	323	98677	1	6	75
Trp	7	66	1	3	23	7	7	42	3	7	49	5	5	22	4	22	5	99681	25	16
Tyr	11	12	30	23	43	11	4	4	136	16	22	5	4	224	6	43	12	11	99371	11
Val	226	9	7	16	13	7	29	35	3	504	161	7	72	24	11	28	67	3	5	98771
π	.077	.051	.043	.052	.020	.041	.062	.074	.023	.052	.091	.059	.024	.040	.051	.069	.059	.014	.032	.066

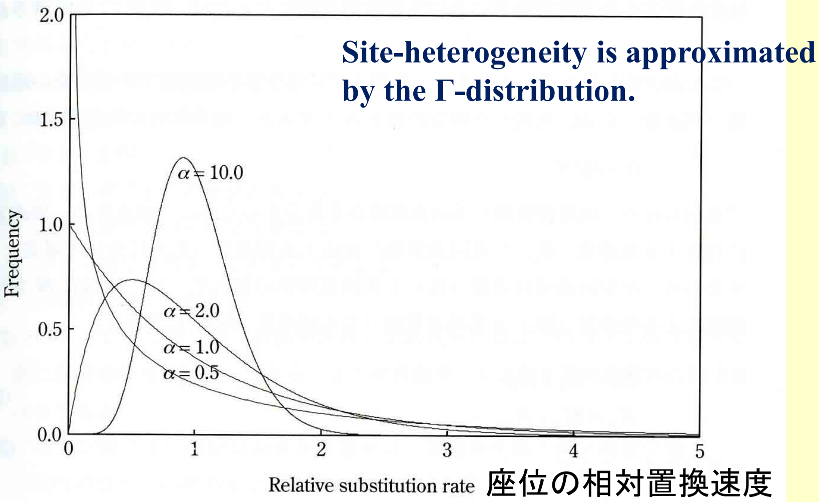
	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile
Ala	9904144	379	904	289	309	41	202	5812	335	7301
Arg	1435	9979648	444	31	533	4750	39	1109	3978	144
Asn	1668	216	9887772	12977	304	3731	1301	2566	11944	2050
Asp	1094	31	26638	9944387	10	1188	12042	2733	2744	328
Cys	3710	1688	1976	31	9933516	1617	39	1479	3406	4747
Gln	118	3610	5820	903	388	9930974	6471	325	14021	631
Glu	605	31	2114	9533	10	6740	9965182	1362	1183	250
Gly	7473	376	1787	927	158	145	584	9977898	46	452
His	860	2699	16637	1862	730	12519	1014	91	9925195	928
Ile	5973	31	909	71	324	179	68	288	295	9838322
Leu	1576	254	508	31	132	853	39	116	277	24900
Lys	518	2310	20411	38	10	10008	6477	1094	3074	1481
Met	8783	31	2193	31	32	1018	39	91	288	39192
Phe	394	77	510	81	365	411	55	91	1159	6406
Pro	3362	386	2458	219	161	2951	265	91	1468	1561
Ser	24011	99	16578	1128	1429	1163	1129	6063	1865	3609
Thr	29760	34	7996	458	928	2041	306	538	1078	27877
Trp	118	359	358	324	173	41	39	526	171	144
Tyr	401	31	6417	347	1314	834	271	154	16134	1892
Val	12076	125	64	31	10	408	436	122	46	92532
	0.072	0.019	0.039	0.019	0.006	0.025	0.024	0.056	0.028	0.088

In vertebrate mitochondria Lys←→Arg substitutions occur in the frequency of 1/10 of that in nuclear genes because of the different code-table。

Transition probability matrix of the mtREV model for 1PAM

x10^7

図 5–4　Γ分布　$f(x) = \beta^{\alpha}\Gamma(\alpha)^{-1}e^{-\beta x}x^{\alpha-1}$ $(x > 0)$



Site-heterogeneity is approximated by the Γ-distribution.

Γ(α) はΓ関数である．α = β の特別な場合のみを用いる．この場合，平均 $E(x) = 1$，分散 $\mathrm{Var}(x) = 1/\alpha$ である．

宮田隆「新しい分子進化学入門」（講談社、2010）

Akaike Information Criterion (AIC) for model selection (Akaike, 1973)

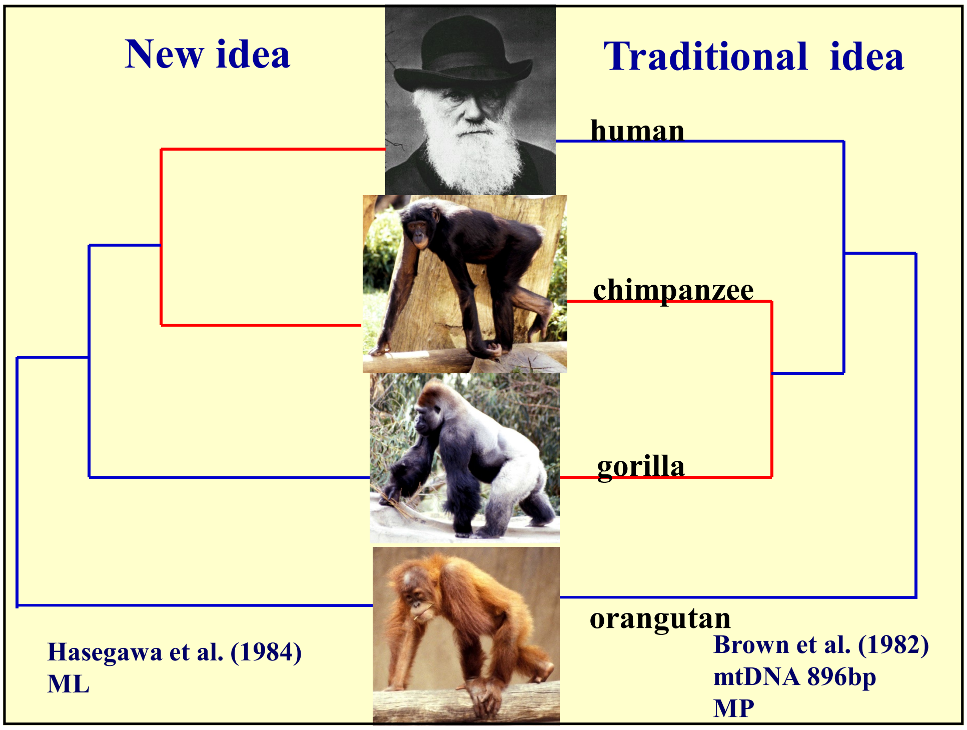
AIC = - 2xln L + 2x #parameters

The better the fitting of the model to the data, the lower is the 1st term. The more complex is the model, the higher is the 2nd term. A model which minimizes the AIC is considered to be the most appropriate model. This implies that, when there are alternative models whose values of ln L are nearly the same, we should choose the one with the smallest number of parameters.

Hirotsugu Akaike

New idea

Traditional idea



Hasegawa et al. (1984)
ML

Brown et al. (1982)
mtDNA 896bp
MP

J Mol Evol (1982) 18:225–239

Journal of Molecular Evolution
© Springer-Verlag 1982

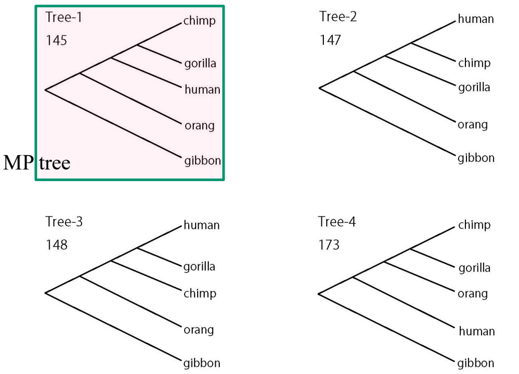
mtDNA 896bp
Portions of NADH4 & NADH5 genes, and 3 tRNA genes (His, Ser & Leu)

Mitochondrial DNA Sequences of Primates: Tempo and Mode of Evolution

Wesley M. Brown[1], Ellen M. Prager, Alice Wang[2], and Allan C. Wilson

Department of Biochemistry, University of California, Berkeley, California 947

The MP method prefers Tree-1.



390 M. HASEGAWA and T. YANO [Vol. 60(B),

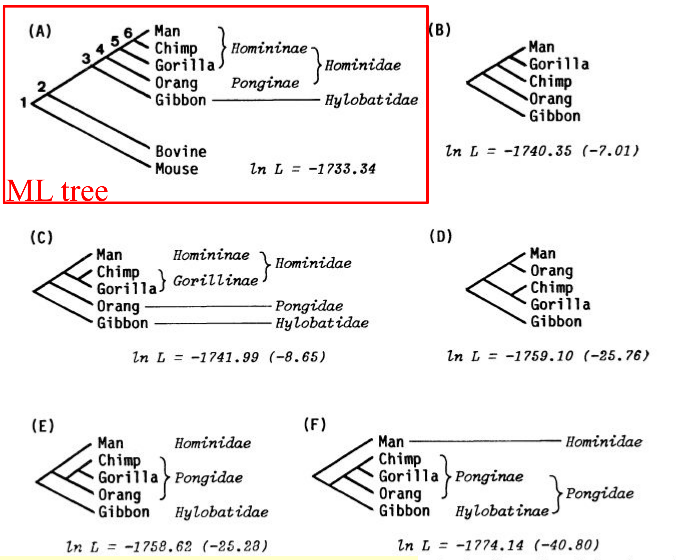


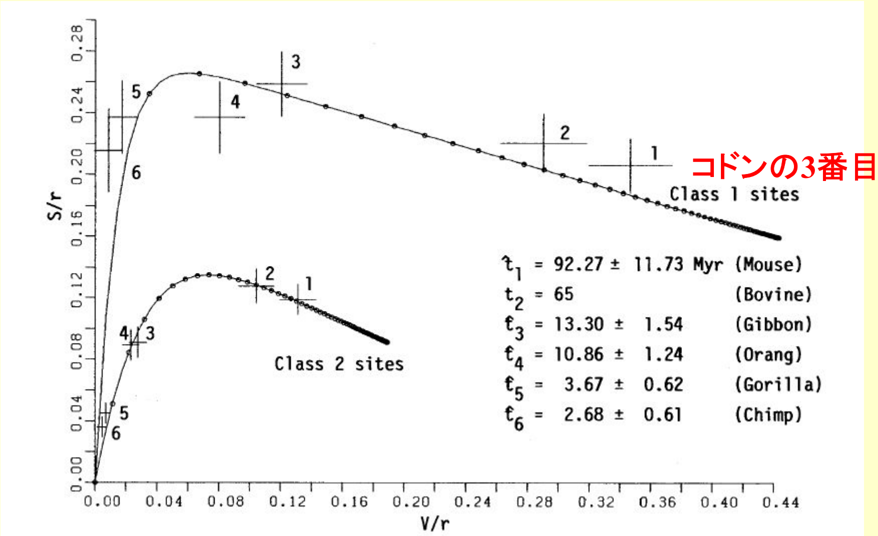
Fig. 1. Alternative phylogenies and classifications of Hominoidea. *ln L* indicates natural logarithmic likelihood of the respective tree of mtDNA including the five species of Hominoidea, bovine, and mouse. Number in parenthesis indicates the difference of *ln L* from that of the maximum likelihood tree (A). Traditional classifications are (E)[25] and (F)[26]. Schwartz proposed the tree (D)[24]. Andrews and Cronin proposed (A) and (C)[23]. Our conclusion is (A).

Why MP and ML method gave different tree?

Transition (upper-half) & transversion (lower-half) differences of 3rd codon positions of mtDNA (Brown et al., 1982)

	mouse	bovine	gibbon	orang	gorilla	chimp	human
mouse		39	53	48	46	50	51
bovine	82		42	44	52	61	57
gibbon	83	71		59	59	64	58
orang	85	65	34		52	60	53
gorilla	77	67	26	18		58	52
chimp	79	67	26	18	4		50
human	77	67	26	20	4	2	

The transition differences of 3rd codon positions do not differ between uman/chimp and human/mouse comparisons→ Multiple transition-type substitutions

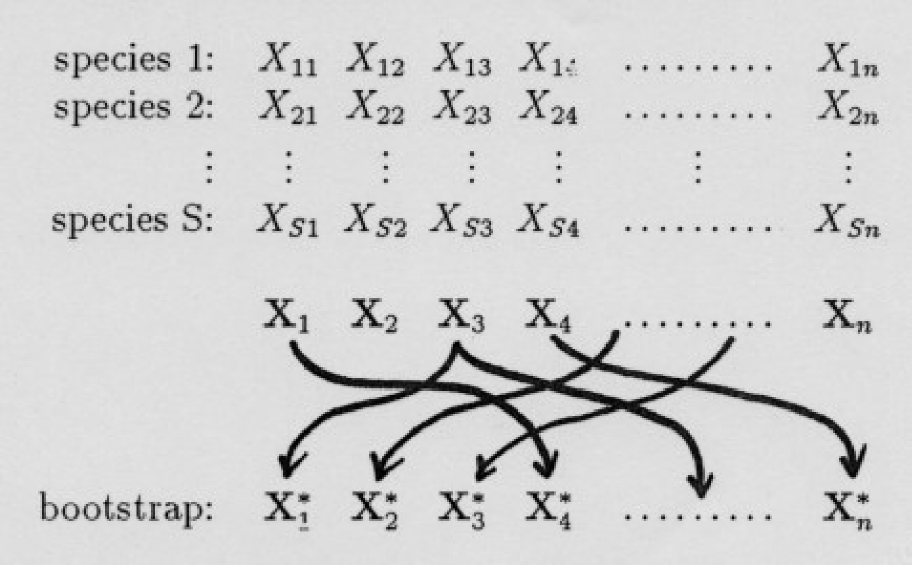


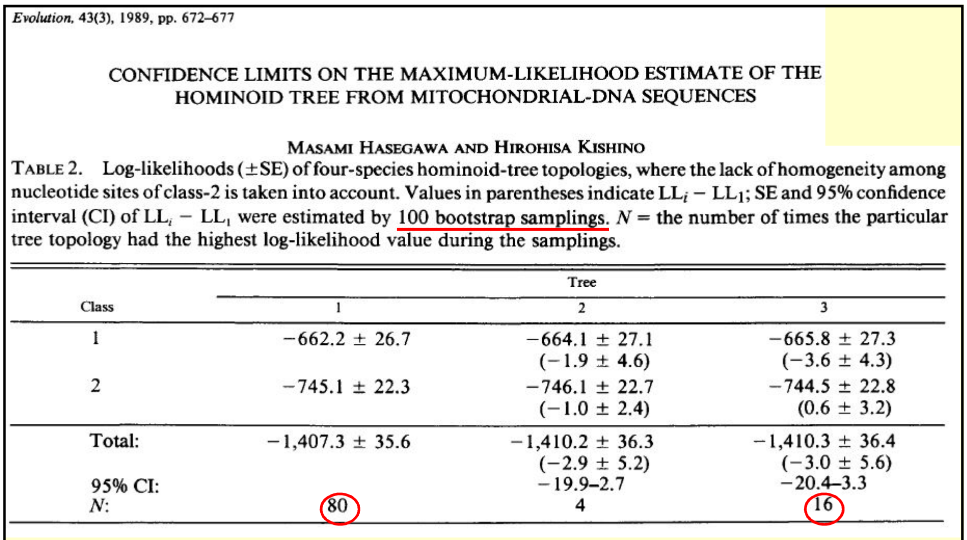
コドンの3番目

V: transversion difference, S: transition difference, r: number of sites. Class 1 sites: 3rd codon positions, Class 2 sites: other sites. 1: Mouse, 2: Bovine, 3: Gibbon, 4: Orang-utan, 5: Gorilla, 6: Chimpanzee (Hasegawa, Kishino & Yano, 1985)

Bootstrap method (Felsenstein, 1985)

(系統樹推定の誤差評価)


species 1: X_{11} X_{12} X_{13} X_{14} .... X_{1n}
species 2: X_{21} X_{22} X_{23} X_{24} .... X_{2n}
species S: X_{S1} X_{S2} X_{S3} X_{S4} .... X_{Sn}
$\mathbf{X}_1$ $\mathbf{X}_2$ $\mathbf{X}_3$ $\mathbf{X}_4$ .... $\mathbf{X}_n$
bootstrap: $\mathbf{X}^*_1$ $\mathbf{X}^*_2$ $\mathbf{X}^*_3$ $\mathbf{X}^*_4$ .... $\mathbf{X}^*_n$



Evolution, 43(3), 1989, pp. 672–677

CONFIDENCE LIMITS ON THE MAXIMUM-LIKELIHOOD ESTIMATE OF THE HOMINOID TREE FROM MITOCHONDRIAL-DNA SEQUENCES

MASAMI HASEGAWA AND HIROHISA KISHINO

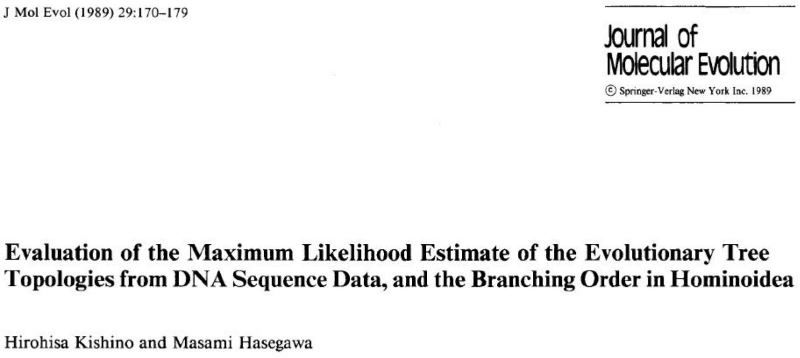
TABLE 2. Log-likelihoods (±SE) of four-species hominoid-tree topologies, where the lack of homogeneity among nucleotide sites of class-2 is taken into account. Values in parentheses indicate $LL_i - LL_1$; SE and 95% confidence interval (CI) of $LL_i - LL_1$ were estimated by 100 bootstrap samplings. N = the number of times the particular tree topology had the highest log-likelihood value during the samplings.

	Tree		
Class	1	2	3
1	−662.2 ± 26.7	−664.1 ± 27.1 (−1.9 ± 4.6)	−665.8 ± 27.3 (−3.6 ± 4.3)
2	−745.1 ± 22.3	−746.1 ± 22.7 (−1.0 ± 2.4)	−744.5 ± 22.8 (0.6 ± 3.2)
Total:	−1,407.3 ± 35.6	−1,410.2 ± 36.3 (−2.9 ± 5.2)	−1,410.3 ± 36.4 (−3.0 ± 5.6)
95% CI:		−19.9–2.7	−20.4–3.3
N:	80	4	16

Tree-1 is the ML tree, but Tree-3 with 16%BP cannot be excluded.

Later Horai et al. (1995) established Tree-1 with the whole mitgenome sequences.

Tree-1:((Human,Chimp),Gorilla)
Tree-2:((Human,Gorilla),Chimp)
Tree-3:((Chimp,Gorilla),Human)



J Mol Evol (1989) 29:170–179

Journal of Molecular Evolution

© Springer-Verlag New York Inc. 1989

Evaluation of the Maximum Likelihood Estimate of the Evolutionary Tree Topologies from DNA Sequence Data, and the Branching Order in Hominoidea

Hirohisa Kishino and Masami Hasegawa

$$l_{(i)}(\hat{\theta}_{(i)} \,|\, \mathbf{X}) = \sum_{h=1}^{n} \log f_{(i)}(\mathbf{X}_h \,|\, \hat{\theta}_{(i)})$$

$$\begin{aligned}\hat{V} &\equiv \hat{\mathrm{V}}\mathrm{ar}[l_{(2)}(\hat{\theta}_{(2)} \,|\, \mathbf{X}) - l_{(1)}(\hat{\theta}_{(1)} \,|\, \mathbf{X})] \\ &= \frac{n}{n-1} \sum_{h=1}^{n} \left\{ \log \frac{f_{(2)}(\mathbf{X}_h \,|\, \hat{\theta}_{(2)})}{f_{(1)}(\mathbf{X}_h \,|\, \hat{\theta}_{(1)})} \right. \\ &\qquad \left. - \frac{1}{n} \sum_{h'=1}^{n} \log \frac{f_{(2)}(\mathbf{X}_{h'} \,|\, \hat{\theta}_{(2)})}{f_{(1)}(\mathbf{X}_{h'} \,|\, \hat{\theta}_{(1)})} \right\}^2\end{aligned}$$

X:Sequence data
θ:Parameters of the model