Molecular Evolution

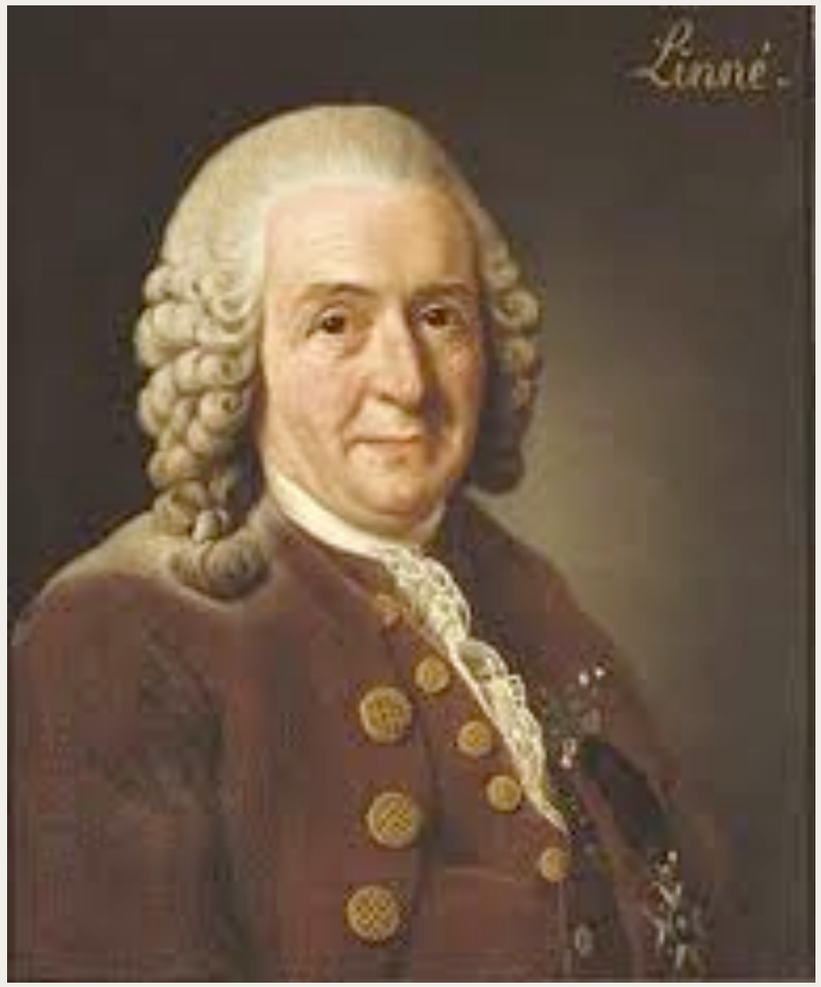
Malay K Basu
Department of Pathology
malay@uab.edu

Ad Verecundiam

Because Dobzhansky says so.

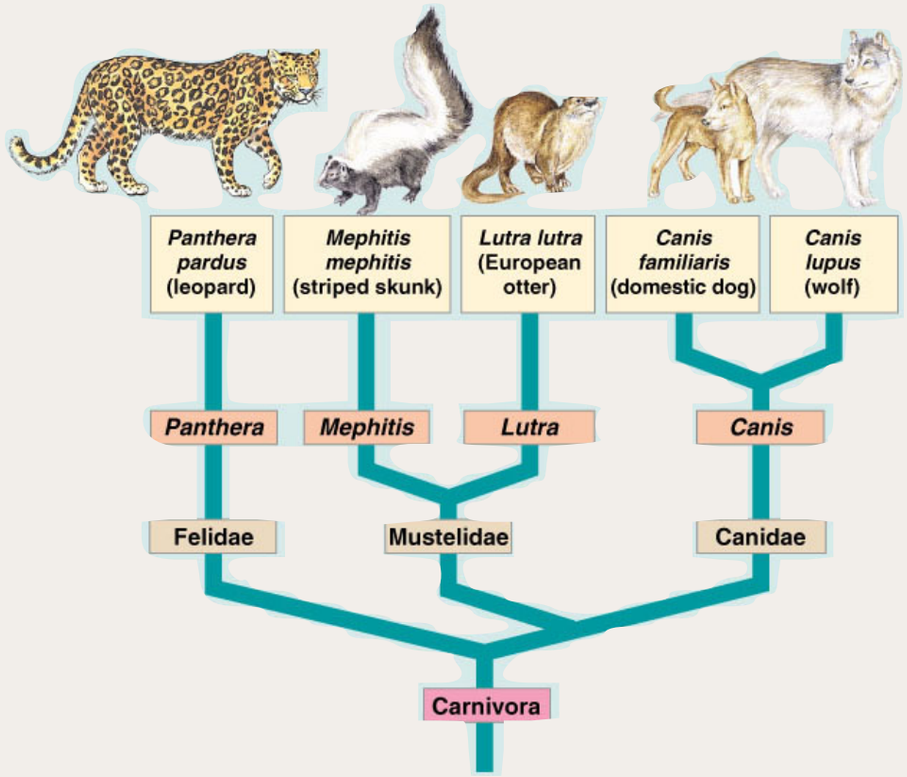
> *Nothing in Biology Makes Sense Except in the Light of Evolution*
>
> - Theodosius Dobzhansky 1973

Classification of all living organisms



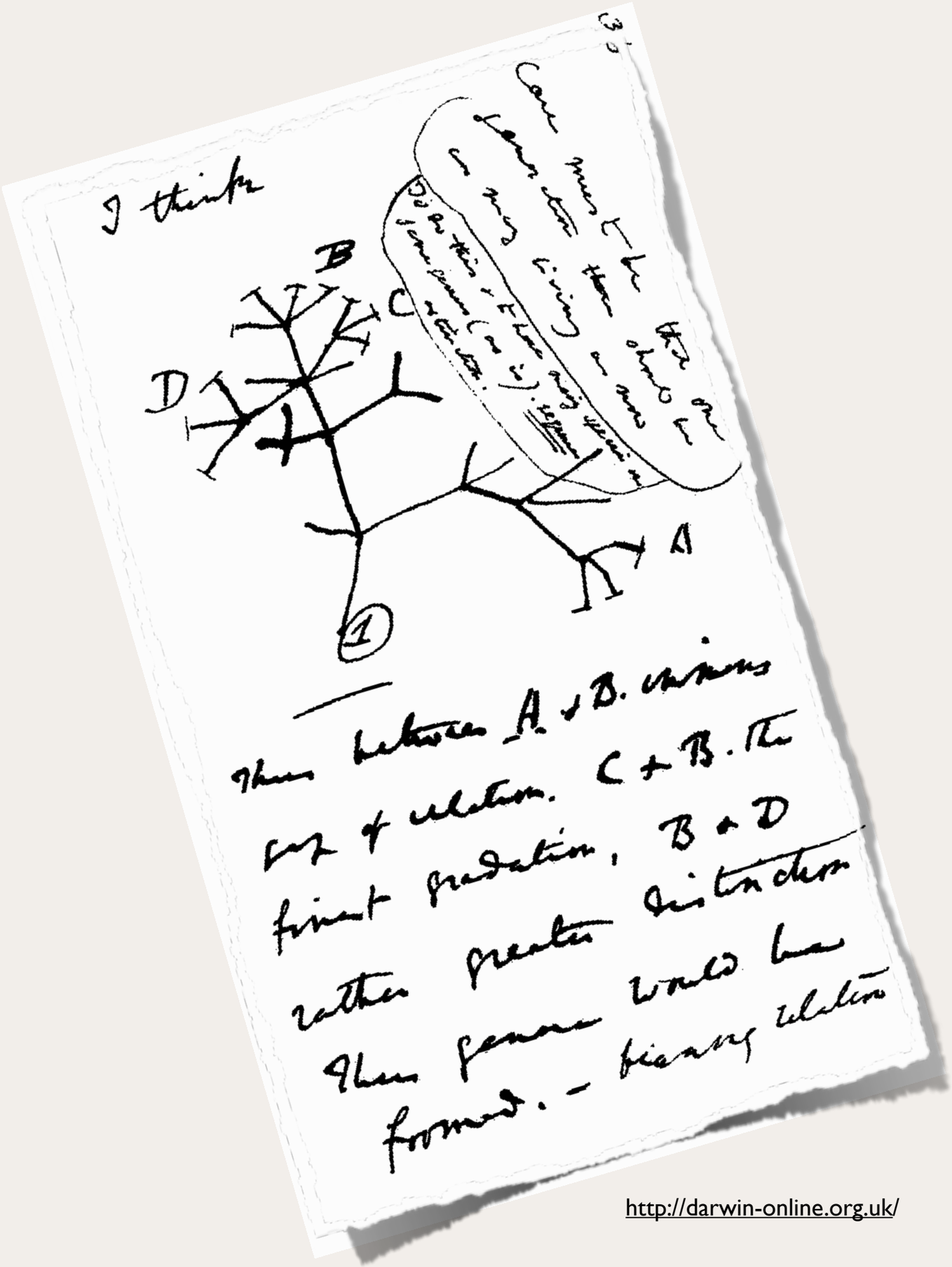
Carl Linnaeus (1707-1778)

Kingdom
Phylum
Class
Order
Family
Genus
Species

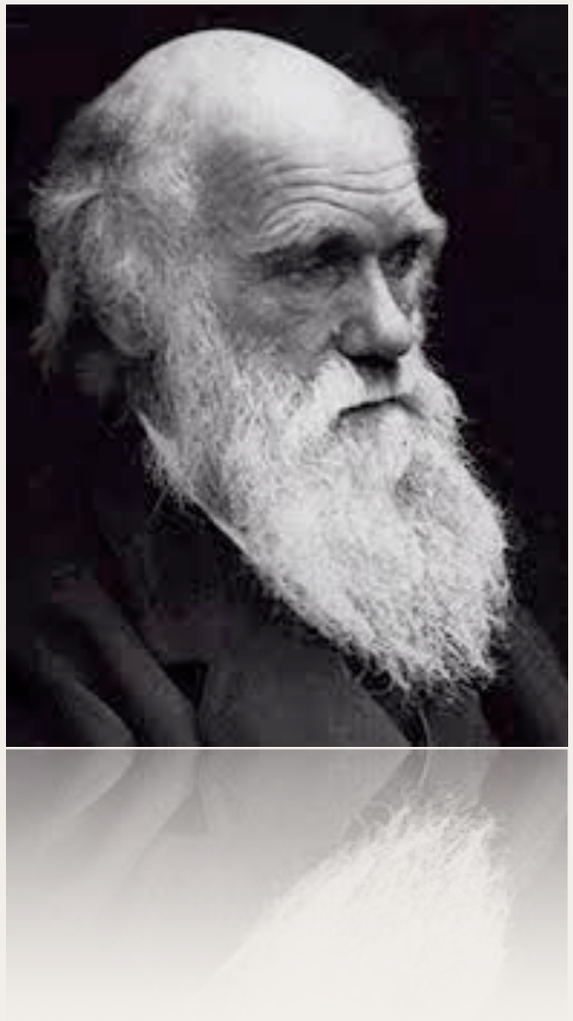


Quiz

Do you "know thyself"?



http://darwin-online.org.uk/

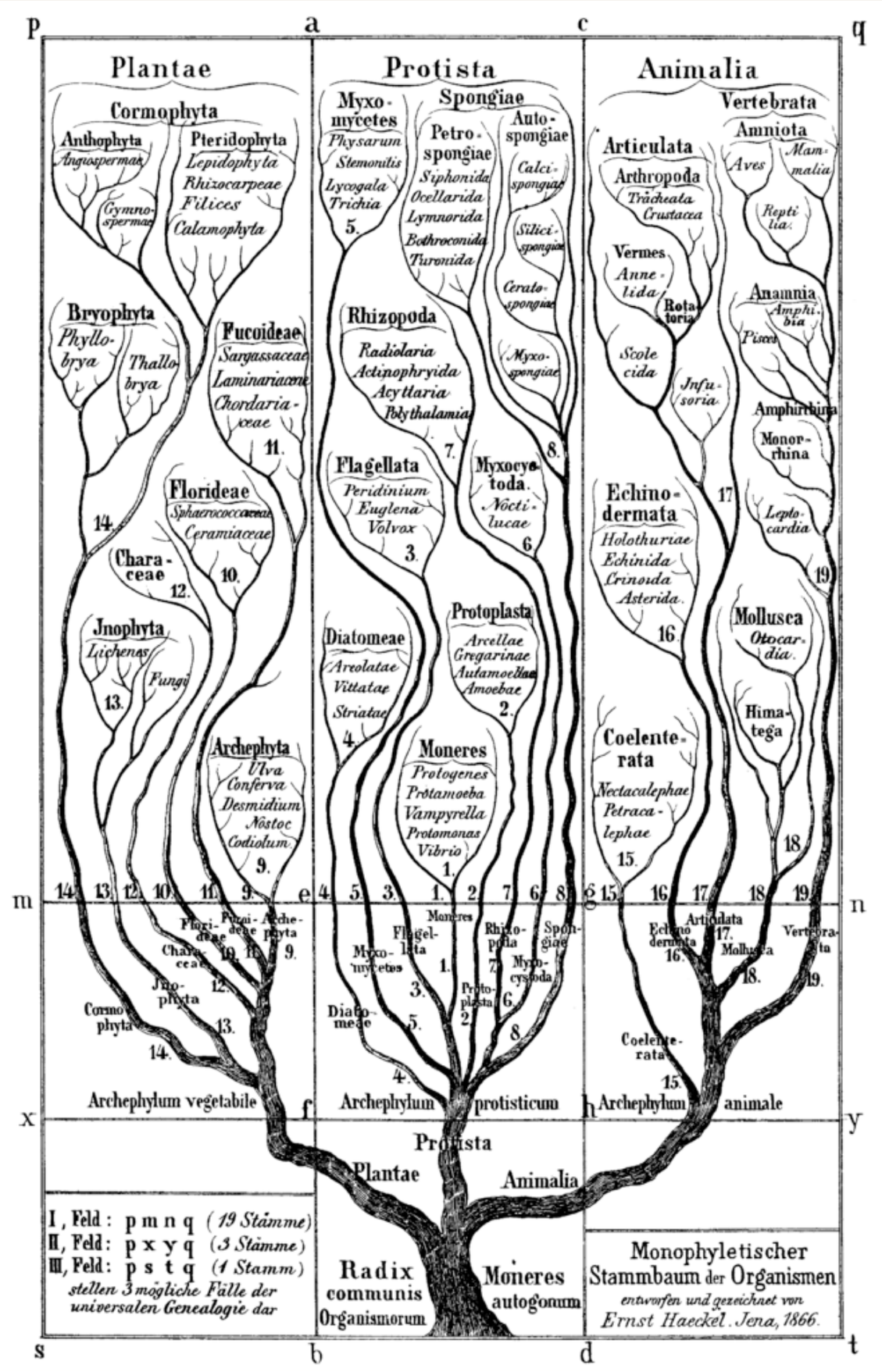


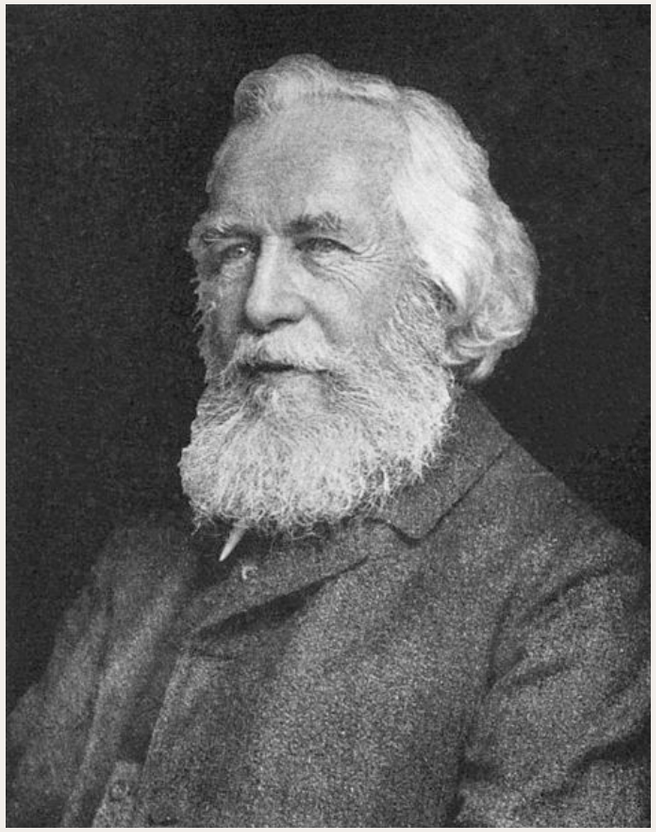
"I think..."

- Charles Darwin
[8]B Notebook p. 37[9]

Unity of Descent
Last Universal Common Ancestor
(LUCA)

"Common descent with modification"

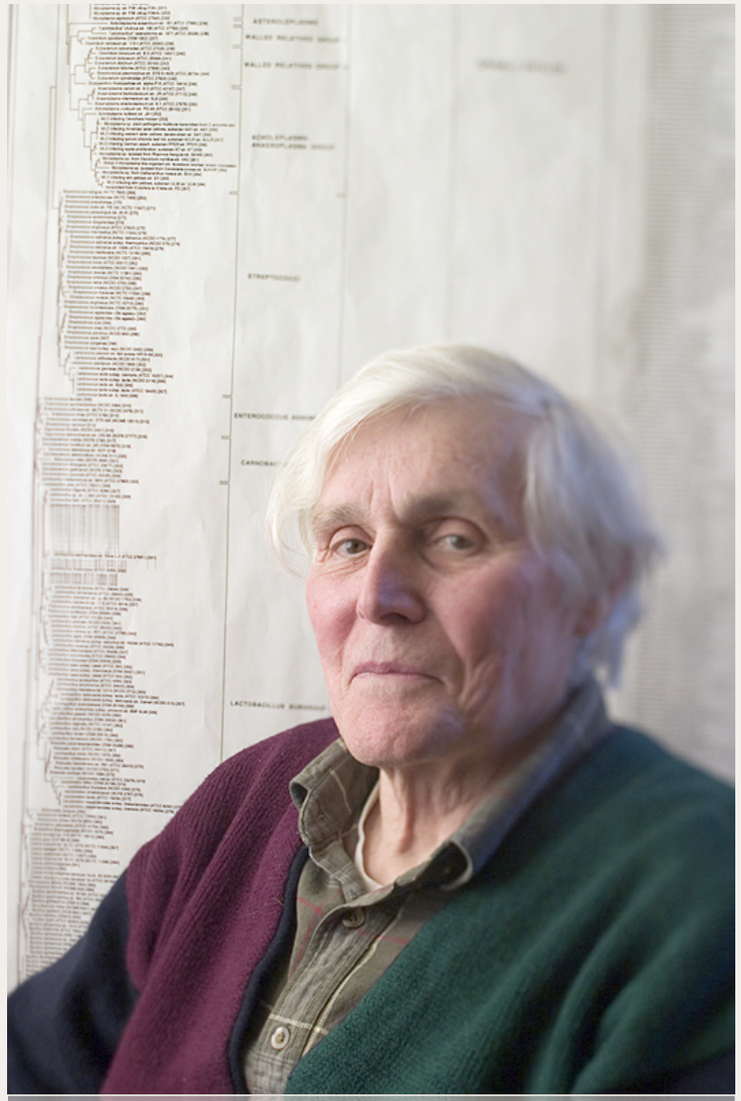




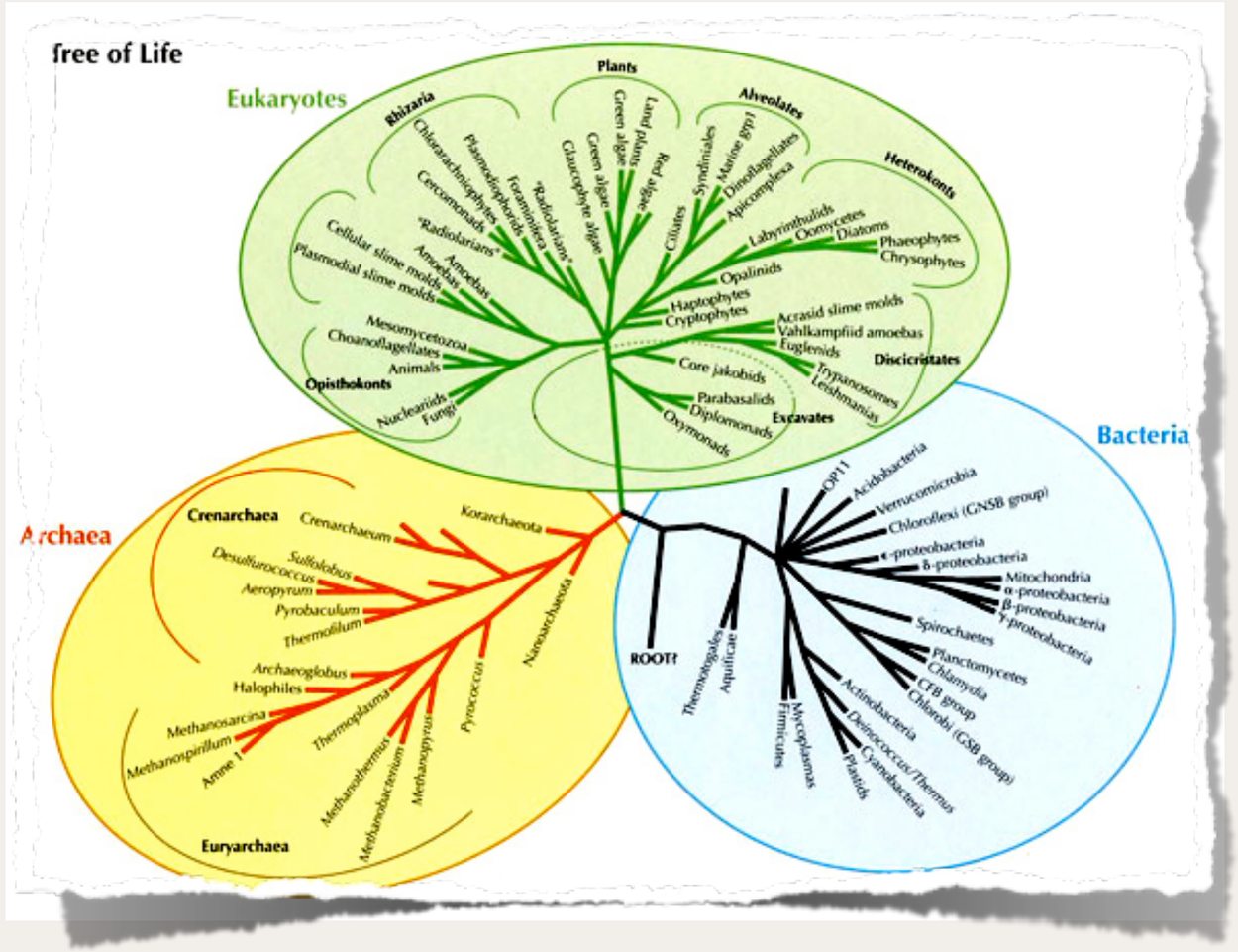
Earnst Haeckel (1834-1919)

http://upload.wikimedia.org/wikipedia/commons/thumb/8/81/Ernst_Haeckel_5.jpg/470px-Ernst_Haeckel_5.jpg

Tree kingdoms of Life



Carl Richard Woese (1928-2012)

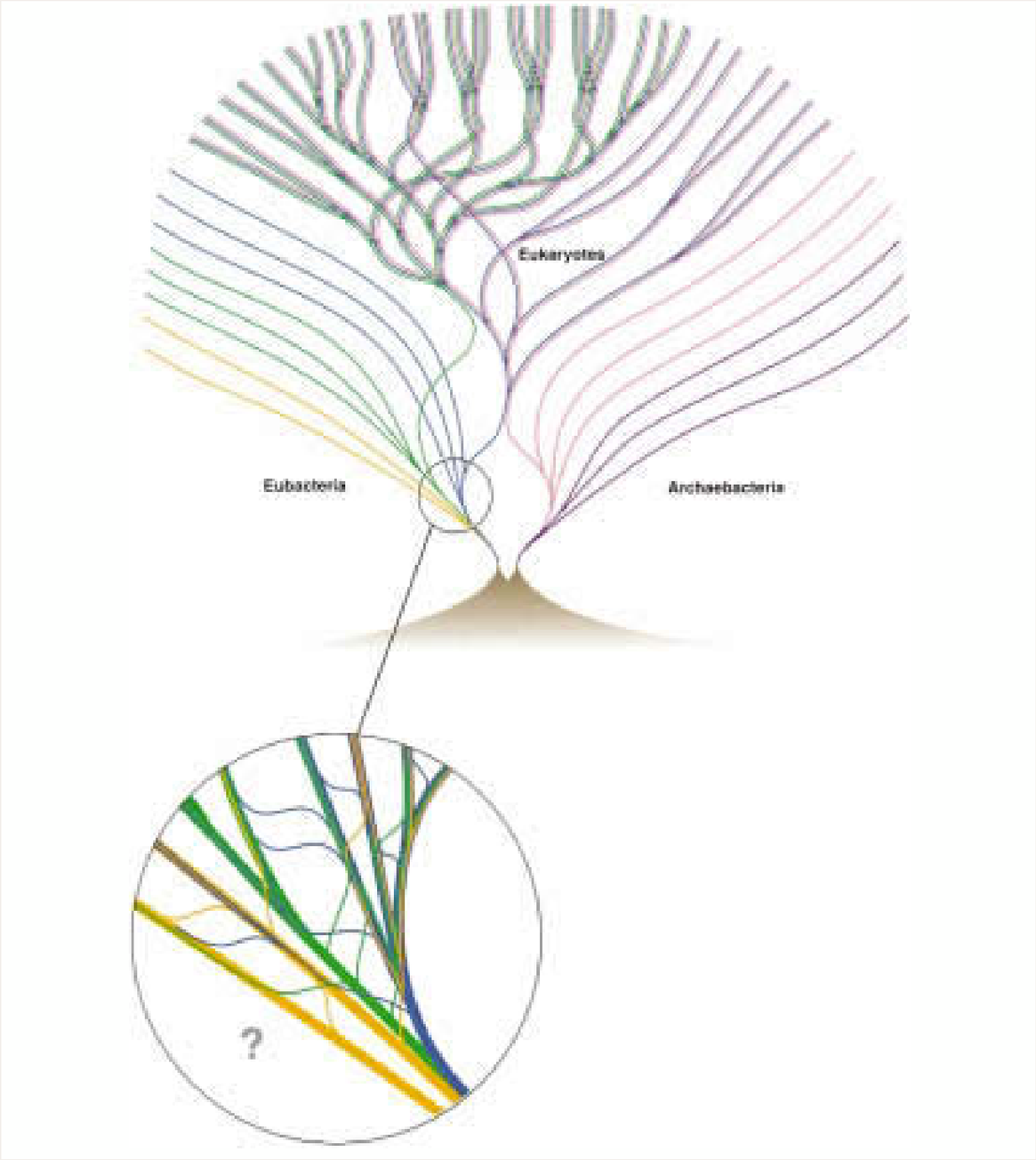


http://home.planet.nl/~gkorthof/korthof88.htm

Horizontal Gene Transfer

Transfer of genetic material from surroundings to genome

Phylogenetic forrest (many trees)



3 pillars of evolution

1. Mutation - Random error in DNA replication
2. Selection - Increase/decrease fitness
3. Drift - Random fluctuation in allele frequency

Bone in classical evolutionary theories

Earnst Haeckel (1924-1994)

Most changes in DNA are "neutral".
Genetic drift is the major cause of evolution

"Molecular" evolution

Evolutionary changes in molecules: DNA and protein sequences

Mutational changes in DNA

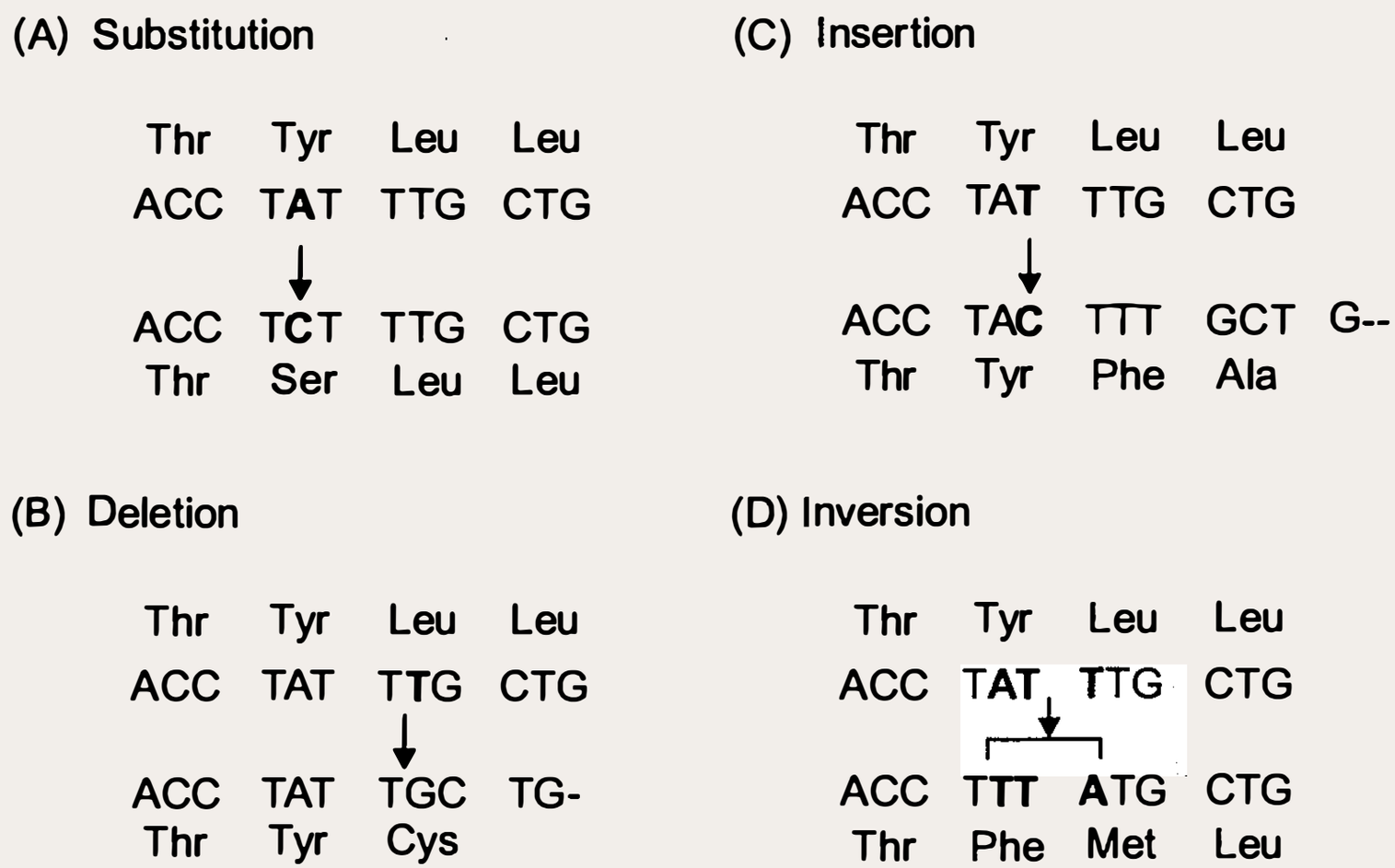


FIGURE 1.2. Four basic types of mutation at the nucleotide level. Nucleotide sequences are presented in units of codons or nucleotide triplets in order to show how the amino acids encoded are affected by the nucleotide changes. The nucleotides affected by the mutational changes are shown in boldface.

Nucleotide substitution

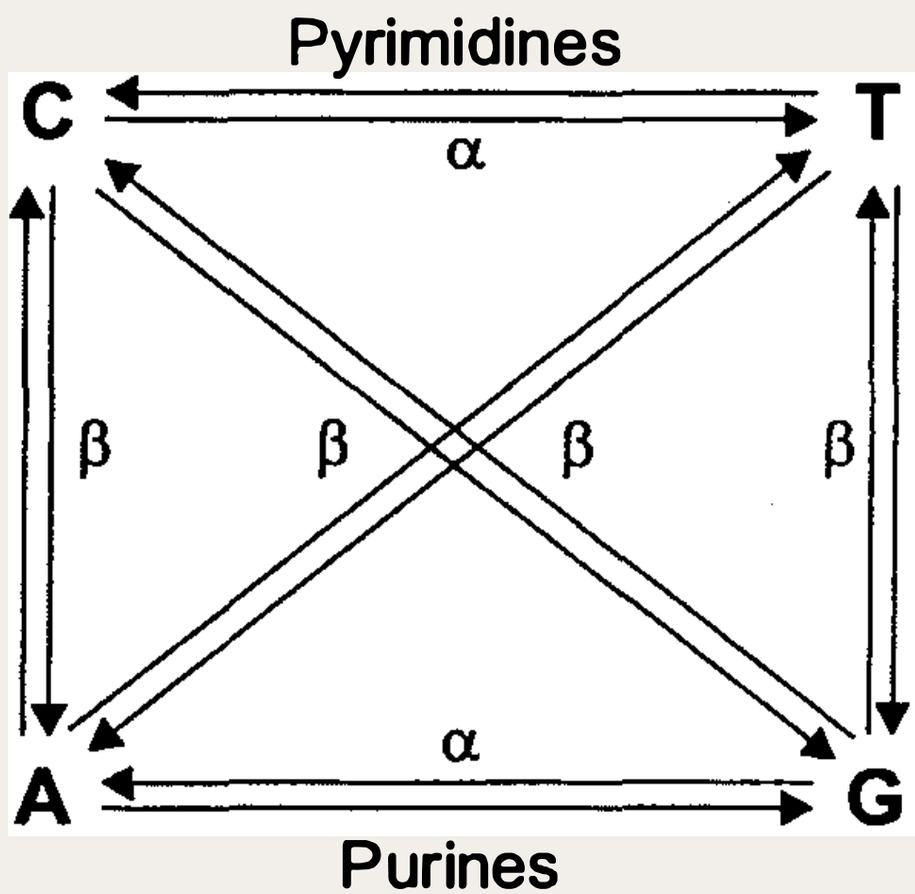


FIGURE 1.3. Transitional (A↔G and T↔C) and transversional (others) nucleotide substitutions. α and β are the rates of transitional and transversional substitutions, respectively.

Standard Genetic Code

TTT	F	Phe		TCT	S	Ser		TAT	Y	Tyr		TGT	C	Cys
TTC	F	Phe		TCC	S	Ser		TAC	Y	Tyr		TGC	C	Cys
TTA	L	Leu		TCA	S	Ser		TAA	*	Ter		TGA	*	Ter
TTG	L	Leu	i	TCG	S	Ser		TAG	*	Ter		TGG	W	Trp
CTT	L	Leu		CCT	P	Pro		CAT	H	His		CGT	R	Arg
CTC	L	Leu		CCC	P	Pro		CAC	H	His		CGC	R	Arg
CTA	L	Leu		CCA	P	Pro		CAA	Q	Gln		CGA	R	Arg
CTG	L	Leu	i	CCG	P	Pro		CAG	Q	Gln		CGG	R	Arg
ATT	I	Ile		ACT	T	Thr		AAT	N	Asn		AGT	S	Ser
ATC	I	Ile		ACC	T	Thr		AAC	N	Asn		AGC	S	Ser
ATA	I	Ile		ACA	T	Thr		AAA	K	Lys		AGA	R	Arg
ATG	M	Met	i	ACG	T	Thr		AAG	K	Lys		AGG	R	Arg
GTT	V	Val		GCT	A	Ala		GAT	D	Asp		GGT	G	Gly
GTC	V	Val		GCC	A	Ala		GAC	D	Asp		GGC	G	Gly
GTA	V	Val		GCA	A	Ala		GAA	E	Glu		GGA	G	Gly
GTG	V	Val		GCG	A	Ala		GAG	E	Glu		GGG	G	Gly

Codon bias

Phe	UUU	15	(0.51)	Ser	UCU	32	(1.86)	Tyr	UAU	18	(0.64)
	UUC	44	(1.49)		UCC	38	(2.21)		**UAC**	38	(1.36)
Leu	UUA	2	(0.07)		UCA	2	(0.12)	Ter	UAA		
	UUG	8	(0.27)		UCG	5	(0.29)	Ter	UAG		
Leu	CUU	11	(0.36)	Pro	CCU	9	(0.48)	His	CAU	5	(0.36)
	CUC	18	(0.60)		CCC	0	(0.00)		CAC	23	(1.64)
	CUA	1	(0.03)		CCA	11	(0.59)	Gln	CAA	15	(0.34)
	CUG	141	(4.67)		**CCG**	55	(2.93)		**CAG**	73	(1.66)
Ile	AUU	29	(0.69)	Thr	**ACU**	19	(0.78)	Asn	AAU	4	(0.11)
	AUC	98	(2.31)		**ACC**	63	(2.57)		**AAC**	66	(1.89)
	AUA	0	(0.00)		ACA	3	(0.12)	Lys	**AAA**	77	(1.35)
Met	AUG	60	(1.00)		ACG	13	(0.53)		AAG	37	(0.65)
Val	**GUU**	55	(1.53)	Ala	GCU	30	(0.94)	Asp	GAU	60	(0.83)
	GUC	21	(0.58)		GCC	19	(0.59)		GAC	85	(1.17)
	GUA	34	(0.94)		**GCA**	30	(0.94)	Glu	**GAA**	147	(1.52)
	GUG	34	(0.94)		**GCG**	49	(1.53)		GAG	46	(0.48)

Cys	UGU	5	(1.00)	Arg	**CGU**	89	(3.93)	Ser	AGU	3	(0.17)
	UGC	5	(1.00)		**CGC**	46	(2.03)		AGC	23	(1.34)
Ter	UGA				CGA	1	(0.04)	Arg	AGA	0	(0.00)
Trp	UGG	8	(1.00)		CGG	0	(0.00)		AGG	0	(0.00)
Gly	**GGU**	78	(2.40)								
	GGC	47	(1.45)								
	GGA	0	(0.00)								
	GGG	5	(0.15)								

FIGURE 1.4. Codon frequencies observed in the RNA polymerase genes (rpo B and D genes) of the bacterium *Escherichia coli*. The codons optimal for the translational system are shown in boldface. Relative synonymous codon usages (RSCU) given in the parentheses were computed by Equation (1.1). Data from Ikemura (1985).

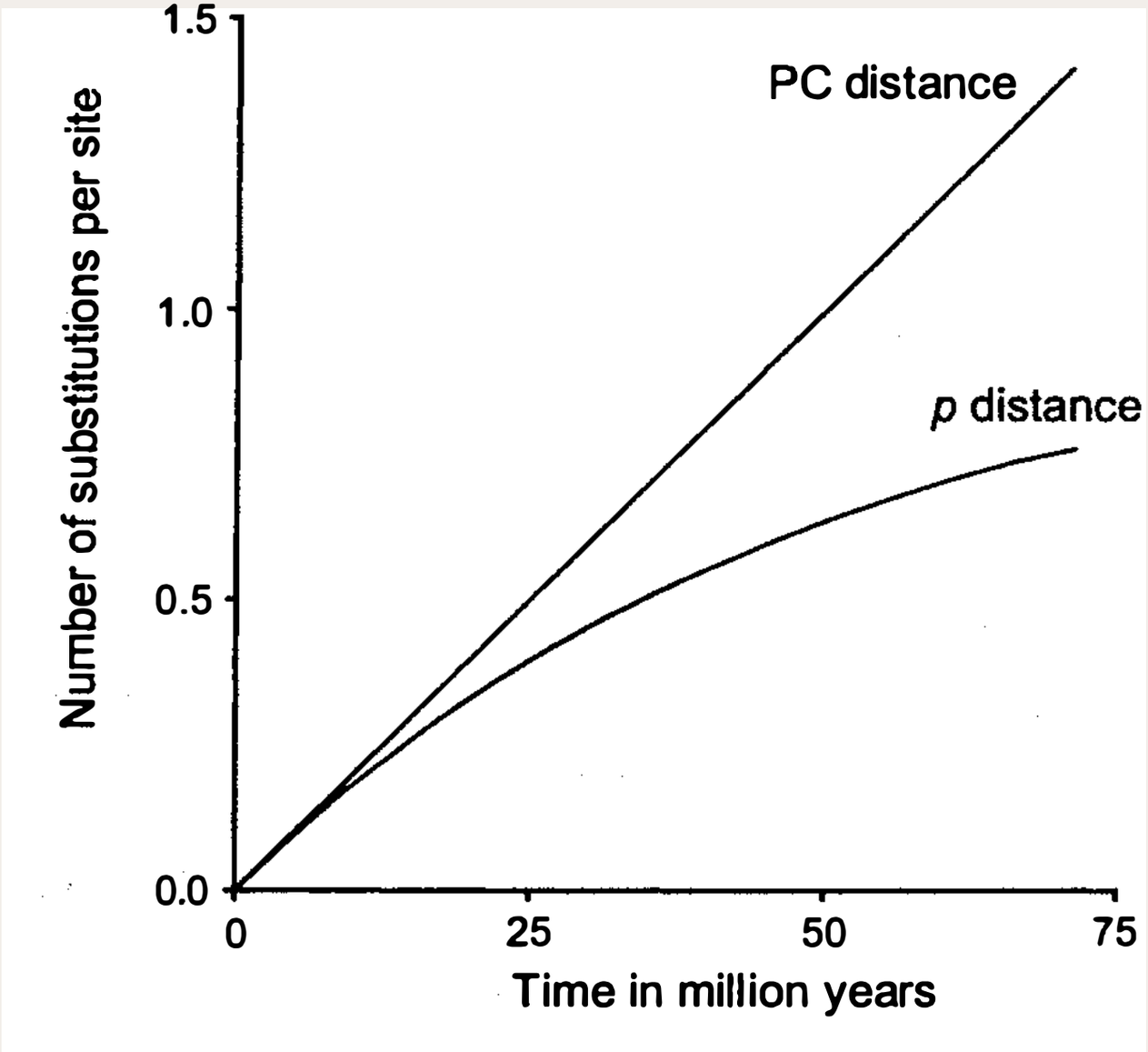
Result of substitution

- Synonymous - Does not change the AA
- Non-synonymous (missense) - Changes the AA
- Nonsense - Creates a stop codon

P Distance

No of times two sequences differ

Sequence divergence with time



Models of nucleotide substitutions

Table 3.2 Models of nucleotide substitution.

	A	T	C	G	A	T	C	G
	(A) Jukes-Cantor model				(E) HKY model			
A	-	α	α	α	-	βg_T	βg_C	αg_G
T	α	-	α	α	βg_A	-	αg_C	βg_G
C	α	α	-	α	βg_A	αg_T	-	βg_G
G	α	α	α	-	αg_A	βg_T	βg_C	-
	(B) Kimura model				(F) Tamura-Nei model			
A	-	β	β	α	-	βg_T	βg_C	$\alpha_1 g_G$
T	β	-	α	β	βg_A	-	$\alpha_2 g_C$	βg_C
C	β	α	-	β	βg_A	$\alpha_2 g_T$	-	βg_G
G	α	β	β	-	$\alpha_1 g_A$	βg_T	βg_C	-
	(C) Equal-input model				(G) General reversible model			
A	-	αg_T	αg_C	αg_G	-	$a g_T$	$b g_C$	$c g_G$
T	αg_A	-	αg_C	αg_G	$a g_A$	-	$d g_C$	$e g_G$
C	αg_A	αg_T	-	αg_G	$b g_A$	$d g_T$	-	$f g_G$
G	αg_A	αg_T	αg_C	-	$c g_A$	$e g_T$	$f g_C$	-
	(D) Tamura model				(H) Unrestricted model			
A	-	$\beta\theta_2$	$\beta\theta_1$	$\alpha\theta_1$	-	a_{12}	a_{13}	a_{14}
T	$\beta\theta_2$	-	$\alpha\theta_1$	$\beta\theta_1$	a_{21}	-	a_{23}	a_{24}
C	$\beta\theta_2$	$\alpha\theta_2$	-	$\beta\theta_1$	a_{31}	a_{32}	-	a_{34}
G	$\alpha\theta_2$	$\beta\theta_2$	$\beta\theta_1$	-	a_{41}	a_{42}	a_{43}	-

Note: An element (e_{ij}) of the above substitution matrices stands for the substitution rate from the nucleotide in the i-th row to the nucleotide in the j-th column. g_A, g_T, g_C, and g_G are the nucleotide frequencies. $\theta_1 = g_G + g_C$. $\theta_2 = g_A + g_T$.

Empirical AA substitution table

PAM (Point Accepted Mutation)

- Created by Margaret Dayhoff
- Different matrix for different evolutionary distance

BLOSUM

BLOSUM62

```
#  Matrix made by matblas from blosum62.iij
#  * column uses minimum score
#  BLOSUM Clustered Scoring Matrix in 1/2 Bit Units
#  Blocks Database = /data/blocks_5.0/blocks.dat
#  Cluster Percentage: >= 62
#  Entropy =   0.6979, Expected =  -0.5209
   A  R  N  D  C  Q  E  G  H  I  L  K  M  F  P  S  T  W  Y  V  B  Z  X  *
A  4 -1 -2 -2  0 -1 -1  0 -2 -1 -1 -1 -1 -2 -1  1  0 -3 -2  0 -2 -1  0 -4
R -1  5  0 -2 -3  1  0 -2  0 -3 -2  2 -1 -3 -2 -1 -1 -3 -2 -3 -1  0 -1 -4
N -2  0  6  1 -3  0  0  0  1 -3 -3  0 -2 -3 -2  1  0 -4 -2 -3  3  0 -1 -4
D -2 -2  1  6 -3  0  2 -1 -1 -3 -4 -1 -3 -3 -1  0 -1 -4 -3 -3  4  1 -1 -4
C  0 -3 -3 -3  9 -3 -4 -3 -3 -1 -1 -3 -1 -2 -3 -1 -1 -2 -2 -1 -3 -3 -2 -4
Q -1  1  0  0 -3  5  2 -2  0 -3 -2  1  0 -3 -1  0 -1 -2 -1 -2  0  3 -1 -4
E -1  0  0  2 -4  2  5 -2  0 -3 -3  1 -2 -3 -1  0 -1 -3 -2 -2  1  4 -1 -4
G  0 -2  0 -1 -3 -2 -2  6 -2 -4 -4 -2 -3 -3 -2  0 -2 -2 -3 -3 -1 -2 -1 -4
H -2  0  1 -1 -3  0  0 -2  8 -3 -3 -1 -2 -1 -2 -1 -2 -2  2 -3  0  0 -1 -4
I -1 -3 -3 -3 -1 -3 -3 -4 -3  4  2 -3  1  0 -3 -2 -1 -3 -1  3 -3 -3 -1 -4
L -1 -2 -3 -4 -1 -2 -3 -4 -3  2  4 -2  2  0 -3 -2 -1 -2 -1  1 -4 -3 -1 -4
K -1  2  0 -1 -3  1  1 -2 -1 -3 -2  5 -1 -3 -1  0 -1 -3 -2 -2  0  1 -1 -4
M -1 -1 -2 -3 -1  0 -2 -3 -2  1  2 -1  5  0 -2 -1 -1 -1 -1  1 -3 -1 -1 -4
F -2 -3 -3 -3 -2 -3 -3 -3 -1  0  0 -3  0  6 -4 -2 -2  1  3 -1 -3 -3 -1 -4
P -1 -2 -2 -1 -3 -1 -1 -2 -2 -3 -3 -1 -2 -4  7 -1 -1 -4 -3 -2 -2 -1 -2 -4
S  1 -1  1  0 -1  0  0  0 -1 -2 -2  0 -1 -2 -1  4  1 -3 -2 -2  0  0  0 -4
T  0 -1  0 -1 -1 -1 -1 -2 -2 -1 -1 -1 -1 -2 -1  1  5 -2 -2  0 -1 -1  0 -4
W -3 -3 -4 -4 -2 -2 -3 -2 -2 -3 -2 -3 -1  1 -4 -3 -2 11  2 -3 -4 -3 -2 -4
Y -2 -2 -2 -3 -2 -1 -2 -3  2 -1 -1 -2 -1  3 -3 -2 -2  2  7 -1 -3 -2 -1 -4
V  0 -3 -3 -3 -1 -2 -2 -3 -3  3  1 -2  1 -1 -2 -2  0 -3 -1  4 -3 -2 -1 -4
B -2 -1  3  4 -3  0  1 -1  0 -3 -4  0 -3 -3 -2  0 -1 -4 -3 -3  4  1 -1 -4
Z -1  0  0  1 -3  3  4 -2  0 -3 -3  1 -1 -3 -1  0 -1 -3 -2 -2  1  4 -1 -4
X  0 -1 -1 -1 -2 -1 -1 -1 -1 -1 -1 -1 -1 -1 -2  0  0 -2 -1 -1 -1 -1 -1 -4
* -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4  1
```

How BLOSUM is calculated

$$S_{ij} = \frac{1}{\lambda} log \left(\frac{f_{ij}}{f_i \times f_j} \right)$$

λ = a scaling parameter

f_{ij} = frequency of number of times one AA changes to another

f_i, f_j = frequency of each AA

Quiz

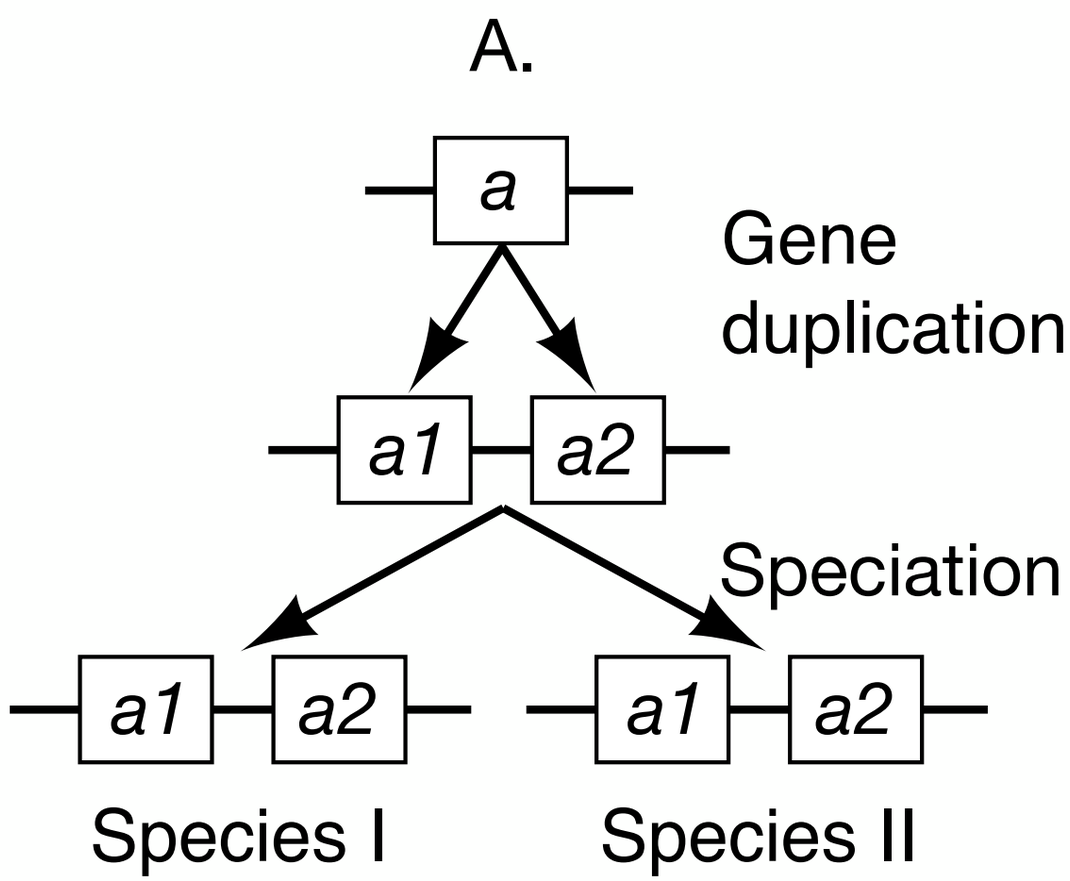
Why the self substitution scores are different for different amino acids?

Find the score of PQG matching PQG using BLOSUM62

Homologs

Genes related by evolution.

Homologs



All 4 : homologs

Species1 (a1,a2) and Species 2 (a1, a2): orthologs

a1 and a2: paralogs

Fitch W. (1970). "Distinguishing homologous from analogous proteins". *Syst Zool* **19** (2): 99–113.

DISTINGUISHING HOMOLOGOUS FROM ANALOGOUS PROTEINS

WALTER M. FITCH

Abstract

Fitch, W. M. (Dept. Physiological Chem., U. Wisconsin, Madison 53706) 1970. Distinguishing homologous from analogous proteins. Syst. Zool., 19:99–113.—This work provides a means by which it is possible to determine whether two groups of related proteins have a common ancestor or are of independent origin. A set of 16 random amino acid sequences were shown to be unrelated by this method. A set of 16 real but presumably unrelated proteins gave a similar result. A set of 24 model proteins which was composed of two independently evolving groups, converging toward the same chemical goal, was correctly shown to be convergently related, with the probability that the result was due to chance being $<10^{-m}$. A set of 24 cytochromes composed of 5 fungi and 19 metazoans was shown to be divergently related, with the probability that the result was due to chance being $< 10^{-6}$. A process was described which leads to the absolute minimum of nucleotide replacements required to account for the divergent descent of a set of genes given a particular topology for the tree depicting their ancestral relations. It was also shown that the convergent processes could realistically lead to amino acid sequences which would produce positive tests for relatedness, not only by a chemical criterion, but by a genetic (nucleotide sequence) criterion as well. Finally, a realistic case is indicated where truly homologous traits, behaving in a perfectly expectable way, may nevertheless lead to a ludicrous phylogeny.

The demonstration that two proteins are related has been attempted using two different criteria. One criterion is to show that their chemical structures are very similar. An early example of this approach was the observation of the relatedness of the oxygen carrying proteins, myoglobin and hemoglobin (Watson and Kendrew, 1961). More recent is the relatedness of two enzymes in carbohydrate metabolism, lysozyme and alpha-lactalbumin (Brew, Vanaman and Hill, 1967). The other criterion is to show that underlying genetic structures of the proteins are more alike than one would expect by chance. This is now possible because our knowledge of the genetic code permits us to determine how many nucleotide positions, at the minimum, must differ in the genes encoding the two presumptively homologous proteins. One then compares the answer obtained to the number of differences one would expect for unrelated proteins. An example of this approach is the observation of the relatedness of plant and bacterial ferredoxins (Matsubara, Jukes and Cantor, 1969) for which added evidence has been produced (Fitch, 1970a). But regardless of the approach, the impulse, too powerful to resist, is to conclude that a particular pair of proteins had a common genic ancestor if they meet whichever criterion the observer uses.

Now two proteins may appear similar because they descend with *di*vergence from a common ancestral gene (i.e., are homologous in a time-honoured meaning dating back at the least to Darwin's *Origin of Species*) or because they descend with *con*vergence from separate ancestral genes (i.e., are analogous). And, if a common genic ancestor is to be the conclusion, a genetic criterion should be superior to a chemical criterion. This is because analogous gene products, although they have no common ancestor, do serve similar functions and may well be expected to have similar chemical structures and thereby be confused with homologous gene products. This danger can only be increased by using a chemical, as opposed to a genetic, criterion.

Sequence similarity is not homology

Homology vs Homoplasy

Detecting selection

Types of selection

Purifying /Negative selection - Does not allow change

Positive/Adaptive selection - Faster change

Neutral

How to measure selection

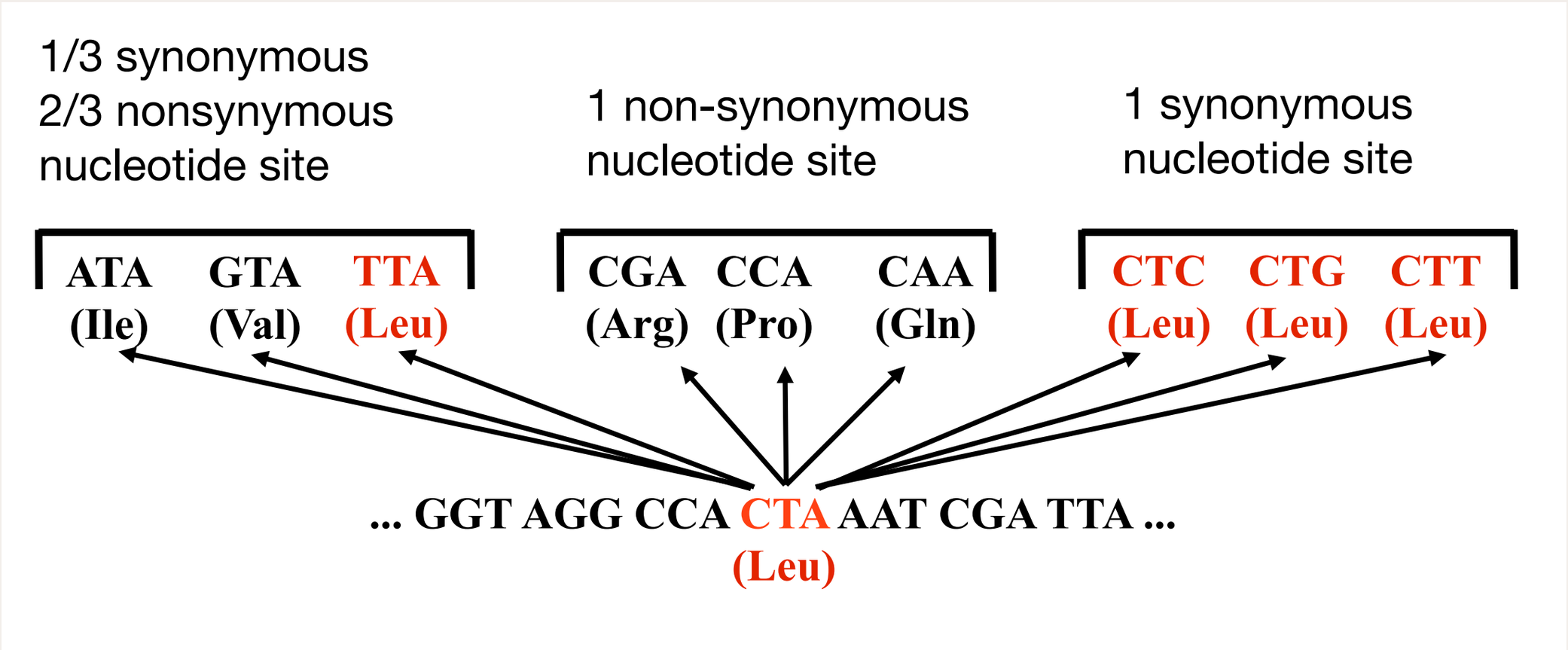
d_n= Non-synonymous substitutions /non-synonymous site

d_s = synonymous substitutions/synonymous site

$d_n/d_s > 1$ = Positive selection

$d_n/d_s < 1$ = Negative selection

$d_n/d_s = 1$ = Neutral selection


1/3 synonymous
2/3 nonsynonymous
nucleotide site
1 non-synonymous
nucleotide site
1 synonymous
nucleotide site
ATA GTA TTA
(Ile) (Val) (Leu)
CGA CCA CAA
(Arg) (Pro) (Gln)
CTC CTG CTT
(Leu) (Leu) (Leu)
... GGT AGG CCA CTA AAT CGA TTA ...
(Leu)

```
 T   P   N   G   A   L   E   L   K   P   V   R
ACTCCGAACGGGGCGTTAGAGTTGAAACCCGTTAGA
  *     *    *     *    *    *   *  *    **
ACGCCGATCGGCGCGATAGGGTTCAAGCTCGTACGA
 T   P   I   G   A   I   G   F   K   L   V   R
```

syn $00100100\frac{1}{2}001001\frac{1}{4}0\frac{1}{2}00\frac{1}{3}\frac{1}{3}0\frac{1}{3}00\frac{1}{3}001001\frac{1}{3}0\frac{2}{3}$ sum $= 7.5833$

non $11011011\frac{1}{2}110110\frac{3}{4}1\frac{1}{2}11\frac{2}{3}\frac{2}{3}1\frac{2}{3}11\frac{2}{3}110100\frac{2}{3}1\frac{1}{3}$ sum $= 28.4167$

$$\mathrm{dN} = \frac{\text{No. non-synonymous substitutions}}{\text{No. non-synonymous sites}} = \frac{5}{28.417} = 0.176$$

$$\mathrm{dS} = \frac{\text{No. synonymous substitutions}}{\text{No. synonymous sites}} = \frac{5}{7.583} = 0.659$$

The ratio is then

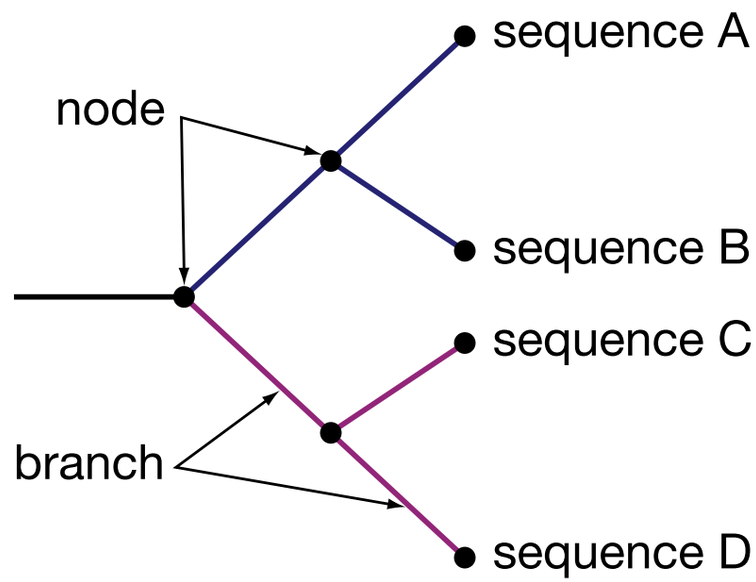
$$\frac{\mathrm{dN}}{\mathrm{dS}} = \frac{0.176}{0.659} = 0.269$$

Phylogeny

Evolutionary Tree

A graph structure showing the relationship amongst species or in case of genes, relationship amongst gene.

A. Rooted tree



B. Unrooted tree

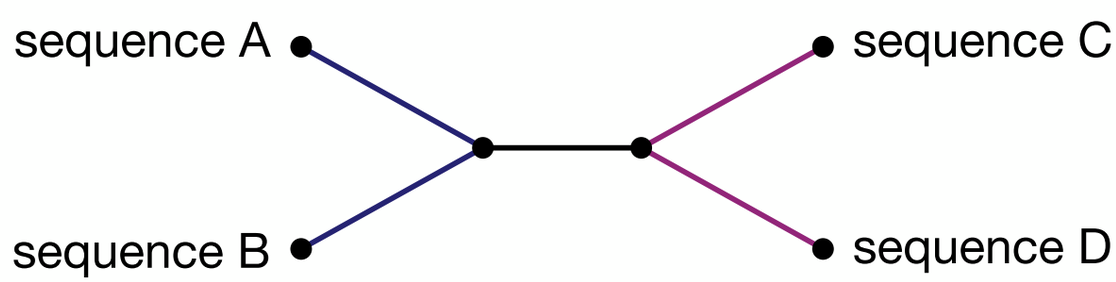


Figure 6.5. Structure of evolutionary trees.

Tree features

Taxon (plural taxa) are atomic units of the tree.

Branch length represent the estimate of the sequence change.

Each internal node represent a speciation even.

Tree features

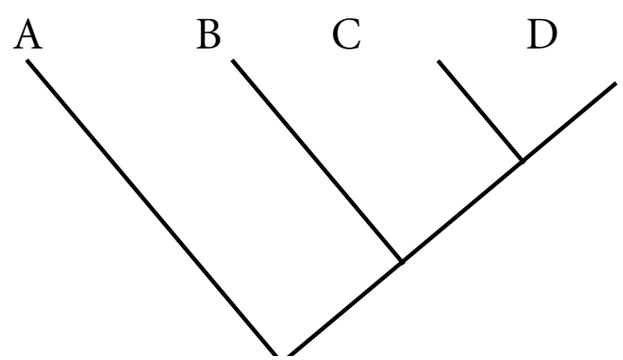
The branch length may differ due to “accelerated evolution” after speciation.

May phylogenetic techniques assume that the brach lengths are same “molecular clock”. Such assumption is only valid for closely related species.

Rooted trees are hard to make

Table 6.2. *Number of possible evolutionary trees to consider as a function of number of sequences*

Taxa or sequence no.	No. of rooted trees	No. of unrooted trees
3	3	1
4	15	3
5	105	15
—	—	—
7	10,395	954



Rooted tree

Root represent common ancestor of all nodes.

In general, root is fixed by a taxon that branched of earlier than the others "outgroup".

Root can also be predicted provided molecular clock assumption holds true.

$3^{1/2}$ Methods

Parsimony

Distance method

Maximum Likelihood

Bayesian

List of phylogenetic software

http://evolution.gs.washington.edu/phylip/software.html

Phylip

http://evolution.genetics.washington.edu/phylip/getme.html

Parsimony

Smallest number of evolutionary changes that explain the observed sequences.

Usually used for ancestral reconstruction using binary characters.

Occam's razor

http://www.reconnections.net/razor.jpg

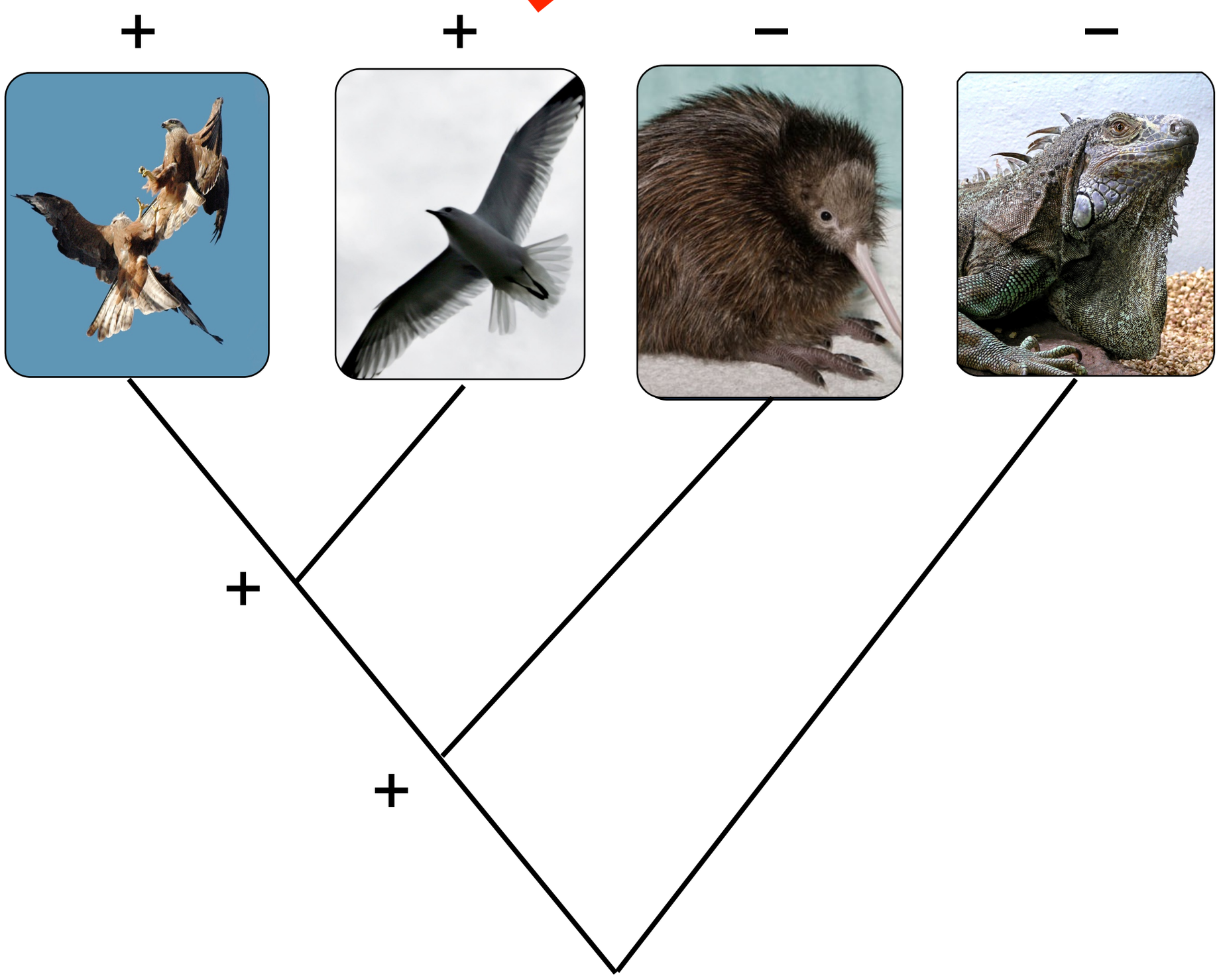
http://upload.wikimedia.org/

William of Ockham

14th Century

Ancestral reconstruction using Parsimony

DOLLO



Main Parsimony programs in phylip

DNAPARS for DNA

PROTPARS for protein

Parsimony

	1	2	3	4
Seq1	A	G	G	A
Seq2	A	G	G	G
Seq3	A	A	C	A
Seq4	A	A	C	G

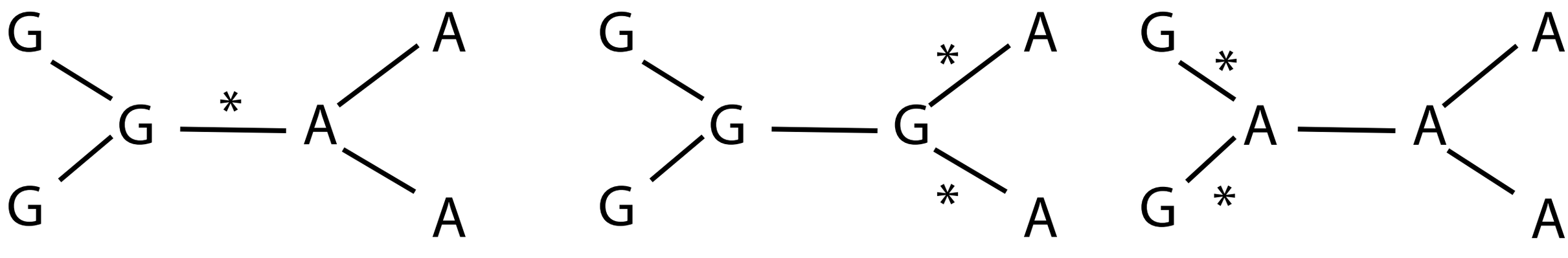
To be informative at least one change is required

Position 1: uninformative

Positions 2-4: informative

Parsimony

3 possible unrooted trees for position 2



Tree 1 is parsimonious tree with just one change

Best tree is the one that explains all the position with least number of changes.

Distance method

Step 1: Calculate distance between all pairs of sequence in a multiple alignment

Step 2: Create a phylogenetic tree from this distance matrix

Creating tree from distance matrix

FITCH: Fitch Margoliash method. No molecular clock.

KITSCH: Fitch Margoliash but under assumption of molecular clock.

NEIGHBOR: Neighbor joining or UPGMA.

NJ trees are unrooted and no assumption of molecular clock.

Align each pair of sequences and calculate distance as (number of mismatches/ number of matches) and create a distance matrix

	A	***B***	***C***	***D***
A	-	$D_{AB} = 20$	$D_{AC} = 25$	$D_{AD} = 37$
B	-	-	$D_{BC} = 45$	$D_{BD} = 42$
C	-	-	-	$D_{CD} = 15$
D	-	-	-	-

Programs to calculate distance matrix in PHYLIP

DNADIST for DNA. Uses various models for DNA

PROTDIST for protein. Uses various models including PAMs.

Creating tree using PHYLIP
Step 1

Create a multiple alignment

```
muscle -in inputfile -phyiout outfile
```

Creating tree using PHYLIP
Step 2

Run a distance program

protdist

Creating tree using PHYLIP
Step 3

Run a distance program

fitch

TABLE 27.11. Neighbor-joining example

Cycle 1

Distance matrix

	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

Step 1

S calculations

$S_x =$ (sum all D_x)/($N - 2$), where N is the # of OTUs in the set.

$S_A = (5+4+7+6+8)/4 = 7.5$
$S_B = (5+7+10+9+11)/4 = 10.5$
$S_C = (4+7+7+6+8)/4 = 8$
$S_D = (7+10+7+5+9)/4 = 9.5$
$S_E = (6+9+6+5+8)/4 = 8.5$
$S_F = (8+11+8+9+8)/4 = 11$

Step 2

Calculate pair with smallest (M), where $M_{ij} = D_{ij} - S_i - S_j$.

Smallest are
$M_{AB} = 5 - 7.5 - 10.5 = -13$
$M_{DE} = 5 - 9.5 - 8.5 = -13$
Choose one of these (AB here).

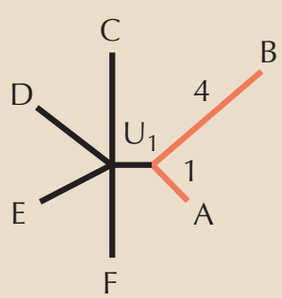
Step 3

Create a node (U) that joins pair with lowest M_{ij} such that $S_{IU} = D_{ij}/2 + (S_i - S_j)/2$.

U_1 joins A and B:
$S_{AU_1} = D_{AB}/2 + (S_A - S_B)/2 = 1$
$S_{BU_1} = D_{AB}/2 + (S_B - S_A)/2 = 4$

Step 4

Join *i* and *j* according to *S* above and make all other taxa in form of a star. Branches in black are of unknown length. Branches in red are of known length.



Step 5

Calculate new distance matrix of all other taxa to U with $D_{xU} = D_{ix} + D_{jx} - D_{ij}$, where *i* and *j* are those selected from above.

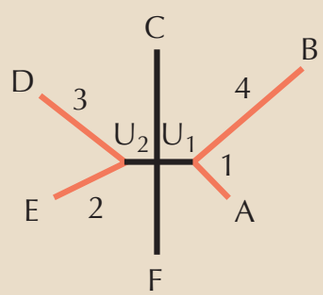
Cycle 2

	U_1	C	D	E
C	3			
D	6	7		
E	5	6	5	
F	7	8	9	8

$S_{U_1} = (3+6+5+7)/3 = 7$
$S_C = (3+7+6=8)/3 = 8$
$S_D = (6+7+5+9)/3 = 9$
$S_E = (5+6+5+8)/3 = 8$
$S_F = (7+8+9+8)/3 = 10.6$

Smallest is
$M_{CU_1} = 3 - 7 - 8 = -12$
$M_{DE} = 5 - 9 - 8 = -12$
Choose one of these (DE here).

U_2 joins D and E:
$S_{DU_2} = D_{DE}/2 + (S_D - S_E)2 = 3$
$S_{EU_2} = D_{DE}/2 + (S_E - S_D)/2 = 2$



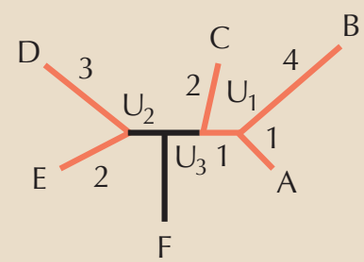
Cycle 3

	U_1	C	U_2
C	3		
U_2	3	4	
F	7	8	6

$S_{U_1} = (3+3+7)/2 = 6.5$
$S_C = (3+4+8)/2 = 7.5$
$S_{U_2} = (3+4+6)/2 = 6.5$
$S_F = (7+8+6)/2 = 10.5$

Smallest is
$M_{CU_1} = 3 - 6.5 - 7.5 = -11$

U_3 joins C and U_1:
$S_{CU_3} = D_{CU_1}/2 + (S_C - S_{U_1})/2 = 2$
$S_{U_1U_3} = D_{CU_1}/2 + (S_{U_1} - S_C)/2 = 1$



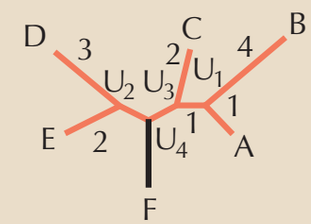
Cycle 4

	U_2	U_3
U_3	2	
F	6	6

$S_{U_2} = (2+6)/1 = 8$
$S_{U_3} = (2+6)/1 = 8$
$S_F = (6+6)/1 = 12$

Smallest is
$M_{U_2F} = 6 - 8 - 12 = -14$
$M_{U_3F} = 6 - 8 - 12 = -14$
$M_{U_2U_3} = 2 - 8 - 8 = -14$
Choose one of these ($M_{U_2U_3}$ here).

U_4 joins U_2 and U_3:
$S_{U_2U_4} = D_{U_2U_3}/2 + (S_{U_2} - S_{U_3})/2 = 1$
$S_{U_3U_4} = D_{U_2U_3}/2 + (S_{U_3} - S_{U_2})/2 = 1$.

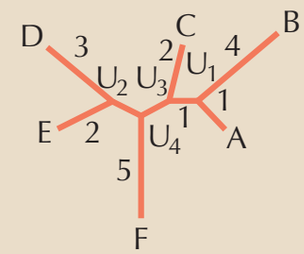


Cycle 5

	U_4
F	5

Because $N - 2 = 0$, we cannot do this calculation.

For last pair, connect U_4 and F with branch length = 5.



Comments

Note this is the same tree we started with (drawn in unrooted form here).

From http://www.icp.ucl.ac.be/~opperd/private/upgma.html.

Output tree format
Newick

```
(P73_HUMAN/:0.16068,((P53_XENLA/:0.18610,((P53_ONCMY/:0.12081,
P53_DANRE/:0.12111):0.02394,P53_HUMAN/:0.22849):0.03528):0.04183,
P53_ORYLA/:0.20291):0.11899,Q27937_LOL:0.48924);
```

Reliability
Bootstrapping

Randomly sample the original alignment

Create many alignments

Create many trees

Create a consensus tree

Bootstrpping in phylip

seqboot

protdist

fitch

consense

Don’t forget to use “multiple” parameters

Maximum Likelihood

Conditional probability

Likelihood

$$Prob(H|D) = \frac{Prob(D|H)Prob(H)}{Prob(D)}$$

$H =$ Hypothesis

D= data

Calculating likelihood

Given a dataset

$$D = D_1, D_2, \ldots, Dn$$

Likelihood

$$L = Prob(D_1|H)Prob(D_2|H)\ldots Prob(D_n|H)$$

Maximum likelihood

Given the likelihood

$$L = Prob(D_1|H)Prob(D_2|H)\ldots Prob(D_n|H)$$

We calculate the likelihood for a set of probabilities of H

The probability of H is “most probably” where the likelihood is maximum.

Let's calculate the probability of heads

HHTTH

$$L = pp(1-p)(1-p)p$$

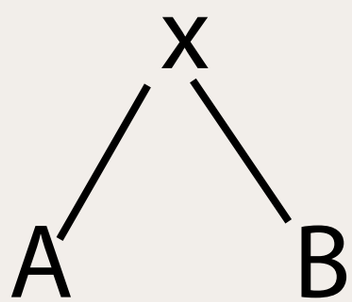
$$= p^3(1-p)^2$$

$$\ln L = \ln p^3 + \ln(1-p)^2$$

$$\frac{d(lnL)}{dp} = \frac{3}{p} - \frac{2}{(1-p)} = 0$$

$$p = \frac{3}{5}$$

Probability of a tree



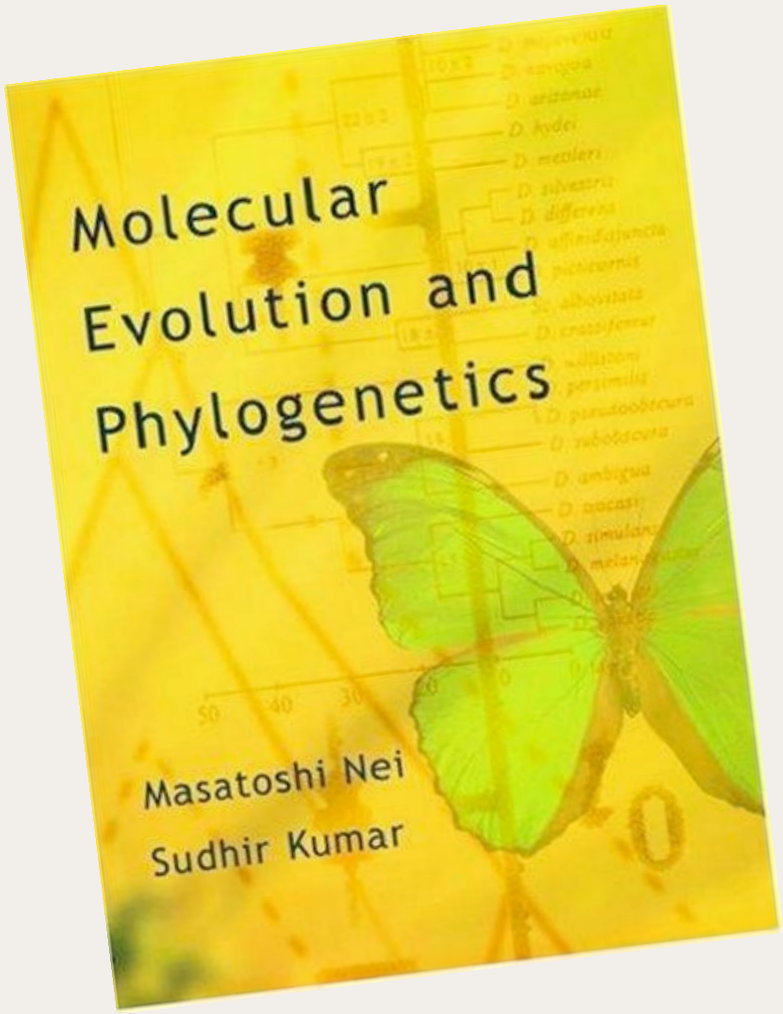
For nucleotide sequence
x = (A, T, G, or C)

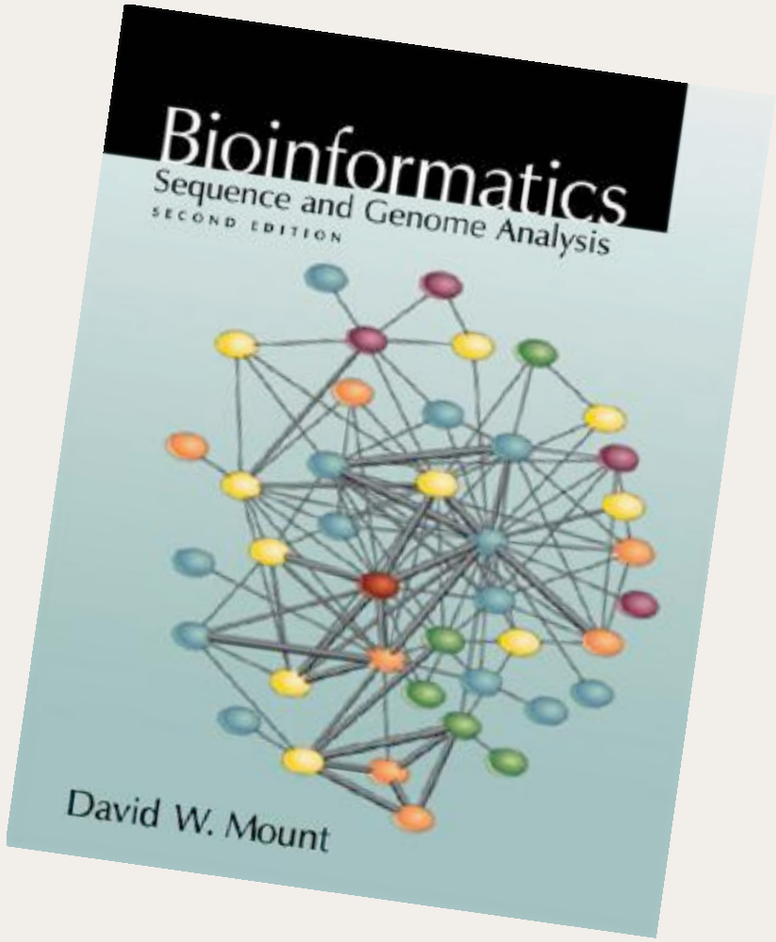
$$L = \sum_x Prob(x) Prob(A|x) Prob(B|x)$$

RAX-ML

http://icwww.epfl.ch/~stamatak/index-Dateien/Page443.htm

Text books


Molecular
Evolution and
Phylogenetics
Masatoshi Nei
Sudhir Kumar


Bioinformatics
Sequence and Genome Analysis
SECOND EDITION
David W. Mount