Introduction to Computational Biology; An Evolutionary Approach: Sequence Variation & Molecular Evolution

Bernhard Haubold & Thomas Wiehe

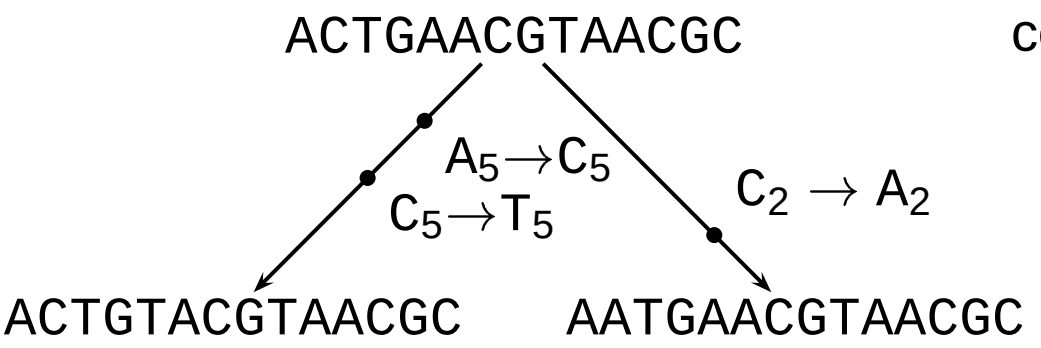
Outline

1 The Record of Past Events

2 Molecular Clock

3 K_a/K_s

Substitutions



common ancestral sequence

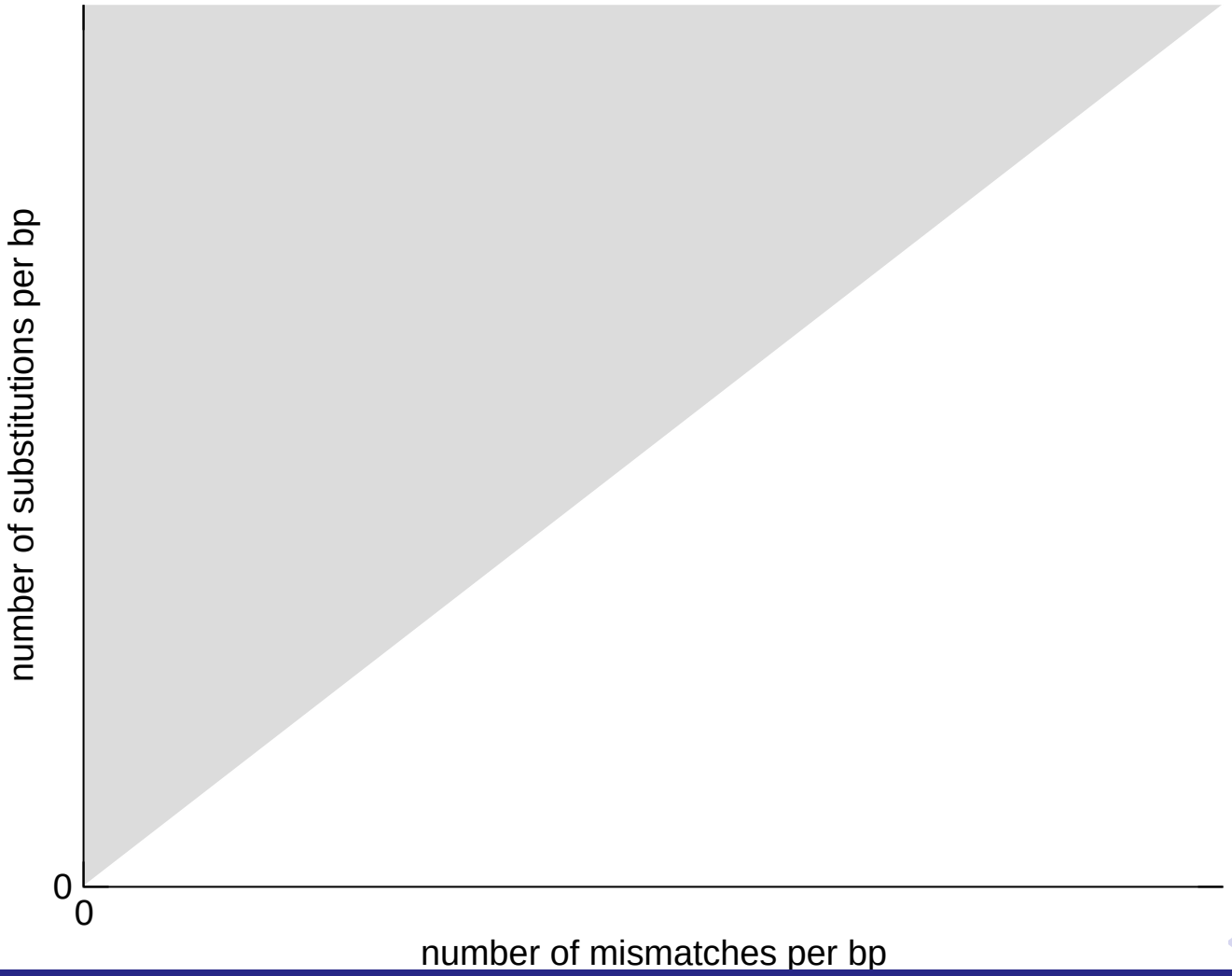
time

present day sequences

```
sequence 1   ACTGTACGTAACGC
             |  ||  |||||||||
sequence 2   AATGAACGTAACGC
```

sequence alignment

Substitutions & Mismatches



Mutation & Substitution Rate

μ number of mutations per unit sequence length per unit time

N number of diploid organisms in population $\Rightarrow 2N\mu$ new mutations per unit time

$1/(2N)$ probability of fixation $\Rightarrow$

$$k = 2N\mu \times \frac{1}{1N} = \mu$$

Comparing Human α_1-Globinlobin

		DNA[2]	AA[3]	time[4]
(1)	human α_2	0.00	0.00	20
(2)	chimp α_1	0.07	0.00	5
(3)	horse α	13.3	12.1	80
(4)	carp α	40.9	52.4	400
(5)	human β	41.2	57.4	450

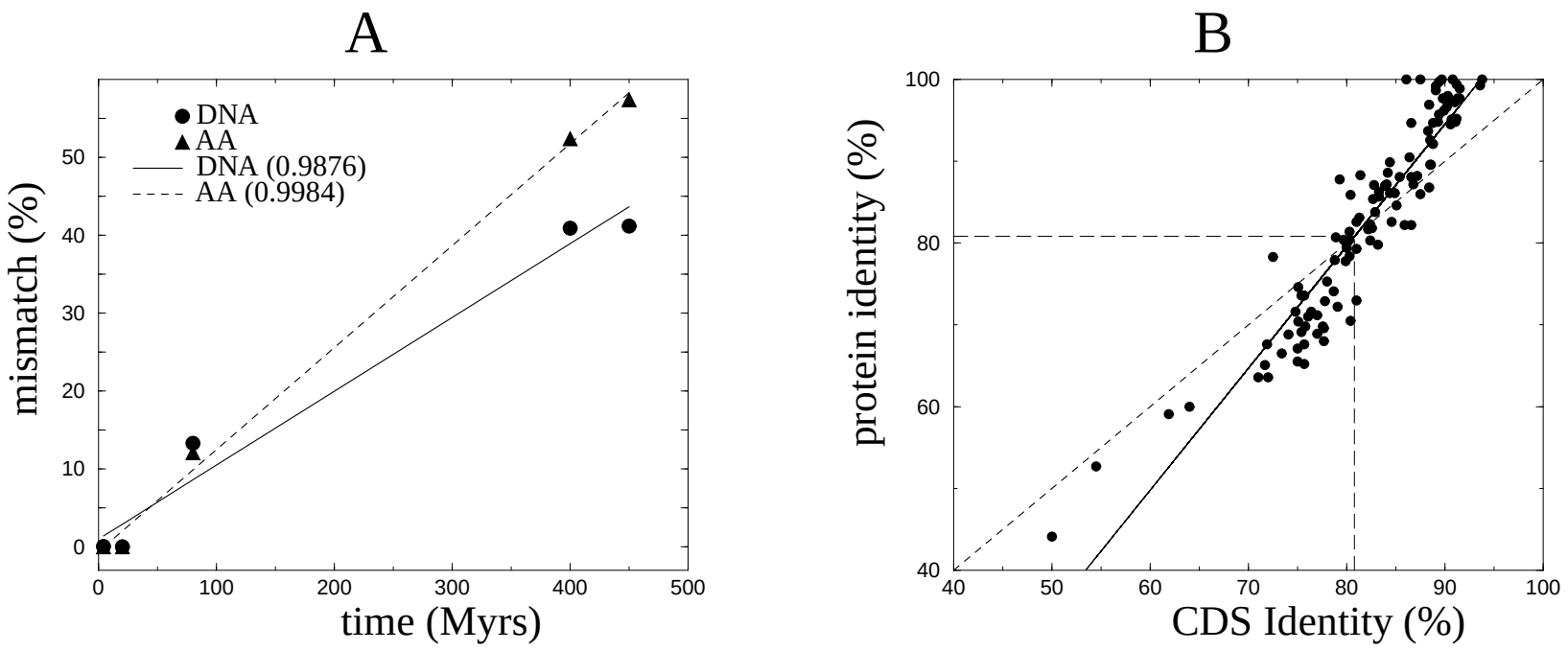
[1] using the program `align` [1] with substitution matrix BLOSUM50

[2] pairwise alignment of the cDNA sequences

[3] pairwise alignment of amino acid sequences

[4] approximate divergence time in million years

Divergence Time and Percent Mismatch



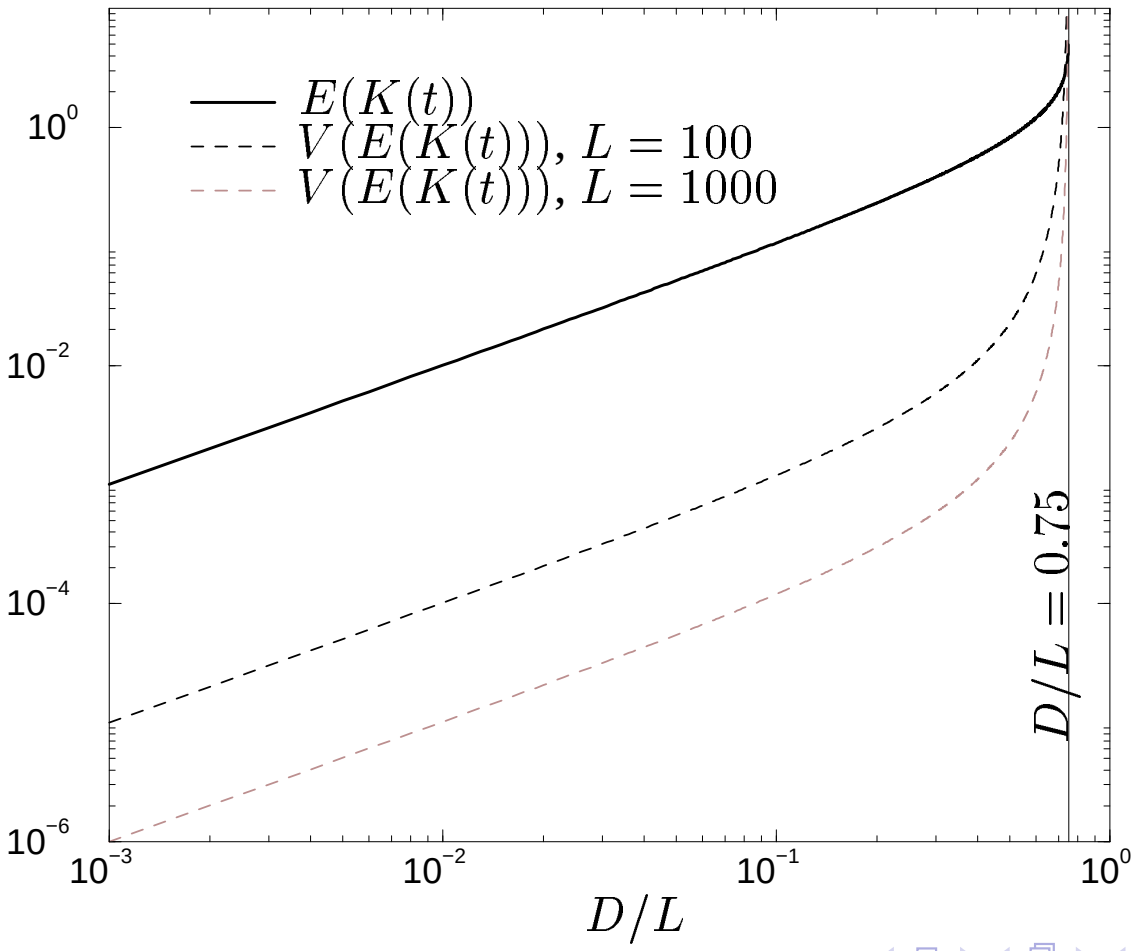
Jukes-Cantor Model

$$M_1 = \begin{array}{c|cccc} & \mathsf{A} & \mathsf{C} & \mathsf{G} & \mathsf{T} \\ \hline \mathsf{A} & 1-3\mu & \mu & \mu & \mu \\ \mathsf{C} & \mu & 1-3\mu & \mu & \mu \\ \mathsf{G} & \mu & \mu & 1-3\mu & \mu \\ \mathsf{T} & \mu & \mu & \mu & 1-3\mu \end{array}$$

Kimura Model

$$M_2 = \begin{array}{c|cccc} & \mathsf{A} & \mathsf{C} & \mathsf{G} & \mathsf{T} \\ \hline \mathsf{A} & 1-\mu_1-2\mu_2 & \mu_2 & \mu_1 & \mu_2 \\ \mathsf{C} & \mu_2 & 1-\mu_1-2\mu_2 & \mu_2 & \mu_1 \\ \mathsf{G} & \mu_1 & \mu_2 & 1-\mu_1-2\mu_2 & \mu_2 \\ \mathsf{T} & \mu_2 & \mu_1 & \mu_2 & 1-\mu_1-2\mu_2 \end{array}$$

$E(K(t))$ vs. D/L


$E(K(t))$
$V(E(K(t)))$, $L = 100$
$V(E(K(t)))$, $L = 1000$
$D/L = 0.75$
10^0
10^{-2}
10^{-4}
10^{-6}
10^{-3}
10^{-2}
10^{-1}
10^0
D/L

Alignment of Coding Sequences

```
              Phe Thr Val Val Asn
sequence 1    TTT ACT GTC GTC AAT
              ::: ::: ::  : : :
sequence 2    TTT ACT GTT GCC ACG
              Phe Thr Val Ala Thr
```

Mutation Pathways

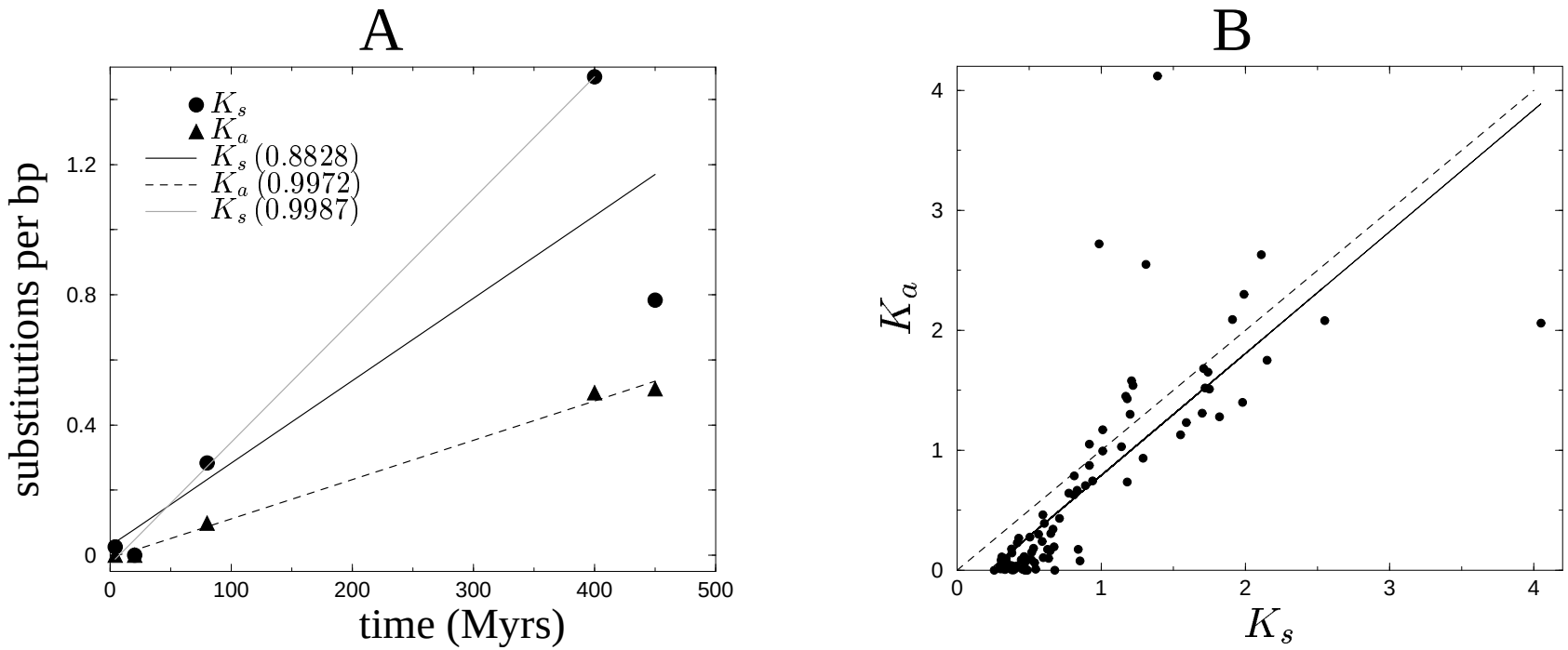
$$\text{pathway 1:} \quad \mathtt{AAT}(\mathrm{Asn}) \xrightarrow{\text{1 non-syn.}} \mathtt{ACT}(\mathrm{Thr}) \xrightarrow{\text{1 syn.}} \mathtt{ACG}(\mathrm{Thr}),$$

$$\text{pathway 2:} \quad \mathtt{AAT}(\mathrm{Asn}) \xrightarrow{\text{1 non-syn.}} \mathtt{AAG}(\mathrm{Lys}) \xrightarrow{\text{1 non-syn.}} \mathtt{ACG}(\mathrm{Thr}).$$

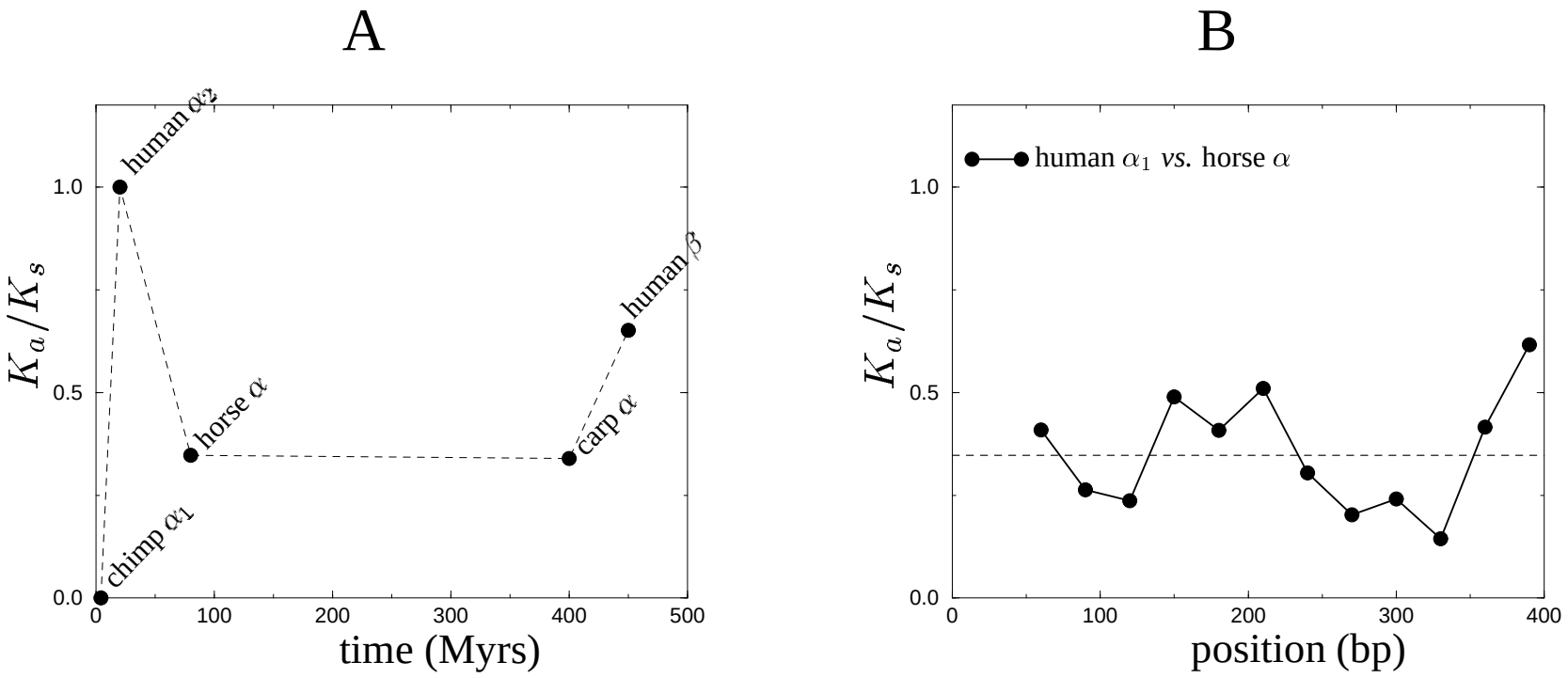
Comparing Human α_1-Globin

	K_s (JC)[2]	K_s (K)	K_a (JC)	K_a (K)	K_a/K_s (JC)	K_a/K_s (K)
human α_2	0.0000	0.0000	0.0000	0.0000	–	–
chimp α_1	0.0256	0.0256	0.0000	0.0000	0.0	0.0
horse α	0.2840	0.2920	0.0987	0.0987	0.3475	0.3380
carp α	1.4700	1.6100	0.4990	0.4990	0.3395	0.3099
human β	0.7840	0.8600	0.5110	0.5120	0.6518	0.5953

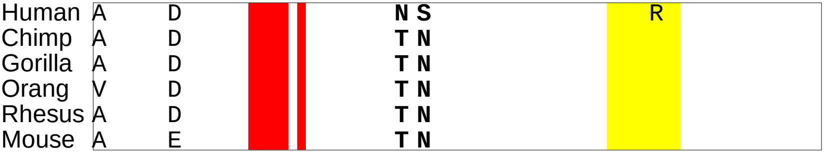
K_a & K_s in Globin Sequences



K_a/K_s in Globin Sequences

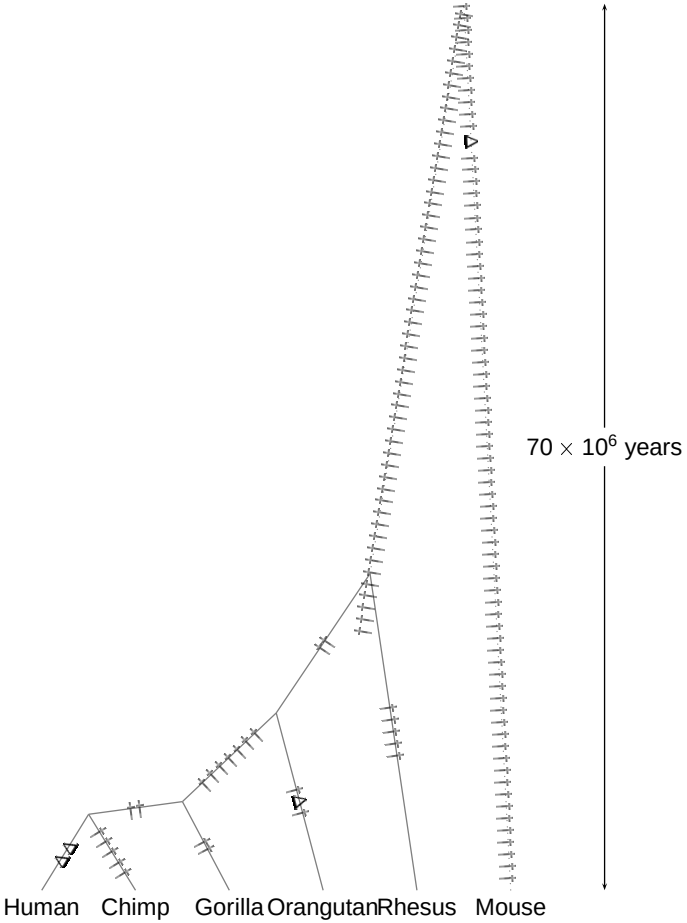


Alignment of FOXP2 Protein Sequences



Human	A	D	**N**	**S**	R
Chimp	A	D	**T**	**N**	
Gorilla	A	D	**T**	**N**	
Orang	V	D	**T**	**N**	
Rhesus	A	D	**T**	**N**	
Mouse	A	E	**T**	**N**	

Substitutions in *FOXP2* Gene



Human Genes under Positive Selection

Gene Name	Function	Human-Chimp Divergence		Human Polymorphism		LR
		D_a	D_s	S_a	S_s	
PRM1	Substitutes for histones in sperm	9	0	0	2	10.1
CMRF35H	Leukocyte membrane antigen	13	0	0	0	9.3
DGAT2L1	Fatty acid synthesis (presumed)	10	1	2	0	6.6
FLJ46156	Unknown	10	1	4	3	6.4
USP26	Testis-specific expression	11	0	1	0	6.2
C15orf2	Testis-specific expression	18	2	12	4	6.1
ABHD1	Unknown	6	0	4	1	5.8
SCML1	Transcriptional repressor, embryonic development (hypot.)	15	1	0	0	5.7
OR2W1	Olfactory receptor	8	0	2	1	5.7
LOC389458	Unknown	8	0	1	0	5.5
APOBEC3F	Antiretroviral factor	11	0	2	1	5.5
MS4A12	Unknown	8	0	1	1	5.4

References

E. W. Myers and W. Miller.
Optimal alignments in linear space.
Computer Applications in the Biosciences, 4:11–17, 1988.