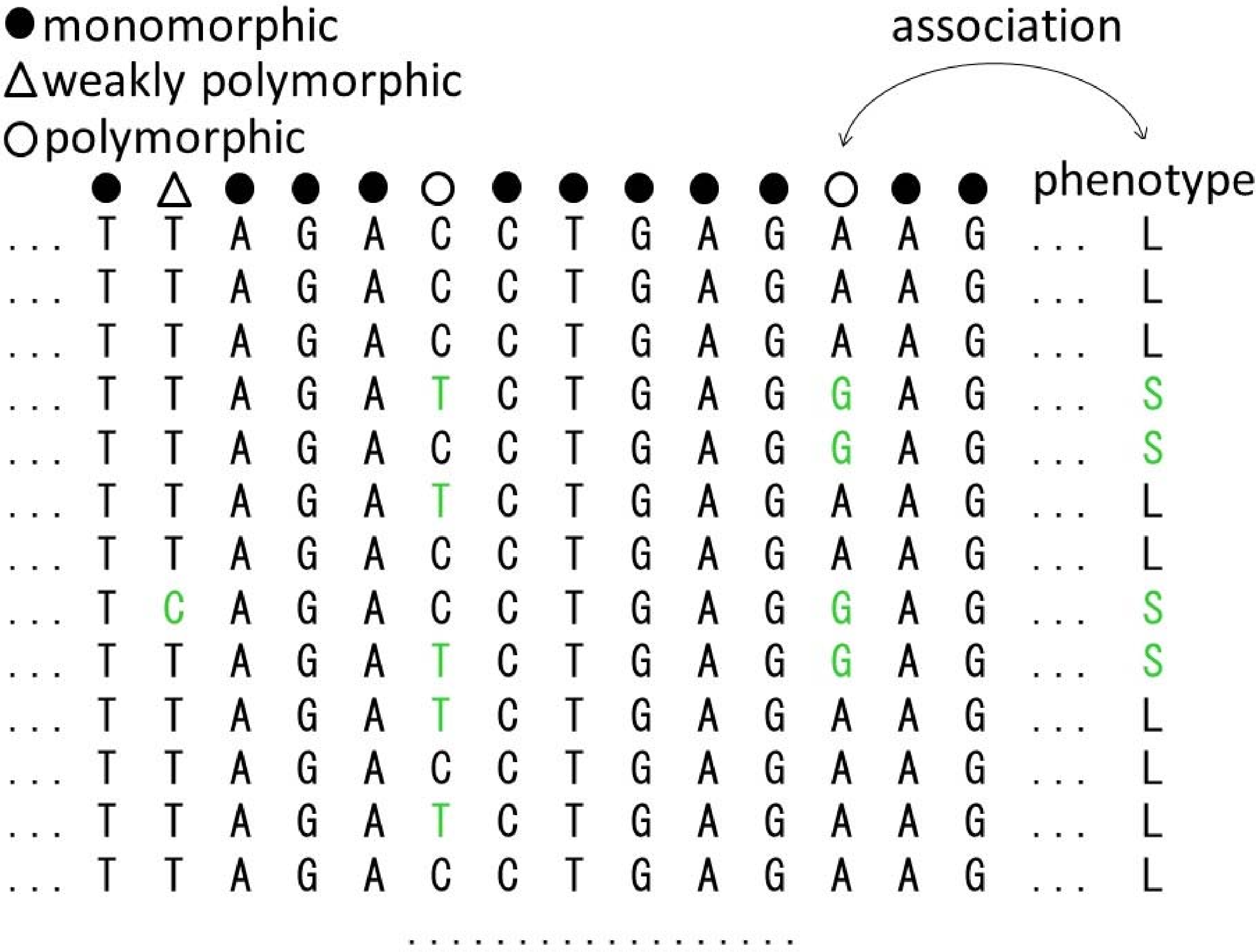
Molecular evolution and phylogenetic tree

Intra-species polymorphism

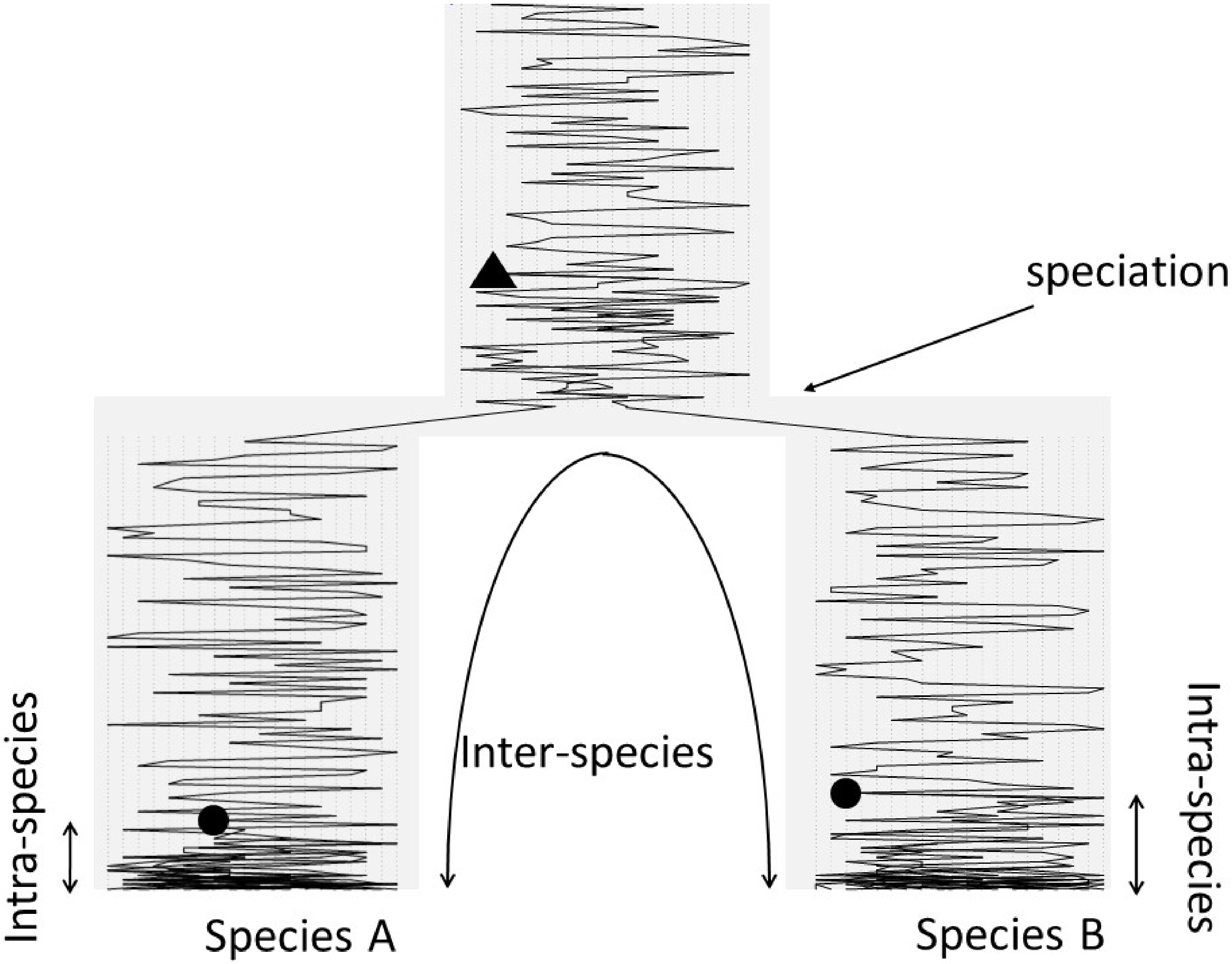


● monomorphic
△ weakly polymorphic
○ polymorphic

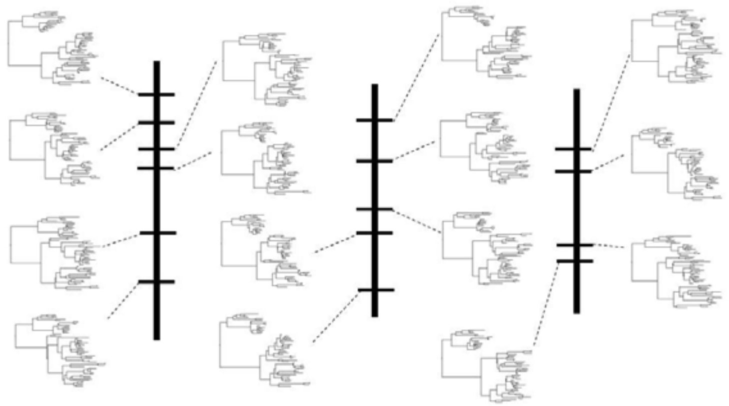
	●	△	●	●	●	○	●	●	●	●	●	○	●	●		phenotype
...	T	T	A	G	A	C	C	T	G	A	G	A	A	G	...	L
...	T	T	A	G	A	C	C	T	G	A	G	A	A	G	...	L
...	T	T	A	G	A	C	C	T	G	A	G	A	A	G	...	L
...	T	T	A	G	A	T	C	T	G	A	G	G	A	G	...	S
...	T	T	A	G	A	C	C	T	G	A	G	G	A	G	...	S
...	T	T	A	G	A	T	C	T	G	A	G	A	A	G	...	L
...	T	T	A	G	A	C	C	T	G	A	G	A	A	G	...	L
...	T	C	A	G	A	C	C	T	G	A	G	G	A	G	...	S
...	T	T	A	G	A	T	C	T	G	A	G	G	A	G	...	S
...	T	T	A	G	A	T	C	T	G	A	G	A	A	G	...	L
...	T	T	A	G	A	C	C	T	G	A	G	A	A	G	...	L
...	T	T	A	G	A	T	C	T	G	A	G	A	A	G	...	L
...	T	T	A	G	A	C	C	T	G	A	G	A	A	G	...	L

....

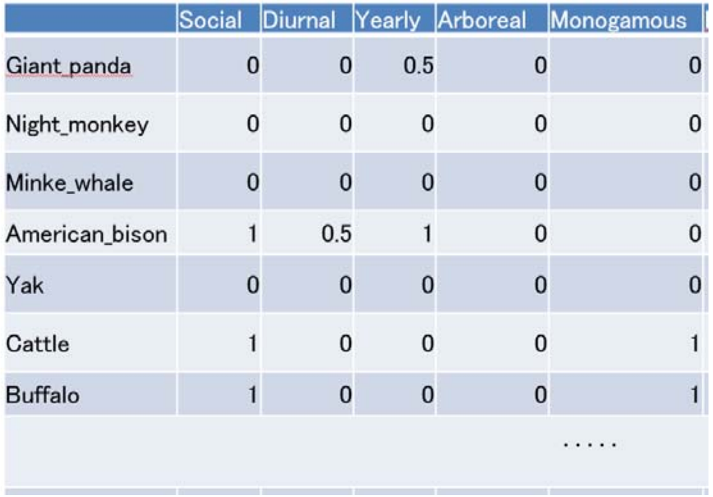
Inter-species diversity and intra-species polymorphism



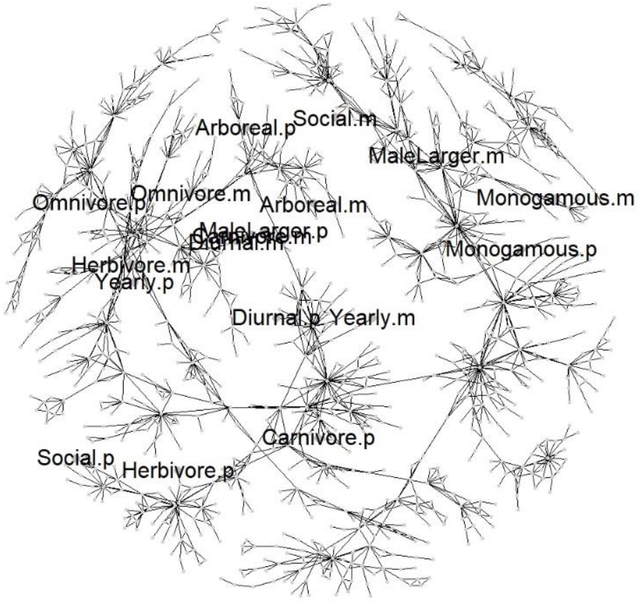
Bridging molecular evolution and phenotypic evolution

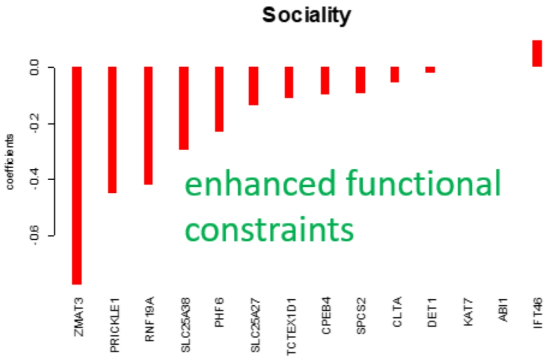


+

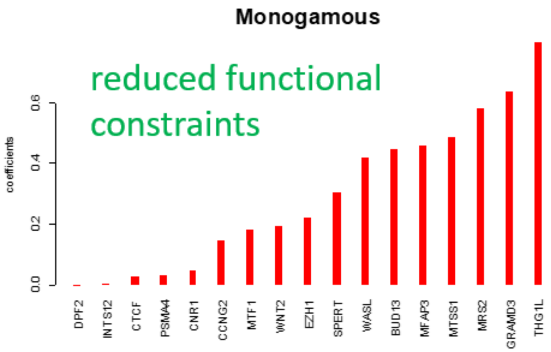


	Social	Diurnal	Yearly	Arboreal	Monogamous
Giant_panda	0	0	0.5	0	0
Night_monkey	0	0	0	0	0
Minke_whale	0	0	0	0	0
American_bison	1	0.5	1	0	0
Yak	0	0	0	0	0
Cattle	1	0	0	0	1
Buffalo	1	0	0	0	1
					…..





enhanced functional constraints



reduced functional constraints

Wu et al (2017) Current Biology **27**: 3025-3033.

Schedule of this course

Dates	Contents
15 May	Molecular evolution and phylogenetic tree
22 May	Rates of molecular evolution
29 May	Population structure and adaptation
5 May	Inferring traits evolution and selection

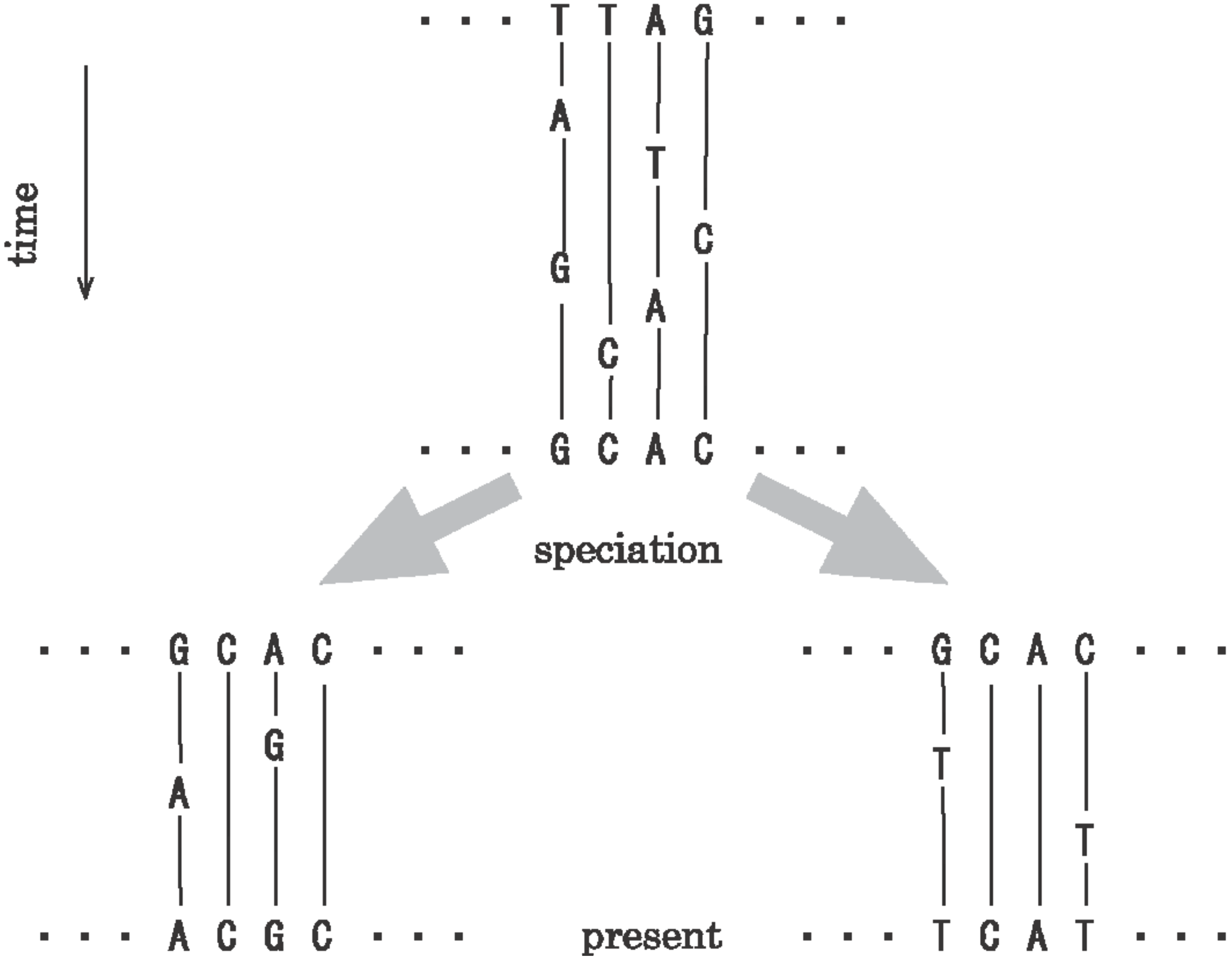
Molecular evolution

The cost of natural selection

Unless selection is very intense, the number of deaths needed to secure the substitution, by natural selection, of one gene for another at a locus, is independent of the intensity of selection. It is about 30 times the number of organisms in a generation. It is suggested that, in horotelic evolution, the mean time taken for each gene substitution is about 300 generations. This accords with the observed slowness of evolution.

Haldane (1957) Genetics. **55**: 511-524

Large collection of molecular evolution provides unbiased estimate of species trees

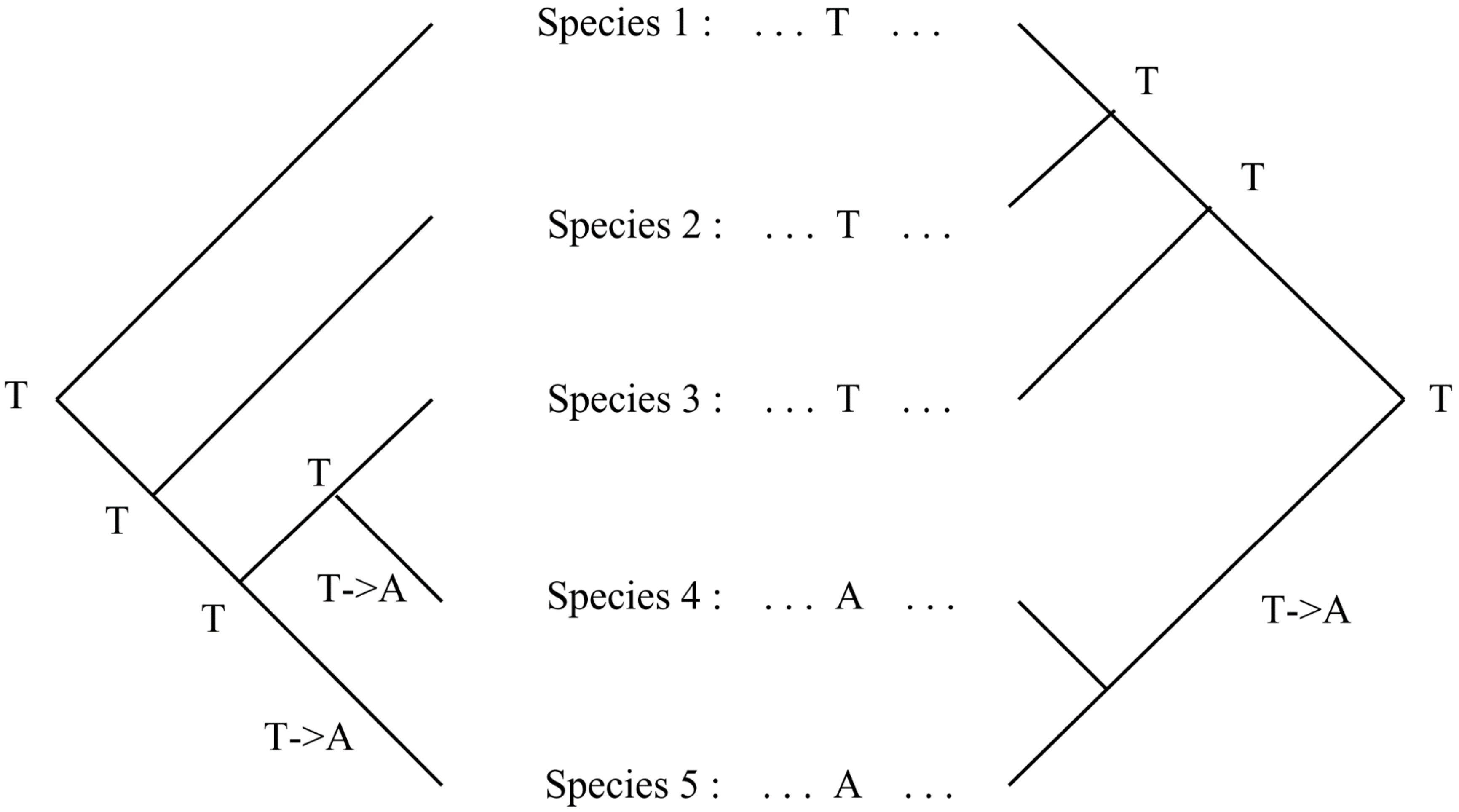


Evolutionary Rate at the Molecular Level

Calculating the rate of evolution in terms of nucleotide substitutions seems to give a value so high that many of the mutations involved must be neutral ones.

Kimura (1968) Nature. **217**: 624-626

Reocnstructing the evolutionary history: Criteria of minimum evolution



Species 1 : . . . T . . .

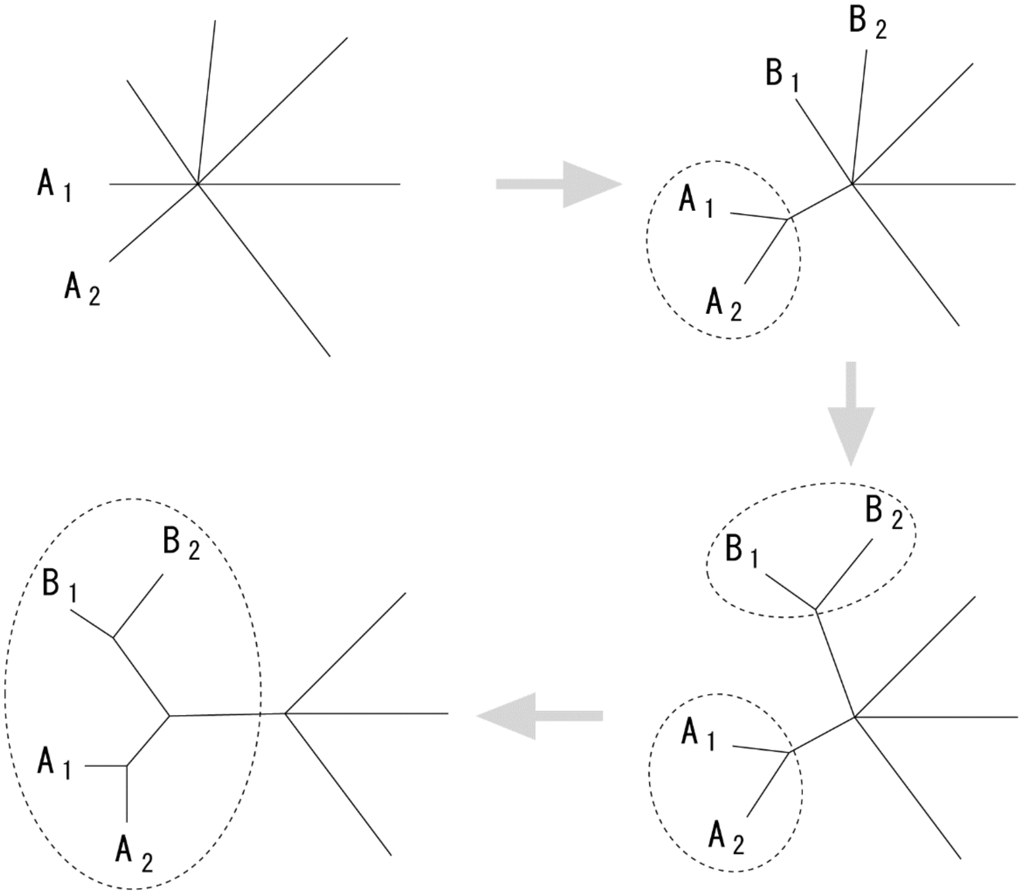
Species 2 : . . . T . . .

Species 3 : . . . T . . .

Species 4 : . . . A . . .

Species 5 : . . . A . . .

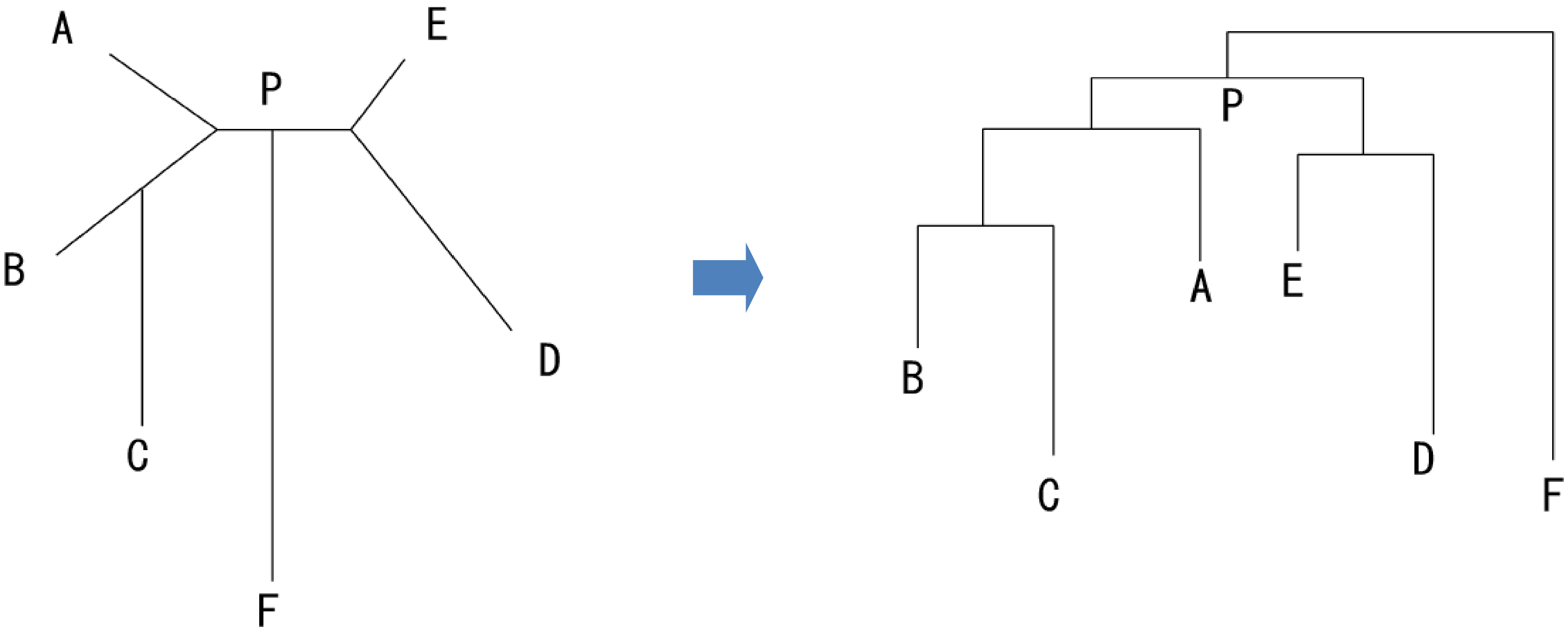
Neighbor joining method



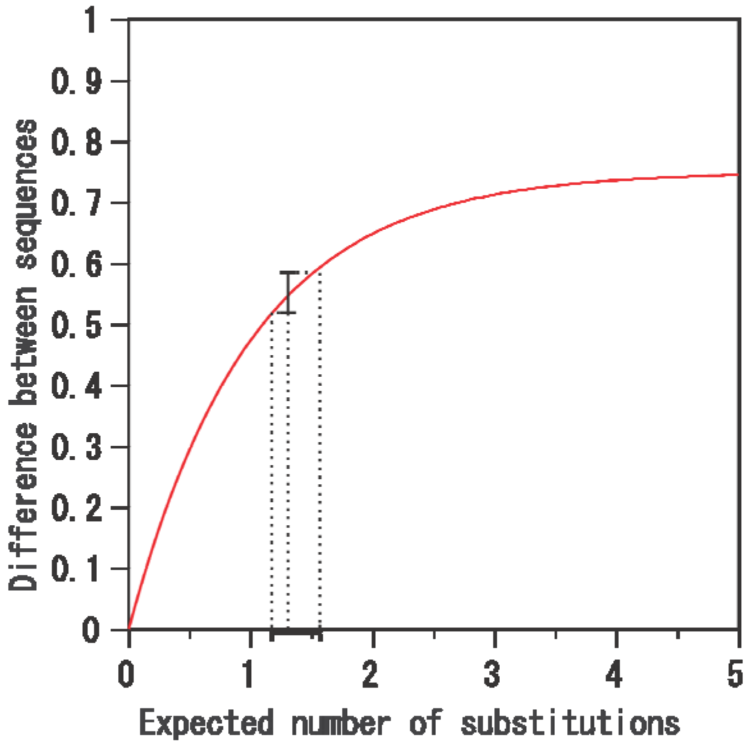
Minimizing the total branch length at each step

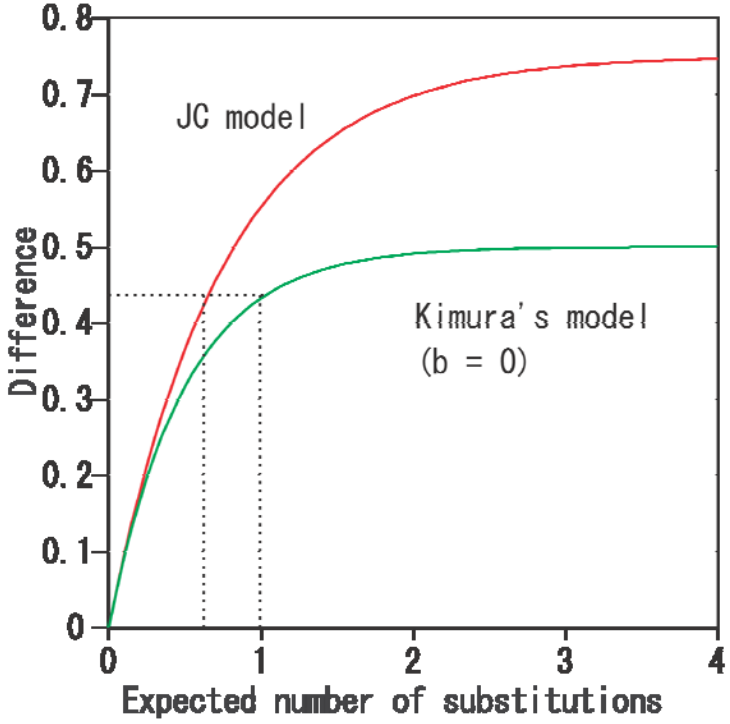
Saitou and Nei (1987) Mol Biol Evol. 4: 406-425.

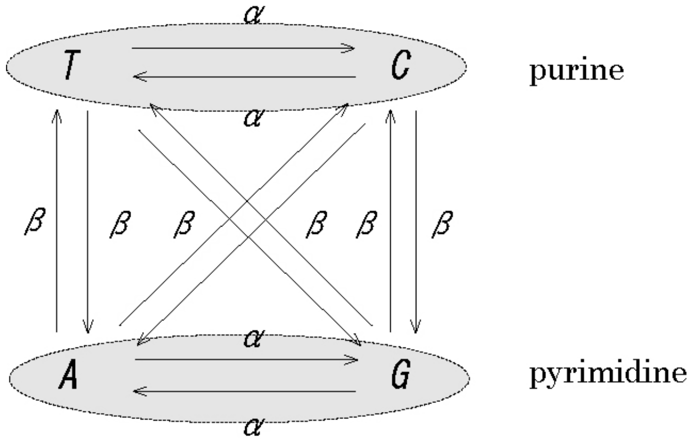
Unrooted tree and rooting by an outgroup



Distance based on statistical model of substitutions







Estimating the pattern of substitutions from sequence comparison

….. GTTCACGGATACTGGTCTAC …..

….. ATTCATGAATACCCGCCTGC …..

$$p_{TT}\,,\,p_{TC}\,,\,p_{TA}\,,\,p_{TG}\,,\,p_{TT}\,,\,p_{CC}\,,\,p_{CA}\,,\,p_{CG}\,,\,p_{AA}\,,\,p_{AG}\,,\,p_{GG}$$

Modeling the proportions of the pattern of sites

$$p_{ij} = p_i \times p_{i\rightarrow j}(t)$$

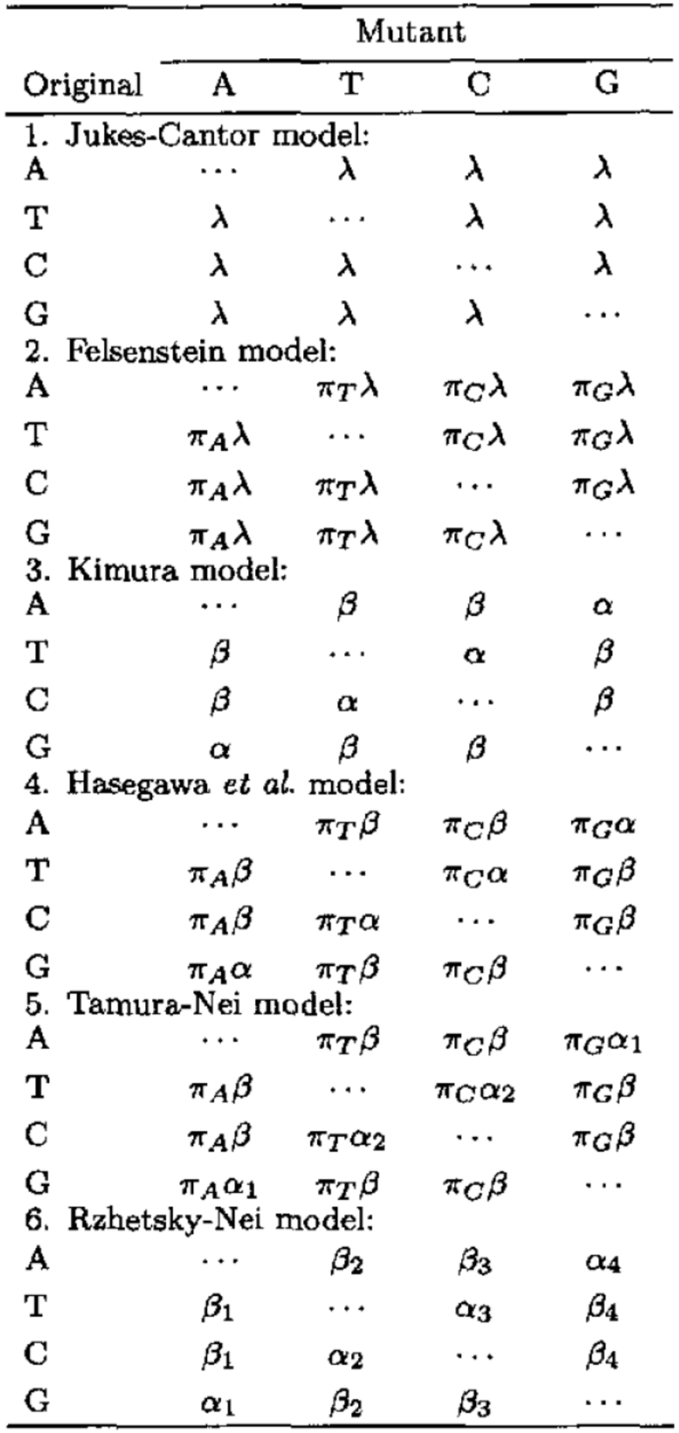
Felsenstein (1981)

$$
\begin{aligned}
p_{ii}(t) &= \exp(-ut) + (1-\exp(-ut))\,\pi_i \\
p_{ij}(t) &= (1-\exp(-ut))\,\pi_j \quad (i \neq j)\ ,
\end{aligned}
$$

Kimura (1981)

$$
\begin{aligned}
p_{TT}(t) = p_{CC}(t) = p_{AA}(t) = p_{GG}(t)& \\
&= \frac{1}{4} + \frac{1}{4}\exp\left(-4\beta t\right) + \frac{1}{2}\exp\left(-2(\alpha+\beta)t\right) \\
p_{TC}(t) = p_{CT}(t) = p_{AG}(t) = p_{GA}(t)& \\
&= \frac{1}{4} + \frac{1}{4}\exp\left(-4\beta t\right) - \frac{1}{2}\exp\left(-2(\alpha+\beta)t\right) \\
p_{TA}(t) = p_{TG}(t) = p_{CA}(t) = p_{CG}(t) = p_{AT}(t) = p_{AC}(t) = p_{GT}(t) = p_{GC}(t)& \\
&= \frac{1}{4} - \frac{1}{4}\exp\left(-4\beta t\right)\ .
\end{aligned}
$$

Models of nucleotide substitutions



	Mutant			
Original	A	T	C	G
1. Jukes-Cantor model:				
A	$\cdots$	λ	λ	λ
T	λ	$\cdots$	λ	λ
C	λ	λ	$\cdots$	λ
G	λ	λ	λ	$\cdots$
2. Felsenstein model:				
A	$\cdots$	$\pi_T\lambda$	$\pi_C\lambda$	$\pi_G\lambda$
T	$\pi_A\lambda$	$\cdots$	$\pi_C\lambda$	$\pi_G\lambda$
C	$\pi_A\lambda$	$\pi_T\lambda$	$\cdots$	$\pi_G\lambda$
G	$\pi_A\lambda$	$\pi_T\lambda$	$\pi_C\lambda$	$\cdots$
3. Kimura model:				
A	$\cdots$	β	β	α
T	β	$\cdots$	α	β
C	β	α	$\cdots$	β
G	α	β	β	$\cdots$
4. Hasegawa *et al.* model:				
A	$\cdots$	$\pi_T\beta$	$\pi_C\beta$	$\pi_G\alpha$
T	$\pi_A\beta$	$\cdots$	$\pi_C\alpha$	$\pi_G\beta$
C	$\pi_A\beta$	$\pi_T\alpha$	$\cdots$	$\pi_G\beta$
G	$\pi_A\alpha$	$\pi_T\beta$	$\pi_C\beta$	$\cdots$
5. Tamura-Nei model:				
A	$\cdots$	$\pi_T\beta$	$\pi_C\beta$	$\pi_G\alpha_1$
T	$\pi_A\beta$	$\cdots$	$\pi_C\alpha_2$	$\pi_G\beta$
C	$\pi_A\beta$	$\pi_T\alpha_2$	$\cdots$	$\pi_G\beta$
G	$\pi_A\alpha_1$	$\pi_T\beta$	$\pi_C\beta$	$\cdots$
6. Rzhetsky-Nei model:				
A	$\cdots$	β_2	β_3	α_4
T	β_1	$\cdots$	α_3	β_4
C	β_1	α_2	$\cdots$	β_4
G	α_1	β_2	β_3	$\cdots$

Modeling the pattern of molecular evolution and the transition probabilities

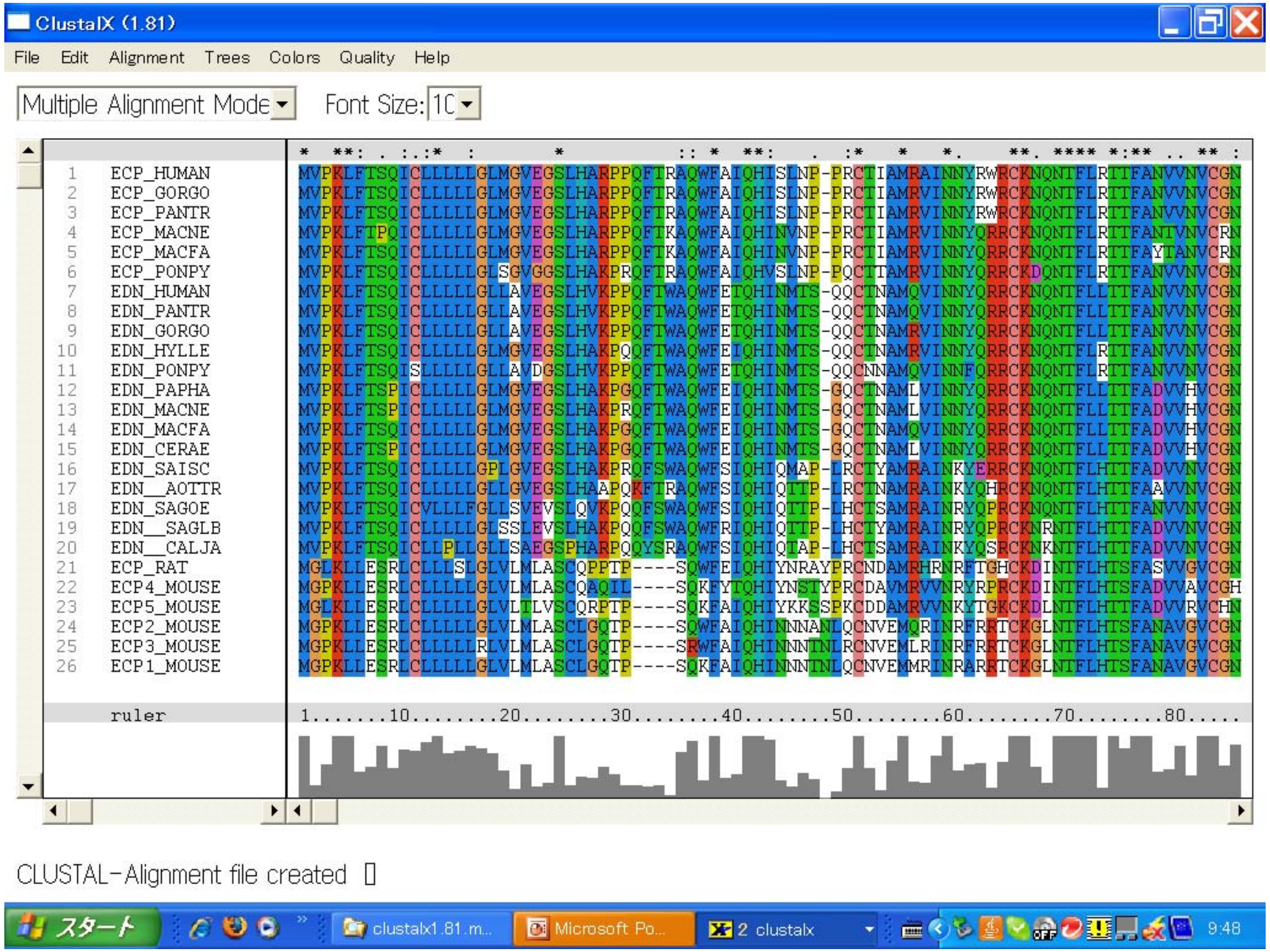
Felsenstein (1981)

$$\begin{aligned} p_{ii}(t) &= \exp(-ut) + (1 - \exp(-ut))\,\pi_i \\ p_{ij}(t) &= (1 - \exp(-ut))\,\pi_j \quad (i \neq j)\ , \end{aligned}$$

Kimura (1980)

$$\begin{aligned} p_{TT}(t) = p_{CC}(t) = p_{AA}(t) = p_{GG}(t) \\ &= \frac{1}{4} + \frac{1}{4}\exp\left(-4\beta t\right) + \frac{1}{2}\exp\left(-2(\alpha+\beta)t\right) \\ p_{TC}(t) = p_{CT}(t) = p_{AG}(t) = p_{GA}(t) \\ &= \frac{1}{4} + \frac{1}{4}\exp\left(-4\beta t\right) - \frac{1}{2}\exp\left(-2(\alpha+\beta)t\right) \\ p_{TA}(t) = p_{TG}(t) = p_{CA}(t) = p_{CG}(t) = p_{AT}(t) = p_{AC}(t) = p_{GT}(t) = p_{GC}(t) \\ &= \frac{1}{4} - \frac{1}{4}\exp\left(-4\beta t\right)\ . \end{aligned}$$

Multiple alignment identifies rate heterogeneity among sites



Multiple Alignment Mode

Font Size: 10

```
                         *  **: . :.:*  :       *          :: *  **:   .  :*   *   *.    **. **** *:** .. ** :
 1   ECP_HUMAN           MVPKLFTSQICLLLLLGLMGVEGSLHARPPQFTRAQWFAIQHISLNP-PRCTIAMRAINNYRWRCKNQNTFLRTTFANVVNVCGN
 2   ECP_GORGO           MVPKLFTSQICLLLLLGLMGVEGSLHARPPQFTRAQWFAIQHISLNP-PRCTIAMRVINNYRWRCKNQNTFLRTTFANVVNVCGN
 3   ECP_PANTR           MVPKLFTSQICLLLLLGLMGVEGSLHARPPQFTRAQWFAIQHISLNP-PRCTIAMRVINNYRWRCKNQNTFLRTTFANVVNVCGN
 4   ECP_MACNE           MVPKLFTPQICLLLLLGLMGVEGSLHARPPQFTKAQWFAIQHINVNP-PRCTIAMRVINNYQRRCKNQNTFLRTTFANTVNVCRN
 5   ECP_MACFA           MVPKLFTSQICLLLLLGLMGVEGSLHARPPQFTKAQWFAIQHINVNP-PRCTIAMRVINNYQRRCKNQNTFLRTTFAYTANVCRN
 6   ECP_PONPY           MVPKLFTSQICLLLLLGLSGVGGSLHAKPRQFTRAQWFAIQHVSLNP-PQCTTAMRVINNYQRRCKDQNTFLRTTFANVVNVCGN
 7   EDN_HUMAN           MVPKLFTSQICLLLLLGLLAVEGSLHVKPPQFTWAQWFETQHINMTS-QQCTNAMQVINNYQRRCKNQNTFLLTTFANVVNVCGN
 8   EDN_PANTR           MVPKLFTSQICLLLLLGLLAVEGSLHVKPPQFTWAQWFETQHINMTS-QQCTNAMQVINNYQRRCKNQNTFLLTTFANVVNVCGN
 9   EDN_GORGO           MVPKLFTSQICLLLLLGLLAVEGSLHVKPPQFTWAQWFETQHINMTS-QQCTNAMRVINNYQRRCKNQNTFLLTTFANVVNVCGN
10   EDN_HYLLE           MVPKLFTSQICLLLLLGLMGVEGSLHAKPQQFTWAQWFEIQHINMTS-QQCTNAMRVINNYQRRCKNQNTFLRTTFANVVNVCGN
11   EDN_PONPY           MVPKLFTSQISLLLLLGLLAVDGSLHVKPPQFTWAQWFETQHINMTS-QQCNNAMQVINNFQRRCKNQNTFLRTTFANVVNVCGN
12   EDN_PAPHA           MVPKLFTSPICLLLLLGLMGVEGSLHAKPGQFTWAQWFEIQHINMTS-GQCTNAMLVINNYQRRCKNQNTFLLTTFADVVHVCGN
13   EDN_MACNE           MVPKLFTSPICLLLLLGLMGVEGSLHAKPRQFTWAQWFEIQHINMTS-GQCTNAMLVINNYQRRCKNQNTFLLTTFADVVHVCGN
14   EDN_MACFA           MVPKLFTSQICLLLLLGLMGVEGSLHAKPGQFTWAQWFEIQHINMTS-GQCTNAMQVINNYQRRCKNQNTFLLTTFADVVHVCGN
15   EDN_CERAE           MVPKLFTSPICLLLLLGLMGVEGSLHAKPGQFTWAQWFEIQHINMTS-GQCTNAMLVINNYQRRCKNQNTFLLTTFADVVHVCGN
16   EDN_SAISC           MVPKLFTSQICLLLLLGPLGVEGSLHAKPRQFSWAQWFSIQHIQMAP-LRCTYAMRAINKYERRCKNQNTFLHTTFADVVNVCGN
17   EDN__AOTTR          MVPKLFTSQICLLLLLGLLGVEGSLHAAPQKFTRAQWFSIQHIQTTP-LRCTNAMRAINKYQHRCKNQNTFLHTTFAAVVNVCGN
18   EDN_SAGOE           MVPKLFTSQICVLLLFGLLSVEVSLQVKPQQFSWAQWFSIQHIQTTP-LHCTSAMRAINRYQPRCKNQNTFLHTTFANVVNVCGN
19   EDN__SAGLB          MVPKLFTSQICLLLLLGLSSLEVSLHAKPQQFSWAQWFRIQHIQTTP-LHCTYAMRAINRYQPRCKNRNTFLHTTFADVVNVCGN
20   EDN__CALJA          MVPKLFTSQICLLPLLGLLSAEGSPHARPQQYSRAQWFSIQHIQTAP-LHCTSAMRAINKYQSRCKNKNTFLHTTFADVVNVCGN
21   ECP_RAT             MGLKLLESRLCLLLSLGLVLMLASCQPPTP----SQWFEIQHIYNRAYPRCNDAMRHRNRFTGHCKDINTFLHTSFASVVGVCGN
22   ECP4_MOUSE          MGPKLLESRLCLLLLLGLVLMLASCQAQIL----SQKFYTQHIYNSTYPRCDAVMRVVNRYRPRCKDINTFLHTSFADVVAVCGH
23   ECP5_MOUSE          MGLKLLESRLCLLLLLGLVLTLVSCQRPTP----SQKFAIQHIYKKSSPKCDDAMRVVNKYTGKCKDLNTFLHTTFADVVRVCHN
24   ECP2_MOUSE          MGPKLLESRLCLLLLLGLVLMLASCLGQTP----SQWFAIQHINNNANLQCNVEMQRINRFRRTCKGLNTFLHTSFANAVGVCGN
25   ECP3_MOUSE          MGPKLLESRLCLLLLLRLVLMLASCLGQTP----SRWFAIQHINNNTNLRCNVEMLRINRFRRTCKGLNTFLHTSFANAVGVCGN
26   ECP1_MOUSE          MGPKLLESRLCLLLLLGLVLMLASCLGQTP----SQKFAIQHINNNTNLQCNVEMMRINRARRTCKGLNTFLHTSFANAVGVCGN

     ruler               1.......10........20........30........40........50........60........70........80.....
```

CLUSTAL-Alignment file created

A note on maximum likelihood and parsimony: the case of complete observation

$n_1, n_2, \dots, n_k$: occurrence (1) / non-occurrence (0) of evolutionary events along the branches 1,2,…,k

- Assuming no multiple evolutionary events along a branch
- Likelihood

$$\mathrm{L}(b|n_1, n_2, \dots, n_k) = \prod_{i=1}^{k} b^{n_i}(1-b)^{1-n_i}$$

$b_1 = b_2 = \cdots = b_k = b$: probability of occurrence along a branch

- Assuming rare occurrence: $\lambda = kb$

$$\mathrm{L}(\lambda|n_T) = e^{-\lambda}\frac{\lambda^{n_T}}{n_T!}$$

$n_T = n_1 + n_2 + \cdots + n_k$:

the number of evolutionary events assuming tree T

A note on maximum likelihood and parsimony: the case of complete observation

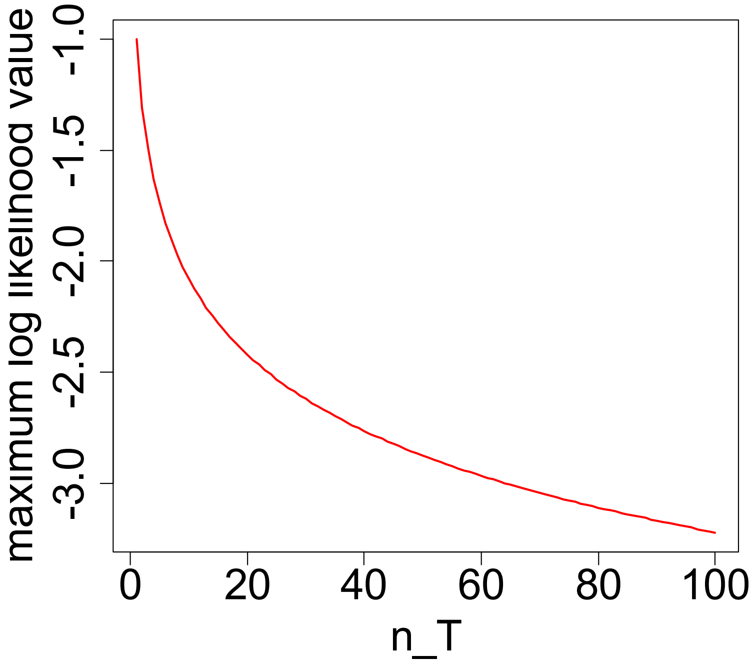
- Maximum likelihood estimate and parsimony

$$\hat{\lambda} = n_T$$

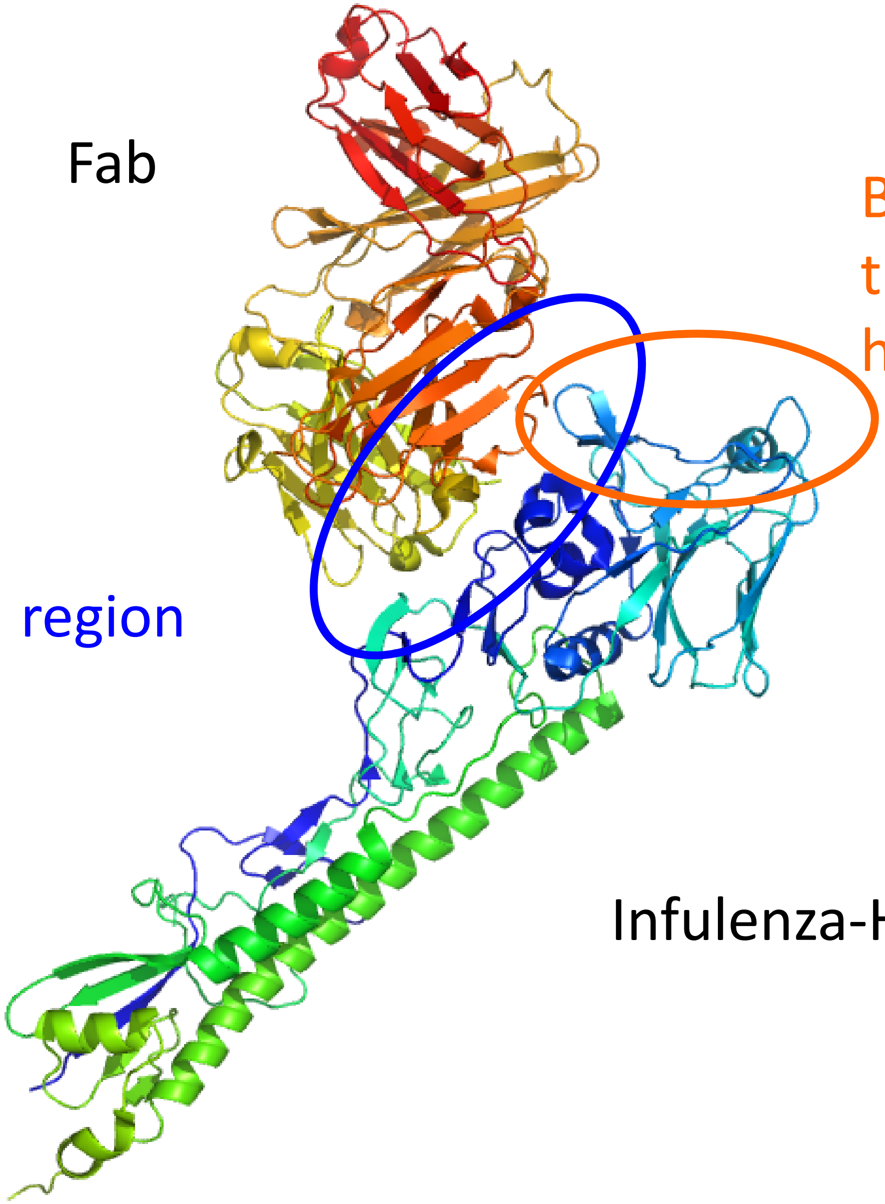
$$\log\mathrm{L}(\hat{\lambda}|n_T) = -\hat{\lambda} + n_T \log \hat{\lambda} - \sum_{i=1}^{n_T} \log i$$

$$= -n_T + n_T \log n_T - \sum_{i=1}^{n_T} \log i$$

Likelihood approach is consistent with the criterion of minimum evolution



Selection pressure varies among the region of the protein



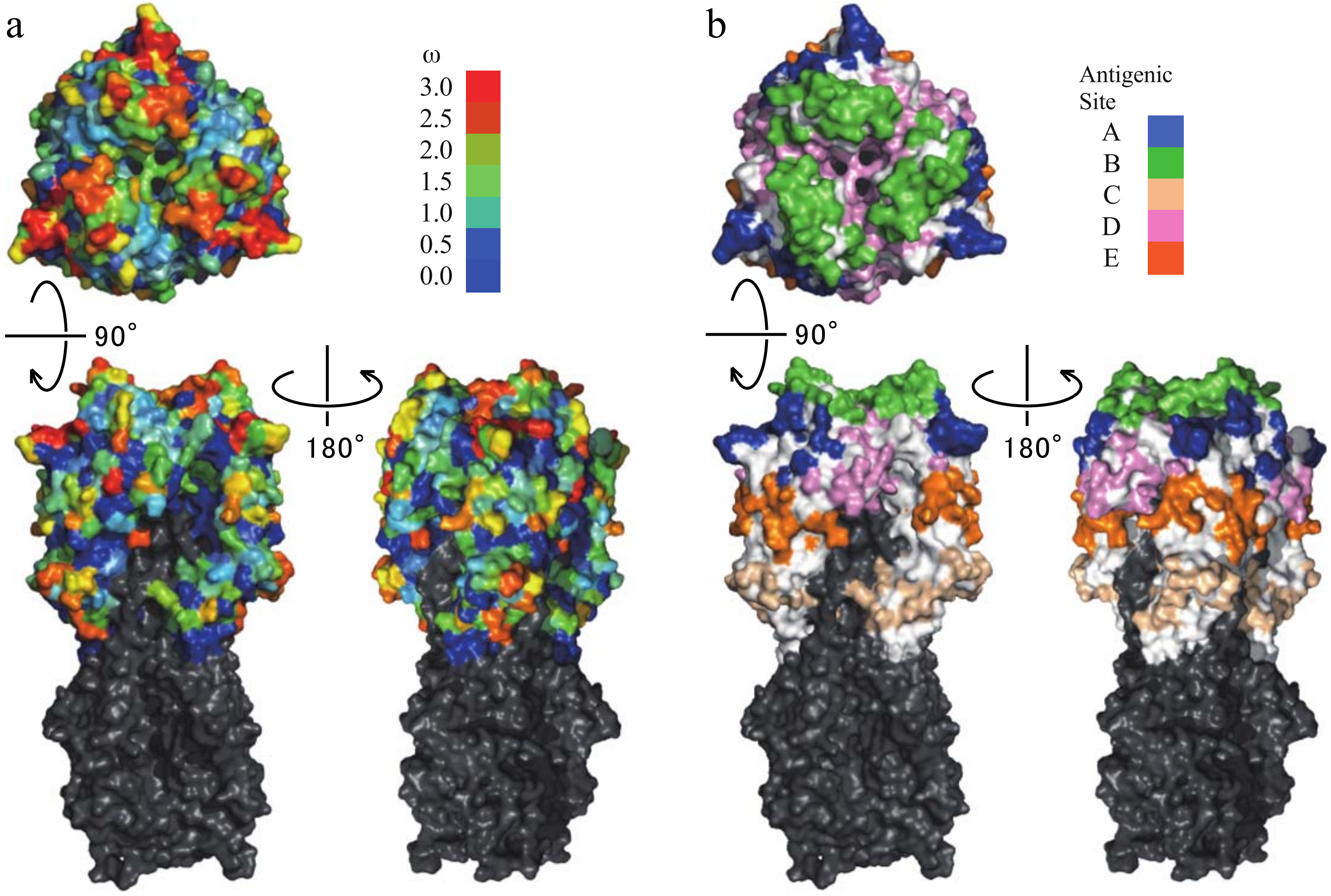
Fab

Binding region to the receptor of the host cell

Binding region to Fab

Infulenza-HA protein

Regions under the diversifying selection and antigenic sites

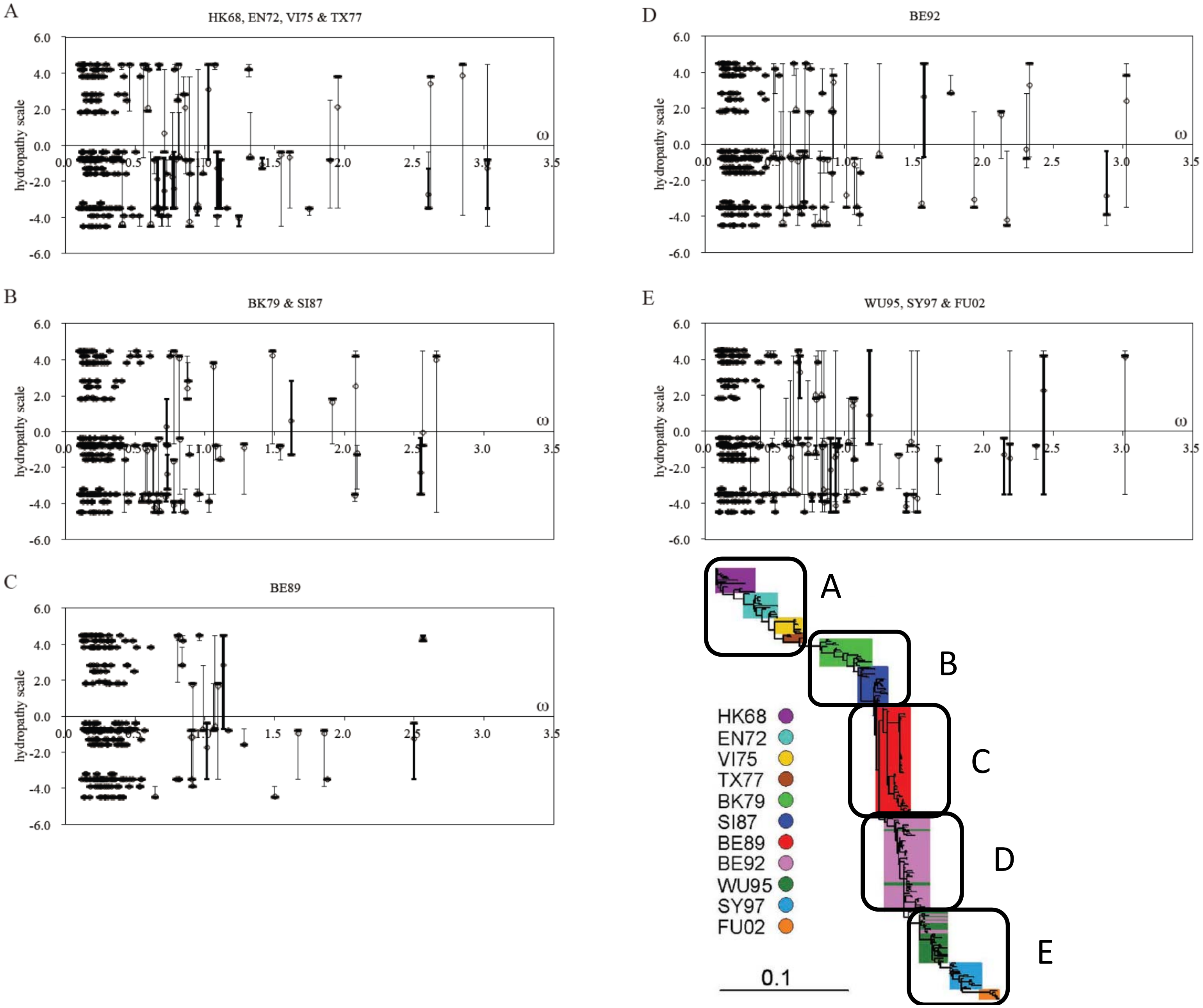


The spatial distribution of selection pressure

The distributions of the antigenic sites A to E.

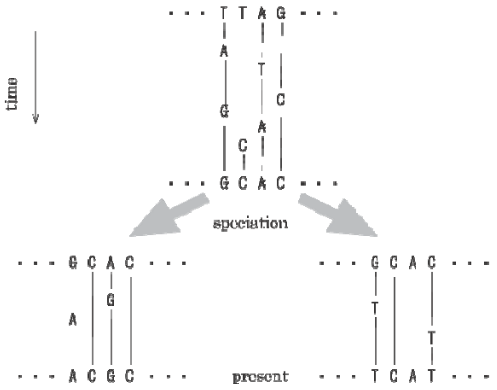
Watabe and Kishino (2013) Mol Biol Evol. 30: 2714-2722.

Changes in chemico-physical feature and the ω values of substitutions



Rate of molecular evolution, mutation rate, and fixation rate

- Fitness of mutations: advantageous (rare), neutral, deleterious
- Mutations deleted from the populations are not observed by comparing the genomes
- Molecular evolution: mutations fixed to the populations



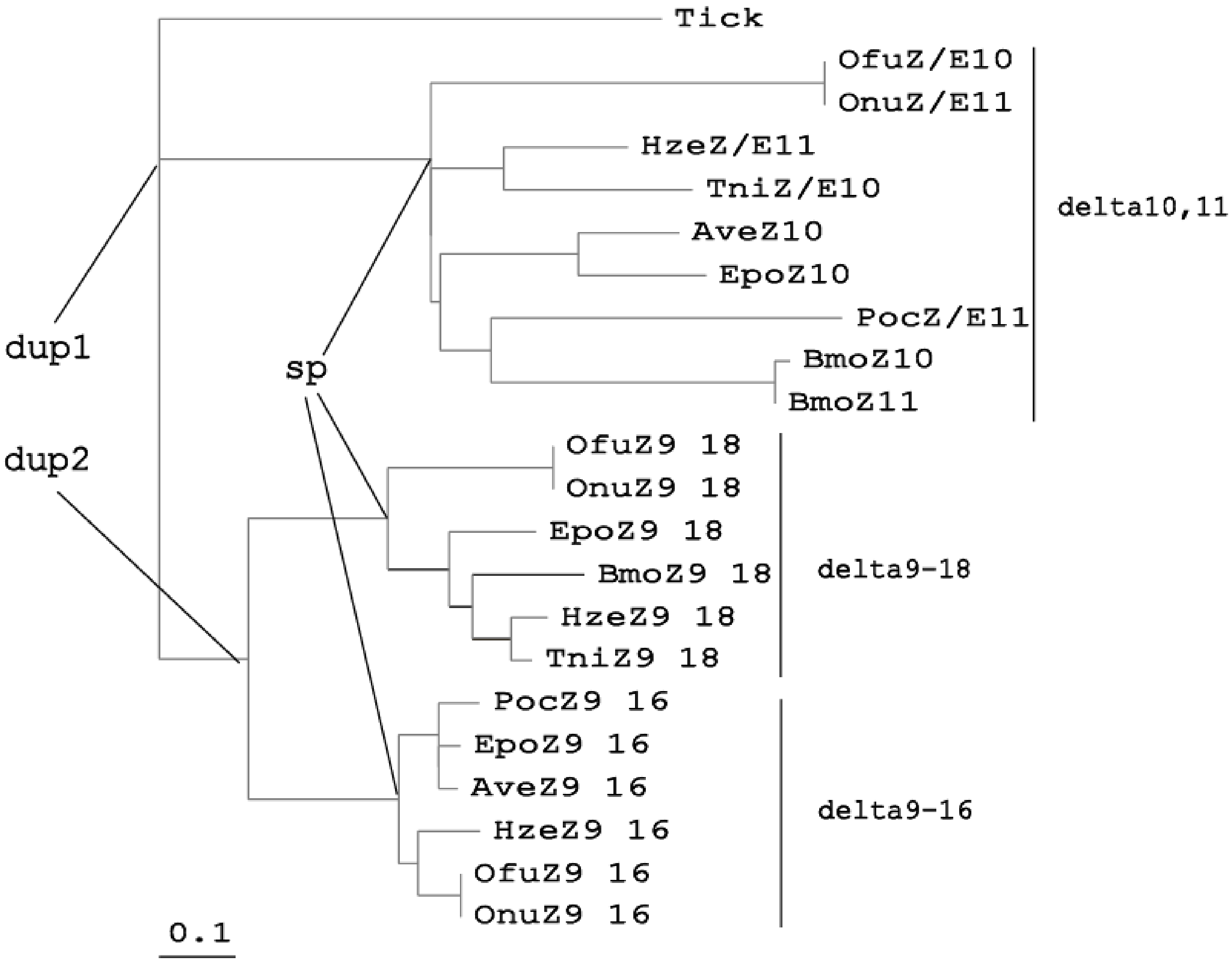
$$
\begin{aligned}
r &= \mu \times 2N \times f \\
&= \mu \times 2N \times p \times \frac{1}{2N} \\
&= \mu p
\end{aligned}
$$

neutral	p
deleterious	$1-p$

- Rate of molecular evolution is the product of mutation rate and proportion of neutral mutations

Generation length, mutagens	Functional constraints, diversifying selection
(affects on the whole genome)	(vary among genes in the genome)

Desaturase and the origin of sex pheromone



Analyzing desaturase sequences

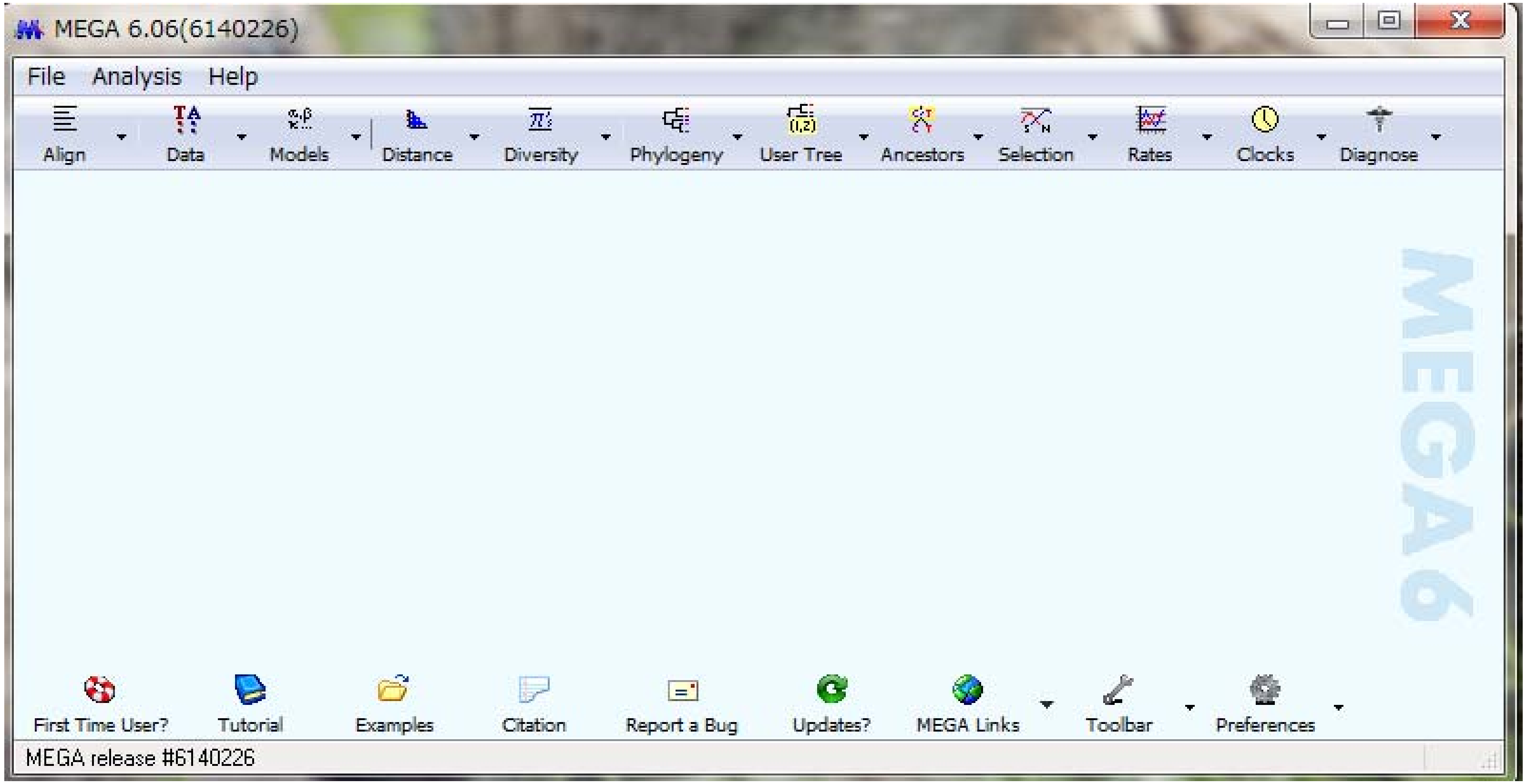
Amino acid sequences of desaturase (fasta format)

```
>HzeZ9_18
MPPQGQTGGSWVLYETDAVNEDTDAPVIVPPSAEKREWKIVWRNVILMGMLHIGGVYGAY
LFLTTAMWRTCIFAVVLYICSGLGITAGAHRLWAHKSYKARLPLRLMLTLFNTLAFQDAV
IDWARDHRMHHKYSETDADPHNATRGFFFAHVGWLLVRKHPQIKAKGHTIDLSDLKSDPI
LRFQKKYYLFLMPLVCFILPCYIPT-LWGESLWNAYFVCSIFRYVYVLNVTWLVNSAAHL
WGAKPYDKNINPVETRPVSLVVLGEGFHNYHHTFPWDYKTAELGDYSLNLTKLFIDTMAA
IGWAYDLKTVSTDVIQKRVKRTGDGSHPVWGWDDHEVHQADKKLAAIINPEKT
>TniZ9_18
MPPQGQTGGSWVLYETDAVNTDTDAPVIVPPSAEKREWKIVWRNVILMGMLHIGGVYGAY
LFLTKAMWLTDLFAFFLYLCSGLGITAGAHRLWAHKSYKARLPLRLLLTLFNTLAFQDAV
IDWARDHRMHHKYSETDADPHNATRGFFFSHVGWLLVRKHPQIKAKGHTIDLSDLKSDPI
LRFQKKYYLTLMPLICFILPSYIPT-LWGESAFNAFFVCSIFRYVYVLNVTWLVNSAAHL
WGSKPYDKNINPVETRPVSLVVLGEGFHNYHHTFPWDYKTAELGDYSLNFTKMFIDFMAS
IGWAYDLKTVSTDVIQKRVKRTGDGSHAVWGWDDHEVHQEDKKLAAIINPEKT
>BmoZ9_18
MPPQRKQEASWVLYEADANNLPEDAPPHVPPSAEKRPWKIVWRNVILFFILHVGGVYGGY
LFLFKAMWRTSIFAIFLYLCSGLGITAGAHRLWAHKSYKARLPLRILLTIFNTIAFQDAV
VDWARDHRMHHKYSETDADPHNATRGFFFSHIGWLLLRKHPEIKAKGHTVDVNELRNDPI
LRFQKKYYQILMPLACFIMPTYVPT-LWGETVWNSFYVCAIFRYVYVLNITWLVNSAAHM
WGSKPYDKNINPVETRPVSLVVLGEGFHNYHHTFPWDYKTAELGDYSLNLSKLFIDFMAK
IDWAYDLKTVSTDVIQKRTKRTGDGSHPVWGYDVGEVATEDKTDTTNLVNSKV
>EpoZ9_18
MPPQGQPPAAWVLEESDATTDDKDVAVAVPPSAEKRKLSIVWRNVILFVLLHTGAVYGGY
LFFTKAMWATKFFAFFLYLCSGLGITAGAHRLWAHKSYKARLPLRILLTLFNTIAFQDSV
LDWARDHRMHHKYSETDADPHNATRGFFFSHVGWLLVRKHPQIKAKGHTIDMSDLCSDPV
LRFQKKYYLTLMPLFCFILPTYIPT-LWGESLWNAYFVAAIFRYCYVLNVTWLVNSAAHK
WGDRPYDKNINPVETKPVSLVVFGEGFHNYHHTFPWDYKTAELGGYSLNISKLFIDTMAK
IGWAYDLKSVSPDIVEKRVKRTGDGSHEVWGWDDKDVPAEQKAAATIINPEKT
. . . . . . .
```

desaturase.fasta

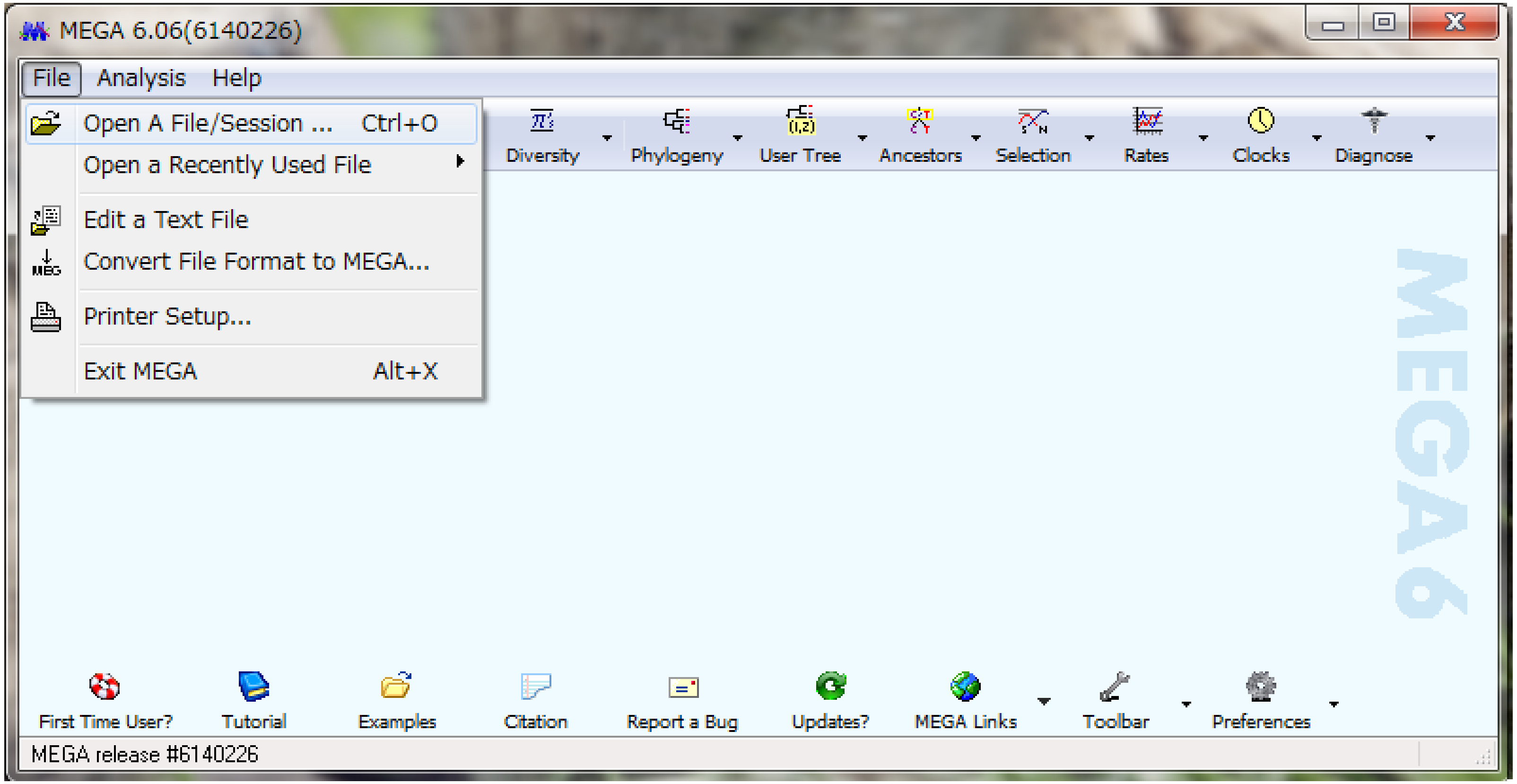
Start MEGA

Double click the shortcut



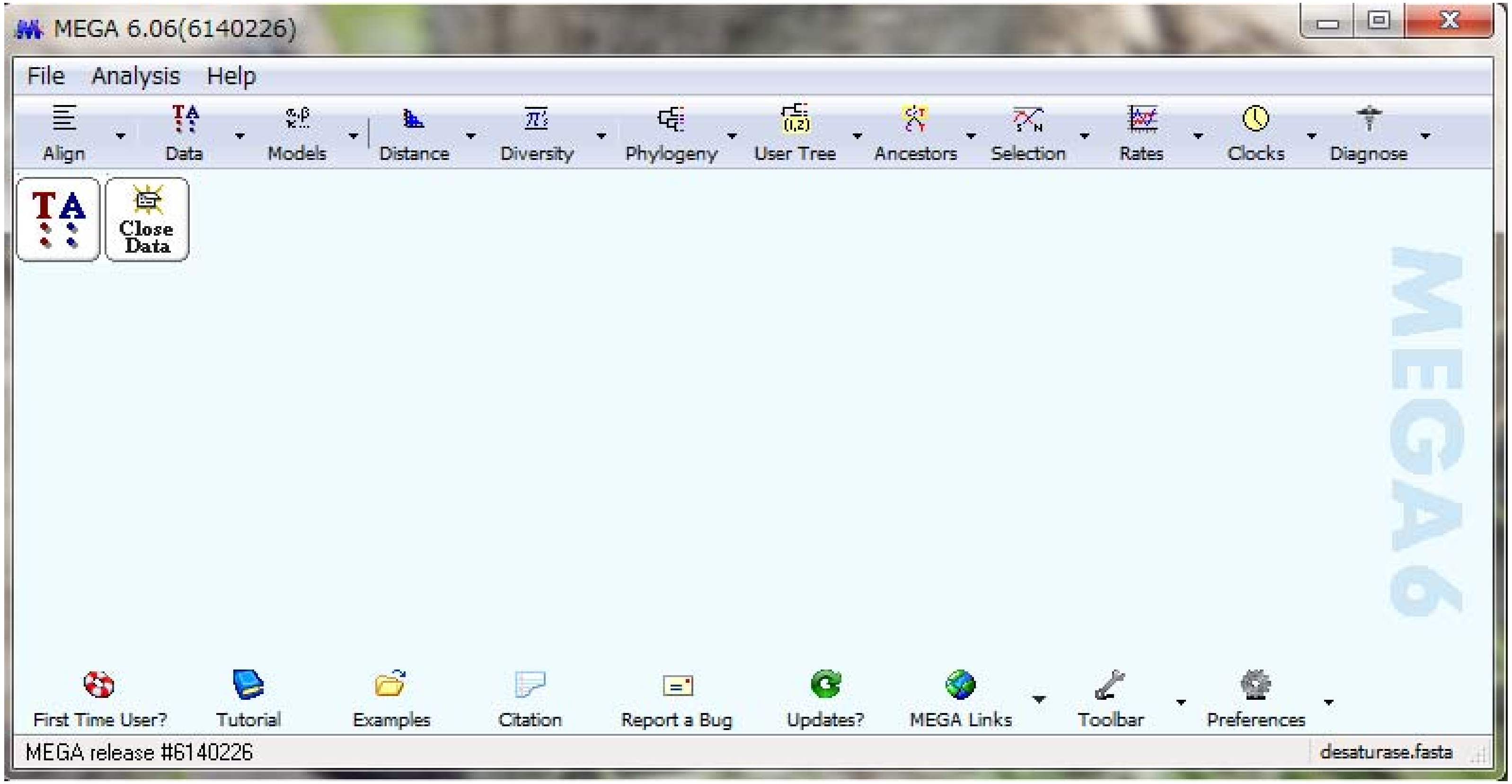
Open the file of amino acid sequences

[File][Open A File/Session]



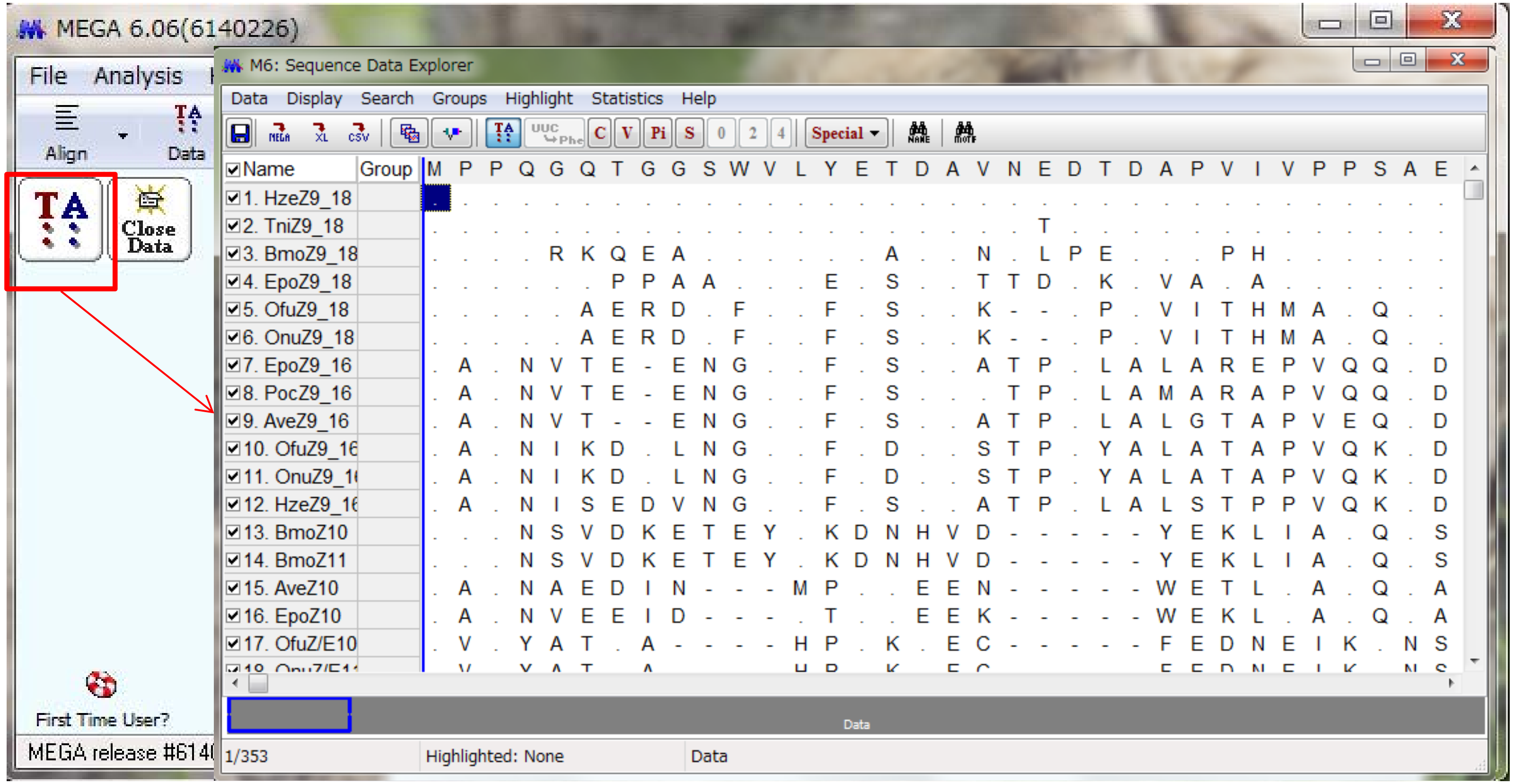
Open the file of amino acid sequences

”Analyze” -> ”Protain Sequence”



View the data

Double click



☑Name	Group	M	P	P	Q	G	Q	T	G	G	S	W	V	L	Y	E	T	D	A	V	N	E	D	T	D	A	P	V	I	V	P	P	S	A	E
☑1. HzeZ9_18		.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
☑2. TniZ9_18		.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.
☑3. BmoZ9_18		.	.	.	.	R	K	Q	E	A	.	.	.	.	.	.	A	.	.	N	.	L	P	E	.	.	.	P	H	.	.	.	.	.	.
☑4. EpoZ9_18		.	.	.	.	.	.	P	P	A	A	.	.	.	E	.	S	.	.	T	T	D	.	K	.	V	A	.	A	.	.	.	.	.	.
☑5. OfuZ9_18		.	.	.	.	.	A	E	R	D	.	F	.	.	F	.	S	.	.	K	-	-	.	P	.	V	I	T	H	M	A	.	Q	.	.
☑6. OnuZ9_18		.	.	.	.	.	A	E	R	D	.	F	.	.	F	.	S	.	.	K	-	-	.	P	.	V	I	T	H	M	A	.	Q	.	.
☑7. EpoZ9_16		.	A	.	N	V	T	E	-	E	N	G	.	.	F	.	S	.	.	A	T	P	.	L	A	L	A	R	E	P	V	Q	Q	.	D
☑8. PocZ9_16		.	A	.	N	V	T	E	-	E	N	G	.	.	F	.	S	.	.	.	T	P	.	L	A	M	A	R	A	P	V	Q	Q	.	D
☑9. AveZ9_16		.	A	.	N	V	T	-	-	E	N	G	.	.	F	.	S	.	.	A	T	P	.	L	A	L	G	T	A	P	V	E	Q	.	D
☑10. OfuZ9_16		.	A	.	N	I	K	D	.	L	N	G	.	.	F	.	D	.	.	S	T	P	.	Y	A	L	A	T	A	P	V	Q	K	.	D
☑11. OnuZ9_1		.	A	.	N	I	K	D	.	L	N	G	.	.	F	.	D	.	.	S	T	P	.	Y	A	L	A	T	A	P	V	Q	K	.	D
☑12. HzeZ9_16		.	A	.	N	I	S	E	D	V	N	G	.	.	F	.	S	.	.	A	T	P	.	L	A	L	S	T	P	P	V	Q	K	.	D
☑13. BmoZ10		.	.	.	N	S	V	D	K	E	T	E	Y	.	K	D	N	H	V	D	-	-	-	-	-	Y	E	K	L	I	A	.	Q	.	S
☑14. BmoZ11		.	.	.	N	S	V	D	K	E	T	E	Y	.	K	D	N	H	V	D	-	-	-	-	-	Y	E	K	L	I	A	.	Q	.	S
☑15. AveZ10		.	A	.	N	A	E	D	I	N	-	-	-	M	P	.	.	E	E	N	-	-	-	-	-	W	E	T	L	.	A	.	Q	.	A
☑16. EpoZ10		.	A	.	N	V	E	E	I	D	-	-	-	.	T	.	.	E	E	K	-	-	-	-	-	W	E	K	L	.	A	.	Q	.	A
☑17. OfuZ/E10		.	V	.	Y	A	T	.	A	-	-	-	-	H	P	.	K	.	E	C	-	-	-	-	-	F	E	D	N	E	I	K	.	N	S

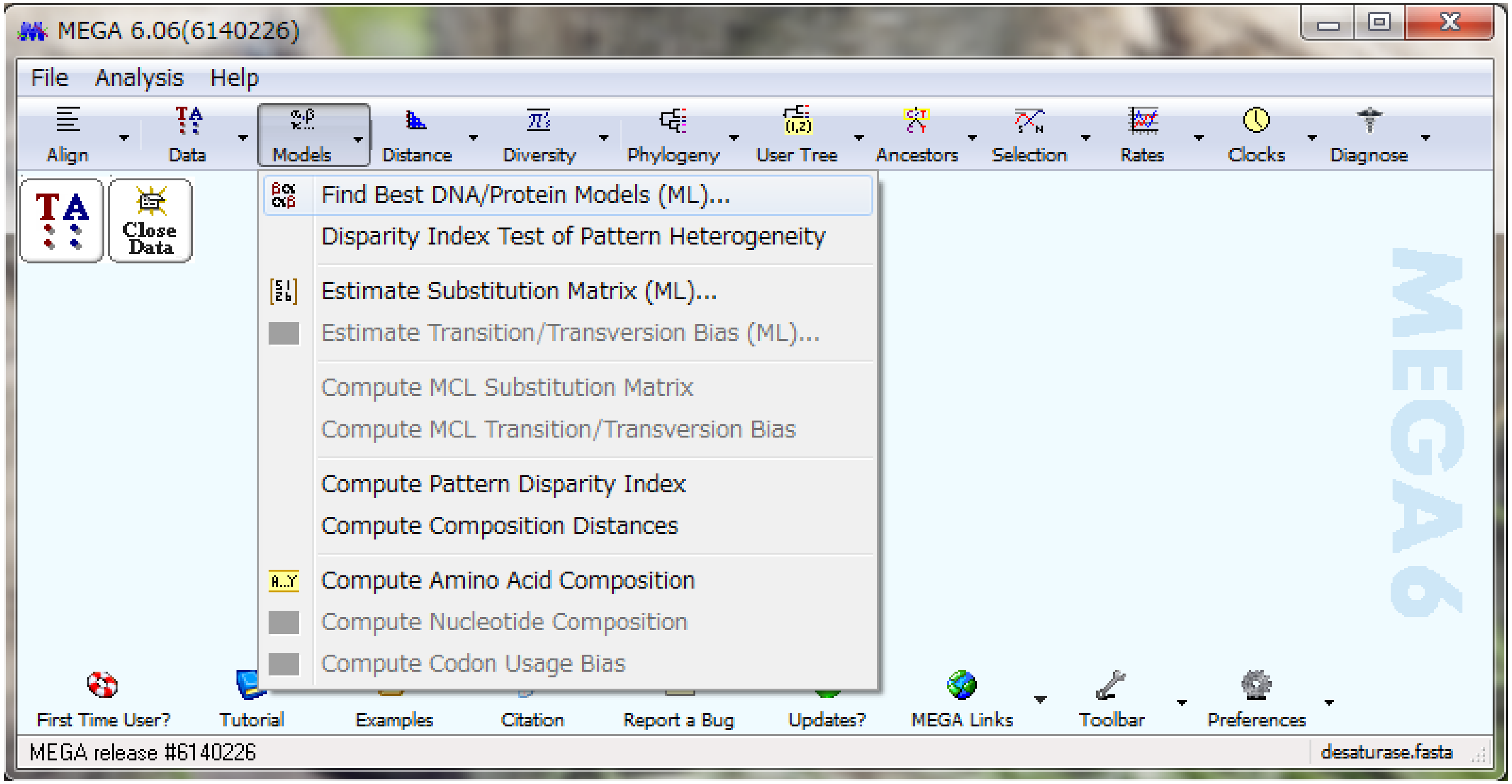
1/353

Highlighted: None

Data

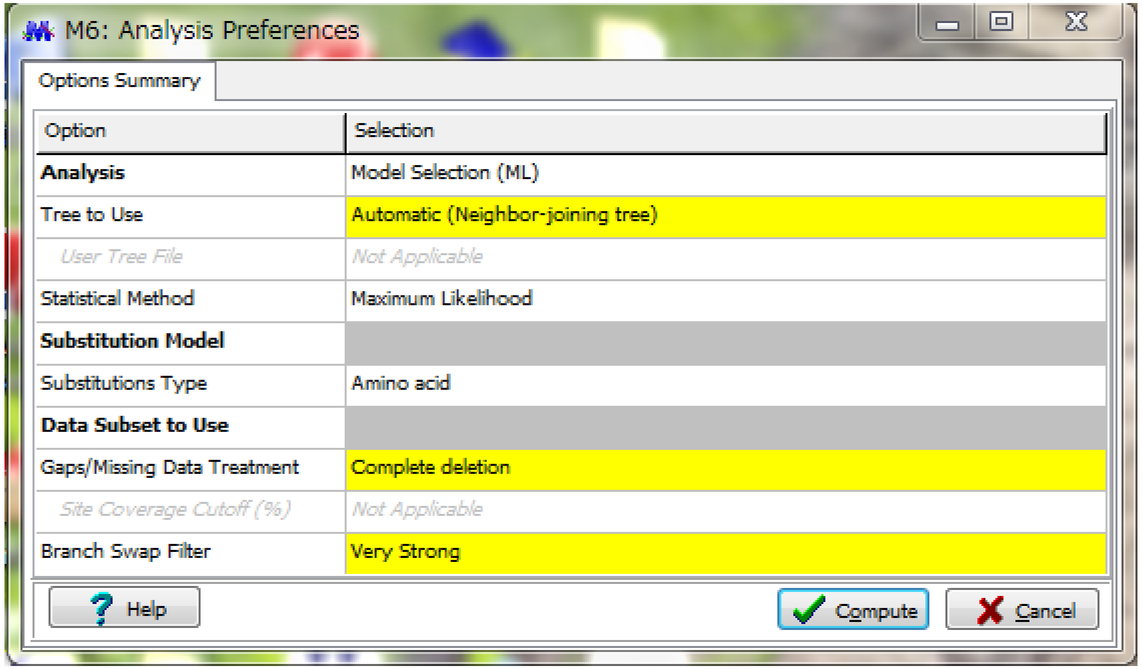
Comparing the models of amino acid replacements

[Models][Find Best DNA/Protein Models (ML)]


MEGA 6.06(6140226)
File
Analysis
Help
Align
Data
Models
Distance
Diversity
Phylogeny
User Tree
Ancestors
Selection
Rates
Clocks
Diagnose
TA
Close Data
Find Best DNA/Protein Models (ML)...
Disparity Index Test of Pattern Heterogeneity
Estimate Substitution Matrix (ML)...
Estimate Transition/Transversion Bias (ML)...
Compute MCL Substitution Matrix
Compute MCL Transition/Transversion Bias
Compute Pattern Disparity Index
Compute Composition Distances
Compute Amino Acid Composition
Compute Nucleotide Composition
Compute Codon Usage Bias
MEGA6
First Time User?
Tutorial
Examples
Citation
Report a Bug
Updates?
MEGA Links
Toolbar
Preferences
MEGA release #6140226
desaturase.fasta

Comparing the models of amino acid replacements

List of options



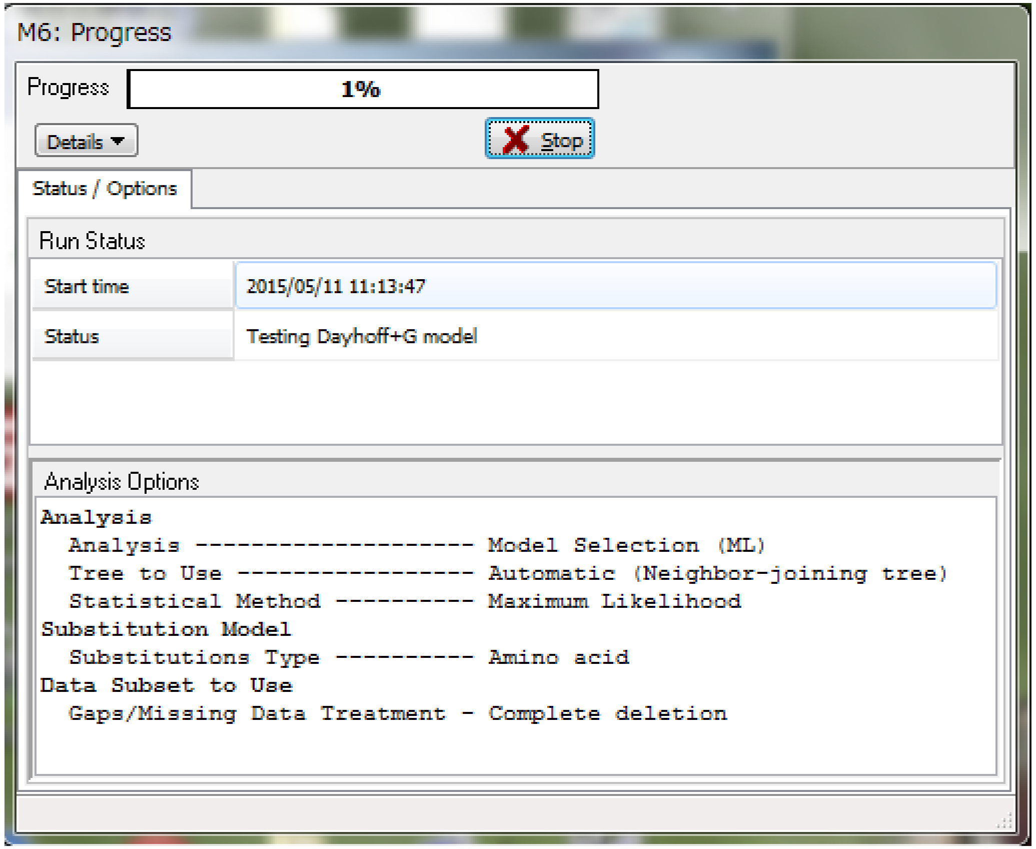
Option	Selection
Analysis	Model Selection (ML)
Tree to Use	Automatic (Neighbor-joining tree)
User Tree File	*Not Applicable*
Statistical Method	Maximum Likelihood
Substitution Model	
Substitutions Type	Amino acid
Data Subset to Use	
Gaps/Missing Data Treatment	Complete deletion
Site Coverage Cutoff (%)	*Not Applicable*
Branch Swap Filter	Very Strong

You can change the options of the items colored yellow.

Here, keep the default values, and click [Compute].

Comparing the models of amino acid replacements

You will see the progress of the analysis



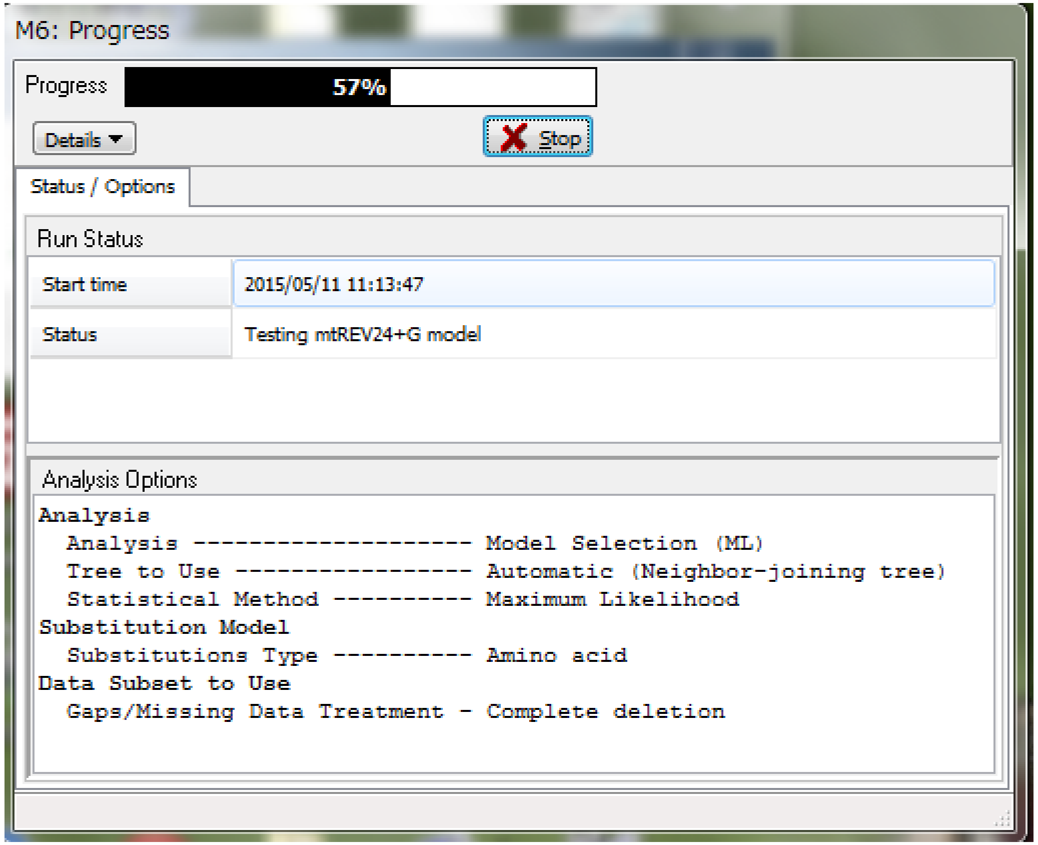
Start time 2015/05/11 11:13:47

Status Testing Dayhoff+G model

```
Analysis
  Analysis -------------------- Model Selection (ML)
  Tree to Use ----------------- Automatic (Neighbor-joining tree)
  Statistical Method ---------- Maximum Likelihood
Substitution Model
  Substitutions Type ---------- Amino acid
Data Subset to Use
  Gaps/Missing Data Treatment - Complete deletion
```

Comparing the models of amino acid replacements

You will see the progress of the analysis



Start time: 2015/05/11 11:13:47

Status: Testing mtREV24+G model

```
Analysis
  Analysis -------------------- Model Selection (ML)
  Tree to Use ----------------- Automatic (Neighbor-joining tree)
  Statistical Method ---------- Maximum Likelihood
Substitution Model
  Substitutions Type ---------- Amino acid
Data Subset to Use
  Gaps/Missing Data Treatment - Complete deletion
```

Because various models of aa replacements are applied, it takes some time.

Be patient..

Comparing the models of amino acid replacements

Once finished, you will see the list of models ranked with BIC.

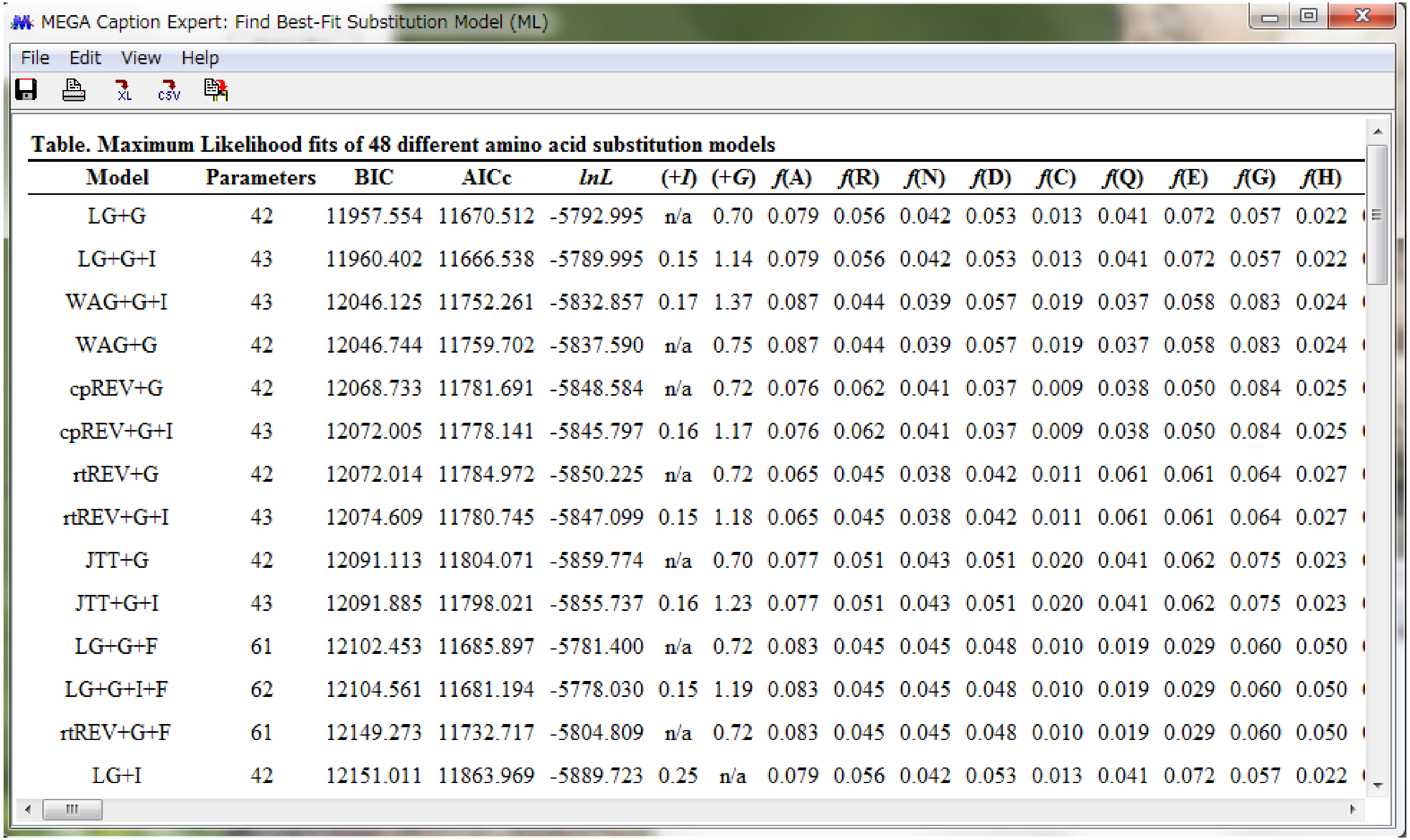
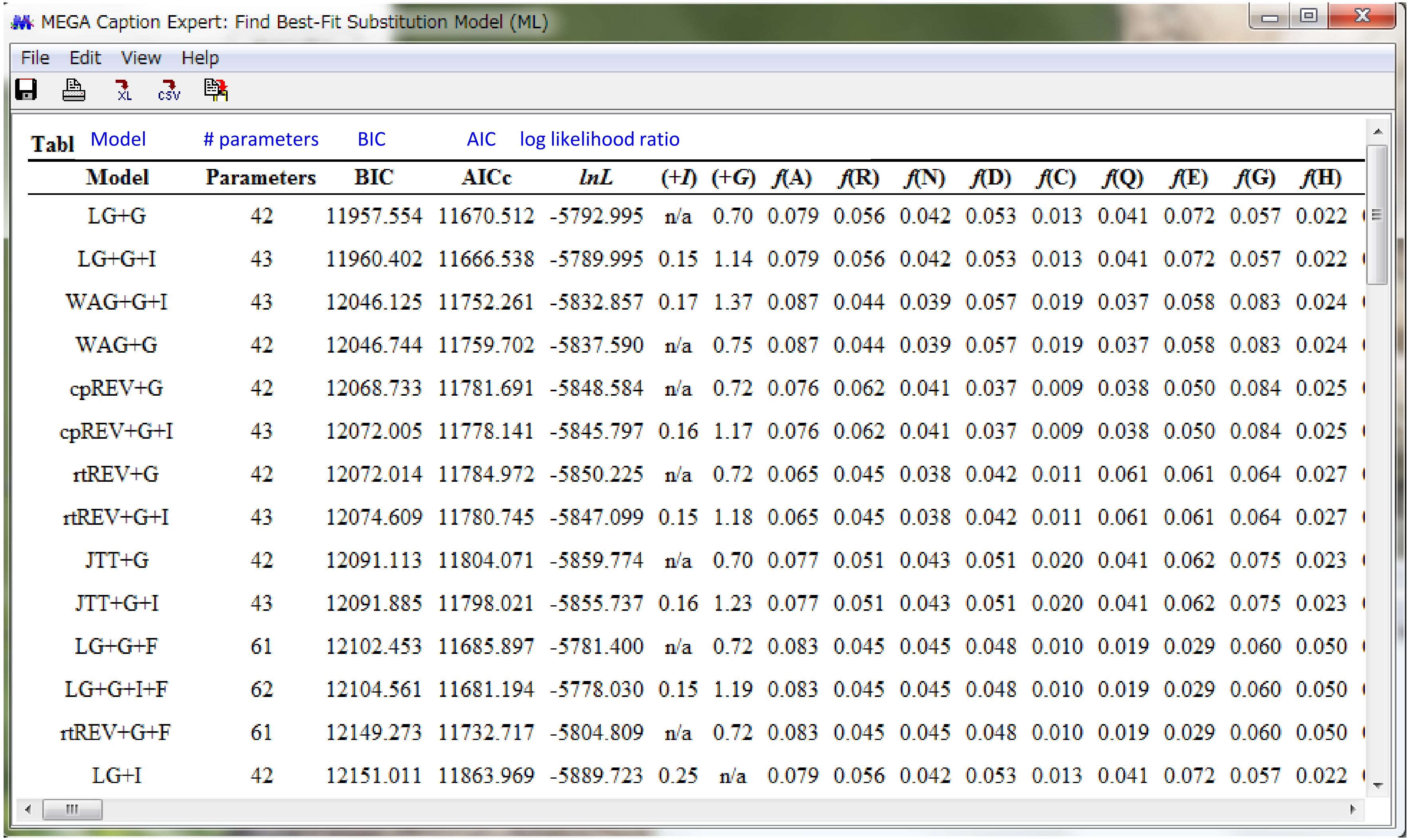


Table. Maximum Likelihood fits of 48 different amino acid substitution models

Model	Parameters	BIC	AICc	*lnL*	(+*I*)	(+*G*)	*f*(A)	*f*(R)	*f*(N)	*f*(D)	*f*(C)	*f*(Q)	*f*(E)	*f*(G)	*f*(H)
LG+G	42	11957.554	11670.512	-5792.995	n/a	0.70	0.079	0.056	0.042	0.053	0.013	0.041	0.072	0.057	0.022
LG+G+I	43	11960.402	11666.538	-5789.995	0.15	1.14	0.079	0.056	0.042	0.053	0.013	0.041	0.072	0.057	0.022
WAG+G+I	43	12046.125	11752.261	-5832.857	0.17	1.37	0.087	0.044	0.039	0.057	0.019	0.037	0.058	0.083	0.024
WAG+G	42	12046.744	11759.702	-5837.590	n/a	0.75	0.087	0.044	0.039	0.057	0.019	0.037	0.058	0.083	0.024
cpREV+G	42	12068.733	11781.691	-5848.584	n/a	0.72	0.076	0.062	0.041	0.037	0.009	0.038	0.050	0.084	0.025
cpREV+G+I	43	12072.005	11778.141	-5845.797	0.16	1.17	0.076	0.062	0.041	0.037	0.009	0.038	0.050	0.084	0.025
rtREV+G	42	12072.014	11784.972	-5850.225	n/a	0.72	0.065	0.045	0.038	0.042	0.011	0.061	0.061	0.064	0.027
rtREV+G+I	43	12074.609	11780.745	-5847.099	0.15	1.18	0.065	0.045	0.038	0.042	0.011	0.061	0.061	0.064	0.027
JTT+G	42	12091.113	11804.071	-5859.774	n/a	0.70	0.077	0.051	0.043	0.051	0.020	0.041	0.062	0.075	0.023
JTT+G+I	43	12091.885	11798.021	-5855.737	0.16	1.23	0.077	0.051	0.043	0.051	0.020	0.041	0.062	0.075	0.023
LG+G+F	61	12102.453	11685.897	-5781.400	n/a	0.72	0.083	0.045	0.045	0.048	0.010	0.019	0.029	0.060	0.050
LG+G+I+F	62	12104.561	11681.194	-5778.030	0.15	1.19	0.083	0.045	0.045	0.048	0.010	0.019	0.029	0.060	0.050
rtREV+G+F	61	12149.273	11732.717	-5804.809	n/a	0.72	0.083	0.045	0.045	0.048	0.010	0.019	0.029	0.060	0.050
LG+I	42	12151.011	11863.969	-5889.723	0.25	n/a	0.079	0.056	0.042	0.053	0.013	0.041	0.072	0.057	0.022

Comparing the models of amino acid replacements

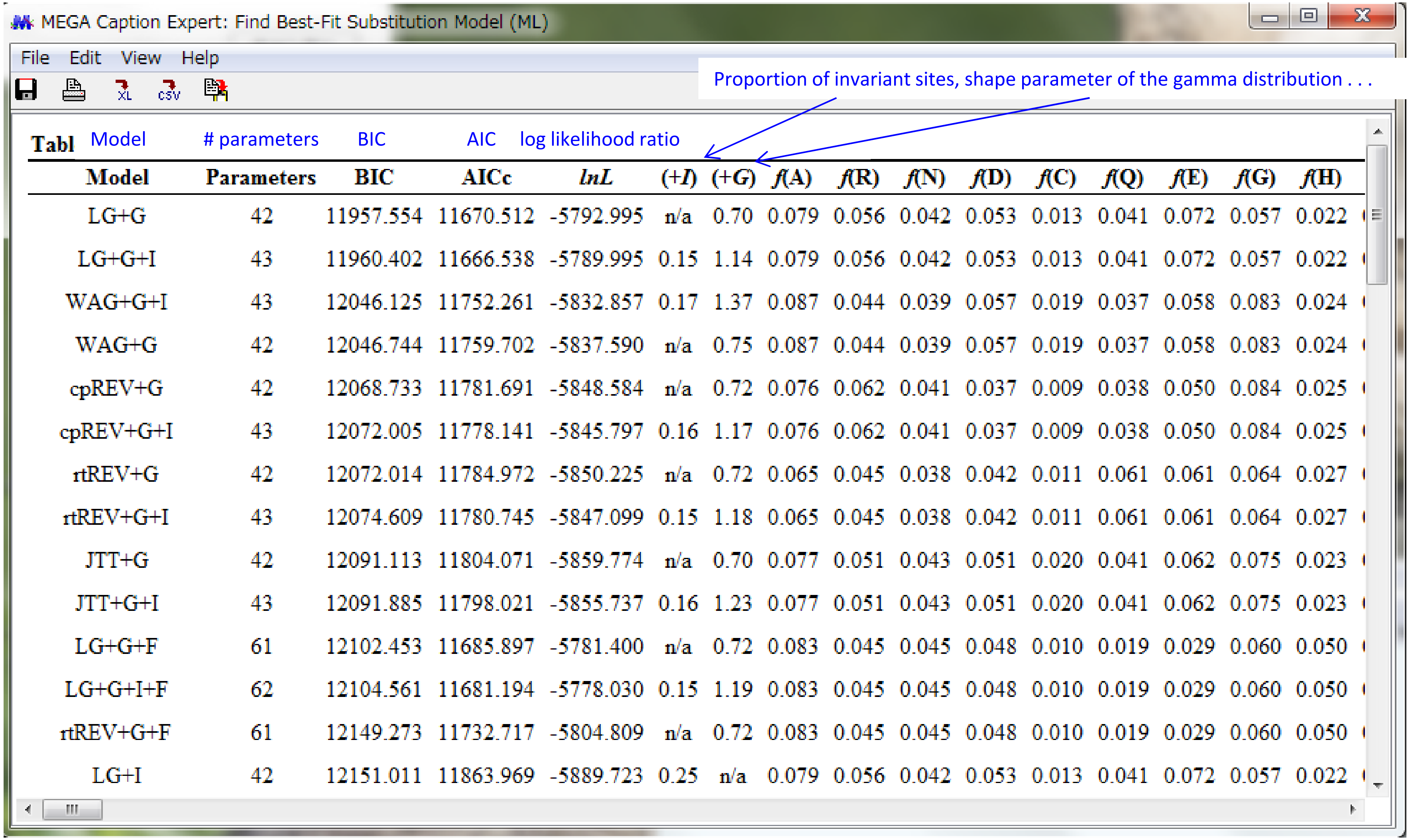
Once finished, you will see the list of models ranked with BIC.



Model	Parameters	BIC	AICc	*lnL*	(+*I*)	(+*G*)	*f*(A)	*f*(R)	*f*(N)	*f*(D)	*f*(C)	*f*(Q)	*f*(E)	*f*(G)	*f*(H)
LG+G	42	11957.554	11670.512	-5792.995	n/a	0.70	0.079	0.056	0.042	0.053	0.013	0.041	0.072	0.057	0.022
LG+G+I	43	11960.402	11666.538	-5789.995	0.15	1.14	0.079	0.056	0.042	0.053	0.013	0.041	0.072	0.057	0.022
WAG+G+I	43	12046.125	11752.261	-5832.857	0.17	1.37	0.087	0.044	0.039	0.057	0.019	0.037	0.058	0.083	0.024
WAG+G	42	12046.744	11759.702	-5837.590	n/a	0.75	0.087	0.044	0.039	0.057	0.019	0.037	0.058	0.083	0.024
cpREV+G	42	12068.733	11781.691	-5848.584	n/a	0.72	0.076	0.062	0.041	0.037	0.009	0.038	0.050	0.084	0.025
cpREV+G+I	43	12072.005	11778.141	-5845.797	0.16	1.17	0.076	0.062	0.041	0.037	0.009	0.038	0.050	0.084	0.025
rtREV+G	42	12072.014	11784.972	-5850.225	n/a	0.72	0.065	0.045	0.038	0.042	0.011	0.061	0.061	0.064	0.027
rtREV+G+I	43	12074.609	11780.745	-5847.099	0.15	1.18	0.065	0.045	0.038	0.042	0.011	0.061	0.061	0.064	0.027
JTT+G	42	12091.113	11804.071	-5859.774	n/a	0.70	0.077	0.051	0.043	0.051	0.020	0.041	0.062	0.075	0.023
JTT+G+I	43	12091.885	11798.021	-5855.737	0.16	1.23	0.077	0.051	0.043	0.051	0.020	0.041	0.062	0.075	0.023
LG+G+F	61	12102.453	11685.897	-5781.400	n/a	0.72	0.083	0.045	0.045	0.048	0.010	0.019	0.029	0.060	0.050
LG+G+I+F	62	12104.561	11681.194	-5778.030	0.15	1.19	0.083	0.045	0.045	0.048	0.010	0.019	0.029	0.060	0.050
rtREV+G+F	61	12149.273	11732.717	-5804.809	n/a	0.72	0.083	0.045	0.045	0.048	0.010	0.019	0.029	0.060	0.050
LG+I	42	12151.011	11863.969	-5889.723	0.25	n/a	0.079	0.056	0.042	0.053	0.013	0.041	0.072	0.057	0.022

Comparing the models of amino acid replacements

Once finished, you will see the list of models ranked with BIC.



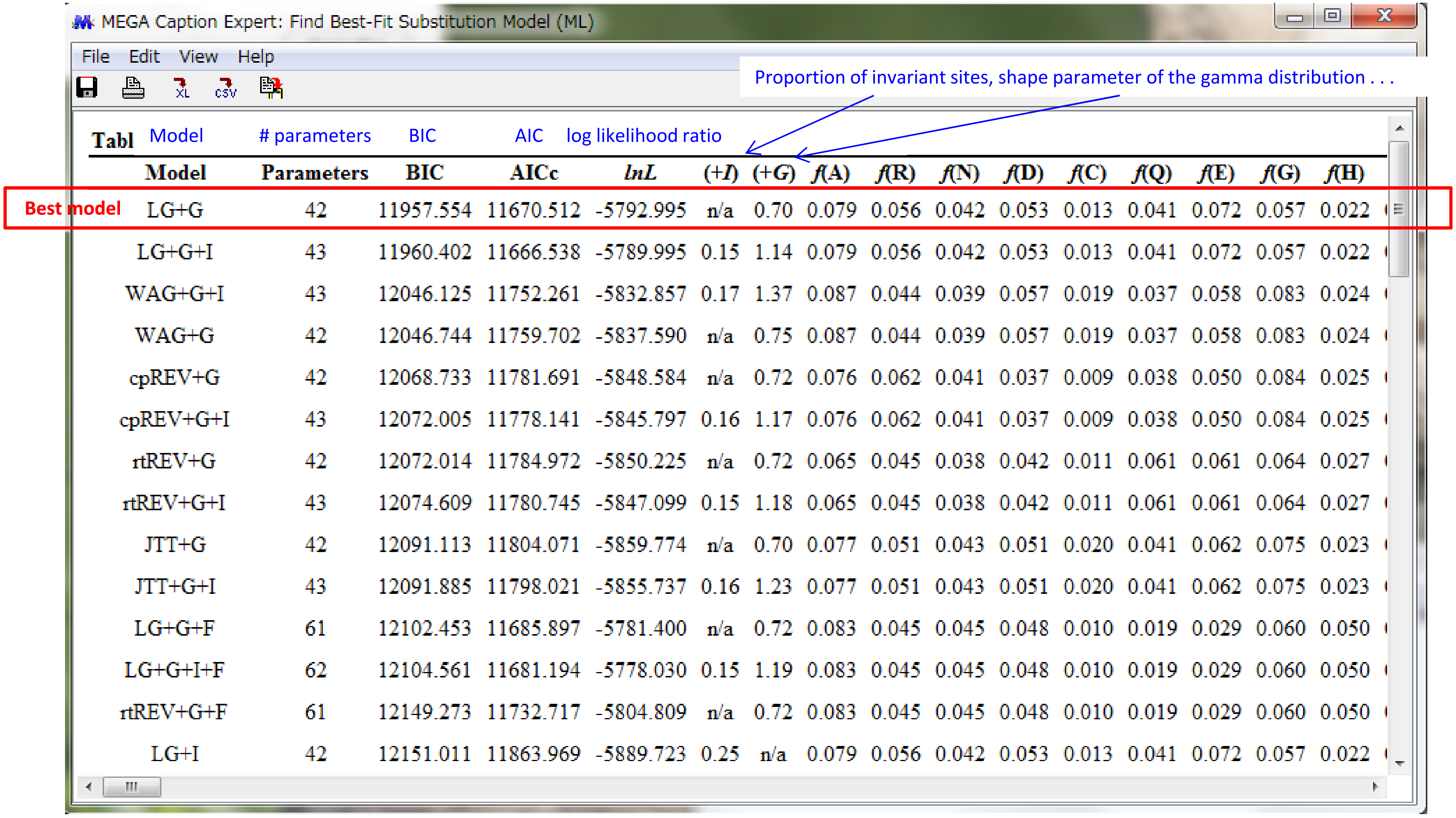
Proportion of invariant sites, shape parameter of the gamma distribution . . .

Model # parameters BIC AIC log likelihood ratio

Model	Parameters	BIC	AICc	*lnL*	(+*I*)	(+*G*)	*f*(A)	*f*(R)	*f*(N)	*f*(D)	*f*(C)	*f*(Q)	*f*(E)	*f*(G)	*f*(H)
LG+G	42	11957.554	11670.512	-5792.995	n/a	0.70	0.079	0.056	0.042	0.053	0.013	0.041	0.072	0.057	0.022
LG+G+I	43	11960.402	11666.538	-5789.995	0.15	1.14	0.079	0.056	0.042	0.053	0.013	0.041	0.072	0.057	0.022
WAG+G+I	43	12046.125	11752.261	-5832.857	0.17	1.37	0.087	0.044	0.039	0.057	0.019	0.037	0.058	0.083	0.024
WAG+G	42	12046.744	11759.702	-5837.590	n/a	0.75	0.087	0.044	0.039	0.057	0.019	0.037	0.058	0.083	0.024
cpREV+G	42	12068.733	11781.691	-5848.584	n/a	0.72	0.076	0.062	0.041	0.037	0.009	0.038	0.050	0.084	0.025
cpREV+G+I	43	12072.005	11778.141	-5845.797	0.16	1.17	0.076	0.062	0.041	0.037	0.009	0.038	0.050	0.084	0.025
rtREV+G	42	12072.014	11784.972	-5850.225	n/a	0.72	0.065	0.045	0.038	0.042	0.011	0.061	0.061	0.064	0.027
rtREV+G+I	43	12074.609	11780.745	-5847.099	0.15	1.18	0.065	0.045	0.038	0.042	0.011	0.061	0.061	0.064	0.027
JTT+G	42	12091.113	11804.071	-5859.774	n/a	0.70	0.077	0.051	0.043	0.051	0.020	0.041	0.062	0.075	0.023
JTT+G+I	43	12091.885	11798.021	-5855.737	0.16	1.23	0.077	0.051	0.043	0.051	0.020	0.041	0.062	0.075	0.023
LG+G+F	61	12102.453	11685.897	-5781.400	n/a	0.72	0.083	0.045	0.045	0.048	0.010	0.019	0.029	0.060	0.050
LG+G+I+F	62	12104.561	11681.194	-5778.030	0.15	1.19	0.083	0.045	0.045	0.048	0.010	0.019	0.029	0.060	0.050
rtREV+G+F	61	12149.273	11732.717	-5804.809	n/a	0.72	0.083	0.045	0.045	0.048	0.010	0.019	0.029	0.060	0.050
LG+I	42	12151.011	11863.969	-5889.723	0.25	n/a	0.079	0.056	0.042	0.053	0.013	0.041	0.072	0.057	0.022

Comparing the models of amino acid replacements

Once finished, you will see the list of models ranked with BIC.



Proportion of invariant sites, shape parameter of the gamma distribution . . .

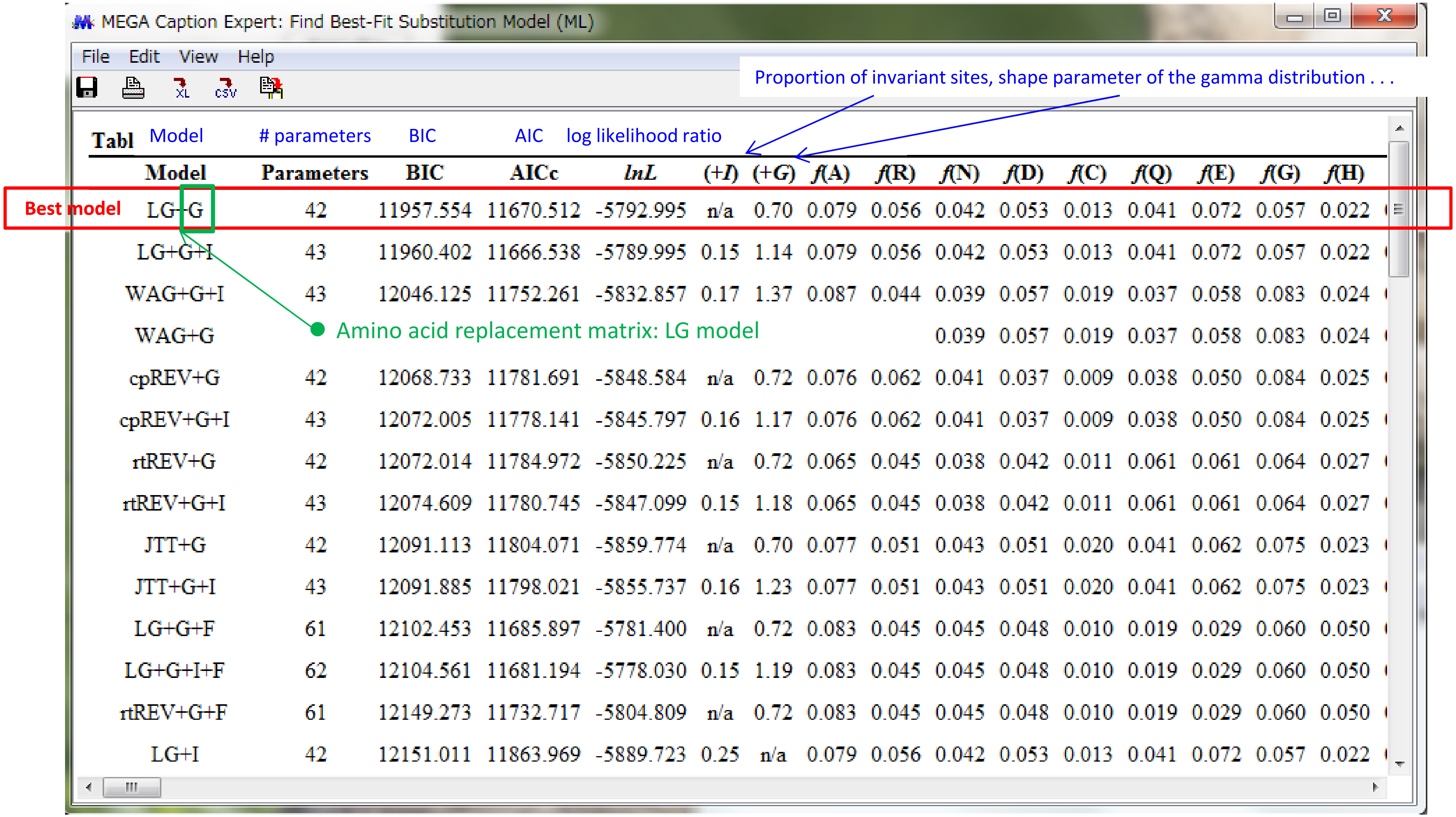
Model # parameters BIC AIC log likelihood ratio

Best model

Model	Parameters	BIC	AICc	lnL	(+I)	(+G)	f(A)	f(R)	f(N)	f(D)	f(C)	f(Q)	f(E)	f(G)	f(H)
LG+G	42	11957.554	11670.512	-5792.995	n/a	0.70	0.079	0.056	0.042	0.053	0.013	0.041	0.072	0.057	0.022
LG+G+I	43	11960.402	11666.538	-5789.995	0.15	1.14	0.079	0.056	0.042	0.053	0.013	0.041	0.072	0.057	0.022
WAG+G+I	43	12046.125	11752.261	-5832.857	0.17	1.37	0.087	0.044	0.039	0.057	0.019	0.037	0.058	0.083	0.024
WAG+G	42	12046.744	11759.702	-5837.590	n/a	0.75	0.087	0.044	0.039	0.057	0.019	0.037	0.058	0.083	0.024
cpREV+G	42	12068.733	11781.691	-5848.584	n/a	0.72	0.076	0.062	0.041	0.037	0.009	0.038	0.050	0.084	0.025
cpREV+G+I	43	12072.005	11778.141	-5845.797	0.16	1.17	0.076	0.062	0.041	0.037	0.009	0.038	0.050	0.084	0.025
rtREV+G	42	12072.014	11784.972	-5850.225	n/a	0.72	0.065	0.045	0.038	0.042	0.011	0.061	0.061	0.064	0.027
rtREV+G+I	43	12074.609	11780.745	-5847.099	0.15	1.18	0.065	0.045	0.038	0.042	0.011	0.061	0.061	0.064	0.027
JTT+G	42	12091.113	11804.071	-5859.774	n/a	0.70	0.077	0.051	0.043	0.051	0.020	0.041	0.062	0.075	0.023
JTT+G+I	43	12091.885	11798.021	-5855.737	0.16	1.23	0.077	0.051	0.043	0.051	0.020	0.041	0.062	0.075	0.023
LG+G+F	61	12102.453	11685.897	-5781.400	n/a	0.72	0.083	0.045	0.045	0.048	0.010	0.019	0.029	0.060	0.050
LG+G+I+F	62	12104.561	11681.194	-5778.030	0.15	1.19	0.083	0.045	0.045	0.048	0.010	0.019	0.029	0.060	0.050
rtREV+G+F	61	12149.273	11732.717	-5804.809	n/a	0.72	0.083	0.045	0.045	0.048	0.010	0.019	0.029	0.060	0.050
LG+I	42	12151.011	11863.969	-5889.723	0.25	n/a	0.079	0.056	0.042	0.053	0.013	0.041	0.072	0.057	0.022

Comparing the models of amino acid replacements

Once finished, you will see the list of models ranked with BIC.



Proportion of invariant sites, shape parameter of the gamma distribution . . .

Model # parameters BIC AIC log likelihood ratio

Best model

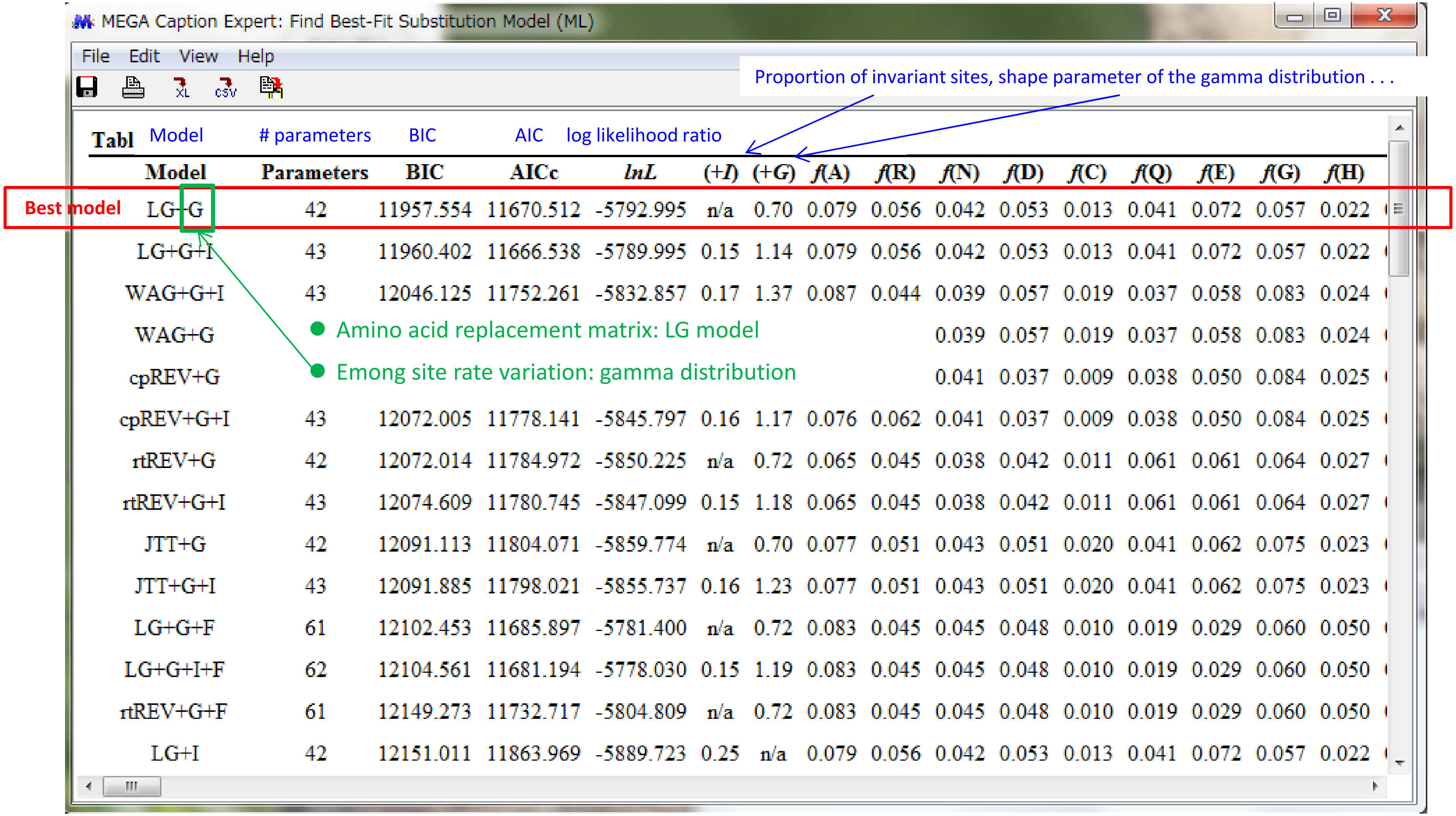
Amino acid replacement matrix: LG model

Tabl

Model	Parameters	BIC	AICc	*lnL*	(+*I*)	(+*G*)	*f*(A)	*f*(R)	*f*(N)	*f*(D)	*f*(C)	*f*(Q)	*f*(E)	*f*(G)	*f*(H)
LG+G	42	11957.554	11670.512	-5792.995	n/a	0.70	0.079	0.056	0.042	0.053	0.013	0.041	0.072	0.057	0.022
LG+G+I	43	11960.402	11666.538	-5789.995	0.15	1.14	0.079	0.056	0.042	0.053	0.013	0.041	0.072	0.057	0.022
WAG+G+I	43	12046.125	11752.261	-5832.857	0.17	1.37	0.087	0.044	0.039	0.057	0.019	0.037	0.058	0.083	0.024
WAG+G									0.039	0.057	0.019	0.037	0.058	0.083	0.024
cpREV+G	42	12068.733	11781.691	-5848.584	n/a	0.72	0.076	0.062	0.041	0.037	0.009	0.038	0.050	0.084	0.025
cpREV+G+I	43	12072.005	11778.141	-5845.797	0.16	1.17	0.076	0.062	0.041	0.037	0.009	0.038	0.050	0.084	0.025
rtREV+G	42	12072.014	11784.972	-5850.225	n/a	0.72	0.065	0.045	0.038	0.042	0.011	0.061	0.061	0.064	0.027
rtREV+G+I	43	12074.609	11780.745	-5847.099	0.15	1.18	0.065	0.045	0.038	0.042	0.011	0.061	0.061	0.064	0.027
JTT+G	42	12091.113	11804.071	-5859.774	n/a	0.70	0.077	0.051	0.043	0.051	0.020	0.041	0.062	0.075	0.023
JTT+G+I	43	12091.885	11798.021	-5855.737	0.16	1.23	0.077	0.051	0.043	0.051	0.020	0.041	0.062	0.075	0.023
LG+G+F	61	12102.453	11685.897	-5781.400	n/a	0.72	0.083	0.045	0.045	0.048	0.010	0.019	0.029	0.060	0.050
LG+G+I+F	62	12104.561	11681.194	-5778.030	0.15	1.19	0.083	0.045	0.045	0.048	0.010	0.019	0.029	0.060	0.050
rtREV+G+F	61	12149.273	11732.717	-5804.809	n/a	0.72	0.083	0.045	0.045	0.048	0.010	0.019	0.029	0.060	0.050
LG+I	42	12151.011	11863.969	-5889.723	0.25	n/a	0.079	0.056	0.042	0.053	0.013	0.041	0.072	0.057	0.022

Comparing the models of amino acid replacements

Once finished, you will see the list of models ranked with BIC.



Proportion of invariant sites, shape parameter of the gamma distribution . . .

Model # parameters BIC AIC log likelihood ratio

Best model

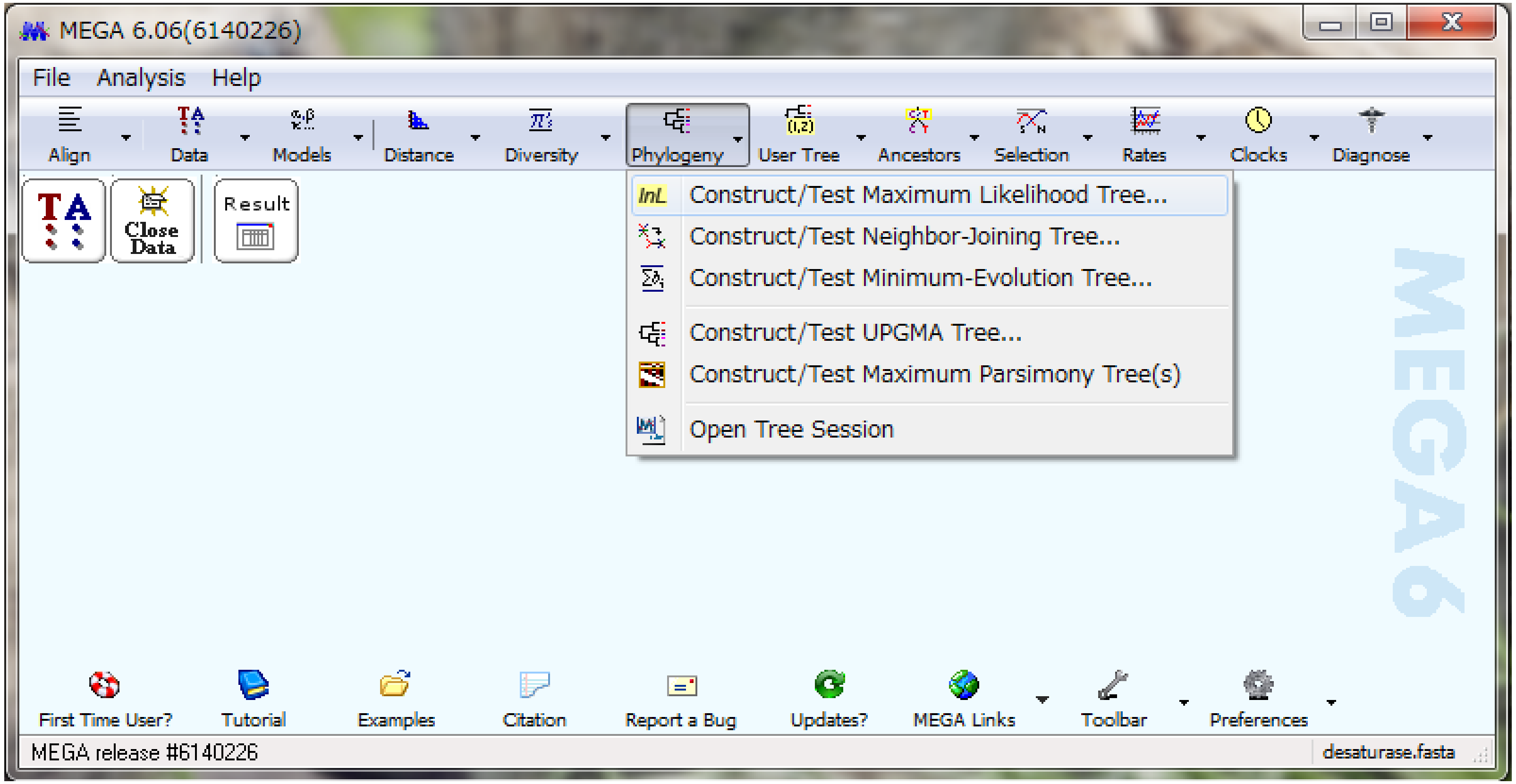
Model	Parameters	BIC	AICc	*lnL*	(+*I*)	(+*G*)	*f*(A)	*f*(R)	*f*(N)	*f*(D)	*f*(C)	*f*(Q)	*f*(E)	*f*(G)	*f*(H)
LG+G	42	11957.554	11670.512	-5792.995	n/a	0.70	0.079	0.056	0.042	0.053	0.013	0.041	0.072	0.057	0.022
LG+G+I	43	11960.402	11666.538	-5789.995	0.15	1.14	0.079	0.056	0.042	0.053	0.013	0.041	0.072	0.057	0.022
WAG+G+I	43	12046.125	11752.261	-5832.857	0.17	1.37	0.087	0.044	0.039	0.057	0.019	0.037	0.058	0.083	0.024
WAG+G									0.039	0.057	0.019	0.037	0.058	0.083	0.024
cpREV+G									0.041	0.037	0.009	0.038	0.050	0.084	0.025
cpREV+G+I	43	12072.005	11778.141	-5845.797	0.16	1.17	0.076	0.062	0.041	0.037	0.009	0.038	0.050	0.084	0.025
rtREV+G	42	12072.014	11784.972	-5850.225	n/a	0.72	0.065	0.045	0.038	0.042	0.011	0.061	0.061	0.064	0.027
rtREV+G+I	43	12074.609	11780.745	-5847.099	0.15	1.18	0.065	0.045	0.038	0.042	0.011	0.061	0.061	0.064	0.027
JTT+G	42	12091.113	11804.071	-5859.774	n/a	0.70	0.077	0.051	0.043	0.051	0.020	0.041	0.062	0.075	0.023
JTT+G+I	43	12091.885	11798.021	-5855.737	0.16	1.23	0.077	0.051	0.043	0.051	0.020	0.041	0.062	0.075	0.023
LG+G+F	61	12102.453	11685.897	-5781.400	n/a	0.72	0.083	0.045	0.045	0.048	0.010	0.019	0.029	0.060	0.050
LG+G+I+F	62	12104.561	11681.194	-5778.030	0.15	1.19	0.083	0.045	0.045	0.048	0.010	0.019	0.029	0.060	0.050
rtREV+G+F	61	12149.273	11732.717	-5804.809	n/a	0.72	0.083	0.045	0.045	0.048	0.010	0.019	0.029	0.060	0.050
LG+I	42	12151.011	11863.969	-5889.723	0.25	n/a	0.079	0.056	0.042	0.053	0.013	0.041	0.072	0.057	0.022

- Amino acid replacement matrix: LG model
- Emong site rate variation: gamma distribution

Maximum likelihood inference

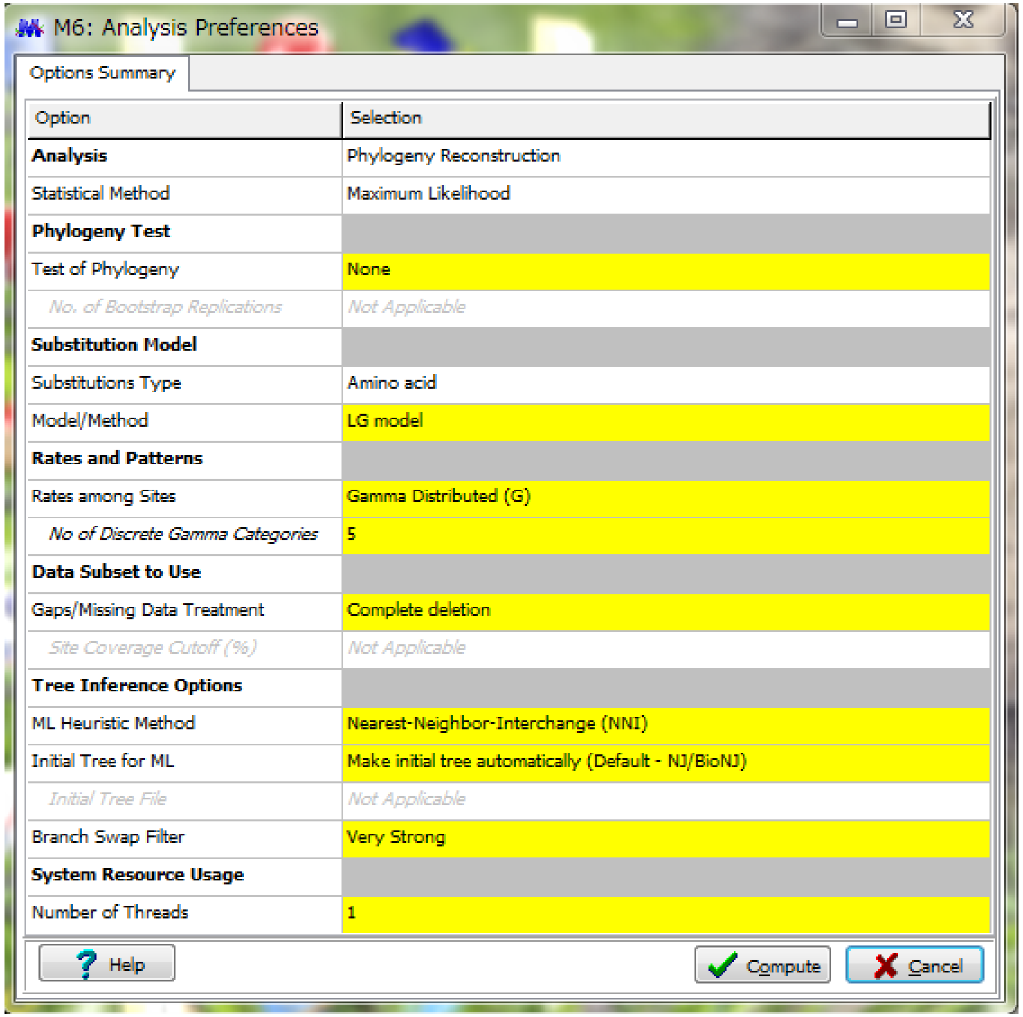
Constructing a phylogenetic tree by the best model

[Phylogeny][Construct/Test Maximum Likelihood Tree]



Constructing a phylogenetic tree by the best model

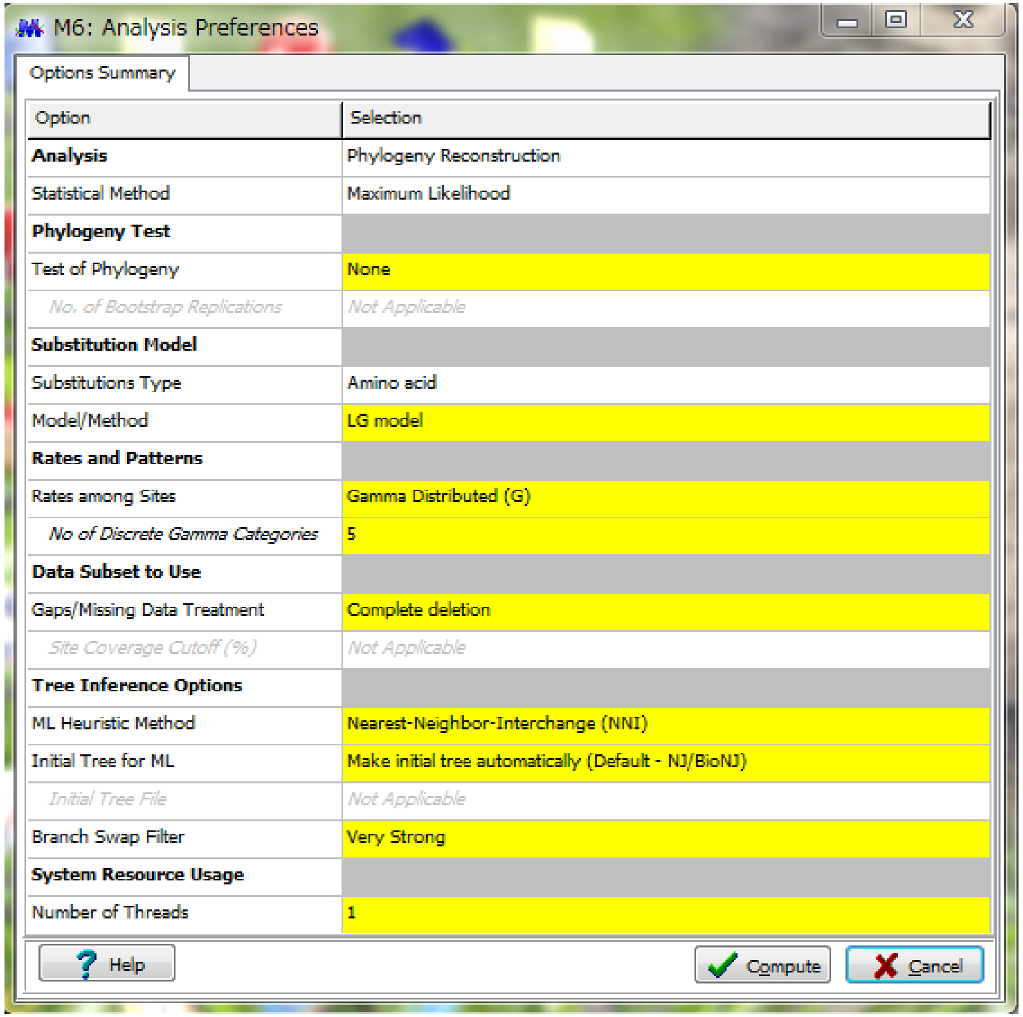
List of options



Option	Selection
Analysis	Phylogeny Reconstruction
Statistical Method	Maximum Likelihood
Phylogeny Test	
Test of Phylogeny	None
No. of Bootstrap Replications	*Not Applicable*
Substitution Model	
Substitutions Type	Amino acid
Model/Method	LG model
Rates and Patterns	
Rates among Sites	Gamma Distributed (G)
No of Discrete Gamma Categories	5
Data Subset to Use	
Gaps/Missing Data Treatment	Complete deletion
Site Coverage Cutoff (%)	*Not Applicable*
Tree Inference Options	
ML Heuristic Method	Nearest-Neighbor-Interchange (NNI)
Initial Tree for ML	Make initial tree automatically (Default - NJ/BioNJ)
Initial Tree File	*Not Applicable*
Branch Swap Filter	Very Strong
System Resource Usage	
Number of Threads	1

Constructing a phylogenetic tree by the best model

List of options



Option	Selection
Analysis	Phylogeny Reconstruction
Statistical Method	Maximum Likelihood
Phylogeny Test	
Test of Phylogeny	None
No. of Bootstrap Replications	*Not Applicable*
Substitution Model	
Substitutions Type	Amino acid
Model/Method	LG model
Rates and Patterns	
Rates among Sites	Gamma Distributed (G)
No of Discrete Gamma Categories	5
Data Subset to Use	
Gaps/Missing Data Treatment	Complete deletion
Site Coverage Cutoff (%)	*Not Applicable*
Tree Inference Options	
ML Heuristic Method	Nearest-Neighbor-Interchange (NNI)
Initial Tree for ML	Make initial tree automatically (Default - NJ/BioNJ)
Initial Tree File	*Not Applicable*
Branch Swap Filter	Very Strong
System Resource Usage	
Number of Threads	1

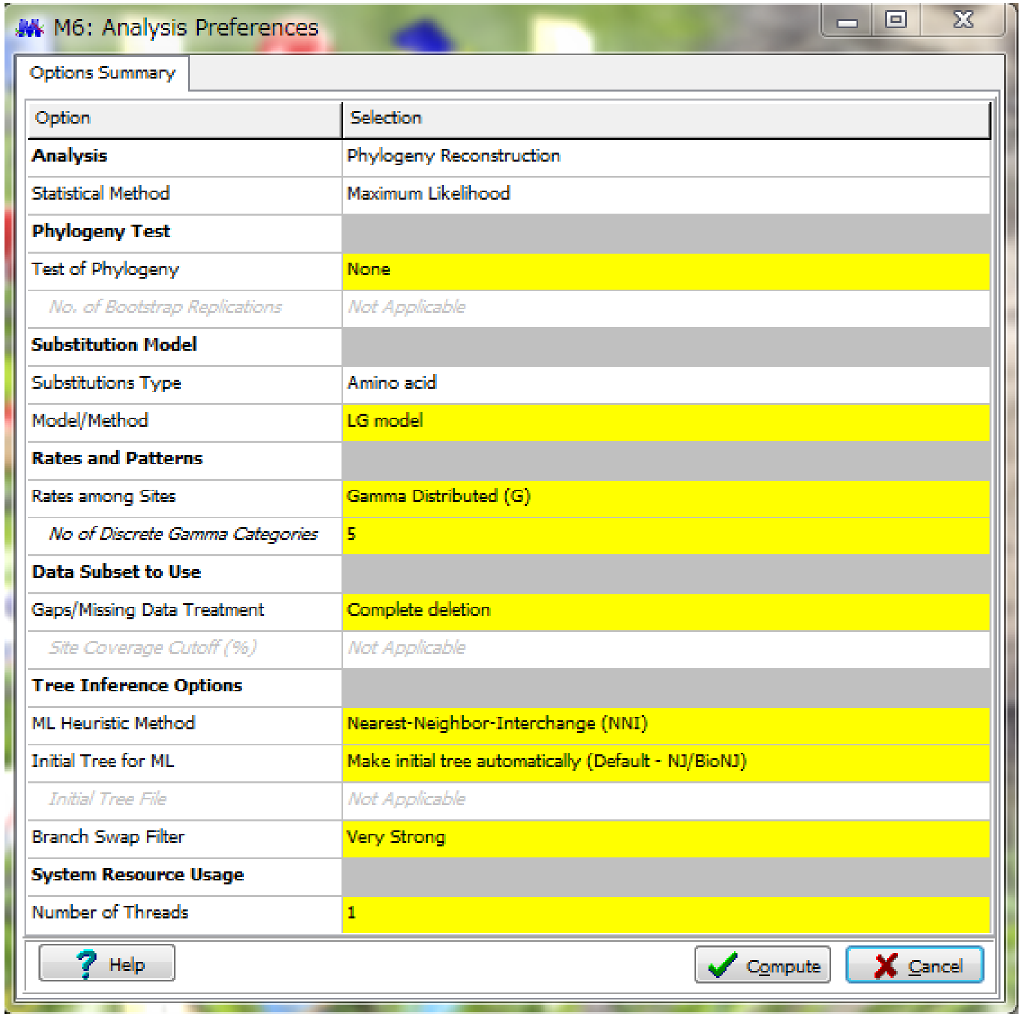
bootstrap (Y/N)

aa replacement matrix

Variable rate among sites

Constructing a phylogenetic tree by the best model

List of options



Option	Selection
Analysis	Phylogeny Reconstruction
Statistical Method	Maximum Likelihood
Phylogeny Test	
Test of Phylogeny	None
No. of Bootstrap Replications	*Not Applicable*
Substitution Model	
Substitutions Type	Amino acid
Model/Method	LG model
Rates and Patterns	
Rates among Sites	Gamma Distributed (G)
No of Discrete Gamma Categories	5
Data Subset to Use	
Gaps/Missing Data Treatment	Complete deletion
Site Coverage Cutoff (%)	*Not Applicable*
Tree Inference Options	
ML Heuristic Method	Nearest-Neighbor-Interchange (NNI)
Initial Tree for ML	Make initial tree automatically (Default - NJ/BioNJ)
Initial Tree File	*Not Applicable*
Branch Swap Filter	Very Strong
System Resource Usage	
Number of Threads	1

bootstrap (Y/N)

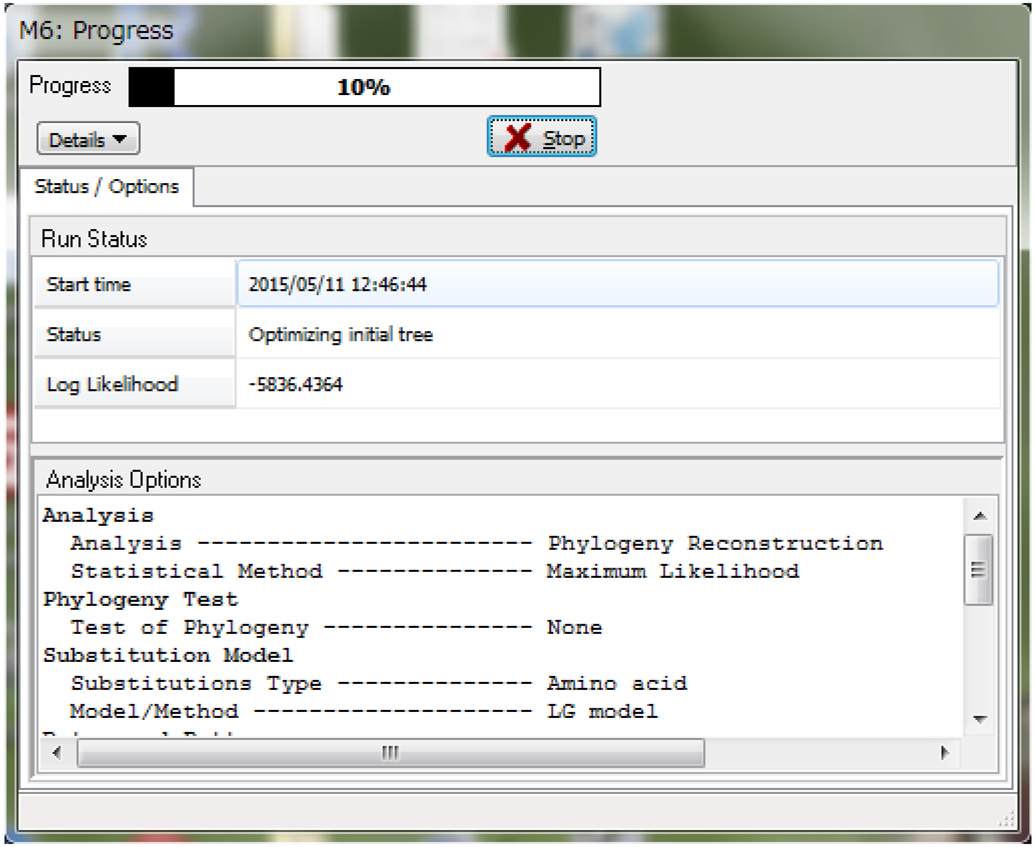
aa replacement matrix

Variable rate among sites

If ready, click Compute].

Constructing a phylogenetic tree by the best model

Computation starts …



Start time: 2015/05/11 12:46:44

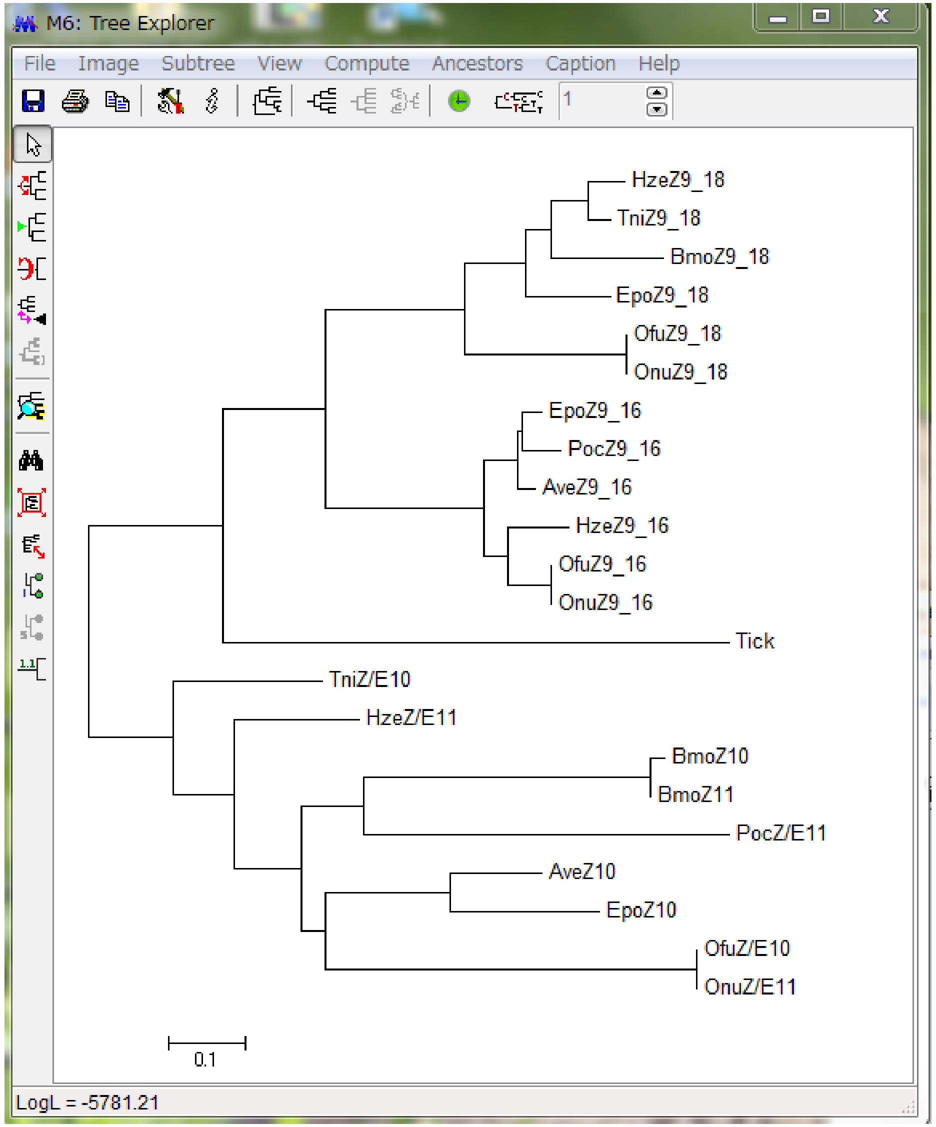
Status: Optimizing initial tree

Log Likelihood: -5836.4364

```
Analysis
  Analysis ------------------------ Phylogeny Reconstruction
  Statistical Method -------------- Maximum Likelihood
Phylogeny Test
  Test of Phylogeny --------------- None
Substitution Model
  Substitutions Type -------------- Amino acid
  Model/Method -------------------- LG model
```

Constructing a phylogenetic tree by the best model

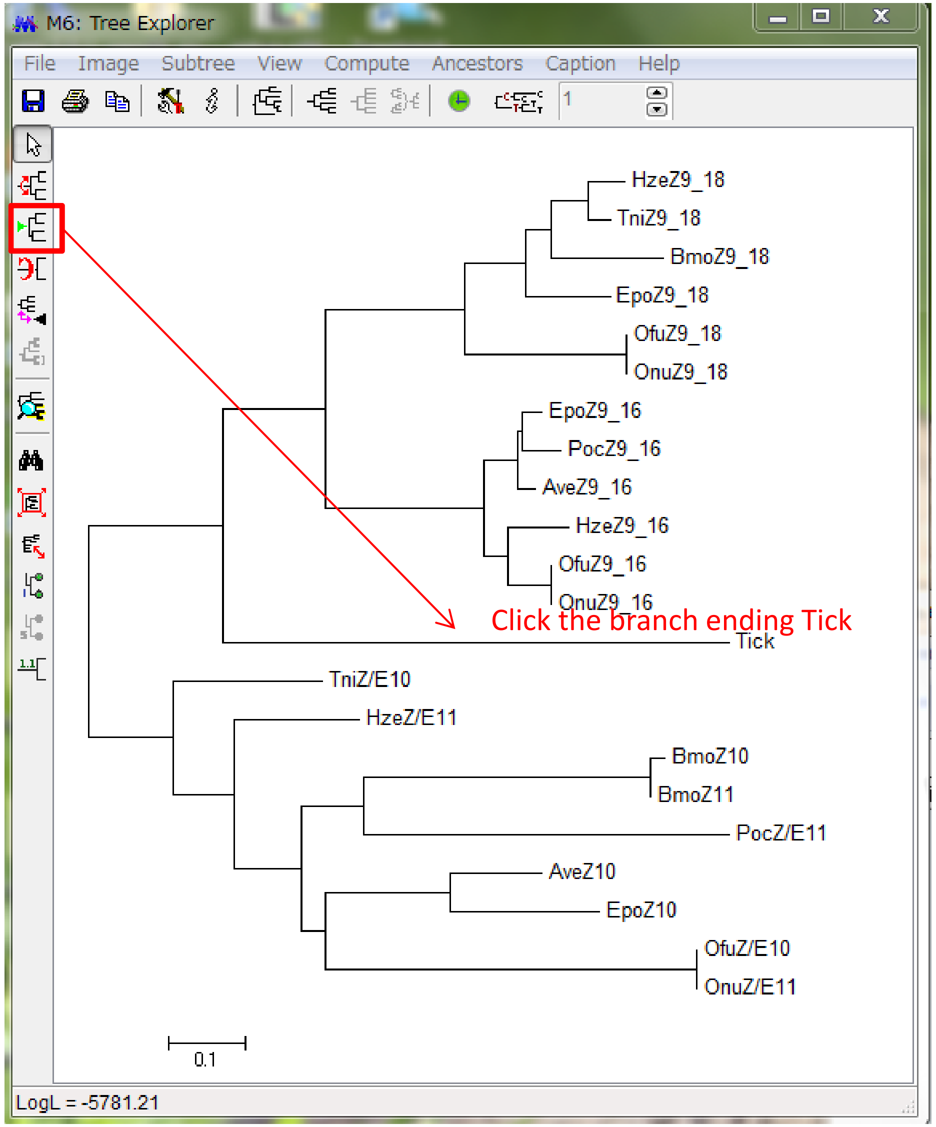
Once the phylogenetic tree is constructed, you will see the figure.



LogL = -5781.21

Constructing a phylogenetic tree by the best model

Click and specify Tick as an outgroup

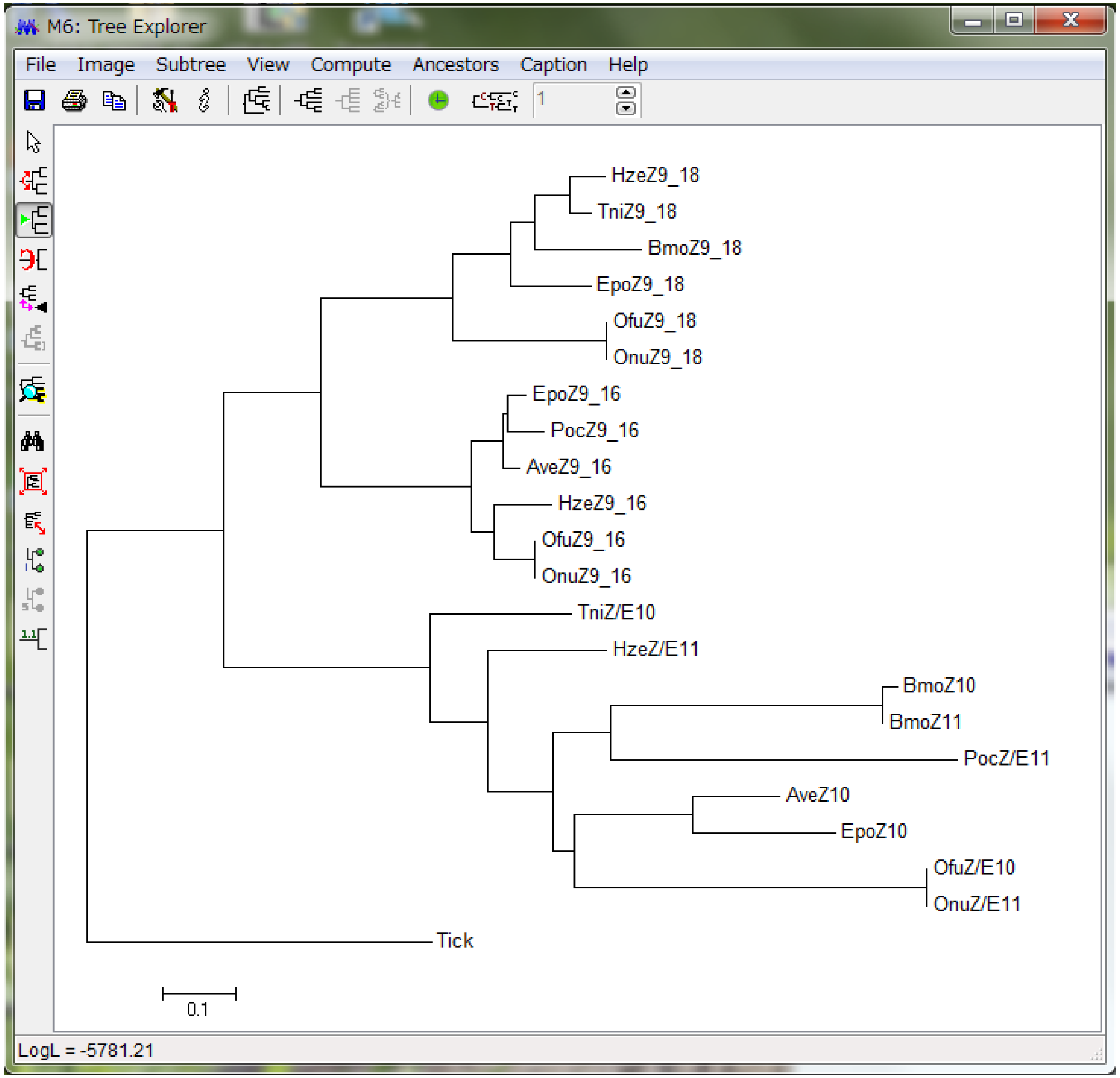


Click the branch ending Tick

LogL = -5781.21

Constructing a phylogenetic tree by the best model

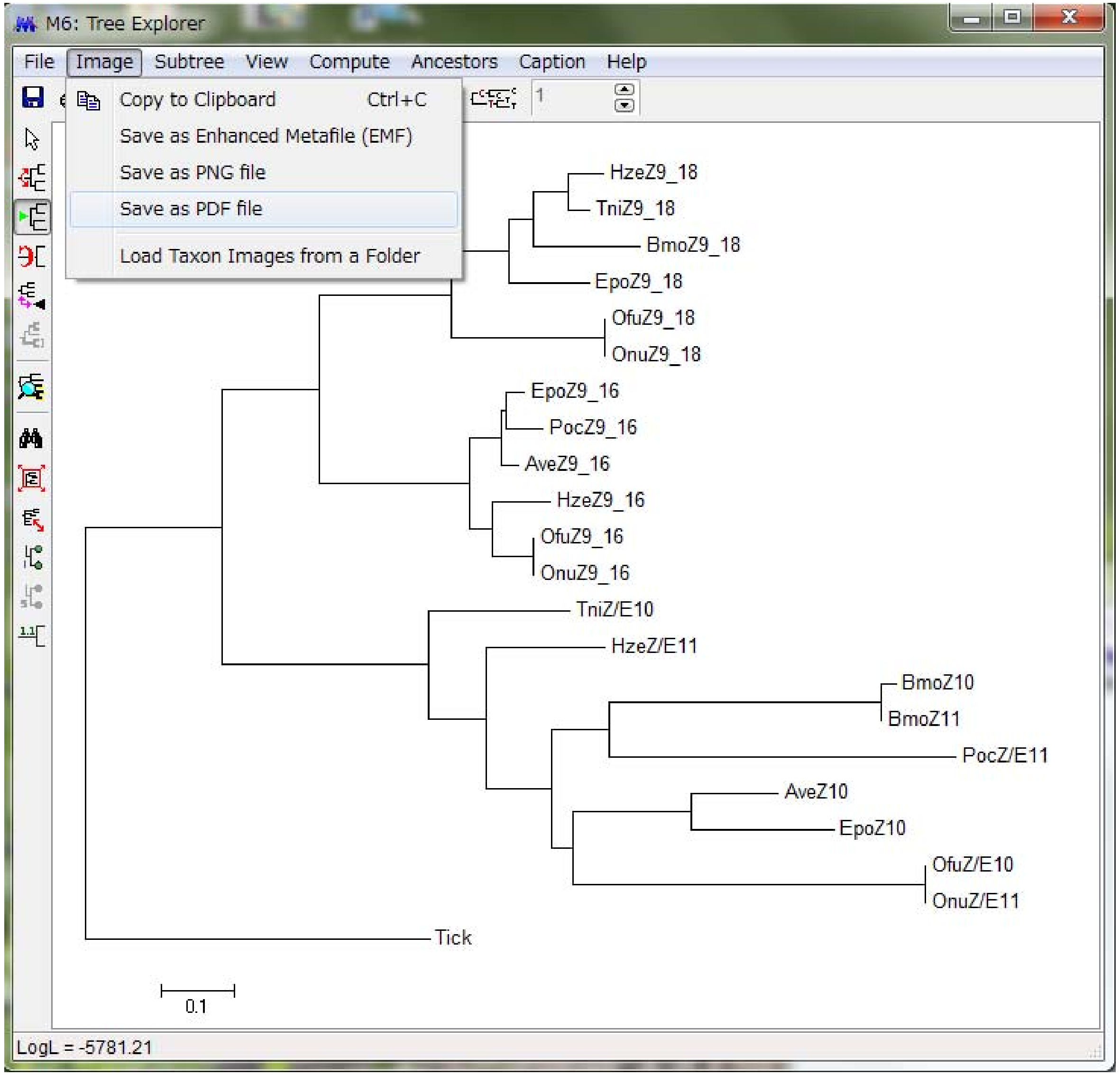
Rooted tree with Tick as an outgroup



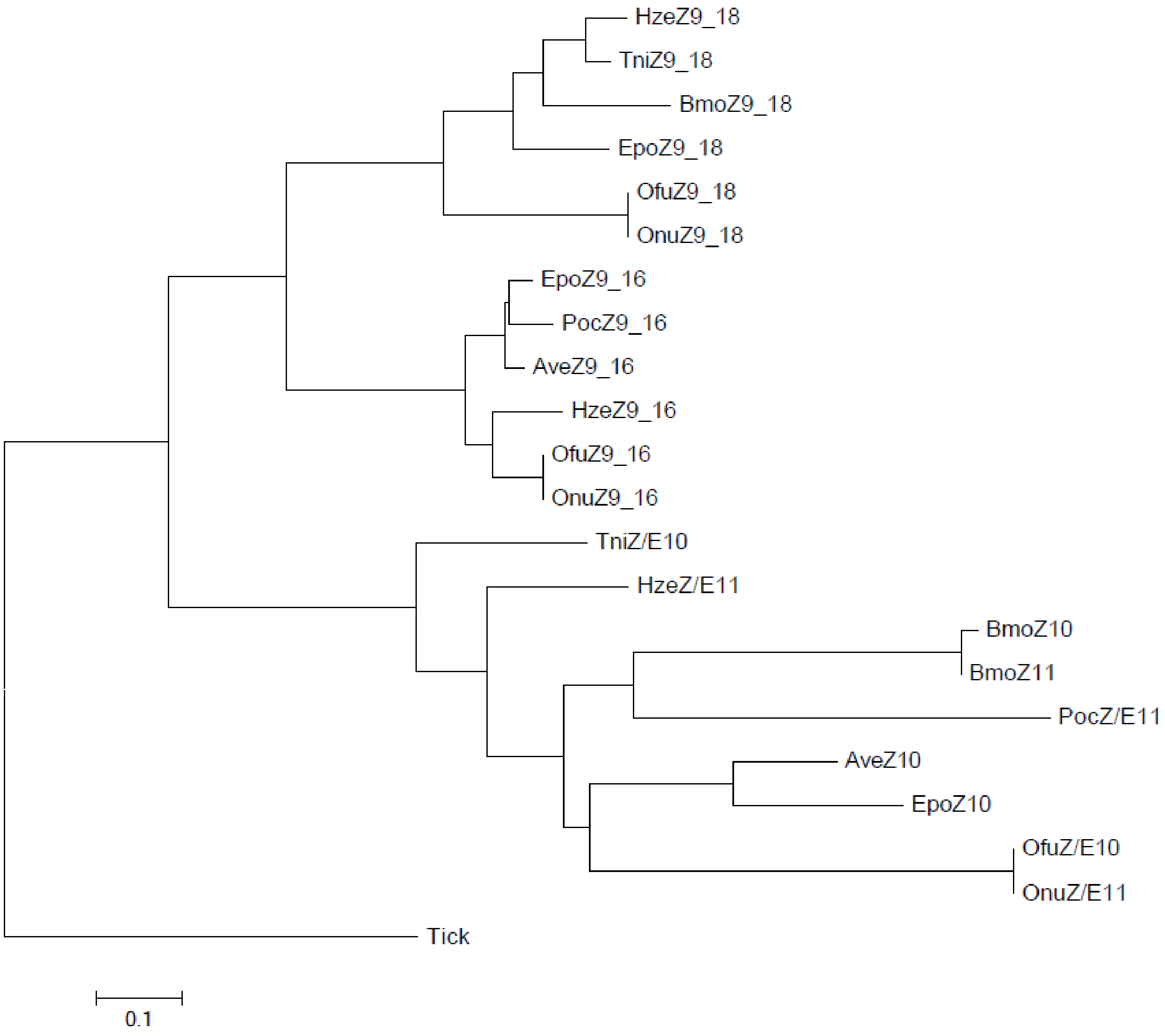
LogL = -5781.21

Save the figure as a file

[Image][Save as PDF file]   (choice of format: clipboard, emf, png, pdf …)

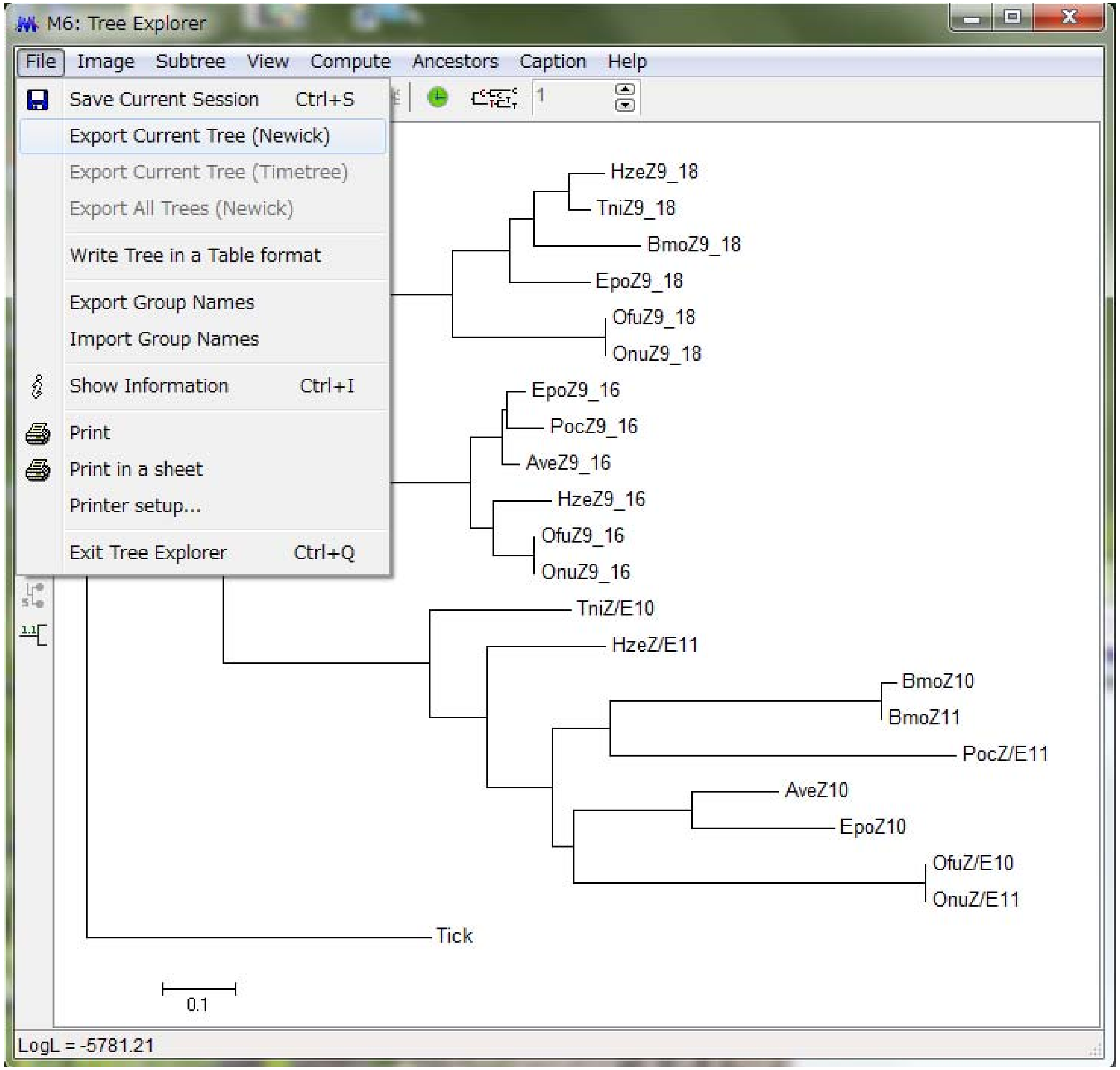


Saved image file



Save the tree information as a newick file

[File][Export Current Tree (Newick)]



Newick file

```
((((((('HzeZ9_18':0.04744891,'TniZ9_18':0.02873954):0.04876
410,'BmoZ9_18':0.14668082):0.03363353,'EpoZ9_18':0.10998
264):0.08067921,('OfuZ9_18':0.00000000,'OnuZ9_18':0.0000
0000):0.21168310):0.18028008,((('EpoZ9_16':0.02681811,'Po
cZ9_16':0.05093039):0.00494556,'AveZ9_16':0.02186046):0.0
4527333,('HzeZ9_16':0.08043957,('OfuZ9_16':0.00000000,'On
uZ9_16':0.00000000):0.05777886):0.03121448):0.20549621):
0.13563580,(TniZ/E10:0.19565117,(HzeZ/E11:0.16265867,(((B
moZ10:0.01952739,BmoZ11:0.00000000):0.37509960,PocZ/E
11:0.47698127):0.08048141,((AveZ10:0.11881414,EpoZ10:0.1
9493515):0.16379172,(OfuZ/E10:0.00000000,OnuZ/E11:0.000
00000):0.48603868):0.03036878):0.08837050):0.08074866):0
.28506824):0.18749402,Tick:0.47416490);
```

More thought on the rate heterogeneity among sites

Heterogeneity/homogeneity among sites

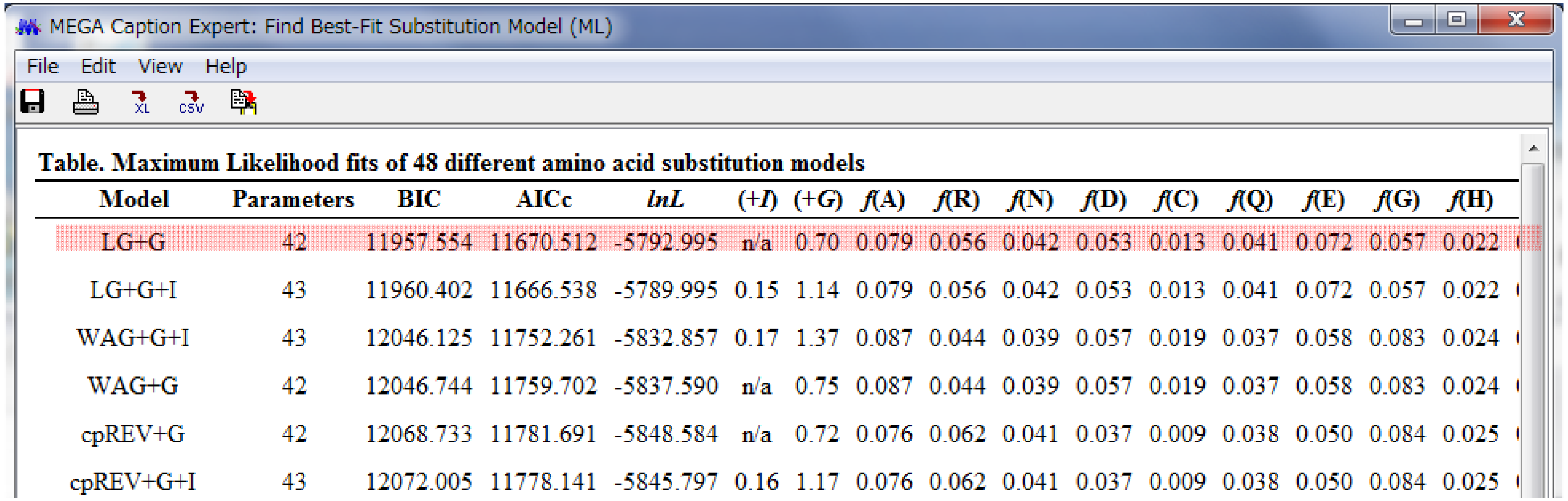
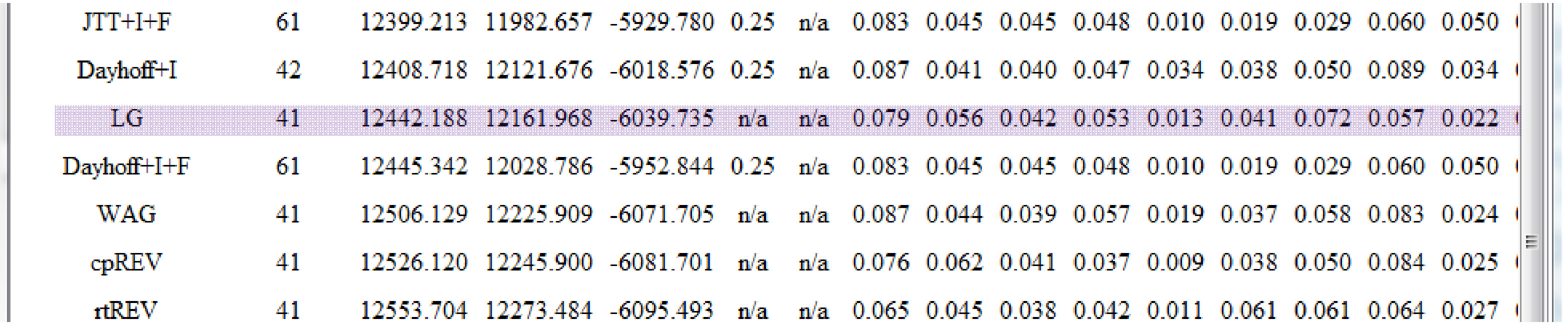


Table. Maximum Likelihood fits of 48 different amino acid substitution models

Model	Parameters	BIC	AICc	*lnL*	(+*I*)	(+*G*)	*f*(A)	*f*(R)	*f*(N)	*f*(D)	*f*(C)	*f*(Q)	*f*(E)	*f*(G)	*f*(H)
LG+G	42	11957.554	11670.512	-5792.995	n/a	0.70	0.079	0.056	0.042	0.053	0.013	0.041	0.072	0.057	0.022
LG+G+I	43	11960.402	11666.538	-5789.995	0.15	1.14	0.079	0.056	0.042	0.053	0.013	0.041	0.072	0.057	0.022
WAG+G+I	43	12046.125	11752.261	-5832.857	0.17	1.37	0.087	0.044	0.039	0.057	0.019	0.037	0.058	0.083	0.024
WAG+G	42	12046.744	11759.702	-5837.590	n/a	0.75	0.087	0.044	0.039	0.057	0.019	0.037	0.058	0.083	0.024
cpREV+G	42	12068.733	11781.691	-5848.584	n/a	0.72	0.076	0.062	0.041	0.037	0.009	0.038	0.050	0.084	0.025
cpREV+G+I	43	12072.005	11778.141	-5845.797	0.16	1.17	0.076	0.062	0.041	0.037	0.009	0.038	0.050	0.084	0.025

Best model

⋮



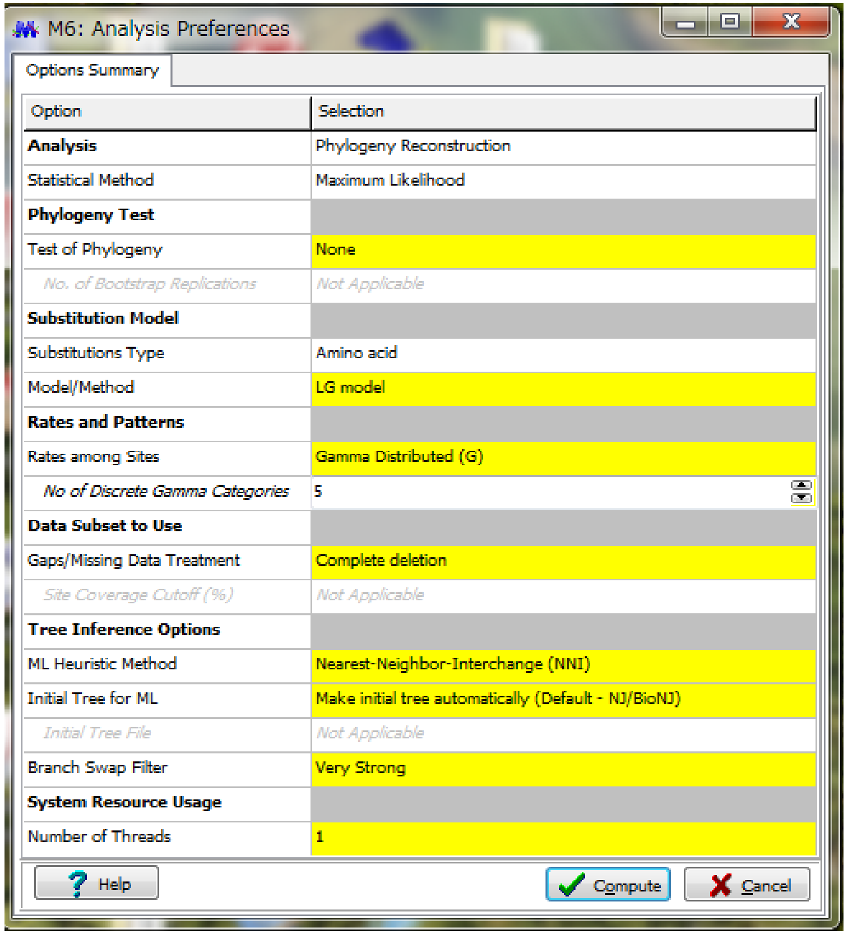
Model	Parameters	BIC	AICc	*lnL*	(+*I*)	(+*G*)	*f*(A)	*f*(R)	*f*(N)	*f*(D)	*f*(C)	*f*(Q)	*f*(E)	*f*(G)	*f*(H)
JTT+I+F	61	12399.213	11982.657	-5929.780	0.25	n/a	0.083	0.045	0.045	0.048	0.010	0.019	0.029	0.060	0.050
Dayhoff+I	42	12408.718	12121.676	-6018.576	0.25	n/a	0.087	0.041	0.040	0.047	0.034	0.038	0.050	0.089	0.034
LG	41	12442.188	12161.968	-6039.735	n/a	n/a	0.079	0.056	0.042	0.053	0.013	0.041	0.072	0.057	0.022
Dayhoff+I+F	61	12445.342	12028.786	-5952.844	0.25	n/a	0.083	0.045	0.045	0.048	0.010	0.019	0.029	0.060	0.050
WAG	41	12506.129	12225.909	-6071.705	n/a	n/a	0.087	0.044	0.039	0.057	0.019	0.037	0.058	0.083	0.024
cpREV	41	12526.120	12245.900	-6081.701	n/a	n/a	0.076	0.062	0.041	0.037	0.009	0.038	0.050	0.084	0.025
rtREV	41	12553.704	12273.484	-6095.493	n/a	n/a	0.065	0.045	0.038	0.042	0.011	0.061	0.061	0.064	0.027

Assuming uniform
rate among sites

⋮

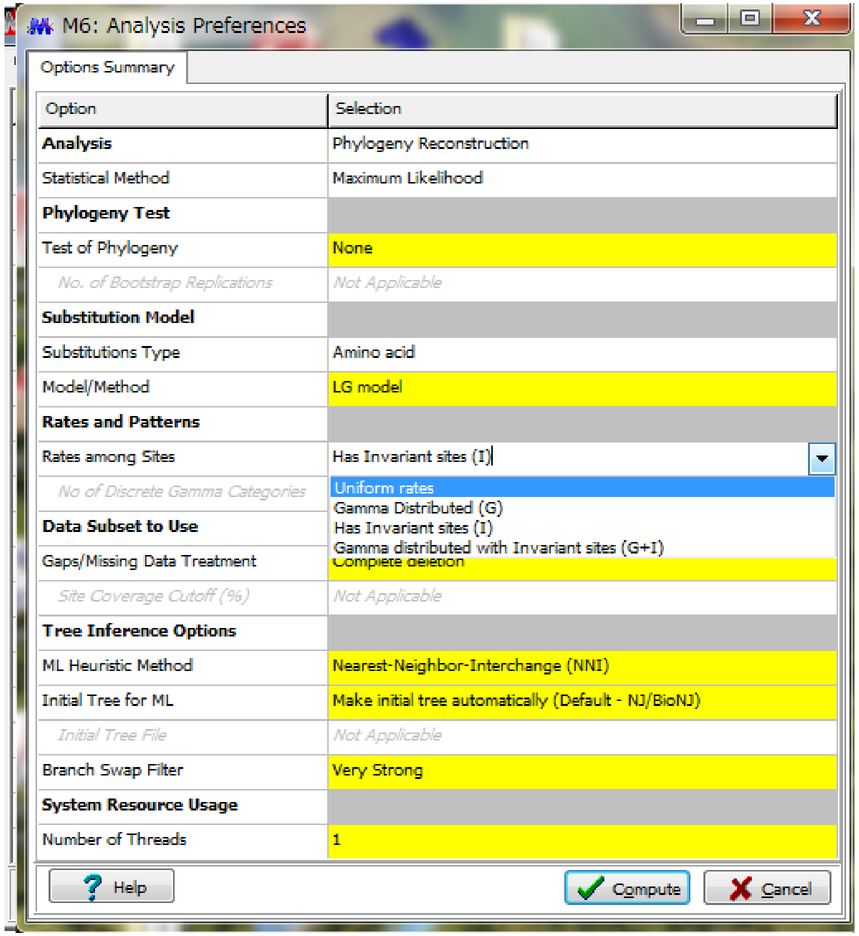
Effect of ignoring rate heterogeneity among sites

[Phylogeny][Construct/Test Maximum Likelihood Tree]
-> [Rates among Sites]: Choose “Uniform rates”



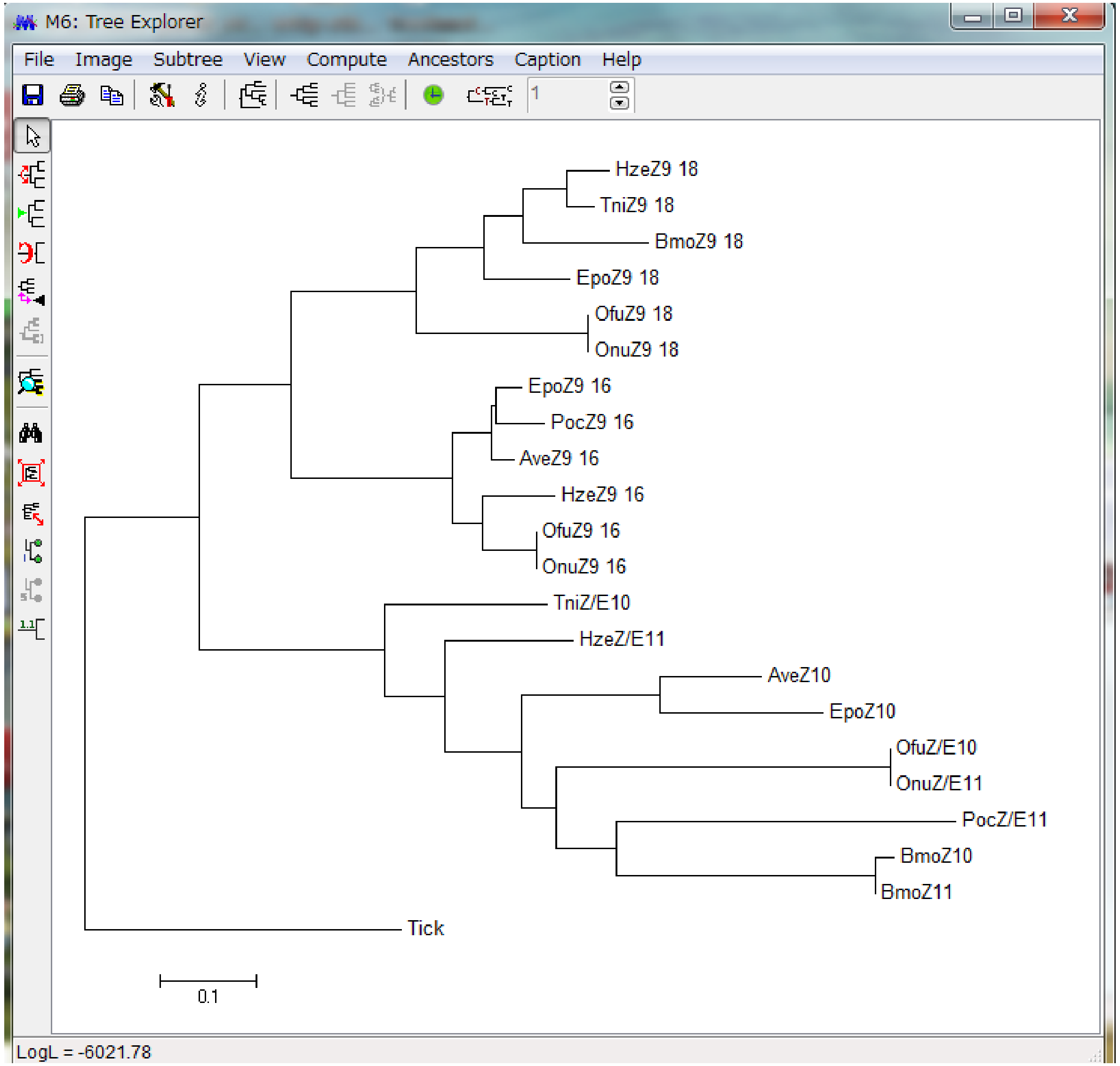
Option	Selection
Analysis	Phylogeny Reconstruction
Statistical Method	Maximum Likelihood
Phylogeny Test	
Test of Phylogeny	None
No. of Bootstrap Replications	*Not Applicable*
Substitution Model	
Substitutions Type	Amino acid
Model/Method	LG model
Rates and Patterns	
Rates among Sites	Gamma Distributed (G)
No of Discrete Gamma Categories	5
Data Subset to Use	
Gaps/Missing Data Treatment	Complete deletion
Site Coverage Cutoff (%)	*Not Applicable*
Tree Inference Options	
ML Heuristic Method	Nearest-Neighbor-Interchange (NNI)
Initial Tree for ML	Make initial tree automatically (Default - NJ/BioNJ)
Initial Tree File	*Not Applicable*
Branch Swap Filter	Very Strong
System Resource Usage	
Number of Threads	1

→



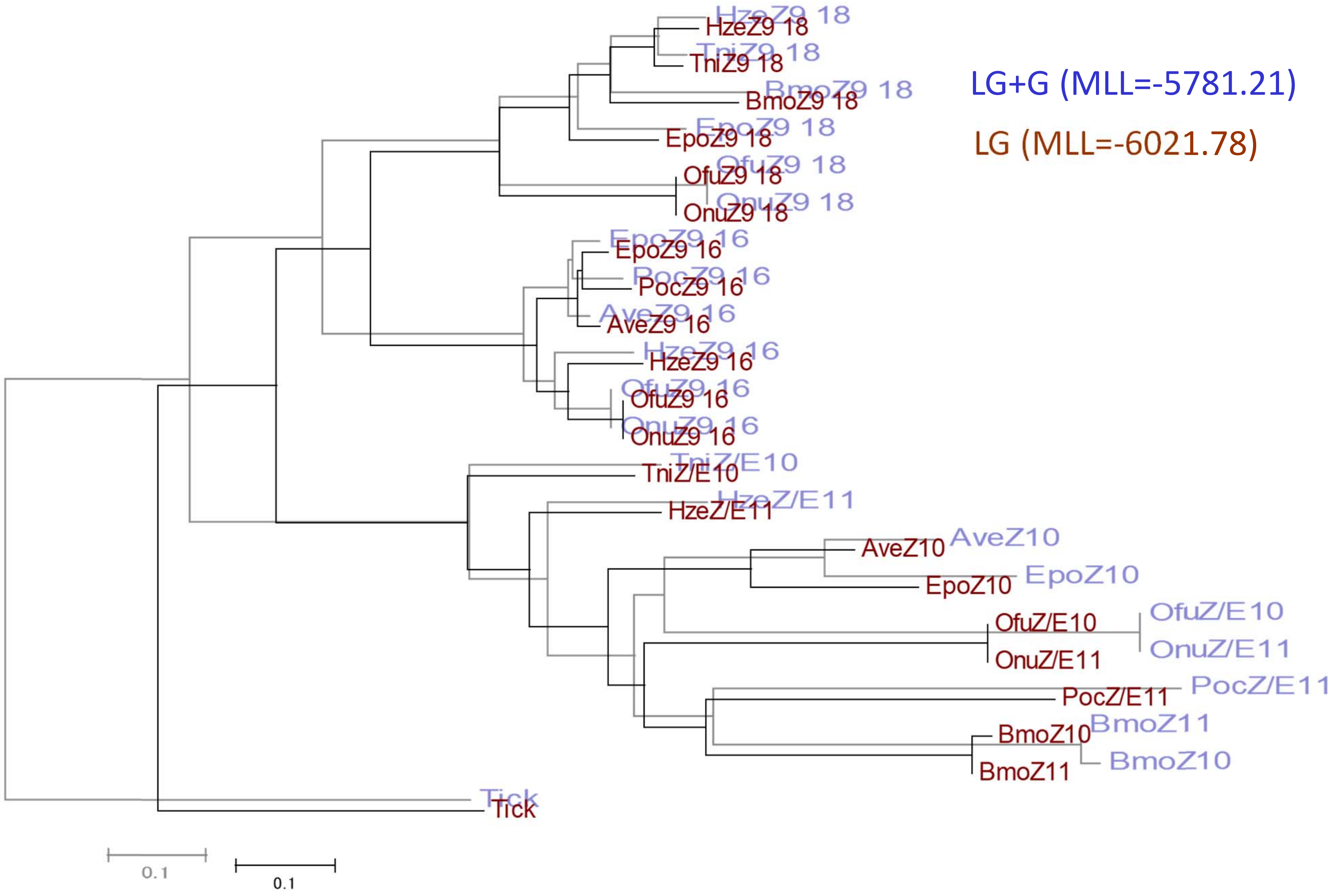
Option	Selection
Analysis	Phylogeny Reconstruction
Statistical Method	Maximum Likelihood
Phylogeny Test	
Test of Phylogeny	None
No. of Bootstrap Replications	*Not Applicable*
Substitution Model	
Substitutions Type	Amino acid
Model/Method	LG model
Rates and Patterns	
Rates among Sites	Has Invariant sites (I)
No of Discrete Gamma Categories	
Data Subset to Use	
Gaps/Missing Data Treatment	Complete deletion
Site Coverage Cutoff (%)	*Not Applicable*
Tree Inference Options	
ML Heuristic Method	Nearest-Neighbor-Interchange (NNI)
Initial Tree for ML	Make initial tree automatically (Default - NJ/BioNJ)
Initial Tree File	*Not Applicable*
Branch Swap Filter	Very Strong
System Resource Usage	
Number of Threads	1

Effect of ignoring rate heterogeneity among sites



LogL = -6021.78

Effect of ignoring rate heterogeneity among sites



LG+G (MLL=-5781.21)

LG (MLL=-6021.78)

[coffee break]

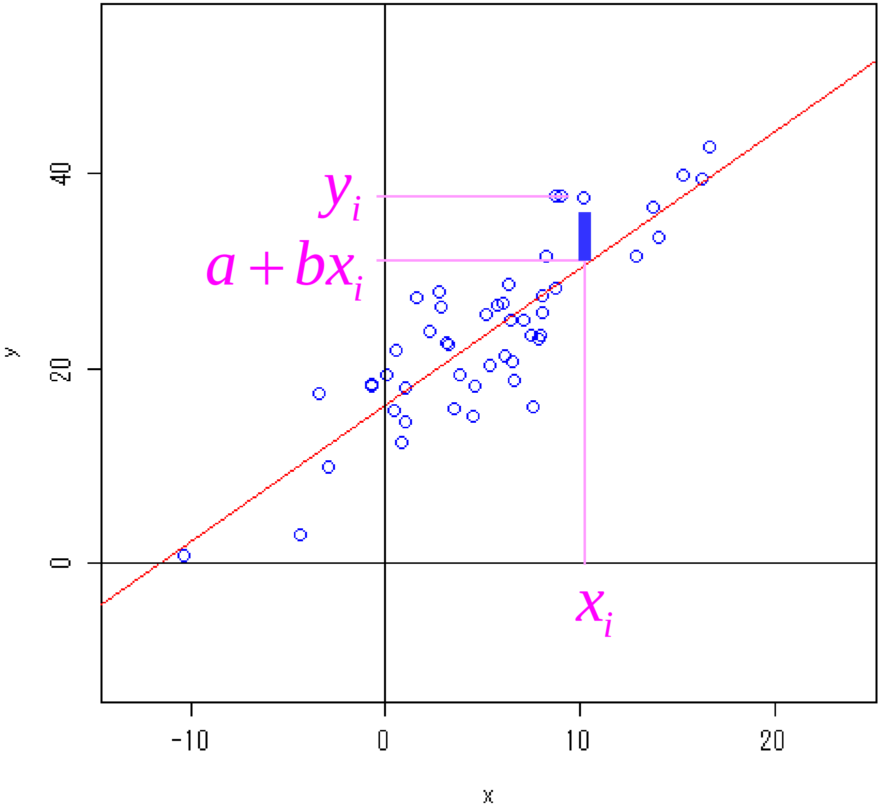
Maximum likelihood inference

Likelihood and maximum likelihood (ML) procedure

- Likelihood describes the probability of observing the data by a statistical model.
- Likelihood function: explicit representation of the likelihood in terms of parameters.
- Log likelihood: log of likelihood
- ML procedure: the method of estimating the parameters by maximizing the log likelihood value
  - MLE: the estimate by ML procedure
  - The variance of the MLE is obtained by the inverse of Fisher information quantity (minus the second derivative (Hessian) of the log likelihood function).

ML procedure for Regression analysis

$$y_i = a + bx_i + \varepsilon_i \quad i = 1,\dots,n$$



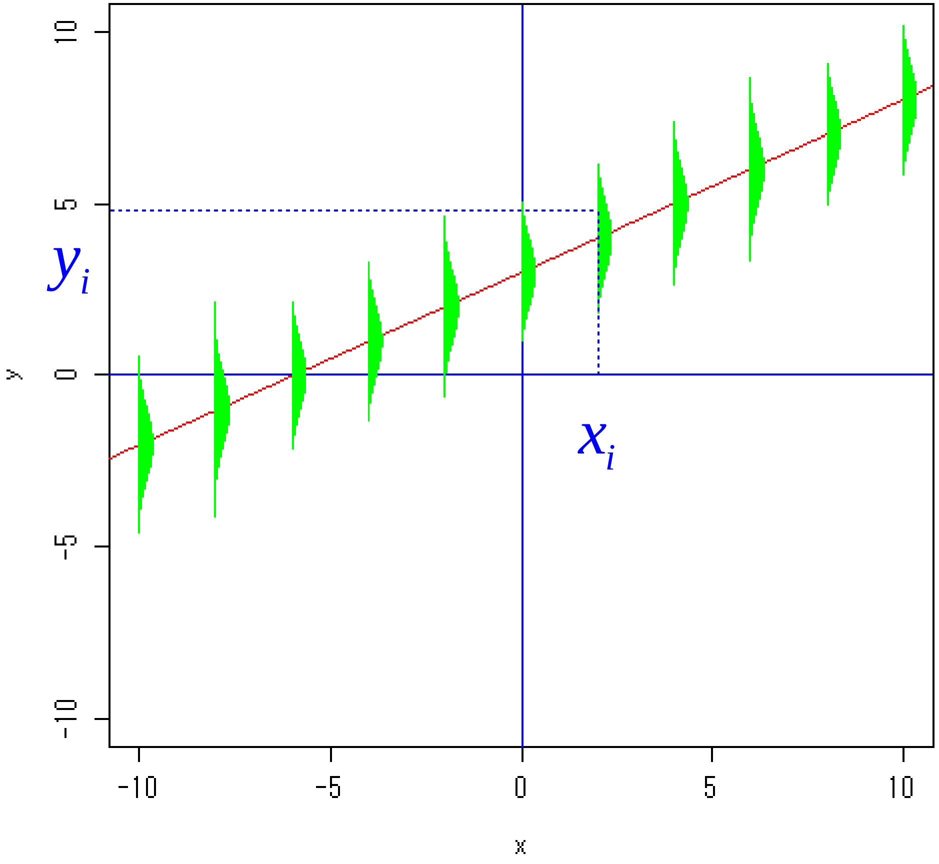
ML procedure for Regression analysis

Statistical model and the likelihood

$$L\left(a,b,s^2 \mid y_1, \cdots\ y_n\right)$$

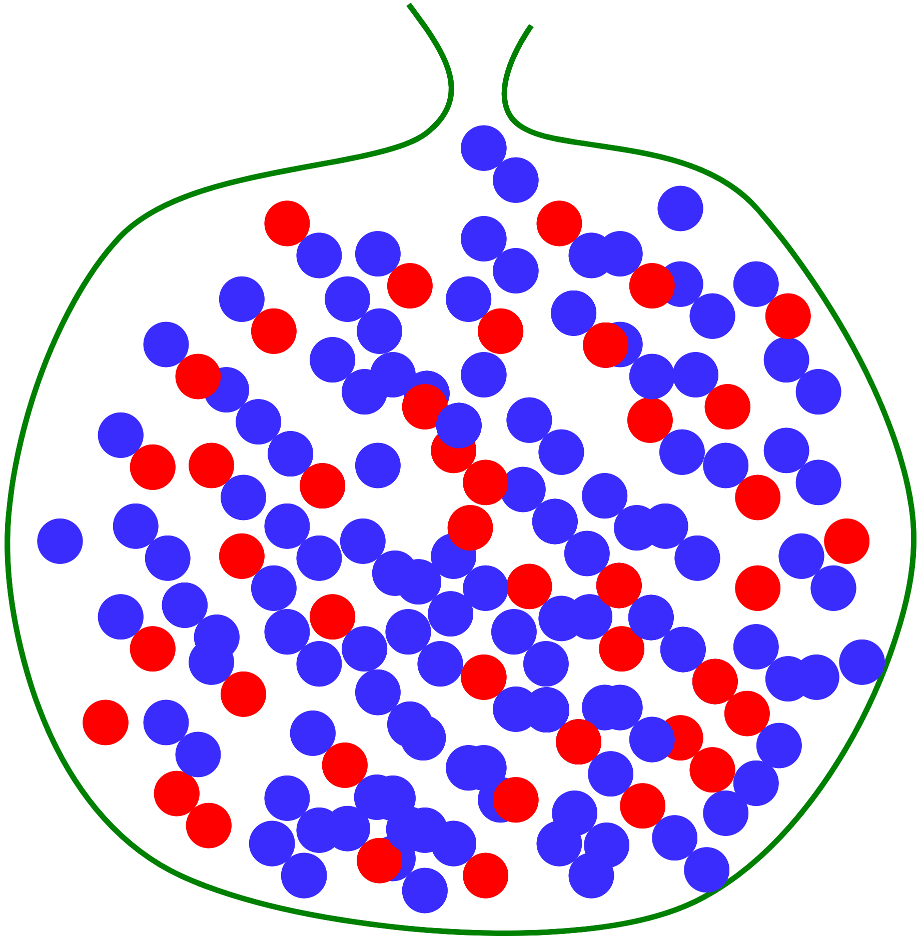
$$= p(y_1)\ \cdots p(y_n)$$

$$= \left(\frac{1}{\sqrt{2\pi s^2}}\right)^n \exp\left[-\sum_{i=1}^{n}\frac{\left(y_i-(a+bx_i)\right)^2}{2s^2}\right]$$

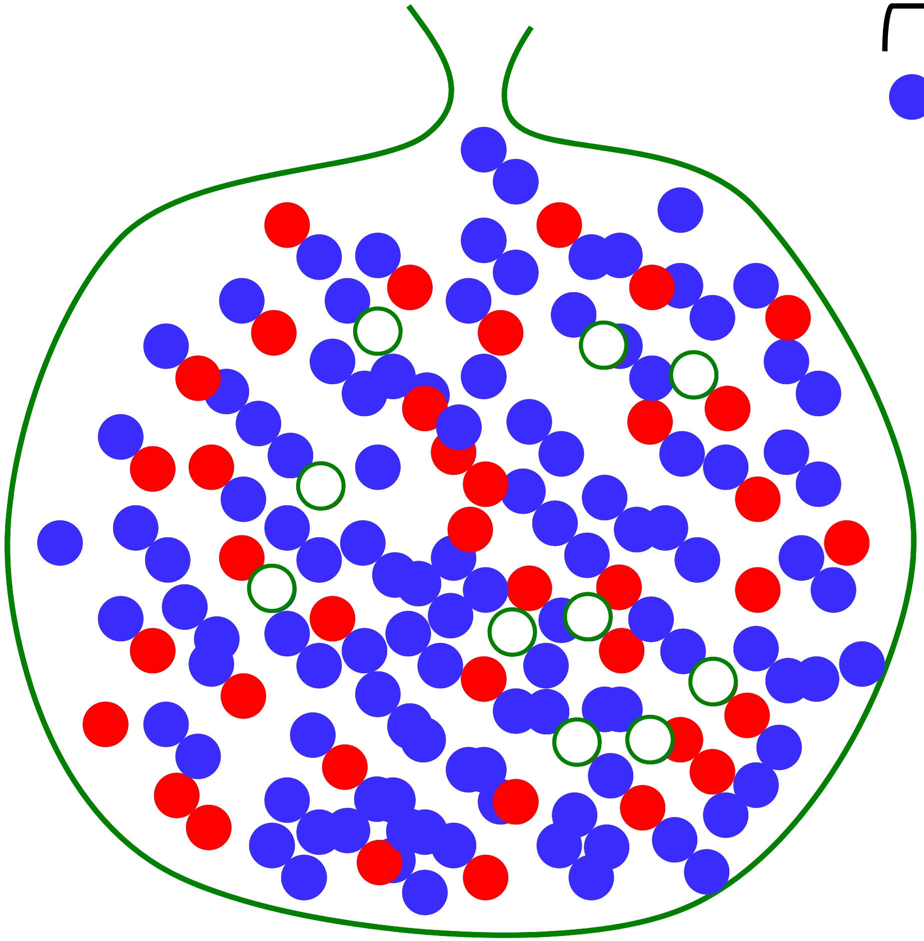


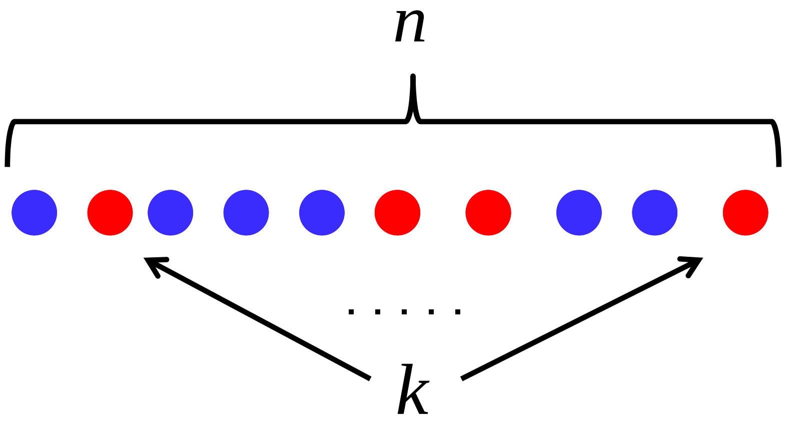
The least squares estimate is the ML estimate assuming normal distribution of error terms.

ML procedure for estimation of the probability



ML procedure for estimation of the probability





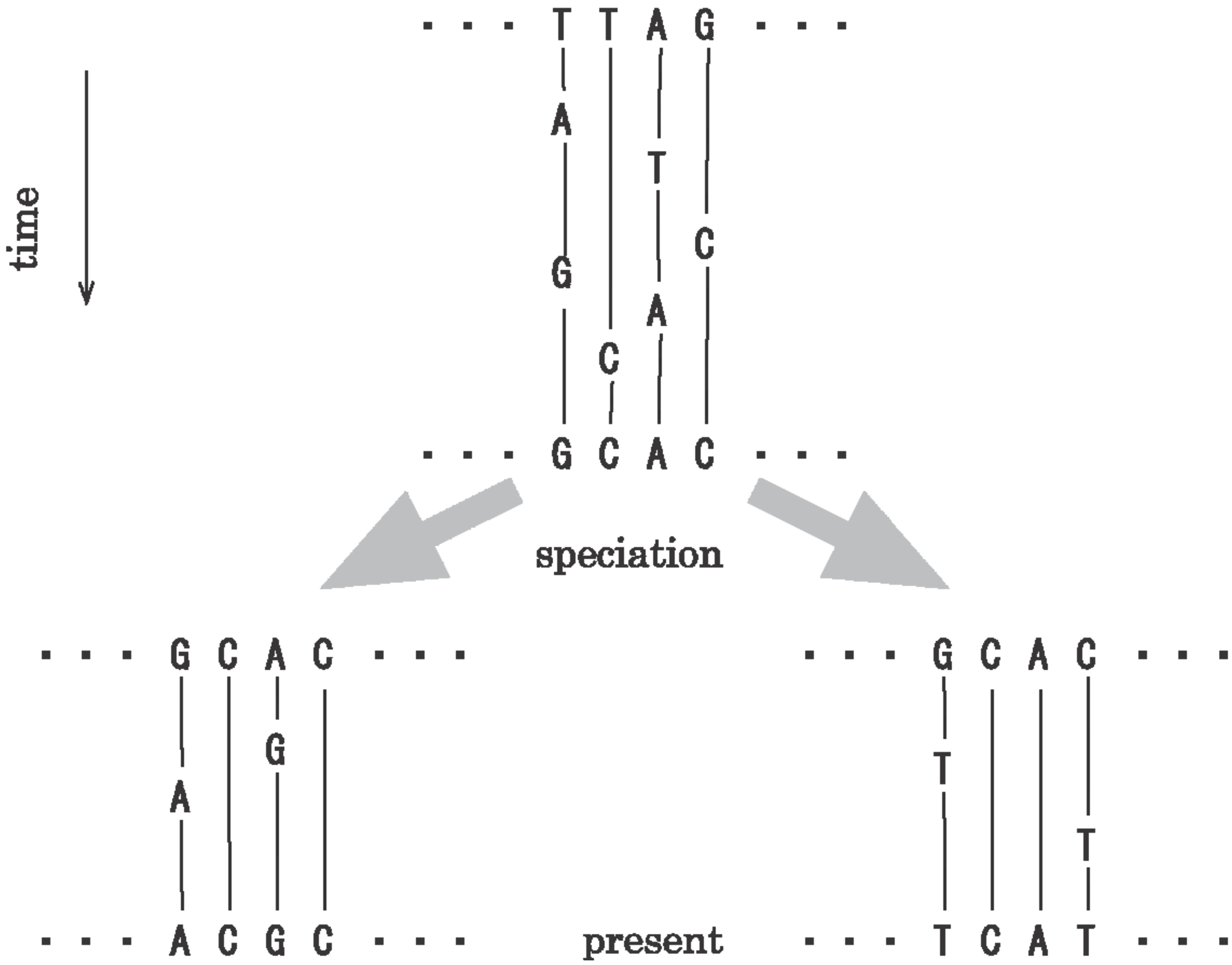
$$L = p^k (1-p)^{n-k}$$

$$\begin{aligned}\lambda &= \log L \\ &= k \log p + (n-k)\log(1-p)\end{aligned}$$

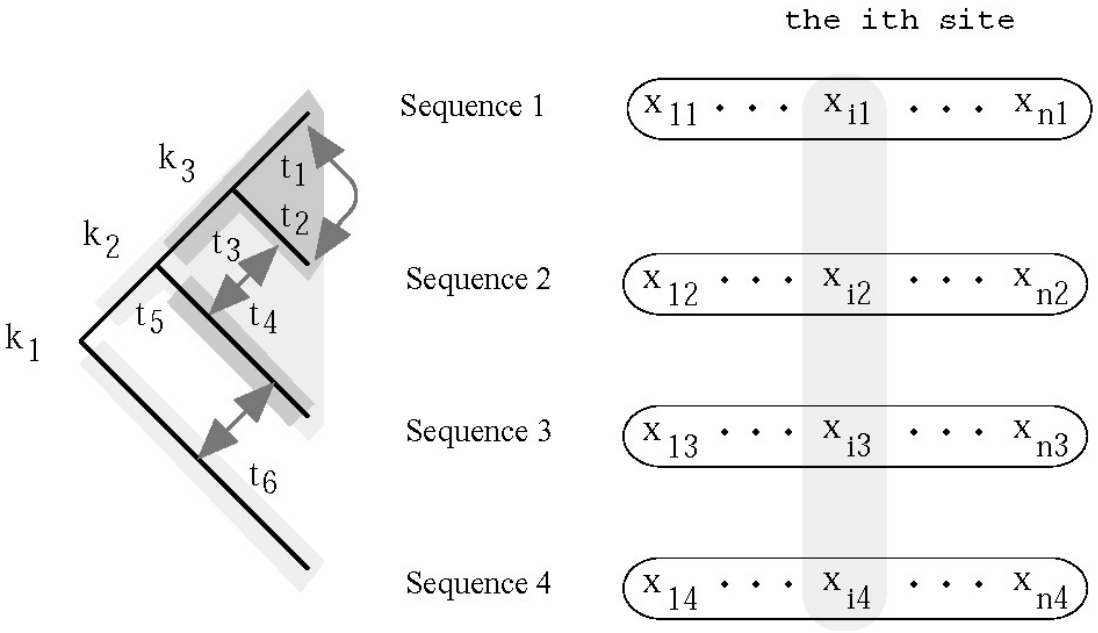
$$\frac{d\lambda}{dp} = \frac{k}{p} + \frac{n-k}{p-1} = 0$$

$$\hat{p} = \frac{k}{n}$$

Statistical model of molecular evolution



Likelihood of a site (alignment column)



$$f(x_{i1},\ldots,x_{i4}|\lambda,t_1,\ldots,t_6)$$

$$= \sum_{k_1}\pi_{k_1}p_{k_1,x_{i4}}(t_6)\left(\sum_{k_2}p_{k_1,k_2}(t_5)\left(p_{k_2,x_{i3}}(t_4)\left(\sum_{k_3}p_{k_2,k_3}(t_3)\left(p_{k_3,x_{i2}}(t_2)p_{k_3,x_{i1}}(t_1)\right)\right)\right)\right)$$

$$l(\boldsymbol{\theta}|\mathbf{X},T)=\sum_{h=1}^{n}\log f(\mathbf{X}_h|\boldsymbol{\theta},T)=\sum_{h=1}^{n}\log\left[\sum_{Z_{i_0}}\pi_{Z_{i_0}}\prod_{j\in node(T)\setminus i_0}\sum_{Z_j}P_{Z_{anc(j)}Z_j}\left(t_{anc(j),j}\right)\right]$$

$$\mathbf{P}(t)=\exp(t\mathbf{R})=\exp(tr\mathbf{R}_0)$$

branch lengths

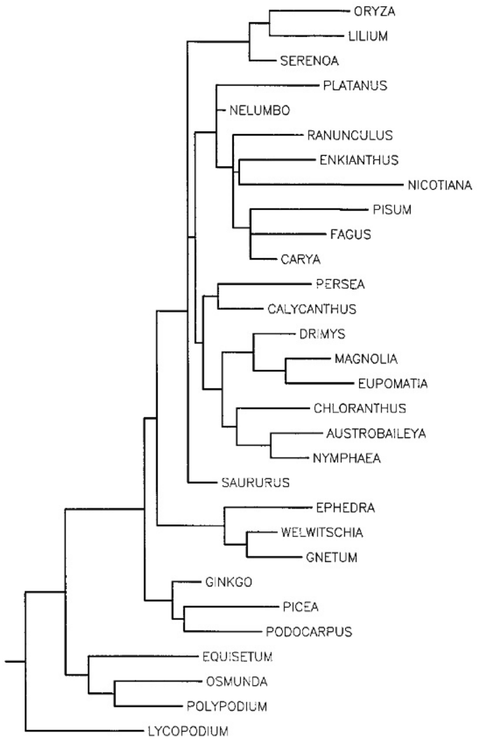
Likelihood of sequences

$$\begin{array}{lccccc}
\text{Species 1} & X_{11} & \cdots & X_{1q} & \cdots & X_{1n} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\text{Species p} & X_{p1} & \cdots & X_{pq} & \cdots & X_{pn} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\text{Species s} & X_{s1} & \cdots & X_{sq} & \cdots & X_{sn}
\end{array}$$

$$l(\mathbf{t},\mathbf{r} \mid \mathbf{X}) = \log f\left(\mathbf{X}_1 \mid \mathbf{t},\mathbf{r}\right) + \ldots + \log f\left(\mathbf{X}_q \mid \mathbf{t},\mathbf{r}\right) + \ldots + \log f\left(\mathbf{X}_n \mid \mathbf{t},\mathbf{r}\right)$$

$$\mathbf{P}(t) = \exp(t\mathbf{R}) = \exp(tr\mathbf{R}_0)$$

branch lengths



Numerical optimization package makes ML easy

An example function

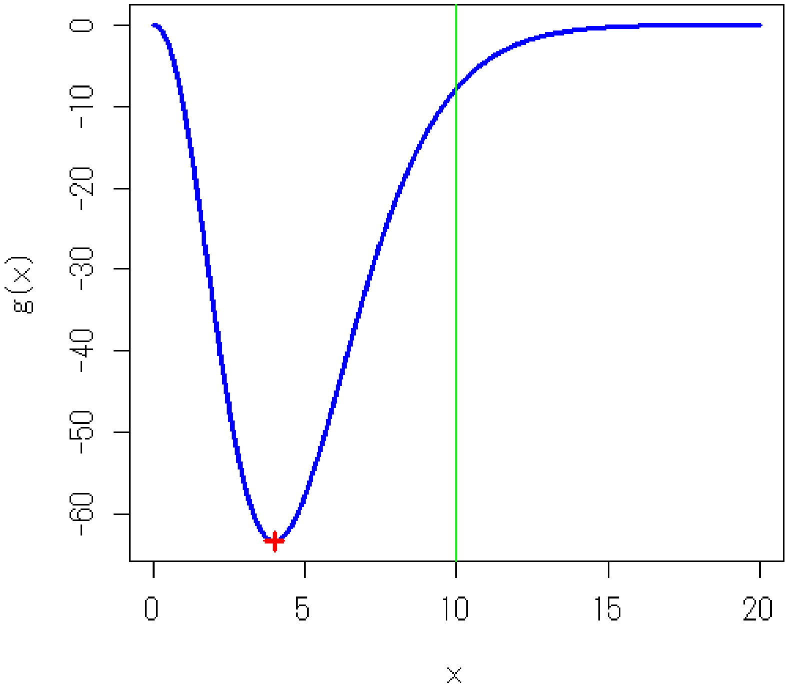
```
g <- function(x)
      { - exp(-x) * x^2*(x+2)^3}

x0 <- 10
xmin <- optim(x0, g, method="BFGS",
                     hessian=T)
xmin
```

```
$par
[1] 4.000001
$value
[1] −63.29885
$hessian
        [,1]
[1,] 13.18725
```

solution

Minimum value

Hessian (second derivative)



Once you describe a likelihood function, the computer calculates the MLE and its variance.

Logistic regression of age at sexual maturity

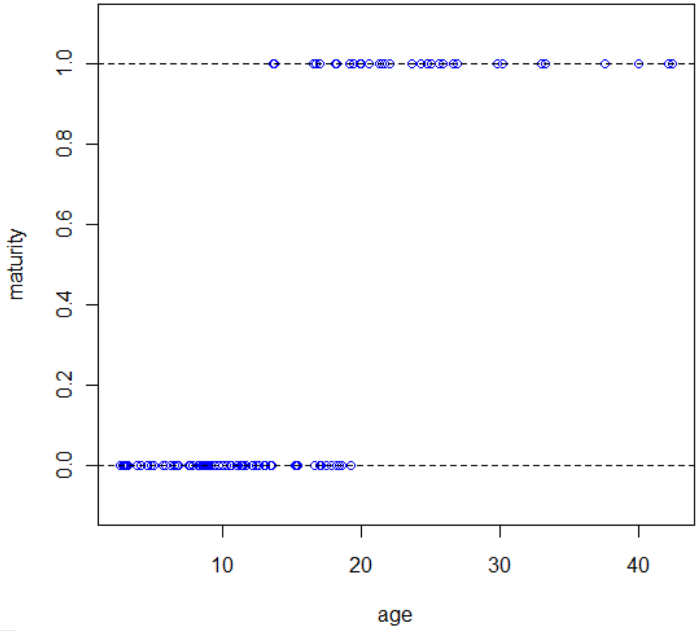
simulated data `mature_age.txt`

```
read.table("mature_age.txt")->mature_age
round(mature_age,3)
```

```
        age maturity
1    39.995        1
2    12.427        0
3    25.069        1
4    20.008        1
5    19.420        1
       . . . . .
```

matured
immatured
matured
matured
matured

```
plot(maturity~age,mature_age)
```



Generalized linear model with binomial distribution

```
maturity.glm <-
  glm(maturity~age,binomial,mature_age)
```

```
summary(maturity.glm)
```

```
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -10.2050     2.4413   -4.18 2.91e-05 ***
age           0.5730     0.1377    4.16 3.18e-05 ***
```

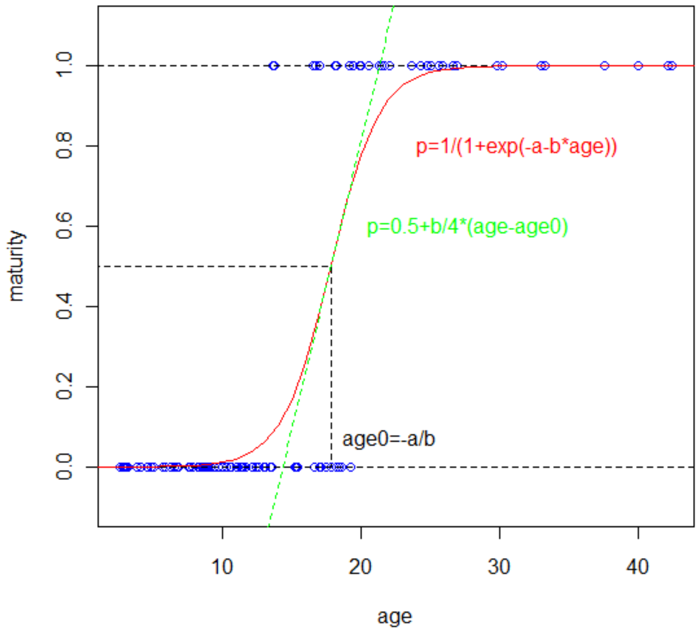
$$p = \frac{1}{1+\exp(10.205-0.573x)}$$

Logit transformation and reparametrization

Modeling the logit-transformed variable

$$\log\frac{p}{1-p} = a + bx$$

$$p = \frac{1}{1+\exp(-a-bx)}$$ Logistic function



Characterization

- $p = 0.5$, when $x = -\frac{a}{b}$
- Well approximated by a line $p = 0.5 + \frac{b}{4}\left(x + \frac{a}{b}\right)$ at $\left(-\frac{a}{b}, 0.5\right)$

Reparametrization `age0=-a/b`, `c=b/4` matches biological characters.

$$p = \frac{1}{1+\exp(-4c(x-\mathrm{age0}))}$$

Maximum likelihood inference by R

```r
nlike.glm <- function(theta,data){
          c <- theta[1]; age0 <- theta[2]
          n <- dim(data)[1]
          x <- data[,1]; y <- data[,2]
          p <- 1/(1+exp(-4*c*(x-age0)))
          loglike <- sum(dbinom(y,size=1,prob=p,log=T))
          return(-loglike)
          }

theta0 <- c(0,0)
theta_est <- optim(theta0,nlike.glm,data=mature_age,hessian=T)
theta_est
est <- theta_est$par
se <- sqrt(diag(solve(theta_est$hessian)))
mle <-
data.frame(estimate=est,standard_error=se,row.names=c("c","age0"))
round(mle,4)
```

```
     estimate standard_error
c      0.1433         0.0344
age0  17.8097         0.6923
```

logistic_ML.R

Fitting K80 model to mtDNA sequences of human and chimpanzee

mtCDNA_human.nuc

```
CTACCCGCCGCAGTACTGATCATTCTATTTCCCCCTCTATTGATCCCCACCTCCAAATAT
CTCATCAACAACCGACTAATTACCACCCAACAATGACTAATCAAACTAACCTCAAAACAA
ATGATAGCCATACACAACACTAAAGGACGAACCTGATCTCTTATACTAGTATCCTTAATC
. . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
```

mtCDNA_chimp.nuc

```
TTACCCGCCGCAGTACTAATCATTCTATTCCCCCCTCTACTGGTCCCCACTTCTAAACAT
CTCATCAACAACCGACTAATTACCACCCAACAATGACTAATTCAACTGACCTCAAAACAA
ATAATAACTATACACAGCACTAAAGGACGAACCTGATCTCTCATACTAGTATCCTTAATC
. . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
```

```
scan("mtCDNA_human.nuc",what="")->human
human <- strsplit(human,split="")
human <- unlist(human)
```

```
 [1] ″C″ ″T″ ″A″ ″C″ ″C″ ″C″ ″G″ ″C″ ″C″ ″G″ ″C″ ″A″ ″G″ ″T″ ″A″ ″C″
[17] ″T″ ″G″ ″A″ ″T″ ″C″ ″A″ ″T″ ″T″ ″C″ ″T″ ″A″ ″T″ ″T″ ″T″ ″C″ ″C″
[33] ″C″ ″C″ ″C″ ″T″ ″C″ ″T″ ″A″ ″T″ ″T″ ″G″ ″A″ ″T″ ″C″ ″C″ ″C″ ″C″
[49] ″A″ ″C″ ″C″ ″T″ ″C″ ″C″ ″A″ ″A″ ″A″ ″T″ ″A″ ″T″ ″C″ ″T″ ″C″ ″A″
```

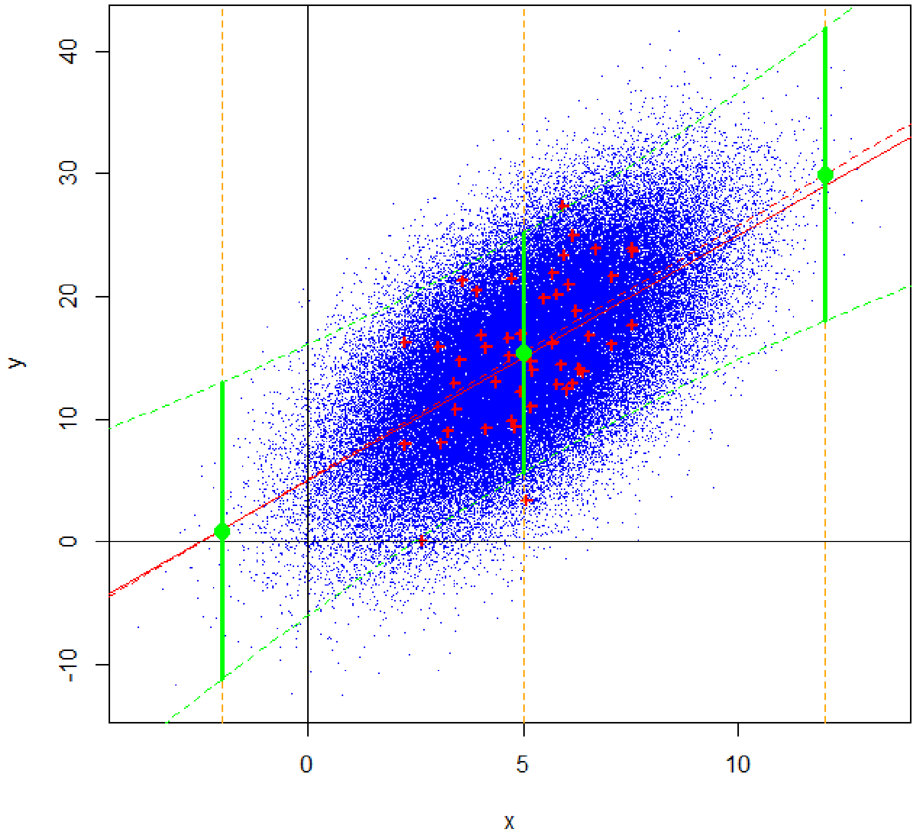
```
table(human,chimp)
```

```
 chimp
human    A    C    G    T
    A 2954   17  141   16
    C   18 3163    4  374
    G  165    5 1110    2
    T   15  310    2 2411
```

```
        estimate standard error
t         0.0037         0.0002
alpha    27.7425         1.9271
```

Kimura80_pair_ML.R

Precision of the prediction and the complexity of the model



Complex models may improve the fitting to the data. However, the variance of the estimates become large, if the model include too many parameters beyond the amount of information in the data.

Model selection based on the precision of the prediction

Variable selection of multiple regression

$$y_i = a + b_1 x_{1i} + \dots + b_k x_{ki} + \varepsilon_i \quad (i = 1, \dots, n)$$

Final prediction error FPE：

$$\mathrm{FPE} = \frac{\mathrm{n+k+1}}{\mathrm{n\text{-}k\text{-}1}} \hat{\sigma}^2$$

$$\hat{\sigma}^2 = \frac{1}{\mathrm{n}} \sum_{i=1}^{n} \left(y_i - \hat{a} - \hat{b}_1 x_{1i} - \dots - \hat{b}_k x_{ki} \right)^2$$

Prediction error is obtained by discounting the sum of squared residuals

By considering the predictive power for general models, Akaike (1974) derived a criteria defined by

$$\mathrm{AIC} = -2 \times (\text{maximum likelihood value}) + 2 \times (\text{number of parameters})$$

In the special case of regression analysis,

$$\mathrm{AIC} = n \log \hat{\sigma}^2 + 2(k+2) \approx \log \mathrm{FPE} \text{ - } \mathrm{nlogn}$$

Plot trees by ape

Read newick file

```
library(ape)
read.tree("desaturase_tree.nwk")->tree
summary(tree)
```

```
Phylogenetic tree: tree

  Number of tips: 22
  Number of nodes: 20
  Branch lengths:
    mean: 0.122944
    variance: 0.02444867
    distribution summary:
      Min.    1st Qu.     Median    3rd Qu.       Max.
0.00000000 0.02457463 0.05800963 0.16716395 0.66505495
  No root edge.
  First ten tip labels: HzeZ9_18
                        TniZ9_18
                        BmoZ9_18
                        EpoZ9_18
                        OfuZ9_18
                        OnuZ9_18
                        EpoZ9_16
                        PocZ9_16
                        AveZ9_16
                        HzeZ9_16
  No node labels.
```

Look at the content

```
names(tree)
```

```
[1] "edge"        "edge.length" "Nnode"       "tip.label"
```

```
tree$edge
```

```
     [,1] [,2]
[1,]   23   24
[2,]   24   25
[3,]   25   26
 . . .
```

```
tree$edge.length
```

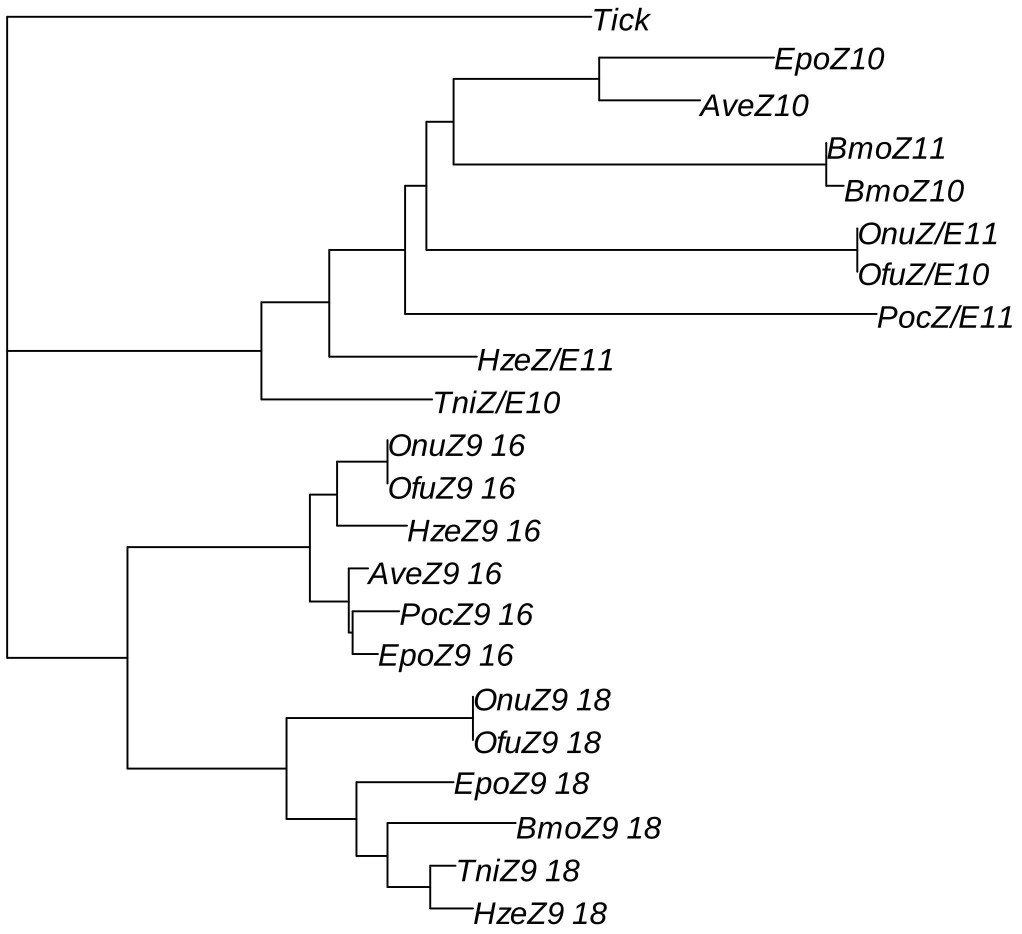
```
 [1] 0.13716450 0.18089948 0.08089457 0.03392684 0.04898512 0.04775221
 [7] 0.02881577 0.14738616 0.11048854 0.21265592 0.00000000 0.00000000
. . .
```

```
tree$tip.label
```

```
 [1] "HzeZ9_18" "TniZ9_18" "BmoZ9_18" "EpoZ9_18" "OfuZ9_18" "OnuZ9_18"
 [7] "EpoZ9_16" "PocZ9_16" "AveZ9_16" "HzeZ9_16" "OfuZ9_16" "OnuZ9_16"
[13] "TniZ/E10" "HzeZ/E11" "PocZ/E11" "OfuZ/E10" "OnuZ/E11" "BmoZ10"
[19] "BmoZ11"   "AveZ10"   "EpoZ10"   "Tick"
```

Plot the tree

```
plot(tree)
```



Check the structure of tip-labels and node labels

```
tree0 <- tree
summary(tree0)
```

```
Phylogenetic tree: tree

  Number of tips: 22
  Number of nodes: 20
  Branch lengths:
    mean: 0.122944
    variance: 0.02444867
    distribution summary:
      Min.    1st Qu.     Median    3rd Qu.       Max. 
0.00000000 0.02457463 0.05800963 0.16716395 0.66505495 
. . .
```

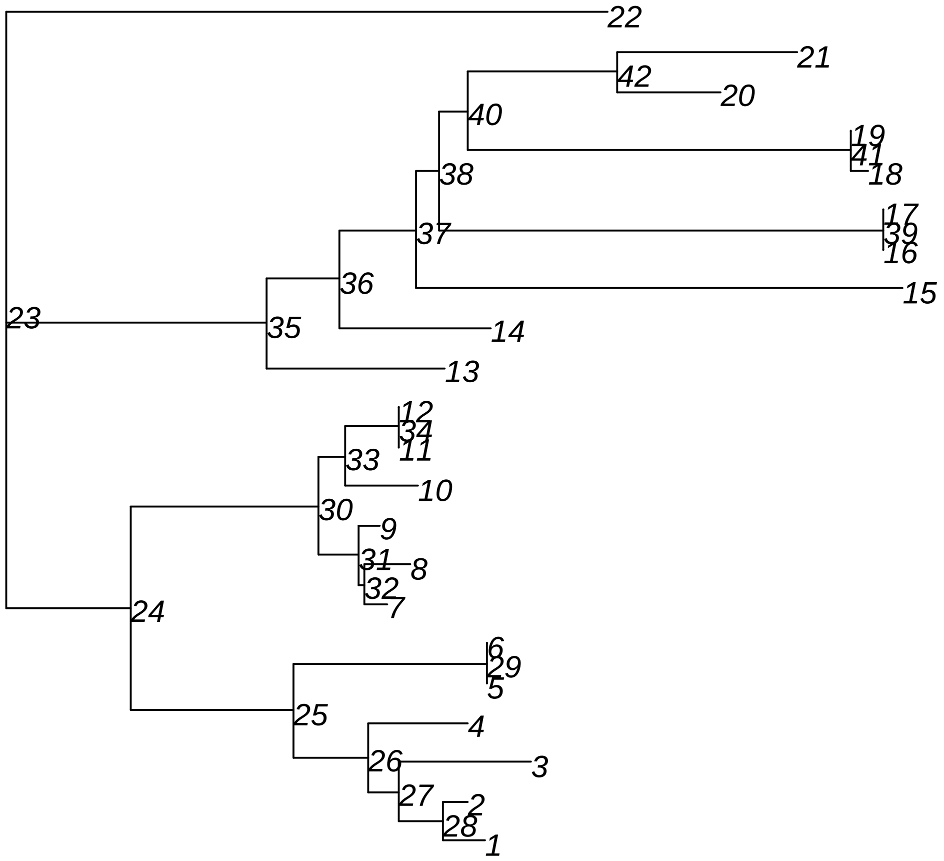
```
tree0$tip.label <- 1:22
tree0$node.label <- 22+1:20
```

Check the structure of edges

```
plot(tree0,show.node.label=T)
```

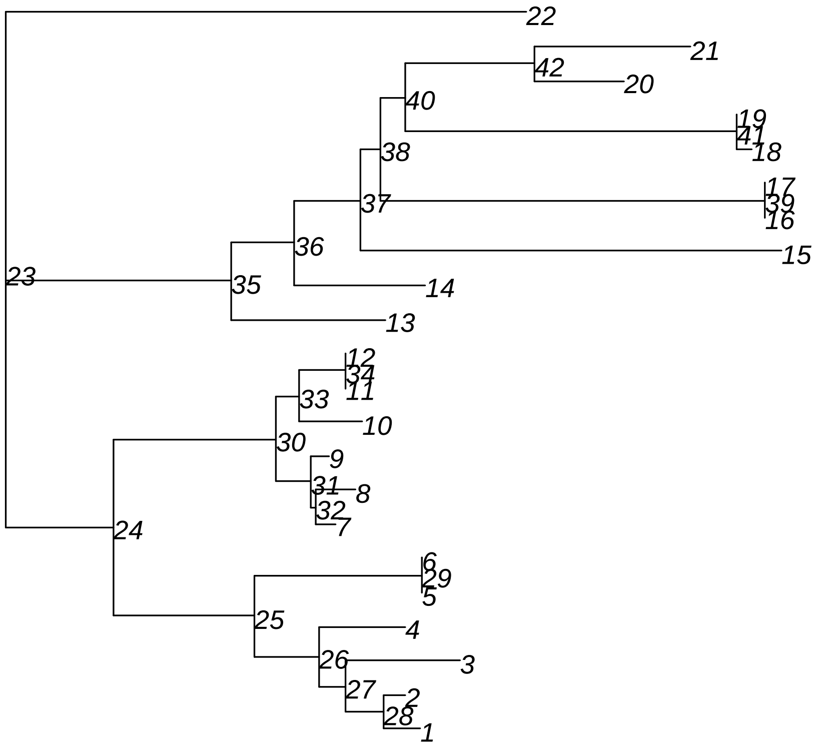
```
tree$edge
```

```
  [,1] [,2]
  [1,]   23   24
  [2,]   24   25
  [3,]   25   26
  [4,]   26   27
  [5,]   27   28
  [6,]   28    1
  [7,]   28    2
  [8,]   27    3
  [9,]   26    4
 [10,]   25   29
 [11,]   29    5
 [12,]   29    6
 [13,]   24   30
 [14,]   30   31
 [15,]   31   32
 [16,]   32    7
 [17,]   32    8
 [18,]   31    9
. . . . .
```



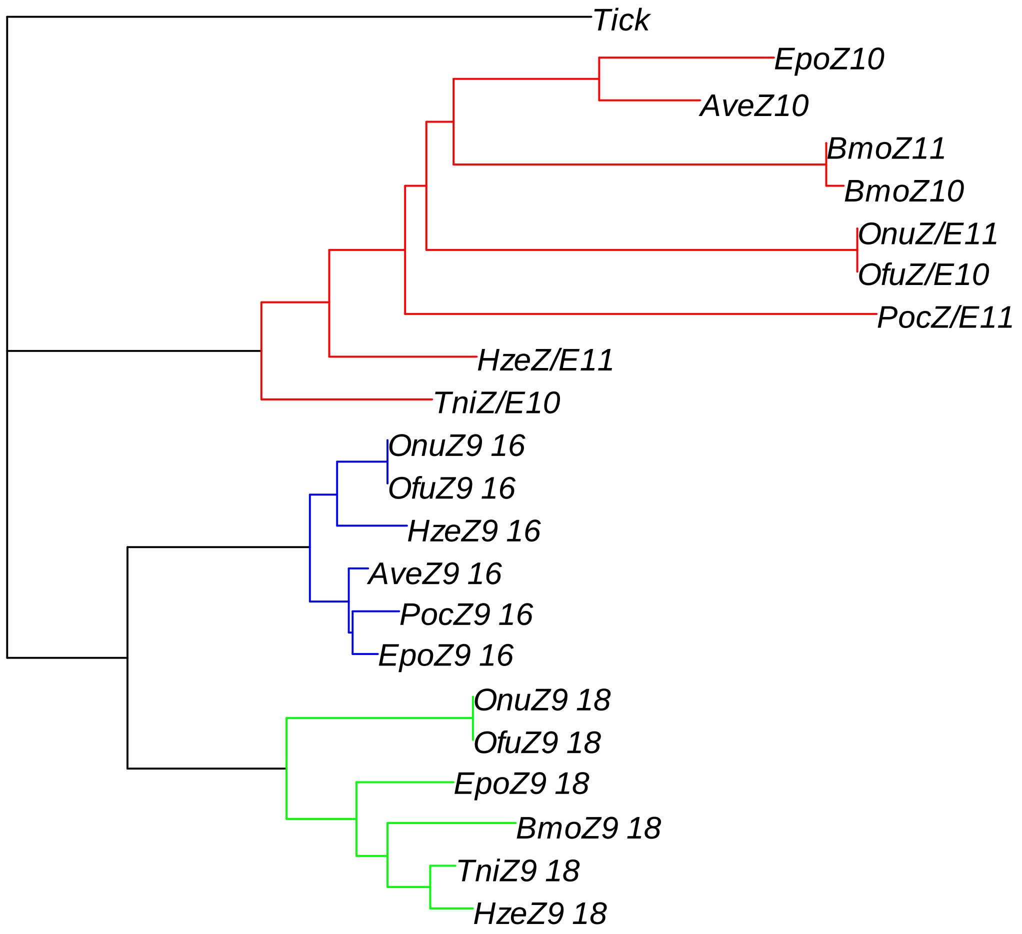
Define the colors of edges for the three groups

```
edges1 <- which.edge(tree, group=tree$tip.label[1:6])
edges2 <- which.edge(tree, group=tree$tip.label[7:12])
edges3 <- which.edge(tree, group=tree$tip.label[13:21])

edge.colors <- rep("black",dim(tree$edge)[1])
edge.colors[edges1] <- "green"
edge.colors[edges2] <- "blue"
edge.colors[edges3] <- "red"
```



Plot the tree with colors for the three groups

```
plot(tree,edge.color=edge.colors)
```



Assignment 1

Construct the phylogenetic tree of desaturase by the models; LG+G, LG, WAG+G, WAG. Compare the values of AICc and BIC, and interpret the difference of the phylogenetic trees. Please also try ape.

Please send the word file / pdf file named **`agri1_name.doc`** / **`agri1_name.pdf`** to:

Hirohisa Kishino (kishino@lbm.ab.a.u-tokyo.ac.jp).

Here, "name" should be replaced by your name.

Deadline: 26 May 2019 (Sunday)