SUBSTITUTION MODELS AND TREE BUILDING – MEGA 11

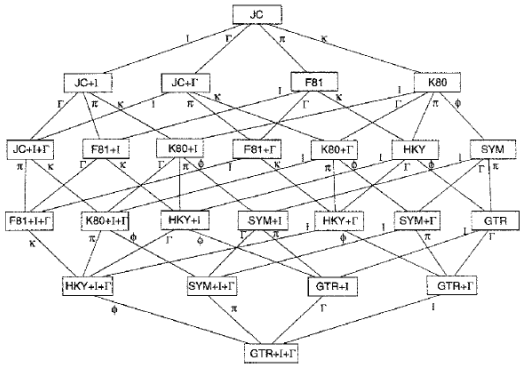
AARHUS
UNIVERSITY
DEPARTMENT OF MOLECULAR BIOLOGY AND GENETICS

CALIN PANTEA
PHD STUDENT

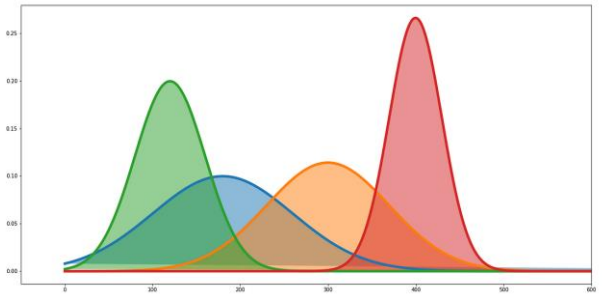
BRIEF RECAP

14:15-14:25

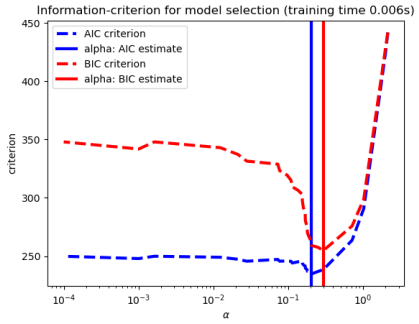
-> substitution models



-> MLE



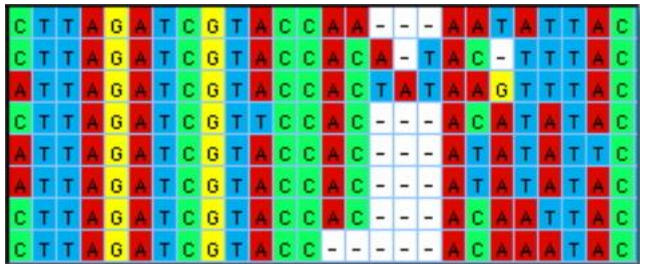
-> BIC and AIC



TREE BUILDING

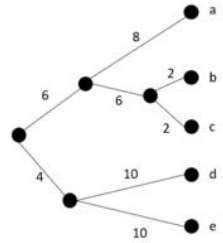
14:15-14:25

- sequence alignment

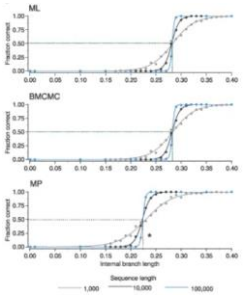


- sequence distance can be stored as a distance matrix and visualised into phylogenetic trees

M	a	b	c	d	e
a	0	16	16	28	28
b	16	0	4	28	28
c	16	4	0	28	28
d	28	28	28	0	20
e	28	28	28	20	0



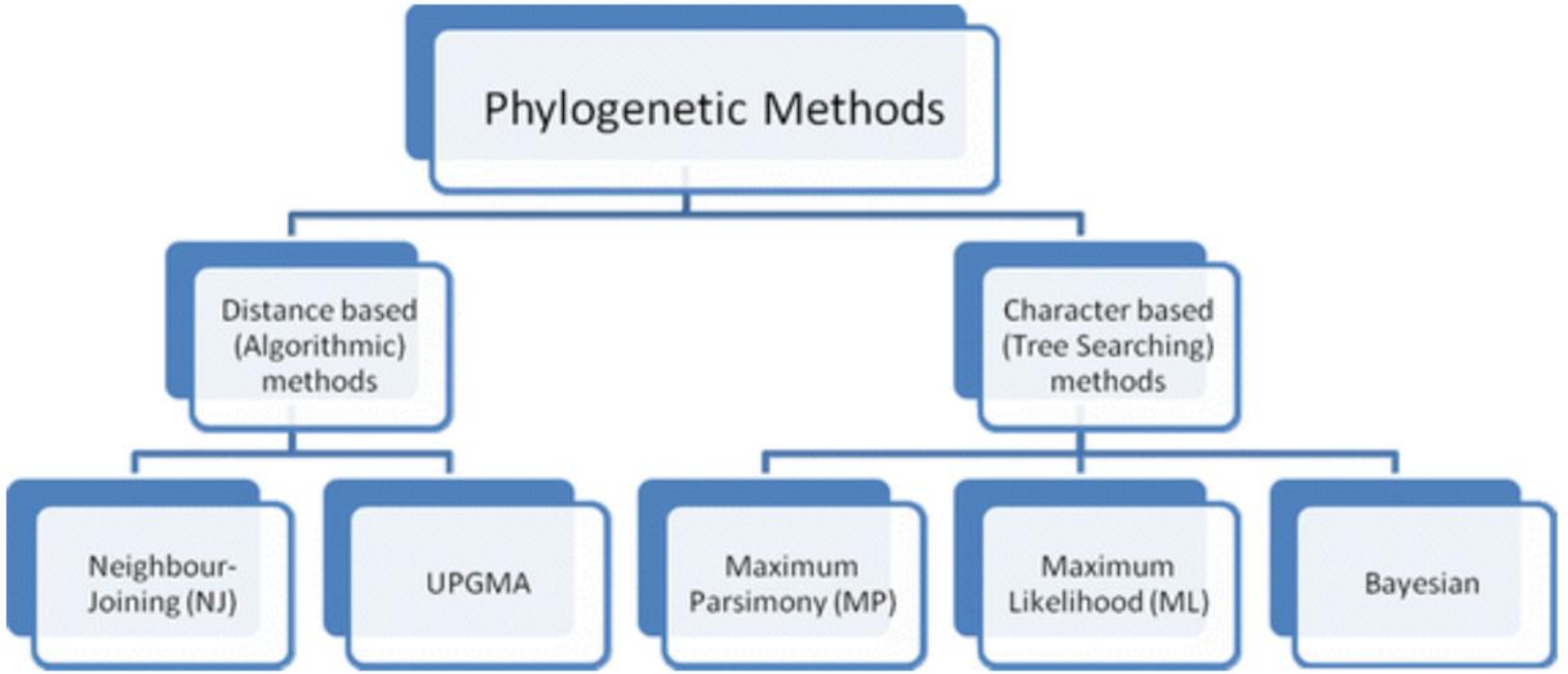
- nucleotide substitution models can be used to infer trees based on their probability to fit the optimal model



- rooted (LUCA node) vs unrooted (only leaf node relatedness)

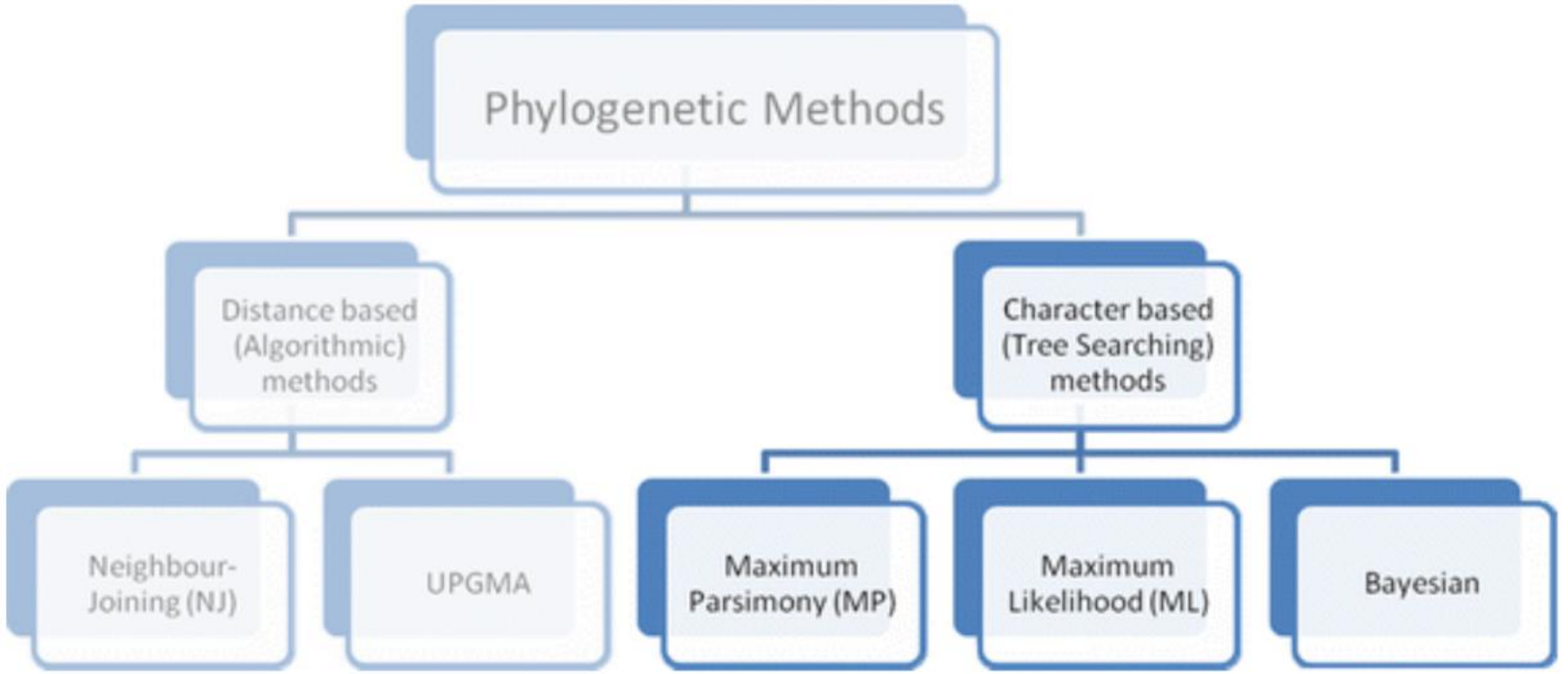
TREE BUILDING

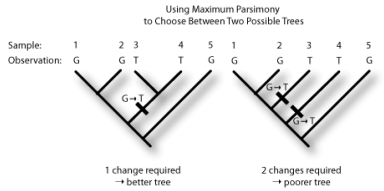
14:15-14:25

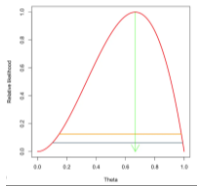


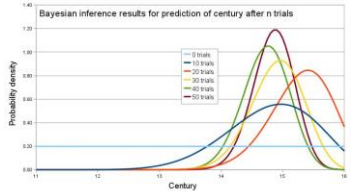
TREE BUILDING

14:15-14:25



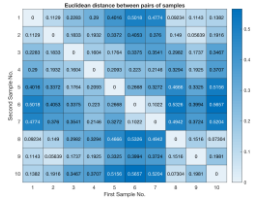


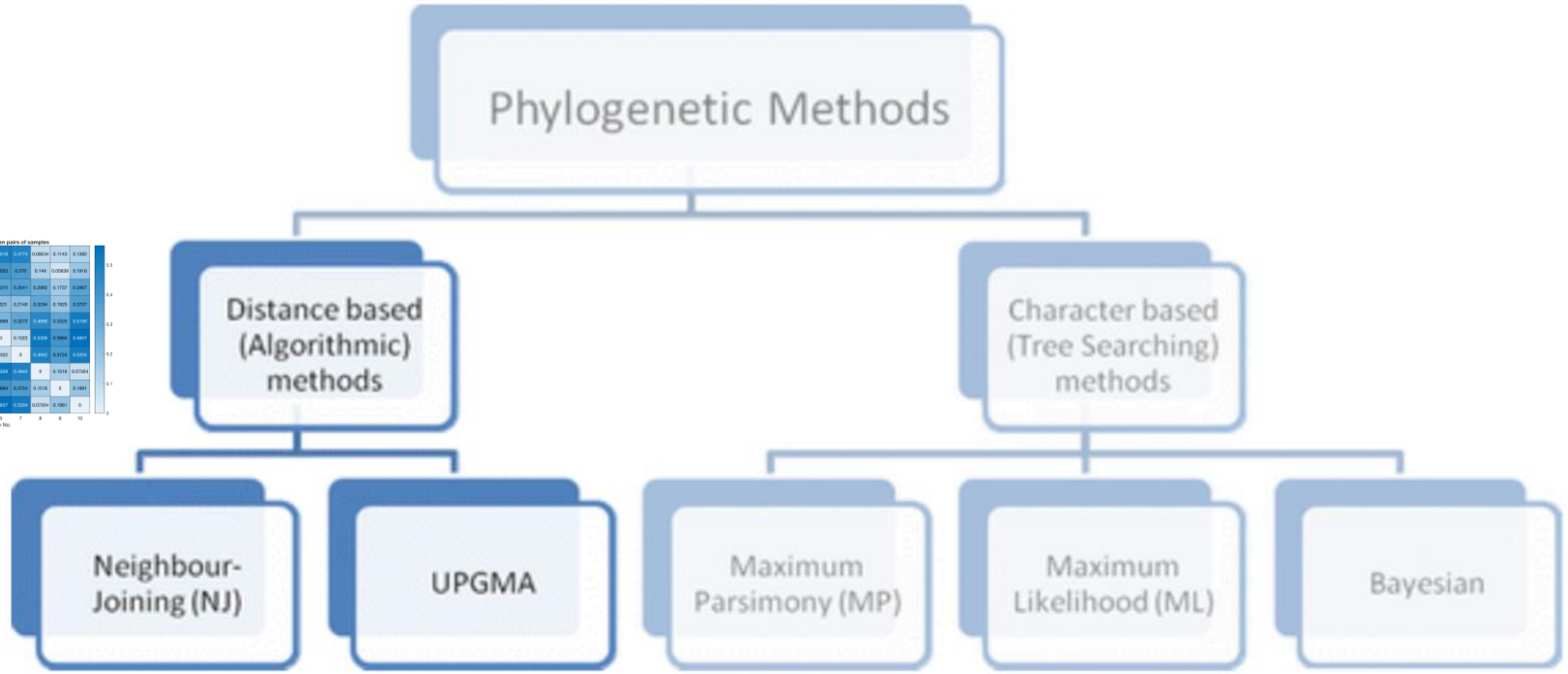


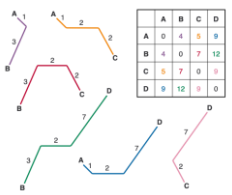


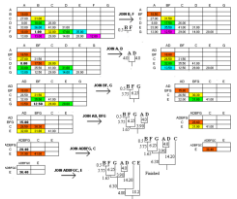
TREE BUILDING

14:15-14:25









TREE BUILDING – NJ

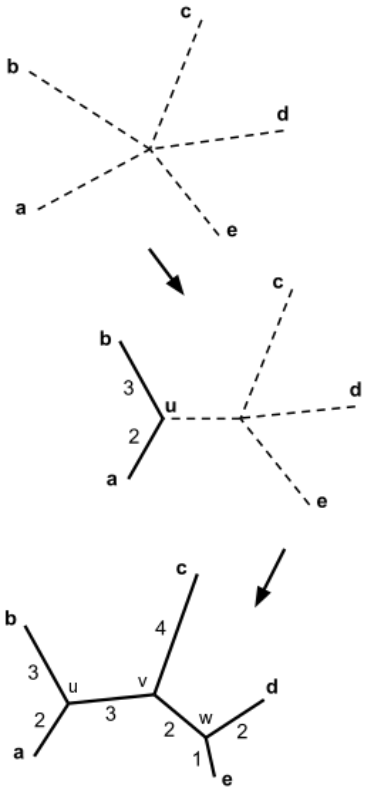
14:15-14:25

Pairwise distance-based clustering

	a	b	c	d	e
a	0	5	9	9	8
b	5	0	10	10	9
c	9	10	0	8	7
d	9	10	8	0	3
e	8	9	7	3	0

$$Q(i,j) = (n-2)d(i,j) - \sum_{k=1}^{n} d(i,k) - \sum_{k=1}^{n} d(j,k)$$

	a	b	c	d	e
a		−50	−38	−34	−34
b	−50		−38	−34	−34
c	−38	−38		−40	−40
d	−34	−34	−40		−48
e	−34	−34	−40	−48	



TREE BUILDING – UPGMA

14:15-14:25

Also pairwise distance-based clustering

http://www.slimsuite.unsw.edu.au/teaching/upgma/

	A	B	C	D	E	F	G
A							
B	19.00						
C	27.00	31.00					
D	8.00	18.00	26.00				
E	33.00	36.00	41.00	31.00			
F	18.00	1.00	32.00	17.00	35.00		
G	13.00	13.00	29.00	14.00	28.00	12.00	

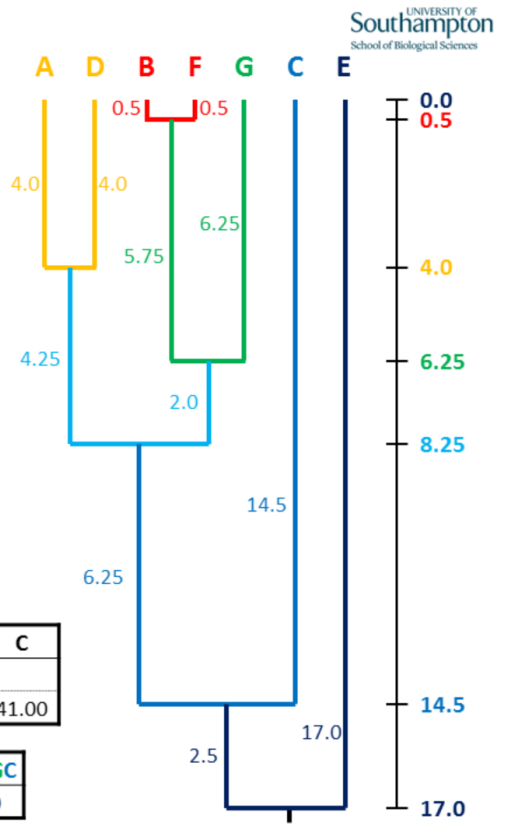
	A	BF	C	D	E
BF	18.50				
C	27.00	31.50			
D	8.00	17.50	26.00		
E	33.00	35.50	41.00	31.00	
G	13.00	12.50	29.00	14.00	28.00

	AD	BF	C	E
BF	18.00			
C	26.50	31.50		
E	32.00	35.50	41.00	
G	13.50	12.50	29.00	28.00

	AD	BFG	C
BFG	16.50		
C	26.50	30.67	
E	32.00	33.00	41.00

	ADBFG	C
C	29.00	
E	32.60	41.00

	ADBFGC
E	34.00



MEGA – MODEL SELECTION AND TREE BUILDING

14:25-14:55
15:10-15:45

https://github.com/Bjarke-M/Evolutionary_Thinking_2023/tree/main/week36/Friday

- Exercise 1: Substitution model testing
- Exercise 2: Tree building – MEGA tutorial
- Exercise 3: Tree building – application

EXERCISES

14:25-14:55
15:10-15:45

~1 hour (with cake break)

BREAK ‘TILL 15:10

14:55-15:10

EXERCISES

14:25-14:55
15:10-15:45

~30’ left (post cake break)

Then short discussion

EXERCISE 1

15:45-15:50

Table. Maximum Likelihood fits of 24 different nucleotide substitution models

Model	Parameter	BIC	AICc	lnL	(+I)	(+G)	R	f(A)	f(T)	f(C)	f(G)	r(AT)	r(AC)	r(AG)	r(TA)	r(TC)	r(TG)	r(CA)	r(CT)	r(CG)	r(GA)	r(GT)	r(GC)
HKY+G	12	5341.97	5265.154	-2620.54	n/a	0.23	10.67	0.311	0.253	0.329	0.107	0.01	0.014	0.098	0.013	0.302	0.004	0.013	0.232	0.004	0.285	0.01	0.014
TN93+G	13	5350.248	5267.037	-2620.48	n/a	0.23	10.54	0.311	0.253	0.329	0.107	0.01	0.014	0.103	0.013	0.29	0.004	0.013	0.223	0.004	0.301	0.01	0.014
HKY+G+I	13	5350.509	5267.298	-2620.61	0.52	1.34	10.39	0.311	0.253	0.329	0.107	0.011	0.014	0.098	0.013	0.301	0.005	0.013	0.232	0.005	0.285	0.011	0.014
TN93+G+I	14	5357.915	5268.309	-2620.11	0.57	2.53	10.37	0.311	0.253	0.329	0.107	0.01	0.013	0.117	0.013	0.261	0.004	0.013	0.201	0.004	0.34	0.01	0.013
GTR+G	16	5369.669	5267.276	-2617.58	n/a	0.25	9.84	0.311	0.253	0.329	0.107	0.008	0.024	0.103	0.01	0.295	0	0.023	0.227	0.003	0.3	0	0.008
HKY+I	12	5376.019	5299.203	-2637.57	0.34	n/a	5.26	0.311	0.253	0.329	0.107	0.02	0.025	0.09	0.024	0.279	0.008	0.024	0.214	0.008	0.263	0.02	0.025
GTR+G+I	17	5377.728	5268.943	-2617.4	0.56	2.87	9.93	0.311	0.253	0.329	0.107	0.009	0.024	0.116	0.011	0.268	0	0.022	0.206	0.001	0.339	0	0.004
TN93+I	13	5384.221	5301.009	-2637.46	0.34	n/a	5.24	0.311	0.253	0.329	0.107	0.019	0.025	0.094	0.024	0.27	0.008	0.024	0.208	0.008	0.275	0.019	0.025
GTR+I	16	5398.716	5296.324	-2632.1	0.34	n/a	5.34	0.311	0.253	0.329	0.107	0.016	0.037	0.096	0.02	0.278	0	0.035	0.213	0.007	0.279	0	0.02
HKY	11	5421.587	5351.167	-2664.55	n/a	n/a	4.47	0.311	0.253	0.329	0.107	0.022	0.029	0.088	0.027	0.271	0.009	0.027	0.208	0.009	0.256	0.022	0.029
TN93	12	5429.942	5353.126	-2664.53	n/a	n/a	4.47	0.311	0.253	0.329	0.107	0.022	0.029	0.09	0.027	0.268	0.009	0.027	0.206	0.009	0.261	0.022	0.029
GTR	15	5440.687	5344.687	-2657.29	n/a	n/a	4.51	0.311	0.253	0.329	0.107	0.019	0.043	0.092	0.023	0.274	0	0.041	0.211	0.007	0.267	0	0.023
T92+G	10	5513.101	5449.078	-2714.52	n/a	0.53	5.39	0.282	0.282	0.218	0.218	0.022	0.017	0.184	0.022	0.184	0.017	0.022	0.238	0.017	0.238	0.022	0.017
T92+I	10	5514.197	5450.175	-2715.06	0.48	n/a	5.09	0.282	0.282	0.218	0.218	0.023	0.018	0.183	0.023	0.183	0.018	0.023	0.236	0.018	0.236	0.023	0.018
T92+G+I	11	5521.504	5451.085	-2714.51	0.09	0.65	5.4	0.282	0.282	0.218	0.218	0.022	0.017	0.184	0.022	0.184	0.017	0.022	0.239	0.017	0.239	0.022	0.017
K2+G	9	5525.136	5467.512	-2724.74	n/a	0.51	5.44	0.25	0.25	0.25	0.25	0.019	0.019	0.211	0.019	0.211	0.019	0.019	0.211	0.019	0.211	0.019	0.019
K2+G+I	10	5533.545	5469.523	-2724.74	0.04	0.55	5.45	0.25	0.25	0.25	0.25	0.019	0.019	0.211	0.019	0.211	0.019	0.019	0.211	0.019	0.211	0.019	0.019
K2+I	9	5534.578	5476.954	-2729.46	0.34	n/a	4.7	0.25	0.25	0.25	0.25	0.022	0.022	0.206	0.022	0.206	0.022	0.022	0.206	0.022	0.206	0.022	0.022
T92	9	5547.364	5489.74	-2735.85	n/a	n/a	4.31	0.282	0.282	0.218	0.218	0.026	0.02	0.178	0.026	0.178	0.02	0.026	0.23	0.02	0.23	0.026	0.02
K2	8	5561.739	5510.514	-2747.24	n/a	n/a	4.31	0.25	0.25	0.25	0.25	0.024	0.024	0.203	0.024	0.203	0.024	0.024	0.203	0.024	0.203	0.024	0.024
JC+I	8	5865.808	5814.583	-2899.28	0.41	n/a	0.5	0.25	0.25	0.25	0.25	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083
JC+G	8	5867.097	5815.872	-2899.92	n/a	0.91	0.5	0.25	0.25	0.25	0.25	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083
JC+G+I	9	5874.225	5816.601	-2899.28	0.4	200	0.5	0.25	0.25	0.25	0.25	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083
JC	7	5883.048	5838.223	-2912.1	n/a	n/a	0.5	0.25	0.25	0.25	0.25	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083

NOTE.-- Models with the lowest BIC scores (Bayesian Information Criterion) are considered to describe the substitution pattern the best. For each model, AICc value (Akaike Information Criterion, corrected), Maximum Likelihood value (lnL), and the number of

Abbreviations: TR: General Time Reversible; HKY: Hasegawa-Kishino-Yano; TN93: Tamura-Nei; T92: Tamura 3-parameter; K2: Kimura 2-parameter; JC: Jukes-Cantor./div>

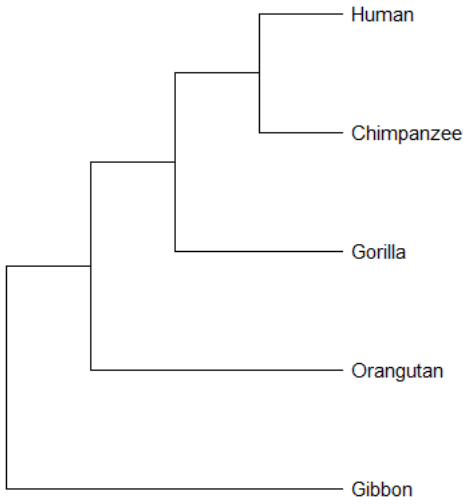
1. Nei M. and Kumar S. (2000). Molecular Evolution and Phylogenetics. Oxford University Press, New York.

2. Tamura K., Stecher G., and Kumar S. (2021). MEGA 11: Molecular Evolutionary Genetics Analysis Version 11. Molecular Biology and Evolution https://doi.org/10.1093/molbev/msab120.

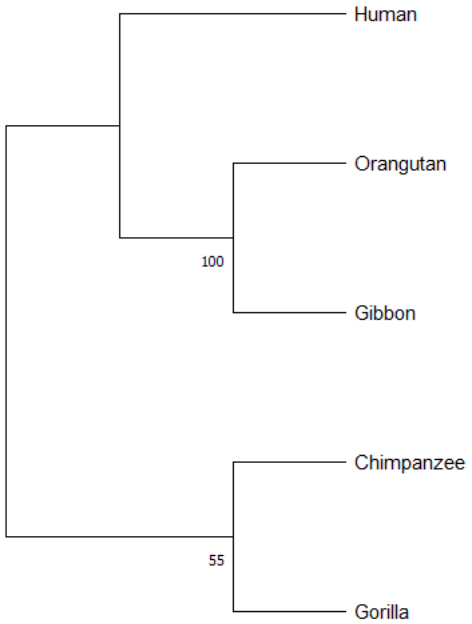
EXERCISE 2

15:45-15:50

NJ



ML

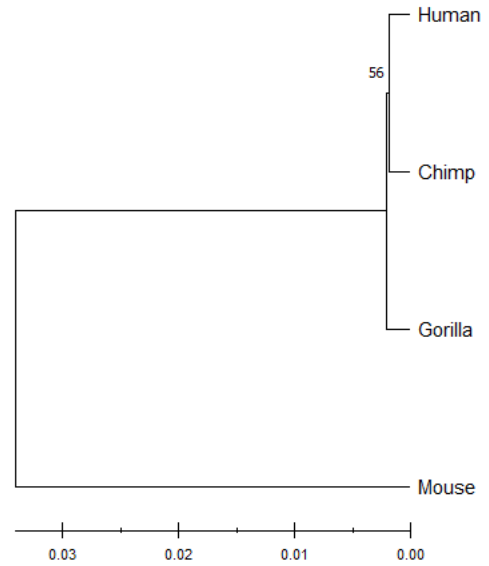


EXERCISE 3

15:45-15:50

Table. Maximum Likelihood fits of 24 different nucleotide substitution models

Model	Parameters	BIC	AICc	*lnL*	(+*I*)	(+*G*)	*R*	*f*(A)	*f*(T)	*f*(C)	*f*(G)	*r*(AT)	*r*(AC)	*r*(AG)	*r*(TA)	*r*(TC)	*r*(TG)	*r*(CA)	*r*(CT)	*r*(CG)	*r*(GA)	*r*(GT)	*r*(GC)
HKY	9	7353.758	7290.258	-3636.118	n/a	n/a	3.70	0.314	0.200	0.264	0.222	0.021	0.028	0.175	0.033	0.208	0.023	0.033	0.158	0.023	0.248	0.021	0.028
TN93	10	7361.503	7290.950	-3635.462	n/a	n/a	3.71	0.314	0.200	0.264	0.222	0.021	0.028	0.193	0.033	0.183	0.023	0.033	0.139	0.023	0.274	0.021	0.028
HKY+I	10	7362.626	7292.073	-3636.023	0.37	n/a	3.83	0.314	0.200	0.264	0.222	0.021	0.027	0.176	0.032	0.210	0.023	0.032	0.159	0.023	0.250	0.021	0.027
HKY+G	10	7362.634	7292.081	-3636.028	n/a	1.32	3.82	0.314	0.200	0.264	0.222	0.021	0.027	0.176	0.032	0.210	0.023	0.032	0.159	0.023	0.250	0.021	0.027
TN93+I	11	7370.298	7292.692	-3635.331	0.43	n/a	3.88	0.314	0.200	0.264	0.222	0.020	0.027	0.196	0.032	0.183	0.023	0.032	0.139	0.023	0.278	0.020	0.027
TN93+G	11	7370.303	7292.697	-3635.333	n/a	0.99	3.87	0.314	0.200	0.264	0.222	0.020	0.027	0.196	0.032	0.183	0.023	0.032	0.139	0.023	0.278	0.020	0.027
HKY+G+I	11	7371.684	7294.078	-3636.023	0.37	200.00	3.83	0.314	0.200	0.264	0.222	0.021	0.027	0.176	0.032	0.210	0.023	0.032	0.159	0.023	0.250	0.021	0.027
GTR	13	7375.954	7284.244	-3629.101	n/a	n/a	3.70	0.314	0.200	0.264	0.222	0.012	0.052	0.195	0.019	0.185	0.025	0.062	0.140	0.004	0.276	0.023	0.005
TN93+G+I	12	7379.360	7294.702	-3635.333	0.26	2.65	3.87	0.314	0.200	0.264	0.222	0.020	0.027	0.196	0.032	0.183	0.023	0.032	0.139	0.023	0.278	0.020	0.027
GTR+I	14	7384.521	7285.759	-3628.855	0.52	n/a	3.94	0.314	0.200	0.264	0.222	0.011	0.052	0.199	0.016	0.184	0.025	0.062	0.140	0.002	0.283	0.023	0.002
GTR+G	14	7384.571	7285.810	-3628.880	n/a	0.39	4.09	0.314	0.200	0.264	0.222	0.009	0.052	0.202	0.015	0.184	0.025	0.063	0.140	0.000	0.287	0.023	0.000
GTR+G+I	15	7395.894	7290.081	-3630.013	0.55	0.81	4.59	0.314	0.200	0.264	0.222	0.000	0.053	0.208	0.000	0.185	0.028	0.063	0.141	0.000	0.295	0.026	0.000
K2	6	7399.031	7356.693	-3672.342	n/a	n/a	3.70	0.250	0.250	0.250	0.250	0.027	0.027	0.197	0.027	0.197	0.027	0.027	0.197	0.027	0.197	0.027	0.027
T92	7	7407.759	7358.367	-3672.177	n/a	n/a	3.70	0.257	0.257	0.243	0.243	0.027	0.026	0.191	0.027	0.191	0.026	0.027	0.203	0.026	0.203	0.027	0.026
K2+I	7	7407.822	7358.430	-3672.208	0.43	n/a	3.86	0.250	0.250	0.250	0.250	0.026	0.026	0.199	0.026	0.199	0.026	0.026	0.199	0.026	0.199	0.026	0.026
K2+G	7	7407.838	7358.445	-3672.216	n/a	1.02	3.85	0.250	0.250	0.250	0.250	0.026	0.026	0.198	0.026	0.198	0.026	0.026	0.198	0.026	0.198	0.026	0.026
T92+I	8	7416.552	7360.105	-3672.044	0.43	n/a	3.86	0.257	0.257	0.243	0.243	0.026	0.025	0.193	0.026	0.193	0.025	0.026	0.204	0.025	0.204	0.026	0.025
T92+G	8	7416.567	7360.121	-3672.052	n/a	1.03	3.85	0.257	0.257	0.243	0.243	0.027	0.025	0.193	0.027	0.193	0.025	0.027	0.204	0.025	0.204	0.027	0.025
K2+G+I	8	7416.880	7360.434	-3672.209	0.43	200.00	3.86	0.250	0.250	0.250	0.250	0.026	0.026	0.199	0.026	0.199	0.026	0.026	0.199	0.026	0.199	0.026	0.026
T92+G+I	9	7425.610	7362.110	-3672.044	0.43	200.00	3.86	0.257	0.257	0.243	0.243	0.026	0.025	0.193	0.026	0.193	0.025	0.026	0.204	0.025	0.204	0.026	0.025
JC	5	7511.431	7476.149	-3733.071	n/a	n/a	0.50	0.250	0.250	0.250	0.250	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083
JC+I	6	7520.377	7478.040	-3733.015	0.29	n/a	0.50	0.250	0.250	0.250	0.250	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083
JC+G	6	7520.378	7478.041	-3733.015	n/a	1.94	0.50	0.250	0.250	0.250	0.250	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083



FEEDBACK - MENTI

15:50-15:55

Go to

www.menti.com

Enter the code

5167 8855

Or use QR code

NEXT WEEK

15:55-16:00

We will discuss the Zoonomia paper on Wed and work on the first hand-in on Fri

AARHUS
UNIVERSITY