SUBSTITUTION MODELS

SHORT RECAP + MODELTEST PAPER DISCUSSION

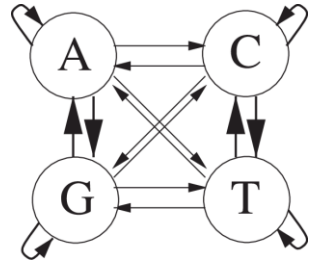
AARHUS
UNIVERSITY
DEPARTMENT OF MOLECULAR BIOLOGY AND GENETICS

EVOLUTIONARY THINKING 2023
WEEK 36

CALIN PANTEA
PHD STUDENT

EVENT MODELLING

- Basic nucleotide substitution models
- Appropriateness
- Gamma distribution
- Model selection (AIC and BIC)



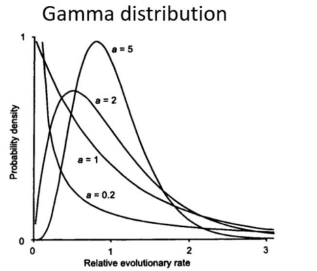
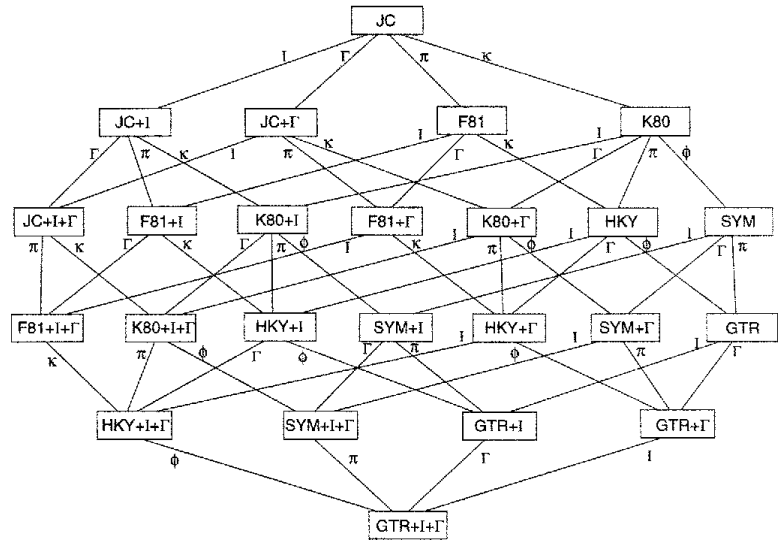


FIGURE 2.3. Gamma distributions of substitution rates among sites for different gamma parameters (α's).

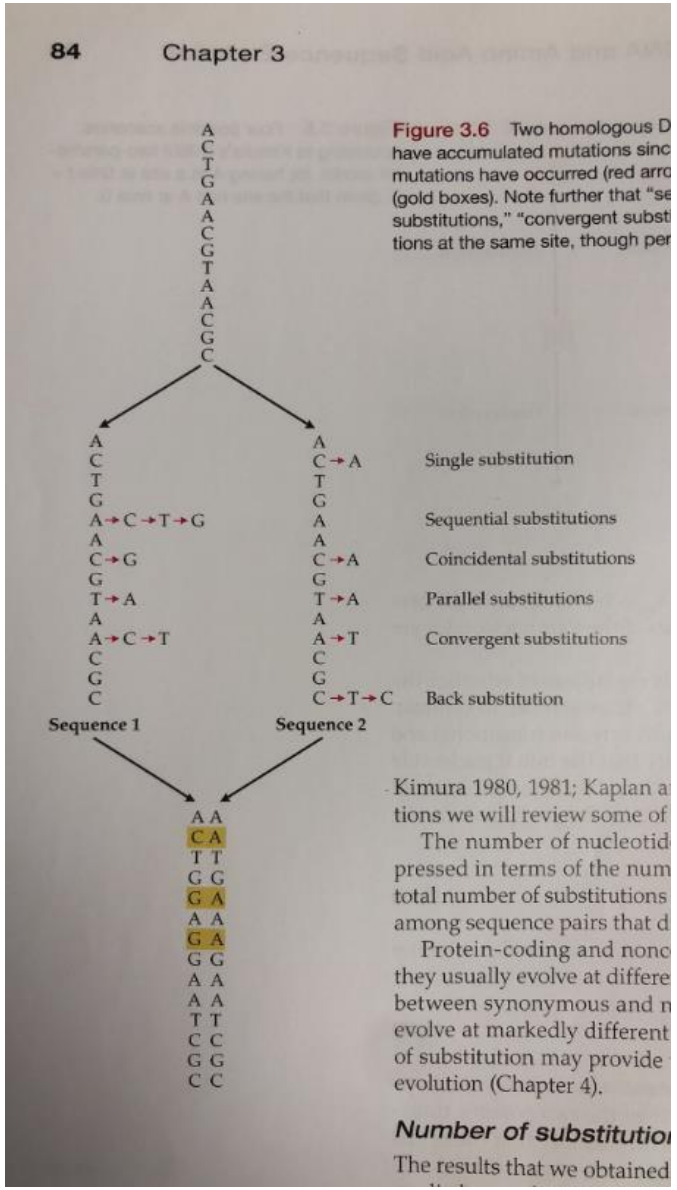


NUCLEOTIDE SUBSTITUTION

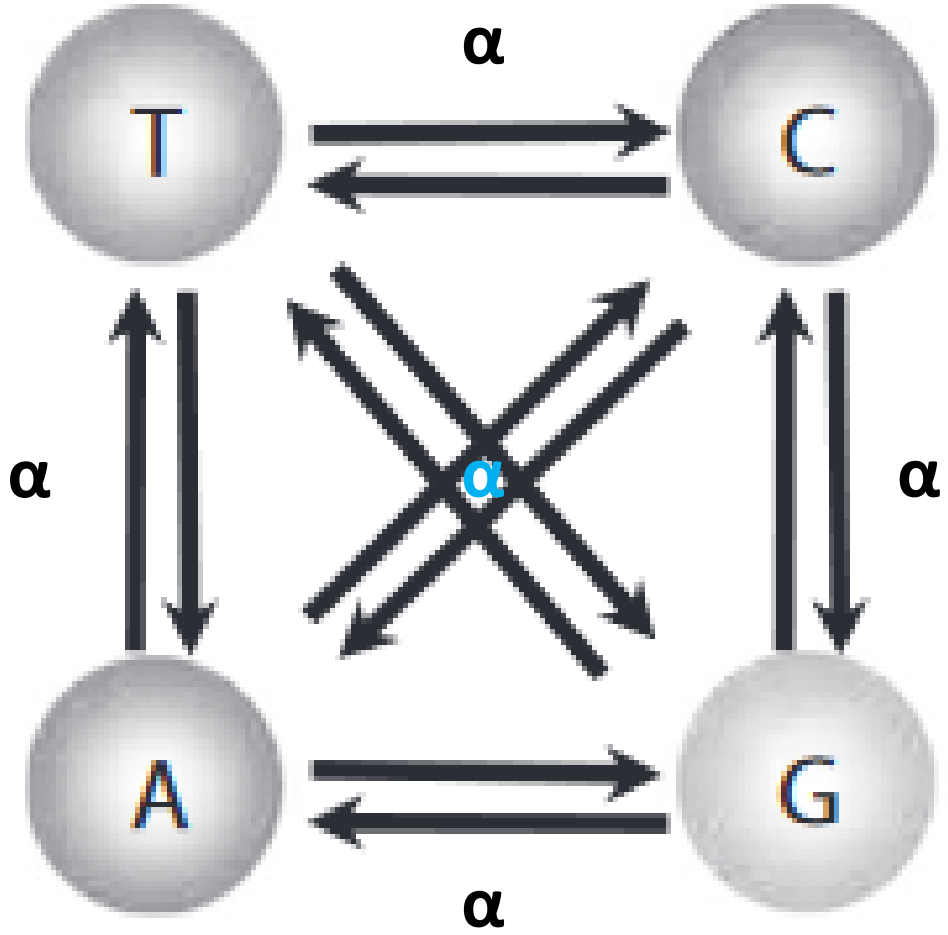
- divergent sequences

AGATGCTAGCATCGACTAGCATCAGCTGACCCCGCGCGCGCAT

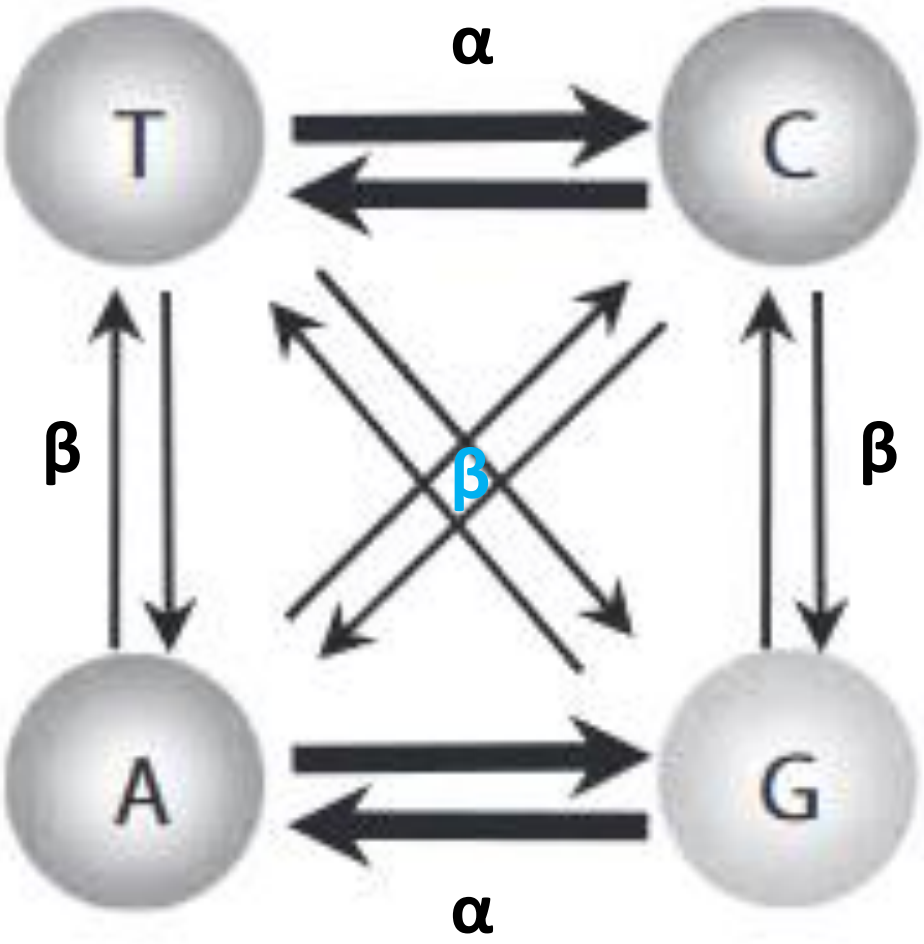
AGATAATAGCATCGACTAGCATCAGCGGACCCCGCGCGCGCAC



JUKES-CANTOR (JC69)



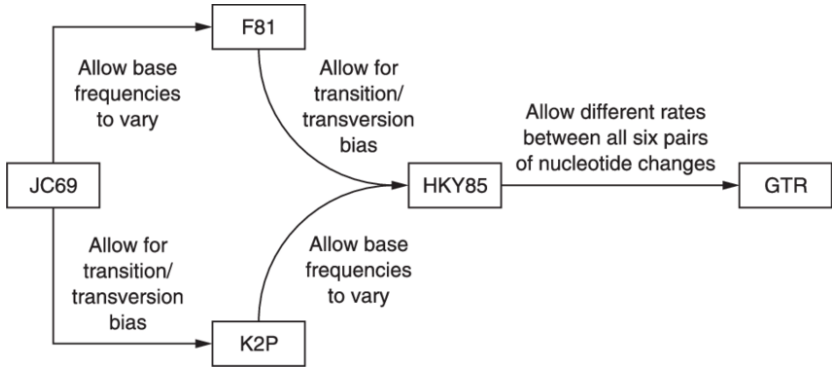
KIMURA (K2P)



OTHER MODELS

Which model is the best?

- Simple models
  - Few parameters (LOW sampling variance)
  - Perform poorly for large distance
- Complex models
  - Many parameters (LARGE sampling variance)
  - Better behaved for large divergence
  - Require LOTS of data
- What is the true patterns of substitution ?



Gamma distribution

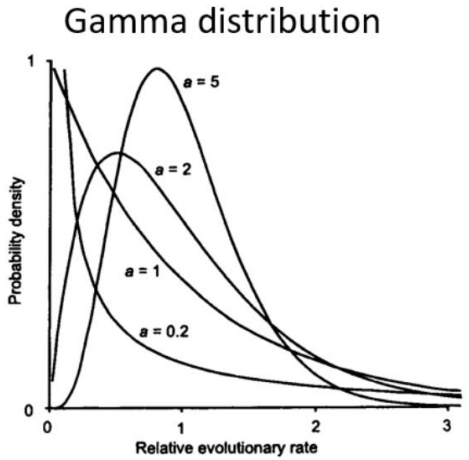
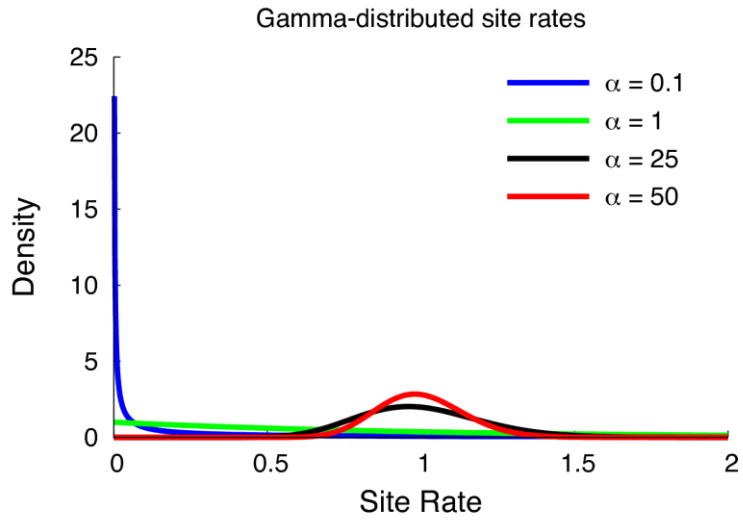


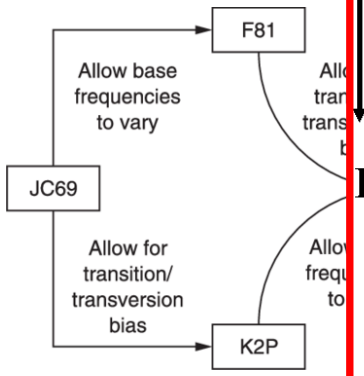
FIGURE 2.3. Gamma distributions of substitution rates among sites for different gamma parameters (*a*'s).



OTHER

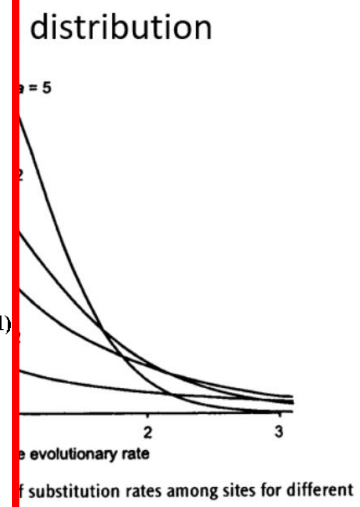
Which

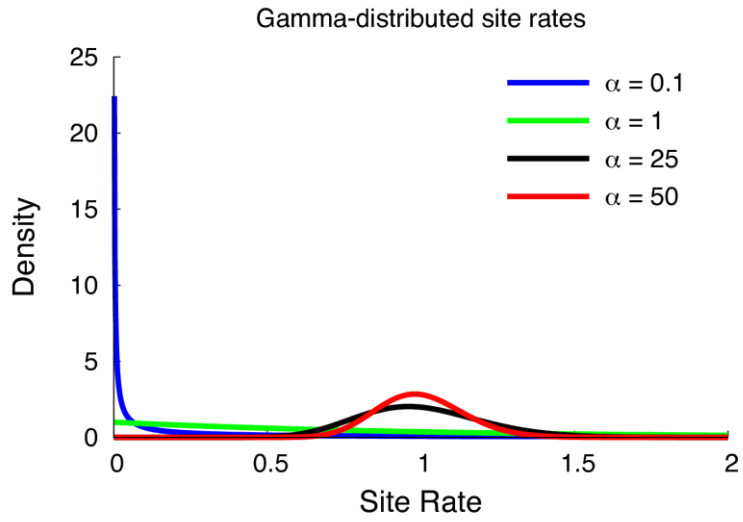
- Simple models
  - Few paramete
  - Perform poorl
- Complex mode
  - Many parame
  - Better behave
  - Require LOTS
- What is the tru



Model	Free parameters in the Q-matrix
GTR	**8** $a, b, c, d, e, \pi_A, \pi_C, \pi_G,$ $(a+b+c+d+e+f=1 \quad \pi_A+\pi_C+\pi_G+\pi_T=1)$
TN93	**5** $a=c=d=f, b=e, \pi_A, \pi_C, \pi_G,$ *(Ti/Tv ratio, Y/R ratio and and three frequencies since* $\pi_A+\pi_C+\pi_G+\pi_T=1$*)*
HKY85 → **F84**	**4** $a=b=c=d=1, \quad b=e=k, \pi_A, \pi_C, \pi_G,$ *(Ti/Tv ratio and only three frequencies since* $\pi_A+\pi_C+\pi_G+\pi_T=1$*)*
F81	**3** $a=c=d=f=b=e=f, \pi_A, \pi_C, \pi_G,$ *(only three nt frequencies since* $\pi_A+\pi_C+\pi_G+\pi_T=1$*)*
K80	**1** $a=b=c=d=1, \quad b=e=k, \quad \pi_A=\pi_C=\pi_G=\pi_T$ *(Ti/Tv ratio)*
JC69	**0** $a=b=c=d=e=f, \quad \pi_A=\pi_C=\pi_G=\pi_T$

Gamma distribution



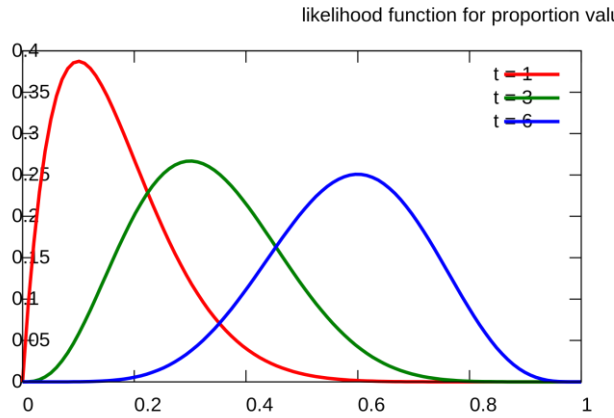


LIKELIHOOD FUNCTION AND MLE

Likelihood function

$\mathcal{L}(\theta \mid x)$

- describes the distribution of probabilities that a model fitted to data describes it well



MLE (maximum likelihood estimation) – maximum value of the likelihood function for model describing studied data

- basically – highest possible probability for the fitted model to really described the data

-> but more complex models are more prone to overfitting!

AIC AND BIC

k = number of model parametres; n = number of observations

Bayesian information criterion (BIC)

$$\mathrm{BIC} = k\ln(n) - 2\ln(\widehat{L})$$

- depends on sample size ‘n’
- can only be reliable as long as n >> k

Akaike information criterion (AIC)

$$\mathrm{AIC} = 2k - 2\ln(\hat{L})$$

- somewhat more permissive
- sometimes more reliable
- better choice than BIC for n!>>k

MODELTEST PAPER

- Introduces an early solution for substitution model choice: simple model likelihood pairwise comparison
- Old and plain; but still a solid early paper discussion resource
- Why?
  - Practice paper discussion
  - Insight into scientific thinking
  - Useful for exam mindset <and> future research work

Course week	Week number	Lecture Monday	Readings	TA Wednesday	TA Friday	Hand-ins
1	35	Origin of life substitution models	Yang, pp 1-23 and originoflife paper	Tree pretest exercise (pen and paper) Installing MEGA11	MEGA11 exercises: retrieve sequences, alignment and distance matrix	
2	36	Phylogenetic tree reconstruction	Yang pp 23-50, graur 165-178	Paper discussion: Modeltest paper, Yang problems	MEGA11 exercises: substitution models	
3	37	Trees continued plus Molecular clocks	Graur 165-209	Paper discussion: Zoonomia paper	Work on Hand in I	Hand-in I: Phylogenetic tree analysis
4	38	Student prestentations and work on hand in I	Yang 361-368	Paper discussion: HIV infection on a dental clinic	**Introduction to R using phylogenetics**	Hand-in I deadline
5	39	Basics of population genetics	NS 2,3, Charlesworth	Problems NS 1,2,3 Hand-in I correction	Wright-Fisher and Coalescent simulator exercises (R)	
6	40	The coalescent continued	NS 4,5,6, Ellegren paper	Problems NS 4,5,6	Problems NS 4,5,6	
7	41	Tests of neutrality	NS 7,8,9, Vicoso paper	Problems NS 7,8	**R exercise: detecting positive selection using dn/ds. Jeopardy**	Hand-in II: NORK
-	42	Autumn Break		Autumn Break	Autumn Break	Autumn Break
8	43	Incomplete lineage sorting	Mailund et al.	Paper discussion:Iker Rivas Gonzalez et al	Hand-in II correction	Hand-in II deadline
9	44	Comparative genomics and X chromosomes	Review	Paper discussion:	**Chatgbt driven phylogenetics exercise in R**	
10	45	Genome scans and primate genomics,	Marques-Bonet et al and Kuderna et al	Paper discussion: Gao et al.	**R exercise**	
11	46	Ancient DNA	Rasmussen et al.	Paper discussion: Green et al.	**Update: MEGA11 exercise with data**	
12	47	Genome evolution, mutation		Paper discussion: Bergeron et al	Ancient DNA exercise Work on Hand-in III	Hand-in III: Ancient DNA
13	48	Genome evolution, TE		Paper discussion: Hallast et al. Self-test	Self-test	Hand-in III deadline
14	49	Evaluation	none	Hand-in III correction/discussion	Summary of the exercises/discussion	

PAPER QUESTIONS

Group 1

Last Name, First Name, Username, Org Defined ID
Barenys, Felix, au752670, 752670
Jørgensen, Jonas Riber, au683396, 683396
Martensen, René, au684546, 684546
Petersen, Simon, au649982, 649982

Group 3

Last Name, First Name, Username, Org Defined ID
Dahl, Astrid Klitgaard, au612796, 612796
Grand, Christina Espeseth, au682054, 682054
Hansen, Johanne Gottenborg Høyer, au671624, 671624
Vargas Fichera, Anna Gabriela, au752667, 752667

Group 5

Last Name, First Name, Username, Org Defined ID
Fledelius, William, au682687, 682687
Saha, Anik, au746885, 746885
Thingholm, Malthe Skovmand, au599620, 599620

Group 7

Last Name, First Name, Username, Org Defined ID
Ezekiel, Blessing, au709133, 709133
Grandpierre, Nimród, au752683, 752683
Jensen, Sebastian Kjellerup Godske, au664538, 664538
Jørgensen, Søren, au643843, 643843

Group 9

Last Name, First Name, Username, Org Defined ID
DE SILVA, HENDA HEWA IRESHA MIHIRA., au746824, 746824
Holt, Karl Immanuel, au724387, 724387
Namsaraeva, Altana, au746258, 746258
Skov, Anna Kathrine Jensen, au646861, 646861

Group 11

Last Name, First Name, Username, Org Defined ID
Bader, Liliane Zoe, au752655, 752655
Calandra, Elías, au752669, 752669
Charlton, Chester Henry, au747556, 747556
Danger, Stinna Weisberg, au643227, 643227

1-6

7-11

Group 2

Last Name, First Name, Username, Org Defined ID
Chatzivasileiou, Savvas, au752684, 752684
Chen, Chongming, au746597, 746597
Kongsgaard, Viktor, au661991, 661991
von Vegesack, Hans Carlo, au752686, 752686

Group 4

Last Name, First Name, Username, Org Defined ID
Groth, Thorbjørn Sejr, au665701, 665701
Jensen, Ane Naur, au672628, 672628
Sucipto, Abby Kurniawati, au747615, 747615
Wahab, Adam Faraldon, au667387, 667387

Group 6

Last Name, First Name, Username, Org Defined ID
Andersen, Søren Villum Poulsgaard, au635621, 635621
Coto Escofet, Alba, au731539, 731539
Guo, Jiahui, au747088, 747088
Lütken, Gorm Hjorth, au681865, 681865

Group 8

Last Name, First Name, Username, Org Defined ID
Dam, Peter Konrad Alling, au610627, 610627
Martín Pestaña, David, au752685, 752685
Nikolaidou, Eleni, au752666, 752666
Svenningsen, Laurids, au632957, 632957

Group 10

Last Name, First Name, Username, Org Defined ID
Fur, Laura Marie Hagen, au640803, 640803
Gálfalvy, Nóra, au752687, 752687
Jørgensen, Carolina Ferreira, au590004, 590004
Ulstrup, Johan Christensen, au664823, 664823

YOUR QUESTIONS ABOUT PAPER QUESTIONS

Group 1

Last Name, First Name, Username, Org Defined ID
Barenys, Felix, au752670, 752670
Jørgensen, Jonas Riber, au683396, 683396
Martensen, René, au684546, 684546
Petersen, Simon, au649982, 649982

Group 3

Last Name, First Name, Username, Org Defined ID
Dahl, Astrid Klitgaard, au612796, 612796
Grand, Christina Espeseth, au682054, 682054
Hansen, Johanne Gottenborg Høyer, au671624, 671624
Vargas Fichera, Anna Gabriela, au752667, 752667

Group 5

Last Name, First Name, Username, Org Defined ID
Fledelius, William, au682687, 682687
Saha, Anik, au746885, 746885
Thingholm, Malthe Skovmand, au599620, 599620

Group 7

Last Name, First Name, Username, Org Defined ID
Ezekiel, Blessing, au709133, 709133
Grandpierre, Nimród, au752683, 752683
Jensen, Sebastian Kjellerup Godske, au664538, 664538
Jørgensen, Søren, au643843, 643843

Group 9

Last Name, First Name, Username, Org Defined ID
DE SILVA, HENDA HEWA IRESHA MIHIRA., au746824, 746824
Holt, Karl Immanuel, au724387, 724387
Namsaraeva, Altana, au746258, 746258
Skov, Anna Kathrine Jensen, au646861, 646861

Group 11

Last Name, First Name, Username, Org Defined ID
Bader, Liliane Zoe, au752655, 752655
Calandra, Elías, au752669, 752669
Charlton, Chester Henry, au747556, 747556
Danger, Stinna Weisberg, au643227, 643227

1-6

7-11

Group 2

Last Name, First Name, Username, Org Defined ID
Chatzivasileiou, Savvas, au752684, 752684
Chen, Chongming, au746597, 746597
Kongsgaard, Viktor, au661991, 661991
von Vegesack, Hans Carlo, au752686, 752686

Group 4

Last Name, First Name, Username, Org Defined ID
Groth, Thorbjørn Sejr, au665701, 665701
Jensen, Ane Naur, au672628, 672628
Sucipto, Abby Kurniawati, au747615, 747615
Wahab, Adam Faraldon, au667387, 667387

Group 6

Last Name, First Name, Username, Org Defined ID
Andersen, Søren Villum Poulsgaard, au635621, 635621
Coto Escofet, Alba, au731539, 731539
Guo, Jiahui, au747088, 747088
Lütken, Gorm Hjorth, au681865, 681865

Group 8

Last Name, First Name, Username, Org Defined ID
Dam, Peter Konrad Alling, au610627, 610627
Martín Pestaña, David, au752685, 752685
Nikolaidou, Eleni, au752666, 752666
Svenningsen, Laurids, au632957, 632957

Group 10

Last Name, First Name, Username, Org Defined ID
Fur, Laura Marie Hagen, au640803, 640803
Gálfalvy, Nóra, au752687, 752687
Jørgensen, Carolina Ferreira, au590004, 590004
Ulstrup, Johan Christensen, au664823, 664823

QUESTIONS ABOUT PAPER QUESTIONS

Google Docs

QUIZ TIME!

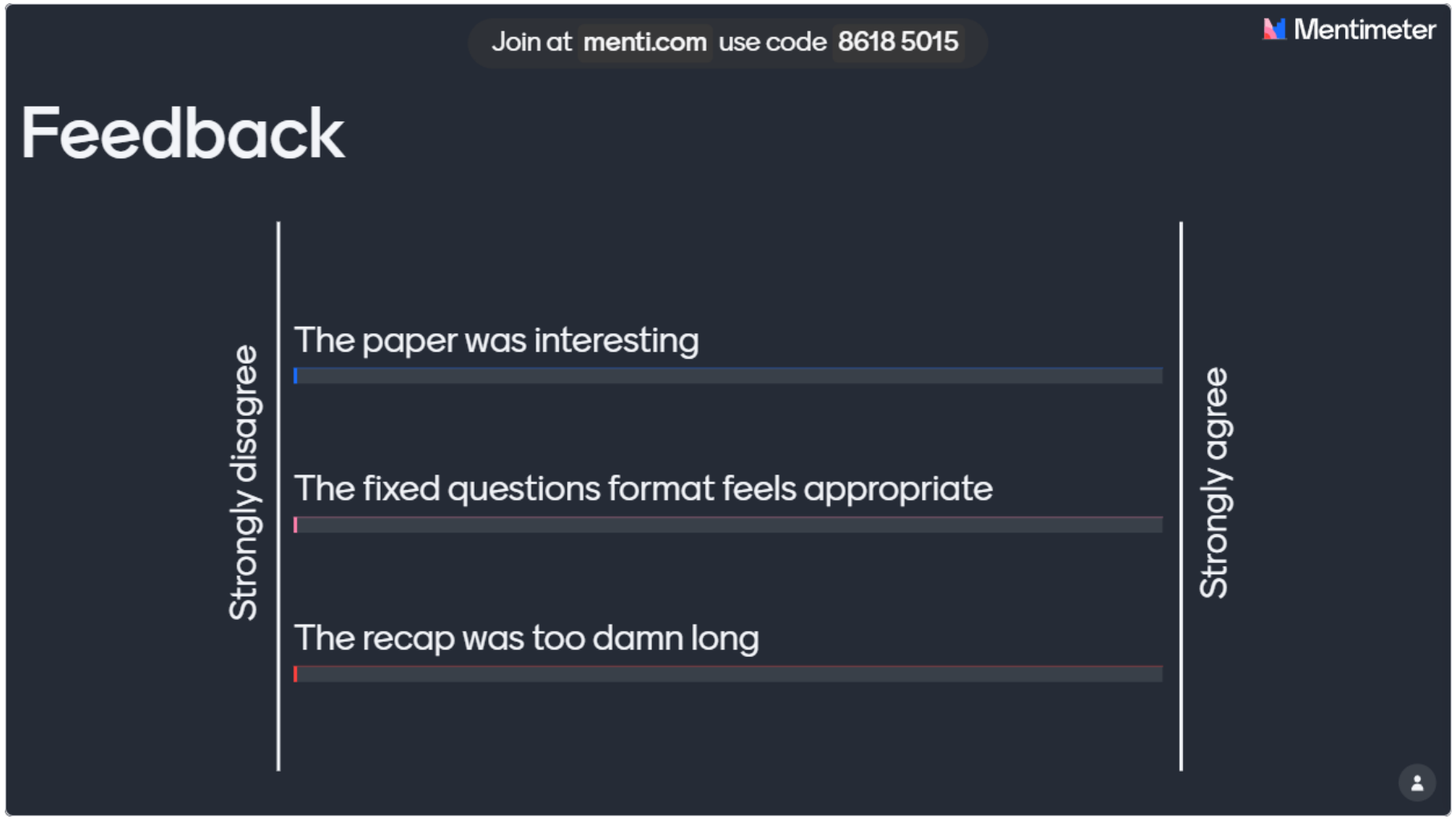
questions

and

feedback

GO TO
menti.com

ENTER THE CODE
8618 5015



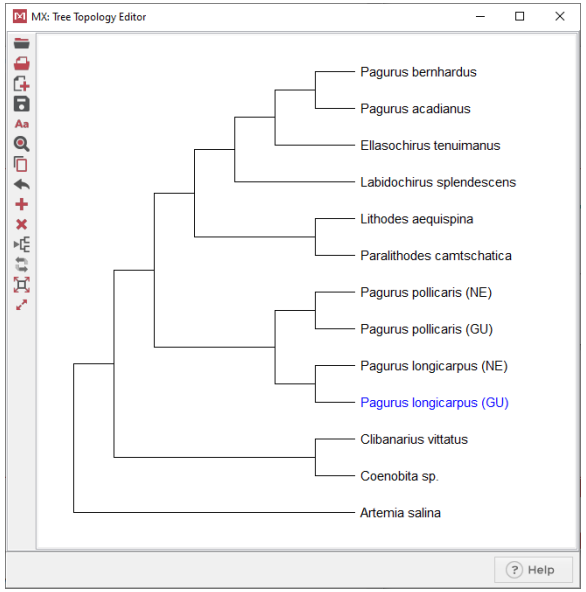
POTENTIAL LECTURE CHANGES

Monday 18th Sept (Week 38) – No lecture (Mikkel is away)

Possibly moved the week before, on Friday the 15th Sept!

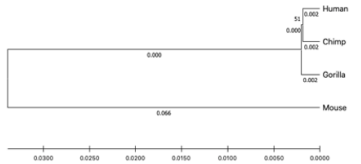
SEE YOU ON FRIDAY FOR SOME MORE DELIGHTFUL MEGA SHENANIGANS

MEGA tree building!

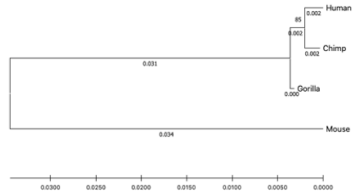


FOXP2

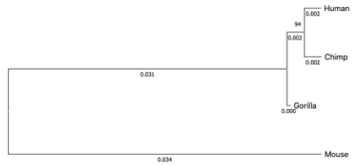
UPGMA



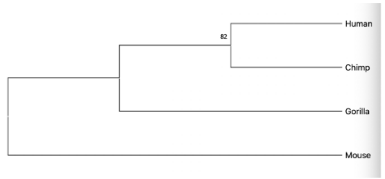
NJ



ML



MP



AARHUS
UNIVERSITY