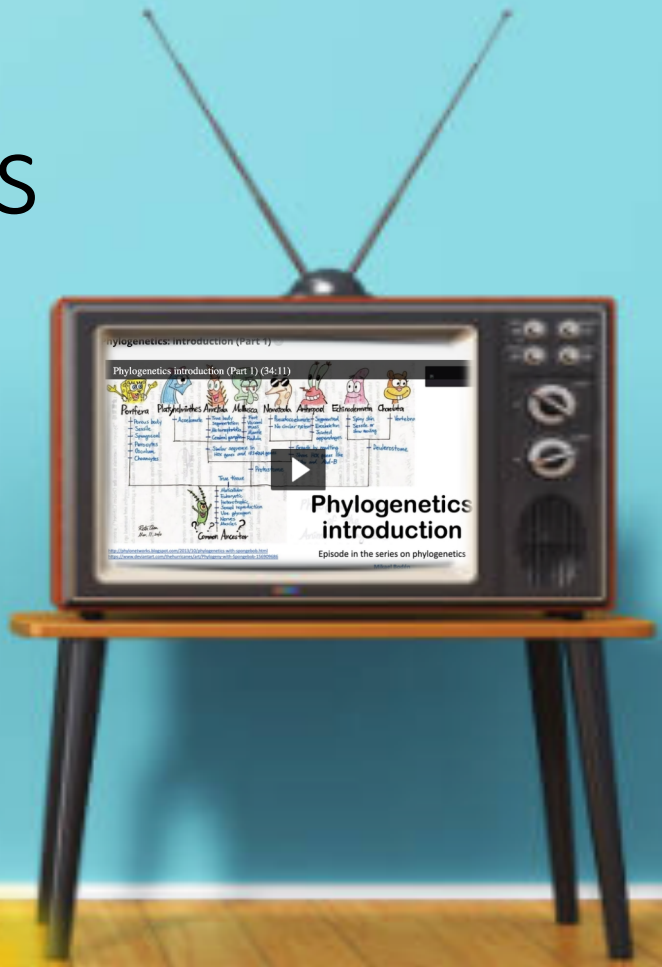
Phylogenetics 1: week 7

Watch the recordings

Phylogenetics: introduction

Phylogenetics: quantifying evolution



Mikael Bodén

About me…

Mikael Bodén

PhD in Computer Science (Exeter/UK)
m.boden@uq.edu.au

that guy

Associate Professor
School of Chemistry and Molecular Biosciences
The University of Queensland

Research Group Leader Bioinformatics
http://bioinf.scmb.uq.edu.au

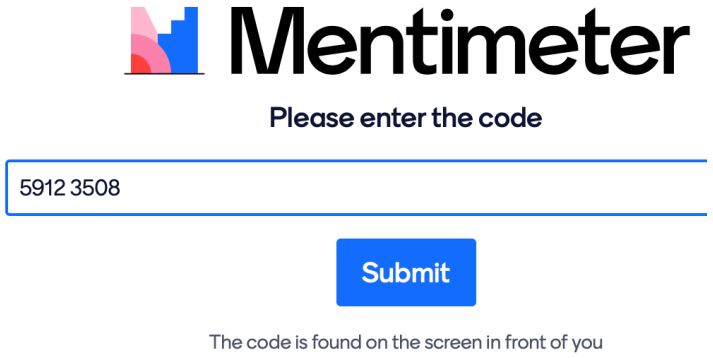
Program Director Bioinformatics

Course co-ordinator
SCIE3100/BINF7000
Bioinformatics 2: Development and Research
formerly
BIOL3014
Advanced Bioinformatics

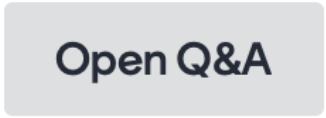
We're using **Mentimeter**

- Go here on your phone/laptop/tablet

https://menti.com



- Respond to question by …
  - typing a phrase, *or*
  - multiple phrases on separate rows
- Q&A turned on



- I'll stop at a few points to check Zoom chat also

Answers are not saved or used to identify you

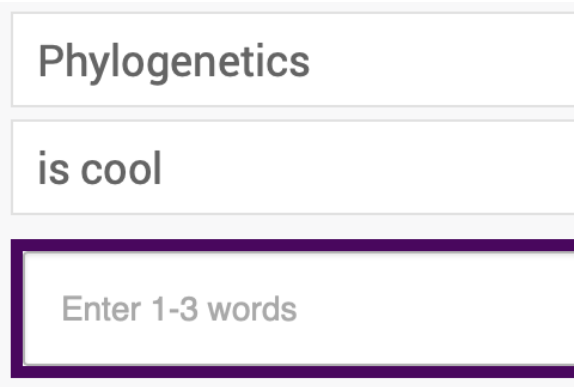
Or … **UQwordcloud**

- Go here on your phone/laptop/tablet

https://apps.elearning.uq.edu.au/wordcloud/48709

- Refresh browser when new question appears

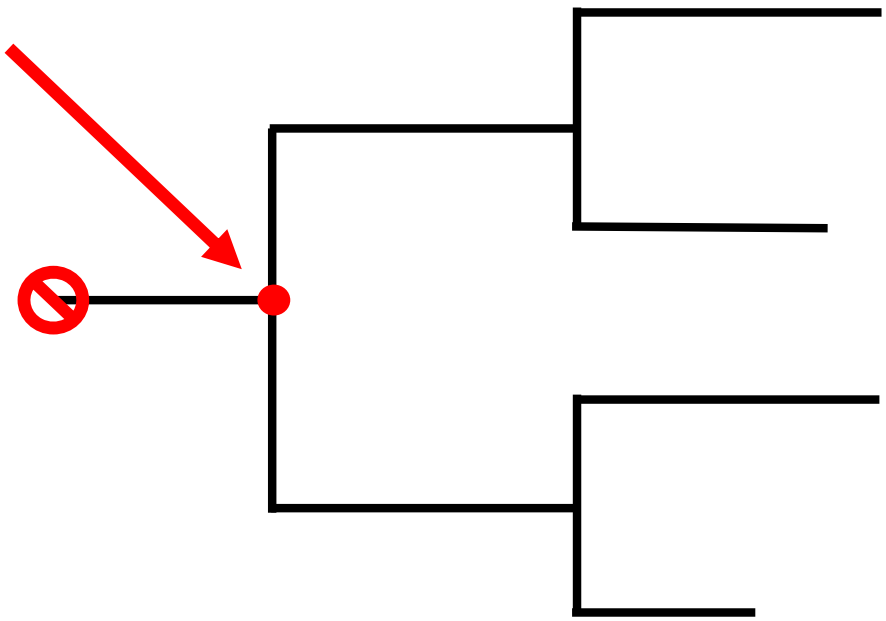
- Respond to question by …
  - typing a phrase, *or*
  - multiple phrases on separate rows



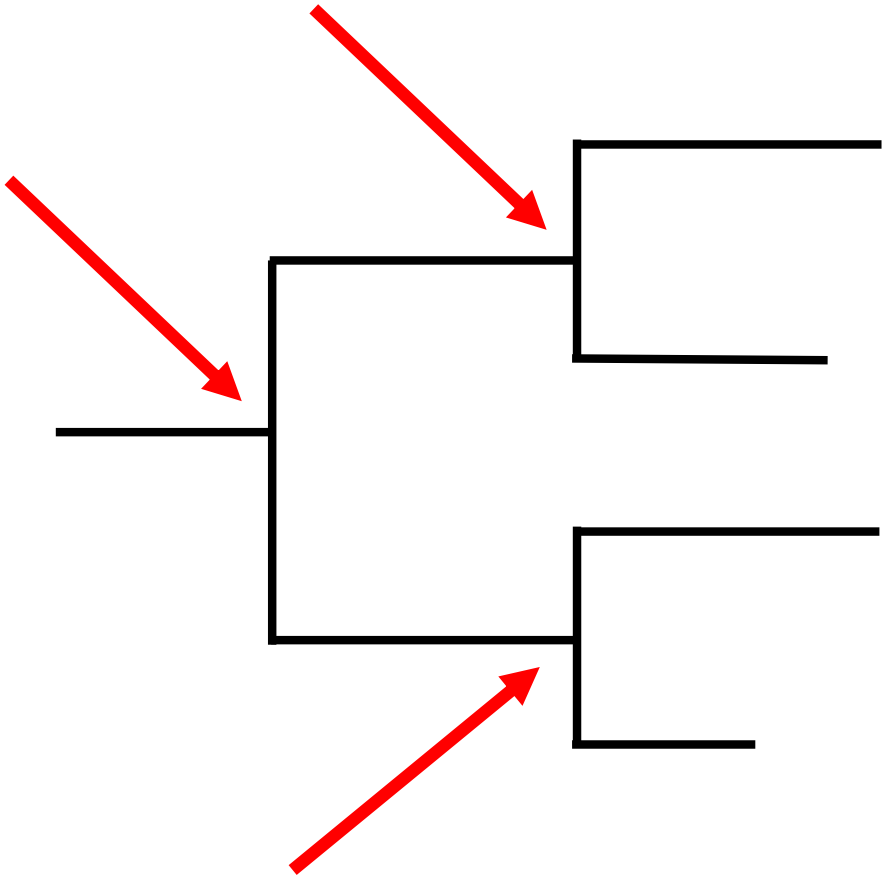
- I'll stop at a few points to check Zoom

Answers are not saved or used to identify you

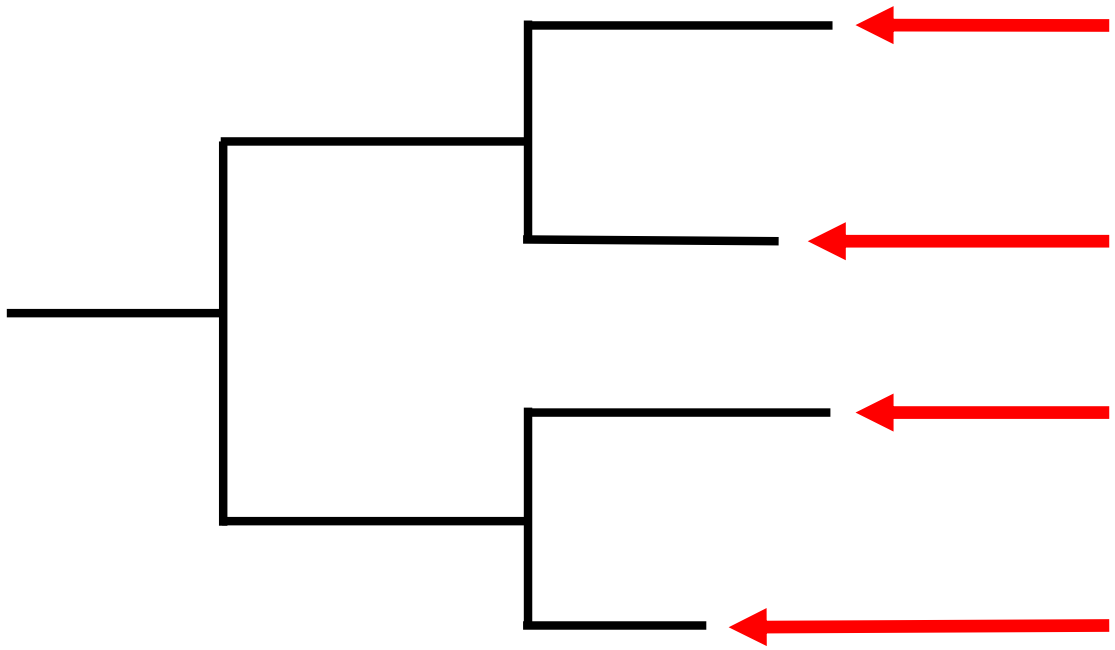
P1: What name would you use to refer to …



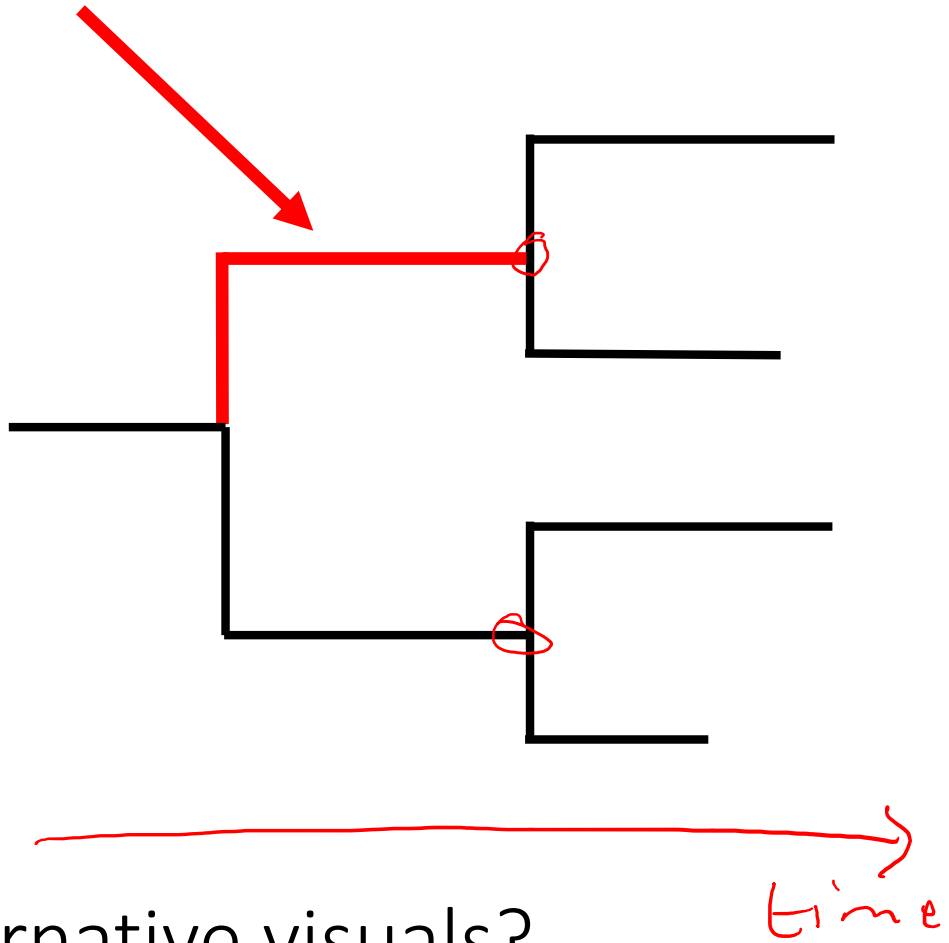
P2: What name would you use to refer to …



P3: What name would you use to refer to …

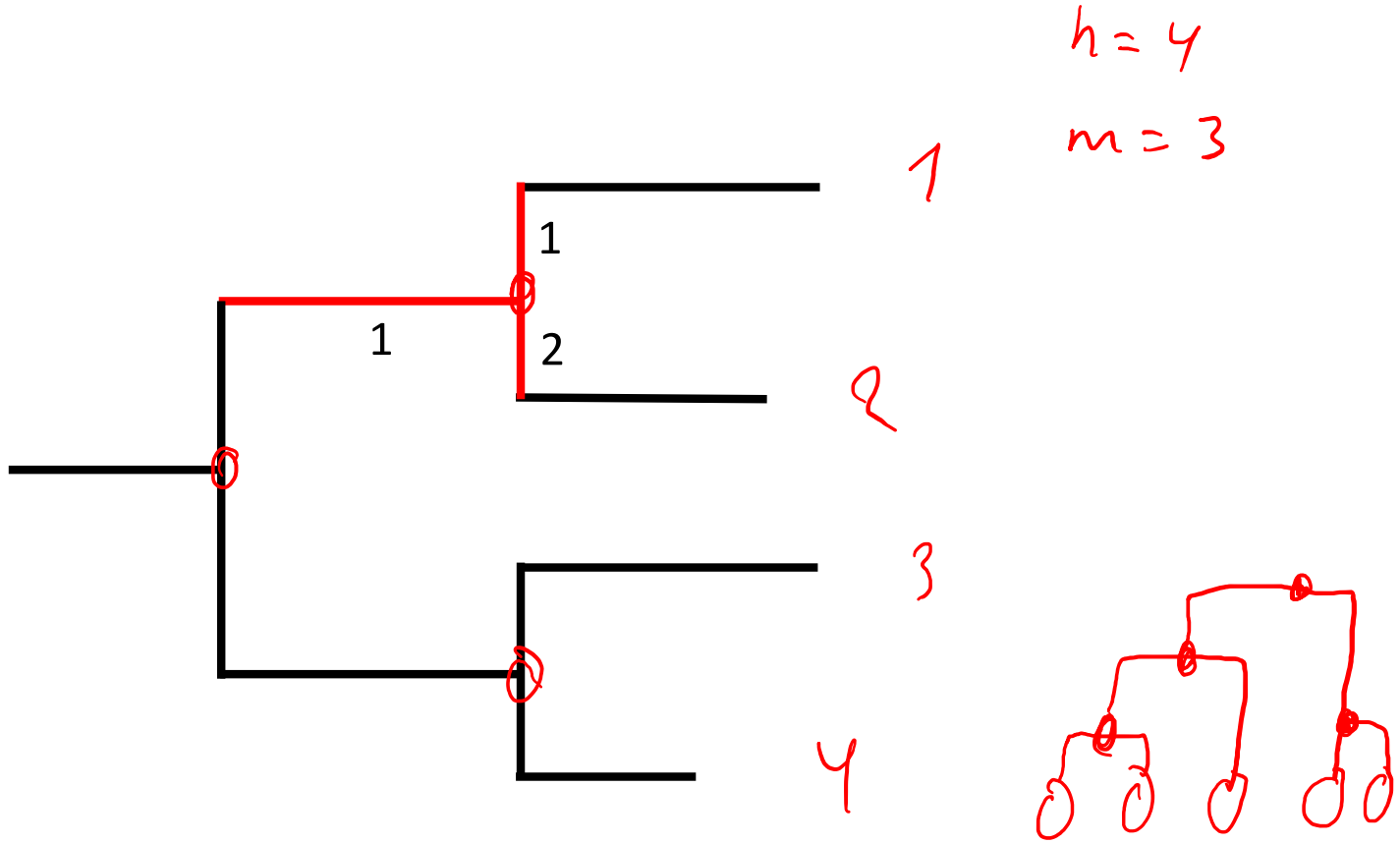


P4: What name would you use to refer to …



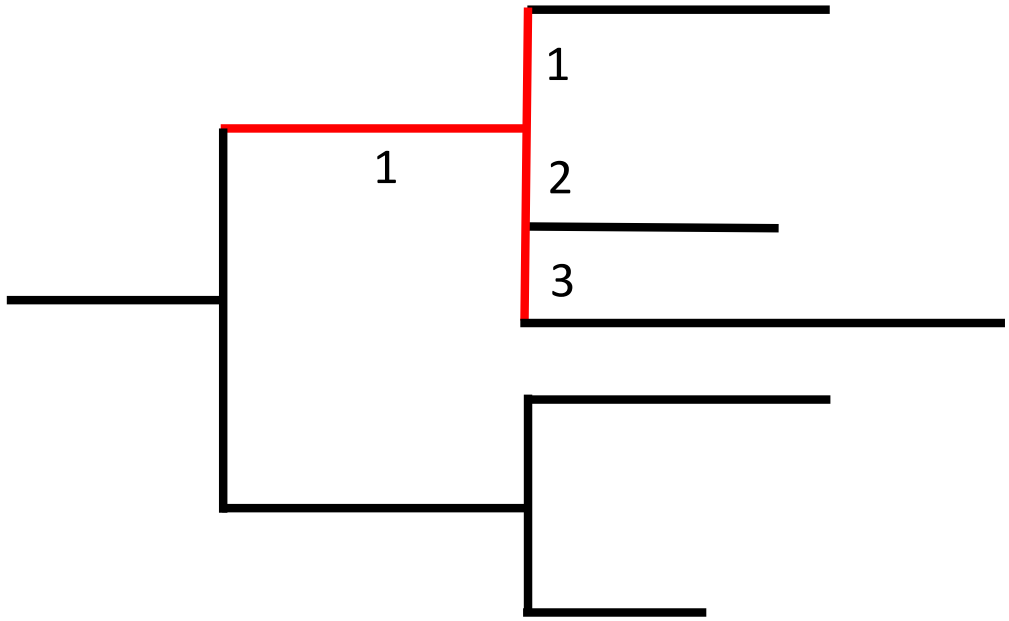
Discuss: alternative visuals?

P5: What is the term to use for a tree when *each* branchpoint has one ingoing and two outgoing branches?

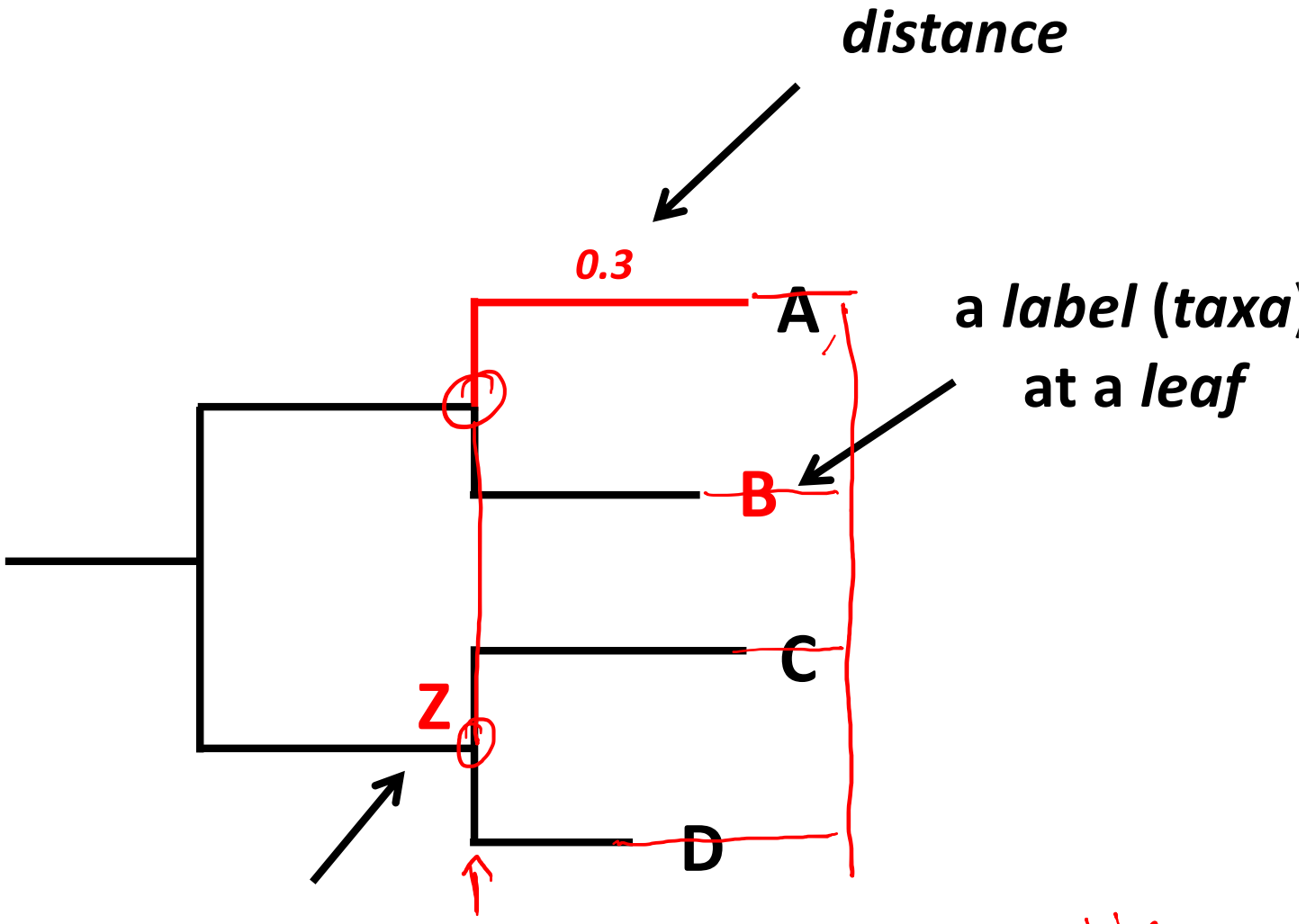


Discuss: number of internal nodes depends on number of leaves, how?

P6: What is the term to use for a tree when *each* branchpoint has one ingoing and many outgoing branches?



Discuss: non-binary evolutionary events?



Discuss: metrics of evolution recover "distances", implication for branch lengths? Types of trees?

P7: Supp Exam question from 2019

To use molecular data to reconstruct evolutionary history requires making a number of reasonable assumptions. Which of the following statements are INCORRECT? Several statements could apply.

A The molecular sequences used in phylogenetic construction are homologous.

B The molecular sequences used in phylogenetic construction share a common origin.

C Phylogenetic divergence cannot be bifurcating.

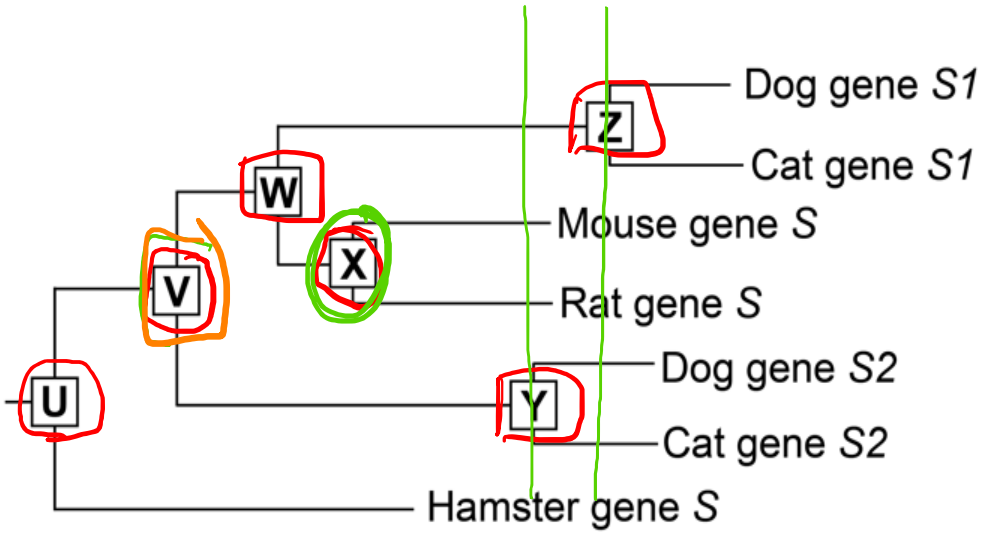
D Parent branch splits into two or more daughter branches at any given point.

E The molecular sequences used in phylogenetic construction cannot be paralogous.

Write each letter on a separate row (or select option if “Choose option/s”)

P8: Example exam question

The phylogenetic tree below shows the evolutionary relationship of gene *S* in dog, cat, and rodents, rooted with the hamster gene *S* as outgroup. Genes *S1* and *S2* in the same organism are paralogs.



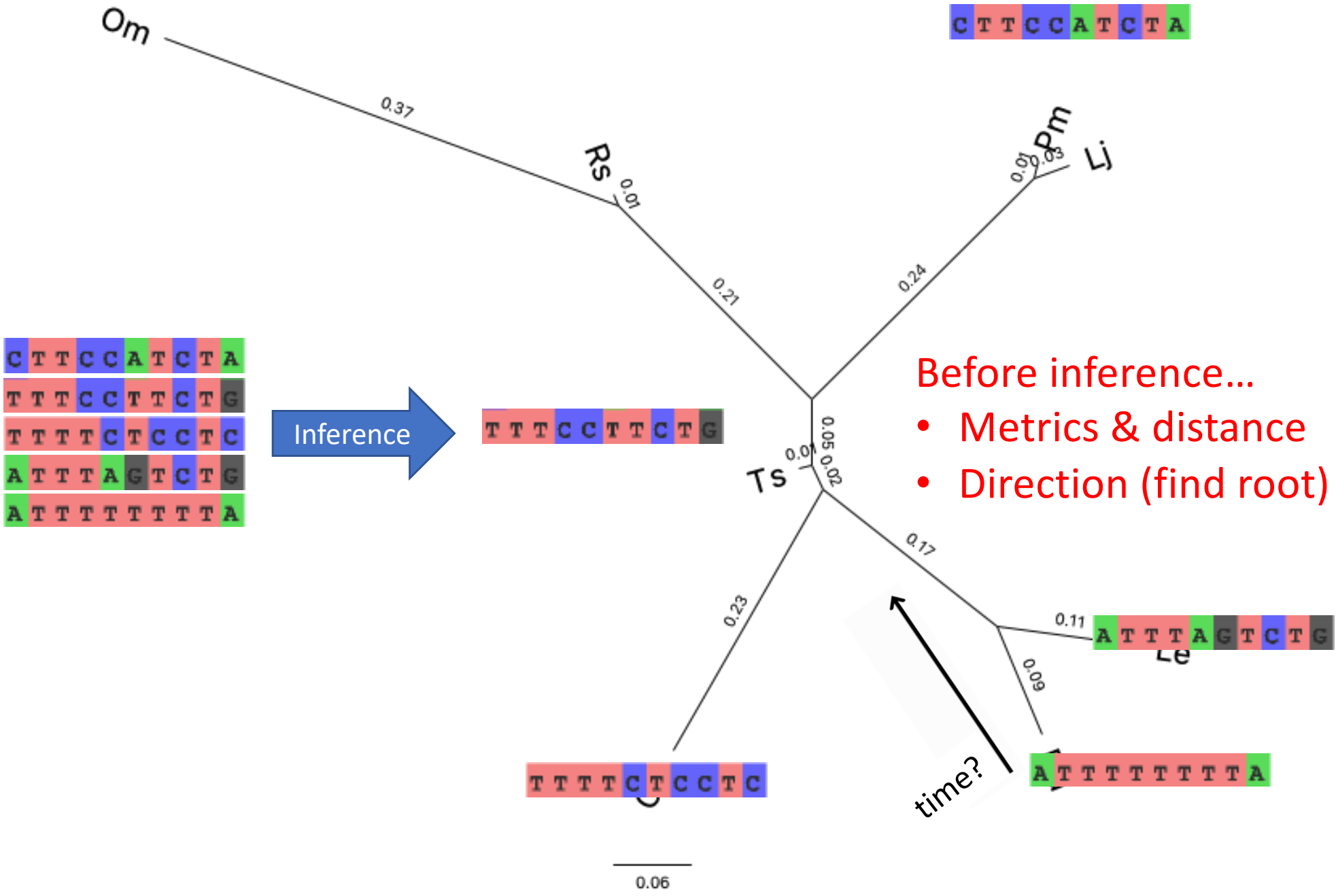
Which of the six nodes (U, V, W, X, Y or Z) describe the following two events?

1. speciation event between Rat and Mouse
2. duplication event between Dog gene *S1* and Dog gene *S2*

Write each letter on a separate row pre-fixed with number, e.g. 1U

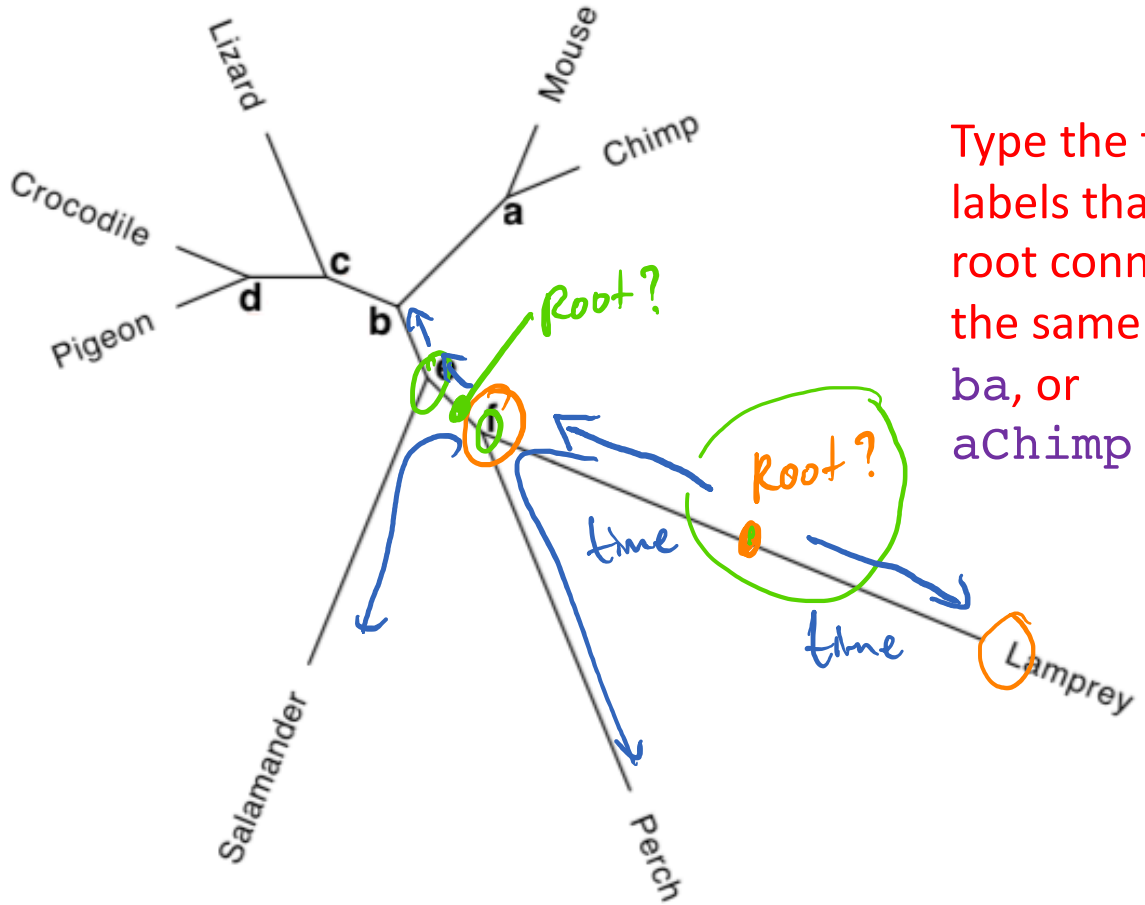
Questions so far?

About trees, homology, etc. in particular


Om
0.37
Rs
0.01
0.21
Pm
Lj
0.01
0.03
0.24
Ts
0.01
0.05
0.02
0.17
0.23
0.11
0.09
Le
time?
0.06
Inference
Before inference…
• Metrics & distance
• Direction (find root)

P8: Exam question from 2019

The unrooted phylogenetic tree below was inferred from eight orthologous sequences representing different species. Each internal branch point is labelled and the leaves are labelled with the species names.

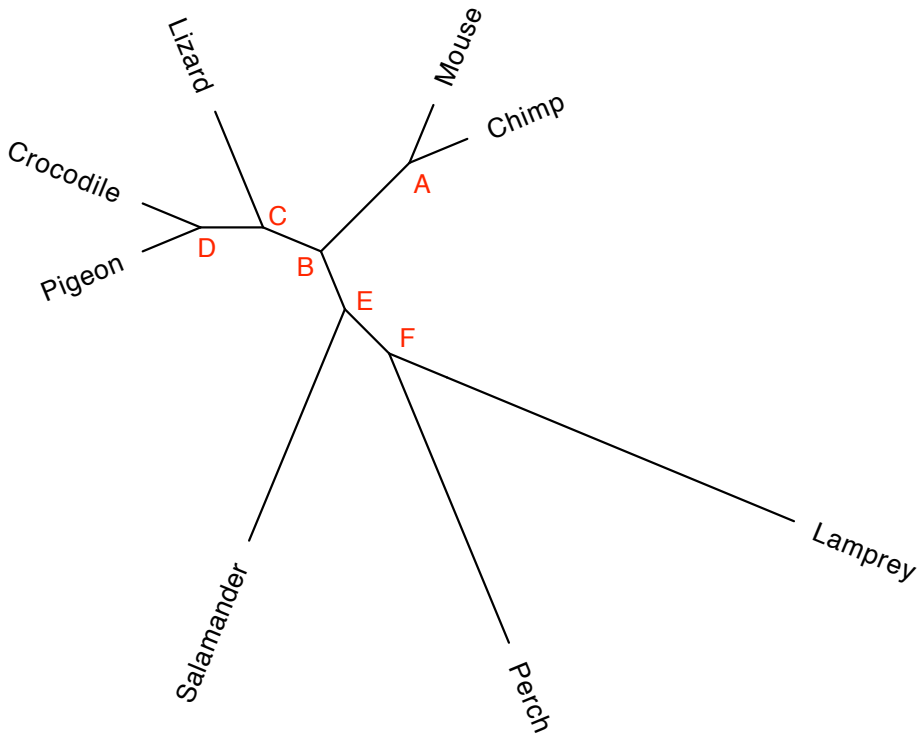


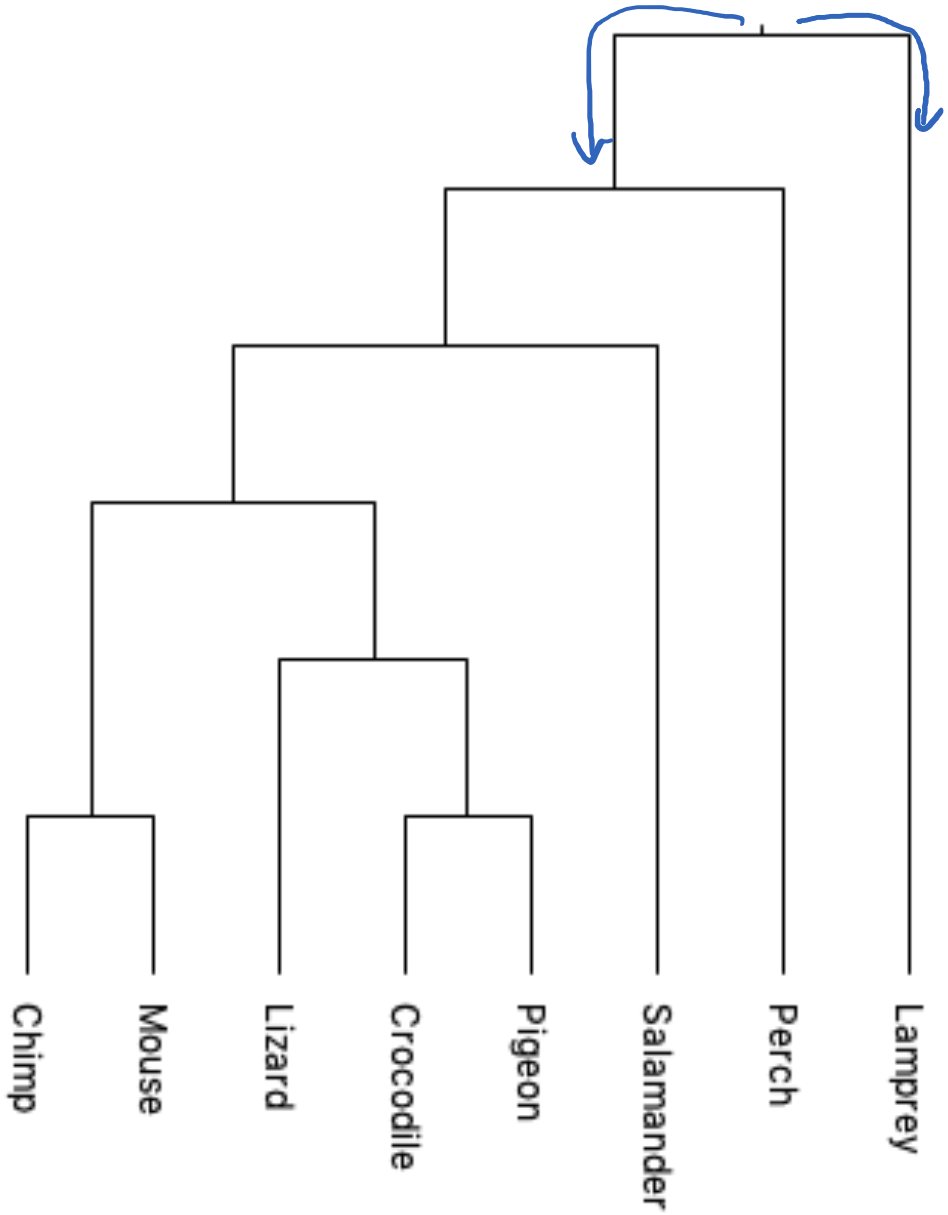
Type the two node labels that the root connects on the same row, e.g. `ba`, or `aChimp`

Based on this tree, answer the following questions.

A. To root the tree with lamprey as an outgroup, on which branch should the root be placed? A branch could be from a to b, or from a to Chimp, etc. (2 marks)

B. Draw the rooted tree in A. (2 marks)





P9: Example exam question

D. A *p*-distance is the proportion of sites at which two sequences differ. Based on the four sequences (W, X, Y and Z) below, complete the following *p*-distance matrix of these sequences. (3 marks)

```
W  CAGCATATG
X  CATCAACTA
Y  CAGCATTTC
Z  CTTGTGAAC
```

	W	X	Y	Z
W	0.00			
X	4/9=0.44	0.00		
Y			0,00	
Z			0.78	0.00

p-distance matrix

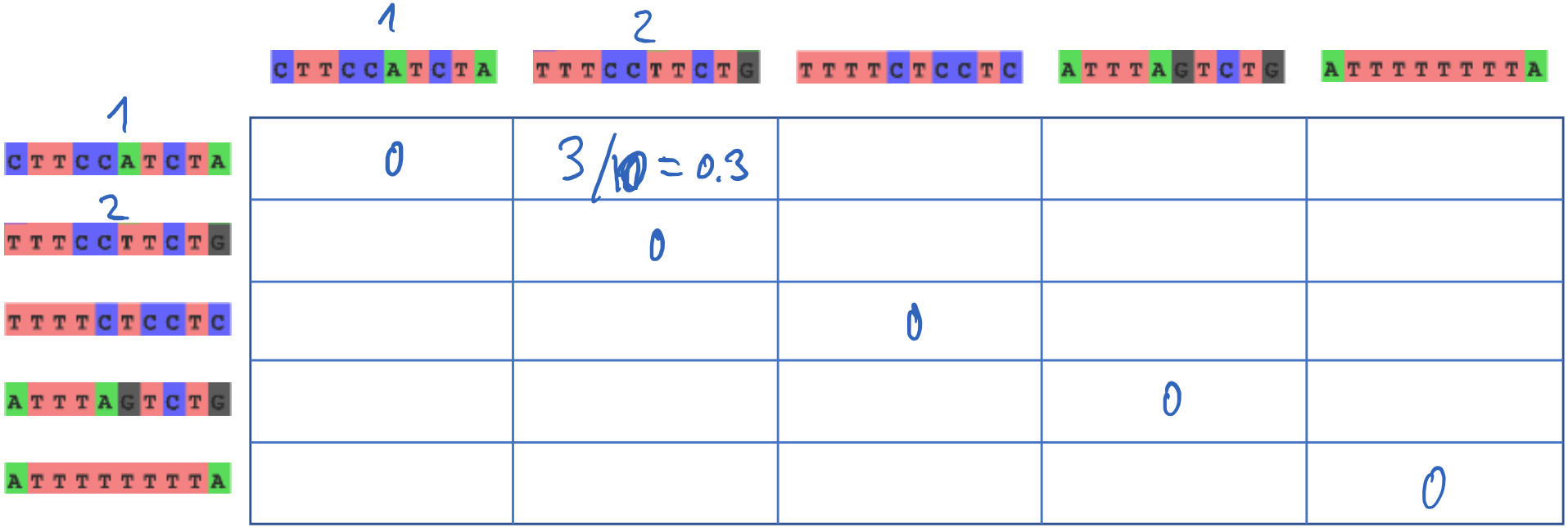
Type for example:
XX=0.00
ZY=0.78
But for the pairs missing

Tip:
1/9 = 0.11
8/9 = 0.89

$$p = \frac{D}{L}$$

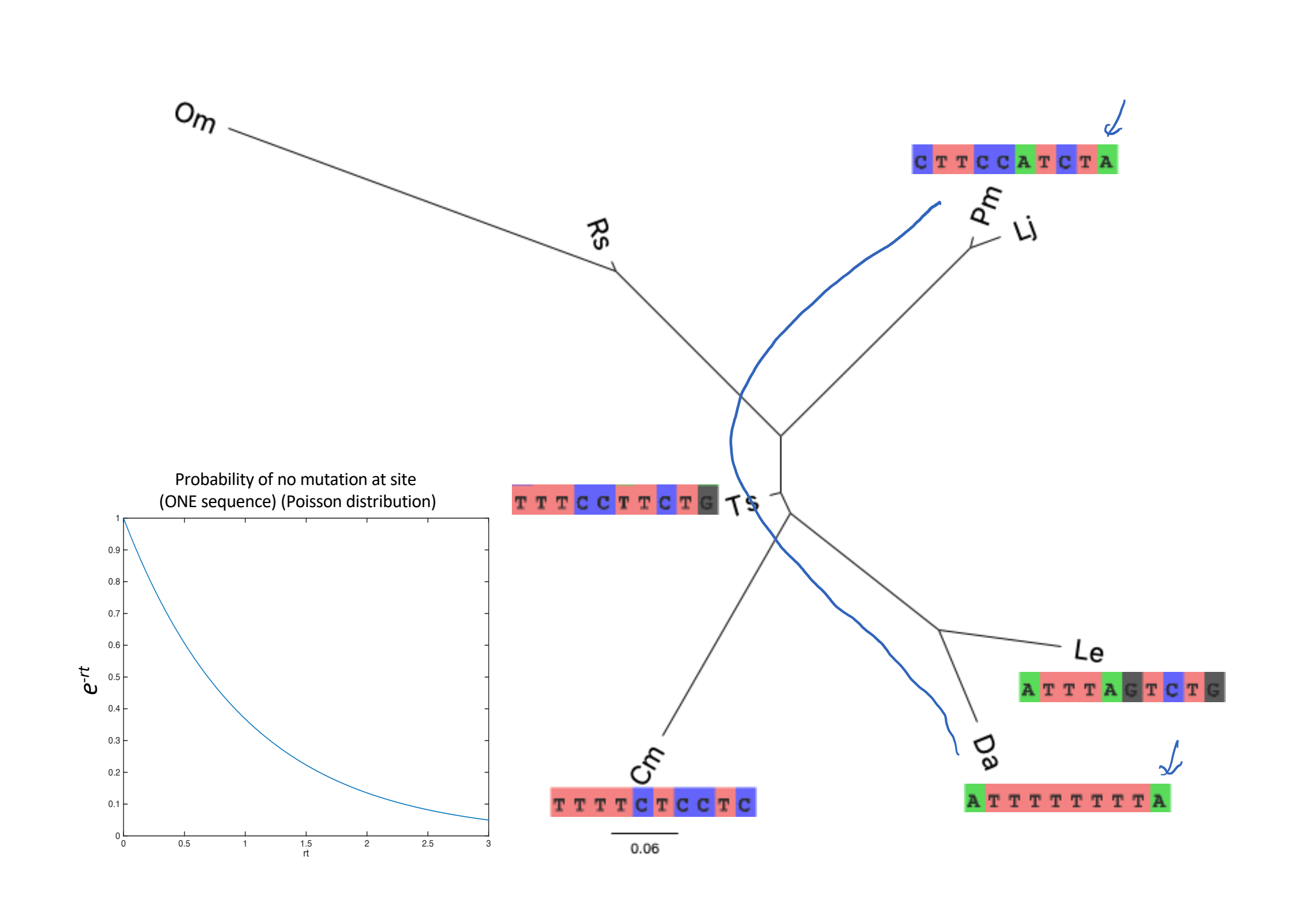
p-distance (aka fractional distance)

Distance matrix

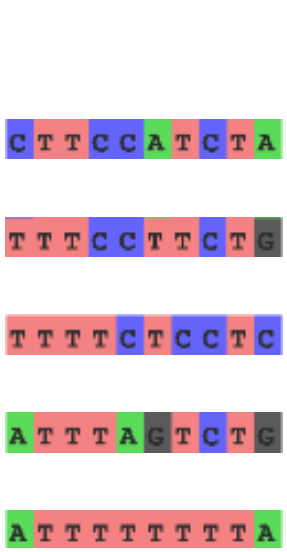


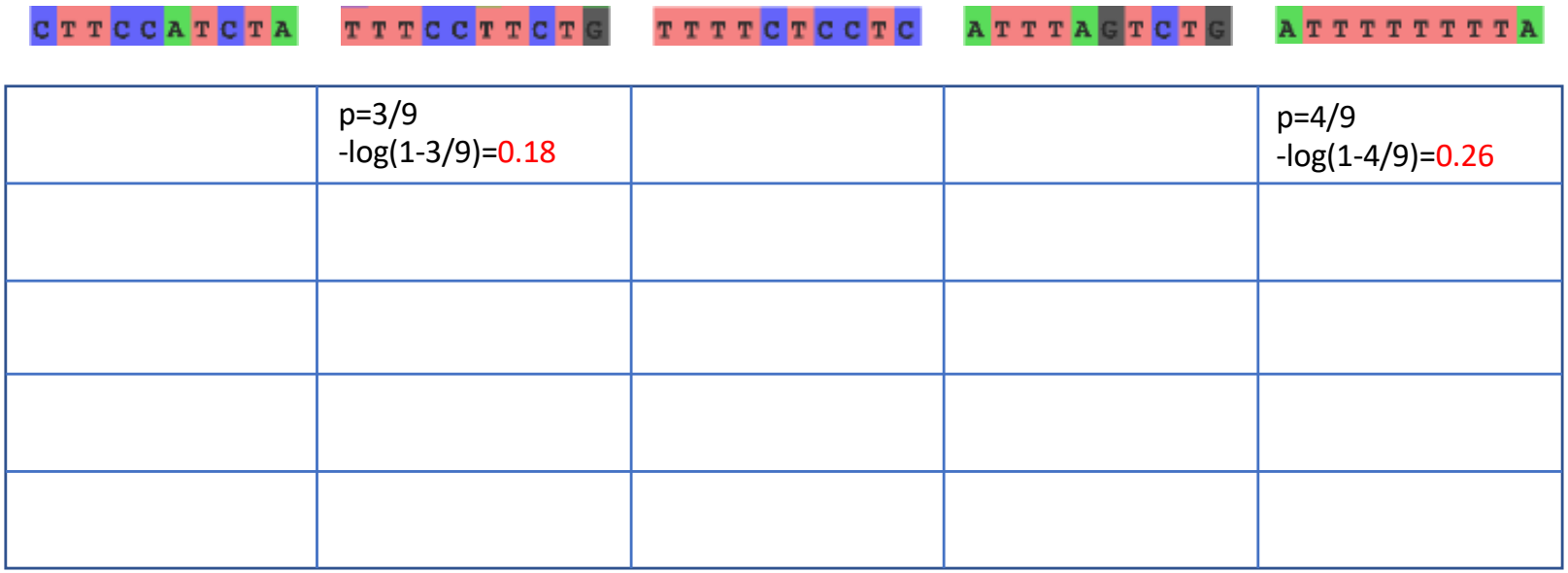
Two conditions for evolutionary time to be proportional to number of changes observed from an alignment:

1. All sequences mutate at a constant rate
2. No position has mutated more than once

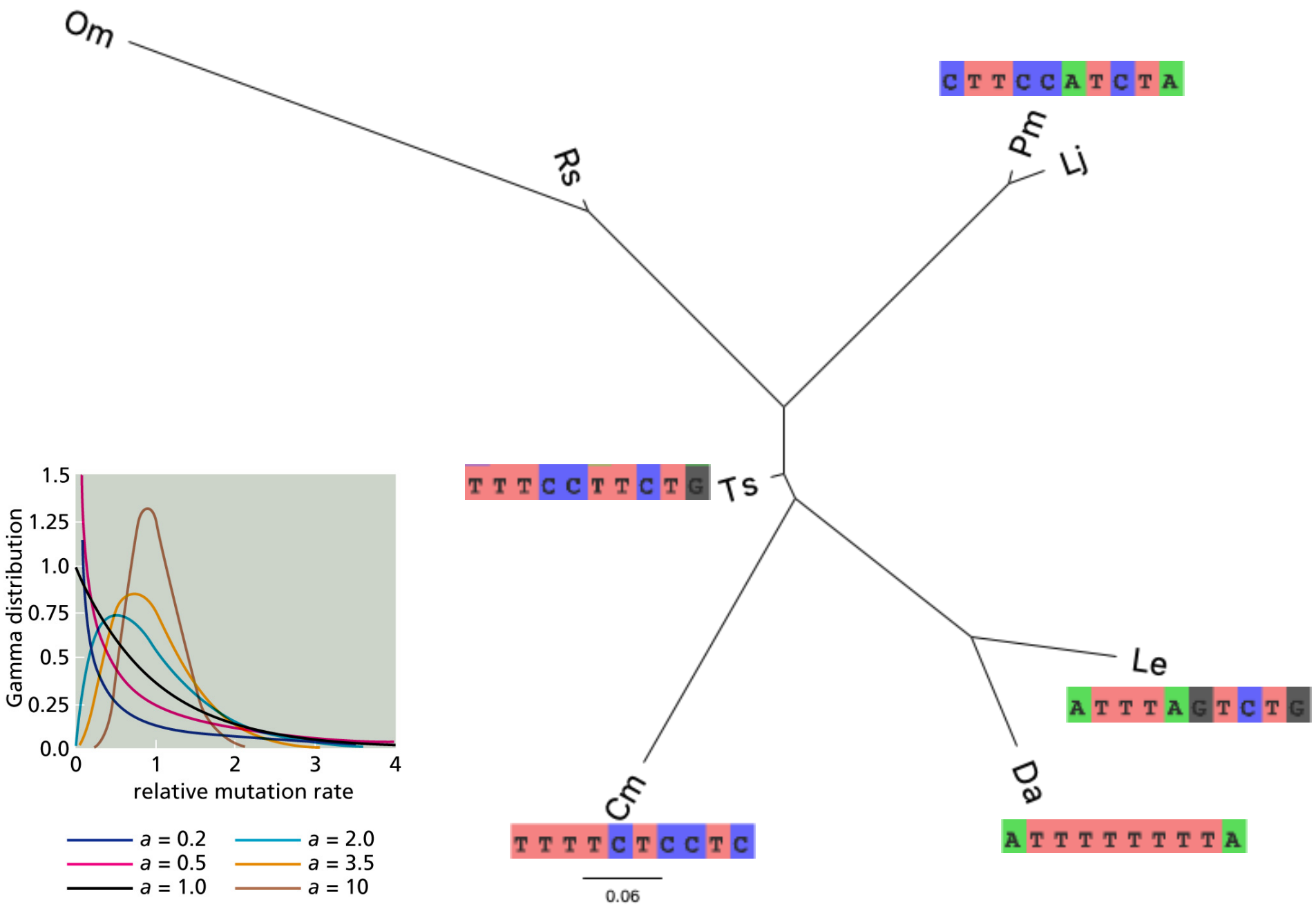

Om
Rs
Pm
Lj
C T T C C A T C T A
T T T C C T T C T G
Ts
Le
A T T T A G T C T G
Cm
T T T T C T C C T C
Da
A T T T T T T T T A
0.06
Probability of no mutation at site
(ONE sequence) (Poisson distribution)
e^{-rt}
rt
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.5
1
1.5
2
2.5
3

Poisson-corrected distance $d_p = -\ln(1-p)$



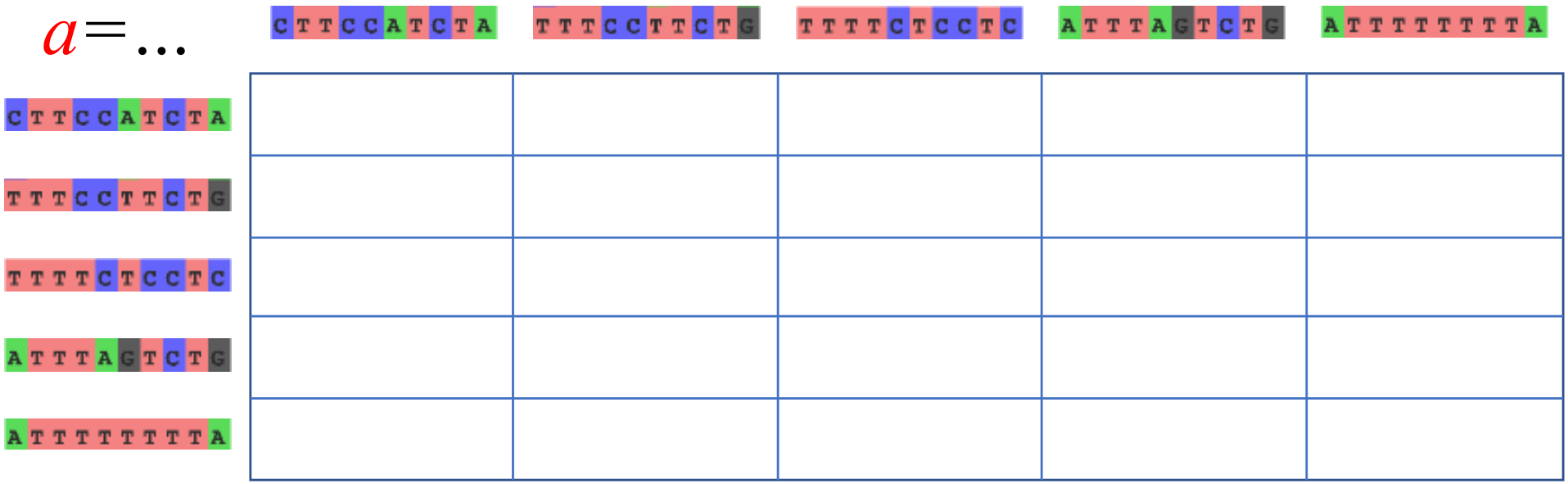


Accounts for multiple mutations at site


Om
Rs
Pm
Lj
C T T C C A T C T A
T T T C C T T C T G
Ts
Le
A T T T A G T C T G
Cm
T T T T C T C C T C
Da
A T T T T T T T T A
0.06
Gamma distribution
1.5
1.25
1.0
0.75
0.5
0.25
0.0
0
1
2
3
4
relative mutation rate
a = 0.2
a = 0.5
a = 1.0
a = 2.0
a = 3.5
a = 10

Gamma-corrected distance

$$d_\Gamma = a[(1-p)^{-1/a} - 1]$$



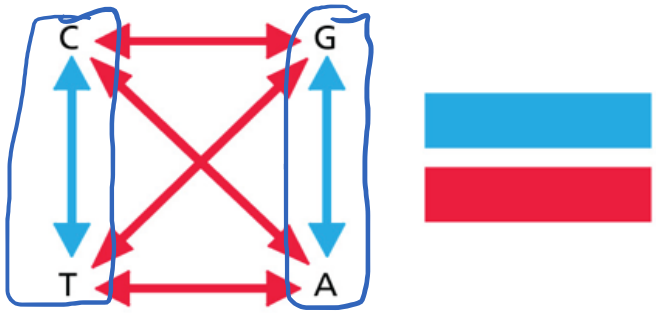
Corrects distance estimate for changes that can be explained by a variable rate

Questions so far?

About distance metrics in particular

P10: Exam question from 2020

To model evolutionary changes between the four DNA bases A, G, C and T, it is possible to distinguish between two classes of substitutions as depicted below: between the two pyrimidines C and T, and between the two purines G and A (in blue), as opposed to all others (in red; changing a pyrimidine to a purine, or vice versa).

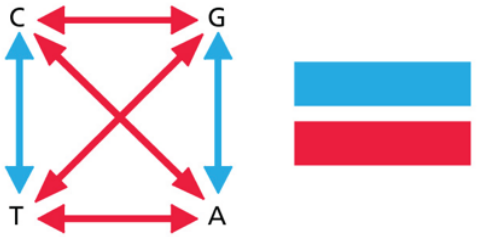


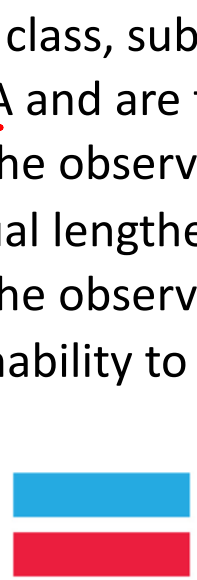
P10A: What name is used to refer to the blue class?

P10B: The red class?

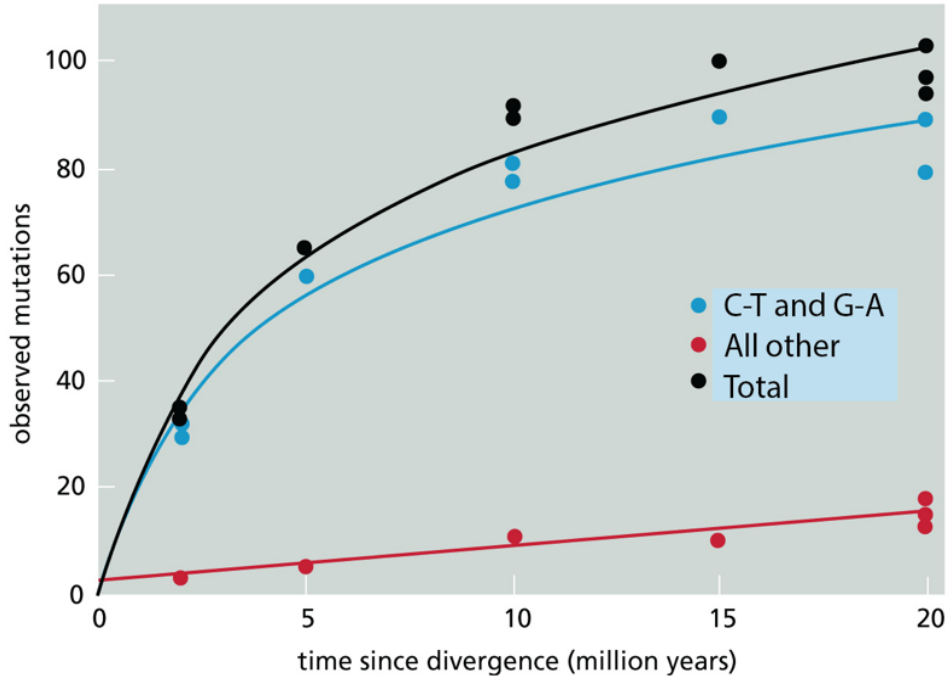
P10C: In the Figure below, observed mutations of the blue class surpass those of the red class over evolutionary time. Identify all statements below that correctly explain the trends:

A. For protein-coding regions of the genome, relative to the blue class, substitutions of the red class are more likely to result in non-synonymous amino acid changes and are therefore less tolerated
B. Relative to the red class, substitutions of the blue class tend to have smaller impact on the fold of the DNA and are therefore less likely to disrupt biological function
C. The saturation of the observed number of substitutions from the blue class over time is explained by gradual lengthening of genome lengths
D. The saturation of the observed number of substitutions from the blue class over time is explained by our inability to count actual changes when they occur over and over





Write each letter on a separate row (or select option if "Choose option/s")



P10D: In the Figure below, three standard DNA rate matrices to model evolutionary change are given (from left to right) JC, K81 and F81; each matrix is specified by parameters as indicated in the elements (asterisks are calculated). (Rows specify source and columns specify target base, ordered as A, G, C and T.) Which of the three matrices has the capacity to distinguish between the red and blue classes of base change?

Write each model acronym on a separate row

$$Q = \begin{array}{c} \\ A \\ G \\ C \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & G & C & T \end{array} \\ \begin{pmatrix} * & \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & * & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & * & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} & * \end{pmatrix} \end{array}$$

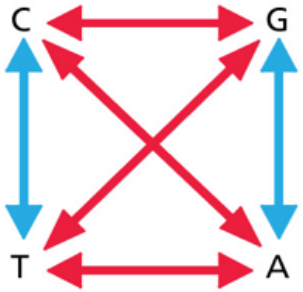
JC

$$Q = \begin{array}{c} \\ A \\ G \\ C \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & G & C & T \end{array} \\ \begin{pmatrix} * & \alpha & \beta & \gamma \\ \alpha & * & \gamma & \beta \\ \beta & \gamma & * & \alpha \\ \gamma & \beta & \alpha & * \end{pmatrix} \end{array}$$

K81

$$Q = \begin{pmatrix} * & \pi_G & \pi_C & \pi_T \\ \pi_A & * & \pi_C & \pi_T \\ \pi_A & \pi_G & * & \pi_T \\ \pi_A & \pi_G & \pi_C & * \end{pmatrix}$$

F81



Questions so far?

About evolutionary models/rate matrices in particular

Reflections (to set you up for what is next)

- *What are the two independence assumptions made for Markov chains, for the purpose of modelling evolution? Hint: (aligned) sequences typically consist of multiple positions, and evolution happens over time.*
- *Challenge: Maximum likelihood for phylogenetic tree inference means what? (a) finding the most likely tree, given the sequence content at the leaves, or (b) finding the tree, that assigns the greatest likelihood to the observable sequence content*

Maximum likelihood finds H

$$\text{argmax}_H \, P(D \mid H)$$

where D is the data (extant states), and H the hypothesis of what happened (tree and/or ancestor states)

Bayesian methods estimate

$$P(H \mid D) \propto P(D \mid H)\, P(H)$$