Phylogenetics 2: week 8

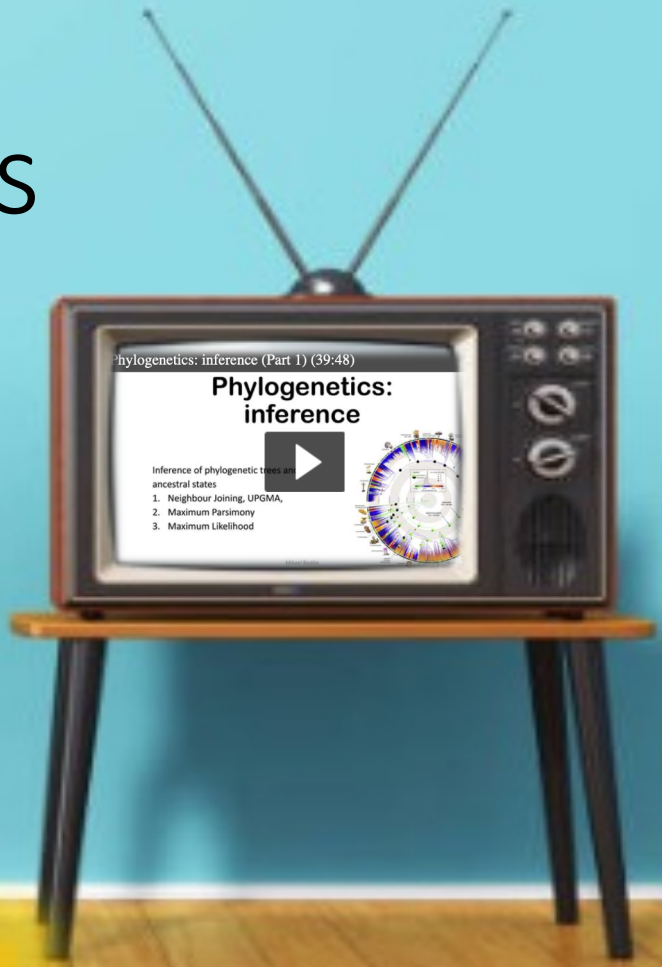
Watch the recordings

Phylogenetics: inference

3 parts + UPGMA demo

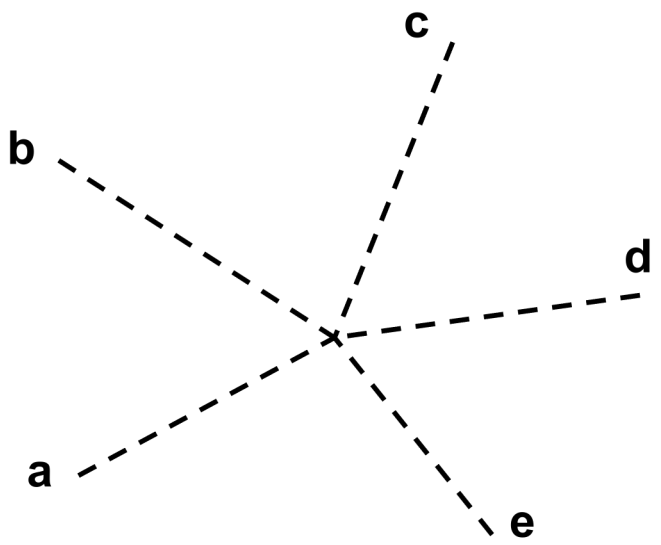
In this session we play with:

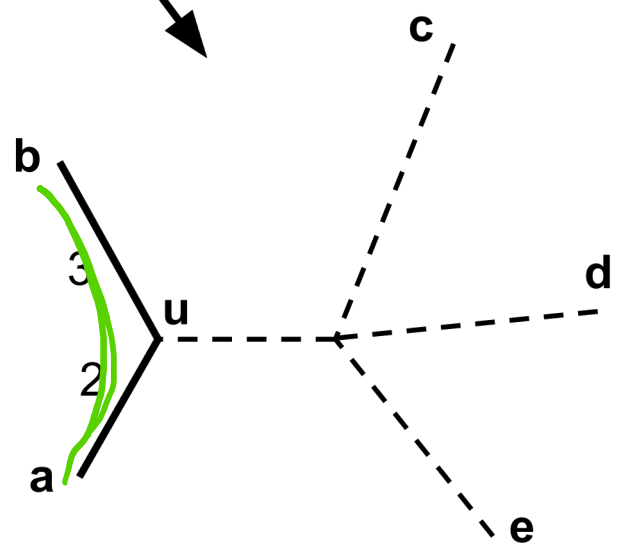
- *Neighbor joining*
- *UPGMA*
- *Maximum parsimony*
- *Maximum likelihood*



Mikael Bodén

By Tomfy at English Wikipedia, CC BY-SA 3.0,
https://commons.wikimedia.org/w/index.php?curid=49046444





$n=5$

	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

"Distance" between a and b d(a,b), d and e d(d,e)...

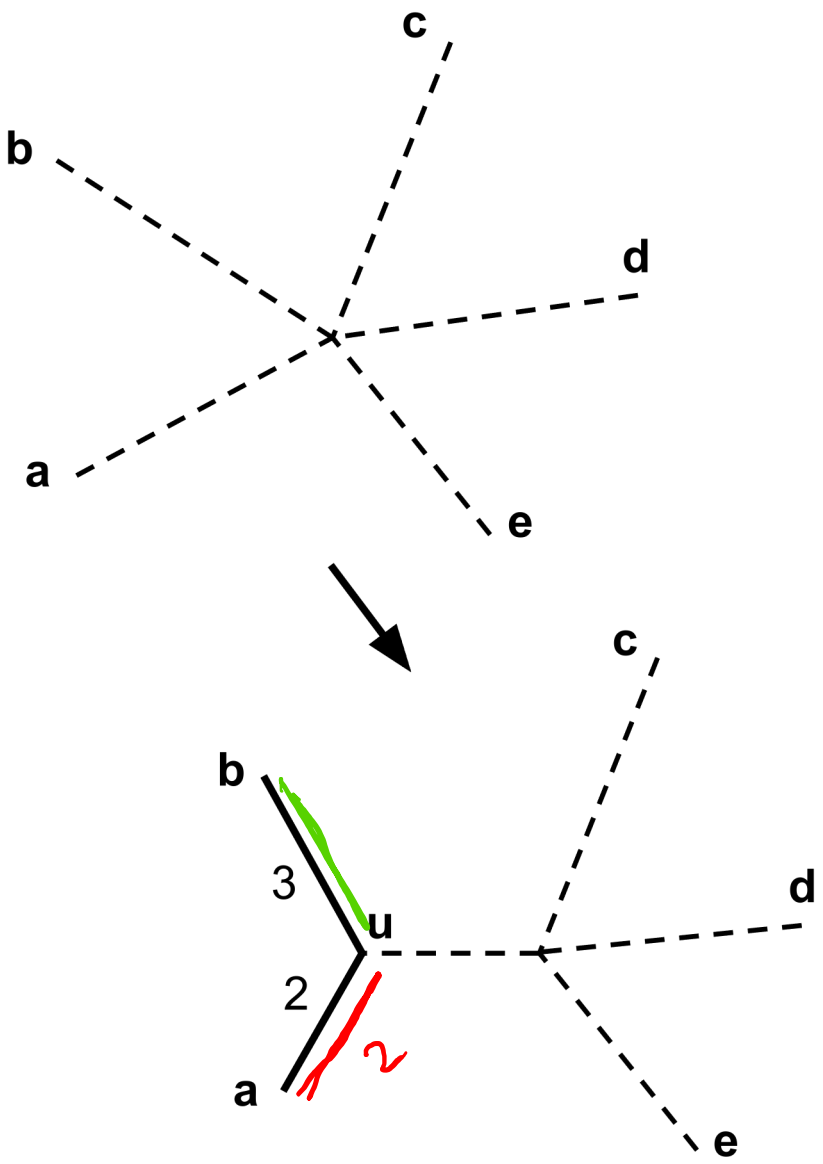
$$Q(i,j) = (n-2)d(i,j) - \sum_{k=1}^{n} d(i,k) - \sum_{k=1}^{n} d(j,k)$$

Looking-in *Looking-out*

$Q(a,b) = 3 \cdot 5 - 31 - 34 = 15 - 31 - 34 = -50$

$Q(d,e) = 3 \cdot 3 - 30 - 27 = 9 - 57 = -48$

By Tomfy at English Wikipedia, CC BY-SA 3.0,
https://commons.wikimedia.org/w/index.php?curid=49046444



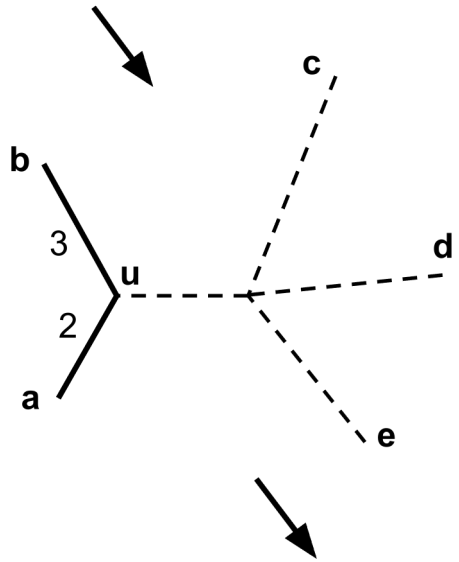
	a	b	c	d	e
a	0	5	9	9	8
b	5	0	10	10	9
c	9	10	0	8	7
d	9	10	8	0	3
e	8	9	7	3	0

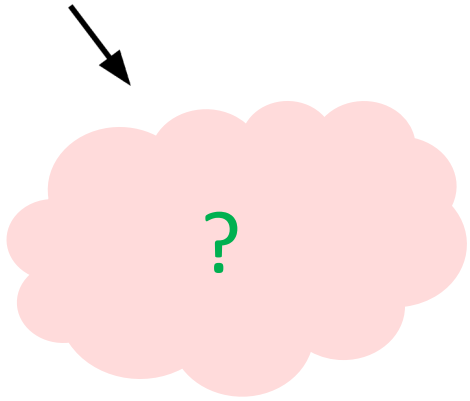
$$\delta(a,u) = \frac{1}{2}d(a,b) + \frac{1}{2(n-2)}\left[\sum_{k=1}^{n} d(a,k) - \sum_{k=1}^{n} d(b,k)\right]$$

$\frac{5}{2} + \frac{1}{6} \cdot (31-34) = 2.5 - 0.5 = 2$

$5-2=3$

By Tomfy at English Wikipedia, CC BY-SA 3.0,
https://commons.wikimedia.org/w/index.php?curid=49046444





n=4

	u	c	d	e
u	0	7	7	6
c	7	0	8	7
d	7	8	0	3
e	6	7	3	0

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

Which pair are we JOINING next?

A. u and c $Q(u,c) = 2\cdot 7 - 20 - 22 = -28$

B. u and d

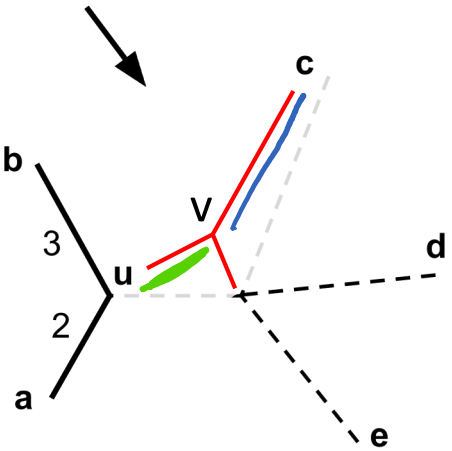
C. u and e

D. c and d

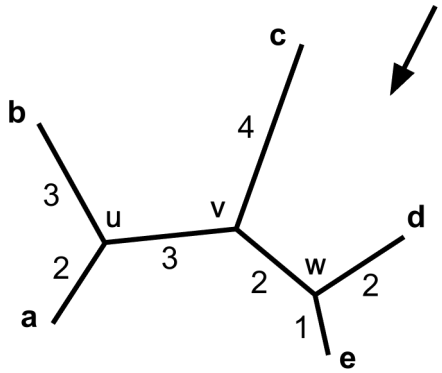
E. c and e

F. d and e $Q(d,e) = 2\cdot 3 - 18 - 16 = -28$

By Tomfy at English Wikipedia, CC BY-SA 3.0,
https://commons.wikimedia.org/w/index.php?curid=49046444



?



https://en.wikipedia.org/wiki/Neighbor_joining

$$\delta(u,v) = \frac{1}{2}d(u,c) + \frac{1}{2(n-2)}\left[\sum_{k=1}^{n} d(u,k) - \sum_{k=1}^{n} d(c,k)\right]$$

And the distance between u and v is 1, 2, 3, 4 or 5?

$$\delta(u,v) = \frac{1}{2}d(u,c) + \frac{1}{2(n-2)}\left[\sum_{k=1}^{n} d(u,k) - \sum_{k=1}^{n} d(c,k)\right]$$

$\delta(u,v) = \frac{7}{2} + \frac{1}{4}(20 - 25) = 3.5 - 0.5$

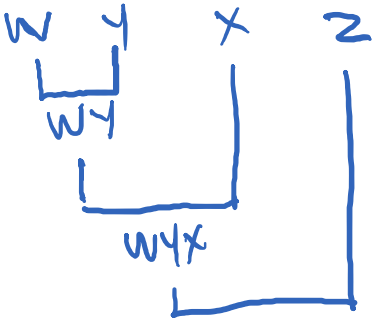
Between c and v?

$\delta(v,c) = 7 - 3 = 4$

Questions so far?

About Neighbor-joining

. A *p*-distance is the proportion of sites at which two sequences differ. Based on the *p*-distance matrix for four sequences (W, X, Y and Z) below, draw a phylogenetic tree structure that shows the relationship among these sequences based on the principles of UPGMA. No calculation is required. (3 marks)

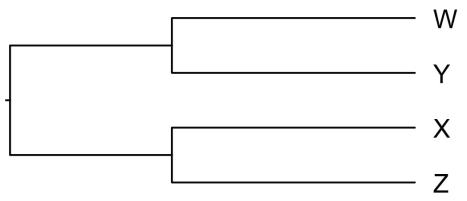


	W	X	Y	Z
W	0.00			
X	0.44	0.00		
Y	0.22	0.44	0.00	
Z	0.78	0.78	0.78	0.00

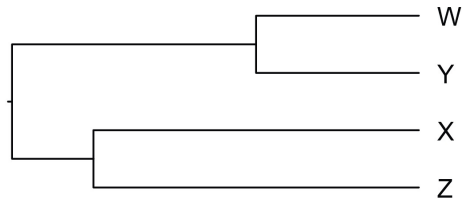
W CAGCATATG
X CATCAACTA
Y CAGCATTTC
Z CTTGTGAAC

Which one is it?

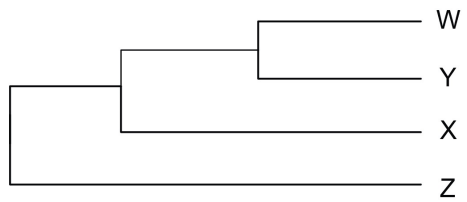
A



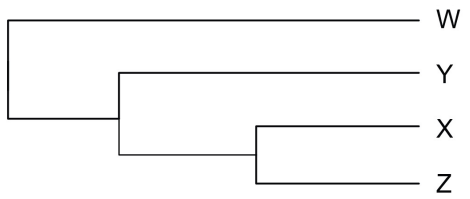
B



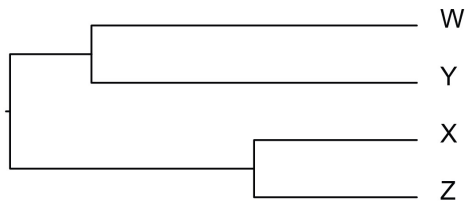
C



D



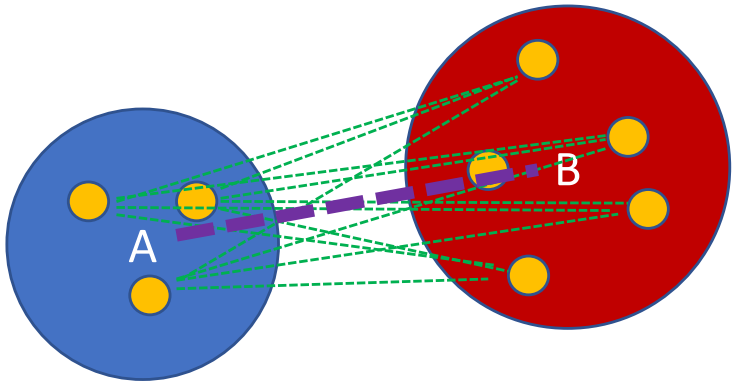
E



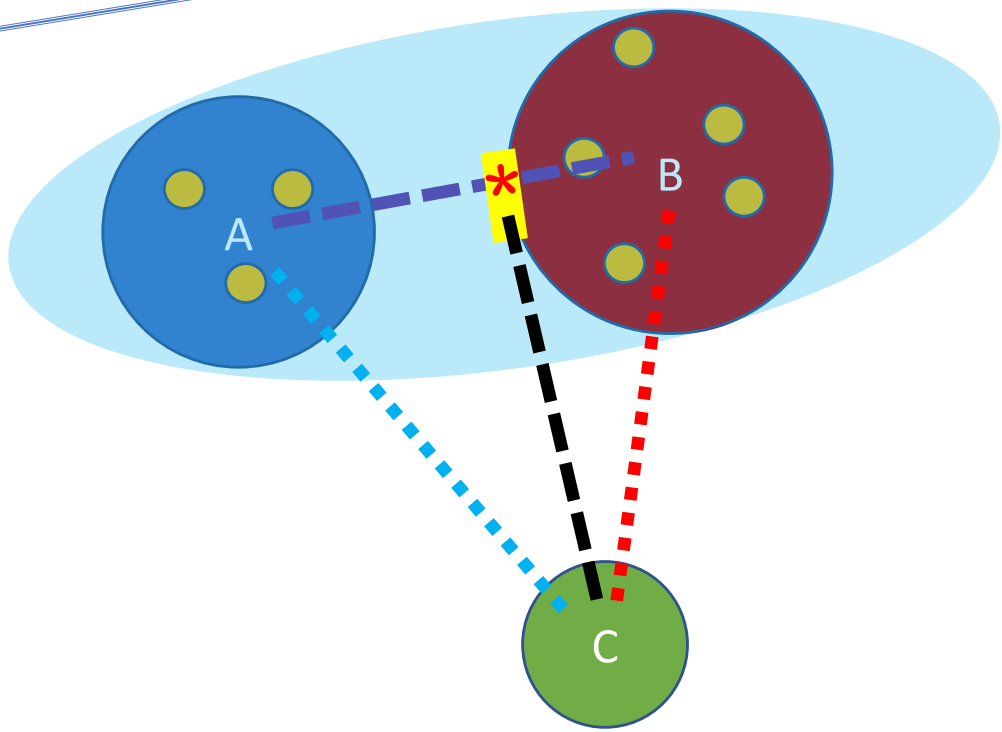
$$d_{AB} = \frac{1}{N_A N_B} \sum_{i \in A, j \in B} d_{ij}$$

A and B are groups containing 1 or more sequences; N_A and N_B are their sizes

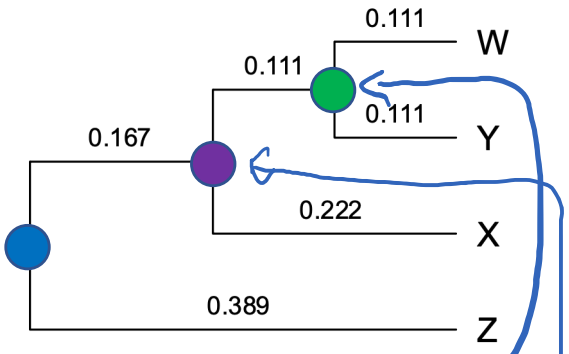
Sum the distances between all pairs of sequences, one from A and one from B; calculate their arithmetic average



$$d_{*C} = \frac{N_A d_{AC} + N_B d_{BC}}{N_A + N_B}$$



```
[0.         0.44444444 0.22222222 0.77777778]
[0.44444444 0.         0.44444444 0.77777778]
[0.22222222 0.44444444 0.         0.77777778]
[0.77777778 0.77777778 0.77777778 0.        ]
```



4 nodes remain

Inspecting "X" and "W" at distance 0.444
Inspecting "W" and "Y" at distance 0.222
Inspecting "X" and "Y" at distance 0.444
Inspecting "Z" and "W" at distance 0.778
Inspecting "Z" and "X" at distance 0.778
Inspecting "Z" and "Y" at distance 0.778
Closest pair is "W" (1) and "Y" (1) at distance 0.222 form new node (W,Y):0.111

(W,Y):0.111 gets distance to "X": (1 * 0.444 + 1 * 0.444) / (1 + 1) = 0.444
(W,Y):0.111 gets distance to "Z": (1 * 0.778 + 1 * 0.778) / (1 + 1) = 0.778

3 nodes remain

Inspecting "Z" and "X" at distance 0.778
Inspecting "(W,Y):0.111" and "X" at distance 0.444
Inspecting "(W,Y):0.111" and "Z" at distance 0.778
Closest pair is "(W,Y):0.111" (2) and "X" (1) at distance 0.444 form new node ((W,Y):0.111,X):0.222

((W,Y):0.111,X):0.222 gets distance to "Z": (2 * 0.778 + 1 * 0.778) / (2 + 1) = 0.778

2 nodes remain

Inspecting "((W,Y):0.111,X):0.222" and "Z" at distance 0.778
Closest pair is "((W,Y):0.111,X):0.222" (3) and "Z" (1) at distance 0.778 form new node (((W,Y):0.111,X):0.222,Z):0.389

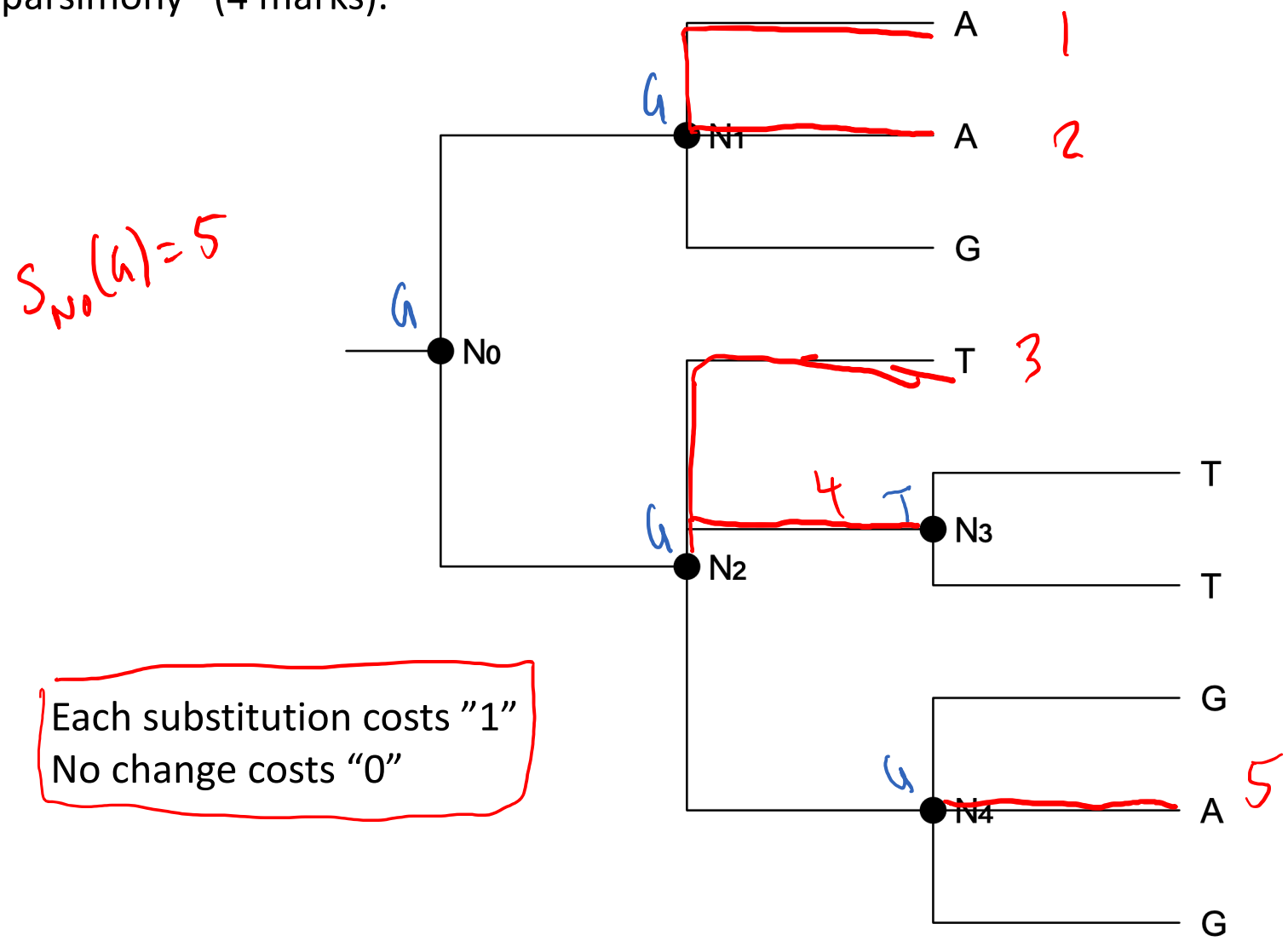
More complex example in UPGMA video

Questions so far?

About UPGMA

Exam 2020: You are provided below with a phylogenetic tree representing a single position in nine genomes (represented by a leaf node). The leaf nodes are labelled with the corresponding base. Ancestor nodes are named N0 through N4; N0 is the root of the tree.

Assign labels to the internal (ancestral) nodes N1 through N4 that give the optimal "parsimony" (4 marks).



There are multiple assignments of N0, each of which form part of the most parsimonious solution. Give all labels of N0 that are optimal (1 mark).

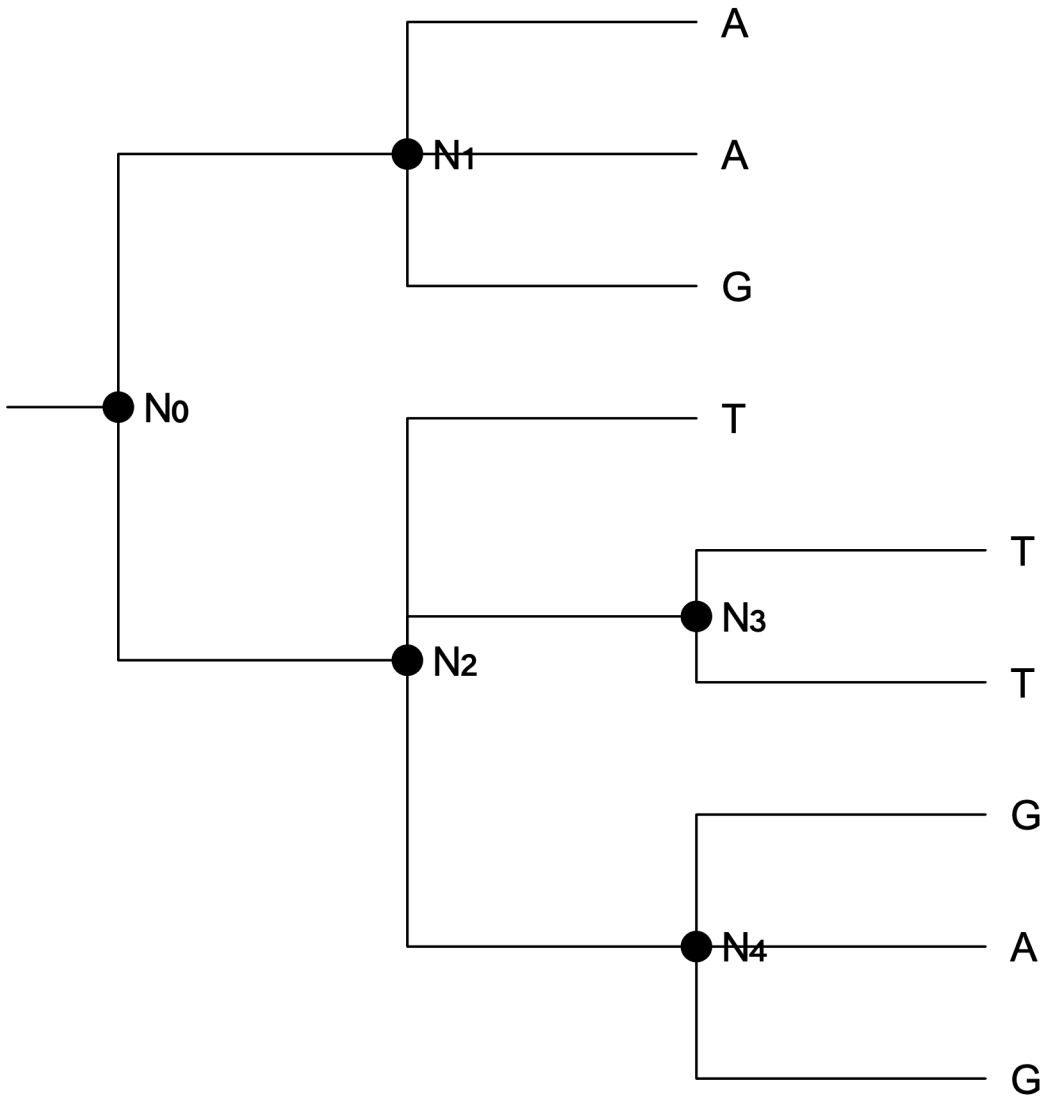
Respond with each possible assignment by typing A, C, G or T

$S_{N0}(A) =$

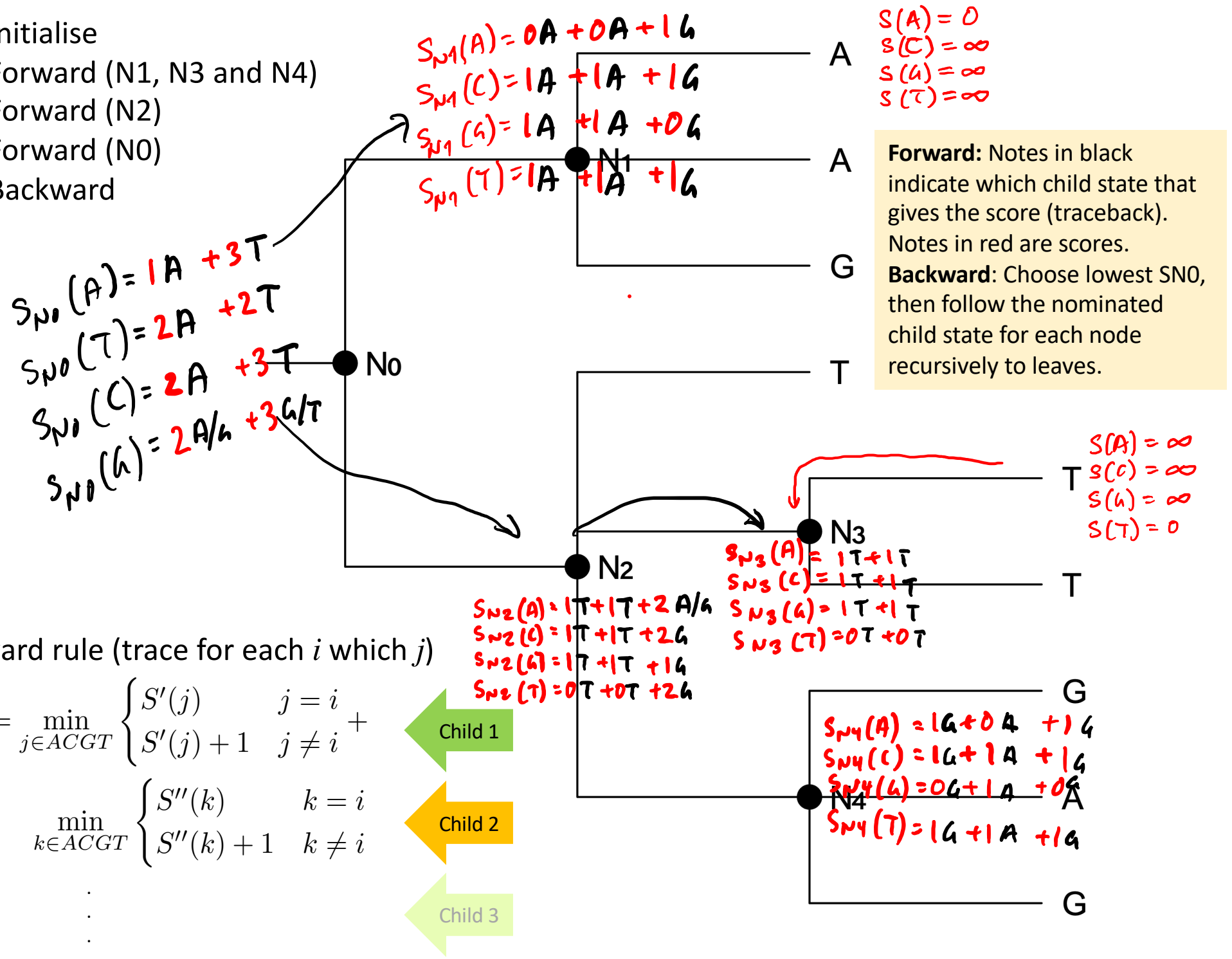
$S_{N0}(C) =$

$S_{N0}(G) = 5$

$S_{N0}(T) =$



- Initialise
- Forward (N1, N3 and N4)
- Forward (N2)
- Forward (N0)
- Backward



Forward: Notes in black indicate which child state that gives the score (traceback). Notes in red are scores.
Backward: Choose lowest SN0, then follow the nominated child state for each node recursively to leaves.

Forward rule (trace for each i which j)

$$S(i) = \min_{j \in ACGT} \begin{cases} S'(j) & j = i \\ S'(j) + 1 & j \neq i \end{cases} +$$

Child 1

$$\min_{k \in ACGT} \begin{cases} S''(k) & k = i \\ S''(k) + 1 & k \neq i \end{cases}$$

Child 2

.
.
.

Child 3

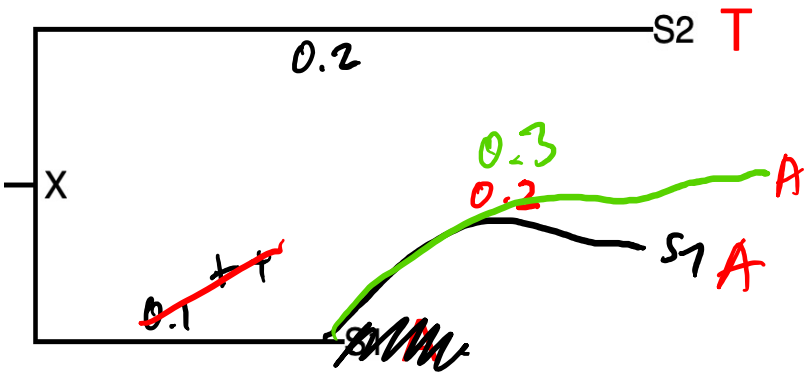
Questions so far?

About Maximum parsimony

Using Jukes-Cantor's evolutionary model of DNA (ACGT) (with substitution probabilities given below) on the phylogenetic tree depicted on the right.

A. which two assignments of the ancestor x are equally probable? Type A, C, G or T

B. true or false: if you increase t on the branch from x to s_1, the probability of x=T increases? Type true or false



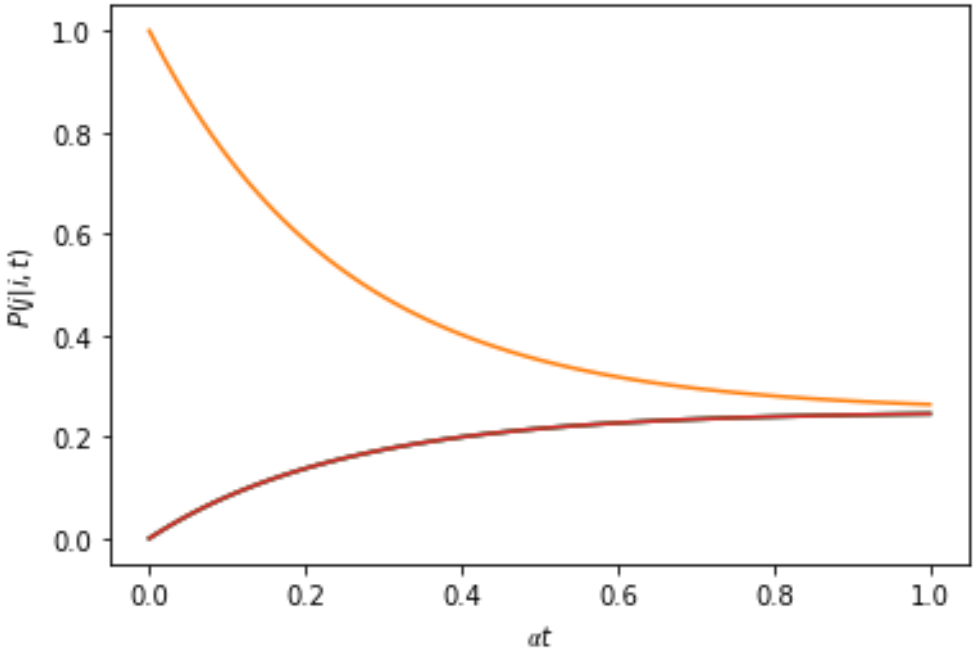
Both questions refer to this quantity:

$$P(x \mid s_1{=}\mathrm{A}, t_1{=}0.1, s_2{=}\mathrm{T}, t_2{=}0.2) = \langle \ldots \rangle$$

Both answers are based on that x's descendants are observed as s_1=A and s_2= T at distances 0.1 and 0.2, respectively. α is set to 1.

$$P(j|i,t) = \begin{cases} \frac{1}{4}(1+3e^{-4\alpha t}) & \text{for } j{=}i \\ \frac{1}{4}(1-e^{-4\alpha t}) \end{cases}$$

Uniform equilibrium frequencies are assumed.



The probability that base i mutates into base j in a time t. The orange/upper line represents $i{=}j$, other lines $i{\neq}j$.

Exam will NOT have ML calculations,
but may probe understanding of principles…

$P(x \mid s_1=\text{A}, t_1=0.1, s_2=\text{T}, t_2=0.2) =$

proportional to the joint probability

$P(x, s_1=\text{A}, t_1=0.1, s_2=\text{T}, t_2=0.2) =$

Exam will not have ML calculations, so this slide is *just* to demonstrate that it is *not* magic…

x	P
A	
C	
G	
T	

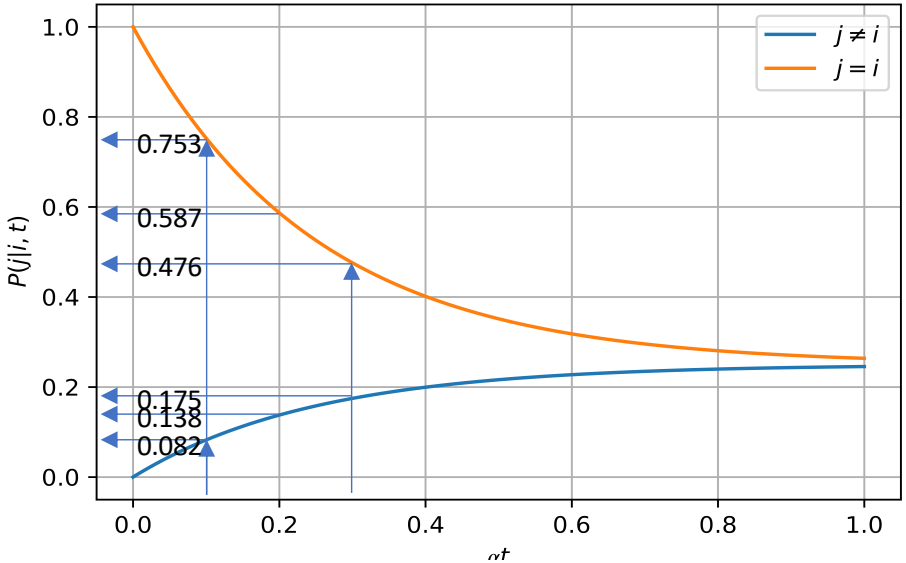
$P(s_1=\text{A} \mid x=\text{A}, t=0.1)\ P(s_2=\text{T} \mid x=\text{A}, t=0.2)\ P(x=\text{A})$

$P(s_1=\text{A} \mid x=\text{C}, t=0.1)\ P(s_2=\text{T} \mid x=\text{C}, t=0.2)\ P(x=\text{C})$

$P(s_1=\text{A} \mid x=\text{G}, t=0.1)\ P(s_2=\text{T} \mid x=\text{G}, t=0.2)\ P(x=\text{G})$

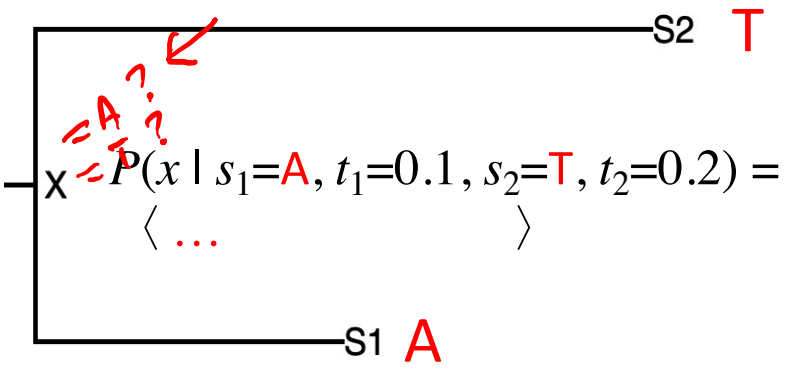
$P(s_1=\text{A} \mid x=\text{T}, t=0.1)\ P(s_2=\text{T} \mid x=\text{T}, t=0.2)\ P(x=\text{T})$

For the curious: the joint probability of multiple random events can be determined as a product of conditional probabilities (a "chain" of the so-called "product rule").
By the "Markov assumption" transitions at a descendant is limited to that of the direct ancestor.



Using Jukes-Cantor's evolutionary model of DNA (ACGT) (with substitution probabilities given below) on the phylogenetic tree depicted on the right.

A. which two assignments of the ancestor x are equally probable?
B. if you increase x on the branch from x to s_1, the probability of x=T increases; true or false?



Both answers are based on that x's descendants are observed as s_1=A and s_2= T at distances 0.1 and 0.2, respectively. α is set to 1.

So… the tree with x=A assigns a greater likelihood than that with x=T to… the observed states at the tips of the tree, and ultimately the data at hand

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)

Questions so far?

About Maximum likelihood