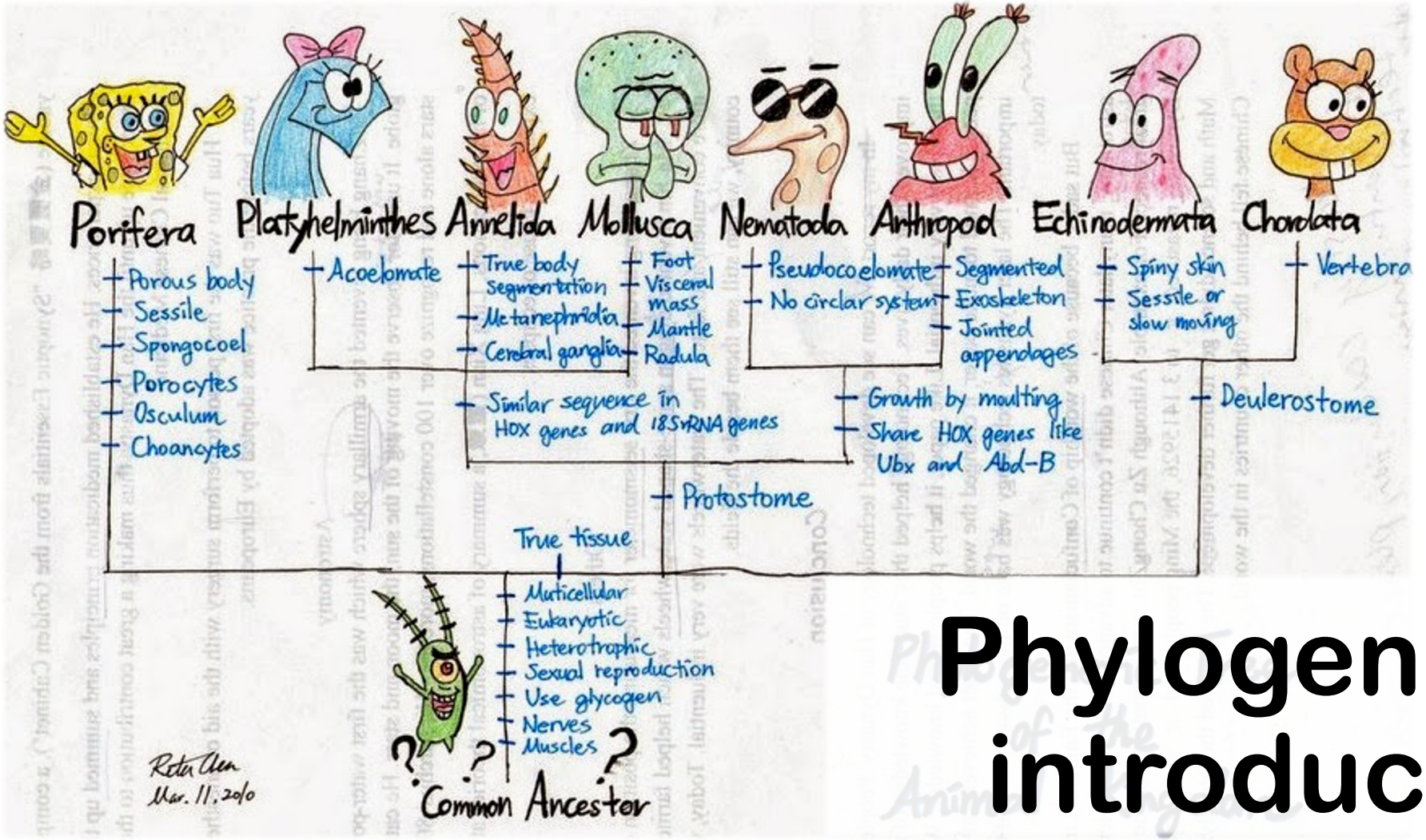


Phylogenetics introduction

http://phylonetworks.blogspot.com/2013/10/phylogenetics-with-spongebob.html
https://www.deviantart.com/thehurricanes/art/Phylogeny-with-Spongebob-156909686

Episode in the series on phylogenetics

Mikael Bodén

Part 1: Phylogenetics introduction

Motivation

Background

Homology

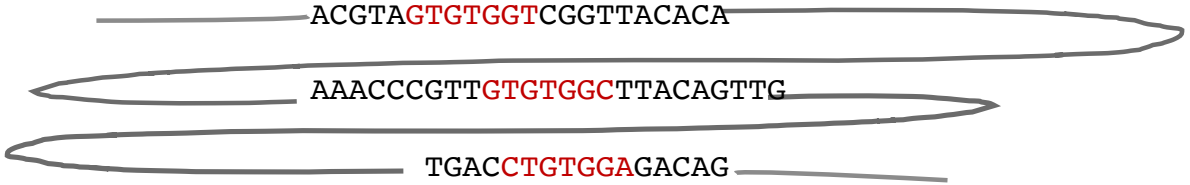
1

Reference …ACGTACGGTTACACAAAACCCGTTTTACGTAGTTGTGACG…

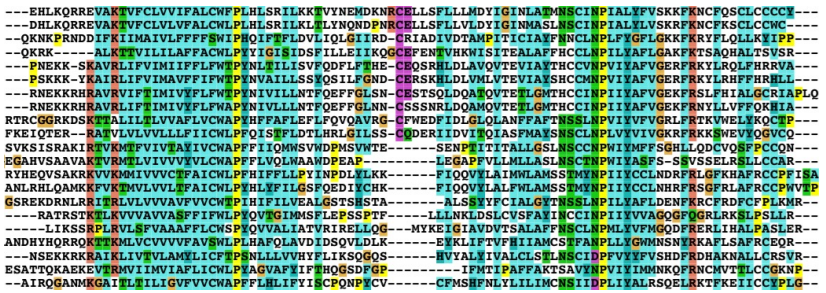
Sequence reads

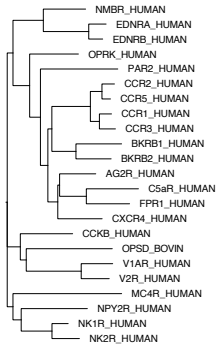
TTACACAAATCCCGTTGCACGTA
TACACAAATCCCGTTGTACGTAG
ACACAAAACCCGTTGCACGTAGT
CAAATCCCGTTGCACGTAGTTGT
AAATCCCGTTGTACGTAGTTGTG

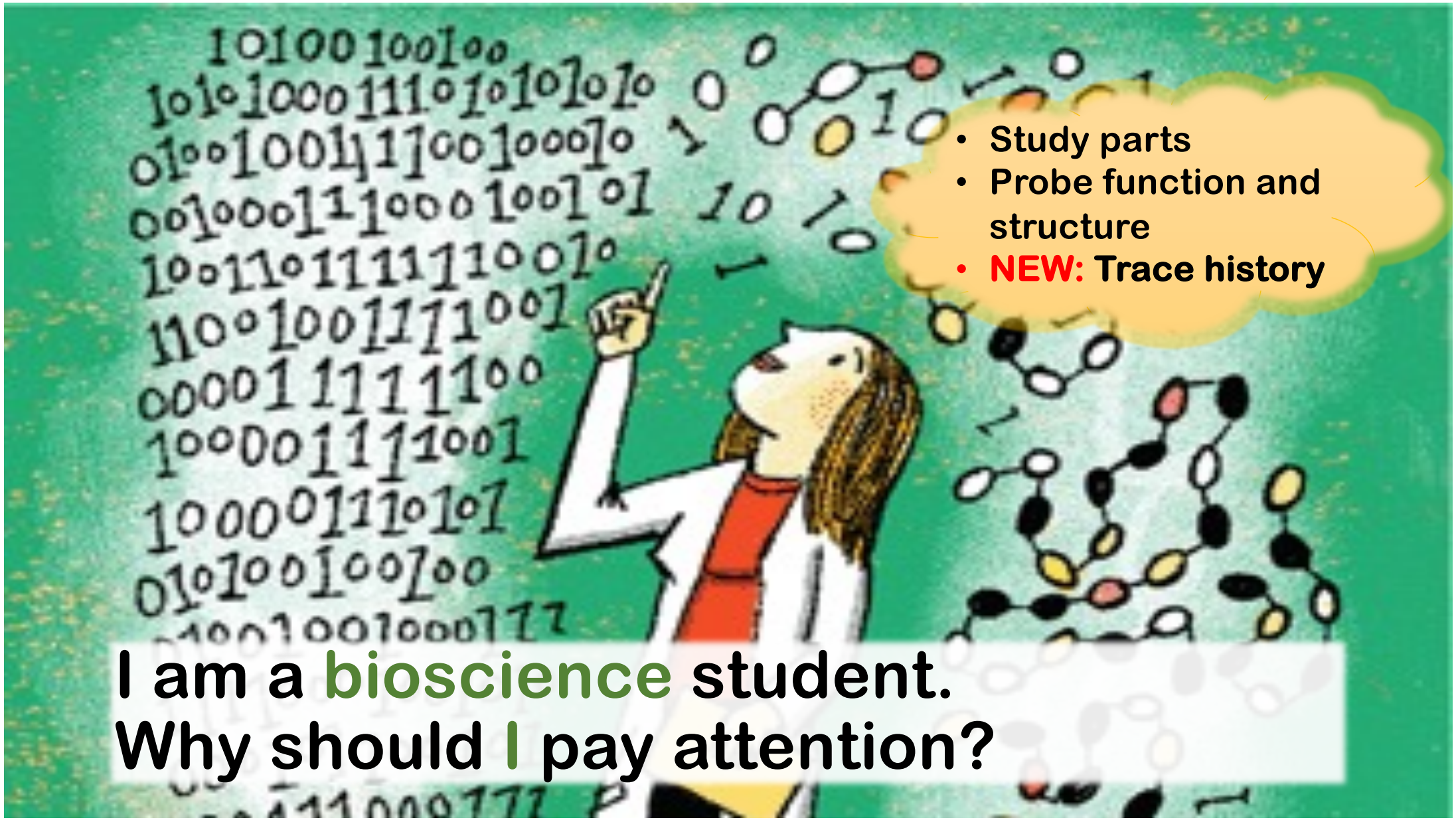
2



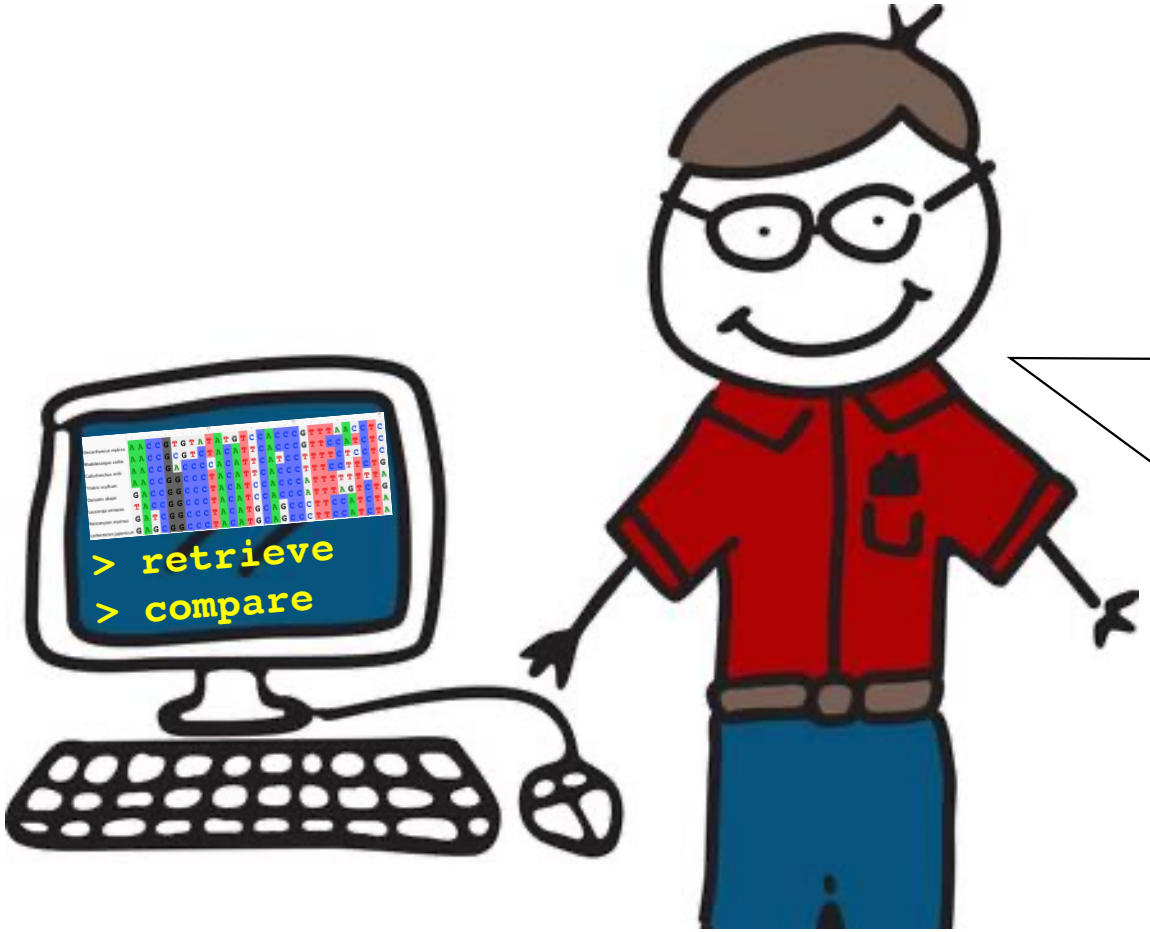
3






• Study parts
• Probe function and structure
• NEW: Trace history
I am a bioscience student.
Why should I pay attention?

I am a computing student. Why should I pay attention?

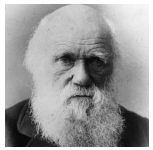


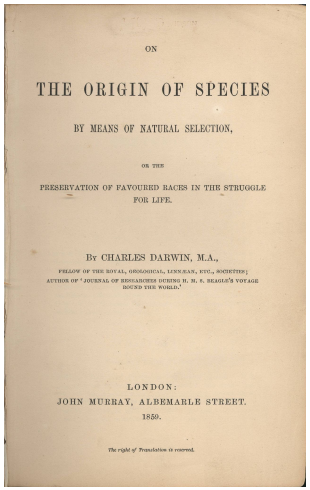
How cool is it that I will learn

- how to determine *when* existing sequences shared an ancestor?
- how to recreate phylogenetic trees, and ancient, probably extinct sequences?
- what it takes to identify the evolutionary events that formed existing sequences?

Phylogeny and common ancestor

- *All* organisms are related by genealogical descent with *modification*, as on a *branching* tree





ON

THE ORIGIN OF SPECIES

BY MEANS OF NATURAL SELECTION,

OR THE

PRESERVATION OF FAVOURED RACES IN THE STRUGGLE FOR LIFE.

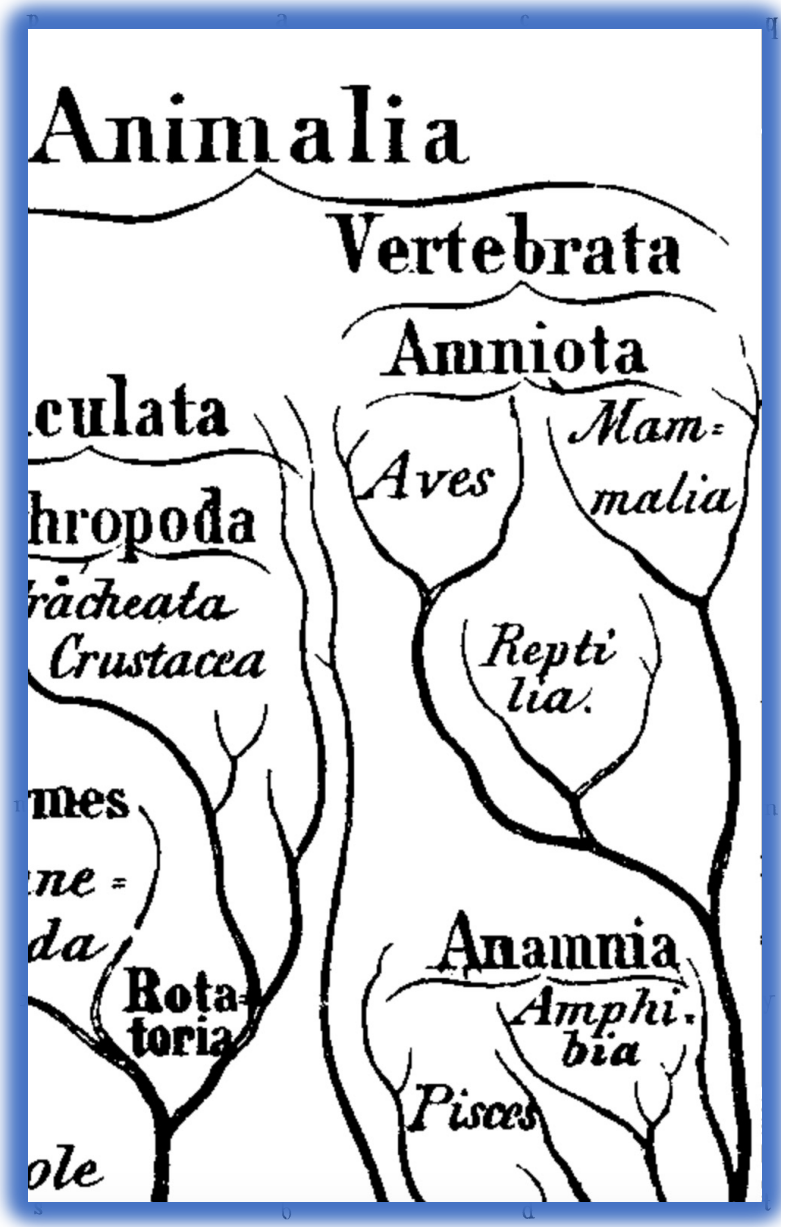
By CHARLES DARWIN, M.A.,

FELLOW OF THE ROYAL, GEOLOGICAL, LINNEAN, ETC., SOCIETIES;
AUTHOR OF 'JOURNAL OF RESEARCHES DURING H. M. S. BEAGLE'S VOYAGE ROUND THE WORLD.'

LONDON:
JOHN MURRAY, ALBEMARLE STREET.
1859.

The right of Translation is reserved.

Phylogenetic tree suggested by Haeckel (1866) →



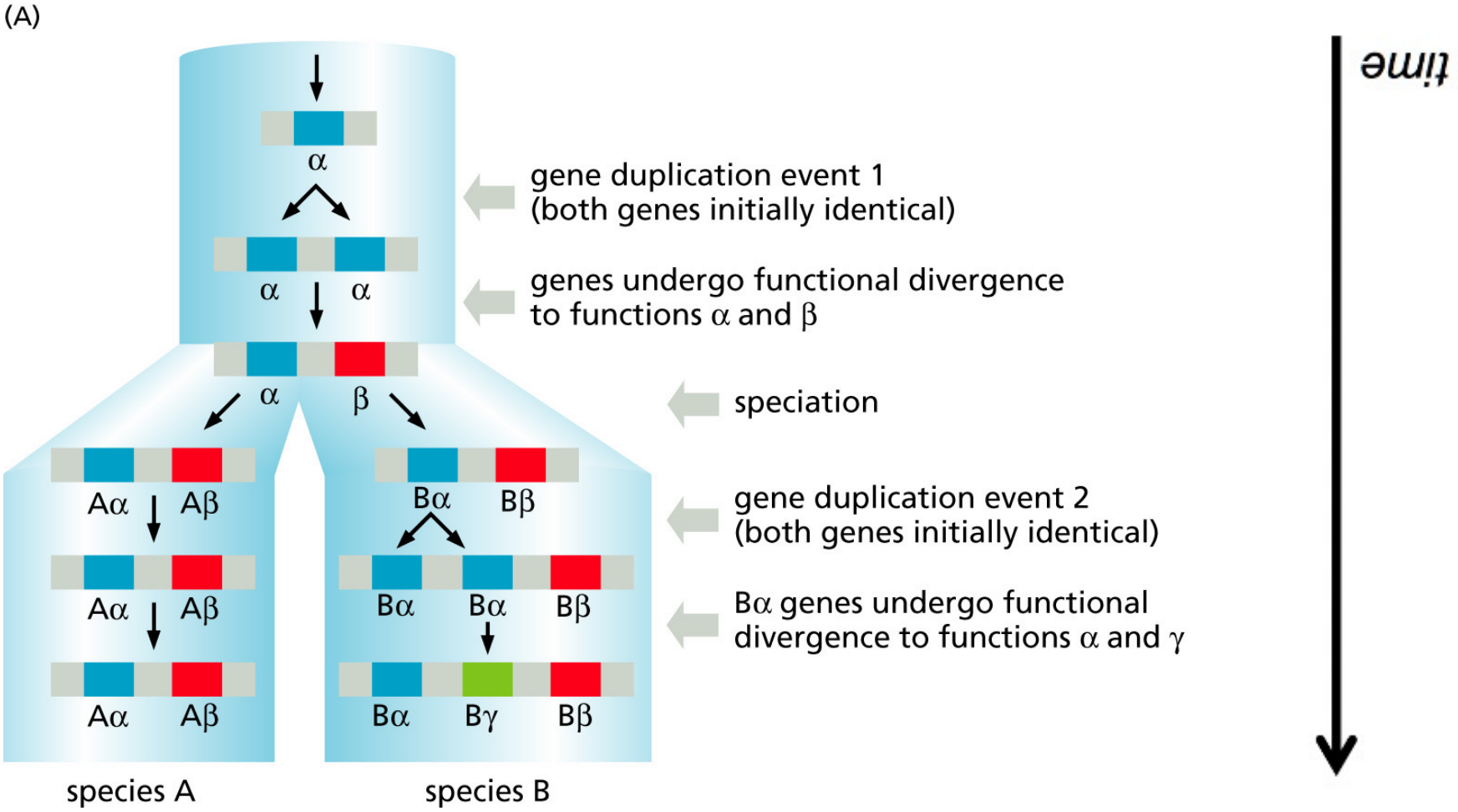
Homology, orthology and paralogy

Homology means descent from a common ancestor

Orthology is a special case of homology, in which the descent of sequences from a common ancestral form is topologically congruent with organismal speciation

Paralogy is a special case of homology, in which the sequences from a common ancestral sequence arose through duplication

Good reads:
Koonin EV (2001) *Genome Biol* 2(4); https://www.ncbi.nlm.nih.gov/pmc/articles/PMC138920/
Jensen RA (2001) *Genome Biol* 2(8); https://www.ncbi.nlm.nih.gov/pmc/articles/PMC138949/



Zvelebil & Baum, *Understanding Bioinformatics* Figure 7.10, p 242.

Part 2: Phylogenetics introduction

Sequence alignment

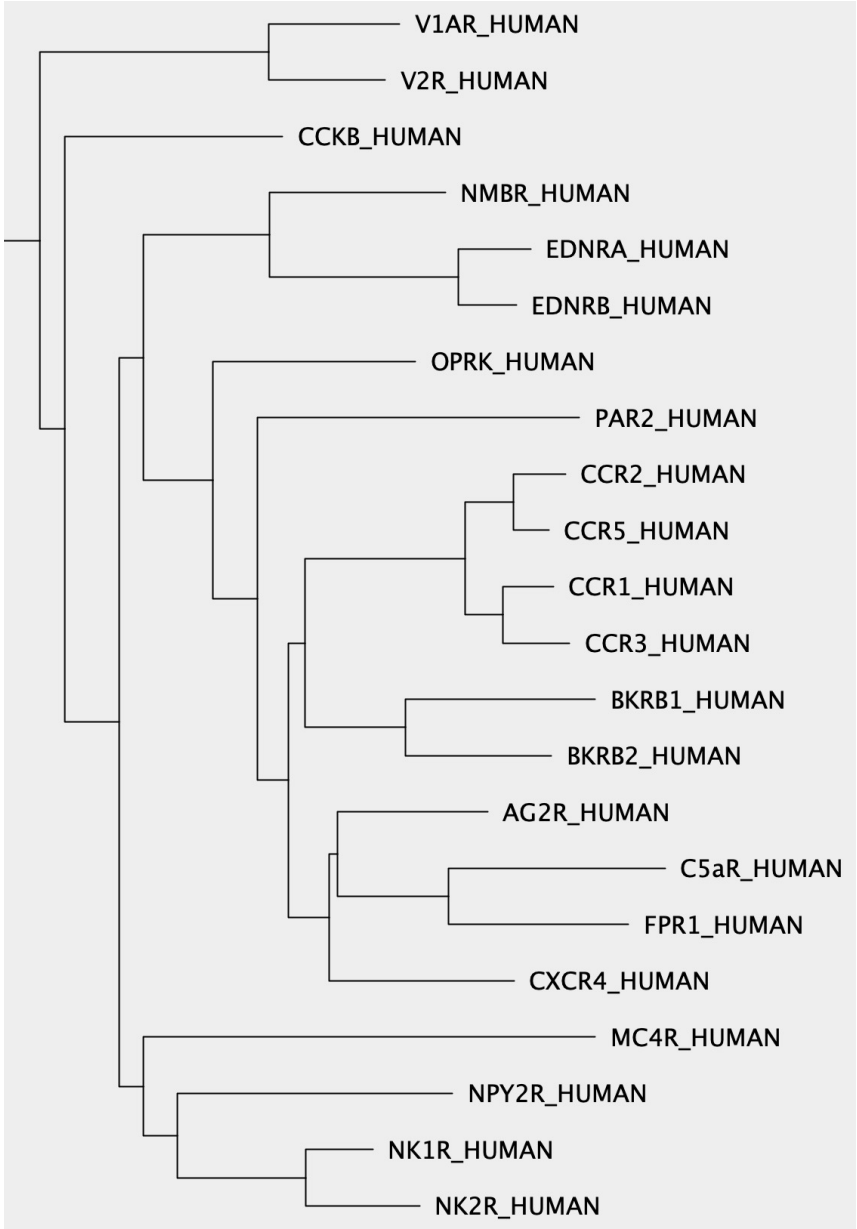
Phylogenetic trees, including terminology

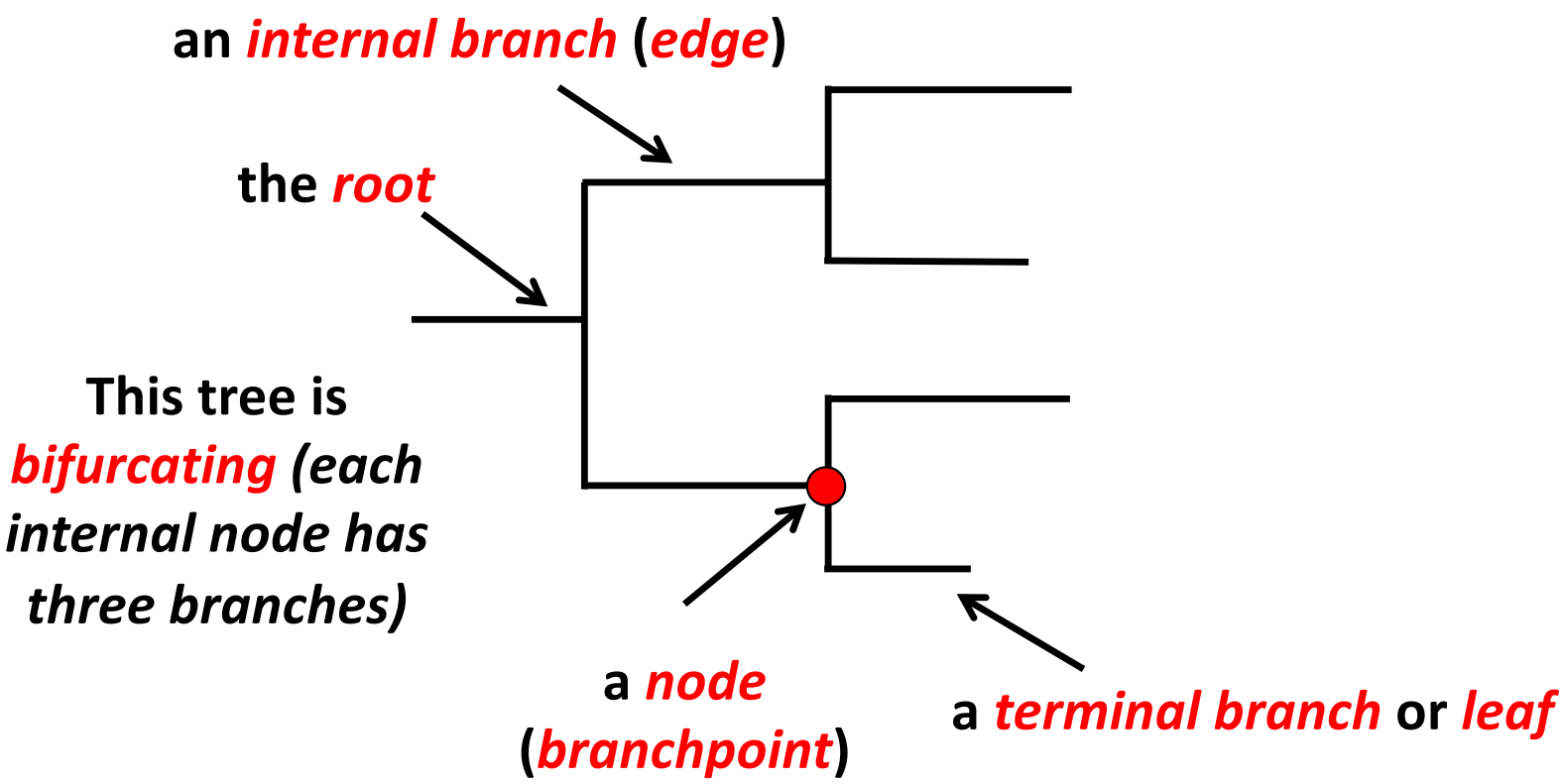
Species v. gene trees

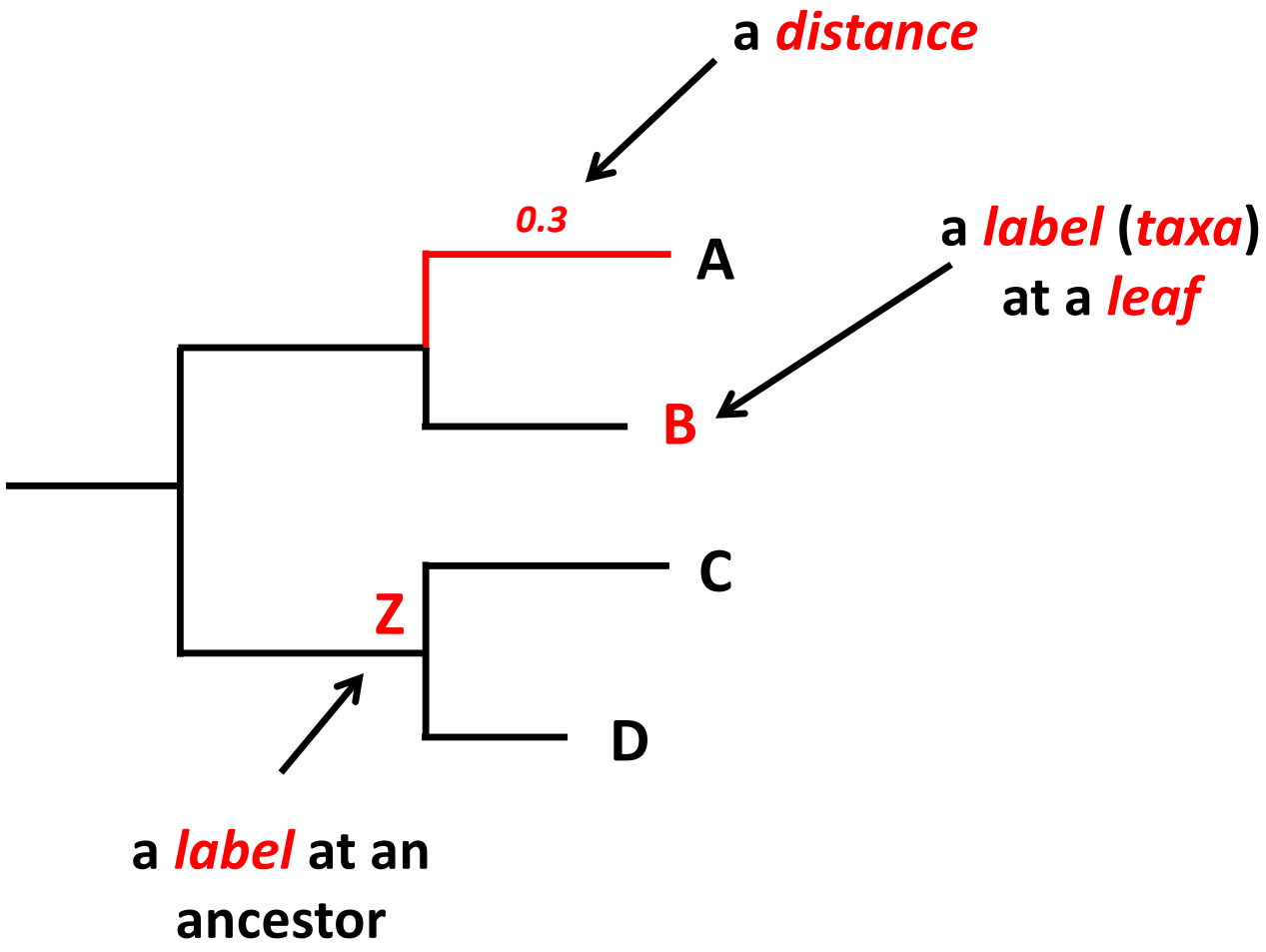
Multiple sequence alignment is input for phylogenetic analysis

Trees

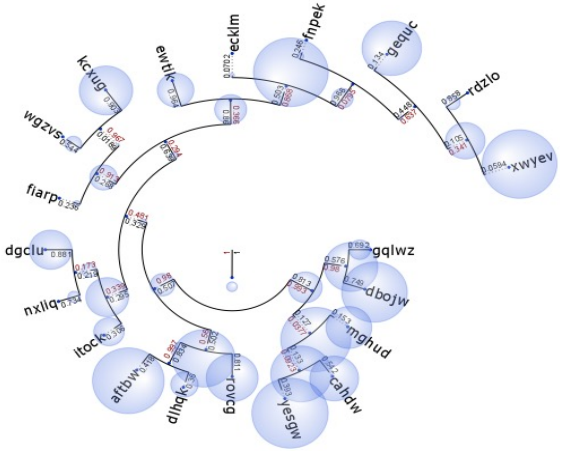
- Trees are connected graphs that do not contain circuits
- Trees are important as data structures
- A phylogenetic tree is a hypothesis about the evolutionary relationships between a set of objects (taxa); interpretation is subject to the type of data

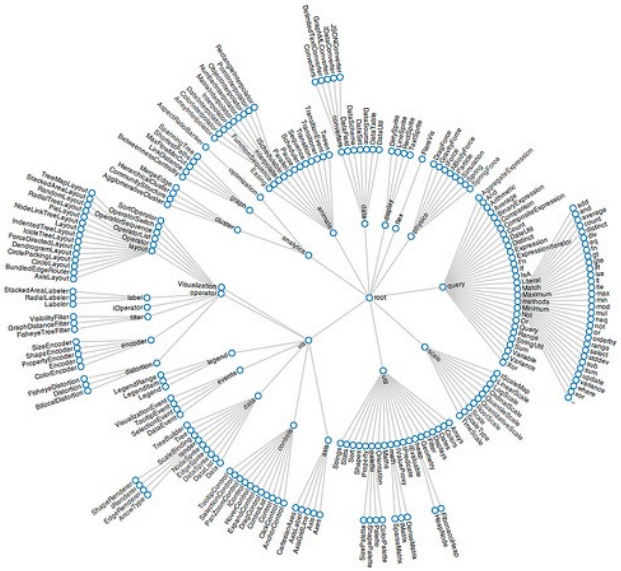


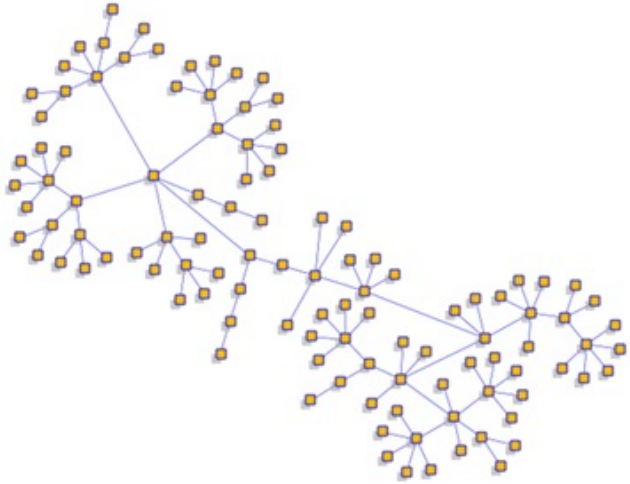

an internal branch (edge)
the root
This tree is bifurcating (each internal node has three branches)
a node (branchpoint)
a terminal branch or leaf

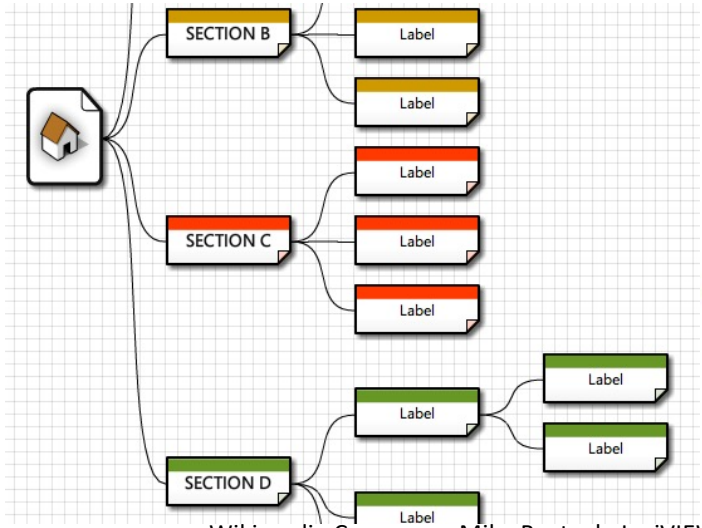


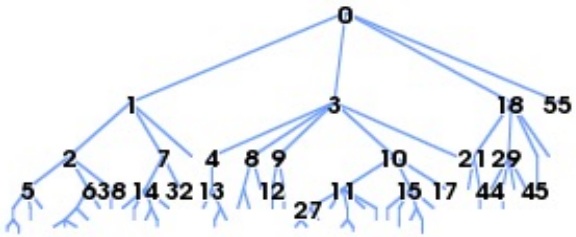
The vertical lines are NOT edges – they're purely for the viewer's convenience

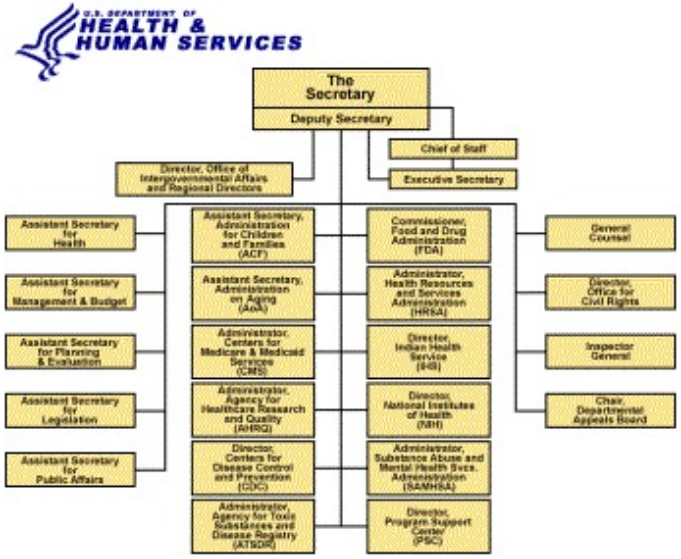








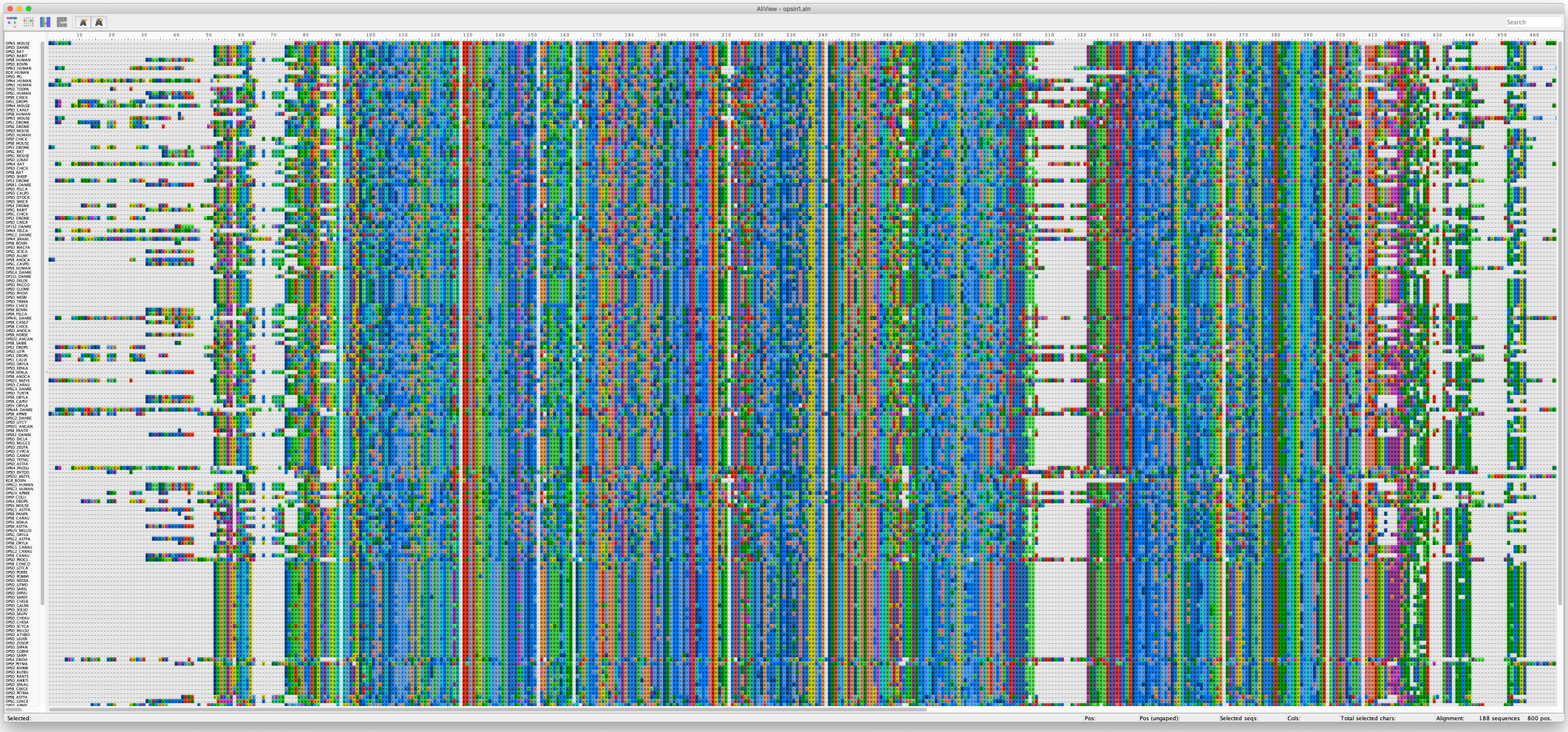


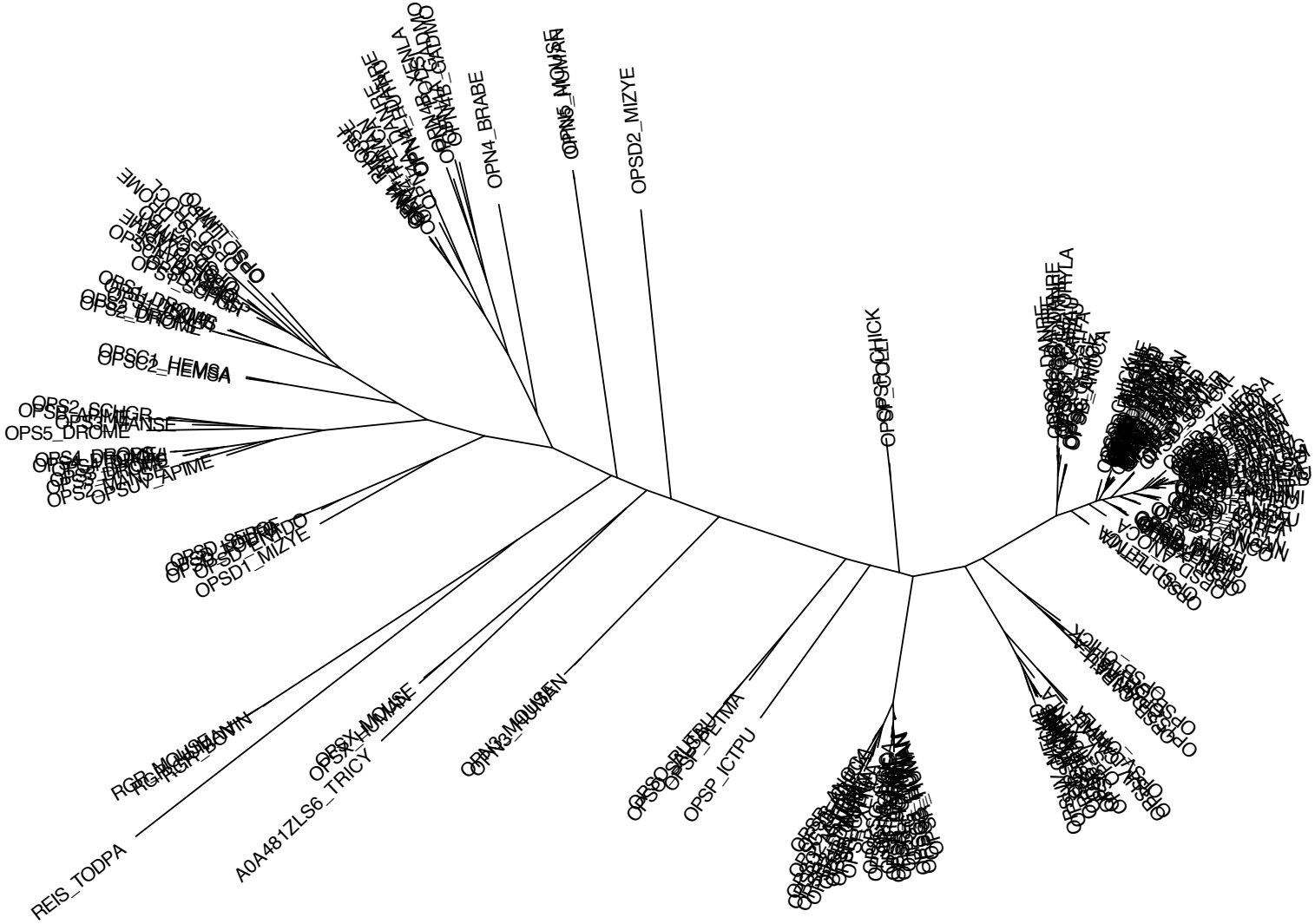


Wikimedia Commons; Mike Bostock; InviVIEW; LovelyCharts; Jaime Huerta-Cepas & Toni Gabaldón; US Dept Health & Human Services

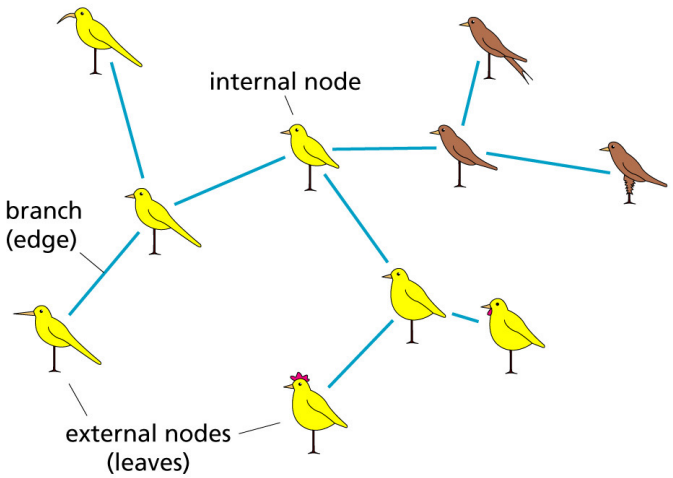
Opsin is a GPCR-1 sub-family

- Bilateria (16464 results)
  - Chordata (14310 results)
    - Branchiostoma (10 results)
      - Branchiostoma belcheri (Amphioxus) (5 results)
      - Branchiostoma floridae (Florida lancelet) (Amphioxus) (4 results)
      - Branchiostoma lanceolatum (Common lancelet) (Amphioxus lanceolatum) (1 results)
    - Vertebrata (14300 results)
      - Cyclostomata (jawless vertebrates) (30 results)
      - Gnathostomata (jawed vertebrates) (14270 results)
  - Protostomia (2154 results)
- Cnidaria (4 results)
  - Pocilloporidae (3 results)
  - Tripedalia cystophora (Jellyfish) (1 results)


AliView - opsin1.aln
Search
Selected:
Pos:
Pos (ungaped):
Selected seqs:
Cols:
Total selected chars:
Alignment:
188 sequences
800 pos.

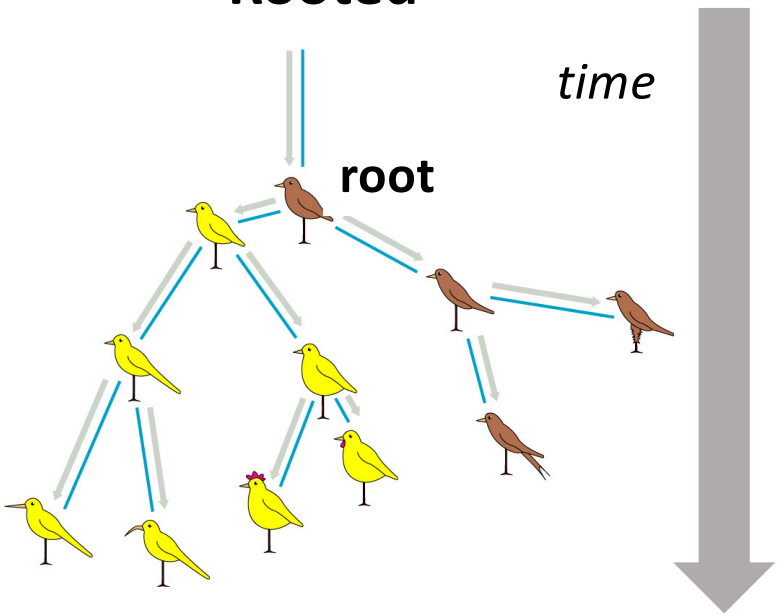

OPN4_BRABE
OPSD2_MIZYE
OPSD_CHICK
OPSP_ICTPU
OPSD_PETMA
OPSX_MOUSE
OPN3_MOUSE
A0A481ZLS6_TRICY
RGR_MOUSE
REIS_TODPA
OPSD_HEMSA
OPSU_APIME
OPS5_DROME

Unrooted



Distances are undirected

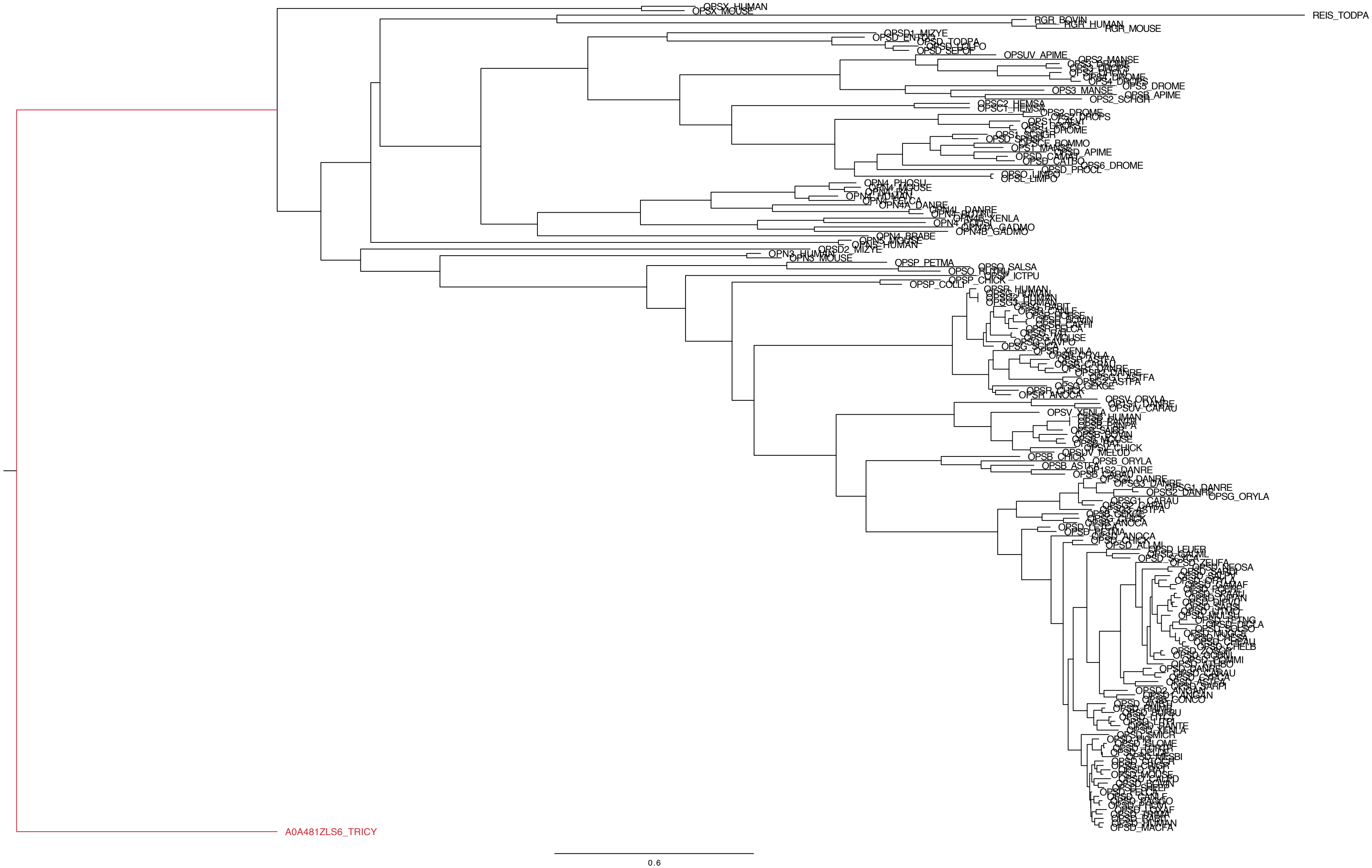
Rooted

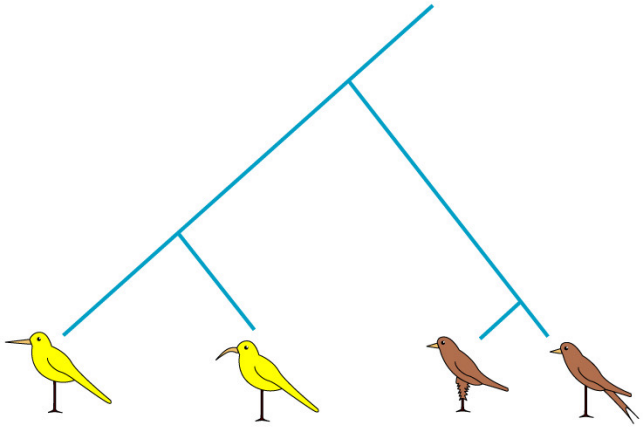


Distances are directed

Zvelebil & Baum, *Understanding Bioinformatics* p 227.

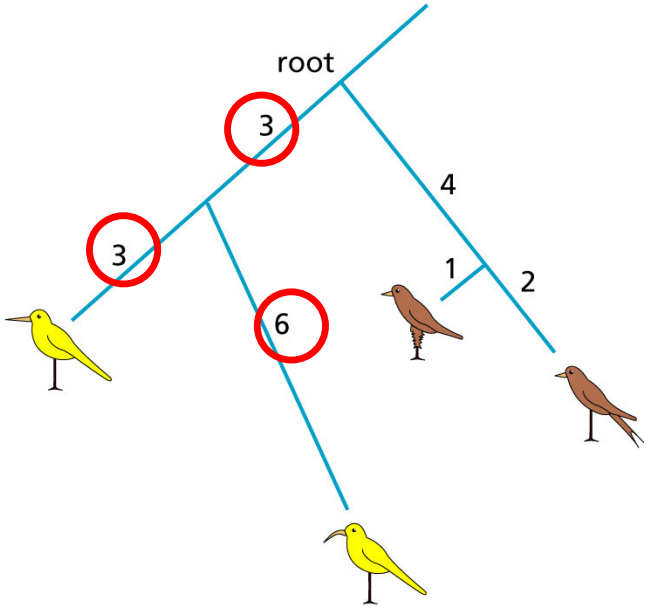

OPSD_FELCA
OPSD_CHICK
OPSD_PETMA
OPSB_CHICK
OPSB_ORYLA
OPSV_ORYLA
OPSV_CANAU
OPSV_MELUD
OPSB_HUMAN
OPSB_MOUSE
OPSB_CAVPO
OPSB_ANOCA
OPSP_ICTTU
OPSP_PETMA
OPSO_SALSA
OPSO_HOTHU
OPN3_HUMAN
OPN3_MOUSE
OPSX_HUMAN
OPSX_MOUSE
A0A481ZLS6_TRICY
REIS_TODPA
RGR_BOVIN
RGR_HUMAN
RGR_MOUSE
OPSD1_MIZYE
OPSD_TODPA
OPSD_SEPOF
OPSUV_APIME
OPS3_MANSE
OPS3_DROME
OPS2_SCHGR
OPSC2_HEMSA
OPSC1_HEMSA
OPS1_DROME
OPSD_PROGL
OPSD_LIMPO
OPN4_RHOSU
OPN4_MOUSE
OPN4_HUMAN
OPN4A_DANRE
OPN4B_DANRE
OPN4_GADMO
OPN4_BRABE
OPSD2_MIZYE
OPSP_CHICK
OPSP_COLLI
OPSG4_DANRE
OPSG3_DANRE
OPSG2_DANRE
OPSG_ORYLA
OPSG1_CANFA
OPSB_ANOCA
OPSD_CHICK
OPSD_ALLMI
OPSD_MACFA
0.6


OPSX_HUMAN
OPSX_MOUSE
REIS_TODPA
RGR_BOVIN
RGR_HUMAN
RGR_MOUSE
OPSD1_MIZYE
OPSD_TODPA
OPSD_SEPOF
OPSUV_APIME
OPS2_MANSE
OPS3_MANSE
OPS5_DROME
OPSB_APIME
OPS2_SCHGR
OPSC2_HEMSA
OPSC1_HEMSA
OPS2_DROME
OPS2_DROPS
OPS1_DROME
OPS1_SCHGR
OPSD_SPHSP
OPSG_APIME
OPSD_CAMHU
OPS6_DROME
OPSD_PROCL
OPSD_LIMPO
OPSL_LIMPO
OPN4_PHOSU
OPN4_MOUSE
OPN4_HUMAN
OPN4A_DANRE
OPN4L_DANRE
OPN4B_XENLA
OPN4A_GADMO
OPN4B_GADMO
OPN4_BRABE
OPN3_HUMAN
OPN3_MOUSE
OPSP_PETMA
OPSO_SALSA
OPSP_ICTPU
OPSP_CHICK
OPSP_COLLI
OPSR_HUMAN
OPSG3_HUMAN
OPSR_ANOCA
OPSV_ORYLA
OPSUV_CARAU
OPSV_XENLA
OPSB_HUMAN
OPSV_CHICK
OPSUV_MELUD
OPSB_CHICK
OPSB_ORYLA
OPSB_ASTFA
OPS2_DANRE
OPSG1_DANRE
OPSG2_DANRE
OPSG_ORYLA
OPSG1_CARAU
OPSD_ANOCA
OPSD_ALLMI
OPSD_LEUER
OPSD_NEOSA
OPSD_ORYLA
OPSD_CARAU
OPSD2_ANGAN
OPSD_CONCO
OPSD_XENLA
OPSD_MESBI
OPSD_MOUSE
OPSD_BOVIN
OPSD_CANLF
OPSD_HUMAN
OPSD_MACFA
A0A481ZLS6_TRICY
0.6



Cladogram

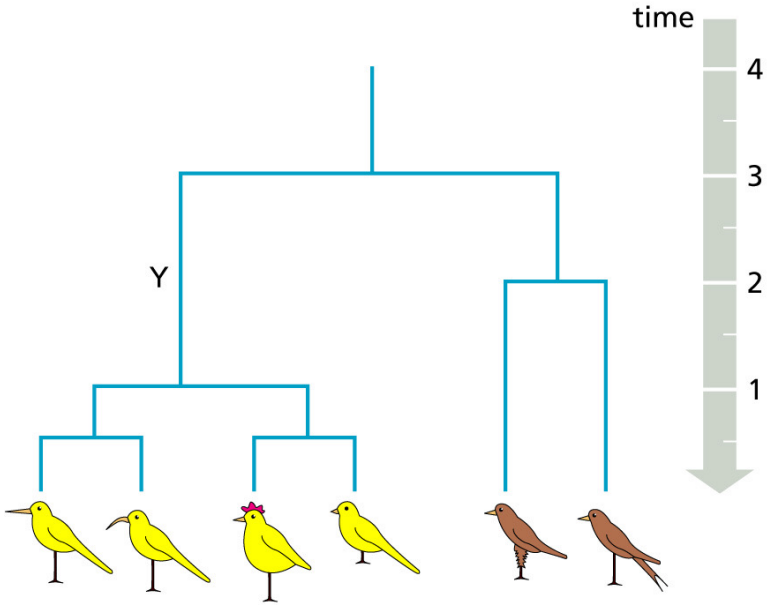
branch lengths have *qualitative*
but not *quantitative* meaning



Additive tree

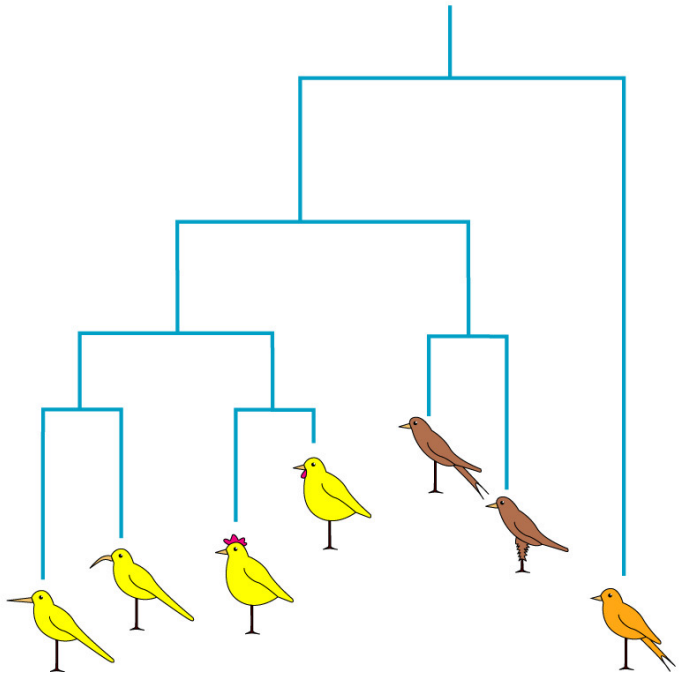
branches quantify the evolutionary
distance, the progress of divergence

Zvelebil & Baum, *Understanding Bioinformatics* p 229.



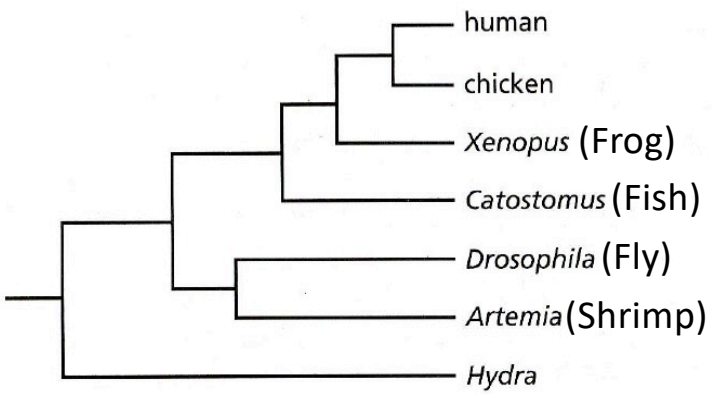
Ultrametric tree

(constant rate of mutation; *molecular clock*)



Additive tree with outgroup

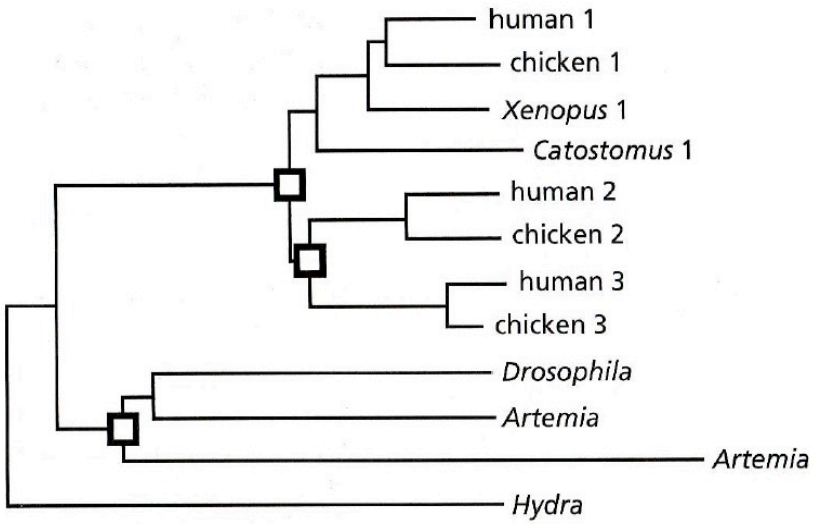
Zvelebil & Baum, *Understanding Bioinformatics* p 229.



Species tree

(evolution of species)

Ideally derived from **strictly orthologous** sequences

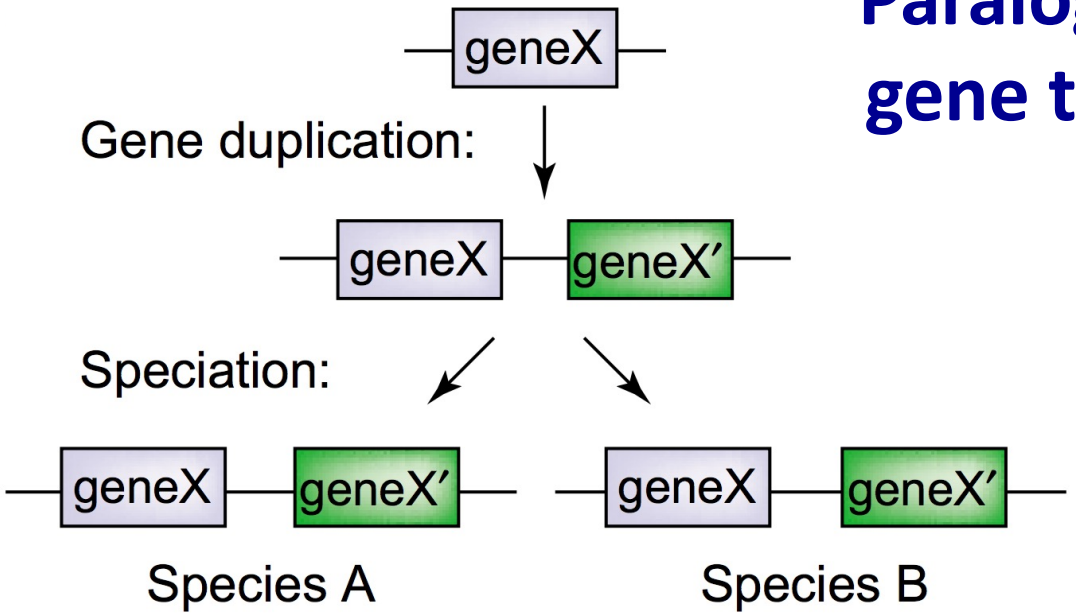


Gene tree

(evolution of homologous genes, in this case a particular family of membrane proteins)

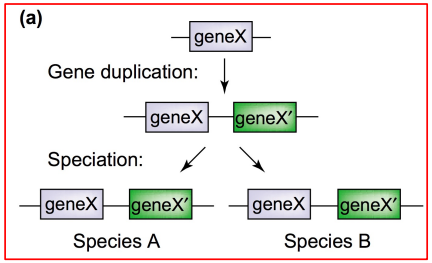
Zvelebil & Baum, *Understanding Bioinformatics* p 231.

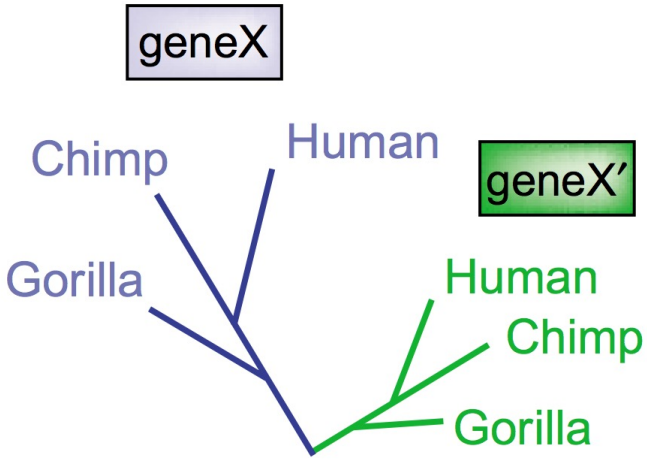
Paralogs in gene trees



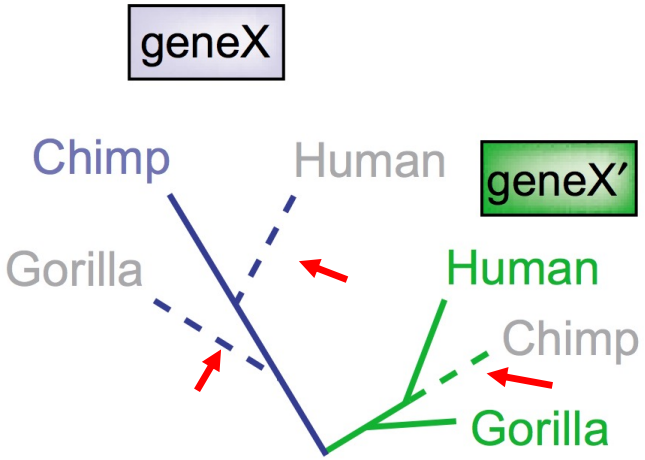
Gene X is duplicated prior to speciation. All subsequent species inherit both copies of the gene (unless one or the other is lost somewhere along the way).

Baldauf, SL (2003) *Trends in Genetics* 19: 345-351.





All sequences of gene X are **orthologs** of each other, and all the sequences of gene X' are **orthologs** of each other. However, X and X' are **paralogs**. Both the X and X' subtrees show the true relationships among the three species.



A tree of the X/X' gene family can be misleading if not all the sequences are included (because of **incomplete sampling** or **gene loss**). If the broken branches are missing, then the true species relationships are misrepresented.

Read more about this?

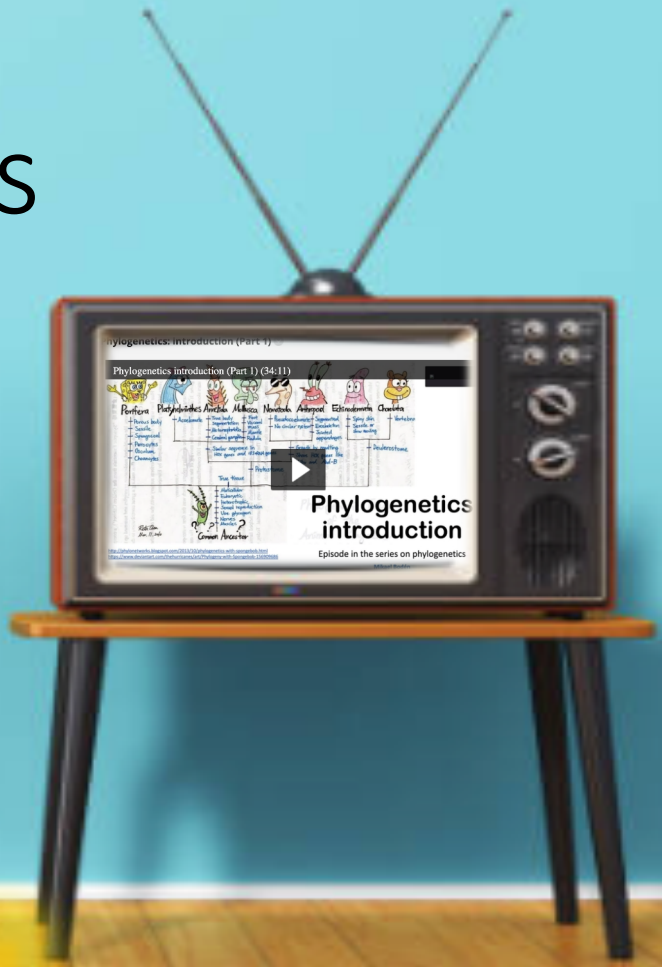
- Some of the material is based on Zvelebil and Baum's textbook "Understanding Bioinformatics" (chapters 7 and 8)

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

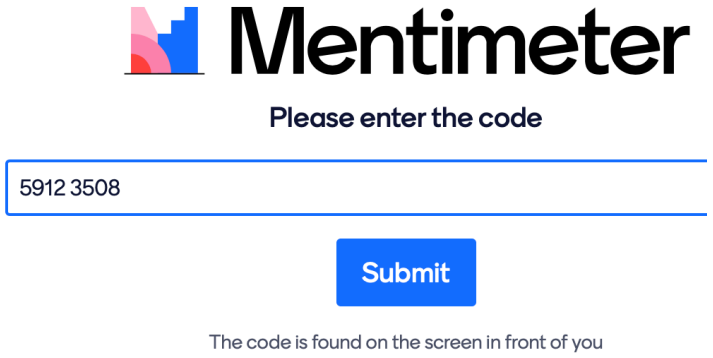
Program Director Bioinformatics

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

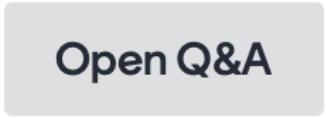
We're using **Mentimeter**

- 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

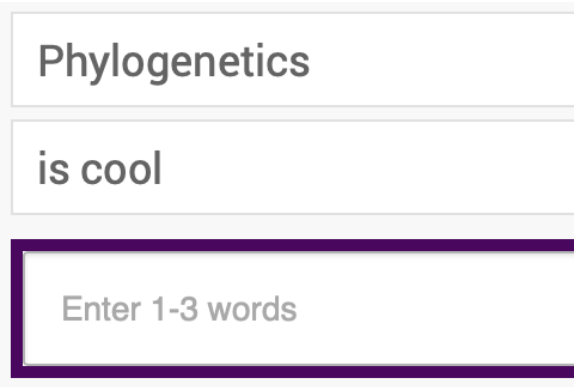
Or … **UQwordcloud**

- Go here on your phone/laptop/tablet

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

- Refresh browser when new question appears

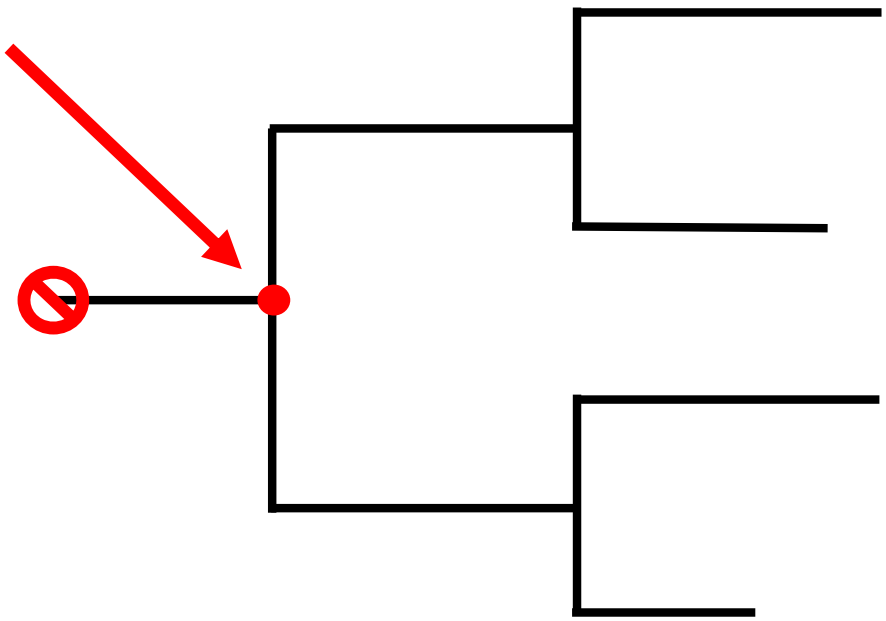
- Respond to question by …
  - typing a phrase, *or*
  - multiple phrases on separate rows



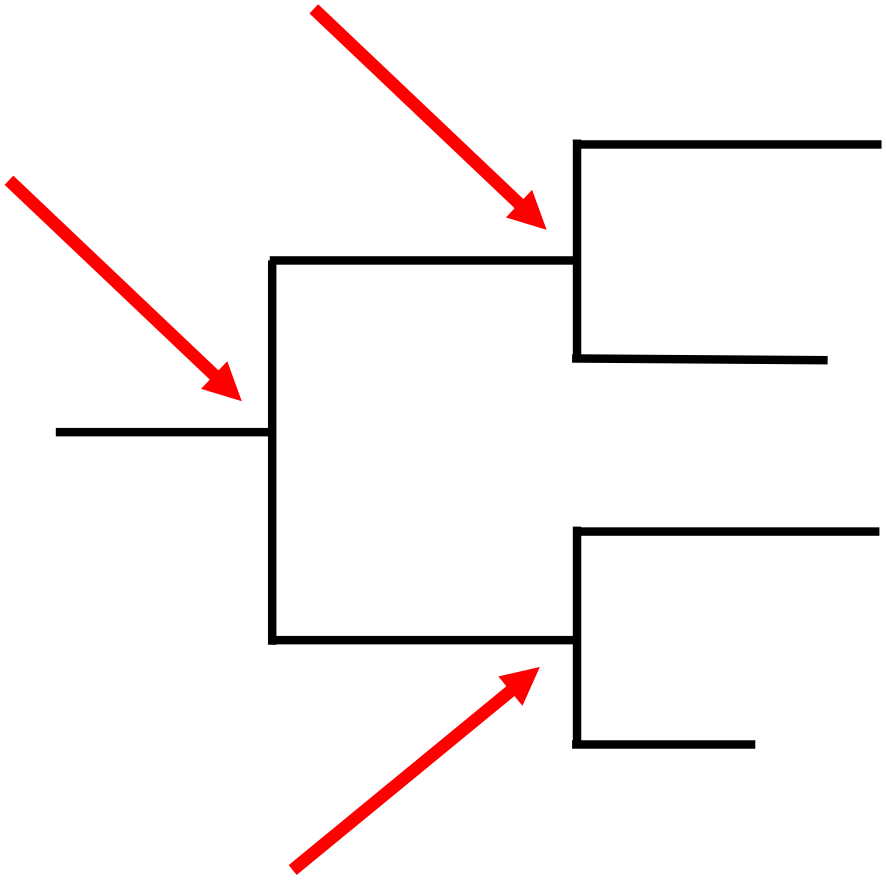
- I'll stop at a few points to check Zoom

Answers are not saved or used to identify you

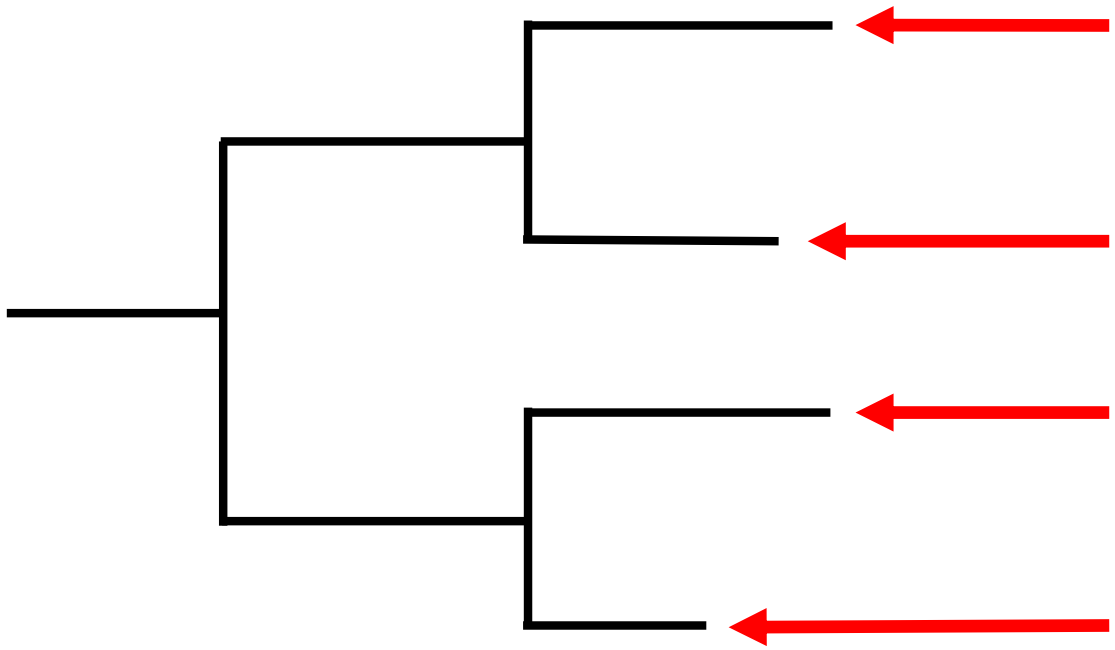
P1: What name would you use to refer to …



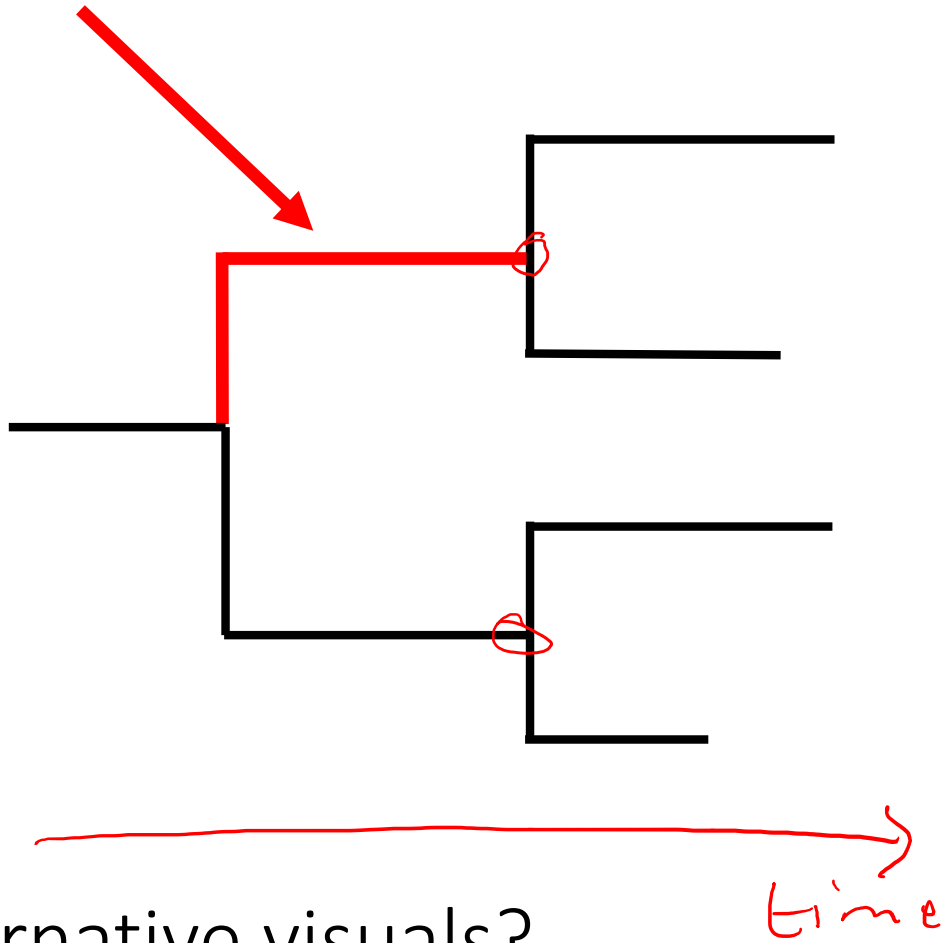
P2: What name would you use to refer to …



P3: What name would you use to refer to …

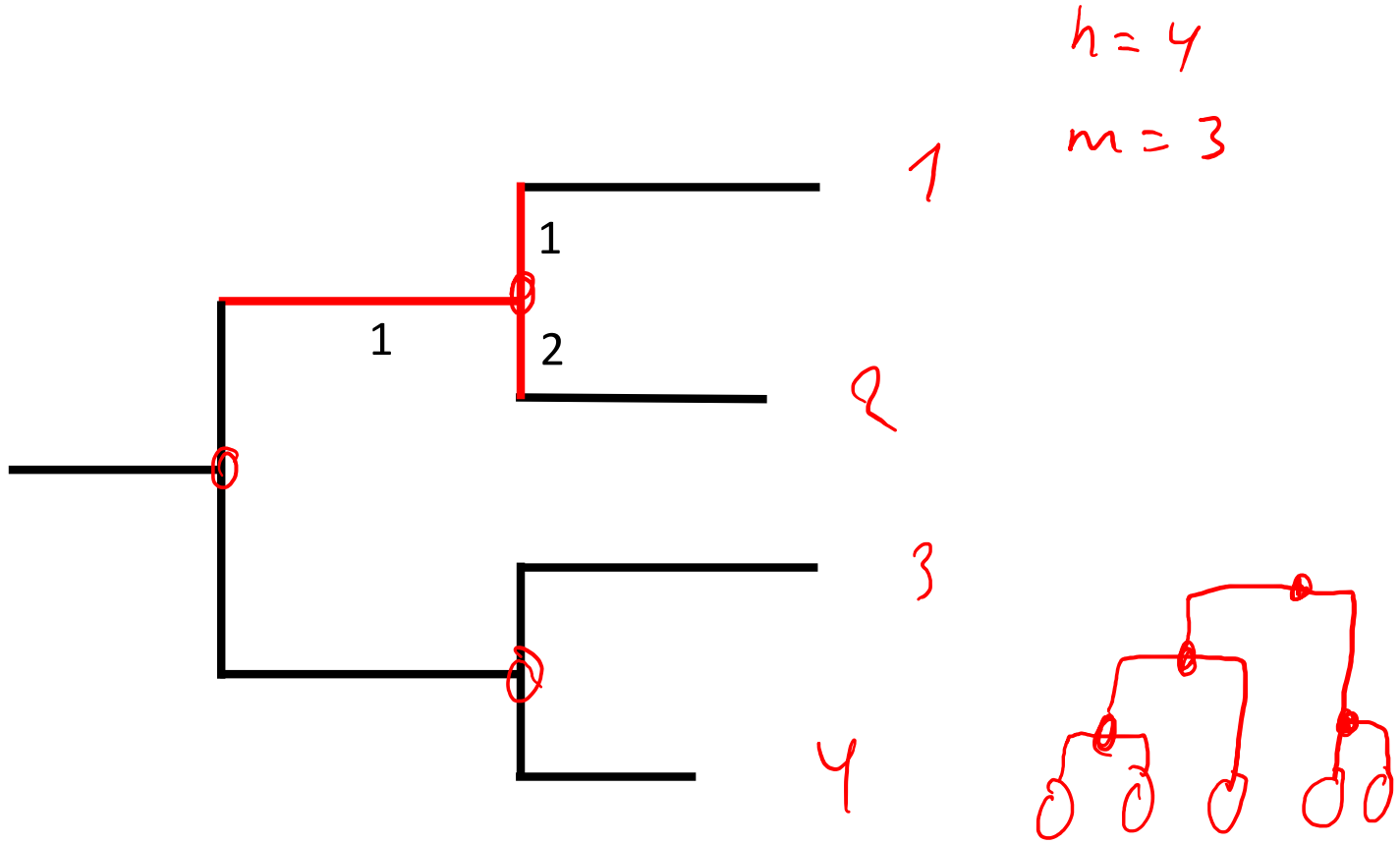


P4: What name would you use to refer to …



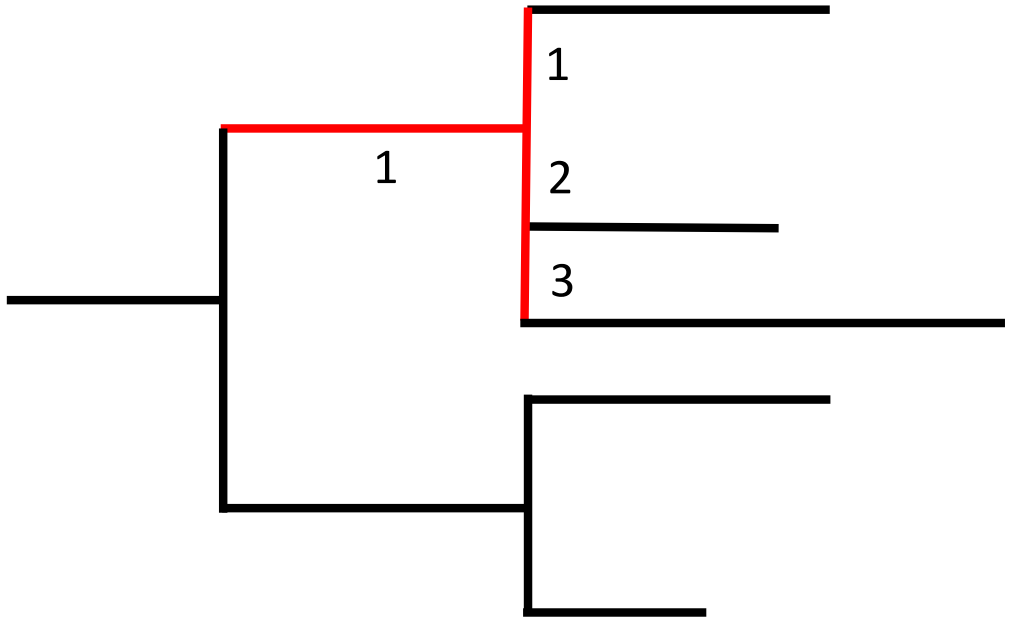
Discuss: alternative visuals?

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

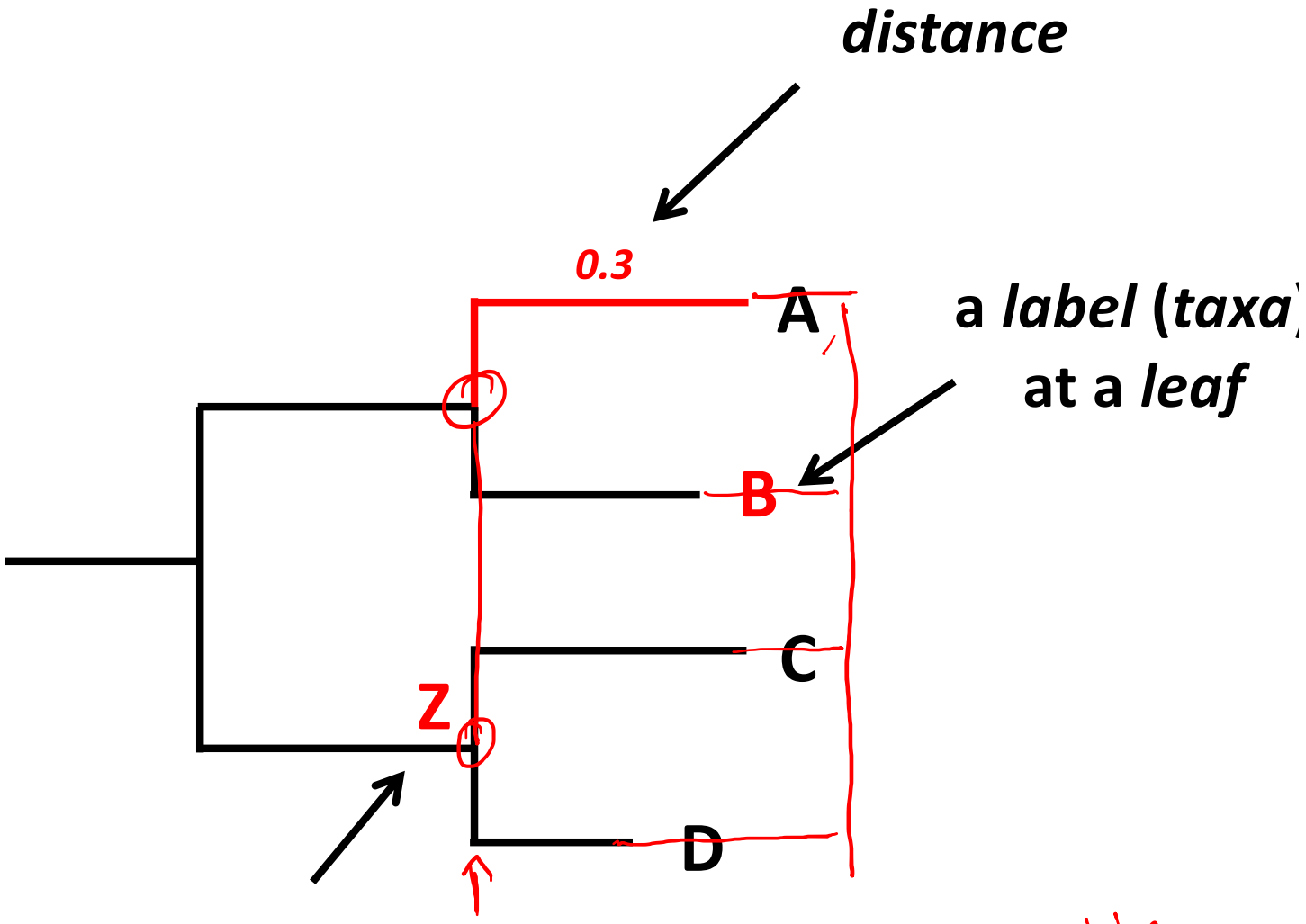


Discuss: number of internal nodes depends on number of leaves, how?

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.

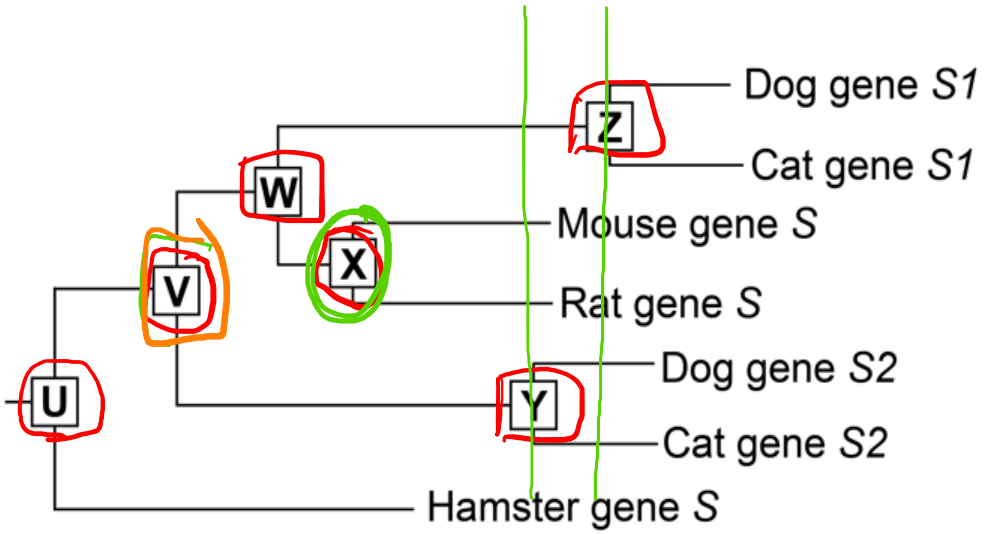
D Parent branch splits into two or more daughter branches at any given point.

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.



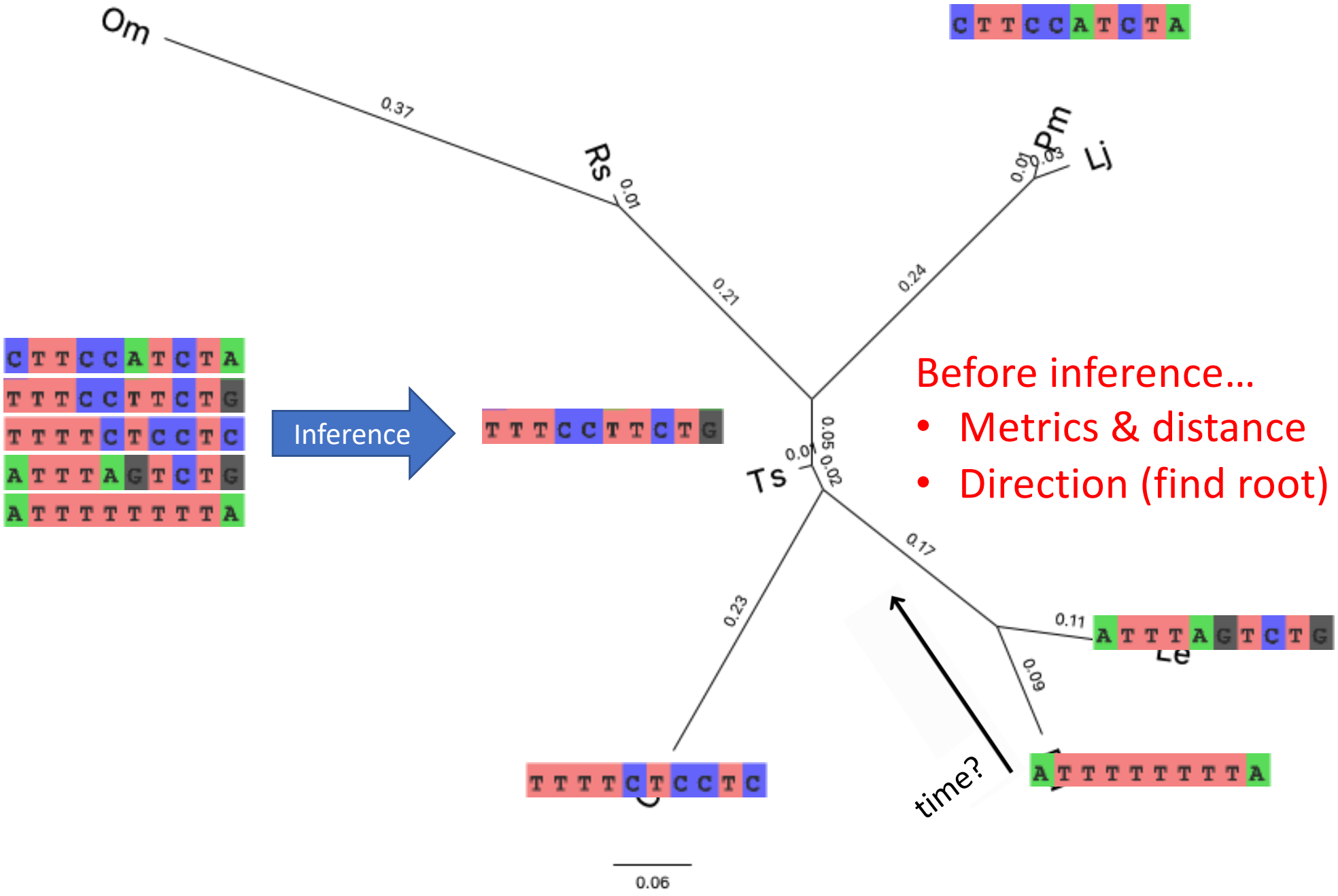
Which of the six nodes (U, V, W, X, Y or Z) describe the following two events?

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

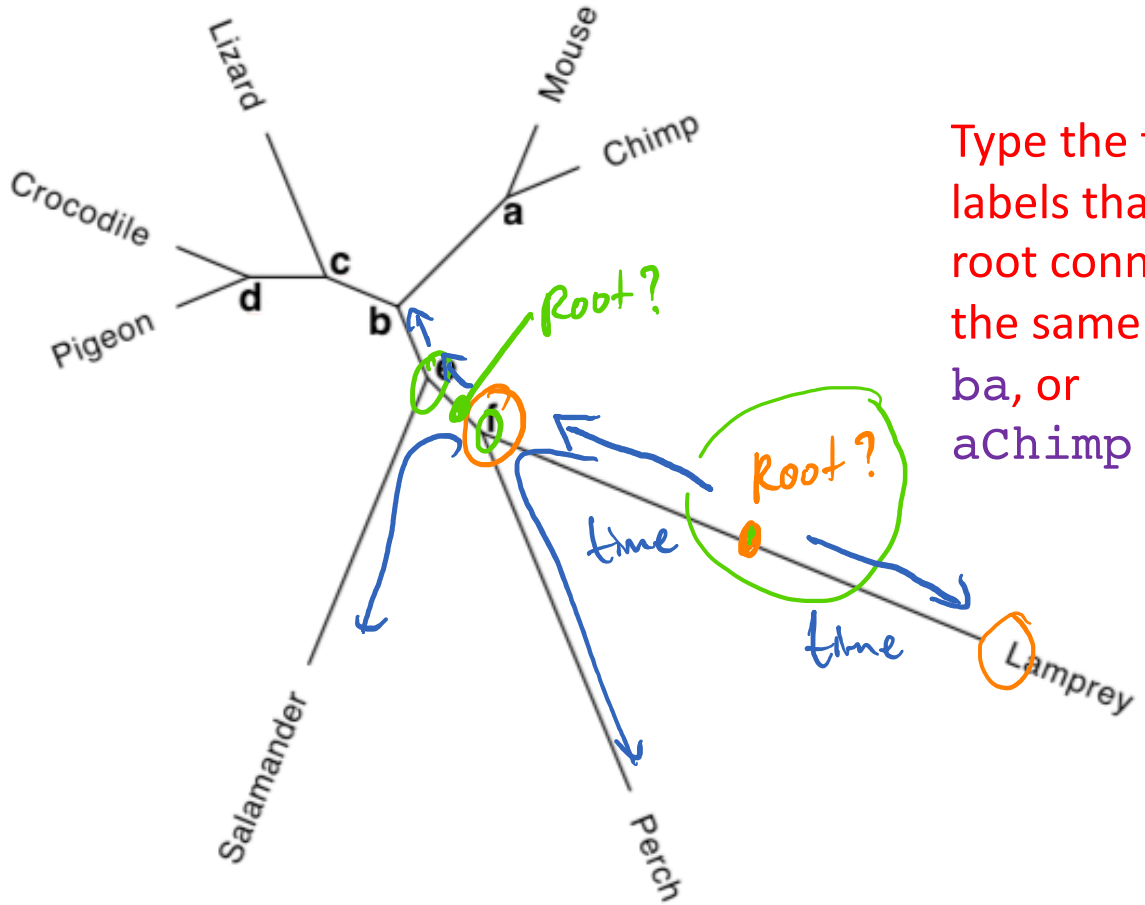
Questions so far?

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.

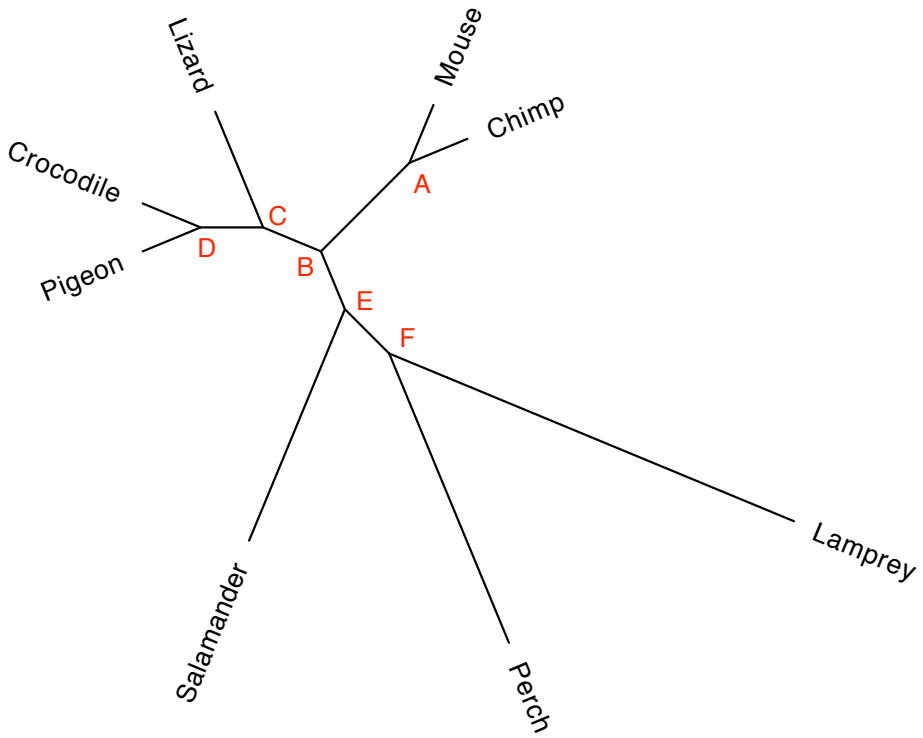


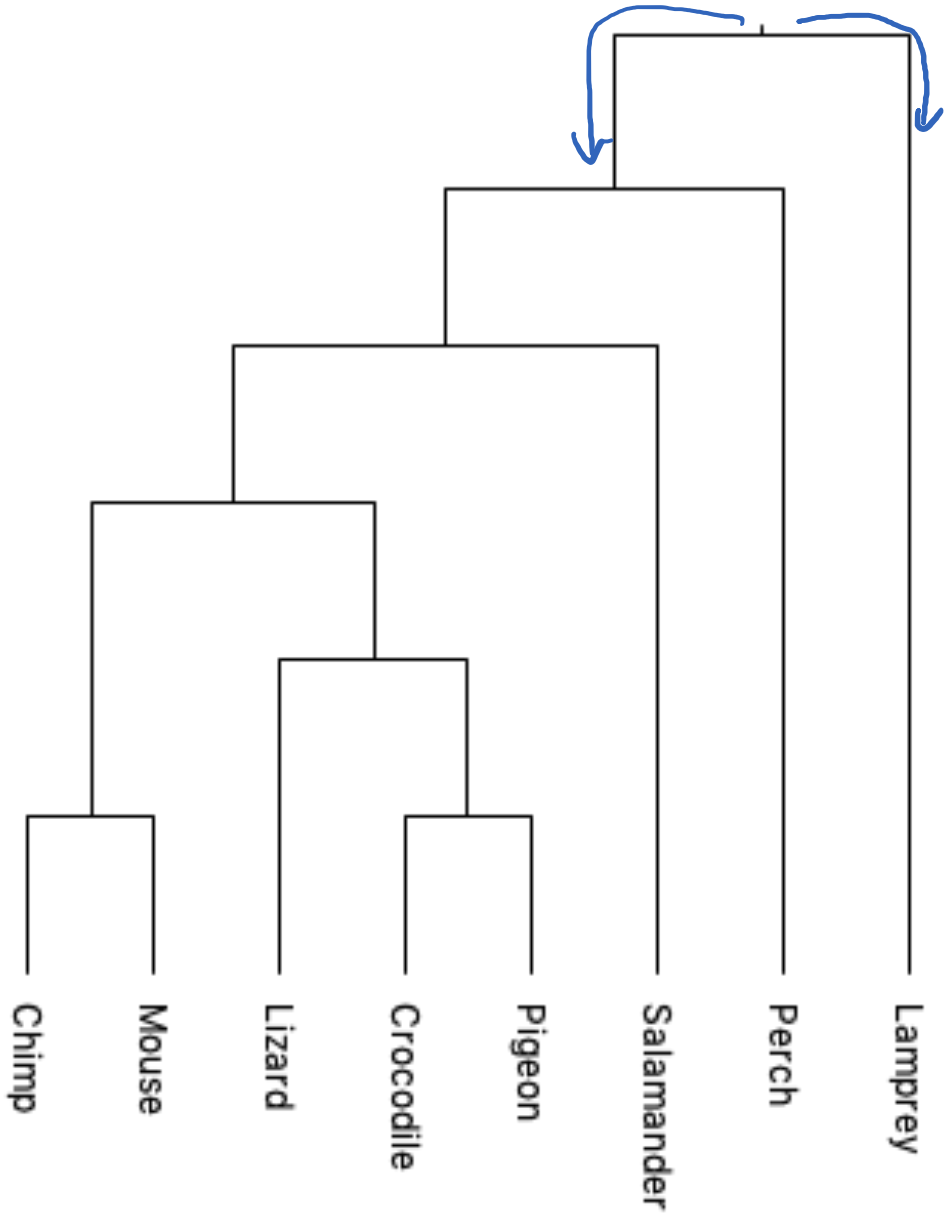
Type the two node labels that the root connects on the same row, e.g. `ba`, or `aChimp`

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

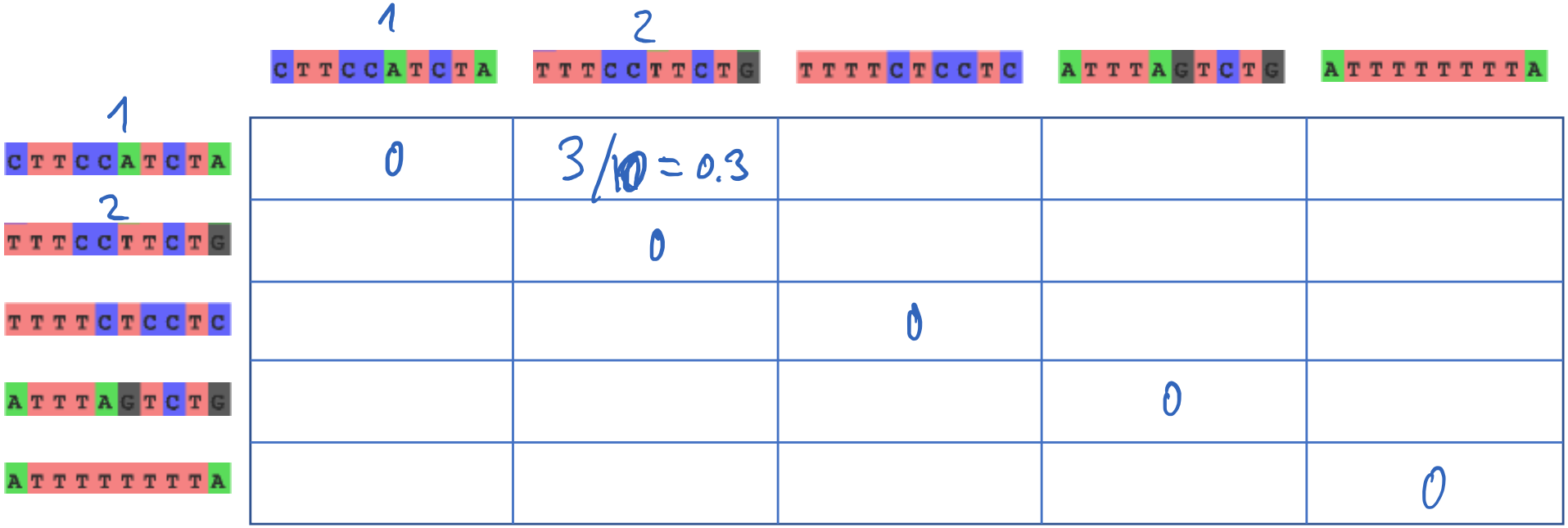
Type for example:
XX=0.00
ZY=0.78
But for the pairs missing

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

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

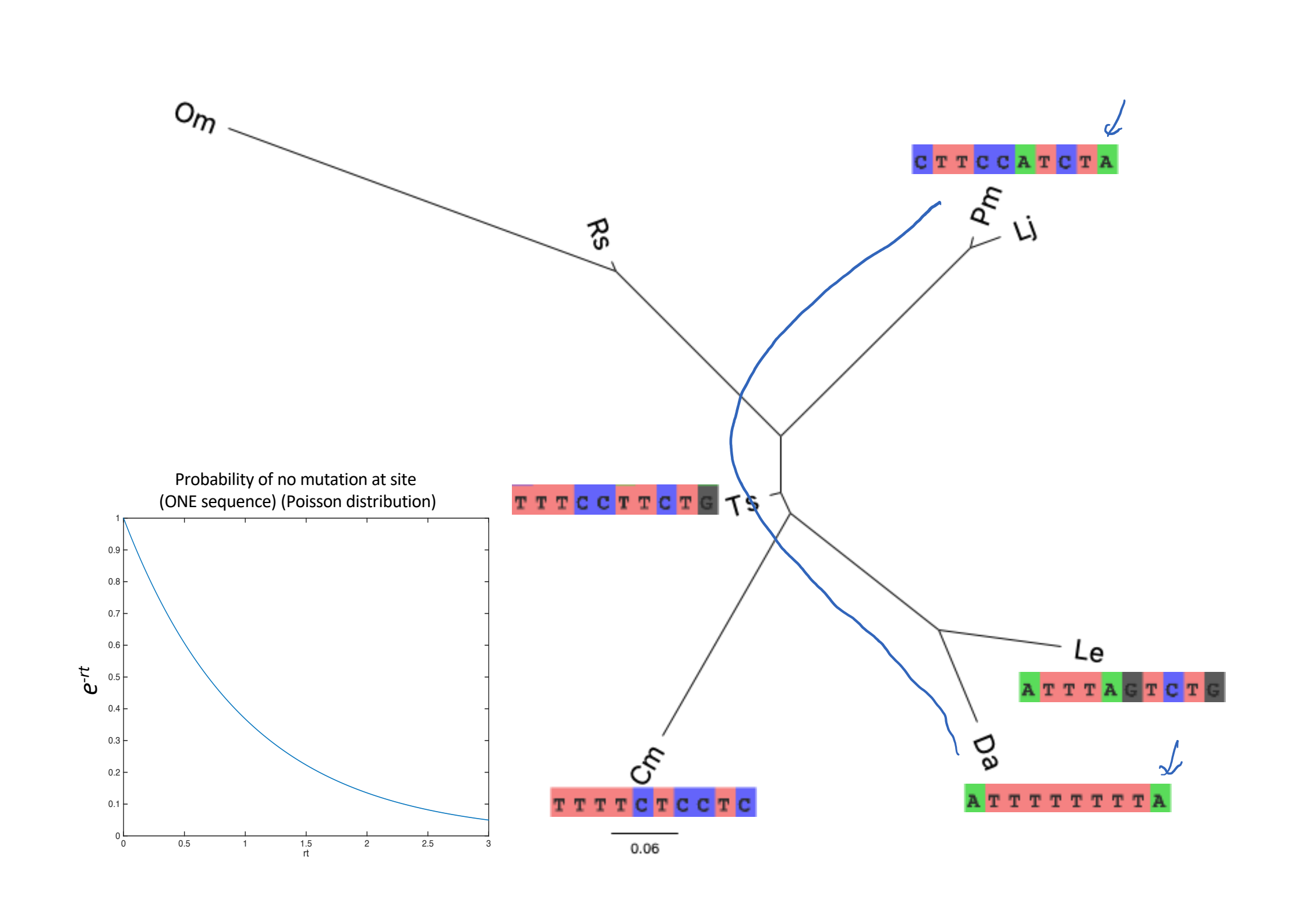
p-distance (aka fractional distance)

Distance matrix

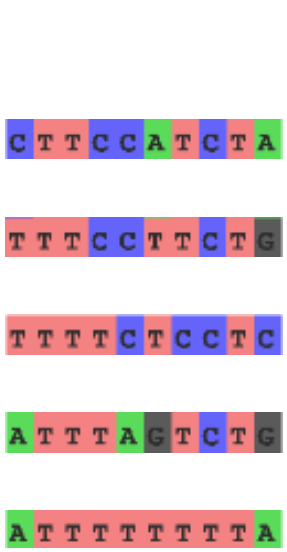


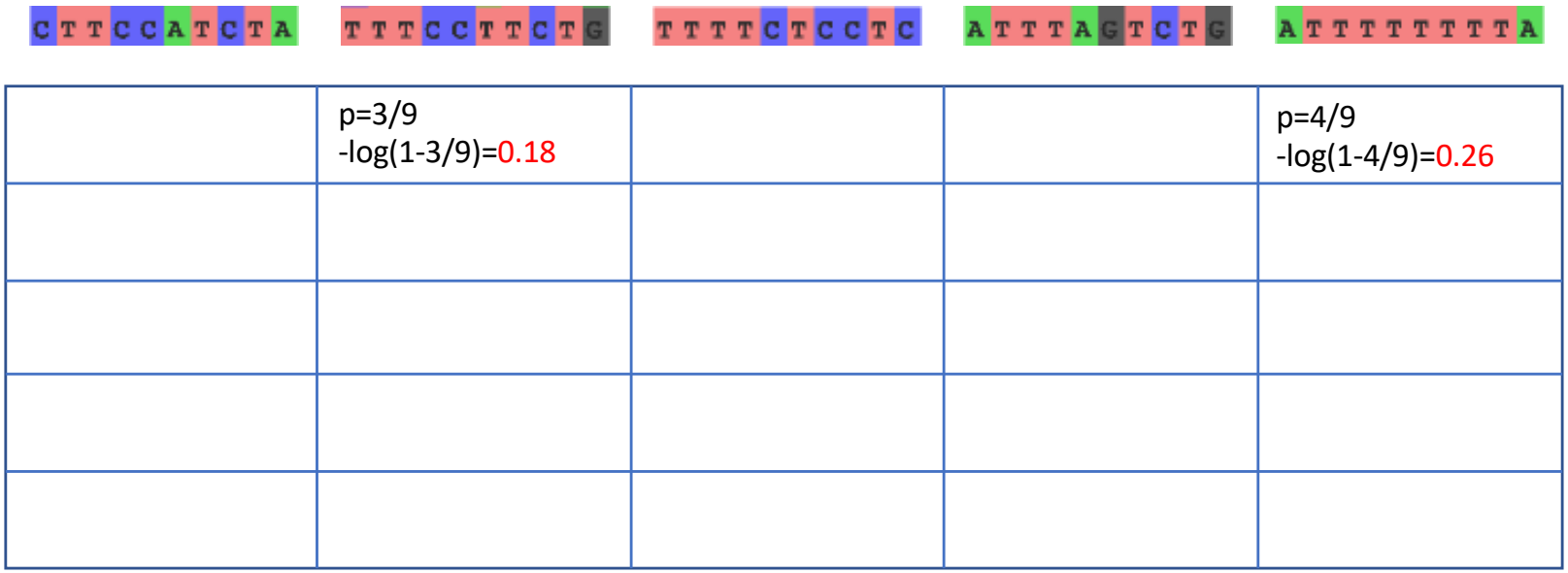
Two conditions for evolutionary time to be proportional to number of changes observed from an alignment:

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

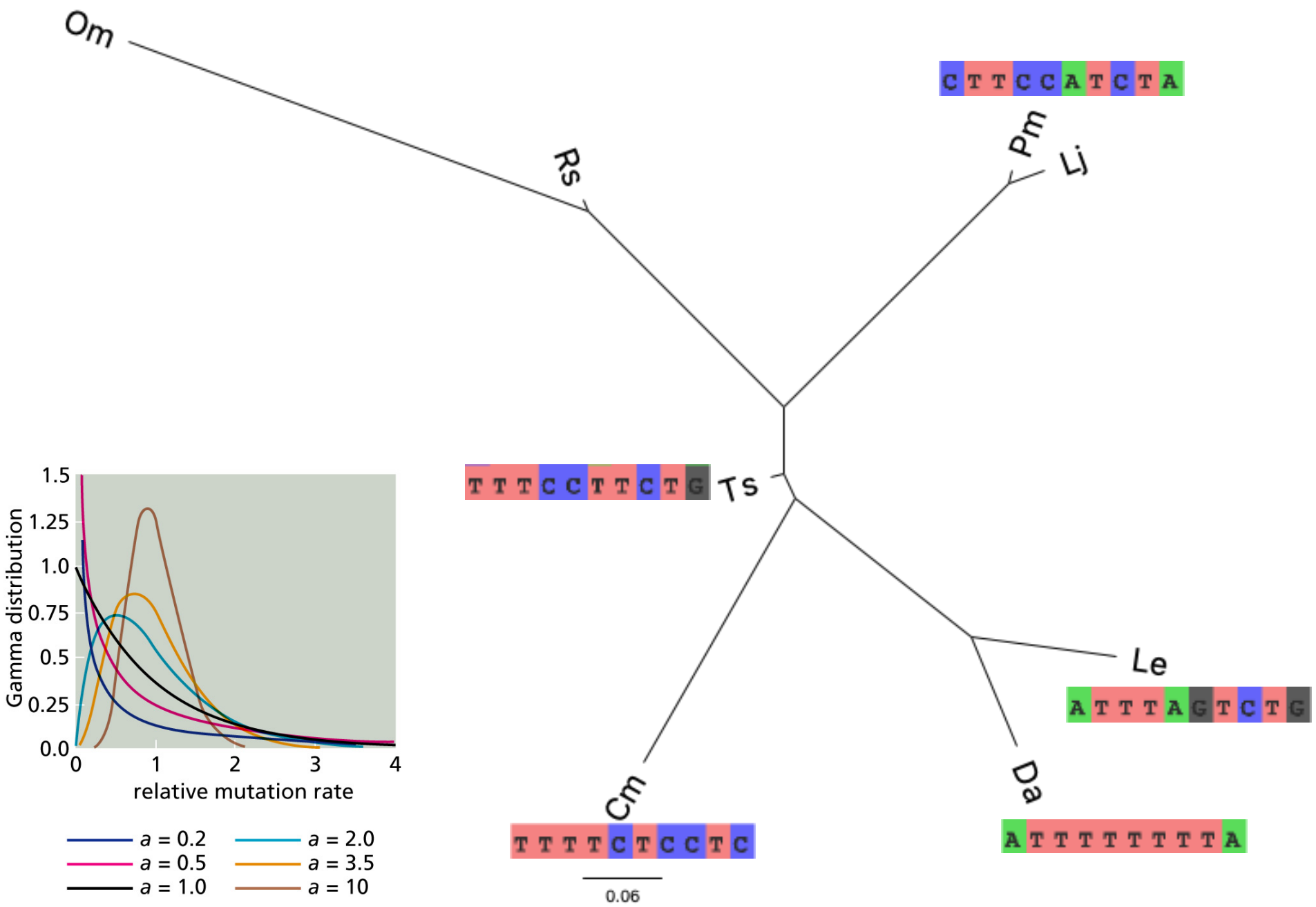

Om
Rs
Pm
Lj
C T T C C A T C T A
T T T C C T T C T G
Ts
Le
A T T T A G T C T G
Cm
T T T T C T C C T C
Da
A T T T T T T T T A
0.06
Probability of no mutation at site
(ONE sequence) (Poisson distribution)
e^{-rt}
rt
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.5
1
1.5
2
2.5
3

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



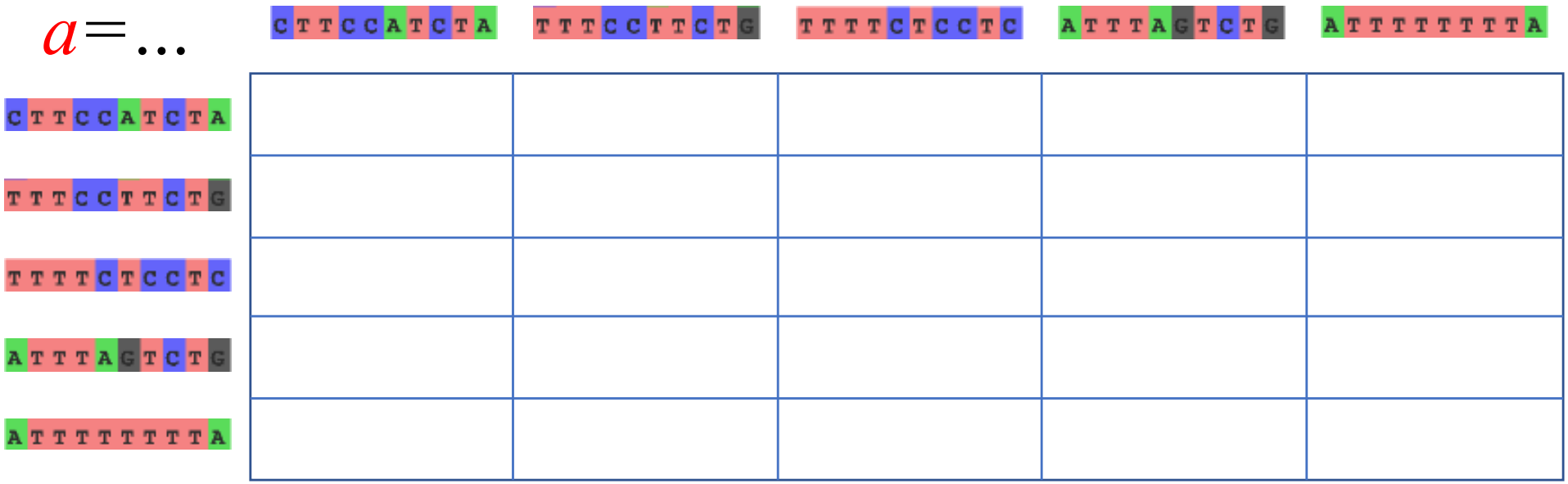


Accounts for multiple mutations at site


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

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



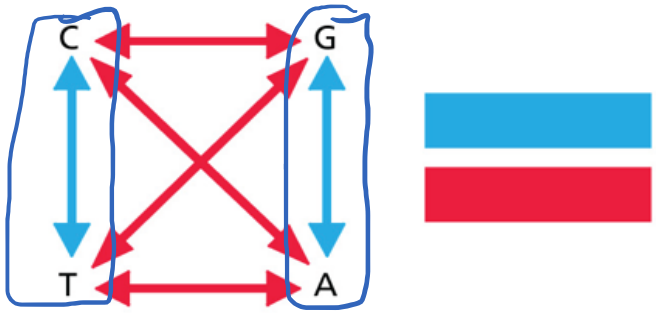
Corrects distance estimate for changes that can be explained by a variable rate

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).

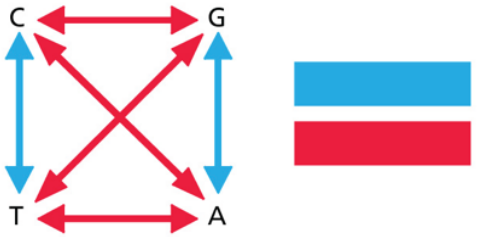


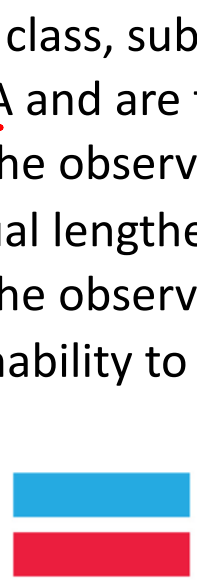
P10A: What name is used to refer to the blue class?

P10B: The red class?

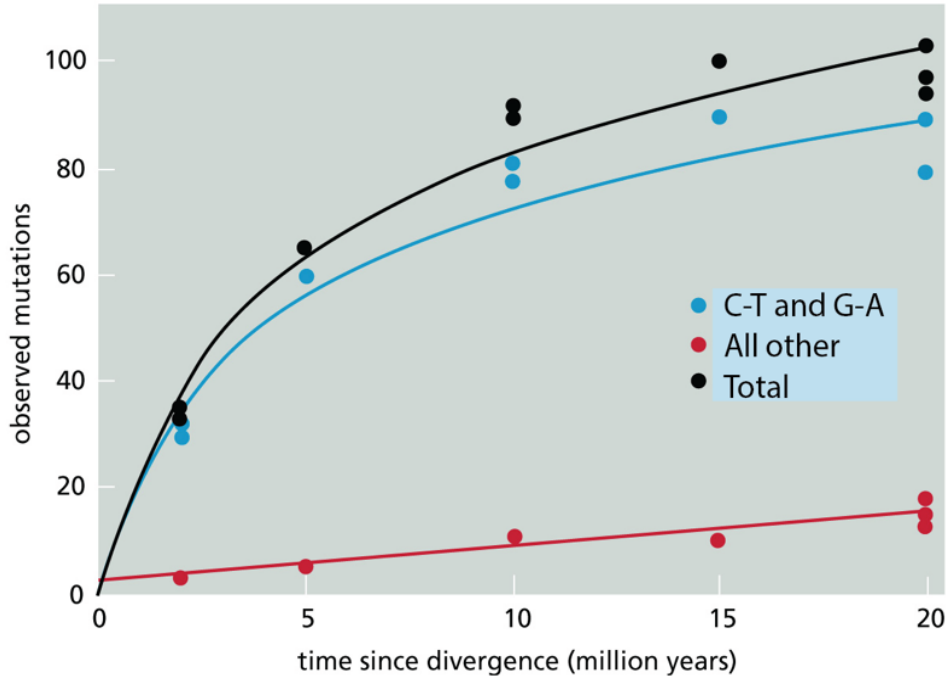
P10C: In the Figure below, observed mutations of the blue class surpass those of the red class over evolutionary time. Identify all statements below that correctly explain the trends:

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

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

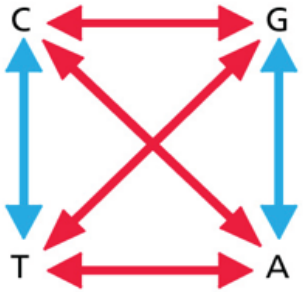
JC

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

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

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

Phylogenetics 2: week 8

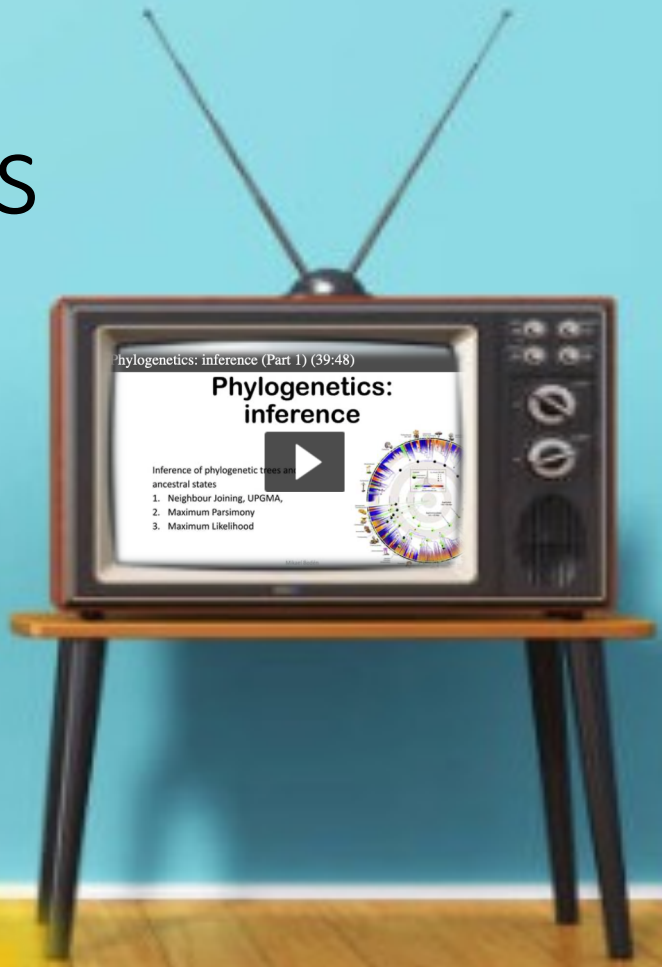
Watch the recordings

Phylogenetics: inference

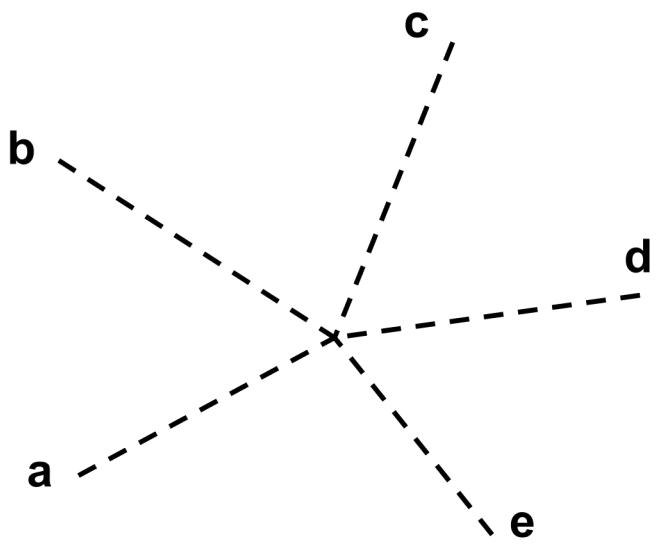
3 parts + UPGMA demo

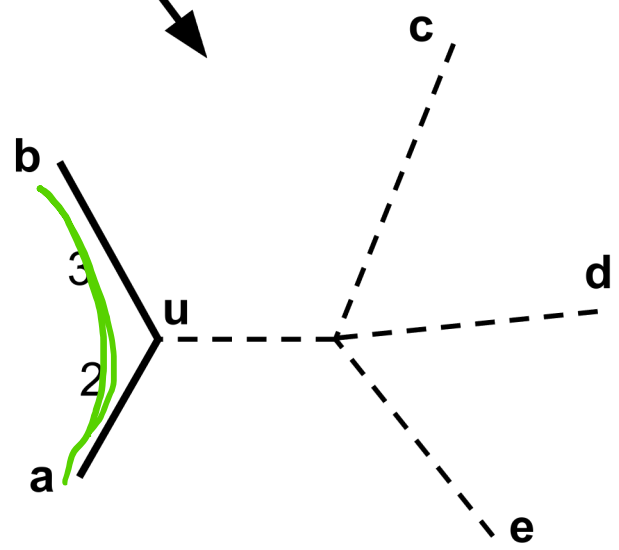
In this session we play with:

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



Mikael Bodén





$n=5$

	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

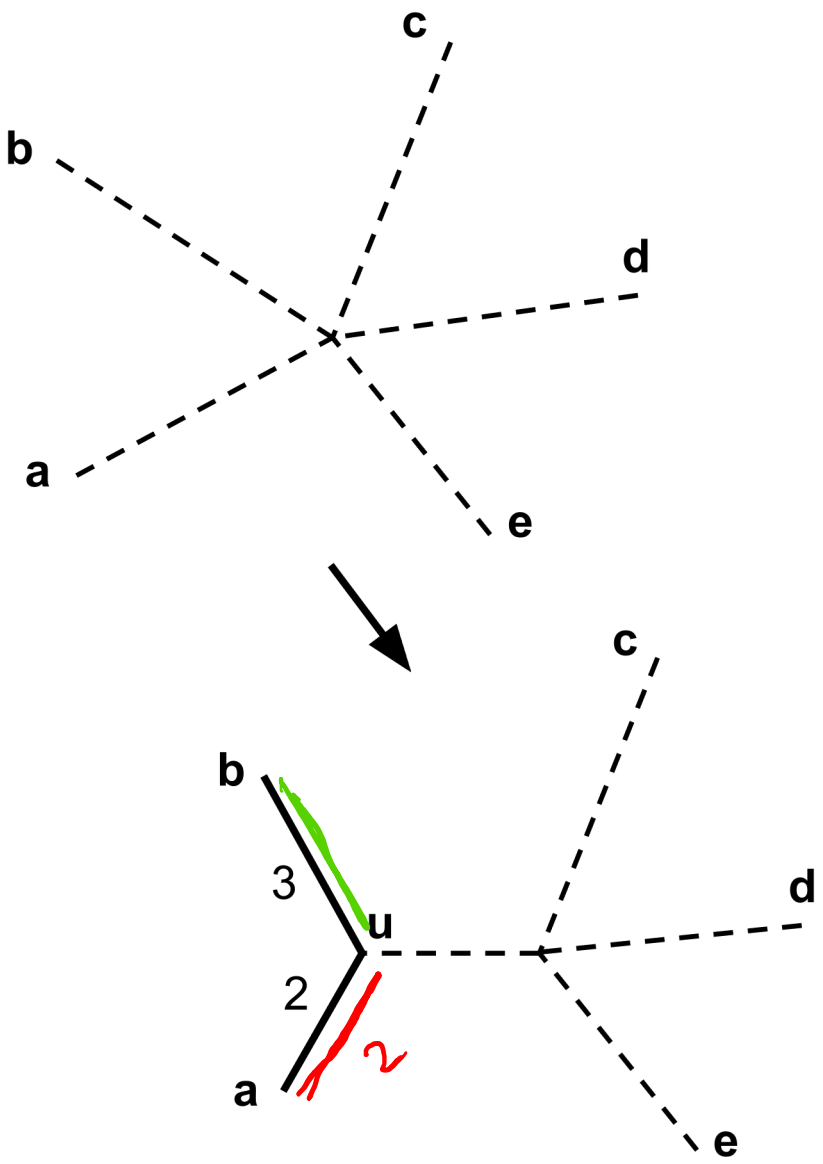
"Distance" between a and b d(a,b), d and e d(d,e)...

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

Looking-in *Looking-out*

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

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

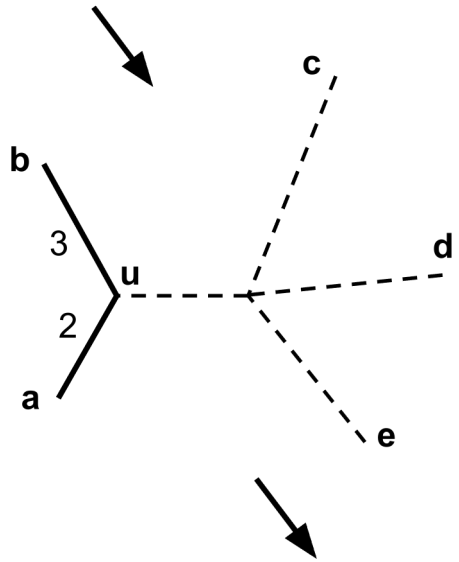


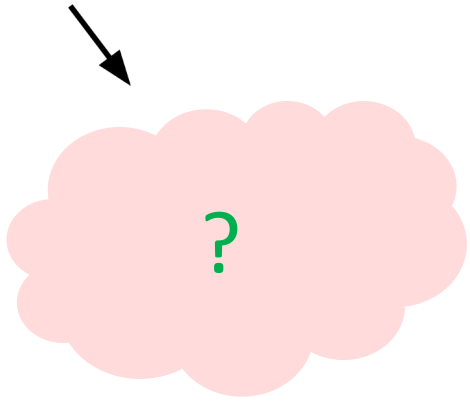
	a	b	c	d	e
a	0	5	9	9	8
b	5	0	10	10	9
c	9	10	0	8	7
d	9	10	8	0	3
e	8	9	7	3	0

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

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

$5-2=3$





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

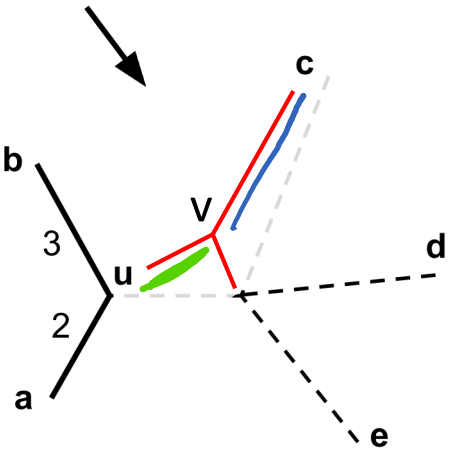
B. u and d

C. u and e

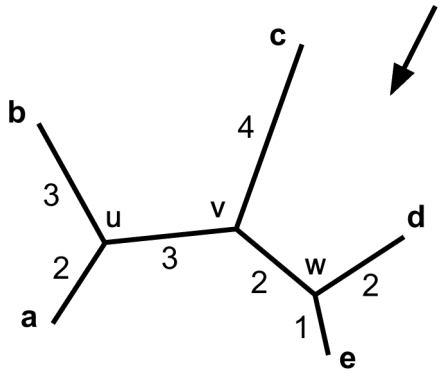
D. c and d

E. c and e

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



?



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?

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

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

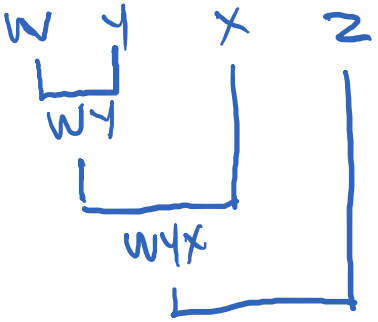
Between c and v?

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

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)

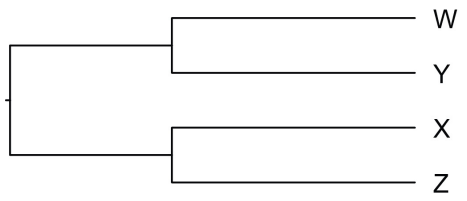


	W	X	Y	Z
W	0.00			
X	0.44	0.00		
Y	0.22	0.44	0.00	
Z	0.78	0.78	0.78	0.00

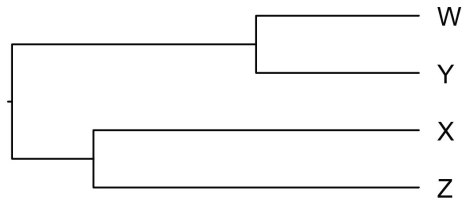
W CAGCATATG
X CATCAACTA
Y CAGCATTTC
Z CTTGTGAAC

Which one is it?

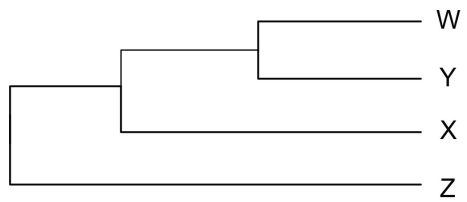
A



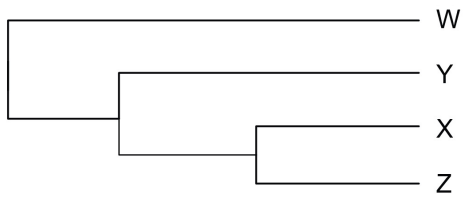
B



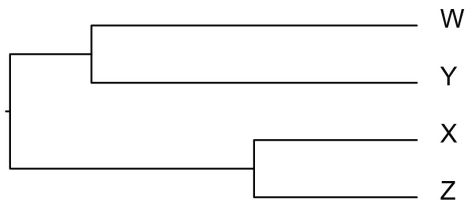
C



D



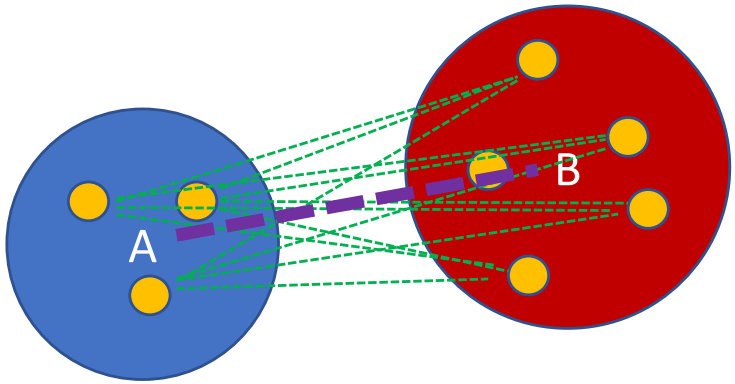
E



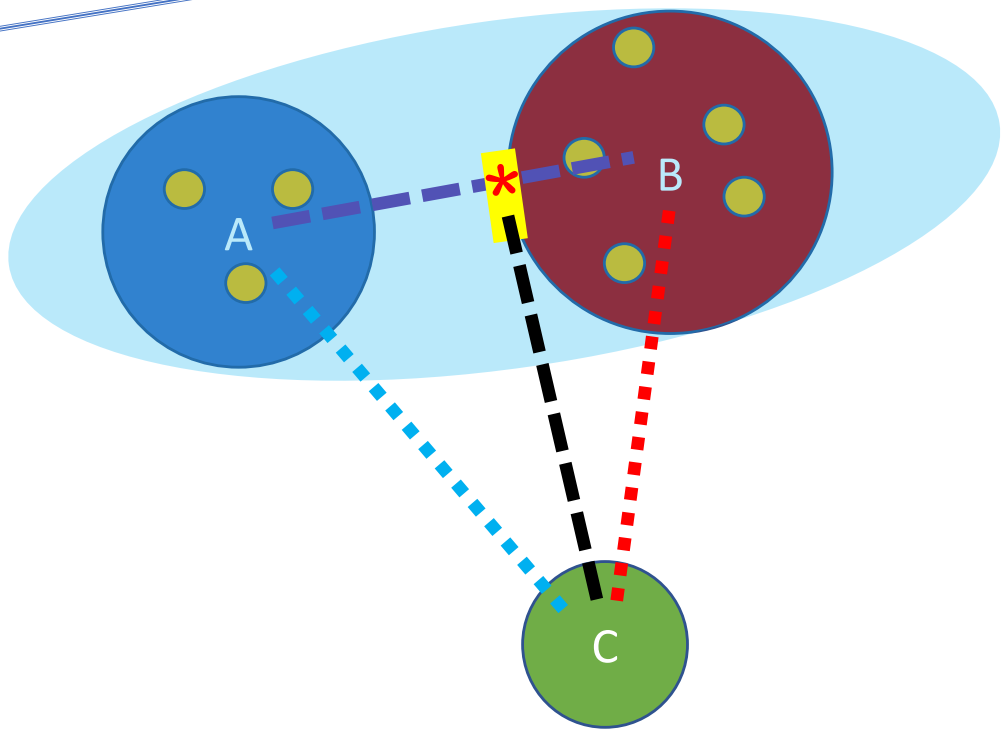
$$d_{AB} = \frac{1}{N_A N_B} \sum_{i \in A, j \in B} d_{ij}$$

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

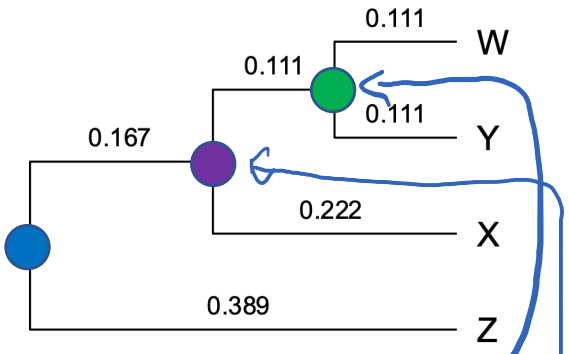
Sum the distances between all pairs of sequences, one from A and one from B; calculate their arithmetic average



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



```
[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

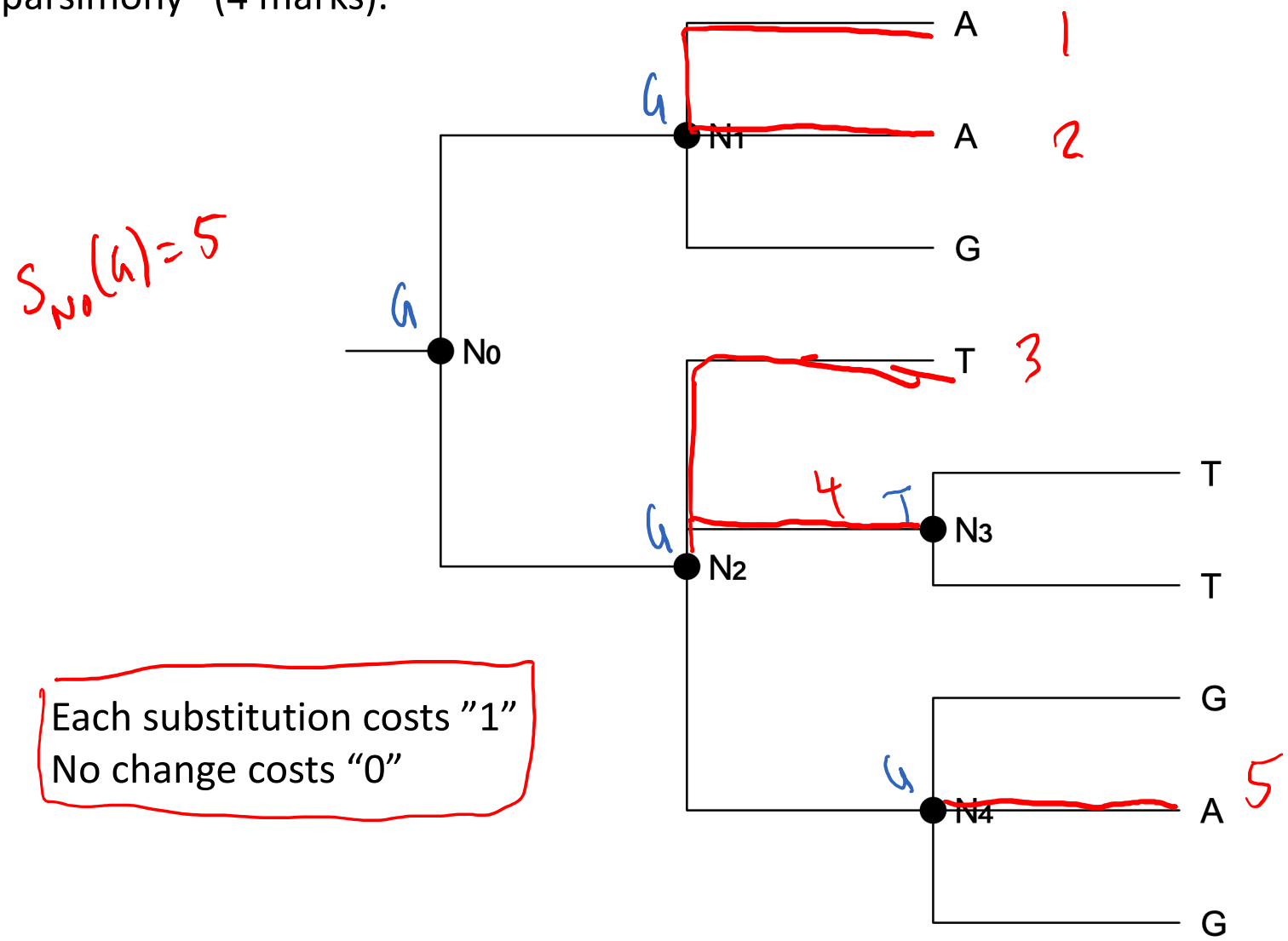
More complex example in UPGMA video

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).

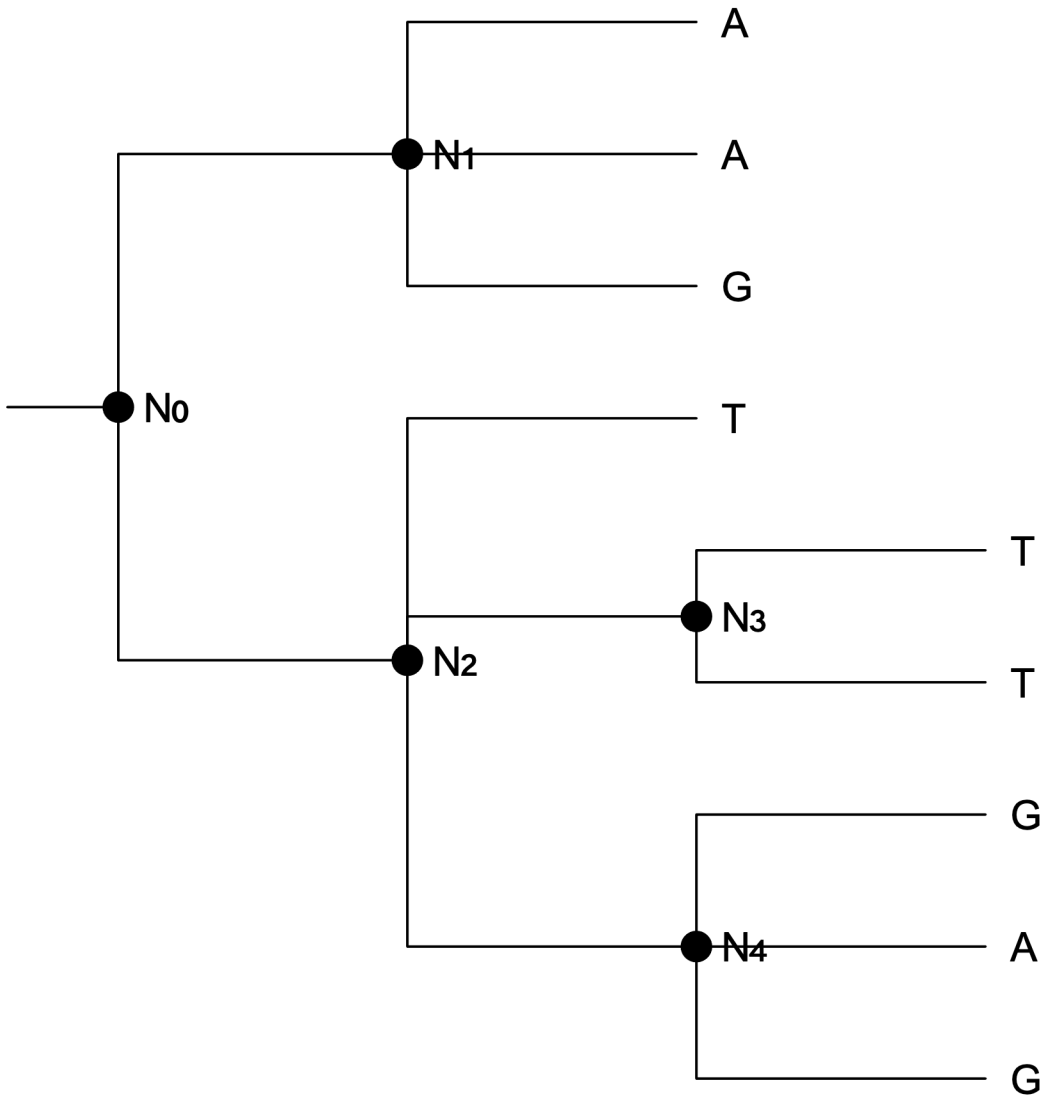
Respond with each possible assignment by typing A, C, G or T

$S_{N0}(A) =$

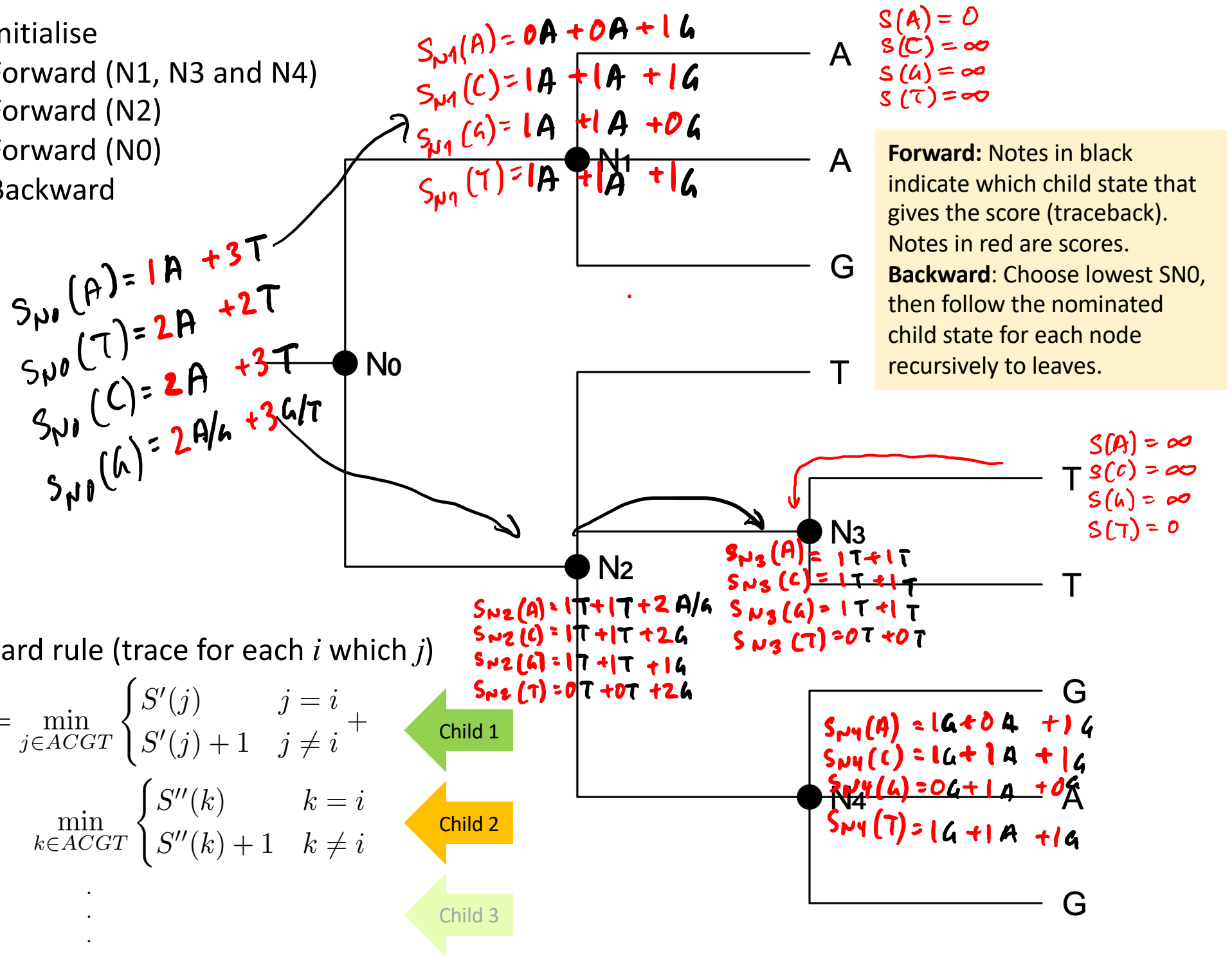
$S_{N0}(C) =$

$S_{N0}(G) = 5$

$S_{N0}(T) =$



- 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

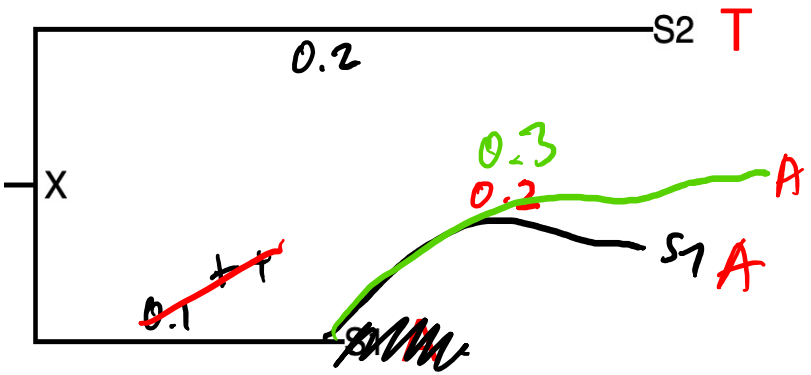
Questions so far?

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



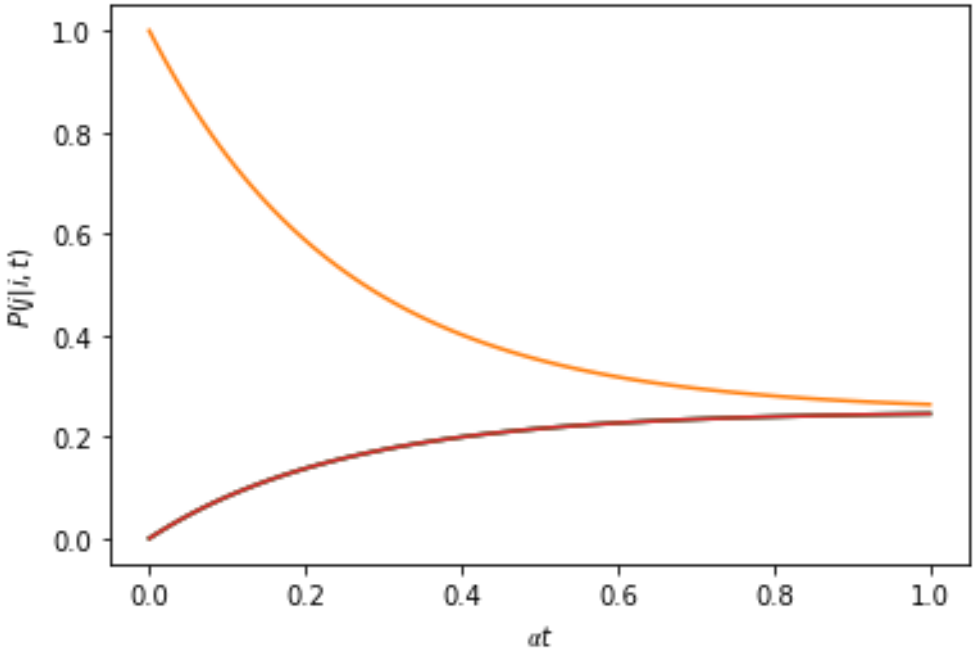
Both questions refer to this quantity:

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

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	

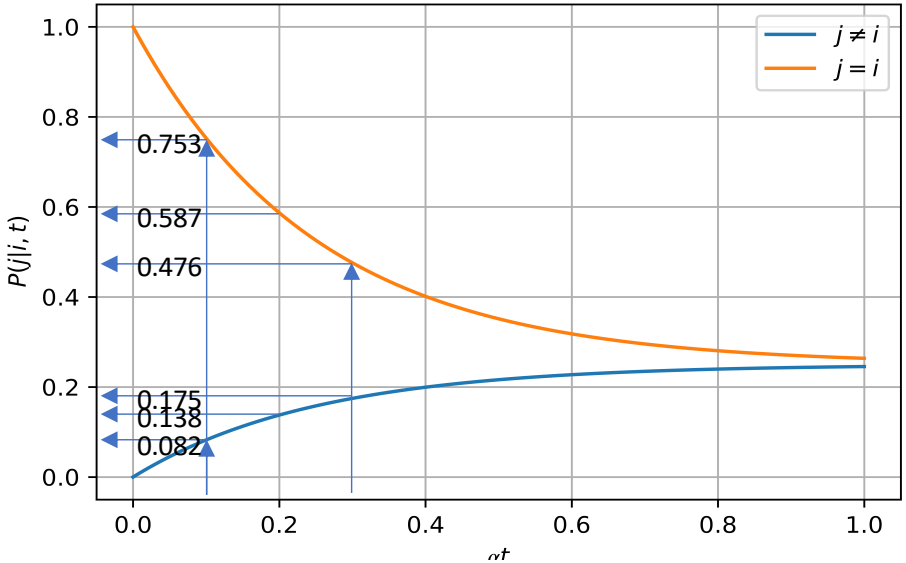
$P(s_1=\text{A} \mid x=\text{A}, t=0.1)\ P(s_2=\text{T} \mid x=\text{A}, t=0.2)\ P(x=\text{A})$

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

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

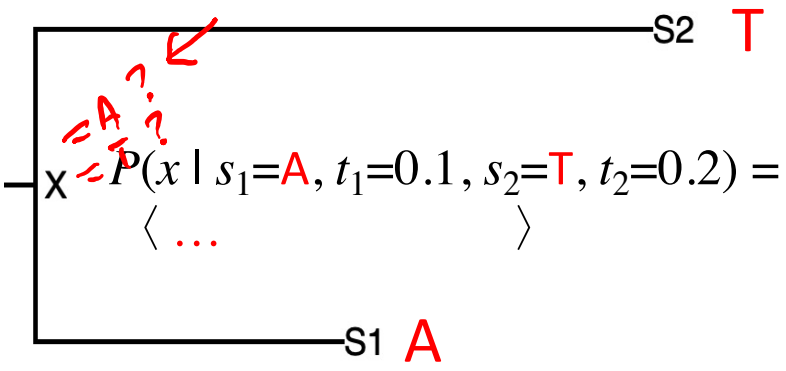
$P(s_1=\text{A} \mid x=\text{T}, t=0.1)\ P(s_2=\text{T} \mid x=\text{T}, t=0.2)\ P(x=\text{T})$

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

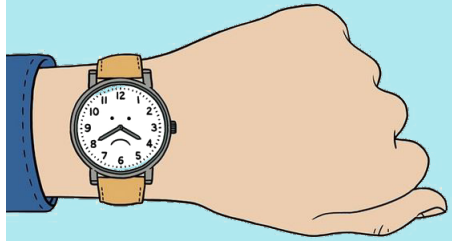
Maximum likelihood finds H

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

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



Phylogenetics: quantifying evolution

Episode in the series on phylogenetics

https://www.seh.ox.ac.uk/blog/mammals-during-the-extinction-of-dinosaurs

Mikael Bodén

Phylogenetics: quantifying evolution (Part 1)

Context: distances and models

Evolutionary distance and corrections (p-distance, Poisson and Gamma)

Molecular clock

Bear in mind…

most related sequences have *many* positions that have mutated, *some* of which have mutated several times

We need to effectively capture such dynamic changes

Metric of distance v. model of change

- The **evolutionary distance** between two sequences is an estimate of the number of mutations that has occurred since they diverged from their common ancestor

- While largely random, general rules may be governing which mutations lead to *changes* over time, imprinted in DNA, RNA and amino acid sequence

- **Evolutionary models** attempt to formalise tendencies of *change* in <INSERT–ALPHABET–HERE> sequences

ALPHABETs
- DNA: A,C,G,T
- RNA: A,C,G,U
- Protein: A,R,N,D,…
- more…

Sets of species/sequences: Distance matrices

- Define sequence distance
- Calculate all pairwise distances

Suppose we have three species i, j and k and a distance metric D

$$D = \begin{bmatrix} D_{i,i} & D_{i,j} & D_{i,k} \\ D_{j,i} & D_{j,j} & D_{j,k} \\ D_{k,i} & D_{k,j} & D_{k,k} \end{bmatrix}$$

p-distance (fractional alignment difference)

The simplest "evolutionary distance" between two sequences is the observed number of mutations since diverged.

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

Positions at which sequences differ

Total number of positions

$$1 - p = \frac{L - D}{L}$$

Positions at which sequences are the same

Total number of positions

The probability of "no change"

p-distance (fractional alignment difference)

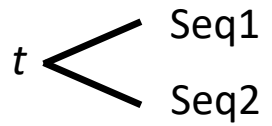
$$p = \frac{D}{L}$$

- Example
  - AAABBA
  - ABABAA
  - p = 2/6 = 1/3 = 0.333

- 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

Poisson distance correction accounts for multiple mutations at site

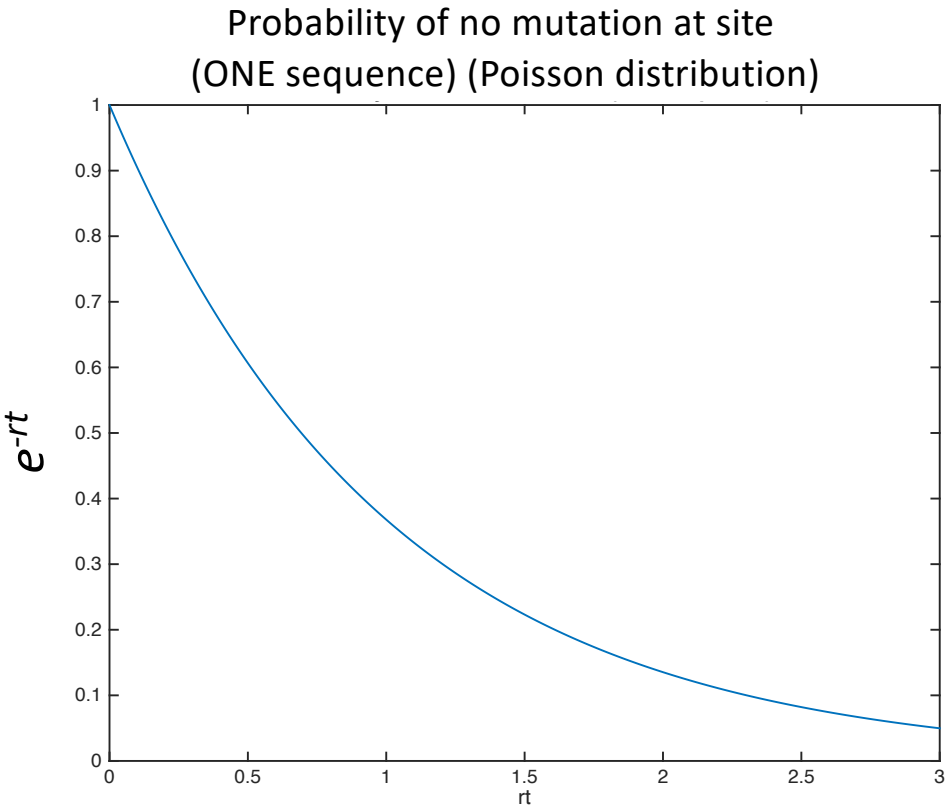
- **Mutation rate** per site: r
- After time t, **expected** number of mutations at site: rt
- *No* mutation at t: e^{-rt} Euler's number (constant)
- Two sequences share ancestor at t, so $2rt$ away



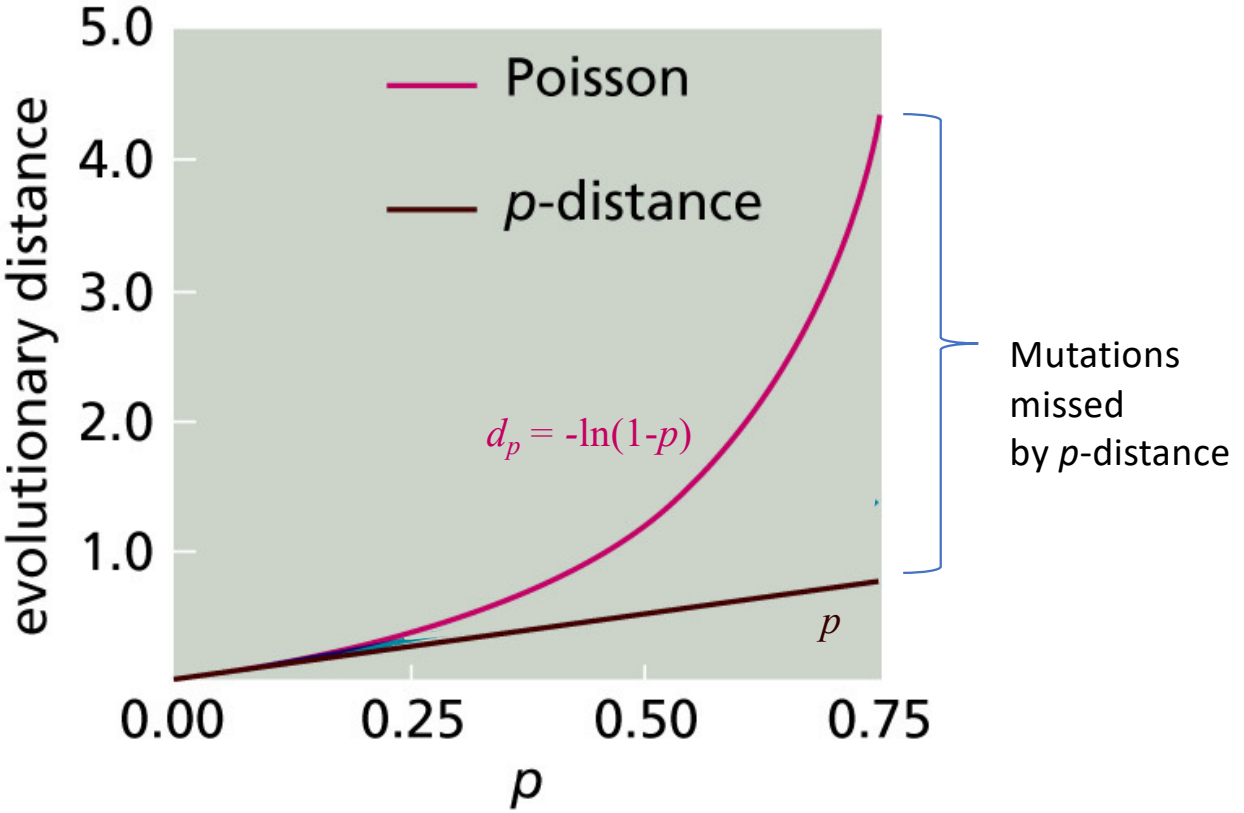
$2rt = d$

$1\text{-}p = e^{-2rt} = e^{-d}$ ← The probability of "no change"

$d_p = -\ln(1\text{-}p)$



Distance varies when *p* is corrected

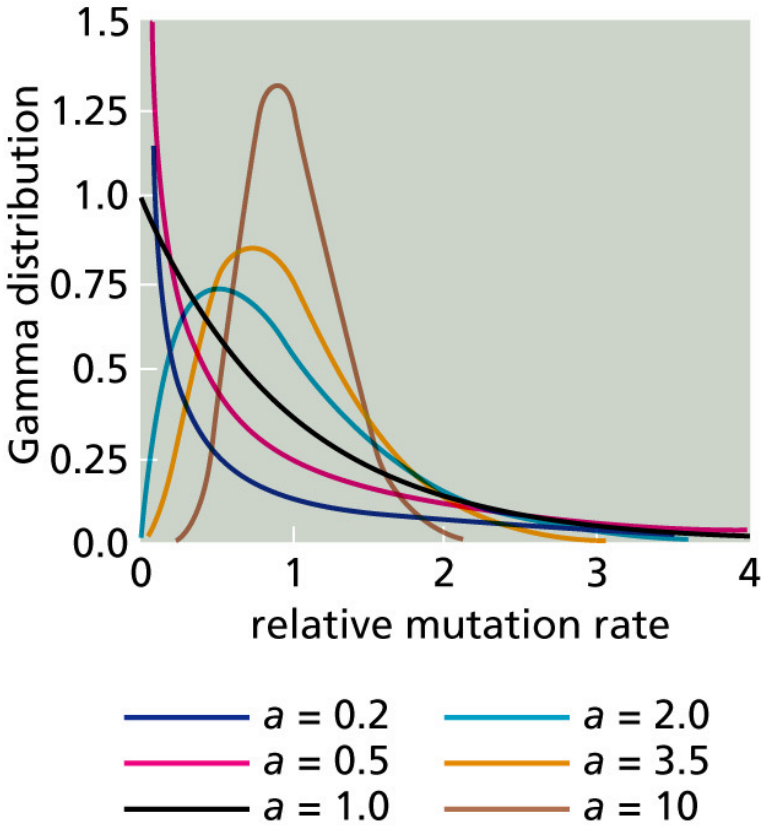


Gamma distance correction
Accounts for site-specific rates

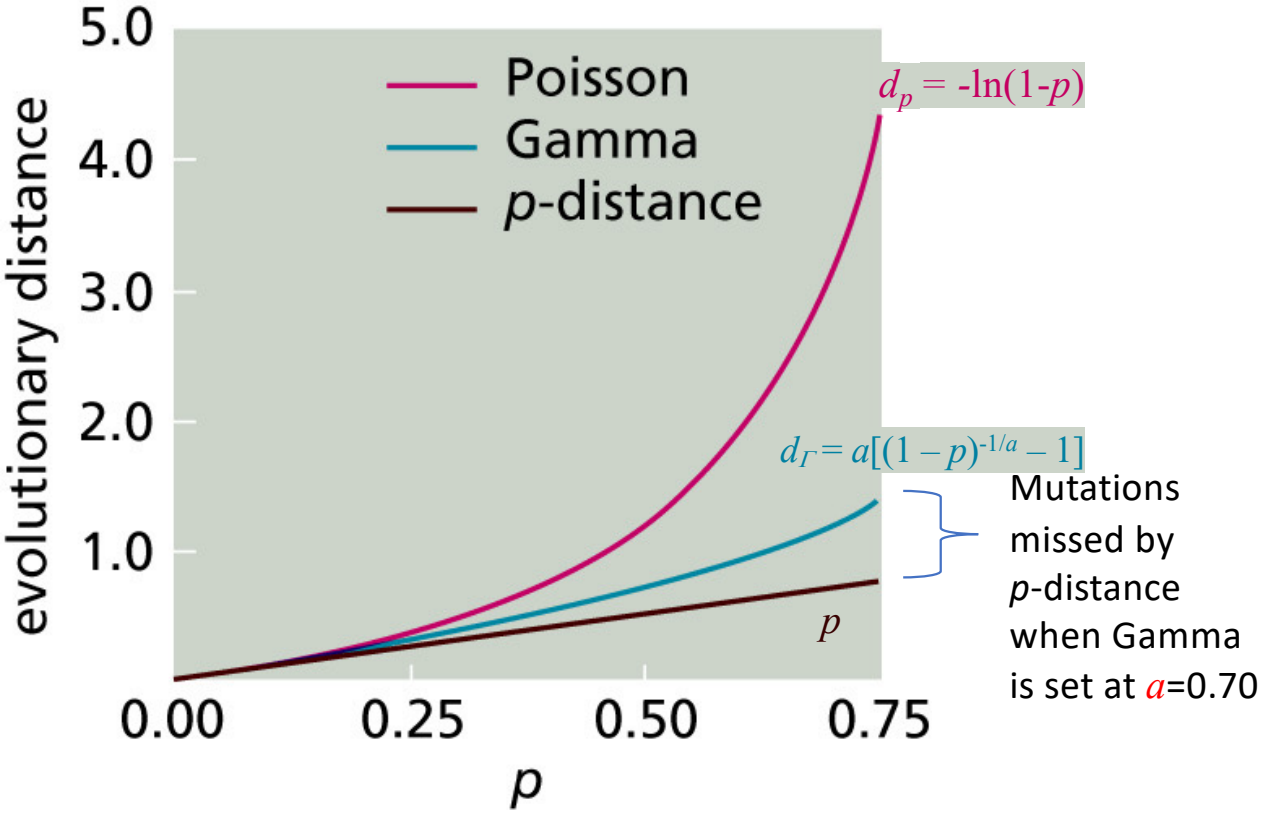
- Poisson only had one r for whole sequence
- The Gamma distribution can model how r varies across sites using a parameter a

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

Gamma corrects distance estimate for changes that can be explained by a variable rate
(a can be found by inspecting relevant data)

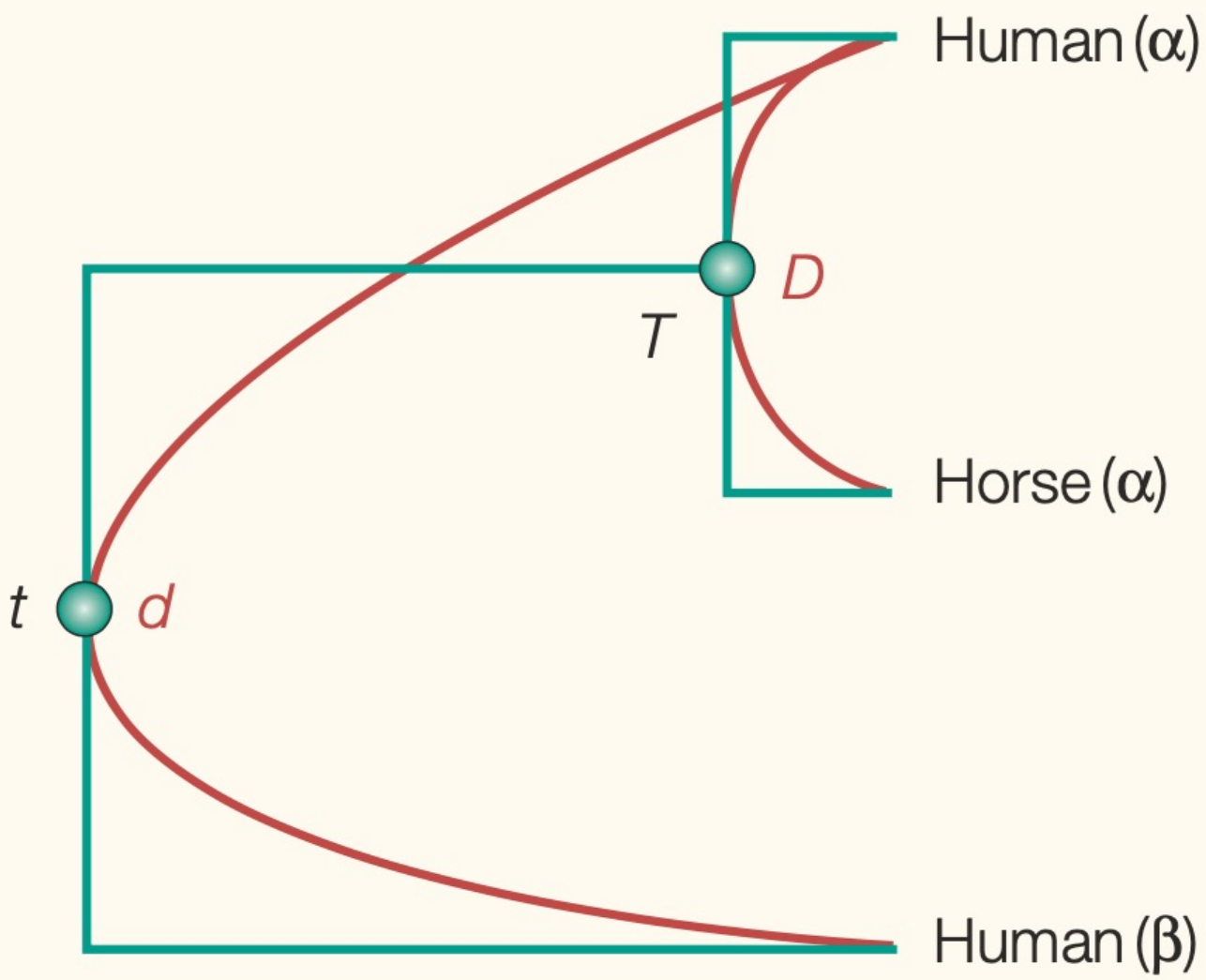


Distance varies when *p* is corrected



Distance = Time? Molecular clock and rate variation

- Zuckerkandl and Pauling noticed that the number of amino acid differences between different lineages changes roughly linearly with time; the rate of evolutionary change was approx. constant over time and over different lineages—this is known as the molecular clock hypothesis
- This is challenged by
  - Changing generation times, population size, species-specific differences (metabolism, ecology, etc), change in function and in the intensity of natural selection



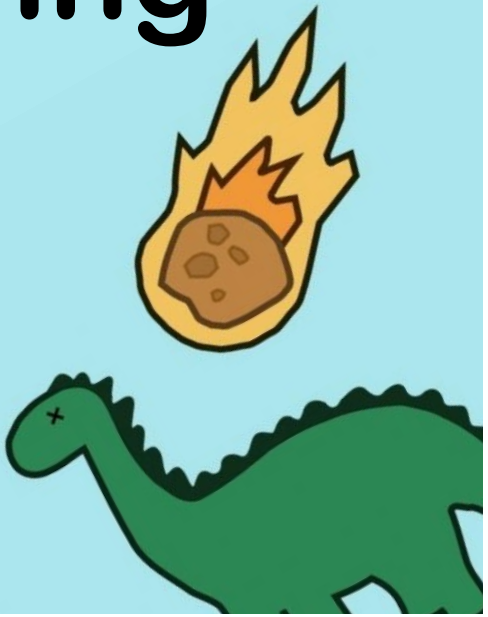
Kumar, S. Nature Reviews Genetics 6:654–662(2005)

Phylogenetics: quantifying evolution (Part 2)

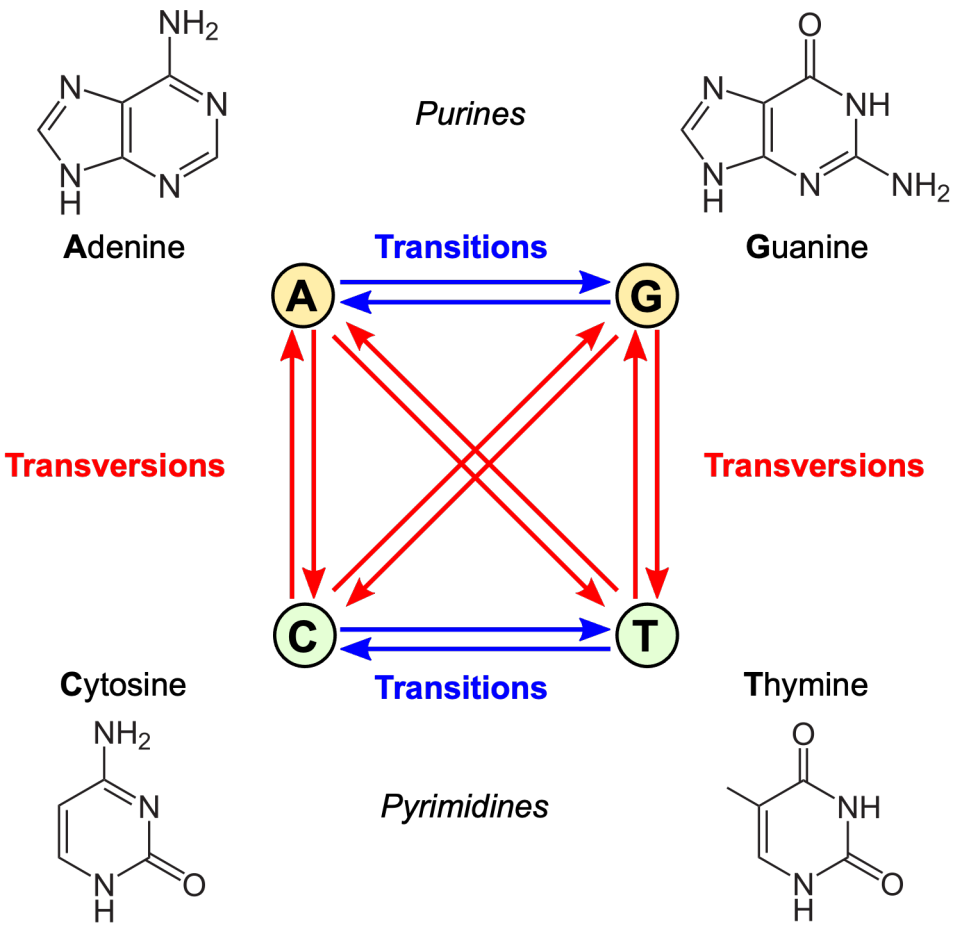
DNA models

Evolutionary models and rate matrices

Probabilistic meaning

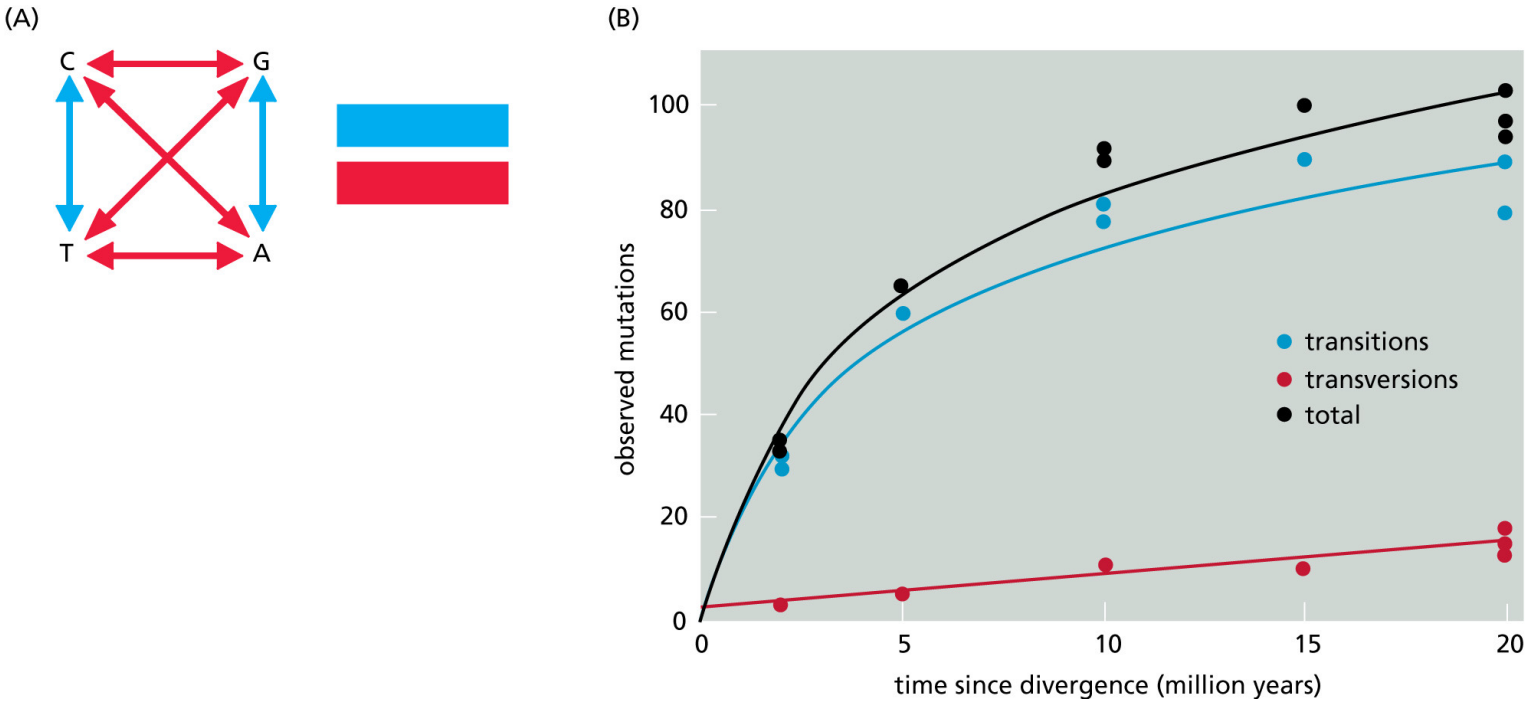


Transition vs. transversion



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

Transition vs. transversion



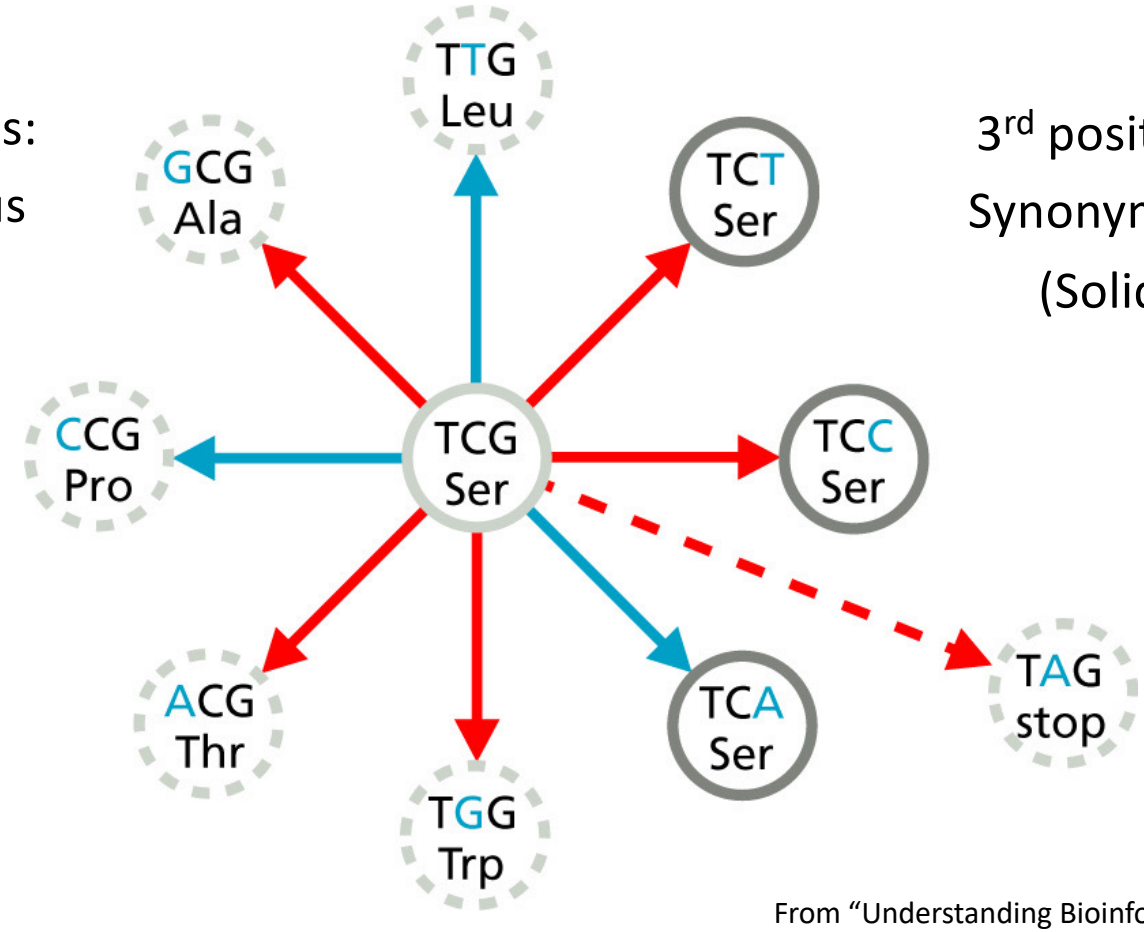
From “Understanding Bioinformatics”, Zvelebil & Baum, p238.

Different codon positions have different mutation rates

1st & 2nd positions:

Non-synonymous

(Dashed)



3rd position:

Synonymous

(Solid)

From “Understanding Bioinformatics”, Zvelebil & Baum, p241.

Evolutionary models

Model name	Base Composition	Different transition and transversion rates	All transition rates identical	All transversion rates identical	Reference
Jukes-Cantor (JC69)	1:1:1:1	No	Yes	Yes	Jukes and Cantor (1969)
Felsenstein 81 (F81)	Variable	No	Yes	Yes	Felsenstein (1981)
Kimura 2 Param (K80)	1:1:1:1	Yes	Yes	Yes	Kimura (1980)
HKY85	Variable	Yes	No	No	Hasegawa et al. (1985)
Tamura-Nei (TN)	Variable	Yes	No	Yes	Tamura and Nei (1993)
K3P (K81)	Variable	Yes	No	Yes	Kimura (1981)
SYM	1:1:1:1	Yes	No	No	Zharkikh (1994)
REV (GTR)	Variable	Yes	No	No	Rodriguez et al. (1990)

From “Understanding Bioinformatics”, Zvelebil & Baum, p253.

Models of (DNA) evolution (based on rate)

$$Q = \begin{pmatrix} * & \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & * & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & * & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} & * \end{pmatrix}$$

JC69 model (Jukes and Cantor 1969)

$$Q = \begin{pmatrix} * & \alpha & \beta & \gamma \\ \alpha & * & \gamma & \beta \\ \beta & \gamma & * & \alpha \\ \gamma & \beta & \alpha & * \end{pmatrix}$$

K81 model (Kimura 1981)

$$Q = \begin{array}{c} \begin{matrix} \texttt{A} & \texttt{G} & \texttt{C} & \texttt{T} \end{matrix} \\ \begin{pmatrix} * & \pi_G & \pi_C & \pi_T \\ \pi_A & * & \pi_C & \pi_T \\ \pi_A & \pi_G & * & \pi_T \\ \pi_A & \pi_G & \pi_C & * \end{pmatrix} \end{array} \begin{matrix} \texttt{A} \\ \texttt{G} \\ \texttt{C} \\ \texttt{T} \end{matrix}$$

F81 model (Felsenstein 1981)

$$Q = \begin{pmatrix} -(\alpha\pi_G + \beta\pi_C + \gamma\pi_T) & \alpha\pi_G & \beta\pi_C & \gamma\pi_T \\ \alpha\pi_A & -(\alpha\pi_A + \delta\pi_C + \epsilon\pi_T) & \delta\pi_C & \epsilon\pi_T \\ \beta\pi_A & \delta\pi_G & -(\beta\pi_A + \delta\pi_G + \eta\pi_T) & \eta\pi_T \\ \gamma\pi_A & \epsilon\pi_G & \eta\pi_C & -(\gamma\pi_A + \epsilon\pi_G + \eta\pi_C) \end{pmatrix}$$

GTR model (Tavaré 1986)

Models of (DNA) evolution: as rate matrix

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

Imaginary (DNA) lineage over (discrete) time

A ← Ancient time
A
A
A
G
G
C
C
G
G
G ← Present time

		To			
		A	**C**	**G**	**T**
From	**A**	-0.3	0.1	0.1	0.1
	C	0.1	-0.3	0.1	0.1
	G	0.1	0.1	-0.3	0.1
	T	0.1	0.1	0.1	-0.3

Note: Theory of continuous time Markov chain (CTMC)
- models change of state of a *single* discrete random variable
- defines probabilities of state changes, satisfying the *Markov property* (i.e. decision of future state depends only on current state)

The Jukes-Cantor (JC) model

- Treats substitutions uniformly
- Sites have identical rates, but depend on nucleotide identity

$$\begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{bmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{bmatrix} \end{array}$$

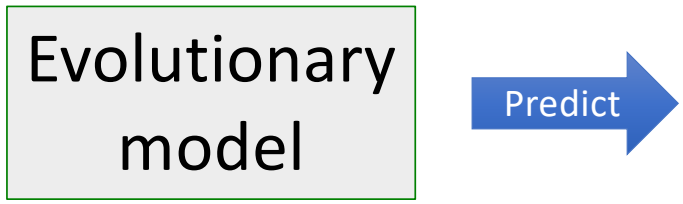
Re-running evolution (forward)

P(A) = ? ← Ancient time

← Present time

Re-running evolution (forward)

P(A) = 0.25 ← Ancient time



← Present time

Re-running evolution (forward)

P(A) = 0.25 ← Ancient time

Evolutionary model

P(C|A) = ? ← T = 0.4

← Present time

Re-running evolution (forward)

P(A) = 0.25 ← Ancient time

Evolutionary model

P(C|A) = ? ← T = 1.0

← Present time

Probabilities come from the model

$P_{ij}(T)$ can be written as a matrix $\mathbf{P}(T)$

T=1

	A	C	G	T
A	0.75	0.08	0.08	0.08
C	0.08	0.75	0.08	0.08
G	0.08	0.08	0.75	0.08
T	0.08	0.08	0.08	0.75

In discrete time:

$$\mathbf{P}(T + dT) =$$

$$\mathbf{P}(T)(\mathbf{I} + \mathbf{Q}dT)$$

T=0

	A	C	G	T
A	1	0	0	0
C	0	1	0	0
G	0	0	1	0
T	0	0	0	1

Q=

	A	C	G	T
A	-0.3	0.1	0.1	0.1
C	0.1	-0.3	0.1	0.1
G	0.1	0.1	-0.3	0.1
T	0.1	0.1	0.1	-0.3

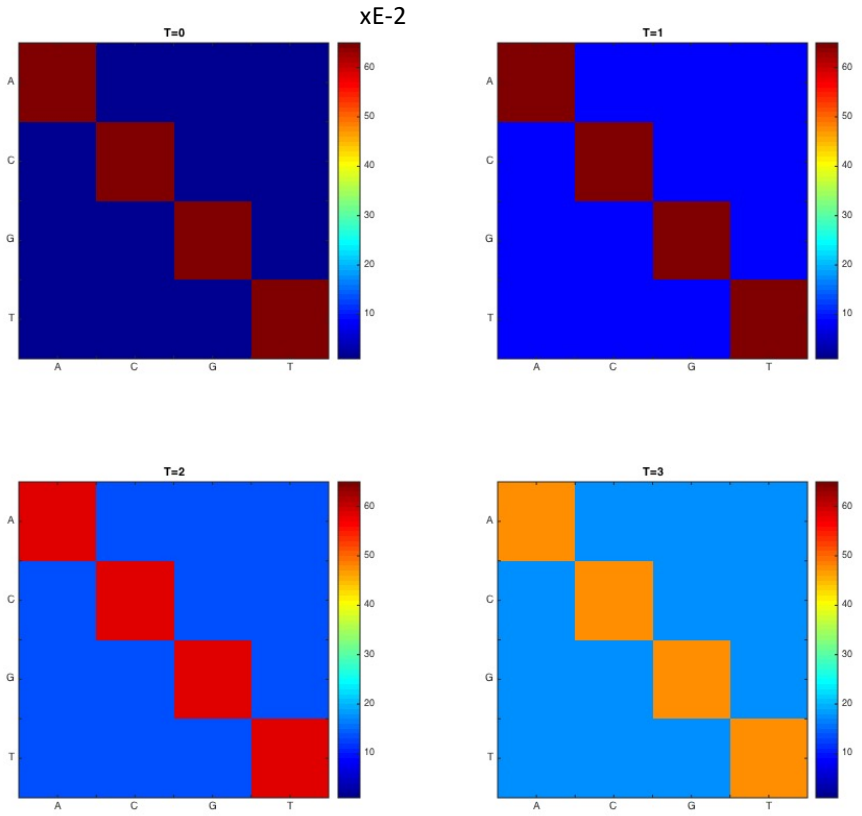
Probabilities come from the model

$P_{ij}(T)$ can be written as a matrix $\mathbf{P}(T)$

In discrete time:

$$\mathbf{P}(T + dT) =$$

$$\mathbf{P}(T)(\mathbf{I} + \mathbf{Q}dT)$$



Rate matrix for proteins

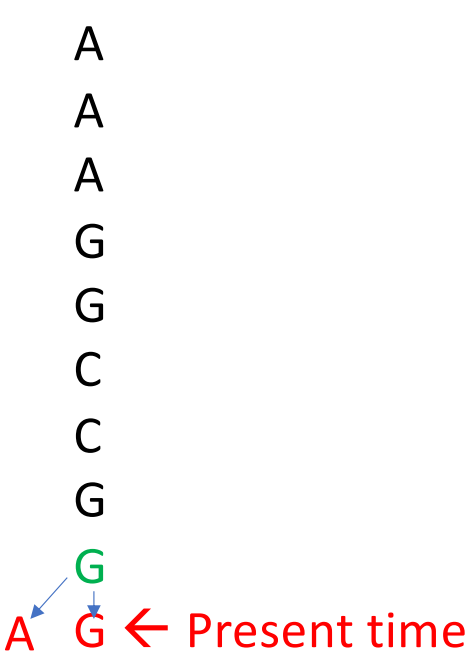
Dayhoff (remember PAM)

Zvelebil and Baum, sec 5.1

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	W
A	**-1.33**	0.01	0.04	0.06	0.01	0.03	0.10	0.21	0.01	0.02	0.04	0.02	0.01	0.01	0.13	0.28	0.22	0.00	0.01	0.13
R	0.02	**-0.87**	0.01	0.00	0.01	0.09	0.00	0.01	0.08	0.02	0.01	0.37	0.01	0.01	0.05	0.11	0.02	0.02	0.00	0.02
N	0.09	0.01	**-1.78**	0.42	0.00	0.04	0.07	0.12	0.18	0.03	0.03	0.25	0.00	0.01	0.02	0.34	0.13	0.00	0.03	0.01
D	0.11	0.00	0.36	**-1.41**	0.00	0.05	0.56	0.11	0.03	0.01	0.00	0.06	0.00	0.00	0.01	0.07	0.04	0.00	0.00	0.01
C	0.03	0.01	0.00	0.00	**-0.27**	0.00	0.00	0.01	0.01	0.02	0.00	0.00	0.00	0.00	0.01	0.11	0.01	0.00	0.03	0.03
Q	0.08	0.10	0.04	0.06	0.00	**-1.24**	0.35	0.03	0.20	0.01	0.06	0.12	0.02	0.00	0.08	0.04	0.03	0.00	0.00	0.02
E	0.17	0.00	0.06	0.53	0.00	0.27	**-1.36**	0.07	0.02	0.02	0.01	0.07	0.00	0.00	0.03	0.06	0.02	0.00	0.01	0.02
G	0.21	0.00	0.06	0.06	0.00	0.01	0.04	**-0.65**	0.00	0.00	0.01	0.02	0.00	0.01	0.02	0.16	0.02	0.00	0.00	0.04
H	0.02	0.10	0.21	0.04	0.01	0.23	0.02	0.01	**-0.88**	0.00	0.04	0.02	0.00	0.02	0.05	0.03	0.01	0.00	0.04	0.03
I	0.06	0.03	0.03	0.01	0.02	0.01	0.03	0.00	0.00	**-1.28**	0.22	0.04	0.05	0.08	0.01	0.02	0.11	0.00	0.01	0.57
L	0.04	0.01	0.01	0.00	0.00	0.03	0.01	0.01	0.02	0.09	**-0.53**	0.02	0.08	0.06	0.02	0.01	0.02	0.01	0.01	0.11
K	0.02	0.19	0.13	0.03	0.00	0.06	0.04	0.02	0.01	0.02	0.02	**-0.75**	0.04	0.00	0.02	0.07	0.08	0.00	0.00	0.01
M	0.06	0.04	0.00	0.00	0.00	0.04	0.02	0.02	0.00	0.12	0.45	0.19	**-1.25**	0.04	0.01	0.04	0.06	0.00	0.00	0.17
F	0.02	0.01	0.01	0.00	0.00	0.00	0.00	0.01	0.02	0.07	0.13	0.00	0.01	**-0.55**	0.01	0.03	0.01	0.01	0.21	0.01
P	0.22	0.04	0.02	0.01	0.01	0.06	0.03	0.03	0.03	0.00	0.03	0.03	0.00	0.00	**-0.75**	0.17	0.05	0.00	0.00	0.03
S	0.35	0.06	0.20	0.05	0.05	0.02	0.04	0.21	0.01	0.01	0.02	0.08	0.01	0.02	0.12	**-1.60**	0.32	0.01	0.01	0.02
T	0.32	0.01	0.09	0.03	0.01	0.02	0.02	0.03	0.01	0.07	0.03	0.11	0.02	0.01	0.04	0.38	**-1.29**	0.00	0.01	0.10
W	0.00	0.08	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.04	0.00	0.00	0.03	0.00	0.05	0.00	**-0.24**	0.02	0.00
Y	0.02	0.00	0.04	0.00	0.03	0.00	0.01	0.00	0.04	0.01	0.02	0.01	0.00	0.28	0.00	0.02	0.02	0.01	**-0.55**	0.02
W	0.18	0.01	0.01	0.01	0.02	0.01	0.02	0.05	0.02	0.33	0.15	0.01	0.04	0.01	0.03	0.02	0.09	0.00	0.01	**-0.99**

On *what* are protein models based?

```
DYISWWQQQ
DYISSWQEQ
DYISLWQEQ
DYISLWQDD
```



- Answer: Counts of character pairs from alignments of closely related sequences
- When sequences are *really* close (in time) the rates and probabilities of change are approximately linear
- Counts scaled based on sequence divergence, and averaging across many alignments

Probabilities come from the model

$P_{ij}(T)$ can be written as a matrix $\mathbf{P}(T)$

In discrete time:

$$\mathbf{P}(T + dT) =$$

$$\mathbf{P}(T)(\mathbf{I} + \mathbf{Q}dT)$$

$\mathbf{P}_{T=1} =$

0.37	0.02	0.02	0.03	0.01	0.02	0.05	0.07	0.01	0.02	0.03	0.04	0.01	0.01	0.04	0.1
0.04	0.40	0.03	0.02	0.01	0.05	0.03	0.04	0.03	0.01	0.03	0.14	0.01	0.01	0.02	0.0
0.05	0.03	0.27	0.12	0.01	0.03	0.05	0.06	0.04	0.02	0.02	0.08	0.01	0.01	0.02	0.1
0.05	0.01	0.08	0.42	0.00	0.03	0.14	0.05	0.02	0.01	0.01	0.04	0.00	0.00	0.02	0.0
0.06	0.02	0.01	0.01	0.62	0.01	0.01	0.03	0.01	0.02	0.03	0.01	0.01	0.01	0.01	0.0
0.06	0.07	0.04	0.05	0.00	0.28	0.11	0.03	0.04	0.01	0.04	0.10	0.01	0.01	0.03	0.0
0.07	0.03	0.04	0.14	0.00	0.07	0.34	0.04	0.02	0.01	0.02	0.08	0.01	0.01	0.02	0.0
0.07	0.02	0.03	0.04	0.01	0.01	0.03	0.62	0.01	0.01	0.01	0.02	0.00	0.00	0.01	0.0
0.03	0.05	0.06	0.04	0.01	0.06	0.04	0.03	0.38	0.01	0.03	0.05	0.01	0.02	0.02	0.0
0.04	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.35	0.15	0.02	0.04	0.03	0.01	0.0
0.03	0.02	0.01	0.01	0.01	0.02	0.01	0.01	0.01	0.08	0.52	0.02	0.04	0.05	0.02	0.0
0.05	0.10	0.05	0.04	0.00	0.06	0.07	0.03	0.02	0.01	0.03	0.36	0.01	0.01	0.02	0.0
0.05	0.02	0.01	0.01	0.01	0.03	0.02	0.02	0.01	0.09	0.19	0.03	0.28	0.03	0.01	0.0
0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.04	0.11	0.01	0.02	0.51	0.01	0.0
0.08	0.02	0.01	0.02	0.00	0.02	0.03	0.03	0.01	0.01	0.03	0.03	0.01	0.01	0.55	0.0
0.12	0.03	0.06	0.04	0.02	0.03	0.04	0.07	0.02	0.02	0.03	0.04	0.01	0.01	0.04	0.2
0.10	0.02	0.04	0.03	0.01	0.02	0.04	0.03	0.01	0.04	0.03	0.05	0.02	0.01	0.03	0.1
0.02	0.03	0.01	0.01	0.01	0.01	0.01	0.02	0.01	0.01	0.05	0.01	0.01	0.04	0.01	0.0
0.02	0.02	0.03	0.02	0.01	0.01	0.01	0.01	0.05	0.02	0.04	0.01	0.01	0.13	0.01	0.0
0.08	0.01	0.01	0.01	0.01	0.01	0.02	0.02	0.01	0.14	0.10	0.02	0.03	0.02	0.01	0.0

$\mathbf{Q} =$

−1.12	0.03	0.02	0.04	0.02	0.04	0.10	0.12	0.01	0.01	0.04	0.06	0.(
0.05	−0.97	0.03	0.01	0.01	0.12	0.03	0.05	0.06	0.01	0.05	0.35	0.(
0.05	0.03	−1.45	0.32	0.01	0.06	0.06	0.10	0.10	0.03	0.01	0.20	0.(
0.07	0.01	0.22	−0.99	0.00	0.02	0.38	0.08	0.02	0.00	0.01	0.03	0.(
0.09	0.02	0.01	0.00	−0.49	0.00	0.00	0.03	0.01	0.01	0.04	0.01	0.(
0.08	0.14	0.06	0.04	0.00	−1.38	0.33	0.03	0.11	0.01	0.08	0.25	0.(
0.14	0.02	0.04	0.37	0.00	0.21	−1.24	0.05	0.02	0.01	0.01	0.17	0.(
0.13	0.03	0.05	0.05	0.01	0.01	0.04	−0.50	0.01	0.00	0.01	0.02	0.(
0.03	0.10	0.16	0.06	0.01	0.17	0.04	0.02	−0.99	0.01	0.05	0.06	0.(
0.02	0.01	0.02	0.00	0.00	0.01	0.01	0.00	0.00	−1.23	0.29	0.02	0.(
0.04	0.02	0.01	0.01	0.01	0.03	0.01	0.01	0.01	0.16	−0.73	0.02	0.(
0.08	0.25	0.12	0.03	0.00	0.15	0.16	0.03	0.02	0.02	0.02	−1.12	0.(
0.08	0.03	0.01	0.01	0.01	0.06	0.02	0.02	0.01	0.22	0.44	0.06	−1.(
0.02	0.01	0.00	0.00	0.01	0.00	0.01	0.00	0.02	0.05	0.19	0.01	0.(
0.13	0.03	0.01	0.03	0.00	0.04	0.04	0.02	0.02	0.01	0.04	0.04	0.(
0.31	0.06	0.16	0.06	0.03	0.04	0.04	0.12	0.02	0.02	0.03	0.06	0.(
0.19	0.03	0.08	0.02	0.01	0.03	0.05	0.02	0.01	0.07	0.03	0.09	0.(
0.01	0.05	0.00	0.01	0.02	0.01	0.01	0.03	0.01	0.01	0.06	0.01	0.(
0.02	0.02	0.05	0.02	0.01	0.01	0.01	0.01	0.10	0.02	0.04	0.01	0.(
0.18	0.01	0.01	0.01	0.02	0.01	0.04	0.02	0.00	0.40	0.16	0.02	0.(

Probabilities come from the model

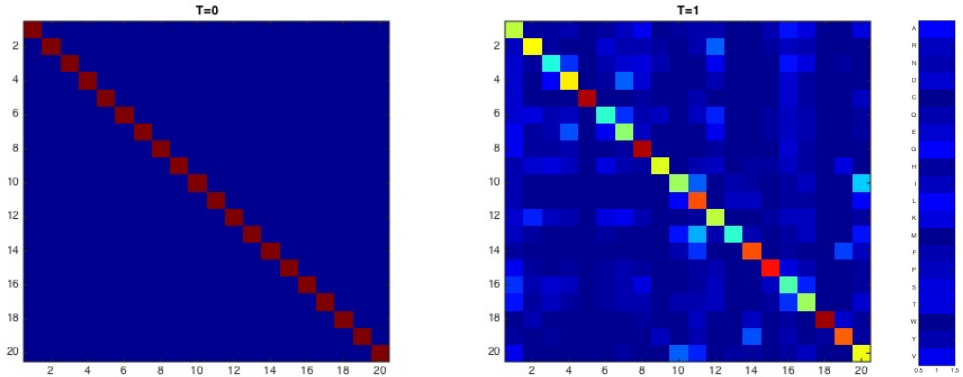
$P_{ij}(T)$ can be written as a matrix $\mathbf{P}(T)$

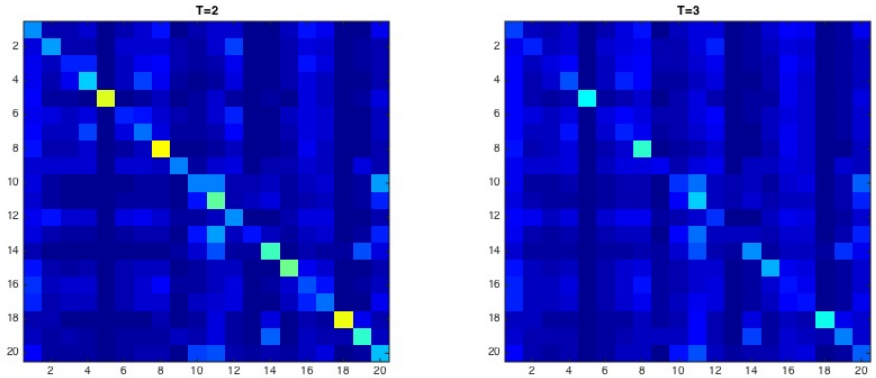
In discrete time:

$$\mathbf{P}(T + dT) =$$

$$\mathbf{P}(T)(\mathbf{I} + \mathbf{Q}dT)$$

So: there is a probability matrix for all possible time lapses.





Rate matrix for proteins

Whelan & Goldman

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	W
A	**-1.12**	0.03	0.02	0.04	0.02	0.04	0.10	0.12	0.01	0.01	0.04	0.06	0.02	0.01	0.07	0.25	0.14	0.00	0.01	0.15
R	0.05	**-0.97**	0.03	0.01	0.01	0.12	0.03	0.05	0.06	0.01	0.05	0.35	0.01	0.00	0.03	0.09	0.04	0.02	0.01	0.02
N	0.05	0.03	**-1.45**	0.32	0.01	0.06	0.06	0.10	0.10	0.03	0.01	0.20	0.00	0.00	0.01	0.29	0.13	0.00	0.04	0.02
D	0.07	0.01	0.22	**-0.99**	0.00	0.02	0.38	0.08	0.02	0.00	0.01	0.03	0.00	0.00	0.02	0.08	0.02	0.00	0.01	0.01
C	0.09	0.02	0.01	0.00	**-0.49**	0.00	0.00	0.03	0.01	0.01	0.04	0.01	0.01	0.02	0.01	0.10	0.03	0.01	0.02	0.08
Q	0.08	0.14	0.06	0.04	0.00	**-1.38**	0.33	0.03	0.11	0.01	0.08	0.25	0.03	0.00	0.05	0.08	0.06	0.00	0.01	0.02
E	0.14	0.02	0.04	0.37	0.00	0.21	**-1.24**	0.05	0.02	0.01	0.01	0.17	0.01	0.00	0.03	0.05	0.05	0.00	0.01	0.04
G	0.13	0.03	0.05	0.05	0.01	0.01	0.04	**-0.50**	0.01	0.00	0.01	0.02	0.00	0.00	0.01	0.10	0.02	0.01	0.00	0.01
H	0.03	0.10	0.16	0.06	0.01	0.17	0.04	0.02	**-0.99**	0.01	0.05	0.06	0.01	0.03	0.03	0.05	0.03	0.00	0.14	0.01
I	0.02	0.01	0.02	0.00	0.00	0.01	0.01	0.00	0.00	**-1.23**	0.29	0.02	0.09	0.04	0.01	0.02	0.09	0.00	0.02	0.58
L	0.04	0.02	0.01	0.01	0.01	0.03	0.01	0.01	0.01	0.16	**-0.73**	0.02	0.10	0.09	0.02	0.03	0.02	0.01	0.02	0.13
K	0.08	0.25	0.12	0.03	0.00	0.15	0.16	0.03	0.02	0.02	0.02	**-1.12**	0.02	0.00	0.03	0.07	0.09	0.00	0.01	0.02
M	0.08	0.03	0.01	0.01	0.01	0.06	0.02	0.02	0.01	0.22	0.44	0.06	**-1.32**	0.05	0.01	0.04	0.10	0.01	0.02	0.15
F	0.02	0.01	0.00	0.00	0.01	0.00	0.01	0.00	0.02	0.05	0.19	0.01	0.02	**-0.72**	0.01	0.04	0.01	0.02	0.24	0.05
P	0.13	0.03	0.01	0.03	0.00	0.04	0.04	0.02	0.02	0.01	0.04	0.04	0.00	0.01	**-0.61**	0.12	0.05	0.00	0.01	0.02
S	0.31	0.06	0.16	0.06	0.03	0.04	0.04	0.12	0.02	0.02	0.03	0.06	0.01	0.02	0.08	**-1.39**	0.28	0.01	0.03	0.02
T	0.19	0.03	0.08	0.02	0.01	0.03	0.05	0.02	0.01	0.07	0.03	0.09	0.03	0.01	0.04	0.32	**-1.16**	0.00	0.01	0.10
W	0.01	0.05	0.00	0.01	0.02	0.01	0.01	0.03	0.01	0.01	0.06	0.01	0.01	0.06	0.01	0.04	0.01	**-0.47**	0.09	0.03
Y	0.02	0.02	0.05	0.02	0.01	0.01	0.01	0.01	0.10	0.02	0.04	0.01	0.01	0.26	0.01	0.06	0.02	0.04	**-0.73**	0.02
W	0.18	0.01	0.01	0.01	0.02	0.01	0.04	0.02	0.00	0.40	0.16	0.02	0.04	0.03	0.02	0.02	0.09	0.01	0.01	**-1.09**

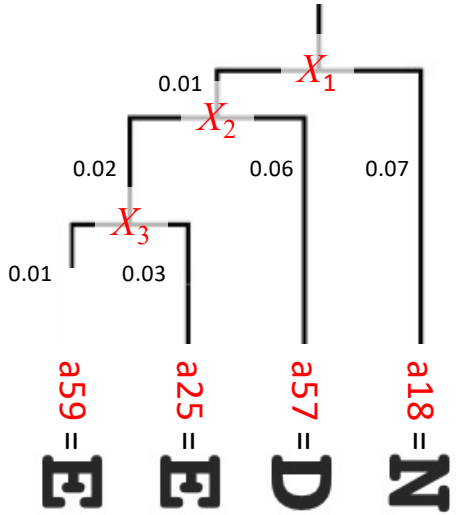
Transition probability matrix (CTMC)

Whelan & Goldman

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	W
A	**0.37**	0.02	0.02	0.03	0.01	0.02	0.05	0.07	0.01	0.02	0.03	0.04	0.01	0.01	0.04	0.10	0.07	0.00	0.01	0.07
R	0.04	**0.40**	0.03	0.02	0.01	0.05	0.03	0.04	0.03	0.01	0.03	0.14	0.01	0.01	0.02	0.05	0.03	0.01	0.01	0.02
N	0.05	0.03	**0.27**	0.12	0.01	0.03	0.05	0.06	0.04	0.02	0.02	0.08	0.01	0.01	0.02	0.10	0.06	0.00	0.02	0.02
D	0.05	0.01	0.08	**0.42**	0.00	0.03	0.14	0.05	0.02	0.01	0.01	0.04	0.00	0.00	0.02	0.05	0.03	0.00	0.01	0.02
C	0.06	0.02	0.01	0.01	**0.62**	0.01	0.01	0.03	0.01	0.02	0.03	0.01	0.01	0.01	0.01	0.06	0.03	0.01	0.02	0.05
Q	0.06	0.07	0.04	0.05	0.00	**0.28**	0.11	0.03	0.04	0.01	0.04	0.10	0.01	0.01	0.03	0.05	0.04	0.00	0.01	0.02
E	0.07	0.03	0.04	0.14	0.00	0.07	**0.34**	0.04	0.02	0.01	0.02	0.08	0.01	0.01	0.02	0.04	0.04	0.00	0.01	0.03
G	0.07	0.02	0.03	0.04	0.01	0.01	0.03	**0.62**	0.01	0.01	0.01	0.02	0.00	0.00	0.01	0.06	0.02	0.00	0.01	0.02
H	0.03	0.05	0.06	0.04	0.01	0.06	0.04	0.03	**0.38**	0.01	0.03	0.05	0.01	0.02	0.02	0.04	0.03	0.01	0.07	0.02
I	0.04	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	**0.35**	0.15	0.02	0.04	0.03	0.01	0.02	0.05	0.00	0.01	0.21
L	0.03	0.02	0.01	0.01	0.01	0.02	0.01	0.01	0.01	0.08	**0.52**	0.02	0.04	0.05	0.02	0.02	0.02	0.01	0.02	0.09
K	0.05	0.10	0.05	0.04	0.00	0.06	0.07	0.03	0.02	0.01	0.03	**0.36**	0.01	0.01	0.02	0.05	0.05	0.00	0.01	0.02
M	0.05	0.02	0.01	0.01	0.01	0.03	0.02	0.02	0.01	0.09	0.19	0.03	**0.28**	0.03	0.01	0.03	0.05	0.01	0.01	0.09
F	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.04	0.11	0.01	0.02	**0.51**	0.01	0.03	0.02	0.02	0.12	0.04
P	0.08	0.02	0.01	0.02	0.00	0.02	0.03	0.03	0.01	0.01	0.03	0.03	0.01	0.01	**0.55**	0.06	0.04	0.00	0.01	0.02
S	0.12	0.03	0.06	0.04	0.02	0.03	0.04	0.07	0.02	0.02	0.03	0.04	0.01	0.01	0.04	**0.29**	0.10	0.01	0.02	0.03
T	0.10	0.02	0.04	0.03	0.01	0.02	0.04	0.03	0.01	0.04	0.03	0.05	0.02	0.01	0.03	0.12	**0.35**	0.00	0.01	0.06
W	0.02	0.03	0.01	0.01	0.01	0.01	0.01	0.02	0.01	0.01	0.05	0.01	0.01	0.04	0.01	0.03	0.01	**0.63**	0.06	0.02
Y	0.02	0.02	0.03	0.02	0.01	0.01	0.01	0.01	0.05	0.02	0.04	0.01	0.01	0.13	0.01	0.03	0.02	0.02	**0.51**	0.02
W	0.08	0.01	0.01	0.01	0.01	0.01	0.02	0.02	0.01	0.14	0.10	0.02	0.03	0.02	0.01	0.03	0.05	0.01	0.01	**0.40**

$t = 1$

Maximum likelihood can be used to determine the tree and ancestors



- Consider a single site (independent of all others)

Summary

- Various distance metrics available to quantify sequence similarity
  - Fractional (*p*-distance)
  - Poisson corrected
  - Gamma corrected
- Also need to account for chemical nature of sequence
  - Transitions/transversion
  - Codon dynamics
  - “Fixed” DNA models
- Evolutionary models based on real data capture similar trends
- Transition rate matrices help model evolution probabilistically

Reading

- Zvelebil & Baum (2008) *Understanding Bioinformatics*
  - ***Chapter 7*** *(7.1-3)*
  - ***Chapter 8*** *(8.1)*
- Kelley and Didulo, *Computational Biology: A Hypertextbook*
  - ***Chapter 6***

UPGMA—**U**nweighted **P**air-**G**roup **M**ethod using arithmetic **A**verages

Principle: the two sequences i and j with the shortest distance d_{ij} must have have been the last to diverge; their branches the same length, at *half* their distance

$$d_{XY} = \frac{1}{N_X N_Y} \sum_{i \in X, j \in Y} d_{ij}$$

UPGMA—**U**nweighted **P**air-**G**roup **M**ethod using arithmetic **A**verages

Principle: the two sequences i and j with the shortest distance d_{ij} must have have been the last to diverge; their branches the same length, at *half* their distance

$$d_{XY} = \frac{1}{N_X N_Y} \sum_{i \in X, j \in Y} d_{ij}$$

X and Y are clusters containing 1 or more sequences; N_X and N_Y are their sizes

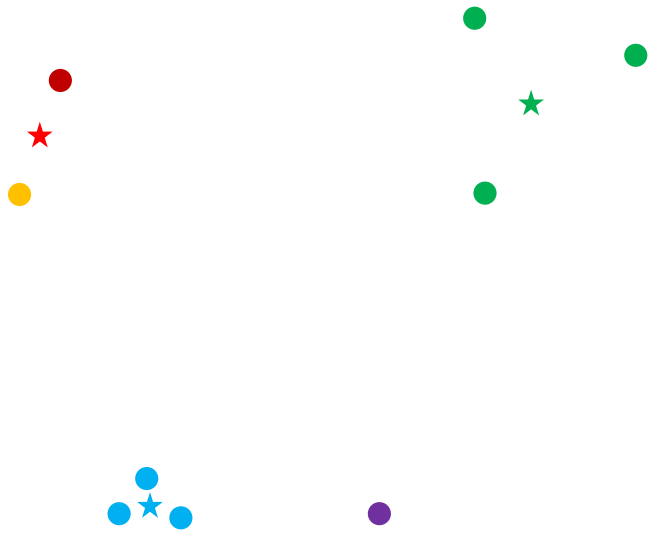
Sum the distances between all pairs of sequences, one from X and one from Y; calculate their arithmetic average

UPGMA—Unweighted Pair-Group Method using arithmetic Averages

Distances between clusters can be determined based on the distances of the two which are merged (computationally more efficient)

$$d_{ZW} = \frac{N_X d_{XW} + N_Y d_{YW}}{N_X + N_Y}$$

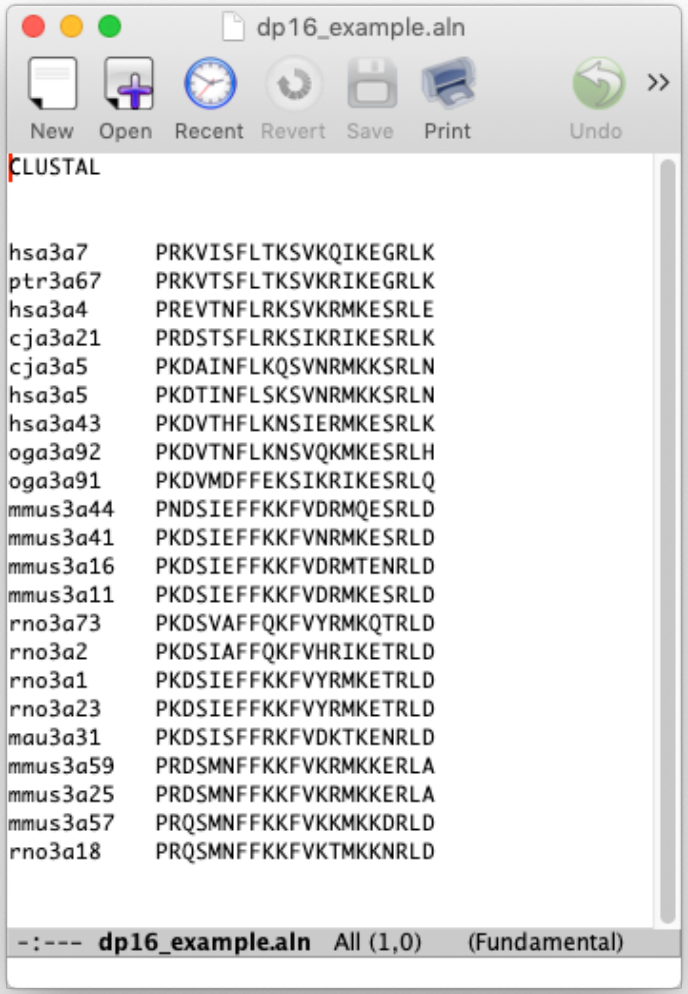
Merge X and Y into new cluster Z,
based on X and Y's distances to others (W)



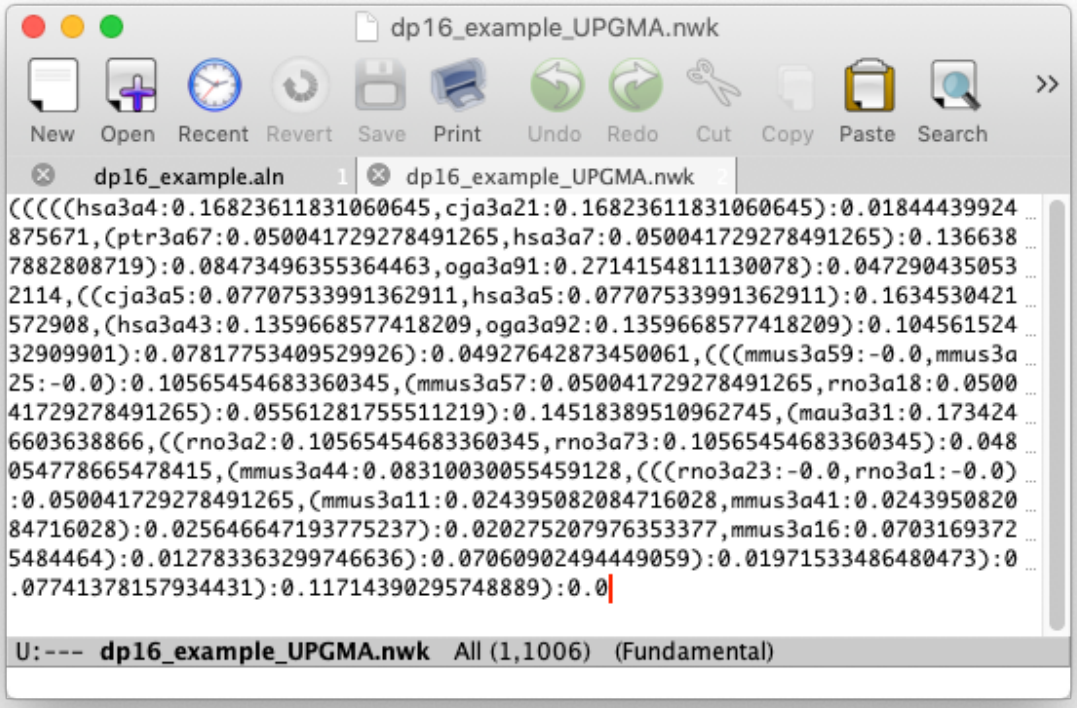
From `phylo.py` (also in `guide.py`)

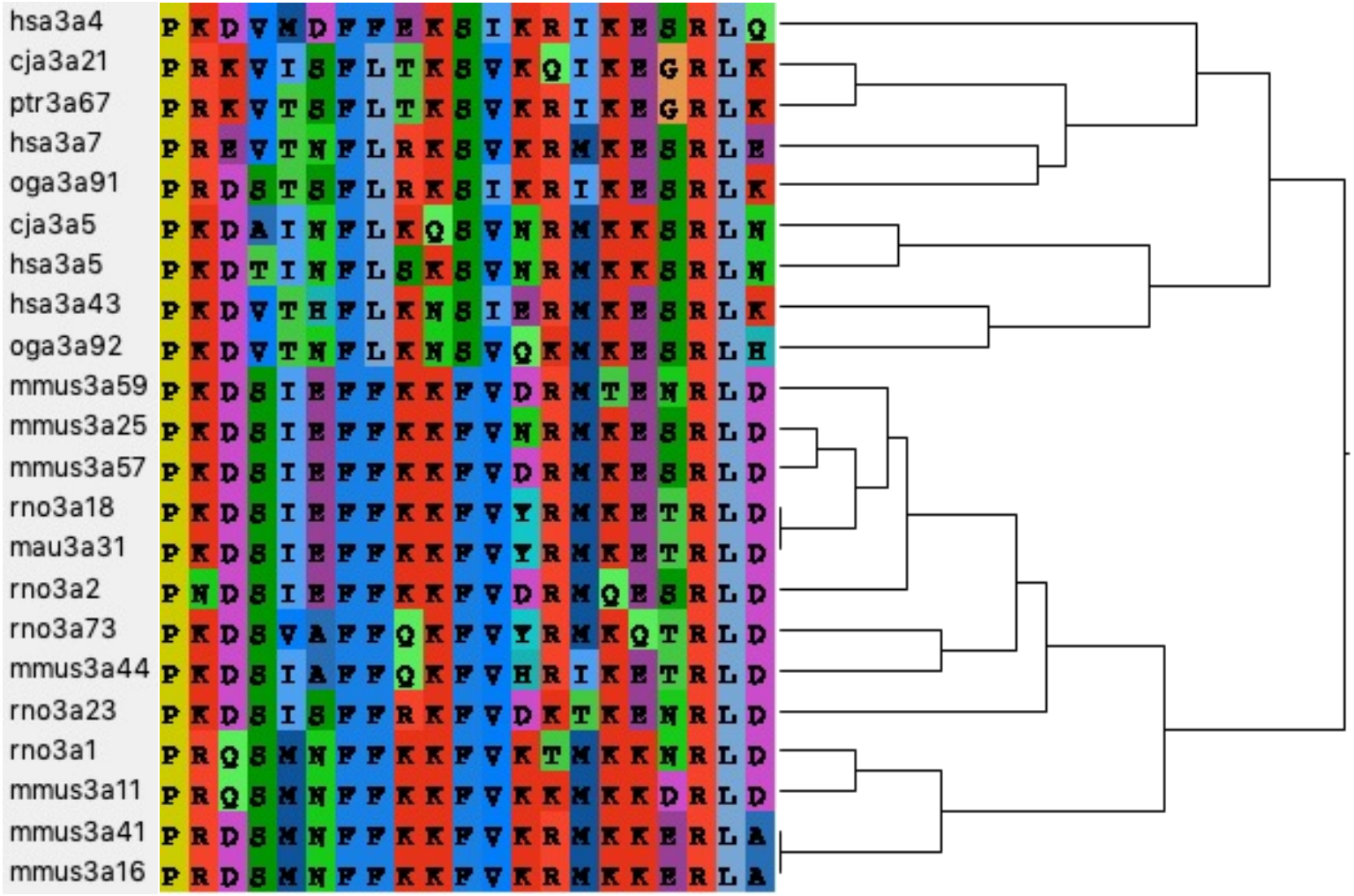
```
""" ------------------------------------------------------------------------------
    Methods for generating a single tree by clustering, here UPGMA Zvelebil and Baum p. 278
    ------------------------------------------------------------------------------"""

def runUPGMA(aln, measure, absoluteDistances=False):
    """ Generate an ultra-metric, bifurcating, rooted tree from an alignment based on pairwise distances.
        Use specified distance metric (see sequence.calcDistances).
        If absoluteDistances is True, the tree will be assigned the total distance from provided species.
        Otherwise, the relative addition at each path will be assigned."""
    D = {}
    N = {}  # The number of sequences in each node
    M = aln.calcDistances(measure)  # determine all pairwise distances
    nodes = [PhyloNode(label=seq.name) for seq in aln.seqs]  # construct all leaf nodes
    """ For each node-pair, assign the distance between them. """
    for i in range(len(nodes)):
        nodes[i].sequence = aln.seqs[i]
        nodes[i].dist = 0.0
        N[nodes[i]] = 1  # each cluster contains a single sequence
        for j in range(0, i):
            D[frozenset([nodes[i], nodes[j]])] = M[i, j]
```

```
""" Treat each node as a cluster, until there is only one cluster left, find the *closest*
    pair of clusters, and merge that pair into a new cluster (to replace the two that merged).
    In each case, the new cluster is represented by the (phylo)node that is formed. """
while len(N) > 1:  # N will contain all "live" clusters, to be reduced to a single below
    closest_pair = (None, None)  # The two nodes that are closest to one another according to supplied metric
    closest_dist = None  # The distance between them
    for pair in D:  # check all pairs which should be merged
        dist = D[pair]
        if closest_dist == None or dist < closest_dist:
            closest_dist = dist
            closest_pair = list(pair)
    # So we know the closest, now we need to merge...
    x = closest_pair[0]  # See Zvelebil and Baum p. 278 for notation
    y = closest_pair[1]
    z = PhyloNode()  # create a new node for the cluster z
    z.dist = D.pop(frozenset([x, y])) / 2.0  # assign the absolute distance, change to relative distance later
    Nx = N.pop(x)  # find number of sequences in x, remove the cluster from list N
    Ny = N.pop(y)  # find number of sequences in y, remove the cluster from list N
    dz = {}  # new distances to cluster z
    for w in N:  # for each node w ...
        # we will merge x and y into a new cluster z, so need to consider w (which is not x or y)
        dxw = D.pop(frozenset([x, w]))  # retrieve and remove distance from D: x to w
        dyw = D.pop(frozenset([y, w]))  # retrieve and remove distance from D: y to w
        dz[w] = (Nx * dxw + Ny * dyw) / (Nx + Ny)  # distance: z to w
    N[z] = Nx + Ny  # total number of sequences in new cluster, insert new cluster in list N
    for w in dz:  # we have to run through the nodes again, now not including the removed x and y
        D[frozenset([z, w])] = dz[w]  # for each "other" cluster, update distance per EQ8.16 (Z&B p. 278)
    x.parent = z
    y.parent = z
    z.children = [x, y]
    nodes.append(z)
```




```
aln = sequence.readClustalFile('dp16_example.aln',
sequence.Protein_Alphabet)
tree = phylo.runUPGMA(aln, "poisson")
phylo.writeNewickFile('dp16_example_UPGMA.nwk', tree)
```




hsa3a4 PKDVMDFFEKSIKRIKESRLQ
cja3a21 PRKVISFLTKSVKQIKEGRLK
ptr3a67 PRKVTSFLTKSVKRIKEGRLK
hsa3a7 PREVTNFLRKSVKRMKESRLE
oga3a91 PRDSTSFLRKSIKRIKESRLK
cja3a5 PKDAINFLKQSVNRMKKSRLN
hsa3a5 PKDTINFLSKSVNRMKKSRLN
hsa3a43 PKDVTHFLKNSIERMKESRLK
oga3a92 PKDVTNFLKNSVQKMKESRLH
mmus3a59 PKDSIEFFKKFVDRMTENRLD
mmus3a25 PKDSIEFFKKFVNRMKESRLD
mmus3a57 PKDSIEFFKKFVDRMKESRLD
rno3a18 PKDSIEFFKKFVYRMKETRLD
mau3a31 PKDSIEFFKKFVYRMKETRLD
rno3a2 PNDSIEFFKKFVDRMQESRLD
rno3a73 PKDSVAFFQKFVYRMKQTRLD
mmus3a44 PKDSIAFFQKFVHRIKETRLD
rno3a23 PKDSISFFRKFVDKTKENRLD
rno3a1 PRQSMNFFKKFVKTMKKNRLD
mmus3a11 PRQSMNFFKKFVKKMKKDRLD
mmus3a41 PRDSMNFFKKFVKRMKKERLA
mmus3a16 PRDSMNFFKKFVKRMKKERLA