1. What kind of information does a phylogenetic tree illustrate?

Phylogenetic tree displays the proposed evolutionary relationship between its members based on their similarity(can be based on different features and principles).

2. What is the difference, in terms of the information displayed, between rooted and unrooted trees?

Rooted tree shows a common ancestor node from which all its branches diverge. The rooted tree also provides information on chronological/evolutionary directionality, based on similarity with the ancestor node, while unrooted tree will only provide information based on similarity between its branches.

3. Use the UPGMA method together with the Hamming distance to build the phylogenetic tree for the following DNA sequences: ACTT, AGGG, GATT, TGGG. Show every step you performed when building the tree. What is one of the main disadvantages of this method?

Plot a table/matrix based on hamming distance between the sequence.

	ACTT	AGGG	GATT	TGGG
ACTT	0	3	2	4
AGGG		0	4	1
GATT			0	4
TGGG				0

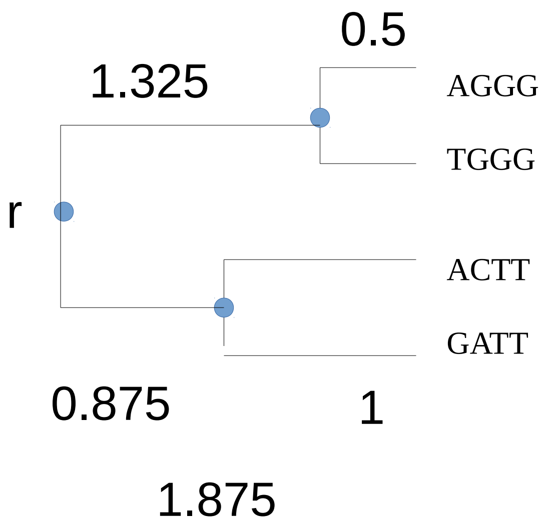
Find the lowest value. This pair is most closely related to each other. The distance to node is 50% of their cumulative distance, = 0.5
Normalize the distance from the pair and the rest of the sequences and collapse the matrix by one row.

	ACTT	R1	GATT
ACTT	0	3.5	2
R1		0	4
GATT			0

Lowest value is between ACTT and GATT and the distance to common node is 1. Normalize the distance and collapse again.

	R1	R2
R1	0	3.75
R2		0

The remaining convergence is between the collapsed nodes (root) and total distance of the tree is 1.875. The branch length between root and R1 node is 1.375. R2 to root is 0.875.



Final tree. Among the most obvious disadvantages is that the collapsed nodes distance to the rest of the tree is normalized and we consider them together when comparing to the rest of the sequences. The tree will not show you how similar the leaves are individually, and distance to root is the same from all leaves(assuming equal rate of molecular clocks and equal sequence age at the time of comparison).

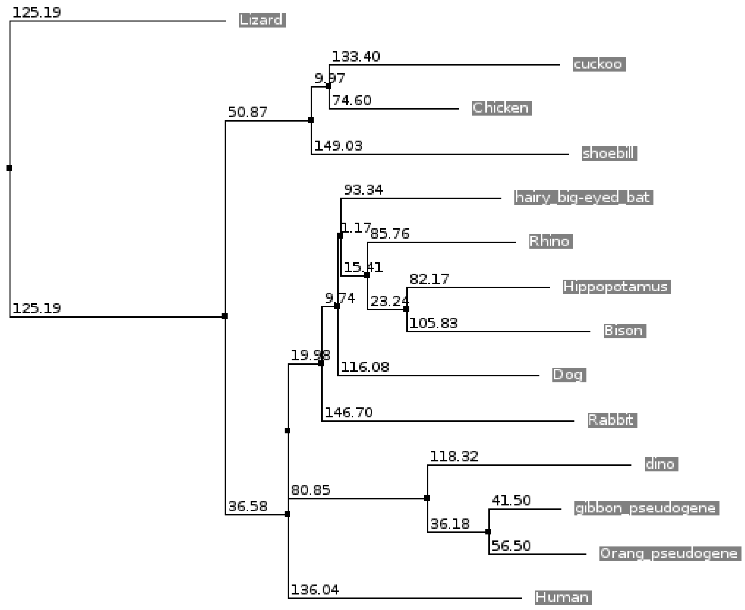
4. Which mouse has the largest evolutionary distance to the common ancestor (tree root)?

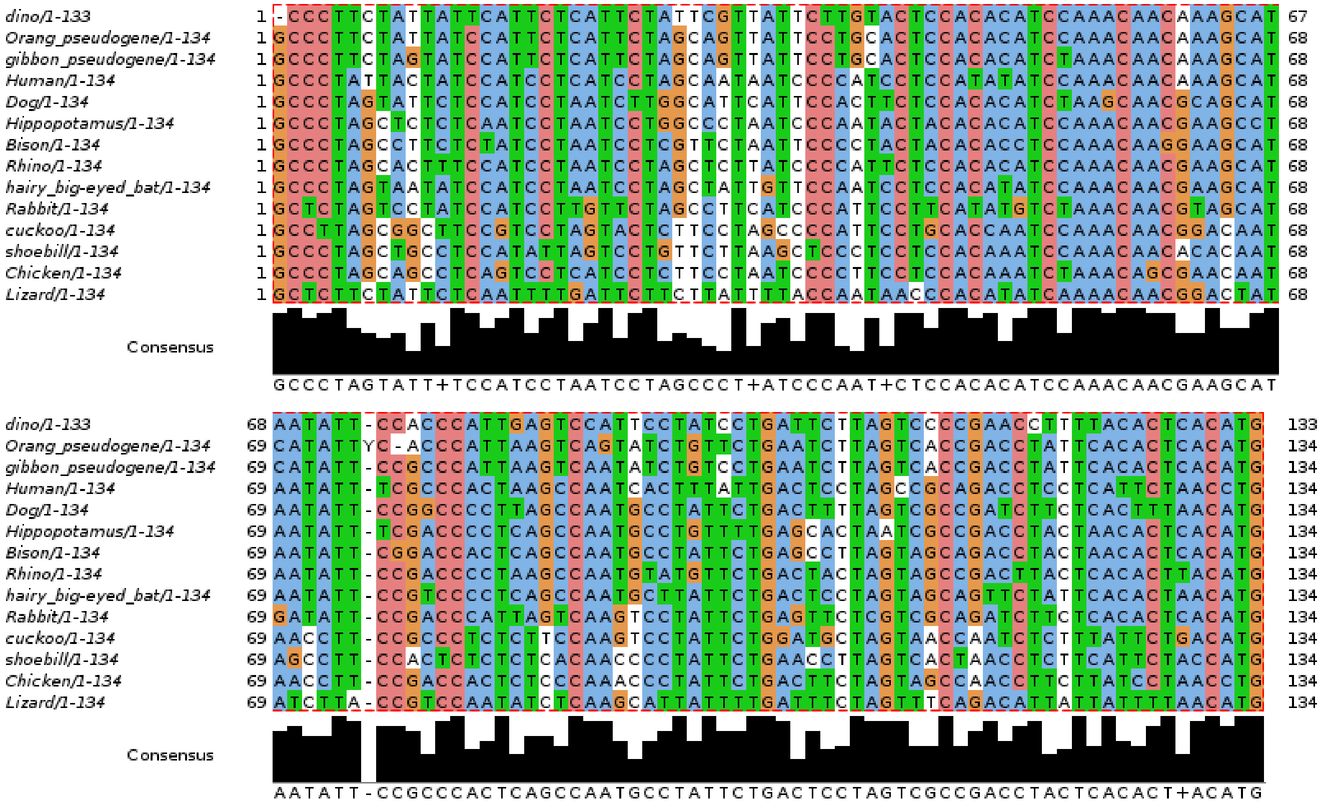
The grey mouse with distance 10.

5. In terms of ordinary time, which mouse diverged first, the grey one or the purple one?

Grey one since it has lowest similarity to others.

6. Use one of the methods shown in the previous lab to align the sequences and construct an evolutionary tree. Create and submit a print screen image of the multiple sequence alignment and the phylogenetic tree.



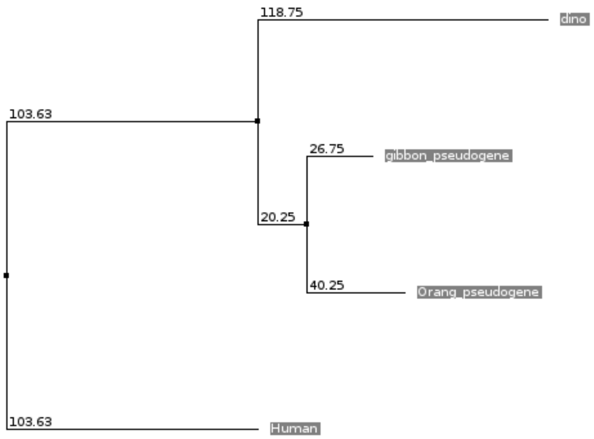


```
dino/1-133                  1 -CCCTTCTATTATTCATTCTCATTCTATTCGTTATTCTTGTACTCCACACATCCAAACAACAAAGCAT 67
Orang_pseudogene/1-134      1 GCCCTTCTATTATCCATTCTCATTCTAGCAGTTATTCCTGCACTCCACACATCCAAACAACAAAGCAT 68
gibbon_pseudogene/1-134     1 GCCCTTCTAGTATCCATTCTCATTCTAGCAGTTATTCCTGCACTCCACACATCTAAACAACAAAGCAT 68
Human/1-134                 1 GCCCTATTACTATCCATCCTCATCCTAGCAATAATCCCATCCTCCATATATCCAAACAACAAAGCAT 68
Dog/1-134                   1 GCCCTAGTATTCTCCATCCTAATCTTGGCATTCATTCCACTTCTCCACACATCTAAGCAACGCAGCAT 68
Hippopotamus/1-134          1 GCCCTAGCTCTCTCAATCCTAATCCTGGCCCTAATCCCAATACTACACACATCCAAACAACGAAGCCT 68
Bison/1-134                 1 GCCCTAGCCTTCTCTATCCTAATCCTCGTTCTAATTCCCCTACTACACACCTCCAAACAAGGAAGCAT 68
Rhino/1-134                 1 GCCCTAGCACTTTCCATCCTAATCCTAGCTCTTATCCCCATTCTCCACACATCCAAACAACGAAGCAT 68
hairy_big-eyed_bat/1-134    1 GCCCTAGTAATATCCATCCTAATCCTAGCTATTGTTCCAATCCTCCACATATCCAAACAACGAAGCAT 68
Rabbit/1-134                1 GCTCTAGTCCTATCCATCCTTGTTCTAGCCTTCATCCCATTCCTTCATATGTCTAAACAACGTAGCAT 68
cuckoo/1-134                1 GCCTTAGCGGCTTCCGTCCTAGTACTCTTCCTAGCCCCATTCCTGCACCAATCCAAACAACGGACAAT 68
shoebill/1-134              1 GCCCTAGCTGCCTCCATATTAGTCCTGTTCTTAAGCTCCCTCCTCCACAAATCCAAACAACACACAAT 68
Chicken/1-134               1 GCCCTAGCAGCCTCAGTCCTCATCCTCTTCCTAATCCCCTTCCTCCACAAATCTAAACAGCGAACAAT 68
Lizard/1-134                1 GCTCTTCTATTCTCAATTTTGATTCTTCTTATTTTACCAATAACCCACATATCAAAACAACGGACTAT 68

                  Consensus
                              GCCCTAGTATT+TCCATCCTAATCCTAGCCCT+ATCCCAAT+CTCCACACATCCAAACAACGAAGCAT

dino/1-133                 68 AATATT-CCACCCATTGAGTCCATTCCTATCCTGATTCTTAGTCCCCGAACCTTTTACACTCACATG 133
Orang_pseudogene/1-134     69 CATATTYC-ACCCATTAAGTCAGTATCTGTTCTGAATCTTAGTCACCGACCTATTCACACTCACATG 134
gibbon_pseudogene/1-134    69 CATATT-CCGCCCATTAAGTCAATATCTGTCCTGAATCTTAGTCACCGACCTATTCACACTCACATG 134
Human/1-134                69 AATATT-TCGCCCACTAAGCCAATCACTTTATTGACTCCTAGCCGCAGACCTCCTCATTCTAACCTG 134
Dog/1-134                  69 AATATT-CCGGCCCCTTAGCCAATGCCTATTCTGACTTTTAGTCGCCGATCTTCTCACTTTAACATG 134
Hippopotamus/1-134         69 AATATT-TCGACCCCTCAGCCAATGCCTGTTTTGAGCACTAATCGCCGACCTACTAACACTCACATG 134
Bison/1-134                69 AATATT-CGGACCACTCAGCCAATGCCTATTCTGAGCCTTAGTAGCAGACCTACTAACACTCACATG 134
Rhino/1-134                69 AATATT-CCGACCCCTAAGCCAATGTATGTTCTGACTACTAGTAGCCGACTTACTCACACTTACATG 134
hairy_big-eyed_bat/1-134   69 AATATT-CCGTCCCCTCAGCCAATGCTTATTCTGACTCCTAGTAGCAGTTCTATTCACACTAACATG 134
Rabbit/1-134               69 GATATT-CCGACCCATTAGTCAAGTCCTATTCTGAGTTCTCGTCGCAGATCTTCTCACACTCACATG 134
cuckoo/1-134               69 AACCTT-CCGCCCTCTCTTCCAAGTCCTATTCTGGATGCTAGTAACCAATCTCTTTATTCTGACATG 134
shoebill/1-134             69 AGCCTT-CCACTCTCTCTCACAACCCCTATTCTGAACCTTAGTCACTAACCTCTTCATTCTACCATG 134
Chicken/1-134              69 AACCTT-CCGACCACTCTCCCAAACCCTATTCTGACTTCTAGTAGCCAACCTTCTTATCCTAACCTG 134
Lizard/1-134               69 ATCTTA-CCGTCCAATATCTCAAGCATTATTTTGATTTCTAGTTTCAGACATTATTATTTTAACATG 134

                  Consensus
                              AATATT-CCGCCCACTCAGCCAATGCCTATTCTGACTCCTAGTCGCCGACCTACTCACACT+ACATG
```

7. Which three species' sequences are the most related to the dinosaur sequence according to your phylogenetic tree?

Orangutan, Gibbon, Human sequences

8. What can you conclude from your observations, in terms of biology and evolution?



It is unlikely the sample belonged to a dinosaur. The sample does not closely match with human sequence either – so it is also unlikely to be a laboratory contamination. The sample might belong to a primate, who (judging only by the tree) is descendant from common ancestor of orangutan, gibbon and human, but diverged later than human branch. However, this kind of tree assumes all members were compared at the same time, and I cannot exclude that the specimen was a direct ancestor of Orangutan/Gibbon based on just one sequence.

9. What do you think is a protein domain?

A protein domain is a part of the protein (or can be the entire protein) that can function and be folded independently from the rest of the chain and domains.

10. What do you think is a protein fold?

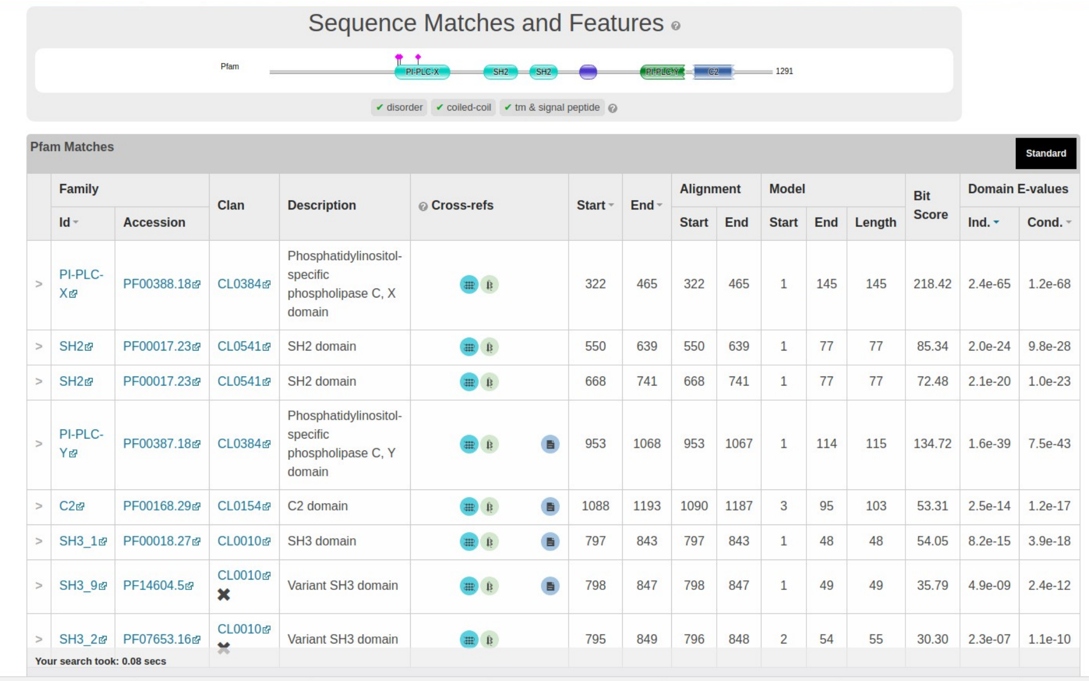
A protein fold is a “stereotypical” tertiary conformation into which an aminoacid chain may collapse. Folds could be of similar structure but of differing aminoacid sequence. A fold could refer to the same part/size of the chain used to describe a protein domain, but those are not synonyms.

11. What procedure does Pfam use to create sequence families and to add new members to existing families? What do you think is the role of HMM profiles in this process?

Pfam organises sequences, sourced from Uniprot database, into families(by sequence similarity) and clans(families with proposed common evolutionary origin). The novel grouping is based on MSA and HMM profiles, but many families and their profiles are based on Prosite data . HMM profiles contain statistically likely motifs for the given family (15 members needed to generate) and can be used to scrutinize the novel candidates. HMMs are also used to organize some related families into higher order clans, if at least two of their HMM profiles could somewhat match the other’s for the same region in the sequences.

12. Find the entry "PLCG1_BOVIN" in the Uniprot (SwissProt) database. Get its peptide sequence and search it using the Pfam Search utility. How many non-overlapping Pfam-A domains does it have? Include a screen shot of the Pfam predictions for this protein in your report.

6 domains.



Sequence Matches and Features

Pfam Matches

	Family		Clan	Description	Cross-refs	Start	End	Alignment		Model			Bit Score	Domain E-values	
	Id	Accession						Start	End	Start	End	Length		Ind.	Cond.
>	PI-PLC-X	PF00388.18	CL0384	Phosphatidylinositol-specific phospholipase C, X domain		322	465	322	465	1	145	145	218.42	2.4e-65	1.2e-68
>	SH2	PF00017.23	CL0541	SH2 domain		550	639	550	639	1	77	77	85.34	2.0e-24	9.8e-28
>	SH2	PF00017.23	CL0541	SH2 domain		668	741	668	741	1	77	77	72.48	2.1e-20	1.0e-23
>	PI-PLC-Y	PF00387.18	CL0384	Phosphatidylinositol-specific phospholipase C, Y domain		953	1068	953	1067	1	114	115	134.72	1.6e-39	7.5e-43
>	C2	PF00168.29	CL0154	C2 domain		1088	1193	1090	1187	3	95	103	53.31	2.5e-14	1.2e-17
>	SH3_1	PF00018.27	CL0010	SH3 domain		797	843	797	843	1	48	48	54.05	8.2e-15	3.9e-18
>	SH3_9	PF14604.5	CL0010 ✖	Variant SH3 domain		798	847	798	847	1	49	49	35.79	4.9e-09	2.4e-12
>	SH3_2	PF07653.16	CL0010 ✖	Variant SH3 domain		795	849	796	848	2	54	55	30.30	2.3e-07	1.1e-10

Your search took: 0.08 secs