1 Форматы файлов

1.1 Task 1

Сконвертируйте sam файл в bam файл и создайте для полученного bam файла индекс. Файл: task_adh1b.sam.

```
samtools view -S -b /srv/common/midterm/task_adh1b.sam > ~/test1/task1.bam
samtools sort task1.bam -o sorted_task1.bam
samtools index sorted_task1.bam
```

1.2 Task 2

Определите сборку bam файла полученного вам в первом задании. В ответе приведите сборку.

```
samtools view -H sorted_task1.bam | grep 'LN:249250621
```

Результат: hg19

1.3 Task 3

Определите, есть ли у человека непереносимость лактозы с помощью IGV по bam: task_lactose.bam
В ответе приведите мутации которые смотрели, их генотип и финальный вывод. Непереносимости

Mutation	REF Sequence	Sample Variants
rs4988235	G/G	G/G
rs41380347	A/A	A/A
rs145946881	C/C	C/C
rs41525747	G/G	G/G
rs121908937	C/C	C/C
rs121908936	-	-

нет

1.4 Task 4

Найдите количество всех ридов, выровненных на хромосому 4, в bam файле: task_adh1b.bam.
В ответе приведите код и количество ридов.

```
samtools view /srv/common/midterm/task_adh1b.bam chr4 | wc -l
```

Результат: 78105

1.5 Task 5

Определите пол человека по bam файлу: task_gender.bam. В ответе приведите код и пол.

```
samtools view /srv/common/midterm/task_gender.bam chrX | wc -l
```

Результат: 592941

```
samtools view /srv/common/midterm/task_gender.bam chrY | wc -l
```

Результат: 380490

1.6 Task 6

Сделайте bed файл включающий только регион гена ADH1B. Посчитайте среднее покрытие данного региона для файла полученного в первом задании. При помощи bedtools определите, какая доля гена ADH1B имеет покрытие x1+. В ответе приведите среднее покрытие, долю гена с покрытием x1+ и код.

1.7 Task 7

Подсчитайте количество позиций в файле chip.vcf.gz. В ответе приведете количество позиций и код.

```
view -H /srv/common/midterm/chip.vcf.gz | wc -l
```

Результат: 1001385

1.8 Task 8

Используя bcftools, извлеките все варианты для интервала 5215000-5233000 хромосомы 21 из VCF файла: chip.vcf.gz. В ответе приведите количество вариантов.

```
bcftools index chip.vcf.gz
bcftools view -r chr21:5215000-5233000 chip.vcf.gz | grep -v "^#" | wc -l
```

Результат: 97

1.9 Task 9

В ответе запишите следующие данные: хромосома, позиция, референсный, альтернативный аллель, генотип для вариантов с гетерозиготой или гомозиготой по альтернативному аллелю у образца NA21135 в том же регионе что в прошлом задании для файла: chip.vcf.gz

```
bcftools view -r chr21:5215000-5233000 chip.vcf.gz | bcftools query -s NA21135 -f
'%CHROM\t%POS\t%REF\t%ALT\t[%GT]\n' | grep '1|1'
bcftools view -r chr21:5215000-5233000 chip.vcf.gz | bcftools query -s NA21135 -f
'%CHROM\t%POS\t%REF\t%ALT\t[%GT]\n' | grep '0|1'
```

Chromosome	Position	Reference Allele	Alternate Allele	Genotype
chr21	5219624	C	A	0|1
chr21	5231730	C	G	0|1
chr21	5225197	G	T	1|1

1.10 Task 10

Отфильтруйте позиции по колонке INFO/AF. Оставьте варианты с частотой не менее 5%. В ответе запишите количество вариантов.

```
bcftools view -r chr21:5215000-5233000 -i 'INFO/AF>=0.05' chip.vcf.gz | grep -v "^#" | wc -l
```

Результат: 5

1.11 Task 11

В ответе запишите генотип всех образцов в буквенном формате (пример: 0/1 стало AG) для позиций
chr21, pos 5231680, REF:T, ALT:C
chr21, pos 5225197, REF:G, ALT:T

2 Глобальные и локальные выравнивания

2.1 Task 1

Выровняйте следующие последовательности с помощью алгоритма Нидлмана-Вунша:
ATGCCCGA
GTCACCC
Используйте следующие параметры для выравнивания: награда за совпадение: +1, штраф за замену: -1, штраф за вставку или удаление: -2.

Формула для заполнения:

$$\text{Score}(i,j) = \max \begin{cases} \text{Score}(i-1, j-1) + \text{match/mismatch}, \\ \text{Score}(i-1, j) + \text{gap penalty}, \\ \text{Score}(i, j-1) + \text{gap penalty} \end{cases}$$

Получим таблицу:

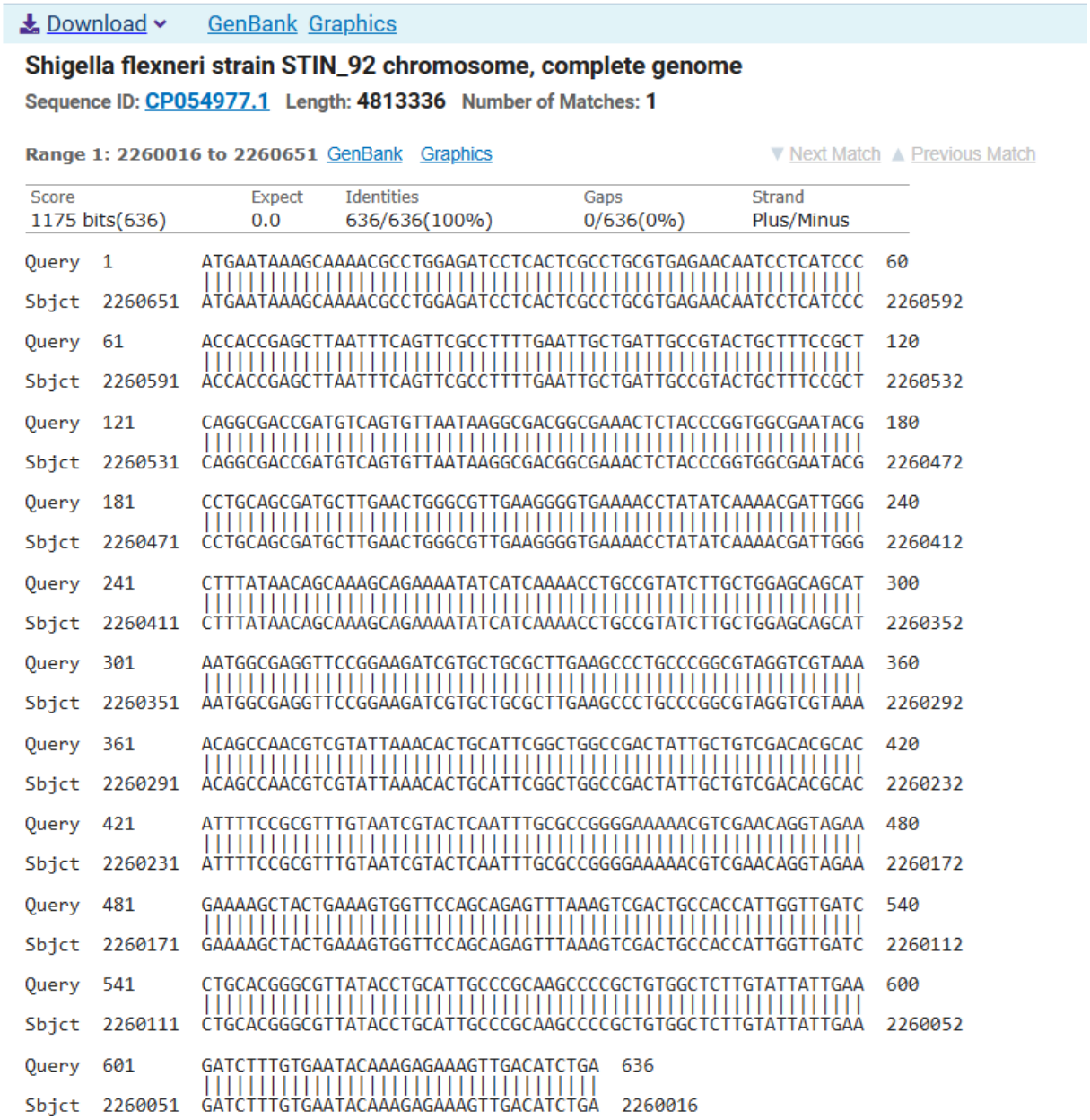
		G	T	C	A	C	C	C
	A	-1	-3	-5	-7	-9	-11	-13
	T	-3	-1	-2	-4	-6	-8	-10
	G	-5	-3	-2	-4	-6	-8	-10
	C	-7	-5	-3	-2	-4	-6	-8
	C	-9	-7	-5	-4	-2	-4	-6
	C	-11	-9	-7	-6	-4	-2	-4
	G	-13	-11	-9	-8	-6	-4	-2
	A	-15	-13	-11	-7	-8	-6	-4

то есть выравнивание выглядит так:
ATGCCCGA
—TC—ACCC

2.2 Task 2

Найдите нуклеотидную последовательность белка эндонуклеазы III (Nth) из бактерии Escherichia coli, штамм K-12, субштамм MG1655, в базе данных NCBI. Проведите соответствующий последовательности BLAST. Напишите какой организм имеет лучший мэтч с данной последовательн помимо Escherichia coli. Выпишите Max Score и E-value. В ответе приведите организм, max score и e-value.

нуклеотидная последовательность белка эндонуклеазы III (Nth) из бактерии Escherichia coli, штамм K-12, субштамм MG1655, бластнув получим Shigella flexneri strain STIN_92 chromosome



Download GenBank Graphics

Shigella flexneri strain STIN_92 chromosome, complete genome

Sequence ID: CP054977.1 Length: 4813336 Number of Matches: 1

Range 1: 2260016 to 2260651 GenBank Graphics ▼ Next Match ▲ Previous Match

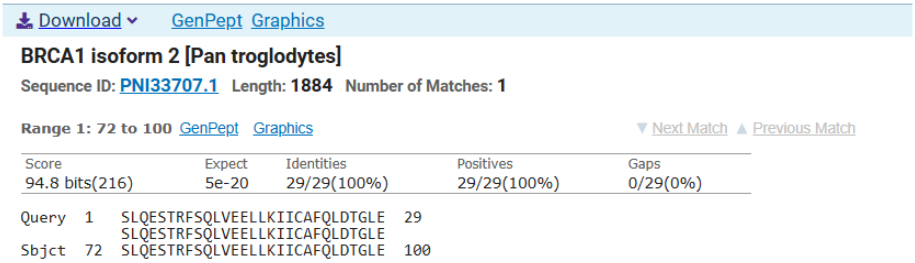
Score	Expect	Identities	Gaps	Strand
1175 bits(636)	0.0	636/636(100%)	0/636(0%)	Plus/Minus

```
Query  1        ATGAATAAAGCAAAACGCCTGGAGATCCTCACTCGCCTGCGTGAGAACAATCCTCATCCC  60
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2260651  ATGAATAAAGCAAAACGCCTGGAGATCCTCACTCGCCTGCGTGAGAACAATCCTCATCCC  2260592

Query  61       ACCACCGAGCTTAATTTCAGTTCGCCTTTTGAATTGCTGATTGCCGTACTGCTTTCCGCT  120
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2260591  ACCACCGAGCTTAATTTCAGTTCGCCTTTTGAATTGCTGATTGCCGTACTGCTTTCCGCT  2260532

Query  121      CAGGCGACCGATGTCAGTGTTAATAAGGCGACGGCGAAACTCTACCCGGTGGCGAATACG  180
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2260531  CAGGCGACCGATGTCAGTGTTAATAAGGCGACGGCGAAACTCTACCCGGTGGCGAATACG  2260472

Query  181      CCTGCAGCGATGCTTGAACTGGGCGTTGAAGGGGTGAAAACCTATATCAAAACGATTGGG  240
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2260471  CCTGCAGCGATGCTTGAACTGGGCGTTGAAGGGGTGAAAACCTATATCAAAACGATTGGG  2260412

Query  241      CTTTATAACAGCAAAGCAGAAAATATCATCAAAACCTGCCGTATCTTGCTGGAGCAGCAT  300
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2260411  CTTTATAACAGCAAAGCAGAAAATATCATCAAAACCTGCCGTATCTTGCTGGAGCAGCAT  2260352

Query  301      AATGGCGAGGTTCCGGAAGATCGTGCTGCGCTTGAAGCCCTGCCCGGCGTAGGTCGTAAA  360
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2260351  AATGGCGAGGTTCCGGAAGATCGTGCTGCGCTTGAAGCCCTGCCCGGCGTAGGTCGTAAA  2260292

Query  361      ACAGCCAACGTCGTATTAAACACTGCATTCGGCTGGCCGACTATTGCTGTCGACACGCAC  420
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2260291  ACAGCCAACGTCGTATTAAACACTGCATTCGGCTGGCCGACTATTGCTGTCGACACGCAC  2260232

Query  421      ATTTTCCGCGTTTGTAATCGTACTCAATTTGCGCCGGGGAAAAACGTCGAACAGGTAGAA  480
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2260231  ATTTTCCGCGTTTGTAATCGTACTCAATTTGCGCCGGGGAAAAACGTCGAACAGGTAGAA  2260172

Query  481      GAAAAGCTACTGAAAGTGGTTCCAGCAGAGTTTAAAGTCGACTGCCACCATTGGTTGATC  540
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2260171  GAAAAGCTACTGAAAGTGGTTCCAGCAGAGTTTAAAGTCGACTGCCACCATTGGTTGATC  2260112

Query  541      CTGCACGGGCGTTATACCTGCATTGCCCGCAAGCCCCGCTGTGGCTCTTGTATTATTGAA  600
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2260111  CTGCACGGGCGTTATACCTGCATTGCCCGCAAGCCCCGCTGTGGCTCTTGTATTATTGAA  2260052

Query  601      GATCTTTGTGAATACAAAGAGAAAGTTGACATCTGA  636
                ||||||||||||||||||||||||||||||||||||
Sbjct  2260051  GATCTTTGTGAATACAAAGAGAAAGTTGACATCTGA  2260016
```

2.3 Task 3

Найдите белковую последовательность человеческого (Homo sapiens) гена BRCA1, в базе данных белков NCBI. Проведите соответствующий последовательности BLAST. Выпишите Max Score и E-value лучшего мэтча с отличным от организма запроса. Выполните парное выравнивание мэтча с исходным запросом и сохраните получившийся dotplot. В ответе приведите. В ответе приведите организм, max score, e-value и dotplot

BRCA1



3 Множественные выравнивания

3.1 Task 1

Возьмите белковую последовательность человеческого гена BRCA1. Выполните белковый BLAST с использованием базы данных Reference proteins database (Refseq protein). Из полученных результатов выберите последовательности для 4ех любых видов (Homo sapiens, Gorilla gorilla gorilla и т.д.). Получите последовательности в формате FASTA. Сократите названия чтобы они содержали только название белка и вид (например, BRCA1_Homo_sapiens).

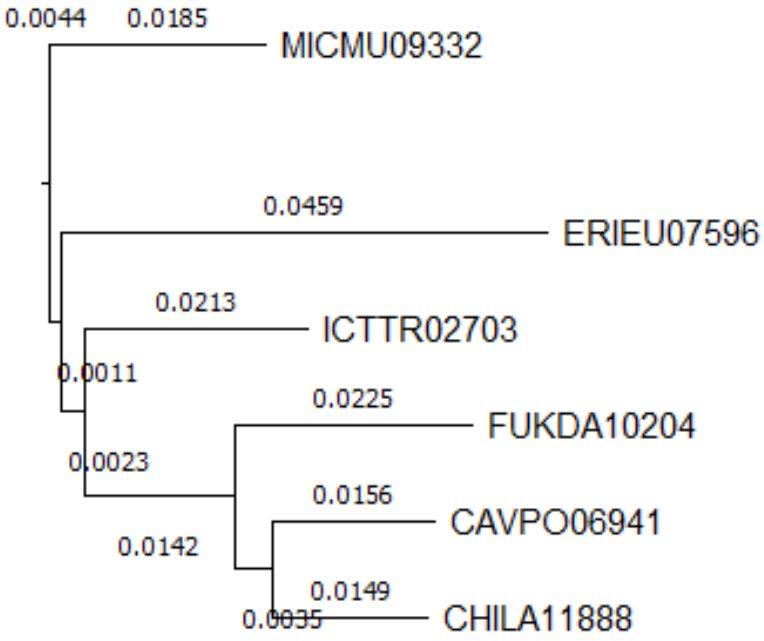
Возьмем:
BRCA1 isoform 2 [Pan troglodytes] Sequence ID: PNI33707.1,
breast cancer type 1 susceptibility protein homolog isoform X1 [Gorilla gorilla gorilla] Sequence ID: XP_030867412.3,
breast cancer type 1 susceptibility protein [Pan paniscus] Sequence ID: NP_001288687.1,
breast cancer type 1 [Gorilla gorilla] Sequence ID: AAT44835.1

```
>BRCA1 Pan troglodytes
MDLSALRVEEVQNVINAMQKILECPICLELIKEPVSTKCDHIFCKFCMLKLLNQKKG
>BRCA1 Gorilla gorilla gorilla
MDLSALRVEEVQNVINAMQKILECPICLELIKEPVSTKCDHIFCKFCMLKLLNQKKG
>BRCA1 Pan paniscus
MDLSALRVEEVQNVINAMQKILECPICLELIKEPVSTKCDHIFCKFCMLKLLNQKKG
>BRCA1 Gorilla gorilla
MDLSALRVEEVQNVINAMQKILECPICLELIKEPVSTKCDHIFCKFCMLKLLNQKKG

```

4 Филогенетика

4.1 Task 1

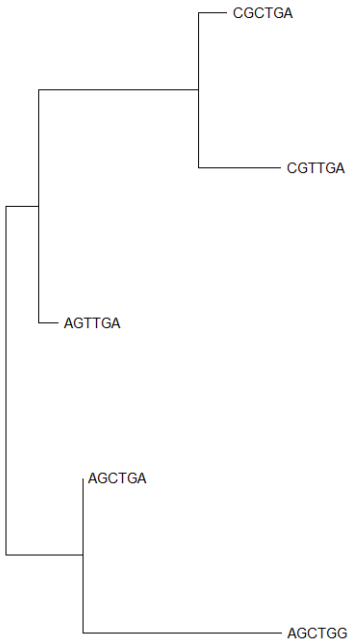
Нарисуйте произвольное ультраметрическое укорененное древо с 6-ю листьями. В ответе приведите фотографию/скриншот.



4.2 Task 2

Постройте филогенетические древо используя расстояние по Хэммингу и посчитав матрицы расстояний, для следующих последовательностей. Приведите матрицы расстояний и древо в качестве ответа

	AGCTGA	AGTTGA	CGCTGA	AGCTGG	CGTTGA
AGCTGA	0	1	1	1	2
AGTTGA	1	0	2	2	1
CGCTGA	1	2	0	2	1
AGCTGG	1	2	2	0	2
CGTTGA	2	1	1	2	0



4.3 Task 3

Постройте дерево записанное в Newick формате (A1:0.1,(A2:0.2,(A3:0.3,A4:0.4):0.2):0.1,A5:0.5);. В ответе приведите древо

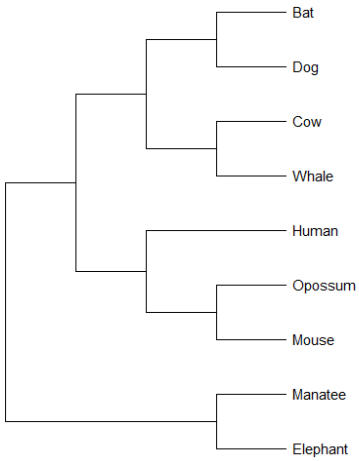
4.4 Task 4

Постройте дерево для файла Mammals.fasta в программе MEGA, методами UPGMA, Neighbourhood Joining, Maximum Parsimony. Укорените деревья в нужном месте на ваш взгляд. Проанализировав ВСЕ деревья в ответе приведите ответы на следующие вопросы:

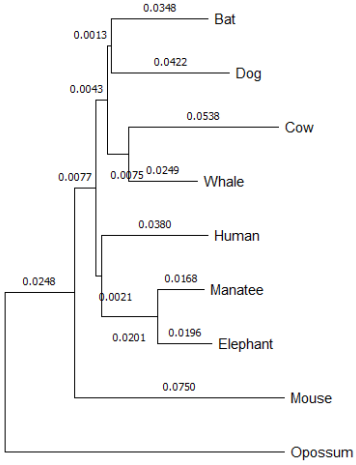
a. Кто ближайший сосед человеку – мышь или собака?

b. Можем ли мы доказать независимое происхождение ламантинов и китообразных?

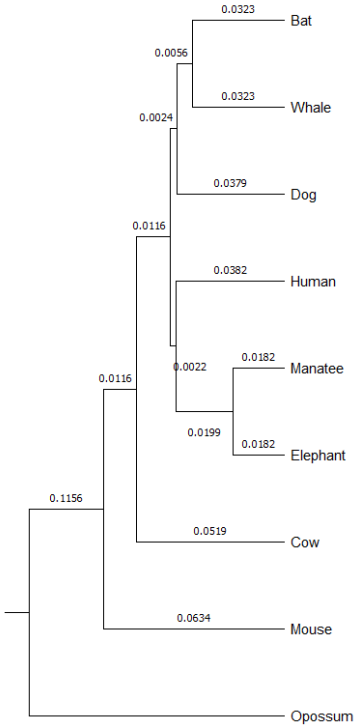
c. Кто ближайший сосед летучим мышам – собака или человек?



MP



NJ



UPGMA

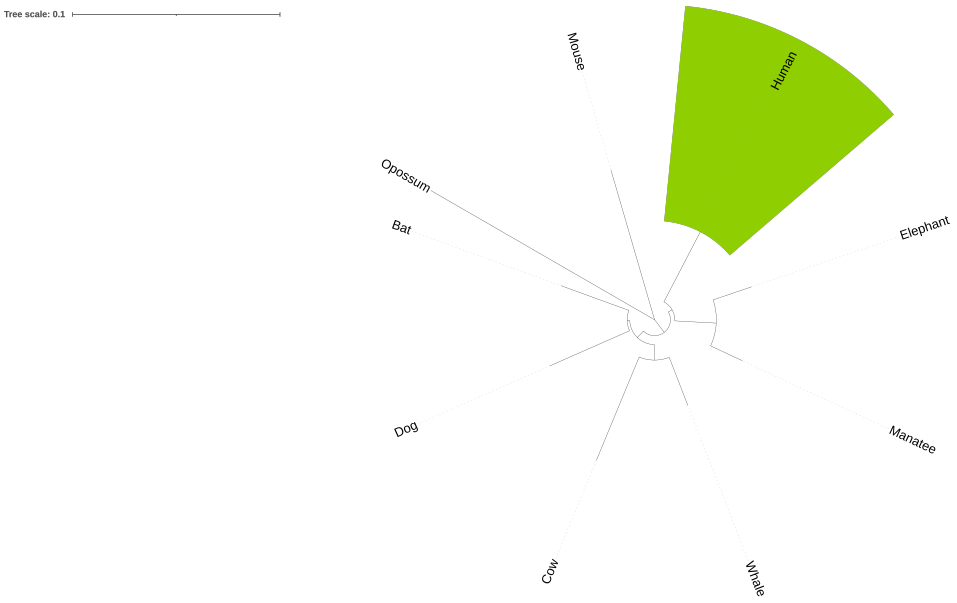
- Мышь
- Да
- Человек

- Собака
- Нет
- Человек

- Собака
- Нет
- Собака

4.5 Task 5

Визуализируйте любое дерево на ваш вкус из задания 4 в программе iTOL, сделав его круговым и покрасив узел с человеком в зеленый цвет. В ответе приведите скриншот.



4.6 Task 6

Используя программу MEGA, постройте древо методом максимального правдоподобия (ultra fast bootstrap with 1000 replicates) для результата множественного выравнивания из задания “Множественные выравнивания”. В ответ приведите дерево.