Отчёт по домашнему заданию номер 6
Иванов Данил. Анализ данных в биологии и медицине

1. Выбрать ген

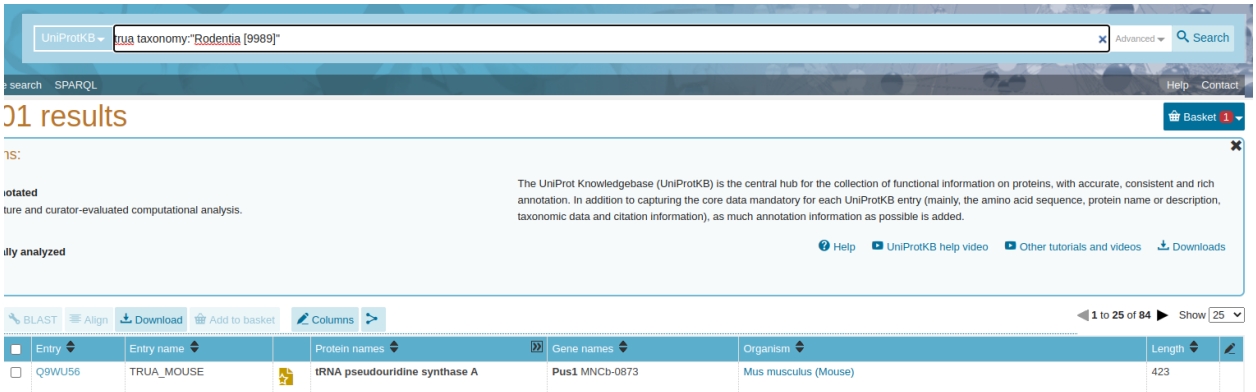
Был выбран ген PUS1 (truA)

2. Найти аминокислотные последовательности генов для нижеперечисленных видов. Составить fasta-файл с последовательностями.

Были включены следующие виды:

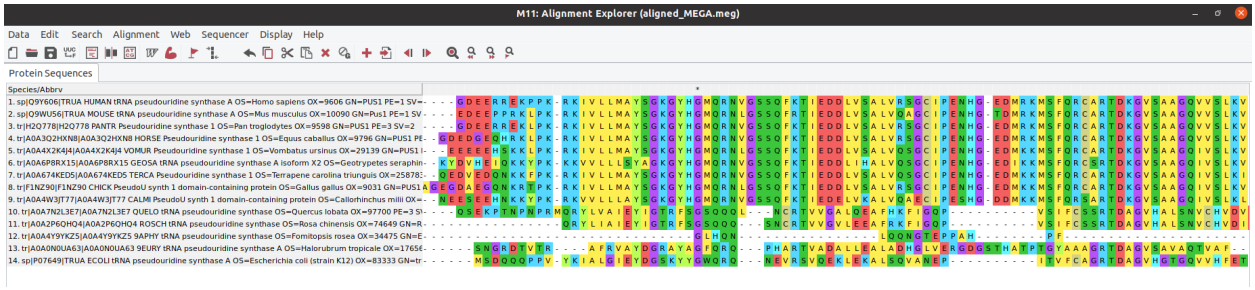
- Человек
- Обезьяна - *Pan troglodytes (Chimpanzee)*
- Грызуны - Mus musculus (Mouse)
- Копытное - *Equus caballus (Horse)*
- Сумчатое - Vombatus ursinus (Common wombat)
- Земноводное - Geotrypetes seraphini (Gaboon caecilian) (Caecilia seraphini)
- Рептилия - Terrapene carolina triunguis (Three-toed box turtle)
- Птица - *Gallus gallus (Chicken)*
- Рыба - *Callorhinchus milii (Ghost shark)*
- Дерево - *Quercus lobata (Valley oak)*
- трава/цветок - *Rosa chinensis (China rose)*
- Грибок - Fomitopsis rosea
- Архея - *Halorubrum tropicale*
- Бактерия - *Escherichia coli*

Для поиска и получения последовательностей использовалась база Uniprot.



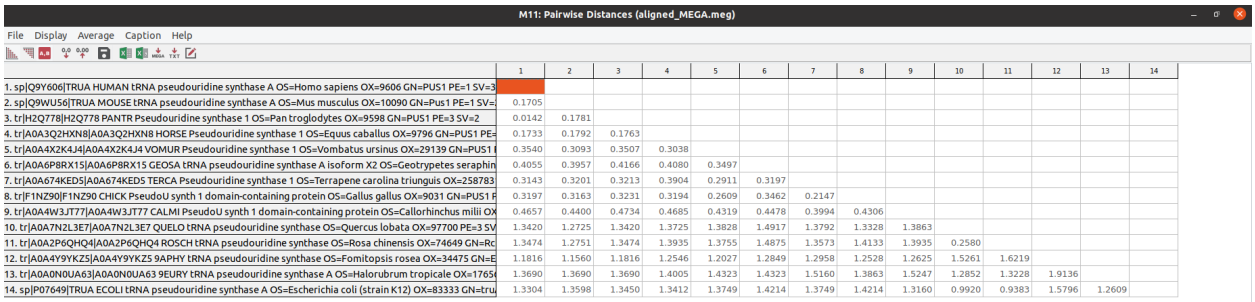
Последовательности в формате FASTA были также получены через UniPROT. Последовательности были скопированы в один файл “seqs_for_MEGA.fa”, файл прилагается.

3. Произведено множественное выравнивание последовательностей алгоритмом Muscle (из Mega). Выравнивание сохранено в файл “aligned_MEGA.meg”, файл прилагается.



4. Построить филогенетическое дерево для аминокислотных последовательностей с бутстрэп-анализом:

На основании выравниваний была построена матрица расстояний:



	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. sp|Q9Y606|TRUA HUMAN tRNA pseudouridine synthase A OS=Homo sapiens OX=9606 GN=PUS1 PE=1 SV=3														
2. sp|Q9WU56|TRUA MOUSE tRNA pseudouridine synthase A OS=Mus musculus OX=10090 GN=Pus1 PE=1 SV=	0.1705													
3. tr|H2Q778|H2Q778 PANTR Pseudouridine synthase 1 OS=Pan troglodytes OX=9598 GN=PUS1 PE=3 SV=2	0.0142	0.1781												
4. tr|A0A3Q2HXN8|A0A3Q2HXN8 HORSE Pseudouridine synthase 1 OS=Equus caballus OX=9796 GN=PUS1 PE=	0.1733	0.1792	0.1763											
5. tr|A0A4X2K4J4|A0A4X2K4J4 VOMUR Pseudouridine synthase 1 OS=Vombatus ursinus OX=29139 GN=PUS1	0.3540	0.3093	0.3507	0.3038										
6. tr|A0A6P8RX15|A0A6P8RX15 GEOSA tRNA pseudouridine synthase A isoform X2 OS=Geotrypetes seraphin	0.4055	0.3957	0.4166	0.4080	0.3497									
7. tr|A0A674KED5|A0A674KED5 TERCA Pseudouridine synthase 1 OS=Terrapene carolina triunguis OX=258783	0.3143	0.3201	0.3213	0.3904	0.2911	0.3197								
8. tr|F1NZ90|F1NZ90 CHICK PseudoU synth 1 domain-containing protein OS=Gallus gallus OX=9031 GN=PUS1 F	0.3197	0.3163	0.3231	0.3194	0.2609	0.3462	0.2147							
9. tr|A0A4W3JT77|A0A4W3JT77 CALMI PseudoU synth 1 domain-containing protein OS=Callorhinchus milii OX	0.4657	0.4400	0.4734	0.4685	0.4319	0.4478	0.3994	0.4306						
10. tr|A0A7N2L3E7|A0A7N2L3E7 QUELO tRNA pseudouridine synthase OS=Quercus lobata OX=97700 PE=3 SV	1.3420	1.2725	1.3420	1.3725	1.3828	1.4917	1.3792	1.3328	1.3863					
11. tr|A0A2P6QHQ4|A0A2P6QHQ4 ROSCH tRNA pseudouridine synthase OS=Rosa chinensis OX=74649 GN=Rc	1.3474	1.2751	1.3474	1.3935	1.3755	1.4875	1.3573	1.4133	1.3935	0.2580				
12. tr|A0A4Y9YKZ5|A0A4Y9YKZ5 9APHY tRNA pseudouridine synthase OS=Fomitopsis rosea OX=34475 GN=E	1.1816	1.1560	1.1816	1.2546	1.2027	1.2849	1.2958	1.2528	1.2625	1.5261	1.6219			
13. tr|A0A0N0UA63|A0A0N0UA63 9EURY tRNA pseudouridine synthase A OS=Halorubrum tropicale OX=1765	1.3690	1.3690	1.3690	1.4005	1.4323	1.4323	1.5160	1.3863	1.5247	1.2852	1.3228	1.9136		
14. sp|P07649|TRUA ECOLI tRNA pseudouridine synthase A OS=Escherichia coli (strain K12) OX=83333 GN=tru	1.3304	1.3598	1.3450	1.3412	1.3749	1.4214	1.3749	1.4214	1.3160	0.9920	0.9383	1.5796	1.2609	

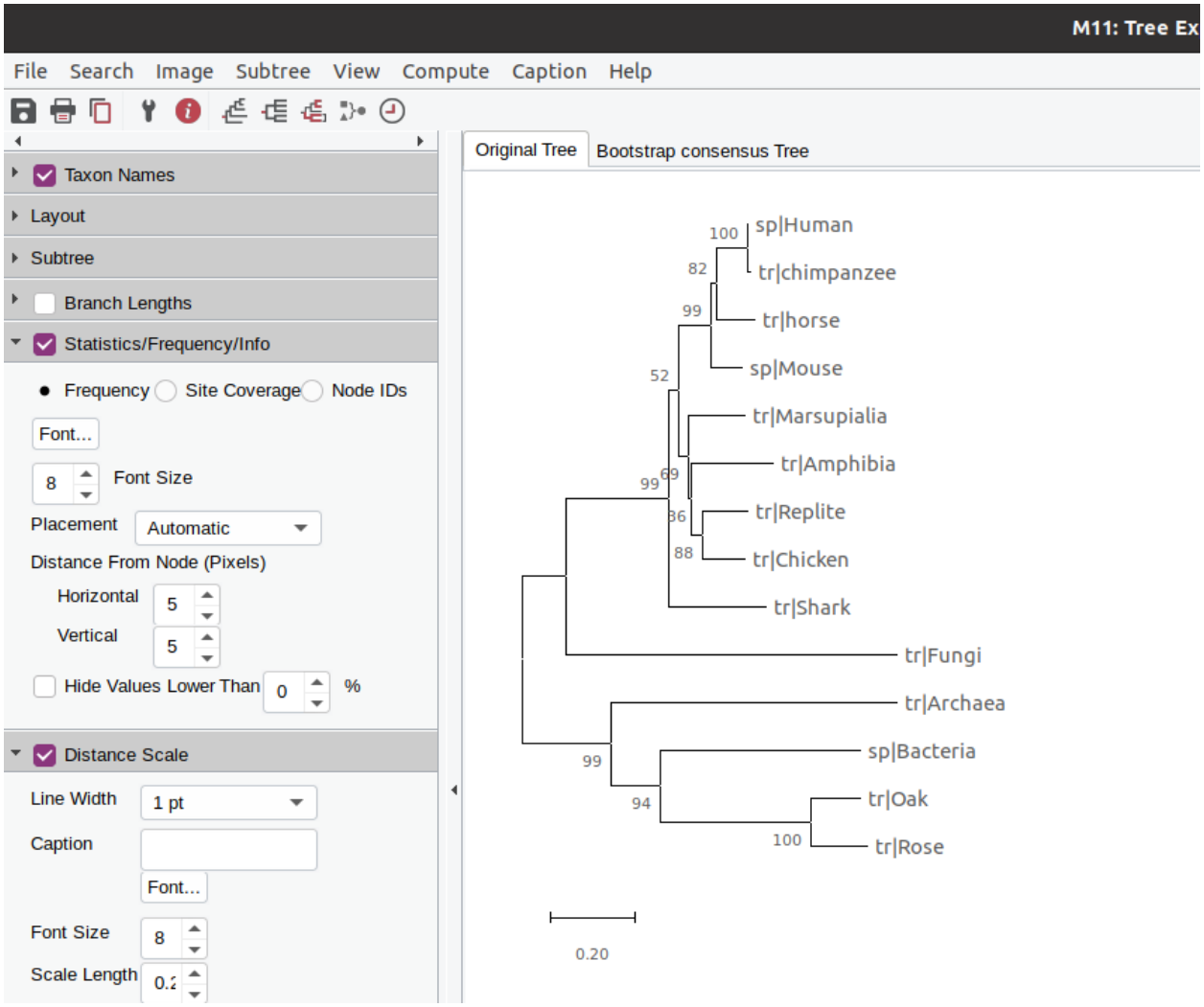
Для простоты в файле .fa виды были переименованы, и построена новая матрица. Деревья будут строиться с упрощёнными названиями.

M11: Pairwise Distances (aligned_MEGA_mod.meg)

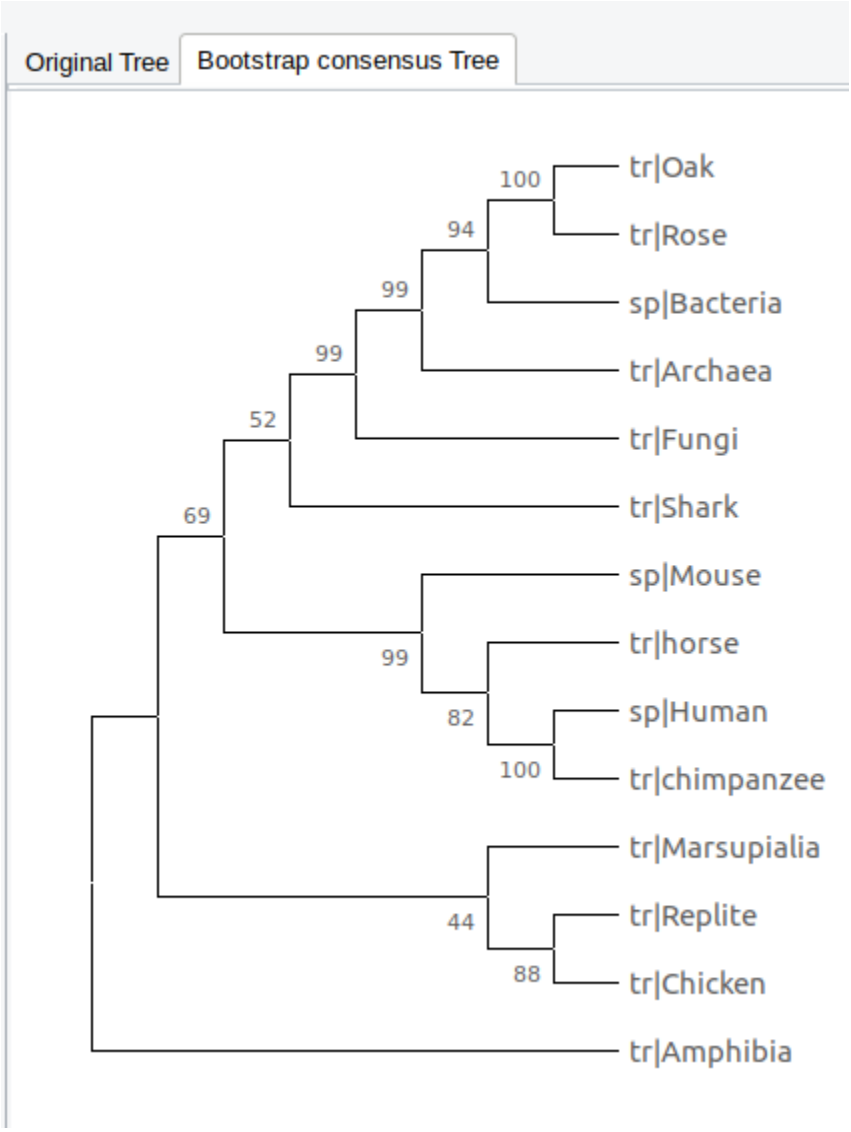
File Display Average Caption Help

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. sp|Human														
2. sp|Mouse	0.1705													
3. tr|chimpanz	0.0142	0.1781												
4. tr|horse	0.1733	0.1792	0.1763											
5. tr|Marsupia	0.3540	0.3093	0.3507	0.3038										
6. tr|Amphibi	0.4055	0.3957	0.4166	0.4080	0.3497									
7. tr|Replite	0.3143	0.3201	0.3213	0.3904	0.2911	0.3197								
8. tr|Chicken	0.3197	0.3163	0.3231	0.3194	0.2609	0.3462	0.2147							
9. tr|Shark	0.4657	0.4400	0.4734	0.4685	0.4319	0.4478	0.3994	0.4306						
10. tr|Oak	1.3420	1.2725	1.3420	1.3725	1.3828	1.4917	1.3792	1.3328	1.3863					
11. tr|Rose	1.3474	1.2751	1.3474	1.3935	1.3755	1.4875	1.3573	1.4133	1.3935	0.2580				
12. tr|Fungi	1.1816	1.1560	1.1816	1.2546	1.2027	1.2849	1.2958	1.2528	1.2625	1.5261	1.6219			
13. tr|Archaea	1.3690	1.3690	1.3690	1.4005	1.4323	1.4323	1.5160	1.3863	1.5247	1.2852	1.3228	1.9136		
14. sp|Bacteri	1.3304	1.3598	1.3450	1.3412	1.3749	1.4214	1.3749	1.4214	1.3160	0.9920	0.9383	1.5796	1.2609	

С помощью метода присоединения соседей было получено дерево 1 (без бутстрапа):

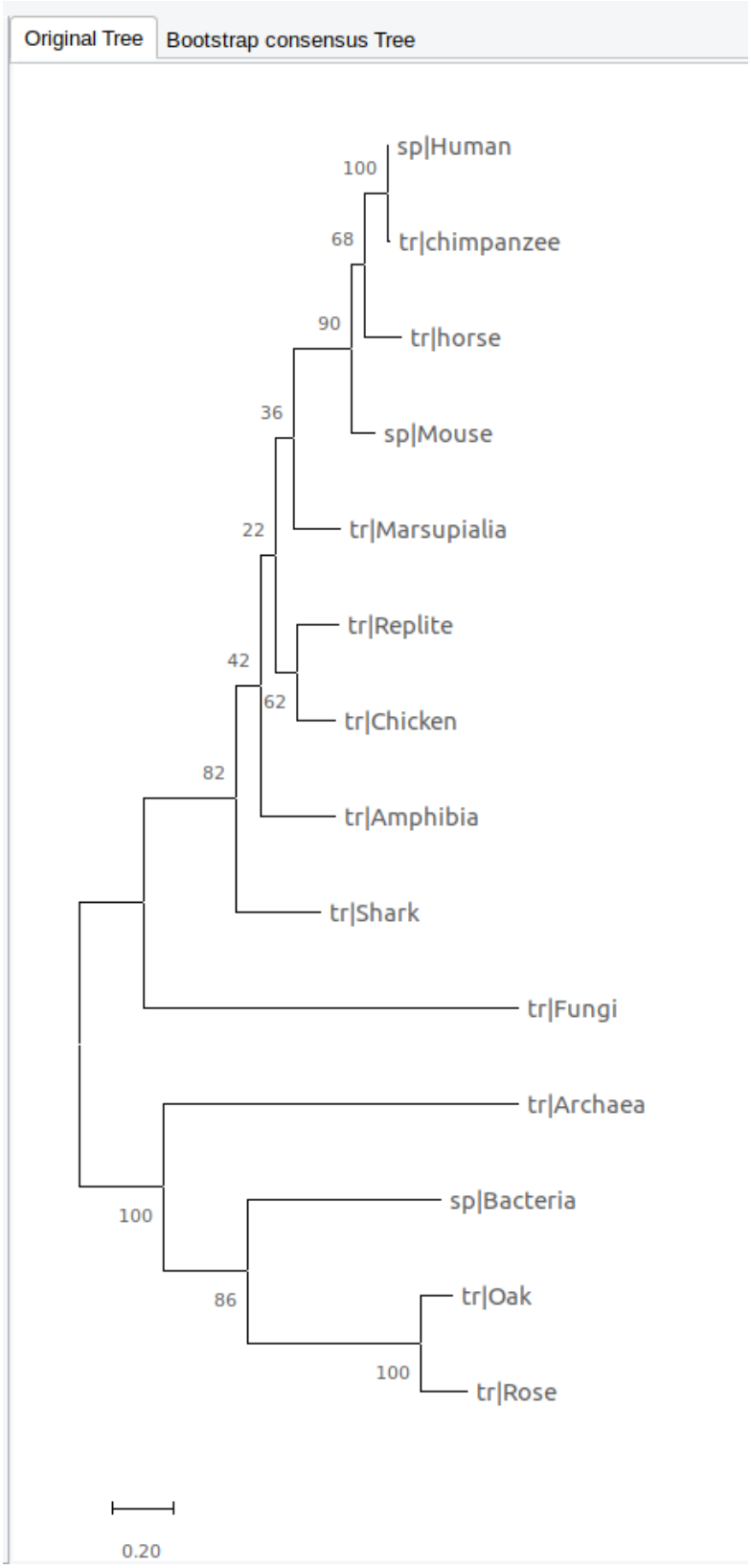


С помощью метода присоединения соседей было получено дерево 2 (с бутстрапом 1000):

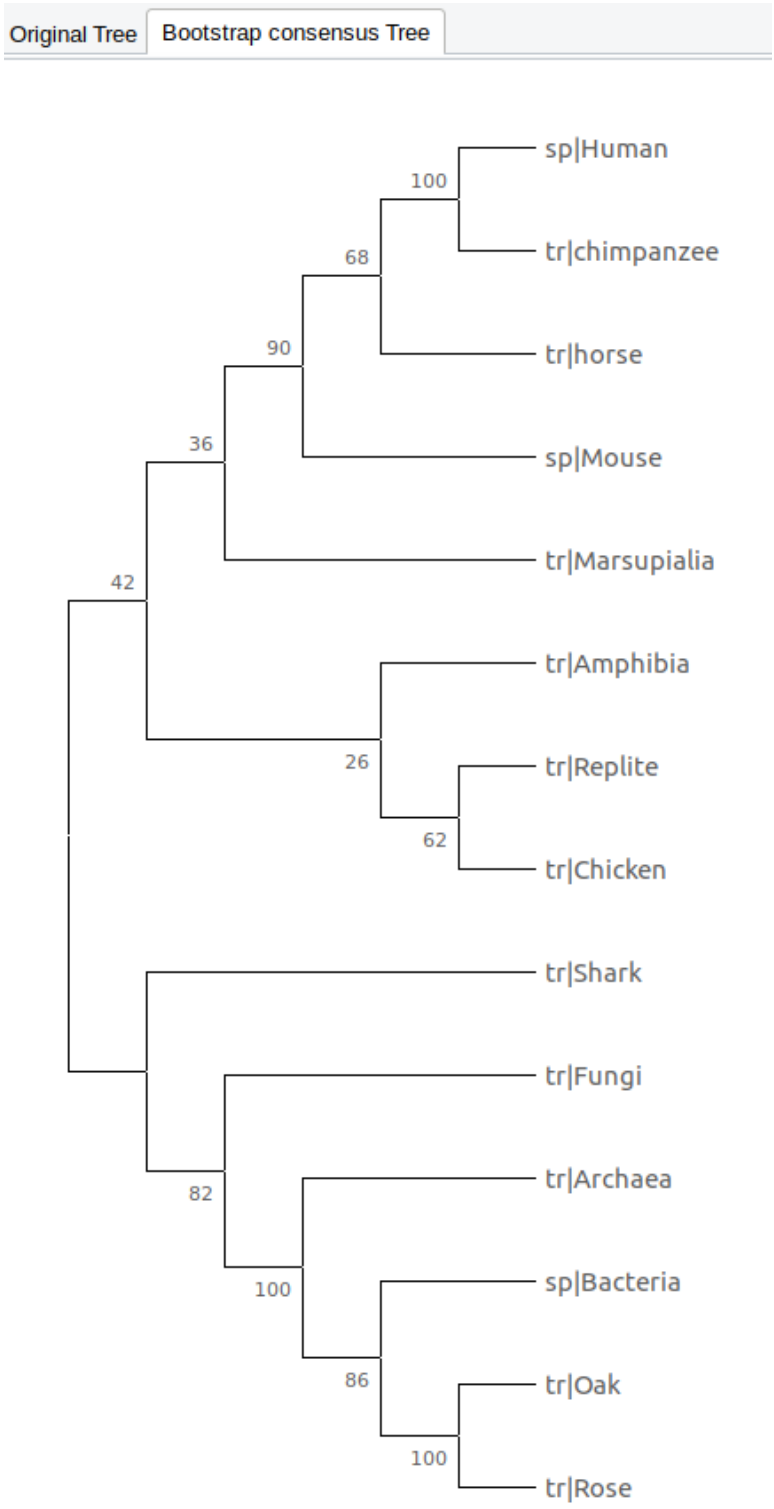


Для метода присоединения соседей использовалась Poisson model эволюции

С помощью ML-метода было получено дерево 3 (без бутстрапа):



С помощью ML-метода было получено дерево 4 (с бутстрапом 50):



Для ML-метода использовалась Jones-Taylor-Thorton модель эволюции.

5.1 Насколько хорошо получившиеся деревья для генов (все 4) соответствуют вашим представлениям о дереве видов? (можно пользоваться https://www.ncbi.nlm.nih.gov/taxonomy)

1) Дерево метода ближайших соседей без бутстрепа: В целом tree соответствует представлениям о дереве видов, за тем исключением, что растения оказались ближе к бактериям. А fungi оказались хоть и близко к животным (как и должно быть), но далеко от растений.
2) Деверо метода ближайших соседей с бутстреп-консенсусом: В целом tree не соответствует представлениям. Например, человек оказывается ближе к бактериям, чем к курице.
3) ML-дерево без бутстрепа. Самое похожее на дерево жизни tree. Но та же проблема с растениями, что в (1)
4) ML-дерево с бутстреп-консенсусом. Акула и фунги сгруппированы вместе с бактериями, что не сильно корректно.

5.2 Если есть различия, то в чем они заключаются и чем можно их объяснить?

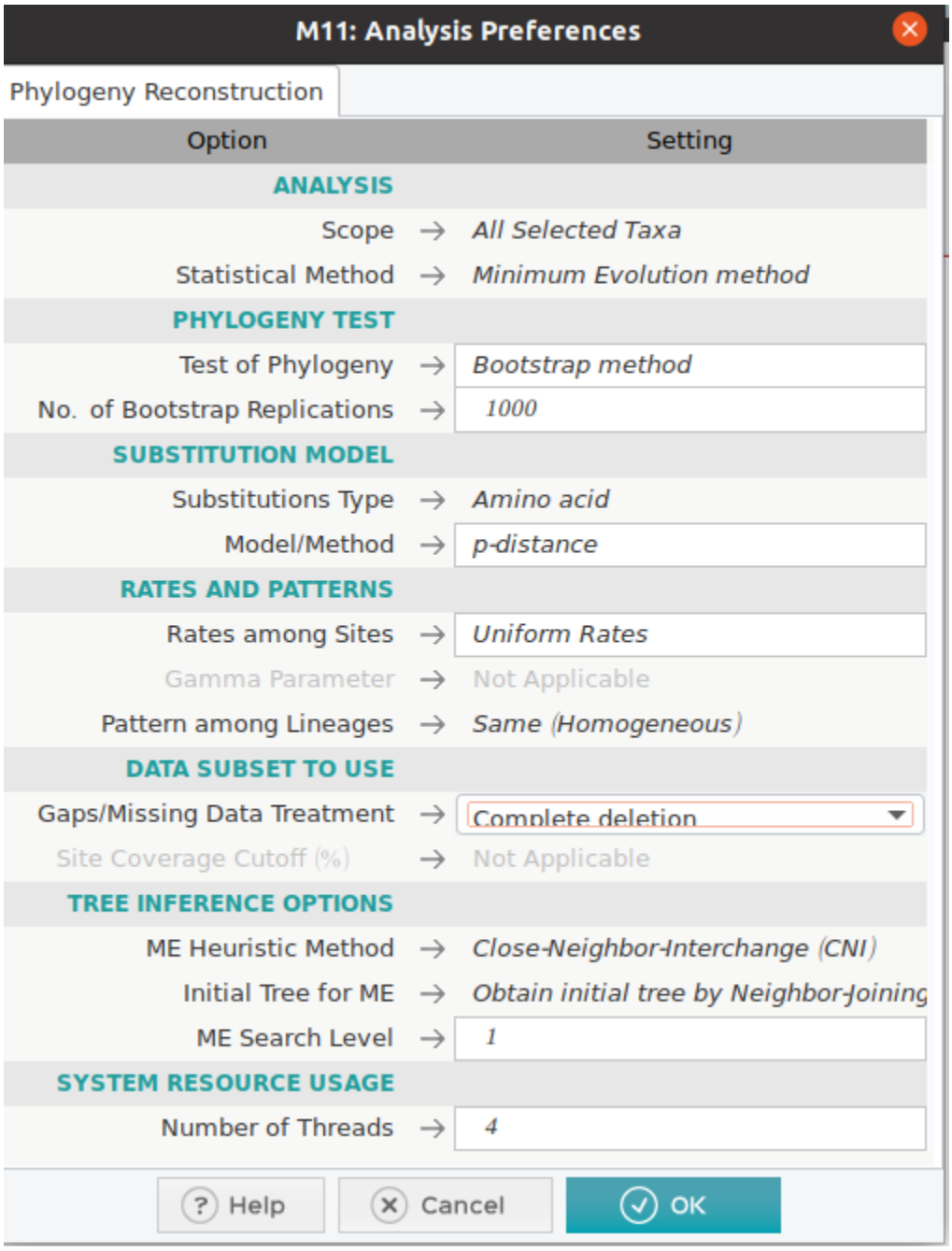
Полагаю, что объяснение вышеуказанных различий в том, что ортологи в растениях сильно отличаются, у них есть дополнительные вставки, и это приводит как к их отдельной классификации, так и к тому, что это сильно влияет на бутстреп-анализ, у которого есть большой шанс выхватить участки выравнивания с большим количеством гэпов.

5.3. В чем смысл бутстреп значений?

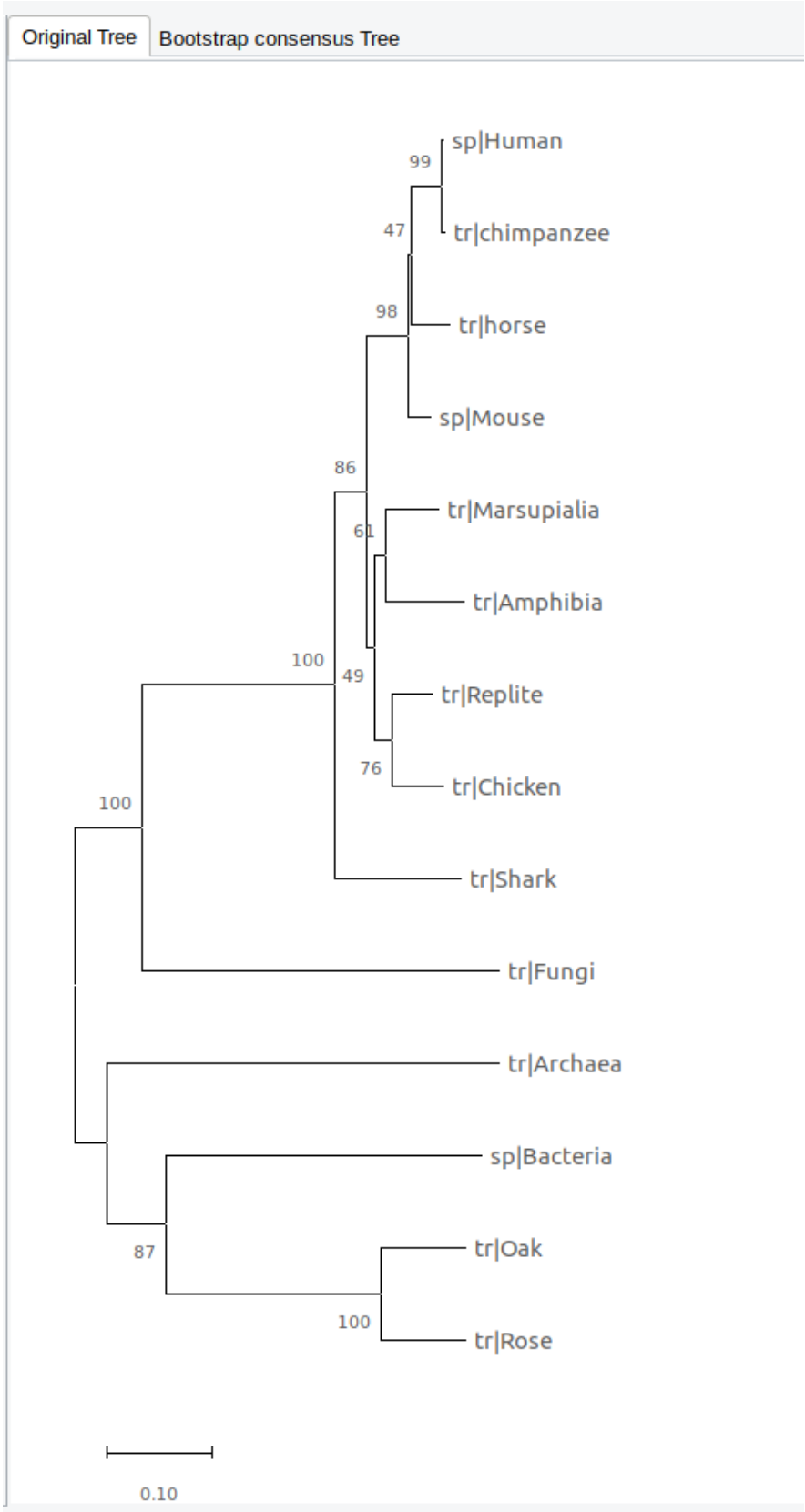
Смысл bootstrap-значения в том, что для статистической оценки мы проводим выборки с возвращением колонок из полученного выравнивания, размер выборки равен числу колонок. Бутстреп-значение - это то, сколько раз мы делаем такую выборку. Если бутстреп 1000, значит мы сделали 1000 таких выборок, и смотрели, как часто была получена именно такая структура в каждом из участков итогового дерева (либо на основании бутстрепа строим консенсусное дерево)

5.4* Попробуйте использовать другие параметры, например, изменить модель замен.

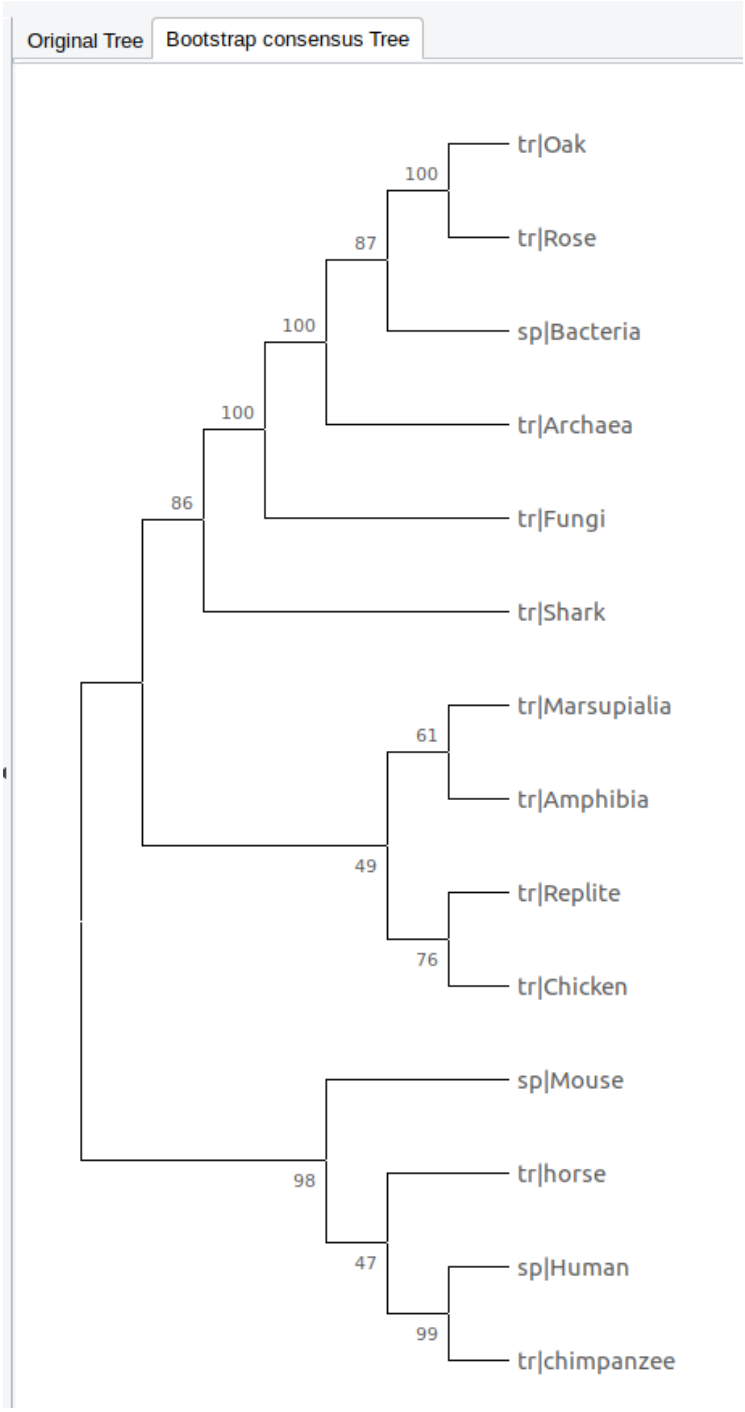
Построено minimum-evolution tree. С параметрами, Bootstrap 1000, модель замен p-distance и полным удалением гэпов.



Было получено дерево, в котором fungi были сгруппированы с животными



И консенсус-дерево



Как видно, удаление гэпов не помогло в бутстрепе, что приводит к предположению, что дело всё-таки в том, что это всего лишь один ген и по нему нельзя хорошо оценивать эволюционное дерево. К тому же этот ген сильно отличается для растений