Филогенетические деревья

Дарья Владимировна Диброва

к.б.н., с.н.с. НИИ ФХБ имени А.Н. Белозерского МГУ имени М.В. Ломоносова,

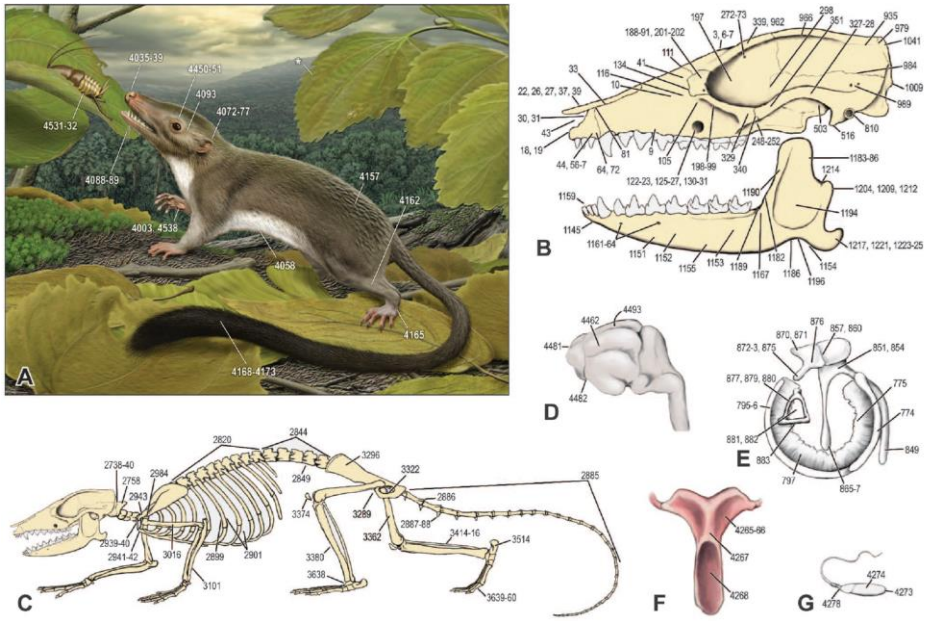
доцент ФББ МГУ имени М.В. Ломоносова

В презентации использованы материалы С.А. Спирина, Ю.А. Алешиной, Е.Д. Рюминой, А.В. Алексеевского

20 февраля 2026 г.

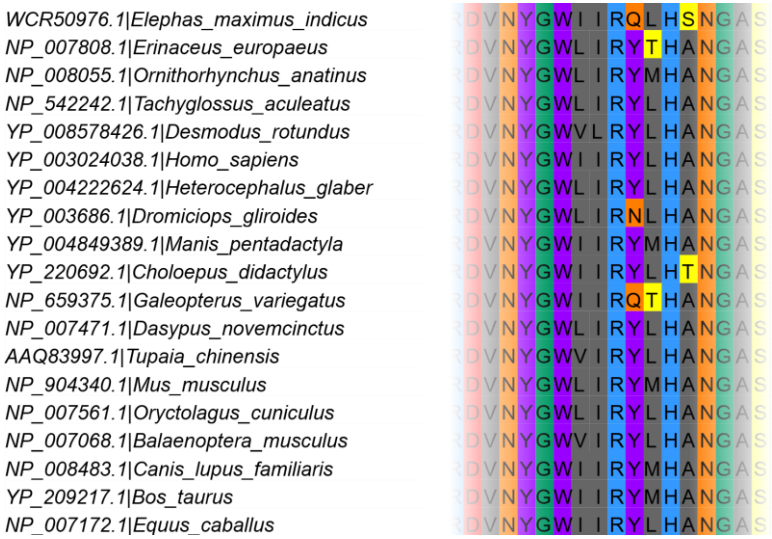
1. Повторение: на основании чего можно строить дерево, топология деревьев

Открытый вопрос: как лучше строить дерево, по признакам – или по последовательностям?



Ваши аргументы – что лучше?

vs.



Источник рисунка вверху слева: O'Leary M.A. *et al.* The Placental Mammal Ancestor and the Post-K-Pg Radiation of Placentals. *Science* 339(6120):662-667, **2013**.

Этапы реконструкции филогенетического дерева по последовательности

Последовательности белков или ДНК/РНК

Muscle, MAFFT, Prank, Clustal Omega...

Множественное выравнивание

Глаза, мозг, м.б. Alnalyser

Корректное множественное выравнивание

Максимальная экономия (Maximum Parsimony), наибольшее правдоподобие (Maximum Likelihood) Bayesian method...

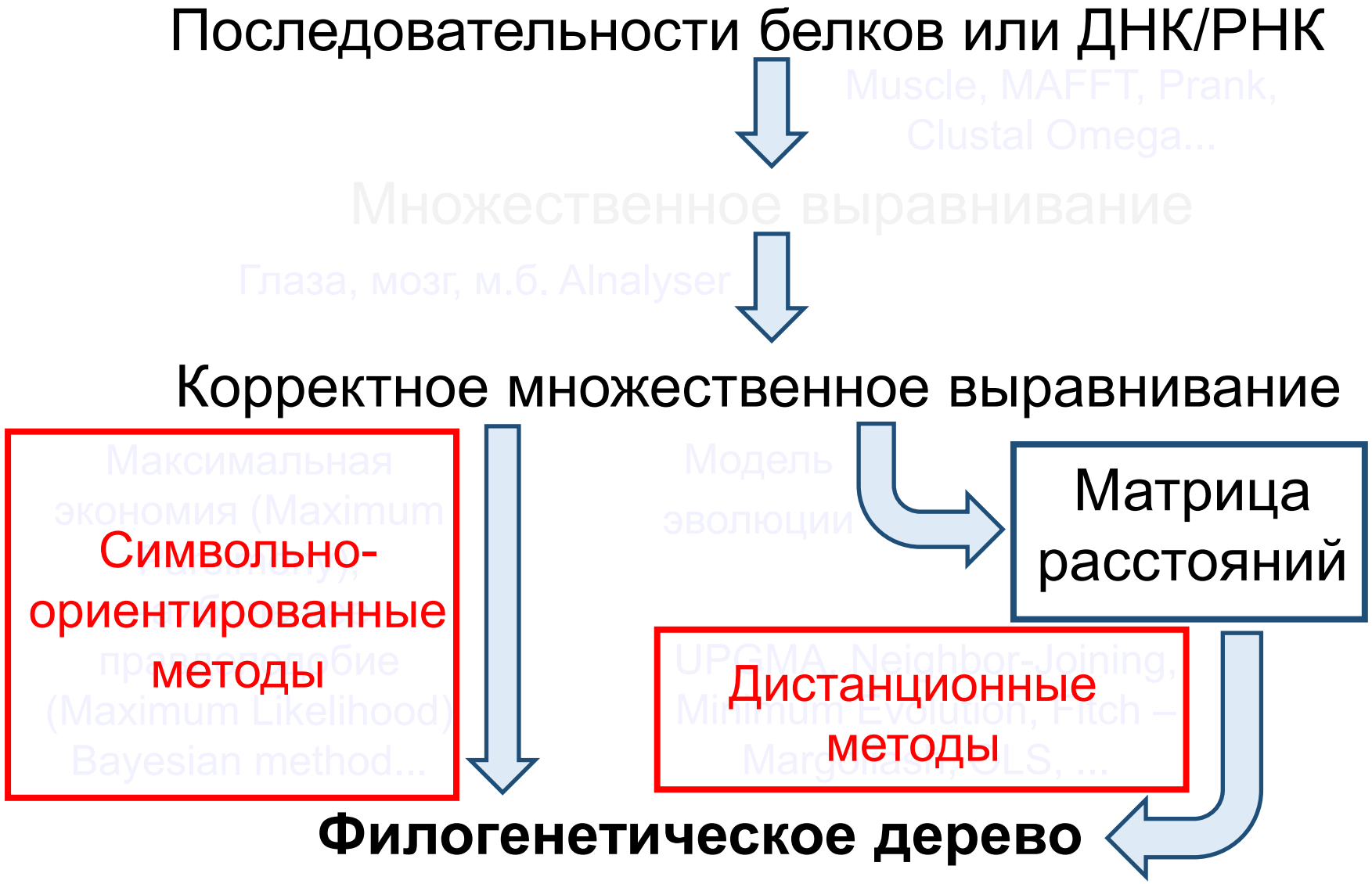
Модель эволюции

Матрица расстояний

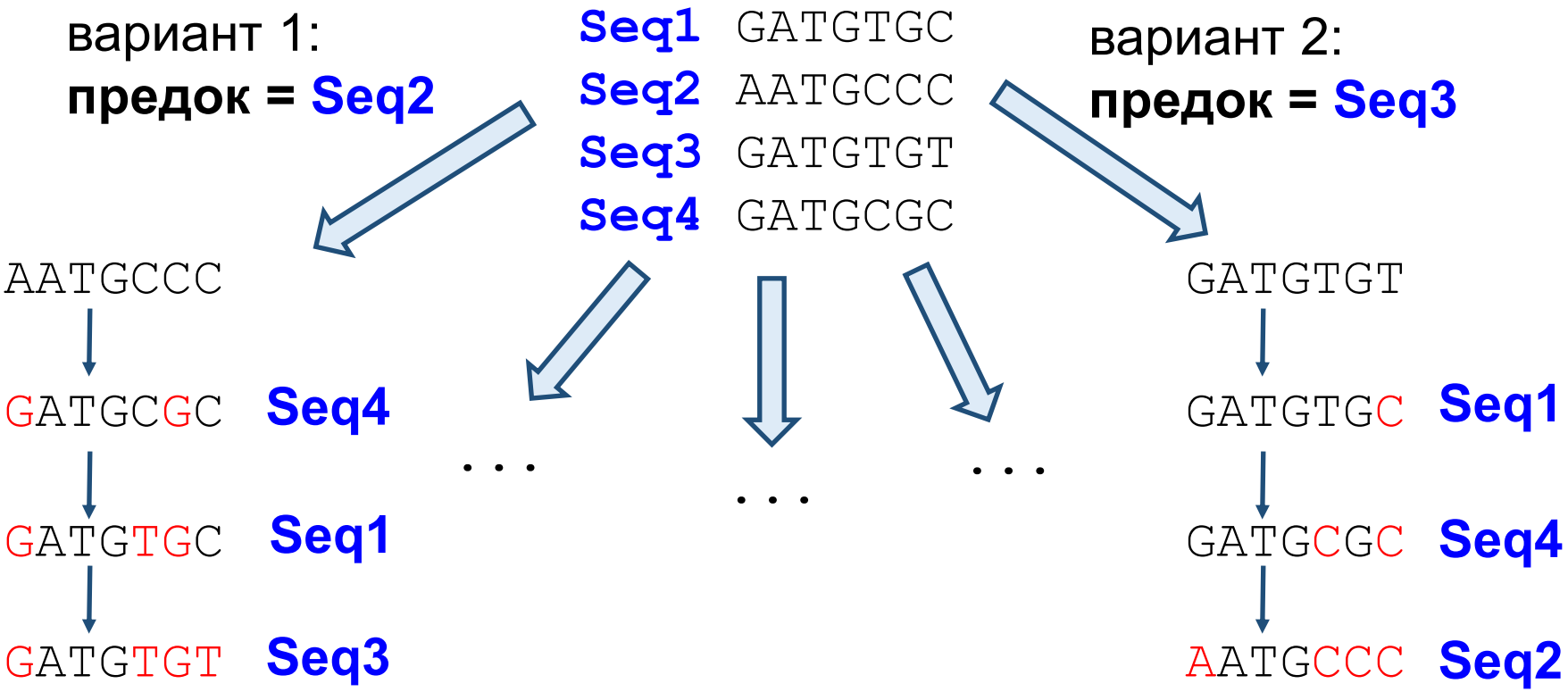
UPGMA, Neighbor-Joining, Minimum Evolution, Fitch – Margoliash, OLS, ...

Филогенетическое дерево

Этапы реконструкции филогенетического дерева по последовательности

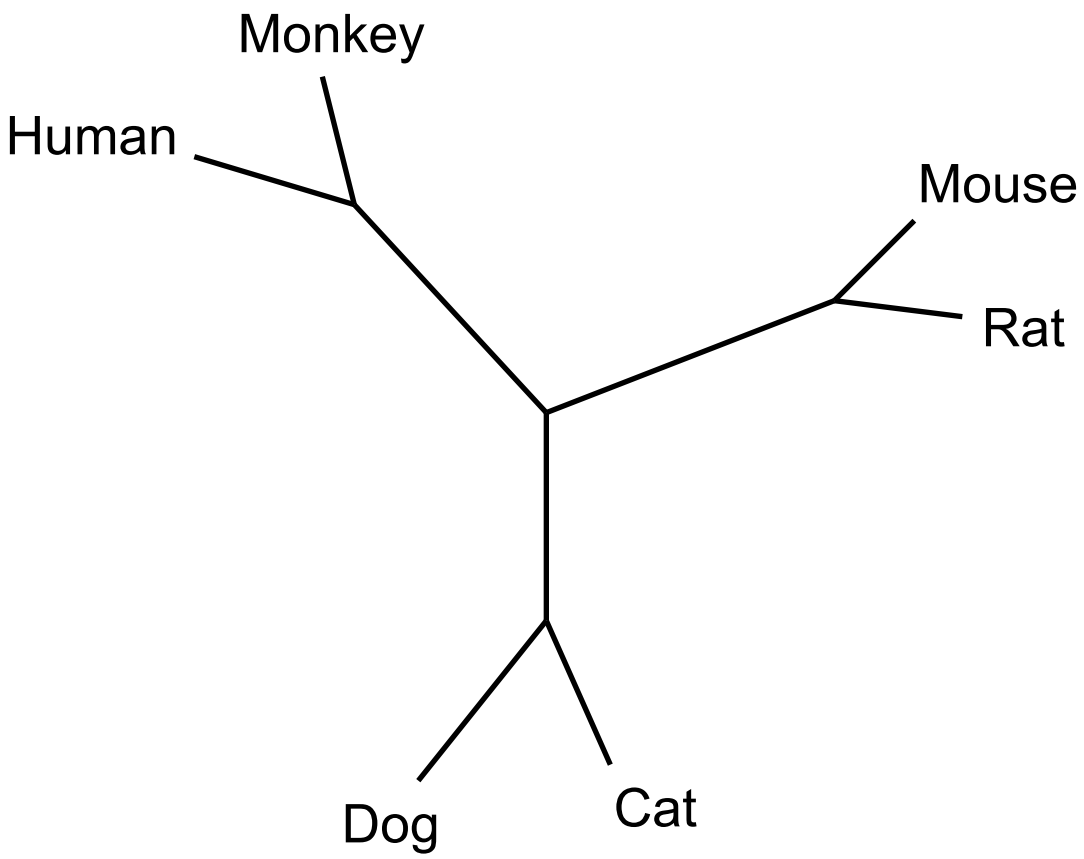


Направление эволюции, предковая и промежуточные последовательности НЕИЗВЕСТНЫ

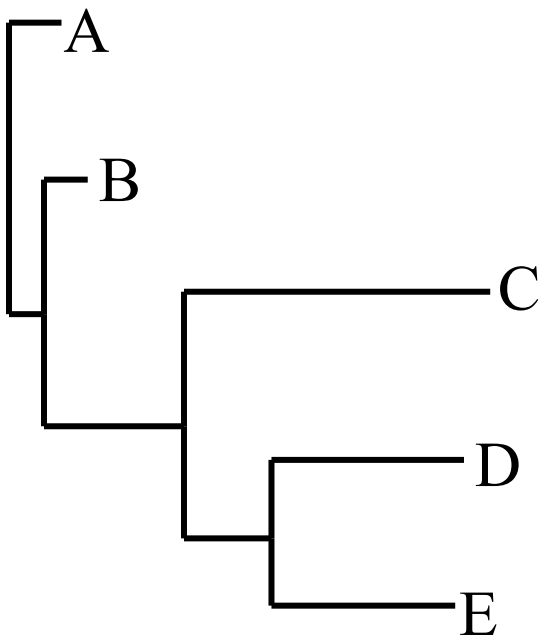


Какие деревья будут соответствовать таким вариантам?

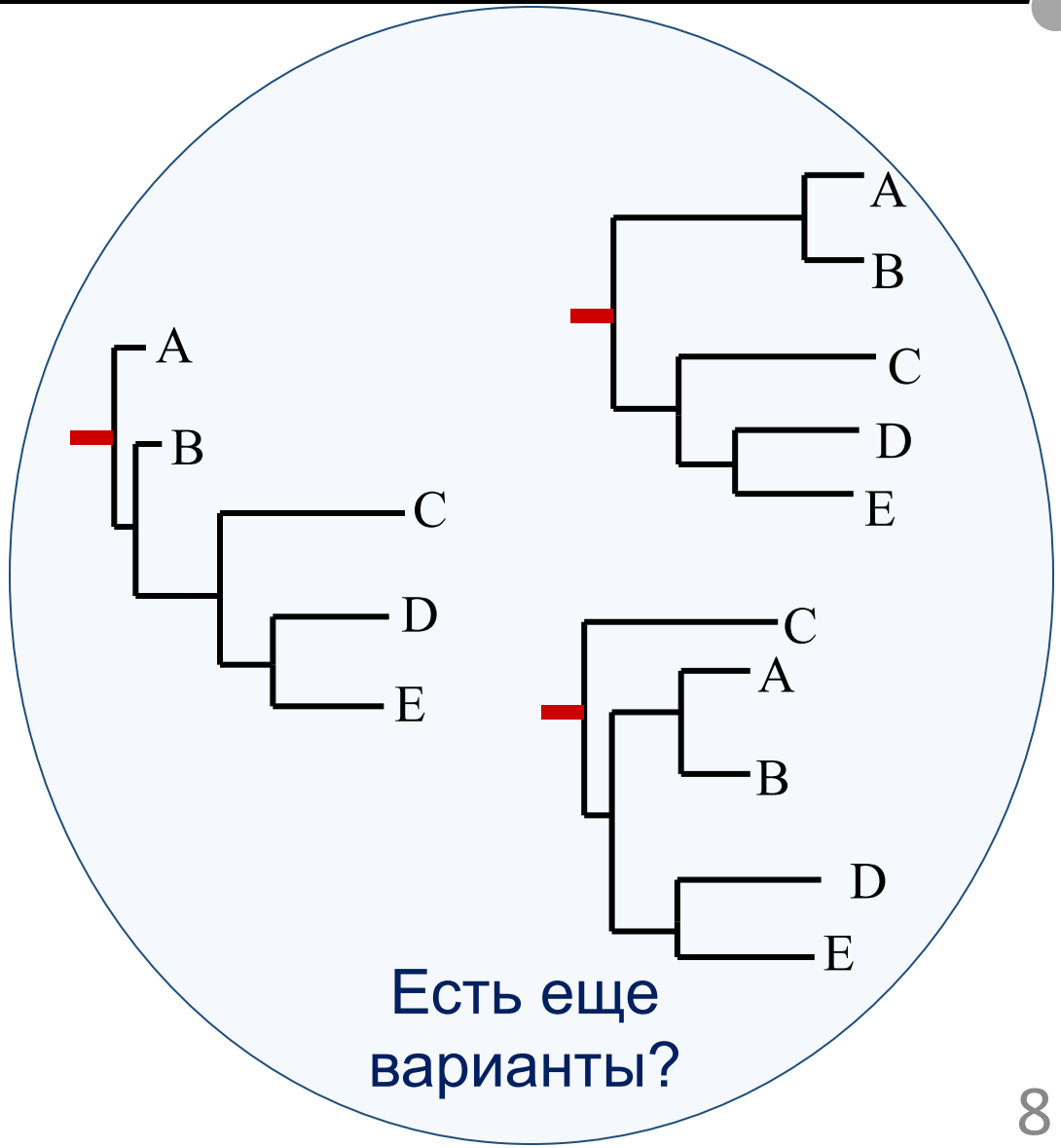
Почти все программы выдают неукоренённое дерево



Неукоренённое дерево следует понимать как множество возможных укоренений

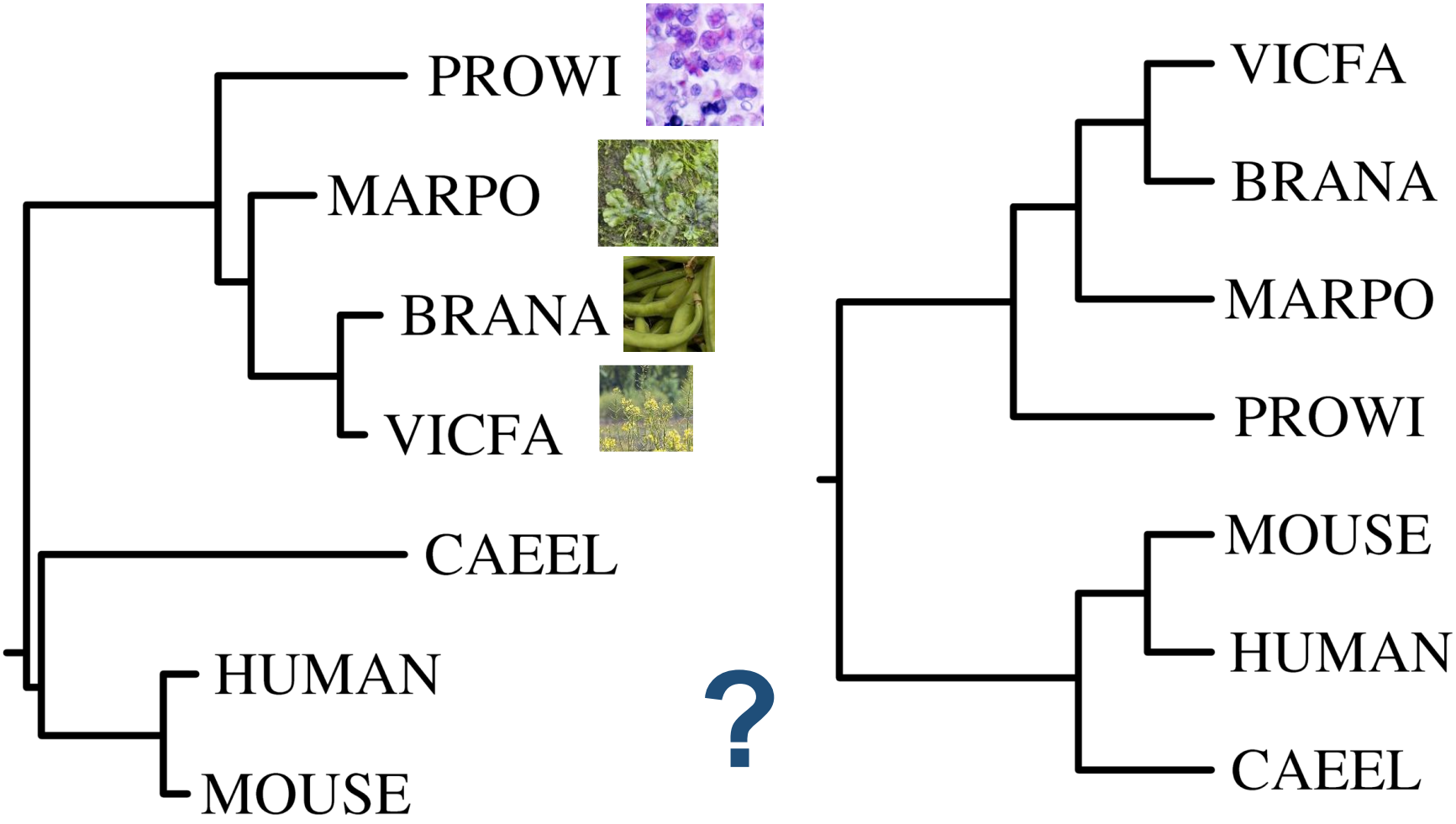


=



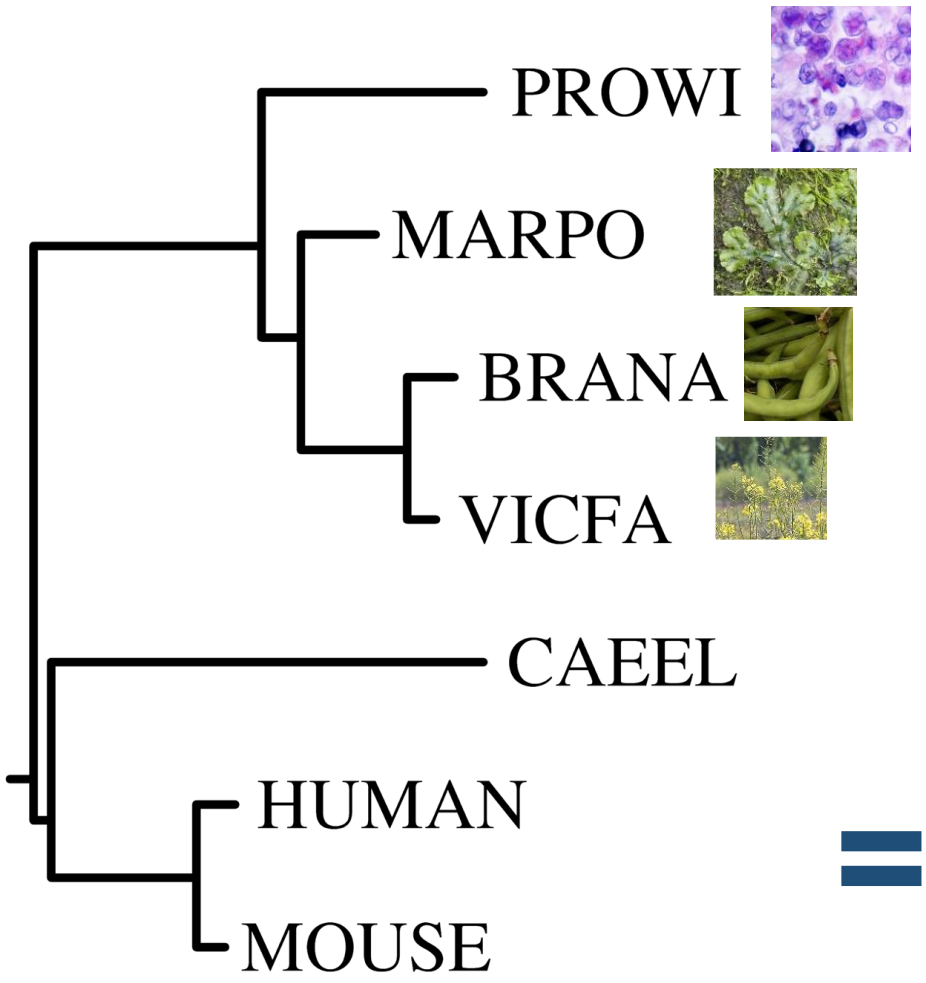
Одинаковая ли у данных деревьев топология?

Источник изображений растений: Википедия

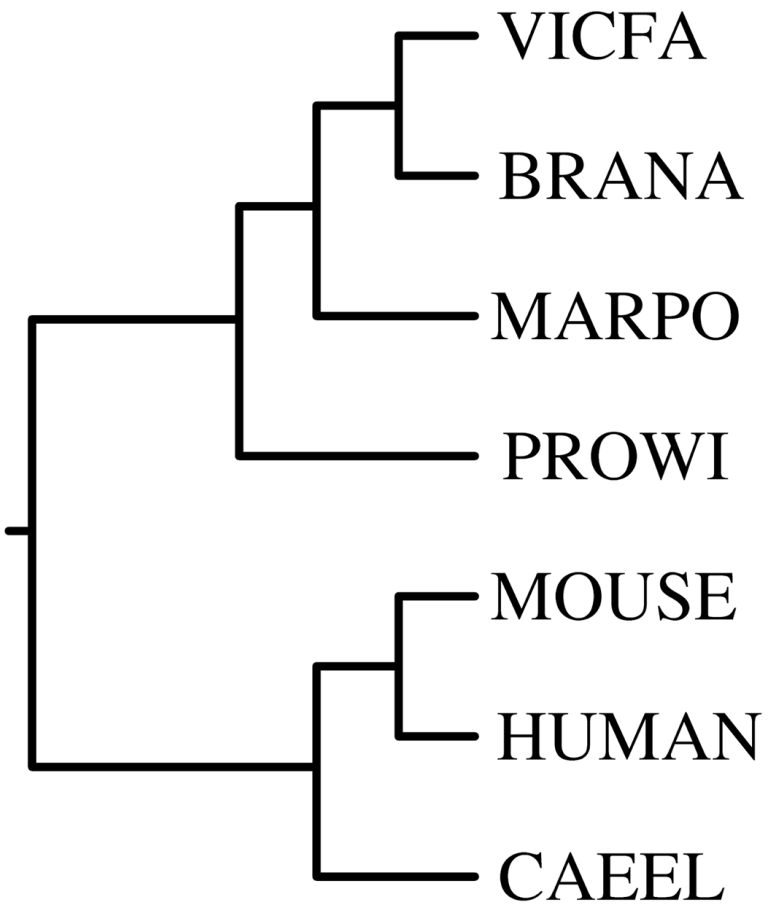


Топология одинаковая, но на правом дереве намного меньше информации

Источник изображений растений: Википедия

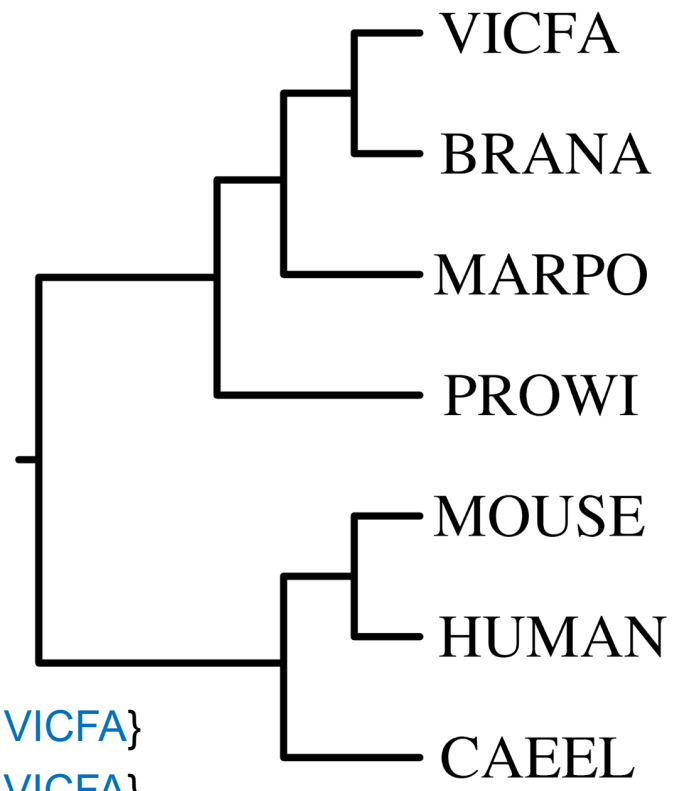


☹ Не показаны длины ветвей!



Топология одинаковая, но на правом дереве намного меньше информации

- Каждая ветвь разбивает множество листьев на два.
- В каждом дереве есть **тривиальные** ветви (отделяющие один лист от всех остальных), они не зависят от топологии.
- Топологию (неукоренённого) дерева можно однозначно записать набором нетривиальных разбиений. Например:



{HUMAN,MOUSE} vs. {CAEEL,PROWI,MARPO,BRANA,VICFA}
{HUMAN,MOUSE,CAEEL} vs. {PROWI,MARPO,BRANA,VICFA}
{HUMAN,MOUSE,CAEEL,PROWI} vs. {MARPO,BRANA,VICFA}
{HUMAN,MOUSE,CAEEL,PROWI,MARPO} vs. {BRANA,VICFA}

2. Методы построения филогенетических деревьев

Классификация основных методов реконструкции деревьев

Название метода	Переборный / прямой	Использует молекулярные часы	Символьный/ дистанционный	Реконструирует длины ветвей
UPGMA	Прямой	Да	Дистанционный	Да
Neighbor-Joining	Прямой	Нет	Дистанционный	Да
OLS **Наименьших квадратов**	Переборный	Может	Дистанционный	Да
Fitch-Margoliash **Фитча – Марголиаша**	Переборный	Может	Дистанционный	Да
Minimum Evolution **Минимальной эволюции**	Переборный	Может	Дистанционный	Да
Maximum Parsimony **Максимальной экономии**	Переборный	Нет	Символьный	Нет ☹
Maximum Likelihood **Наибольшего правдоподобия**	Переборный	Может	Символьный	Да
Bayesian Method **Байесовский**	Переборный	Может	Символьный	Да

Термины в классификации методов

Дистанционные: берут на вход матрицу расстояний (как их считать поговорим дальше).
Символьные: берут на вход выравнивания.

Прямые: строят дерево напрямую, исходя из матрицы расстояний.
Переборные: начинают от дерева, построенного прямым методом, но имеют критерий сравнения двух деревьев («какое лучше») и алгоритм перебора вариантов деревьев.

Вспоминаем, что такое «**молекулярные часы**»:

- С какой скоростью накапливаются мутации в разных ветвях дерева?
- «С одинаковой!» = «молекулярные часы»

Методы, предполагающие молекулярные часы, строят **укоренённые ультраметрические** деревья.
Методы, не предполагающие молекулярные часы, строят **неукоренённые** деревья.

Матрица расстояний

Выравнивание

Нужно вычислять так, чтобы выполнялись аксиомы метрического пространства:

1) d (A, A) = 0
2) d (A, B) > 0, если A ≠ B
3) d (A, B) = d (B, A)
4) d (A,B) ≤ d (A,C) + d (B,C)

Для ультраметрического пространства:

4') d (A,B) ≤ max(d (A,C), d (B,C))

	MUSDO	CHICK	BOVIN	HUMAN
MUSDO	0	9.5	8.9	9.2
CHICK	9.5	0	3.4	2.8
BOVIN	8.9	3.4	0	1.7
HUMAN	9.2	2.8	1.7	0

Как оценить расстояние?

```
Seq1 GATGTGC
Seq2 AATGCCC
Seq3 GATGTGT
Seq4 GATGCGC
```

Как оценить расстояние?

Seq1 G**ATG**TGC
Seq2 A**ATG**CCC
Seq3 G**ATG**TGT
Seq4 G**ATG**CGC

1. Число отличий – чем плохо?

	Seq1	Seq2	Seq3	Seq4
Seq1	0	3	1	1
Seq2	3	0	4	2
Seq3	1	4	0	3
Seq4	1	2	3	0

Как оценить расстояние?

Seq1 G**ATG**TGC
Seq2 A**ATG**CCC
Seq3 G**ATG**TGT
Seq4 G**ATG**CGC

1. Число отличий – чем плохо?
2. Число отличий / длину (**p-distance**) – чем плохо?

	Seq1	Seq2	Seq3	Seq4
Seq1	0	3/7	1/7	1/7
Seq2	3/7	0	4/7	2/7
Seq3	1/7	4/7	0	3/7
Seq4	1/7	2/7	3/7	0

Как оценить расстояние?

```
Seq1  GATGTGC
Seq2  AATGCCC
Seq3  GATGTGT
Seq4  GATGCGC
```

	Seq1	Seq2	Seq3	Seq4
Seq1	0	3/7	1/7	1/7
Seq2	3/7	0	4/7	2/7
Seq3	1/7	4/7	0	3/7
Seq4	1/7	2/7	3/7	0

1. Число отличий – чем плохо?
2. Число отличий / длину (**p-distance**) – чем плохо?

При подсчете расстояния так не учитываются:

- **множественные замены** в одном положении (в том числе обратные);
- **неодинаковая частота разных замен** (например, известно, что транзиции, т.е. замены пурина на пурин и пиримидина на пиримидин, в отличие от трансверсий, происходят чаще; замены аминокислот, как видно из матрицы BLOSUM, тоже сильно разновероятны).

Некоторые модели для нуклеотидов

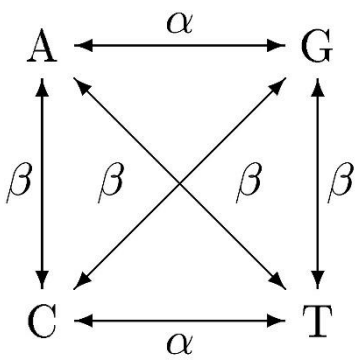
Jukes and Cantor, 1969

$$\left(\pi_A = \pi_G = \pi_C = \pi_T = \frac{1}{4}\right)$$

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

α = β

«rate matrix»



Kimura 2-parameter model (K80), 1980

$\varkappa = f(\alpha, \beta)$

$$Q = \begin{array}{c} \begin{matrix} A & G & C & T \end{matrix} \\ \begin{pmatrix} * & \kappa & 1 & 1 \\ \kappa & * & 1 & 1 \\ 1 & 1 & * & \kappa \\ 1 & 1 & \kappa & * \end{pmatrix} \end{array}$$

α != β

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

Вычисление расстояния между последовательностями

- Почему последовательности различаются? Потому что между ними есть эволюционное расстояние (произошедшие мутации).
- Общая идея **наибольшего правдоподобия**: оцениваем причины по последствиям. Принимаем как наиболее обоснованную ту причину, при которой вероятность наблюдаемых последствий наибольшая.

! Внимание, сейчас речь идет **не о методе построения дерева** методом наибольшего правдоподобия, это другое!

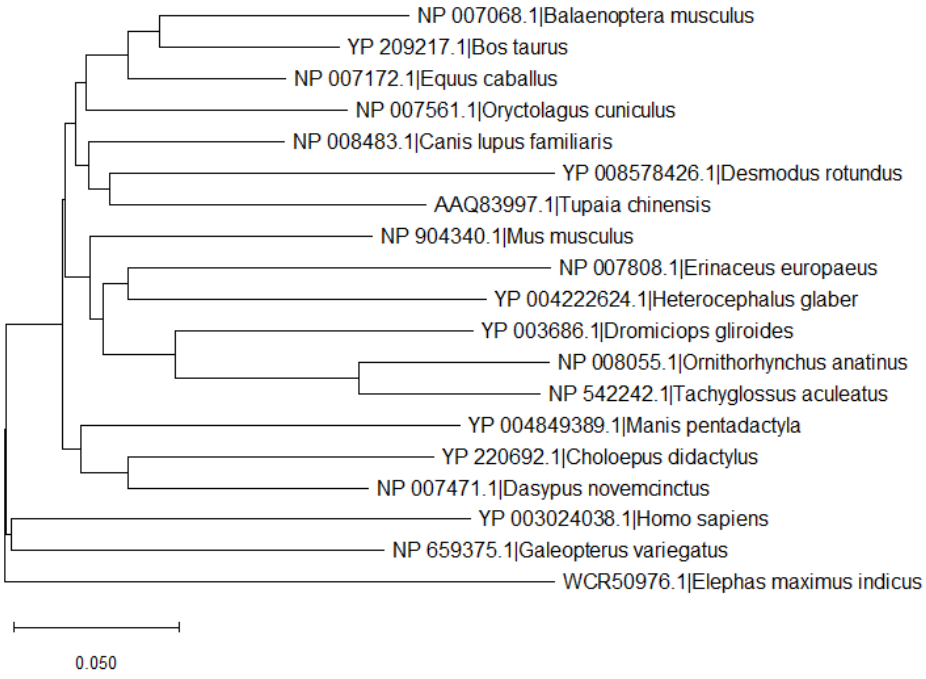
- В нашем случае «причина» – это эволюционное расстояние, а «последствия» – наблюдаемые замены букв. Эволюционная модель (вероятности замен для всех пар букв) предполагается фиксированной.

Моделей много. Для белков наиболее популярны модели **JTT** (по первым буквам фамилий её авторов: Jones, Taylor, Thornton, 1992) и **LG** (Le, Gascuel, 2008)

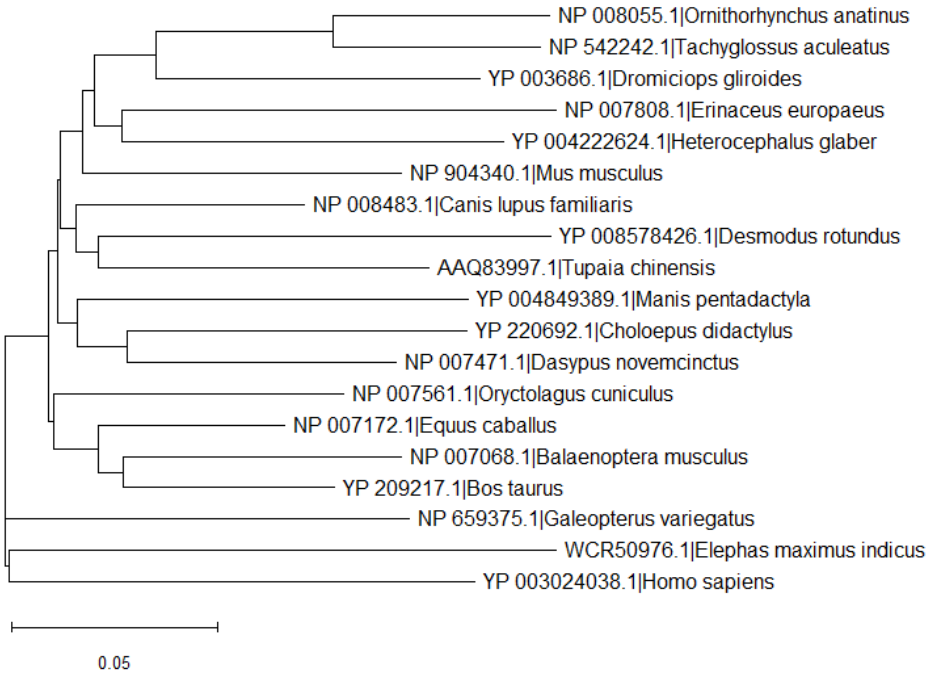
- Для каждого расстояния (= общего числа мутаций) считаем вероятность получить из первой последовательности вторую по модели. За оценку расстояния принимаем то, при котором эта вероятность максимальна.

Способ оценки расстояний влияет на дерево, которое получается

Модель JTT



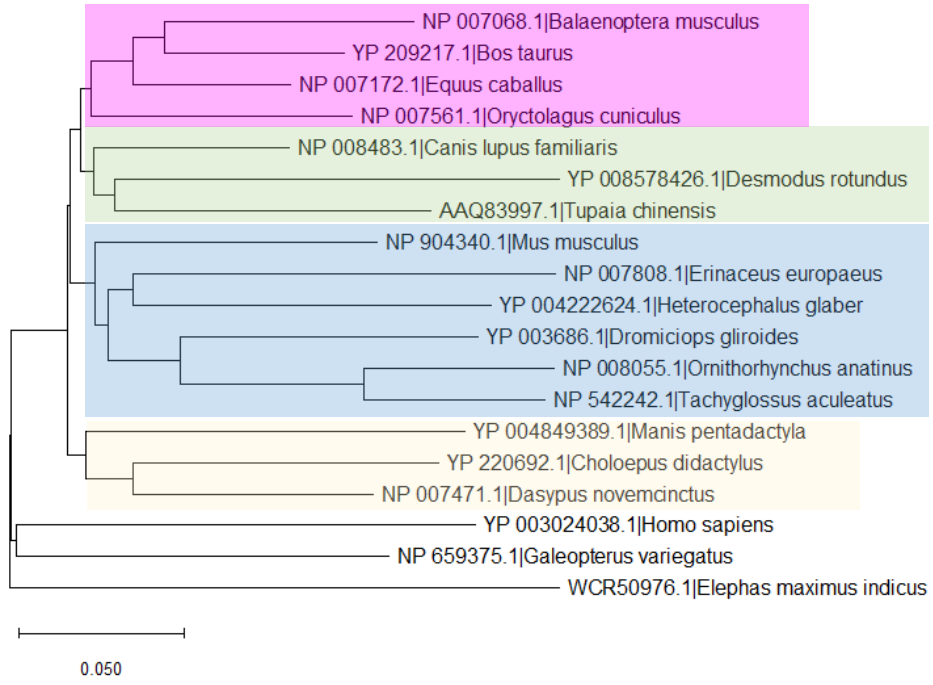
p-distance



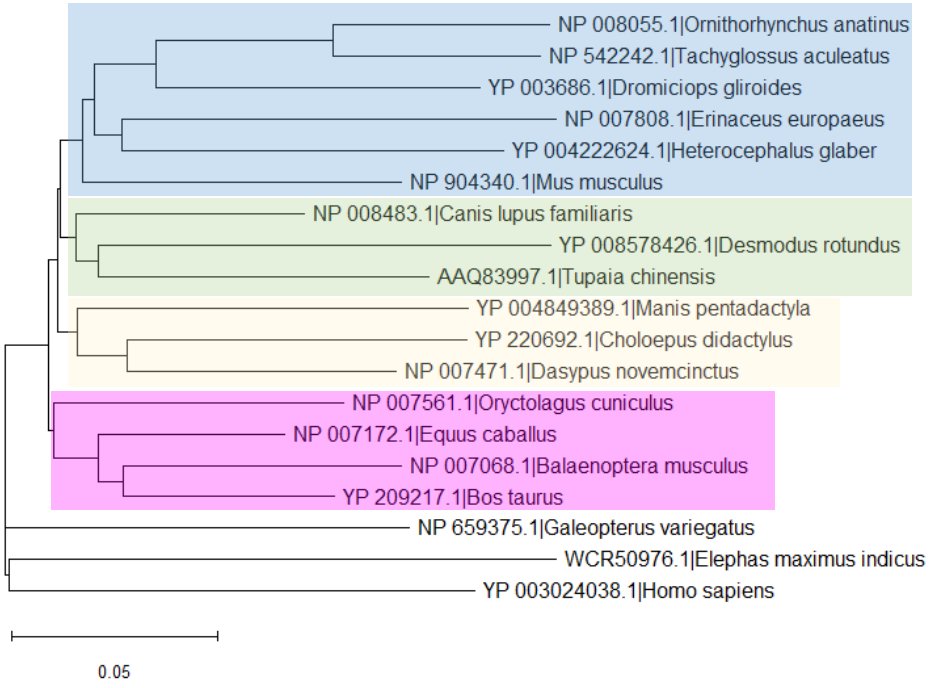
Деревья построены в MEGA12 методом NJ

Способ оценки расстояний влияет на дерево, которое получается

Модель JTT



p-distance



Мораль: если ваше дерево при смене эволюционной модели и/или метода построения радикально меняется, вероятно с данными (т.е. выравниванием) что-то не так

Метод построения дерева UPGMA
(Unweighted Pair Group Method with Arithmetic Mean)

- Найдём в матрице расстояний наименьший элемент, т.е. пару листьев.
- Объединим эти два ближайших листа в кластер (это – узел дерева, соединённый ветвями с листьями, образовавшими его).
- Пересчитаем матрицу расстояний, рассматривая кластер как новый лист. Расстоянием до кластера будем считать как среднее арифметическое расстояний до его элементов (отсюда название метода).
- Повторяем с начала, пока не останется всего два кластера.

- К этому прибавляется способ вычисления длин ветвей. Результат — укоренённое ультраметрическое дерево с длинами ветвей.
- В программе Jalview этот метод реализован под названием «Average distance»

Метод построения дерева UPGMA

(Unweighted Pair Group Method with Arithmetic Mean)

	Dog	Bear	Енот Raccoon	Хорек Weasel
Dog	0	32	48	52
Bear	32	0	26	34
Raccoon	48	26	0	42
Weasel	52	34	42	0

	Dog	B/R	Weasel
Dog	0	40	52
B/R	40	0	38
Weasel	52	38	0

d (B/R, Dog) = ½ * (*d* (Bear, Dog) + *d* (Raccoon, Dog)) = ½ * (32 + 48) = 40

d (B/R, Weasel) = ½ * (*d* (Bear, Weasel) + *d* (Raccoon, Weasel)) = ½ * (34 + 42) = 38

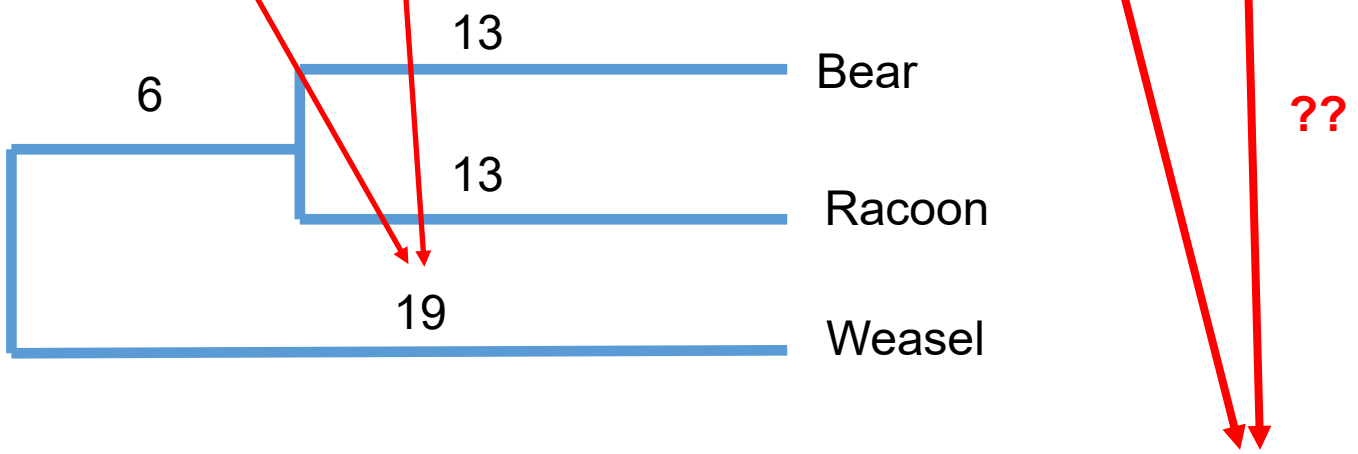
Источник: https://pages.stat.wisc.edu/~larget/Genetics629/lec5-2.pdf

Метод построения дерева UPGMA

(Unweighted Pair Group Method with Arithmetic Mean)

	Dog	B/R	Weasel
Dog	0	40	52
B/R	40	0	38
Weasel	52	38	0

	Dog	B/R/W
Dog	0	44
B/R/W	44	0



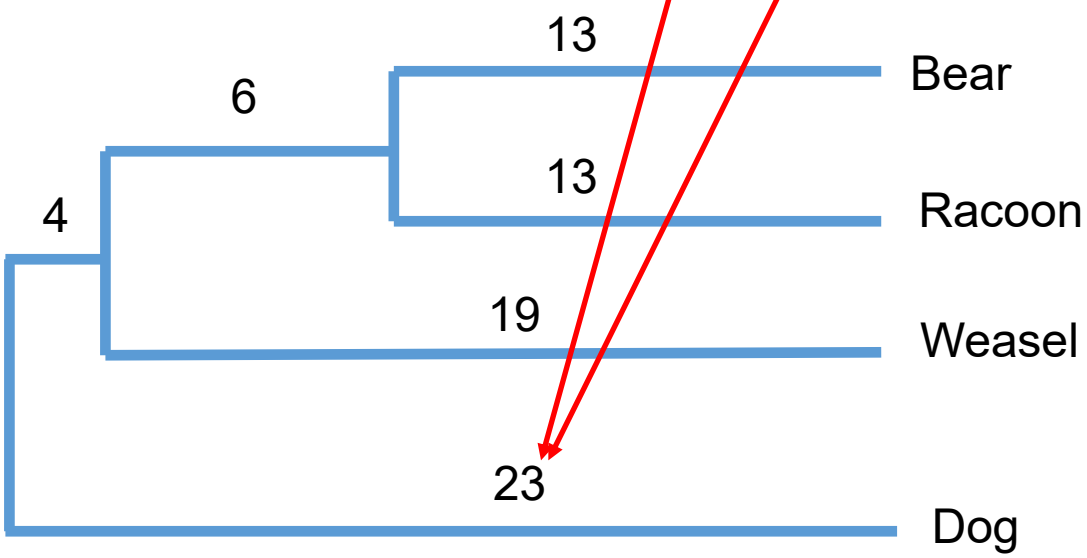
d (B/R/W, Dog) = ½ * (*d* (B/R, Dog) + *d* (Weasel, Dog)) = ½ * (40 + 52) = 46

Источник: https://pages.stat.wisc.edu/~larget/Genetics629/lec5-2.pdf

Метод построения дерева UPGMA

(Unweighted Pair Group Method with Arithmetic Mean)

	Dog	B/R/W
Dog	0	46
B/R/W	46	0



Источник: https://pages.stat.wisc.edu/~larget/Genetics629/lec5-2.pdf

Метод построения дерева NJ
(Neighbor-Joining)

- Выбираем пару последовательностей (A, B) , для которых наименьшее значение имеет величина:

  d(A, B) – s(A) – s(B),
  где d(A, B) — расстояние из входной матрицы, s(A) — сумма расстояний от A до всех остальных последовательностей (аналогично для s(B)).

- Объединяем пару в кластер, с которым далее обращаемся как с одной последовательностью.

- Пересчитываем расстояния (см. пример в доп. слайдах)

См. также: https://pages.stat.wisc.edu/~larget/Genetics629/lec5-2.pdf

Изменение формата выравнивания через BioPython

- Если установлен BioPython:

```
from Bio import AlignIO
in_file = open("input_file.fasta", "r")
out_file = open("output_file.aln", "w")
alignment = AlignIO.parse(in_file, "fasta")
AlignIO.write(alignment, out_file, "phylip-relaxed")
in_file.close()
out_file.close()
```

- Список форматов см. https://biopython.org/wiki/AlignIO
- Он включает форматы stockholm и phylip-relaxed, которых нет в EMBOSS

Доп. слайды

Метод построения дерева NJ
(Neighbor-Joining)

Neighbor-joining Algorithm

1. For each leaf, compute $u_i = \sum_{j \neq i} D_{ij}/(n-2)$.
2. Choose the i and j for which $D_{ij} - u_i - u_j$ is smallest.
3. Join i and j to a new node with lengths $(D_{ij} + u_i - u_j)/2$ to node i and $(D_{ij} + u_j - u_i)/2$ to node j.
4. Compute the distance to the new node (ij) and the other groups as

$$D_{(ij),k} = \frac{D_{ik} + D_{jk} - D_{ij}}{2}$$

5. Delete columns and rows corresponding to i and j and add one for (ij). If there are three or more groups left, go back to the first step. Otherwise, connect the two remaining nodes with their distance.

Источник: https://pages.stat.wisc.edu/~larget/Genetics629/lec5-2.pdf

Метод построения дерева NJ
(Neighbor-Joining)

Example

	D	B	R	W	u_i
Dog	0	32	48	52	66
Bear	32	0	26	34	46
Raccoon	48	26	0	42	58
Weasel	52	34	42	0	64
u_j	66	46	58	64	

	D	B	R	W
Dog	—	−80	−76	−78
Bear	−80	—	−78	−76
Raccoon	−76	−78	—	−80
Weasel	−78	−76	−80	—

- Can choose to join either D/B or R/W because of tie.
- New edge to dog has length $(32 + 66 - 46)/2 = 26$.
- New edge to bear has length $(32 + 46 - 66)/2 = 6$.
- Note these edges sum to 32, but are not equal.

Источник: https://pages.stat.wisc.edu/~larget/Genetics629/lec5-2.pdf

Метод построения дерева NJ
(Neighbor-Joining)

Example

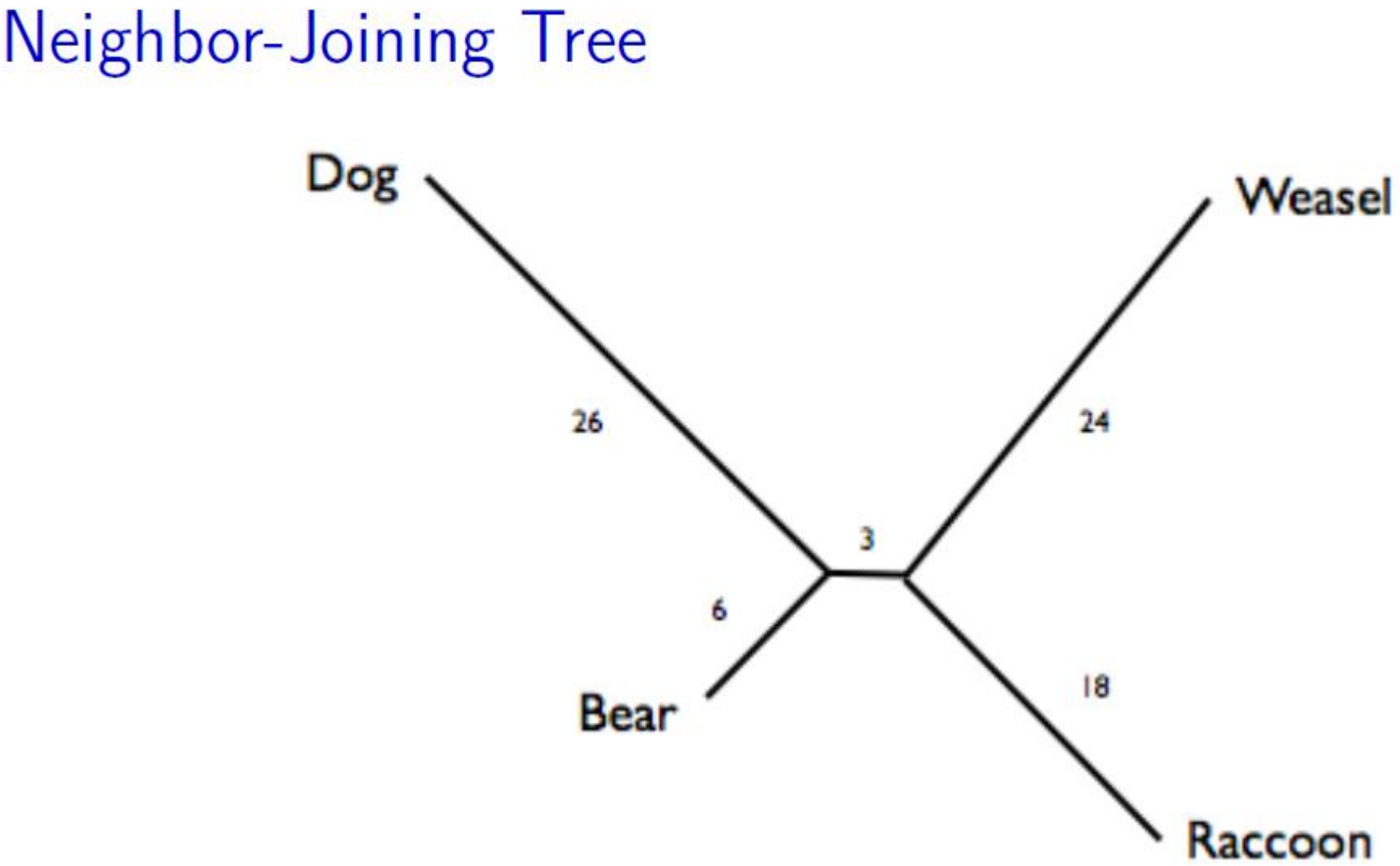
	D/B	R	W	u_i
D/B	0	21	27	48
Raccoon	21	0	42	63
Weasel	27	42	0	69
u_j	48	63	69	

	D/B	R	W
D/B	—	−90	−90
Raccoon	−90	—	−90
Weasel	−90	−90	—

- For the last three, you can always join any pair.
- Simply use the equation from step 4 of the algorthm for the distances.
- Note that in computing u_i, we now use $n = 3$ as there are n groups now.

Источник: https://pages.stat.wisc.edu/~larget/Genetics629/lec5-2.pdf

Метод построения дерева NJ
(Neighbor-Joining)



Источник: https://pages.stat.wisc.edu/~larget/Genetics629/lec5-2.pdf