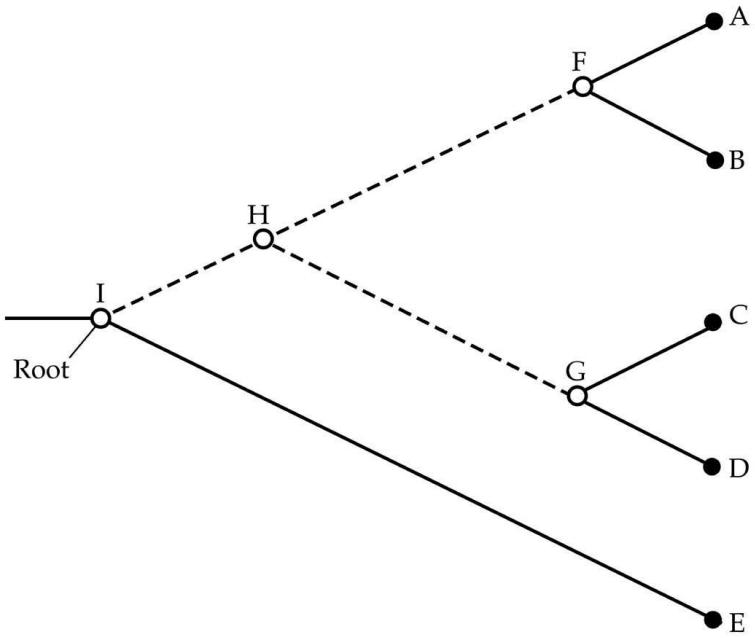
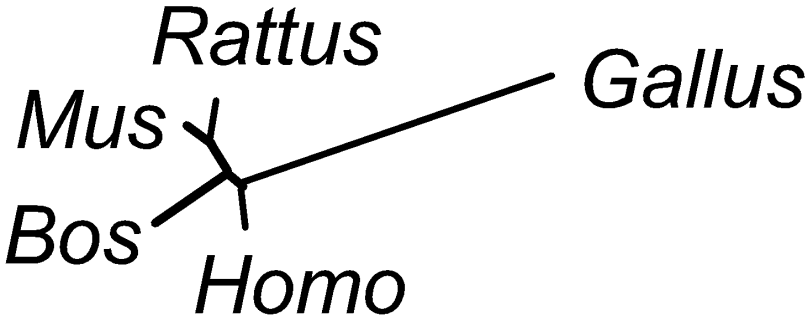
Построение филогенетических деревьев

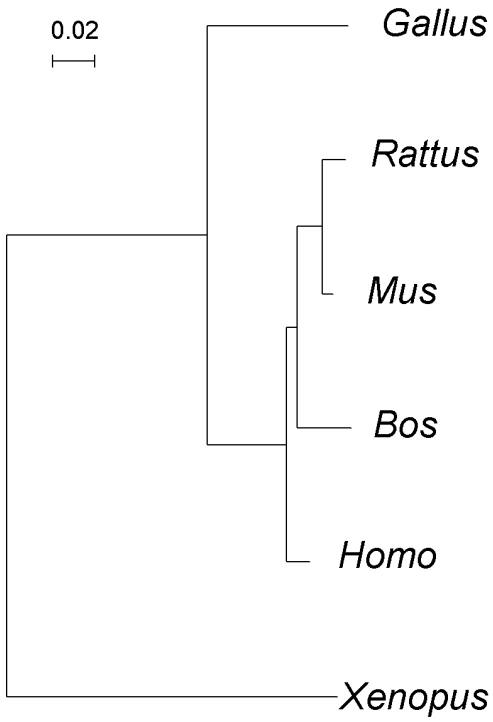
Особенности молекулярной эволюции

1. Скорость эволюции любого белка, выраженная через число аминокислотных замен на сайт в год, приблизительно постоянна и одинакова в разных филогенетических линиях, если только функция и третичная структура этого белка остаются в основном неизменными.

Что такое филогенетические деревья?

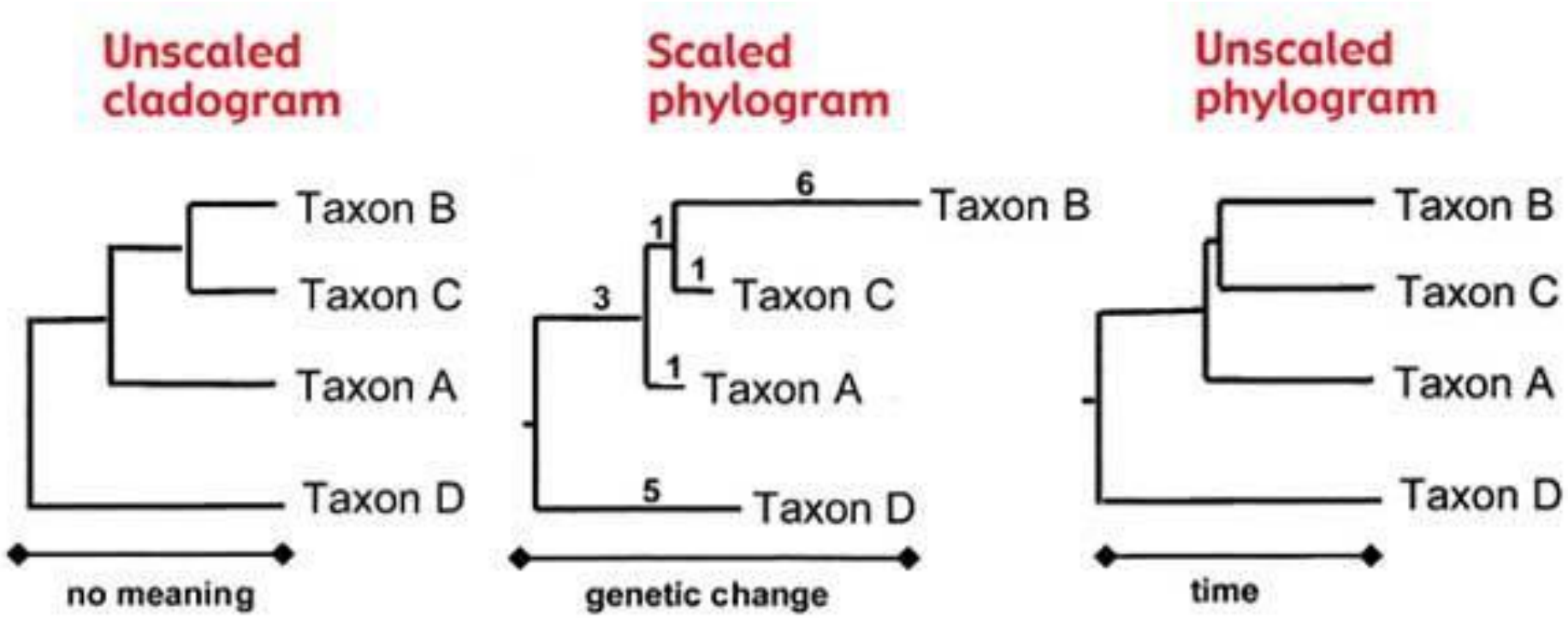






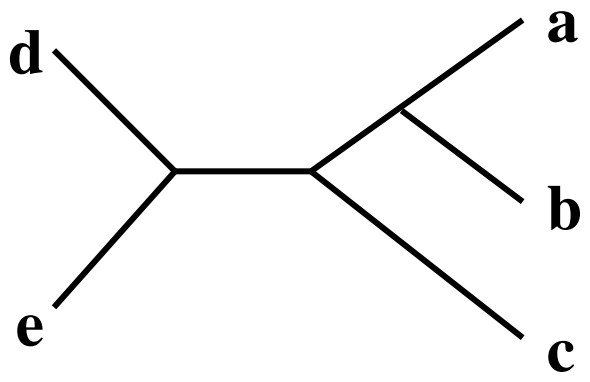
Дерево — это граф, в котором два соседних узла соединены только одним ребром.

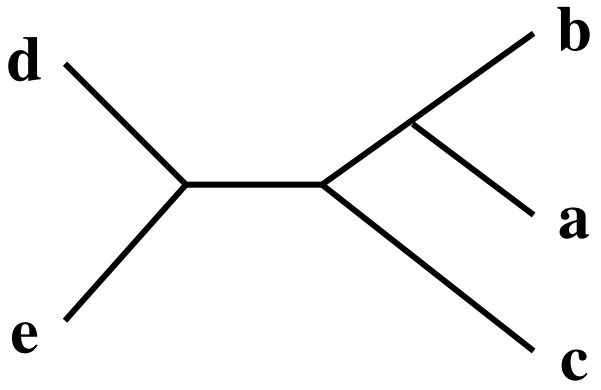
Кладограммы и филограммы

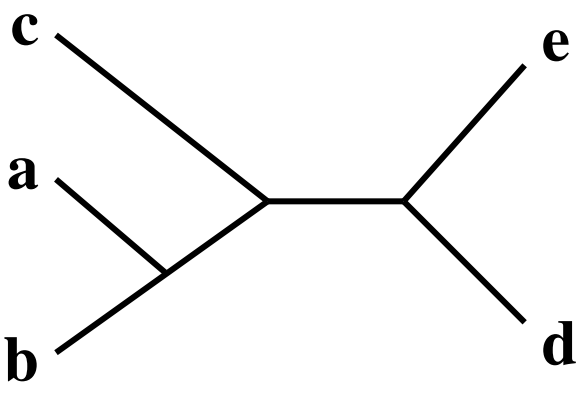


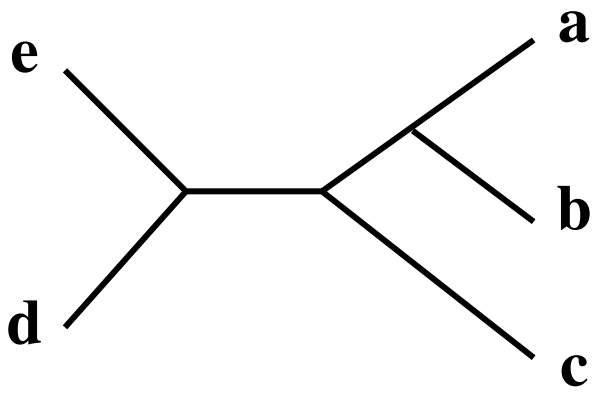
Кладограммы отражают только порядок ветвления, филограммы — ещё и длину ветвей

Сколько здесь разных кладограмм?



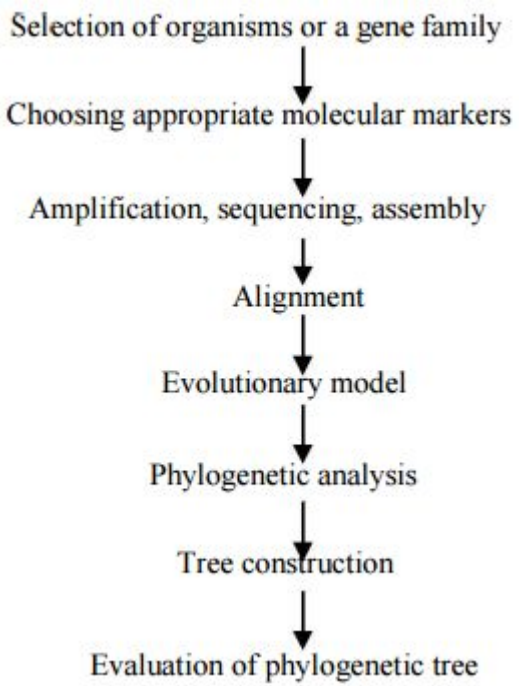




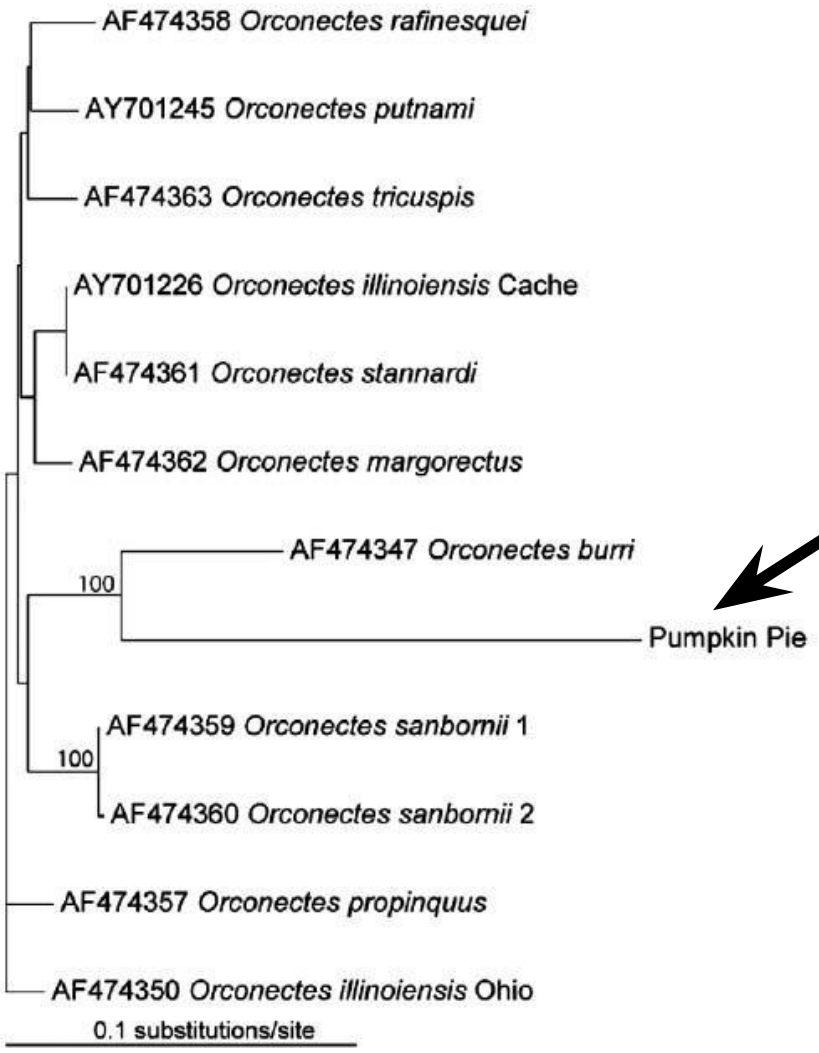


Выбор последовательностей

- Последовательности должны быть гомологичны! Программа выровняет любые последовательности => нужно проверить с помощью Blast
- Затем нужно выровнять последовательности, и по получившемуся выравниванию, определить, какие последовательности включить в анализ

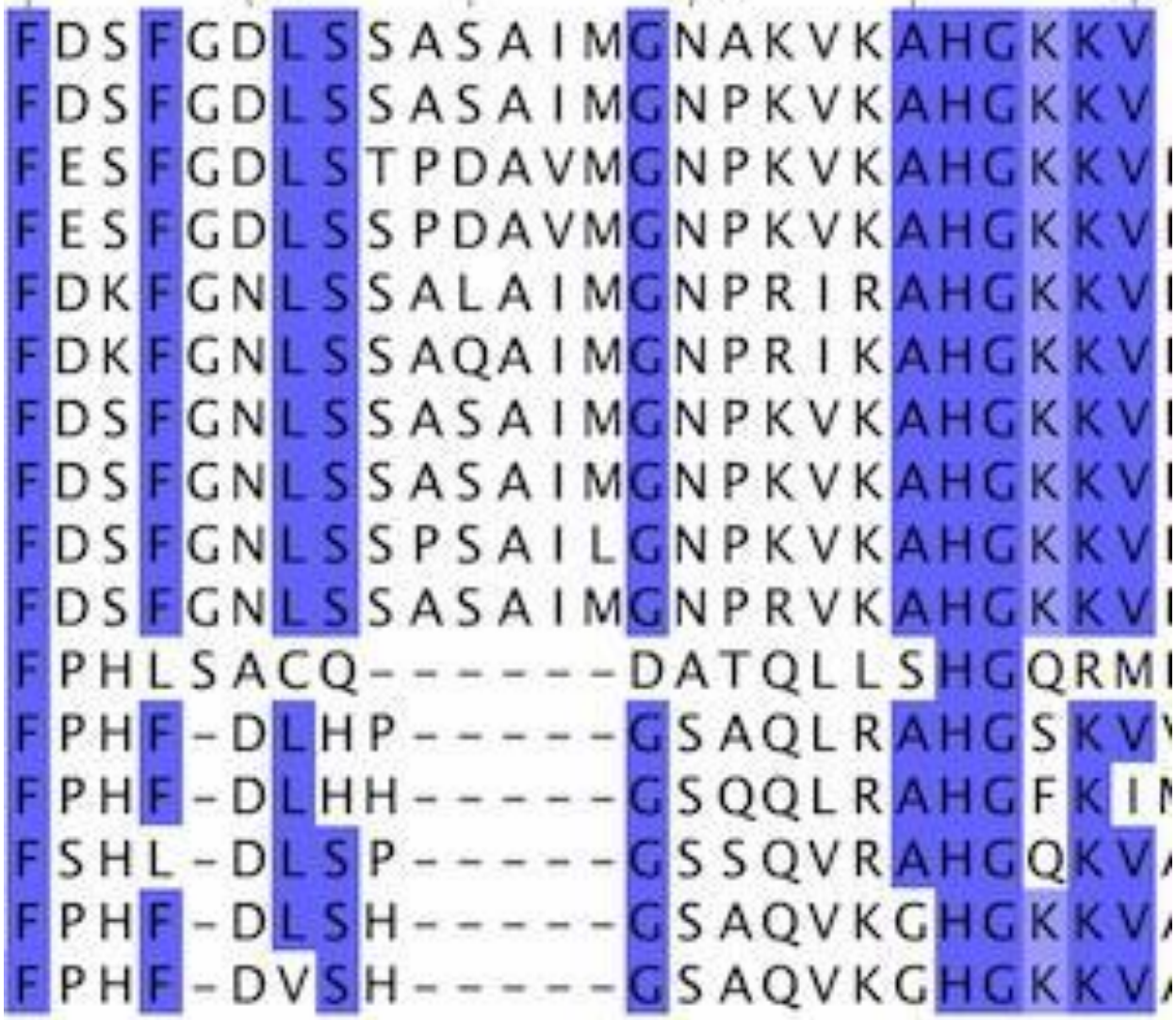


«Эффект тыквенного пирога»



Рецепт тыквенного пирога на филогенетическом дереве креветок.

Выбор последовательностей



Особенности молекулярной эволюции

2. Функционально менее важные молекулы или их части эволюционируют (накапливая эволюционные замены) быстрее, чем более важные

3. Мутационные замены, приводящие к меньшим нарушениям структуры и функции молекулы (консервативные замены), в ходе эволюции происходят чаще тех, которые вызывают существенное нарушение структуры и функции этой

Различия между деревом генов и деревом видов

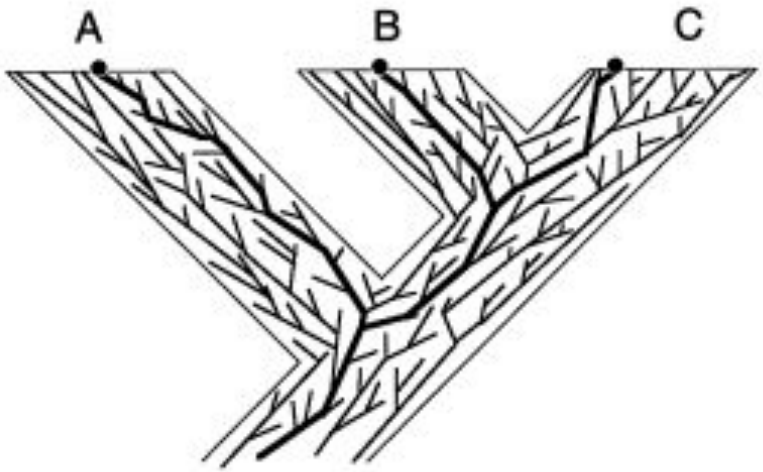


FIGURE 1. A gene tree contained within a species tree leading to three extant species: A, B, and C. Bold branches of gene tree show relationships among the sampled copies of the gene (●). Sampled copies from sister species B and C are sister copies.

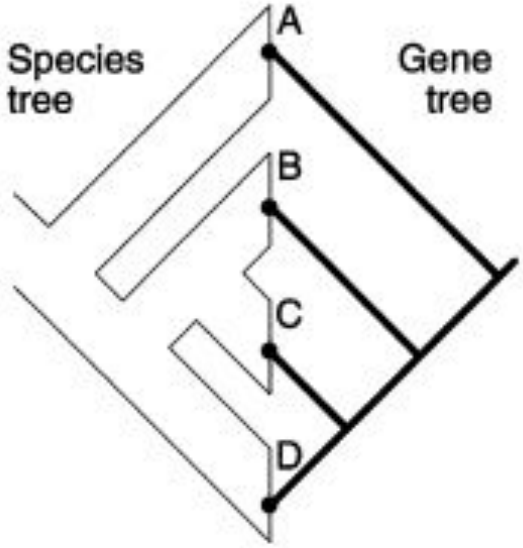
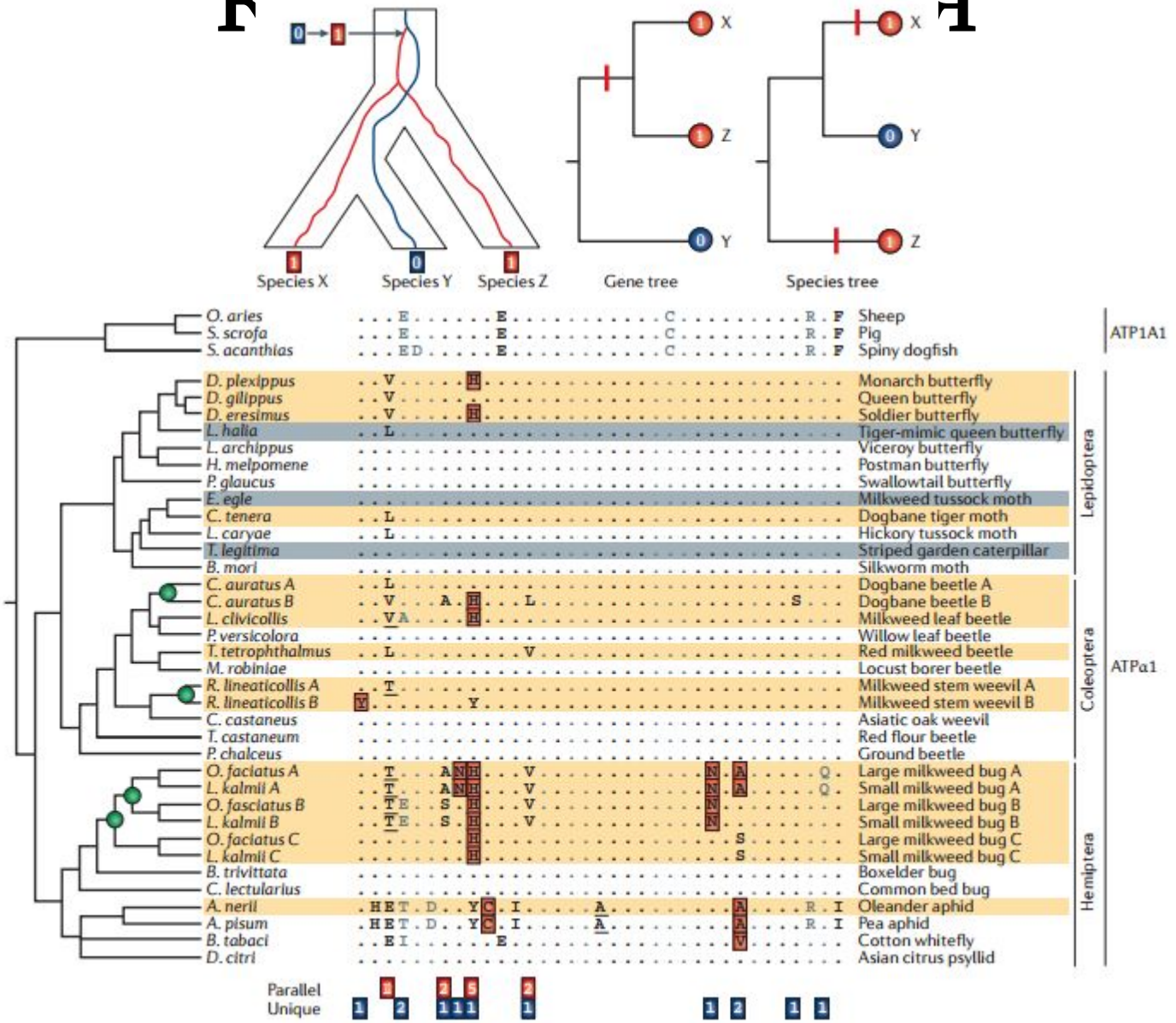


FIGURE 2. Discord between gene and species trees. At left is the species tree of four species, A, B, C, and D, and at right is the tree of a gene sampled one copy per species. Species B and C are sister species, but their gene copies are not sister copies.

Проблема: ортологи и паралоги

Молекулярная конвергенция



Филогенетические маркёры

Свойства:

- Гены, которые представлены одной копией в геноме лучше, чем те, у которых множество копий.
- Длина гена не должна варьировать у разных организмов
- Скорость изменения гена должна соответствовать скорости эволюции таксонов заданного уровня
- Должны легко подбираться специфические праймеры

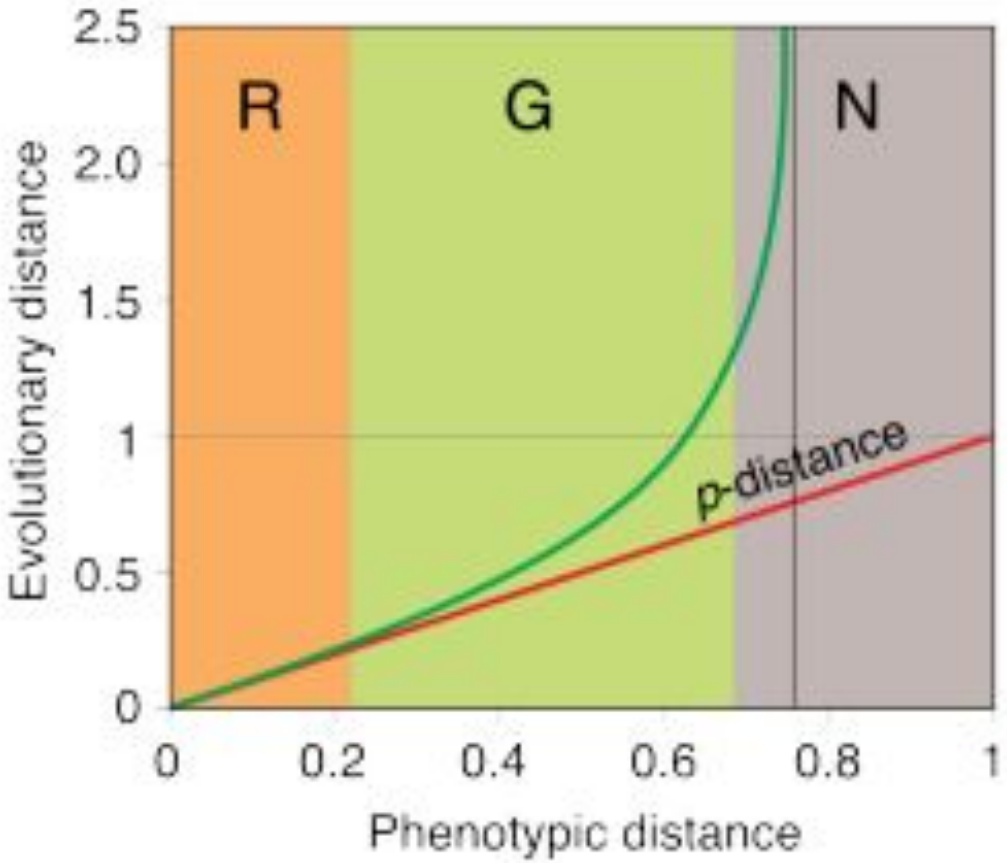


Figure 2: Regions of phenotypic distance corresponding to different estimates of evolutionary distance. R: "Parsimony zone", region where evolutionary distance is accurately approximated by phenotypic distance *d*. G: "Probabilistic zone", region where the *p*-distance under-estimates evolutionary distance. N: "Mutational saturation zone", region where the evolutionary distance cannot be estimated because of loss of phylogenetic information.

Филогенетические маркёры

- Рибосомальные гены
- Митохондриальные гены (COI/II, 12s RNA, cyt b)
- Хлоропластные гены
- Гены домашнего хозяйства и некоторые другие ядерные

Gene	Description	Reference
EF-1α	Elongation factor-1α, Role in protein synthesis.	[52]
rpoA gene	Encoding the alpha subunit of RNA polymerase	[53]
atpB	Encode the beta subunit of ATP synthase	[54]
dnaA	involved in DNA synthesis initiation	[55]
ftsZ	Role in cell division	[56]
gapA	Codes for glyceraldehyde phosphate dehydrogenase	[57]
groEL	Encodes bacterial heat shock protein.	[58]
gltA	Encoding citrate synthase	[59]
ITS	Piece of non-functional RNA situated between structural	[60]
	ribosomal RNAs precursor transcript.	
lux Gene	encode proteins involved in luminescence	[61]
PEPCK	Codes for phosphoenolpyruvate carboxykinase	[62]
pyrH genes	Codes for uridine monophosphate (UMP) kinases	[63]
recA	Role in recombination	[64]
U2 snRNA	Component of the spliceosome	[65]
Wsp gene	Encodes a major cell surface coat protein	[66]
Nuclear H3	Codes for protein which is associated with DNA	[67]
trnH-psbA	Non-coding intergenic spacer region located in plastid genome	[68]
rpoB, rpoC1	Coding region located in plastid genome	[69]

Выбор модели замен

A. Jukes-Cantor model

A	-	λ	λ	λ
T	λ	-	λ	λ
C	λ	λ	-	λ
G	λ	λ	λ	-

λ is the rate of substitution.

C. Kimura 2-parameter model

A	-	ß	ß	α
T	ß	-	α	ß
C	ß	α	-	ß
G	α	ß	ß	-

α and ß are the rates of transitional and transversional substitution, respectively.

E. Hasegawa *et al.* model

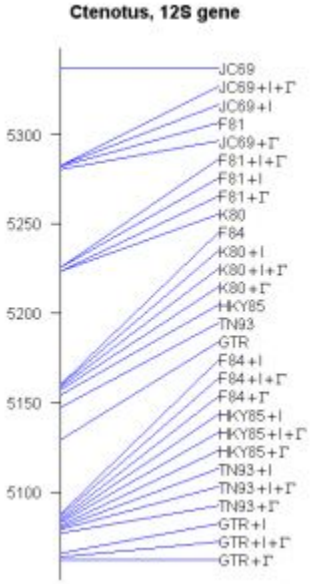
A	-	g_Tß	g_Cß	g_Gα
T	g_Aß	-	g_Cα	g_Gß
C	g_Aß	g_Tα	-	g_Gß
G	g_Aα	g_Tß	g_Cß	-

α and ß are the rates of transitional and transversional substitution, respectively, and g_i denotes the nucleotide frequencies (i=A,T,C,G).

F. Tamura-Nei model

A	-	g_Tß	g_Cß	$g_G\alpha_1$
T	g_Aß	-	$g_C\alpha_2$	g_Gß
C	g_Aß	$g_T\alpha_2$	-	g_Gß
G	$g_A\alpha_1$	g_Tß	g_Cß	-

α_1 and α_2 are the rates of transitional substitution between purines and between pyrimidines, respectively; ß is the rate of transversional substitution; and g_i denotes the nucleotide frequencies (i=A,T,C,G).



Результаты вычисления эволюционных дистанций будут отличаться в зависимости от выбранной модели замен

Выбор модели замен

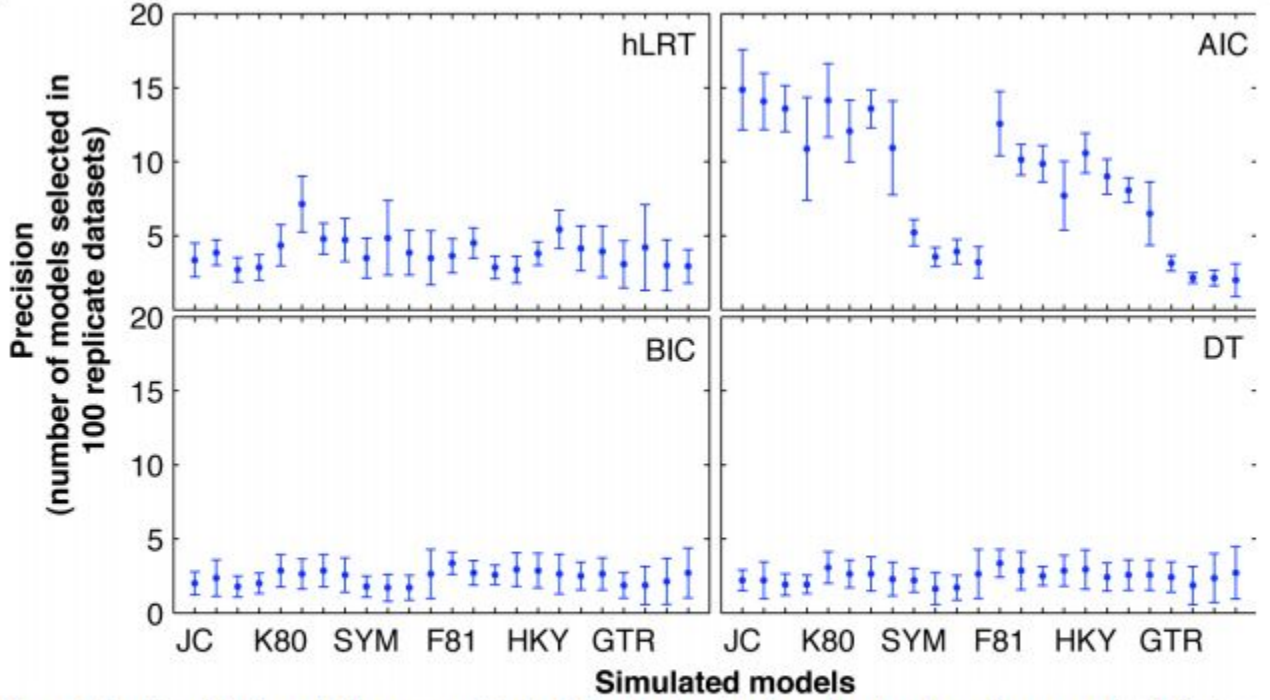


Figure 3 Precision of the four criteria corresponding to 24 simulated models. Categories along the x-axis represent the 24 simulated models. For the sake of clarity, only seven models are labelled, and each one is followed by three similar ones (e.g., JC is followed by JC + I, JC + Γ, and JC + I + Γ). The y-axis represents the means and standard deviations of precision values for each simulated model across the 14 simulations, which are different statistical results from those in Additional file 2. The markers denote the means, while lengths of error bars denote the standard deviation values.

AIC — Akaike's Information Criterion. Быстрее
BIC — Bayesian information criteria. Не «любит» более сложные модели
DT — decision theory
LRT — тест соотношения вероятностей. «Любит» более сложные модели.

Types of data used in phylogenetic inference:

Characters

```
Species A    ATGGCTATTCTTATAGTACG
Species B    ATCGCTAGTCTTATATTACA
Species C    TTCACTAGACCTGTGGTCCA
Species D    TTGACCAGACCTGTGGTCCG
Species E    TTGACCAGTTCTCTAGTTCG
```

Distances

	A	B	C	D	E
Species A	----	0.20	0.50	0.45	0.40
Species B	0.23	----	0.40	0.55	0.50
Species C	0.87	0.59	----	0.15	0.40
Species D	0.73	1.12	0.17	----	0.25
Species E	0.59	0.89	0.61	0.31	----

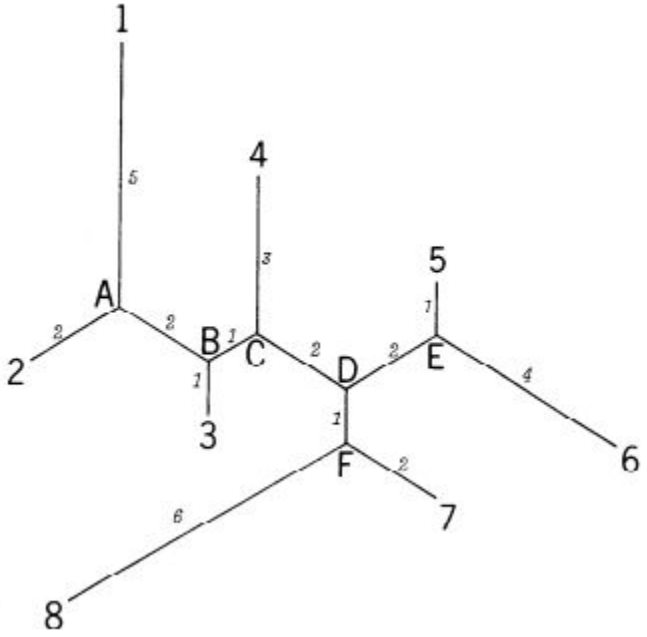
Example 1: Uncorrected "p" distance (=observed percent sequence difference)

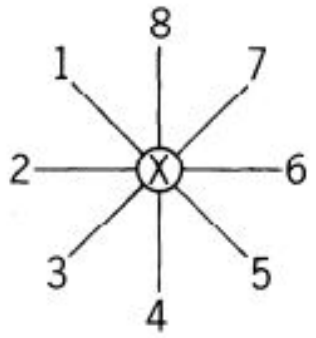
Example 2: Kimura 2-parameter distance (estimate of the true number of substitutions between taxa)

Методы реконструкции филогении

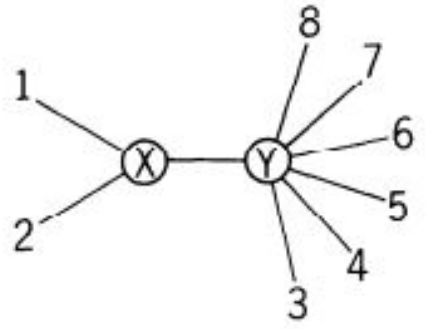
Дистанционные	Максимальной экономии	Максимальной вероятности
Используют только попарные дистанции	Используют только символьные данные	Используют все данные
Минимизация дистанции между ближайшими соседями	Минимизация общей длины дерева (минимизация числа мутаций	Максимизация вероятности заданного дерева с учётом заданных параметров
Очень быстрые	Медленные	***Очень*** медленные
Ищут локальный оптимум вместо глобального	Неверны при быстрой скорости эволюции	Сильно зависят от правильности выбранной модели
Хороши для чернового или предварительного	Лучший выбор для подходящей выборки(<30	Хороши для очень маленьких наборов данных и для оценки

Дистанционные методы Neghbor-joining





(a)



(b)

Начинаем с пары ветвей, которые меньше всего отличаются между собой

$$S_O = \sum_{i=1}^{N} L_{iX} = \frac{1}{N-1} \sum_{i<j} D_{ij},$$

$$L_{1X} + L_{2X} = D_{12},$$

$$L_{XY} = \frac{1}{2(N-2)} \left[\sum_{k=3}^{N} (D_{1k} + D_{2k}) - (N-2)(L_{1X} + L_{2X}) - 2 \sum_{i=3}^{N} L_{iY} \right].$$

$$\sum_{i=3}^{N} L_{iY} = \frac{1}{N-3} \sum_{3 \le i<j} D_{ij}.$$

Дистанционные методы Neghbor-joining

Table 2
S_{ij} Matrices for Two Cycles of the NJ Method for the Data in Table 1

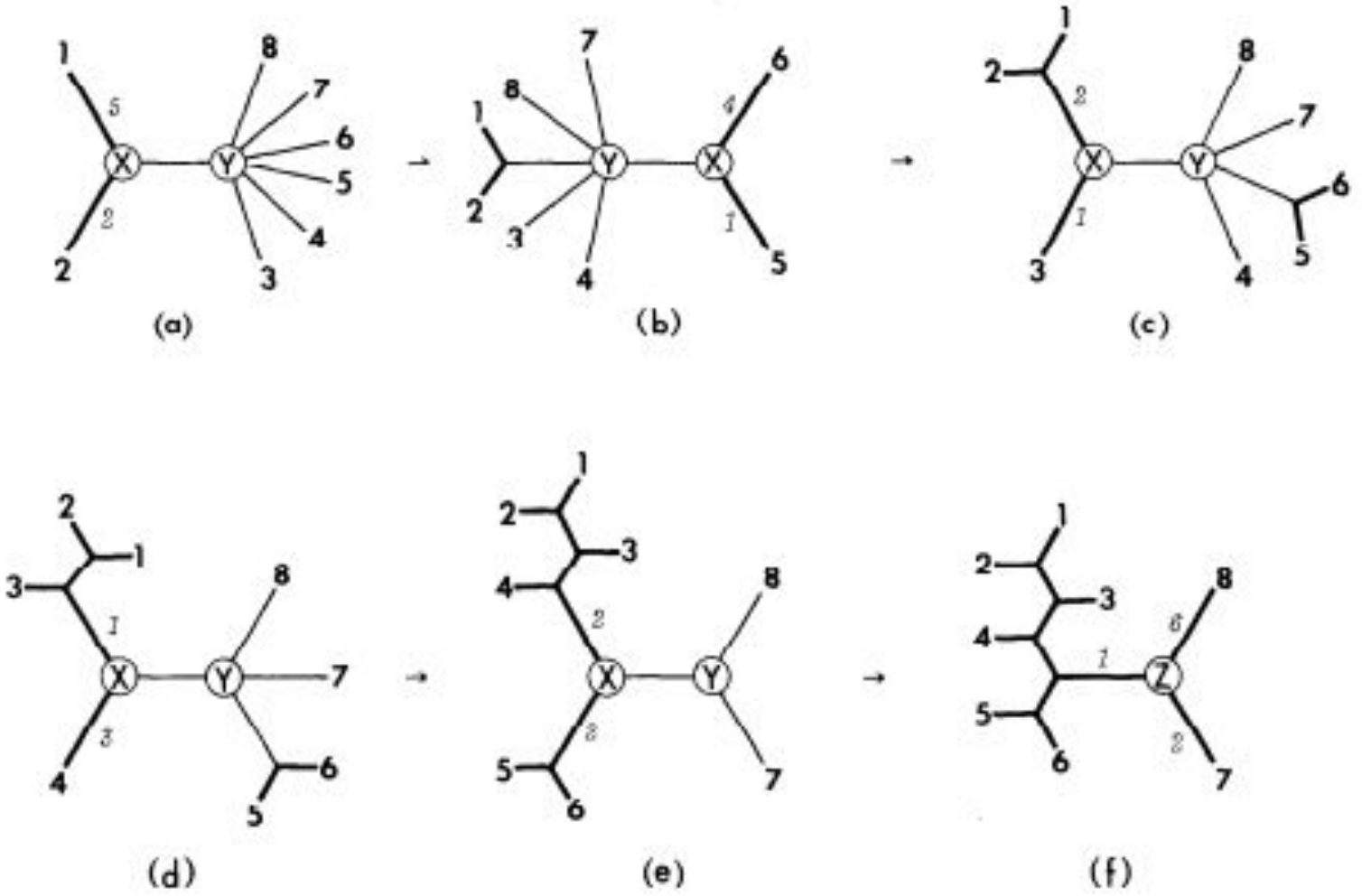
A. Cycle 1: Neighbors = [1, 2]

	OTU						
OTU	1	2	3	4	5	6	7
2 . .	36.67						
3 . .	38.33	38.33					
4 . .	39.00	39.00	38.67				
5 . .	40.33	40.33	40.00	39.67			
6 . .	40.33	40.33	40.00	39.67	37.00		
7 . .	40.17	40.17	39.83	39.50	38.83	38.83	
8 . .	40.17	40.17	39.83	39.50	38.83	38.83	37.67

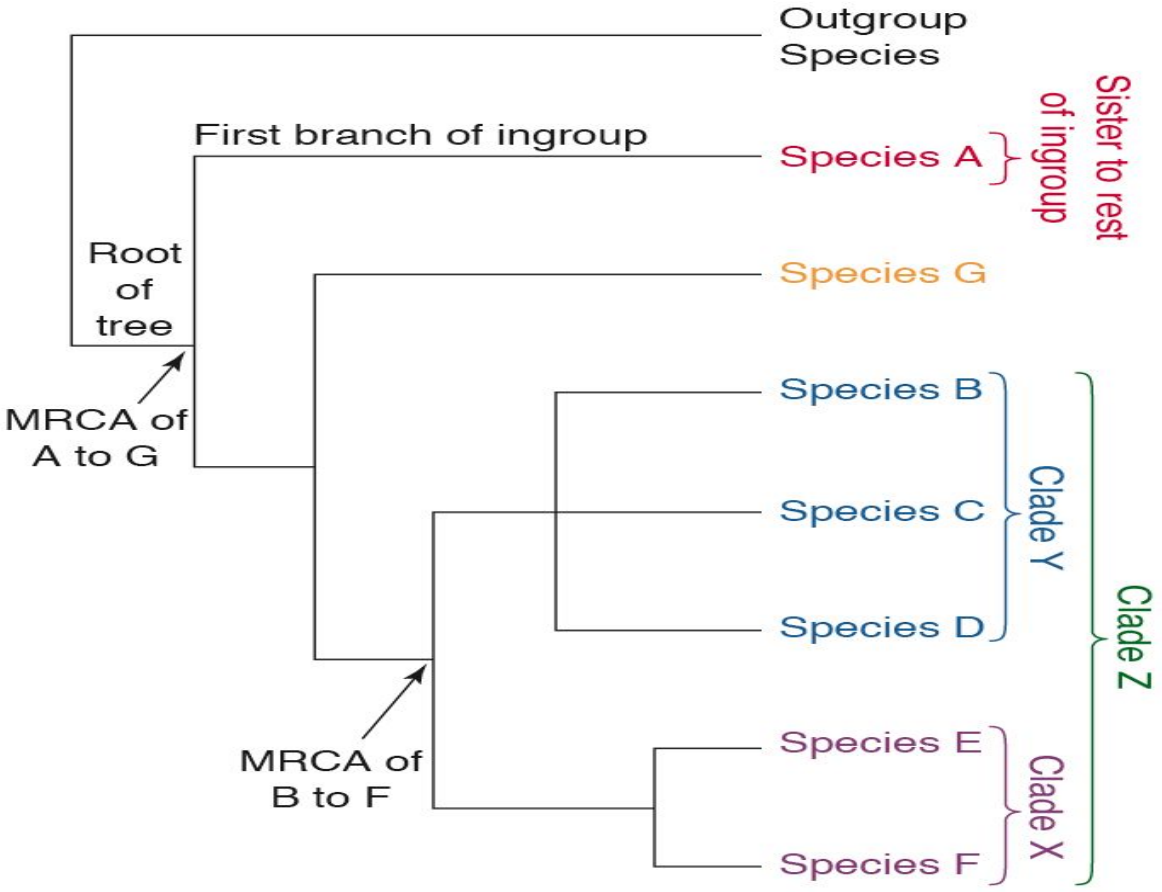
B. Cycle 2: Neighbors = [5, 6]

	OTU					
OTU	1-2	3	4	5	6	7
3 . .	31.50					
4 . .	32.30	32.30				
5 . .	33.90	33.90	33.70			
6 . .	33.90	33.90	33.70	31.30		
7 . .	33.70	33.70	33.50	33.10	33.10	
8 . .	33.70	33.70	33.50	33.10	33.10	31.90

Дистанционные методы
Neghbor-joining

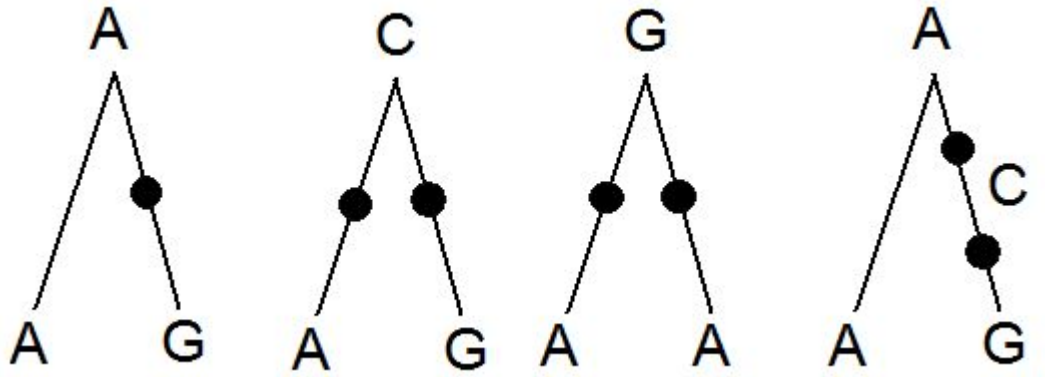


Зачем нужна аутгруппа



Молекулярно-филогенетические методы используют информацию о последовательностях внешней группы (контроля), дистанция от которой для всех остальных последовательностей заведомо выше, чем от других. Таким образом дерево «укореняется», а также внутри дерева убирается

Дистанционные методы Neghbor–joining



•Не учитываются обратные и параллельные замены => Мы считаем не настоящую дистанцию (расстояние), а редакционное расстояние.

•Вычислительно более быстрые.

•В большинстве случаев оценивают только топологию дерева, не воспроизводя исходную последовательность.

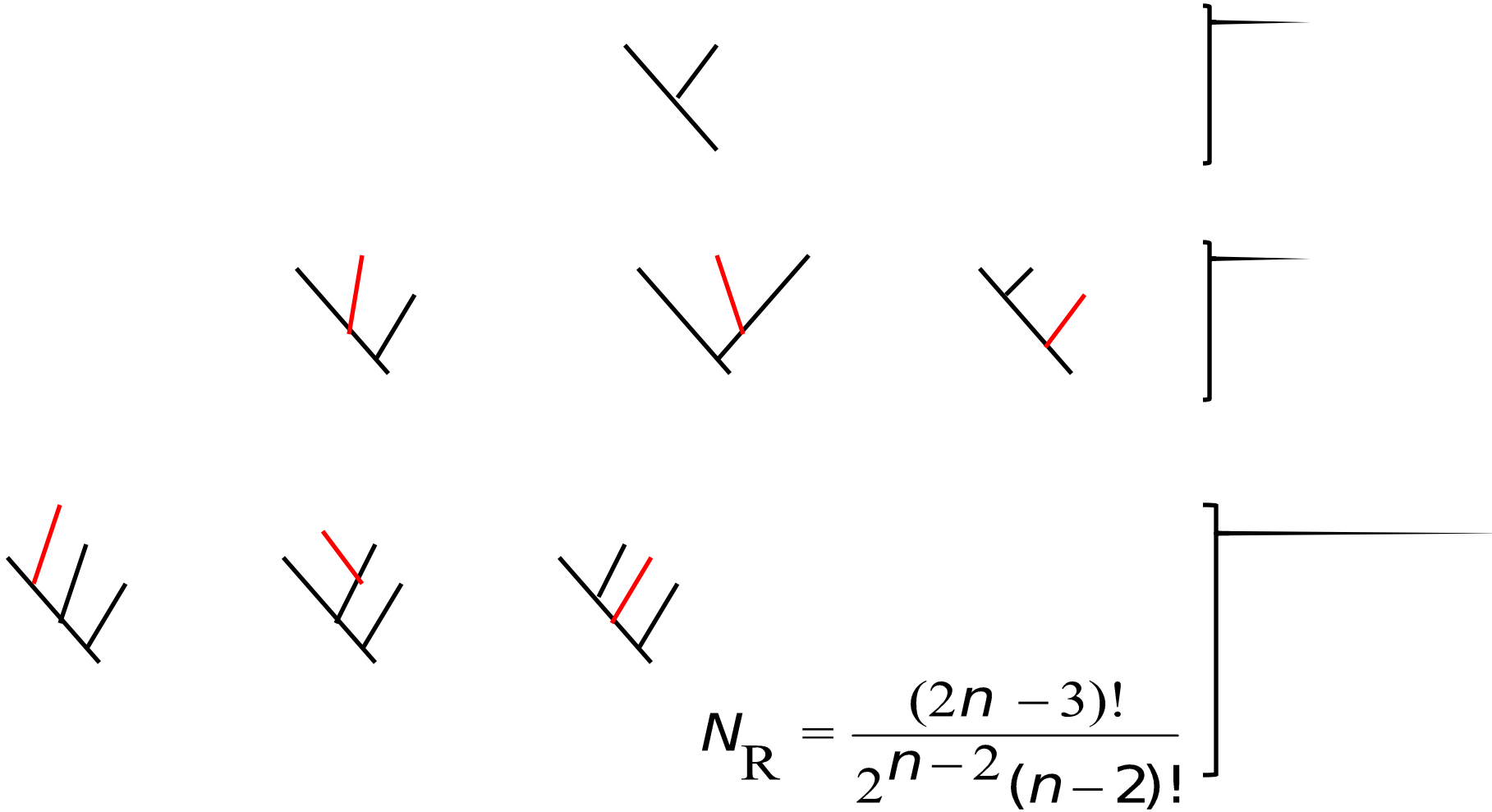
•Если у нас будет бесконечная последовательность, то мы с вероятностью 100% получим истинное

Методы максимальной экономии

- Минимизация числа замен символов
- Всегда реконструируют предковые последовательности
- Лучше работает на небольших наборах последовательностей во многих случаях на больших объёмах данных работает хуж

Let S be a set of n sequences, each of length n, over a fixed alphabet Σ. Let T be a tree leaf-labelled by the set S and with internal nodes labelled by sequences of length n over Σ. The *length* (or *parsimony score*) of T with this labelling is the sum, over all the edges, of the Hamming distances between the labels at the endpoints of the edge. (The Hamming distance between two strings of equal length is just the number of positions in which the two strings differ.) Thus the length of a tree is also the total number of point mutations along the edges of the tree. The *Maximum Parsimony (MP)* problem seeks the tree T leaf-labelled by S with the minimum length. While MP is NP-hard [4], constructing the optimal labeling of the internal nodes of a fixed tree T can be done in polynomial time [3].

Методы максимальной экономии



$$N_R = \frac{(2n-3)!}{2^{n-2}(n-2)!}$$

Число внешних узлов (таксонов)	2	3	4	5	10	20
Число возможных	1	3	15	105	34459425	**8200794532637891559375**

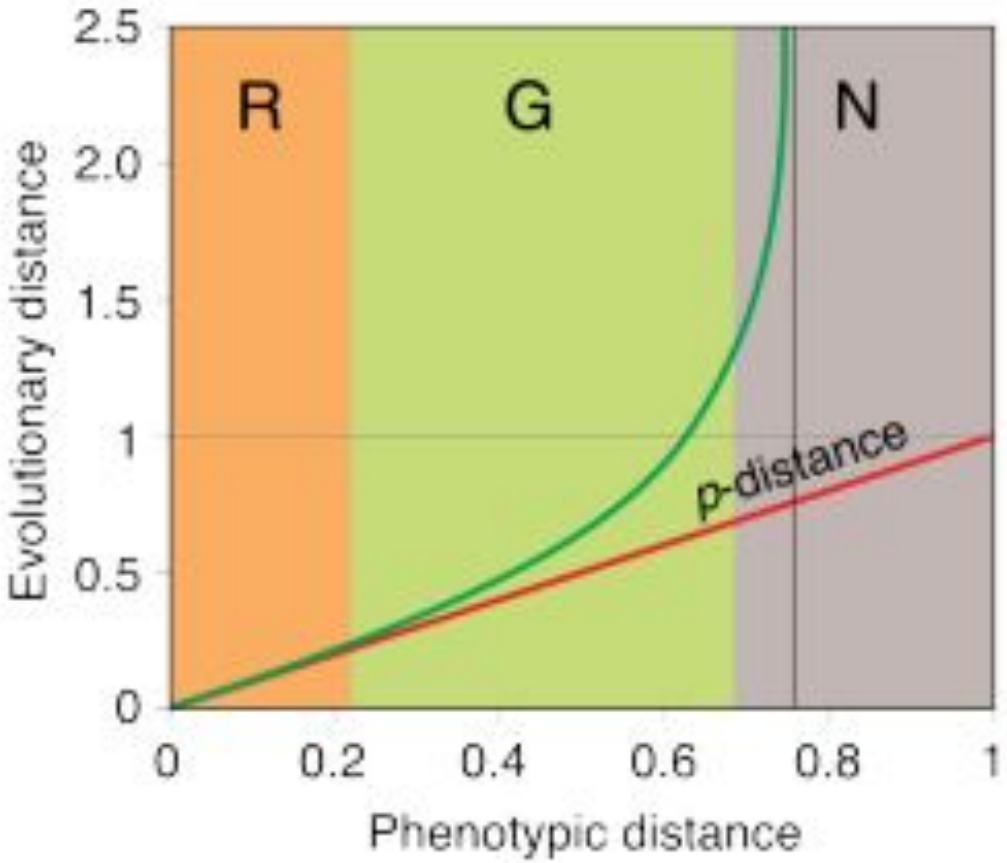
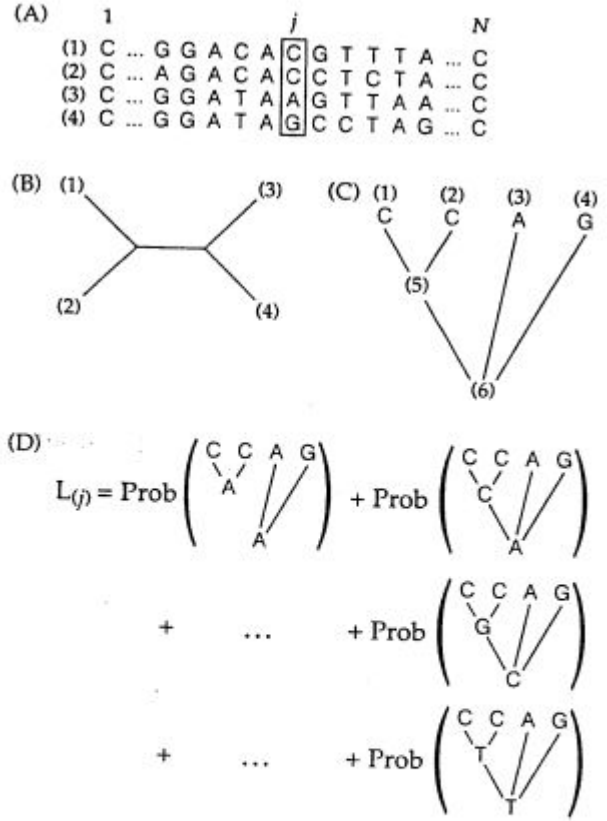


Figure 2: Regions of phenotypic distance corresponding to different estimates of evolutionary distance. R: "Parsimony zone", region where evolutionary distance is accurately approximated by phenotypic distance *d*. G: "Probabilistic zone", region where the *p*-distance under-estimates evolutionary distance. N: "Mutational saturation zone", region where the evolutionary distance cannot be estimated because of loss of phylogenetic information.

Методы максимальной вероятности

- Так же, как и в случае с методами максимальной экономии, генерирует все возможные топологии деревьев
- Предположение особой модели эволюции
- В отличие от метода максимальной экономии может предполагать разную скорость эволюции и скорость замен в разных ветвях дерева
- Поиск дерева с максимальной вероятностью существования, соответствующего данным
- Чем больше последовательность, тем вероятнее найти истинное дерево
- Самые медленные

Методы максимальной вероятности



- В позиции j для каждого внутреннего узла допустимы все четыре нуклеотида, значит всего 4*4=16 возможных деревьев.
- Каждое из деревьев это произведение вероятности возникновения какого-либо основания в корне дерева и вероятность его замены на тот, который в следующем узле. Т.е. частота нуклеотида умноженная на вероятность его мутации, если грубо.

A = 0.25 or средняя частота A в последовательности зависит от модели) f A->C трансверсия = 10^{-6} and A->G транзиции = $2x10^{-6}$ f

Вероятность T1 = 0.25 x 2x10-6 x 10-6 = 5x10-13

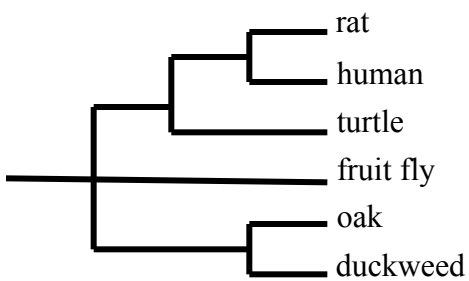
Оценка поддержки дерева

- Bootstrap

```
     0123456789
rat  GAGGCTTATC
human     GTGGCTTATC
turtle    GTGCCCTATG
fruitfly  CTCGCCTTTG
oak  ATCGCTCTTG
duckweed  ATCCCTCCGG
```

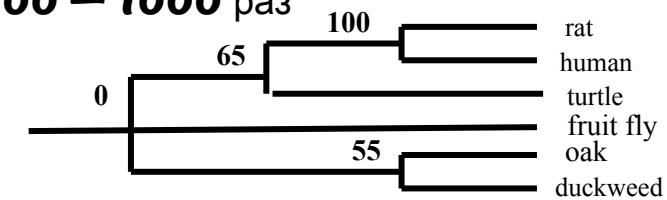
```
     001122234556667
rat  GGAAGGGGCTTTTTA
human     GGTTGGGGCTTTTTA
turtle    GGTTGGGCCCCTTTA
fruitfly  CCTTCCCGCCCTTTT
oak  AATTCCCGCTTCCCT
duckweed  AATTCCCCCTTCCCC
```

```
     445556777888899
rat  CCTTTTAAATTTTCC
human     CCTTTTAAATTTTCC
turtle    CCCCCTAAATTTTGG
fruitfly  CCCCCTTTTTTTTGG
oak  CCTTTCTTTTTTTGG
duckweed  CCTTTCCCCGGGGGG
```



Inferred tree

Повторить перестановку ***100 — 1000*** раз

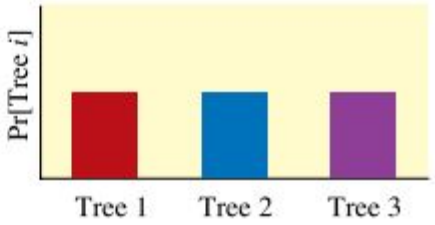


Оценка поддержки дерева

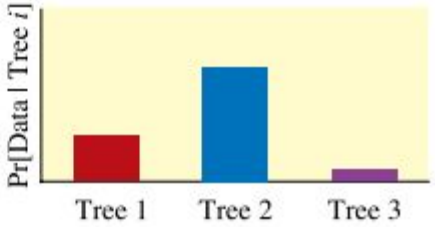
- Bayes inference

$$\Pr[\text{Tree} \mid \text{Data}] = \frac{\Pr[\text{Data} \mid \text{Tree}] \times \Pr[\text{Tree}]}{\Pr[\text{Data}]}$$

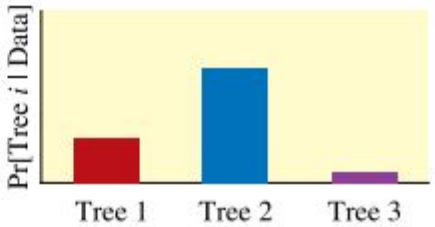
The **prior probability** of a tree represents the probability of the tree before the observations have been made. Typically, all trees are considered equally probable, a priori. However, other information can be used to give some trees more prior probability (e.g., the taxonomy of the group).



The **likelihood** is proportional to the probability of the observations (often an alignment of DNA sequences) conditional on the tree. This probability requires making specific assumptions about the processes generating the observations.



The **posterior probability** of a tree is the probability of the tree conditional on the observations. It is obtained by combining the prior and likelihood for each tree using Bayes' formula.



	BAYES				MLBOOT			
	Correct model		Incorrect model		Correct model		Incorrect model	
Support category (%)	N	% wrong	N	% wrong	N	% wrong	N	% wrong
70.1–80.0	418	17.3	262	54.6	612	10.8	634	15.3
80.1–90.0	542	10.2	337	37.4	713	4.5	622	8.5
90.1–95.0	409	6.2	292	40.4	399	2.0	301	4.7
95.1–100.0	1,216	1.7	2,622	15.6	505	0.2	293	2.0