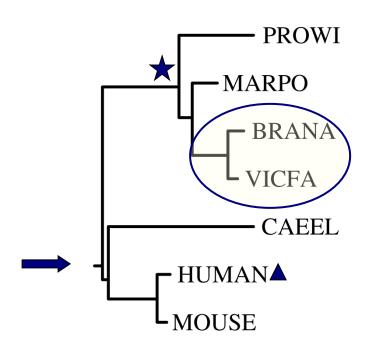
Филогенетические деревья

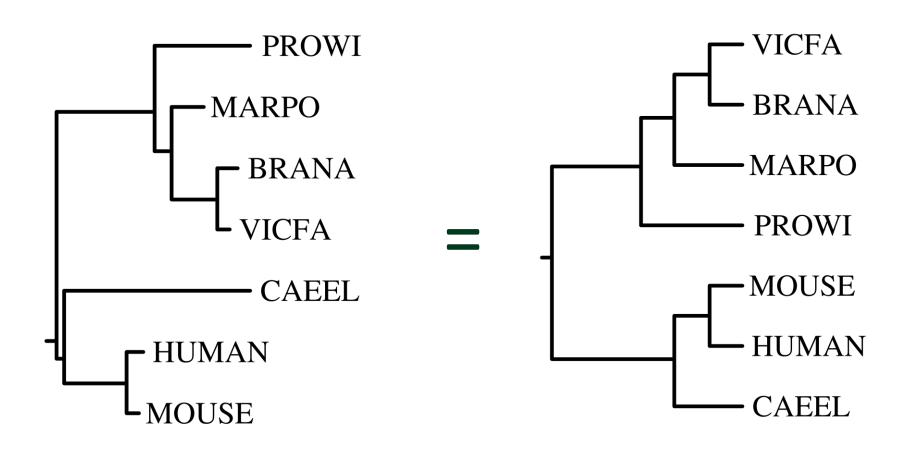
The time will come, I believe, though I shall not live to 2589 - Novosphingobium aro 3089 - Rhodospirillum rubrum ATC 388 - Finodobacter sphaero see it, when we shall have fairly true genealogical trees 2488 - Sinorhizobium meliloti 102 FixK - Agrobacterium tumefacie of each great kingdom of Nature. 635 - Microbulbifer degradans 148 - Carynebacterium glutamicum 525 - Corynebecterium efficiens Y83 **Charles Darwin** 2043 - Rhodopseudomonas palustris NnrR - Rhodobacter sphaeroides 3 Nnr - Rhodobacter sphaeroides ssc 2270 - Rhodobacter sphaeroides NnrR - Rhodobacter sphaeroides 3 NnrR - Rhodobacter sphaeroides 36963 | gi 2360963 602 - Rhodopseudomonas palustris CGA009 | gi 22961327 NnrR - Bradyrhizobium japonicum USDA110 | gi 27382196 NnrR - Brucella suis 1330 | gi 23500037 679 - Sinorhizobium meliloti 1021 | gi 16263132 974 - Agrobacterium tumefaciens C58I gi 15890603 NnrR - Pseudomonas sp. G179 | gi 3925389 Dry - Alcaligenes faecalis S-6 | gi 6978954 II986 - Brucella melitensis 16M | gi 17989331 62 - Brucella suis 1330 | gi 23500019 2234 - Chloroflexus aurantiacus J-10-fi | gi 22972351 3383 - Magnetococcus sp. MC-1 | gl 23001765 2322 - Rhodobacter sphaeroides 2.4.1 gi 22968740 2366 - Rhodopseudomonas palustris CGA009 | gl 22963069 2584 - Magnetospirilium magnetotacticum MS-1| gi 23008552 2313 - Rhodospirilium rubrum ATCC11170 | gi 22967700 7790 - Magnetospirillum magnetotacticum MS-1| gi 23015198 7817 - Magnetospirillum magnetotacticum MS-1| gi 23015226 1661 - Desulfitobacterium hafniense DCB-2| gi 23113143 1156 - Clostridium thermocellum ATCC27405 | gi 23021072 6285 - Magnetospirillum magnetotacticum MS-11 qi 23013640 834 - Deinococcus radiodurans R1 | gi 15806860 5554 - Bradythizobium japonicum USDA110j gi 27380665 1169 - Synechocystis sp. PCC6803 | gl 16330881 1271 - Nostoc punctiforme PCC73102 | gi 23124928 1289 - Nostoc punctiforme PCC73102 gi 23124926 1270 - Nostoc punctiforme PCC73102 | gi 23124927 4381 - Burkholderia fungorum LB400 | gi 22986446 4858 - Ralstonia metalliduransCH34 | gi 22980144 4648 - Burkholderia fungorum LB400 | gi 22986717 7807 - Burkholderia fungorum LB400 | gi 22989856 6063 - Burkholderia fungorum LB400 | gi 22968124 5668 - Burkholderia fungorum LB400 | gi 22986737 2110 - Mesorhizobium loti MAFF303099 | gi 13471969 3112 - Magnetospirilium magnetotacticum MS-1 | gi 23009392 5066 - Magnetospirillum magnetotacticum MS-1 gi 23012191 4644 - Rhodopseudomonas palustris CGA009 | qi 22965311 395 - Bradythizobium japonicum USDA110 | gi 2737550 4600 - Magnetospirillum magnetotacticum MS-1 | gi 23011566 536 - Bradyrhizobium japonicum USDA110 | gi 27375647 75C - Mycobacterium tuberculosis H37Rv | gi 15608813 4604C - Streptomyces coelicolor A3-2 | gi 6491809 346 - Rhodopseudomonas palustris CGA0091 di 22961075 876 - Rhodopseudomonas palustris CGA009 | gi 22961597 2018 - Desulfitobacterium hafniense DCB-2 | gi 23113527 4009 - Bacillus anthracis A2012 | gi 21401384 Yell - Escherichia coli O157-H7-EDL933 | gl 15802719 144C - Streptococcus mutans UA159 gi 2437866 ArcR - Cenococcus ceni | gi 1702627 Bacillus ligheniformis ATCC14580 | gi 8894540 548 - Streptococcus pyogenes M1-GAS-SF370 | gi 15675445 161 - Streptococcus agalactiae 2603V/R | gi 22538295 706 - Enterococcus faecium | gi 22992162 104 - Staphylococous epidermidis ATCC12228 | gi 27467022 2214 - Staphylococcus epidermidis ATCC12228 | gi 27469132 2631 - Staphylococcus aureus ssp. aureus Mu50 | gi 1592562 2132 - Listeria monocytogenes EGDe | gi 16804171 2131 - Listeria monocytogenes EGDe | gi 16804170 2269 - Listeria Innocua CLIP11262 gi 16801333 112 - Listeria monocytogenes EGDe | gi 16802160 159 - Listeria innocua CLIP11262| gi 16799236 2165 - Listeria monocytogenes EGDe | gi 16804204 2268 - Listeria innocua CLIP11262| gi 16801332 597 - Listeria monocytogenes EGDe | gi 16802640 - 606 - Listeria innocua CLIP11262i gi 16799681 1418 - Desulfitobacterium hafnienseDCB-2| gi 23112881 1441 - Clostridium perfringens 13A | gi 18310423 1511 - Clostridium acetobutylicum ATCC824 | gi 15894789 466 - Campylobacter jejuni NCTC11168 | gi 15791830 171 - Thermotoga maritima MSBB | gi 15643927 753 - Listeria monocytogenes EGDe | gi 16802795 2860 - Desulfitobacterium hafniense DCB-2 | gi 23114411 2522 - Clostridium perfringens 13A | gi 18311504

Описание структуры дерева (терминология)

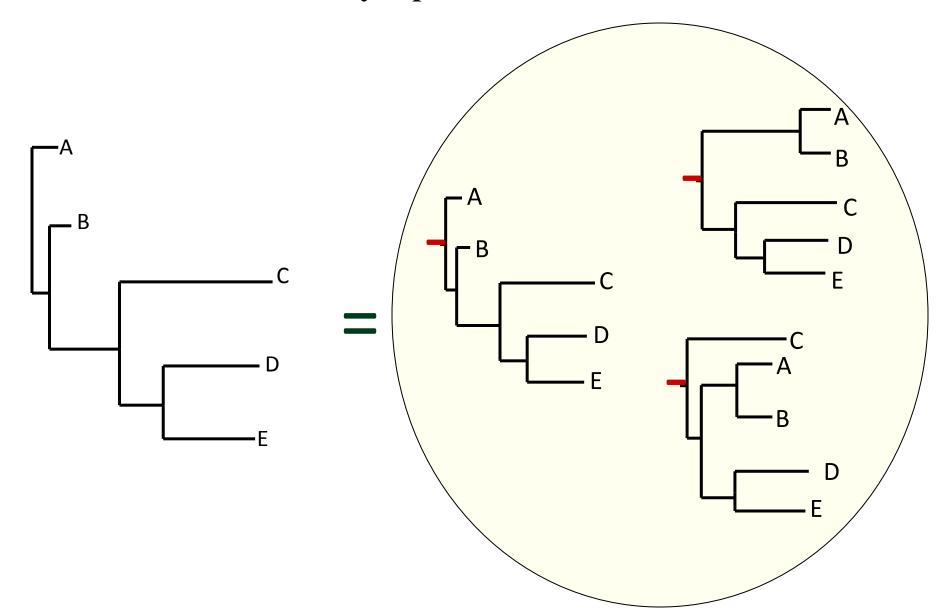
- <u>Узел (node)</u> точка разделения предковой последовательности (вида, популяции) на две независимо эволюционирующие. Соответствует внутренней вершине графа, изображающего эволюцию.
- <u>Лист</u> реальный (современный) объект; внешняя вершина графа. OTU : Operational Taxonomic Unit.
- Ветвь (branch) связь между узлами или между узлом и листом; ребро графа.
- **Корень (root)** гипотетический общий предок.
- <u>Клада</u> группа организмов, которые являются потомками единственного общего предка и всех потомков этого предка.



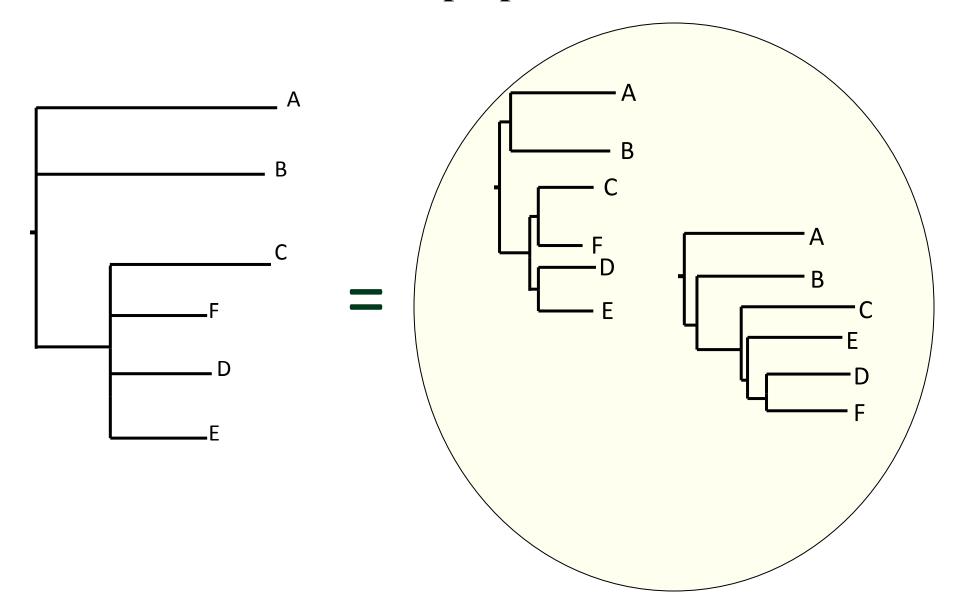
Топология дерева



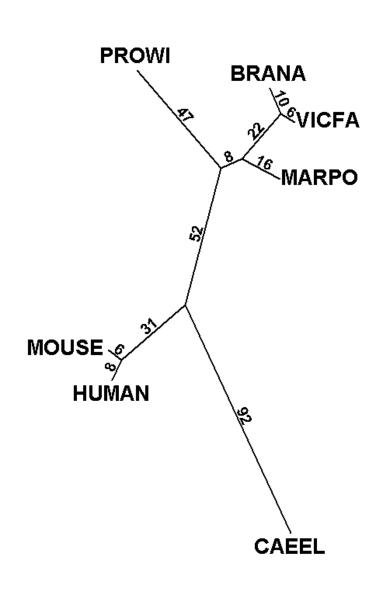
Неукоренённое дерево следует понимать как множество возможных укоренений



Небинарное дерево следует понимать как множество возможных «разрешений»



Расстояния по дереву между листьями



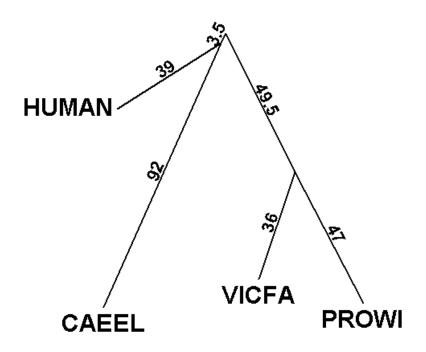
D(MOUSE, CAEEL) = 6+31+92 = 129

Дерево с заданными длинами ветвей порождает **метрическое пространство**, элементами которого являются листья.

Длины ветвей отражают эволюционные расстояния между листьями в данной модели дерева.

Эти расстояния могут численно заметно отличаться от эволюционных расстояний между последовательностями, определенными по Джуксу — Кантору или Кимура.

Скобочная формула



То же дерево, что и на предыдущем слайде, но укоренено в среднюю точку, и часть ветвей обрезана.

Newick Standard:

((HUMAN:39, CAEEL:92):3.5, (VICFA:36, PROWI:47):49.5);

«The reason for the name is that the second and final session of the committee met at Newick's restaurant in Dover, and we enjoyed the meal of lobsters.»

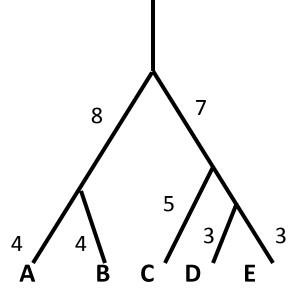
Joseph Felsenstein, http://evolution.genetics.washington.edu/phylip/newicktree.html

Ультраметрические деревья

Если на дереве можно найти точку такую, что расстояния от нее до всех листьев одинаковы, до дерево называется "ультраметрическим".

Ультраметрическое дерево можно однозначно укоренить (в эту самую точку).

Ультраметрическое пространство — особый случай метрического пространства, в котором метрика удовлетворяет усиленному неравенству треугольника: $d(x, z) \le \max(d(x, y), d(y, z))$.



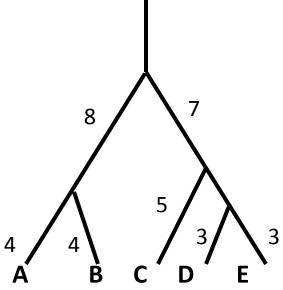
Ультраметрические деревья

Если на дереве можно найти точку такую, что расстояния от нее до всех листьев одинаковы, до дерево называется "ультраметрическим".

Ультраметрическое дерево можно однозначно укоренить (в эту самую точку).

Ультраметрическое пространство — особый случай метрического пространства, в котором метрика удовлетворяет усиленному неравенству треугольника: $d(x, z) \le \max(d(x, y), d(y, z))$.

Задача 19. Доказать равносильность предыдущих утверждений.



Ультраметрические деревья

Если на дереве можно найти точку такую, что расстояния от нее до всех листьев одинаковы, до дерево называется "ультраметрическим".

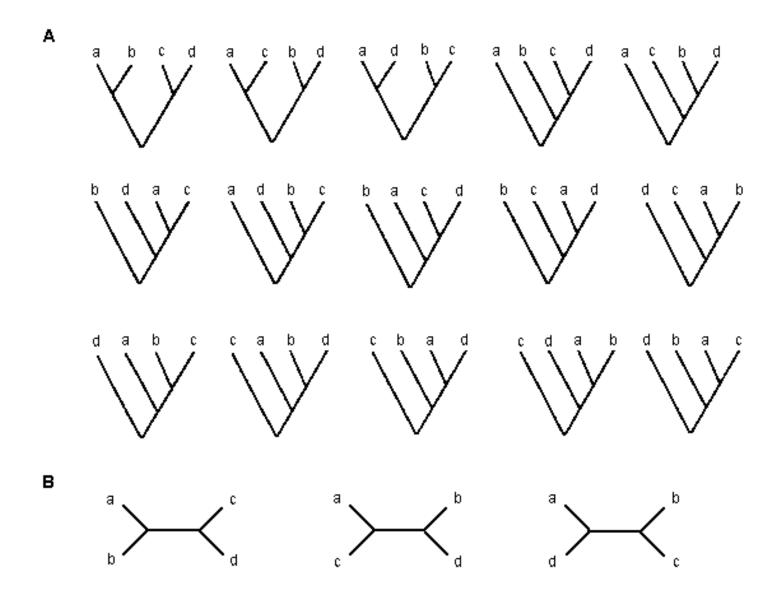
Ультраметрическое дерево можно однозначно укоренить (в эту самую точку).

Ультраметрическое пространство — особый случай метрического пространства, в котором метрика удовлетворяет усиленному неравенству треугольника: $d(x, z) \le \max(d(x, y), d(y, z))$.

Задача 19. Доказать равносильность предыдущих утверждений.

Содержательно ультраметрические деревья соответствуют случаю, когда длины ветвей суть время эволюции, и все последовательности современны (гипотеза молекулярных часов). 4

Задача 20. Сколько существует различных укорененных и неукорененных дихотомических деревьев для *m* листьев?



UPGMA

Unweighted Pair Group Method with Arithmetic Mean

(невзвешенная попарная кластеризация)

Кластерный метод, в котором расстояние между кластерами вычисляется как среднее арифметическое расстояний между их элементами.

K L M N
K 16 16 16
L 8 8
M 4
N

Шаг 1.

- 1. Находим самые близкие объекты.
- 2. Строим первый узел n1:

$$(M,n1) = d(N,n1) = d(M,N) / 2$$

3. Объединяем объекты и пересчитываем матрицу расстояний:

$$d(MN-K) = [d(M-K) + d(N-K)] / 2 = 16$$

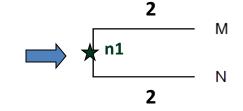
 $d(MN-L) = [d(M-L) + d(N-L)] / 2 = 8$

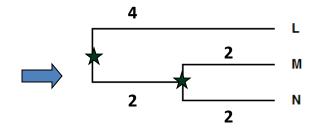
K L MN
K 16 16
L 8
MN

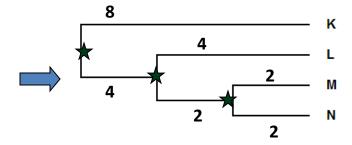
см. шаг 1.

Шаг 2.









UPGMA

2

2

M

Ν

2

Ν

Unweighted Pair Group Method with Arithmetic Mean

(невзвешенная попарная кластеризация)

Кластерный метод, в котором расстояние между кластерами вычисляется как среднее арифметическое расстояний между их элементами.

	K	L	M	N
K		16	16	16
L			8	8
M				4
N				

Шаг 1.

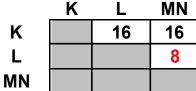
- 1. Находим самые близкие объекты.
- 2. Строим первый узел n1:

$$(M,n1) = d(N,n1) = d(M,N) / 2$$

3. Объединяем объекты и пересчитываем матрицу расстояний:

$$d(MN-K) = [d(M-K) + d(N-K)] / 2 = 16$$

 $d(MN-L) = [d(M-L) + d(N-L)] / 2 = 8$



Шаг 2.

см. шаг 1.

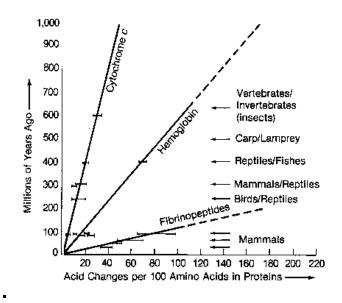


Получаем единственное укорененное ультраметрическое дерево

Гипотеза «молекулярных часов» (E.Zuckerkandl, L.Pauling, 1962)

За равное время во всех ветвях накапливается равное число мутаций.

Если гипотеза молекулярных часов принимается, число различий между выровненными последовательностями можно считать примерно пропорциональным времени.

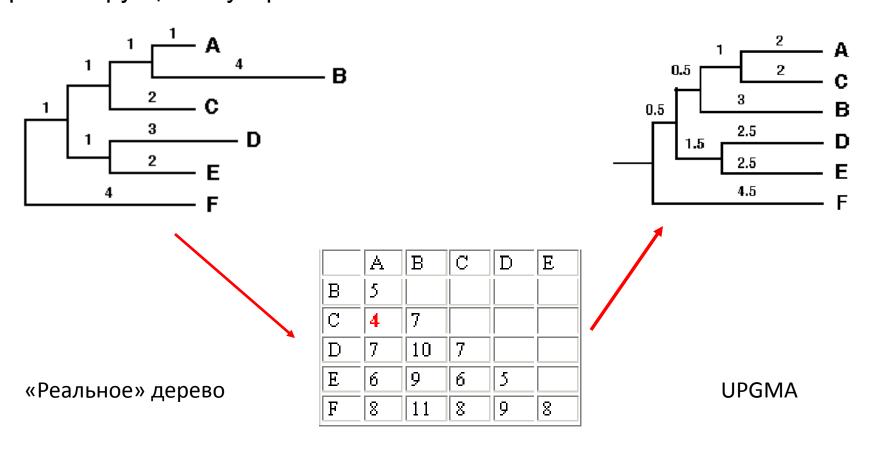


Отклонения от ультраметричности можно считать случайными. Эволюция реконструируется в виде **ультраметрического** дерева.

Если данные таковы, что гипотеза молекулярных часов не проходит, то реконструкция укорененного дерева намного менее надёжна, чем реконструкция неукоренённого

UPGMA

Если данные таковы, что гипотеза молекулярных часов не проходит, то реконструкция укорененного дерева намного менее надёжна, чем реконструкция неукоренённого



Задача 21. Написать программу, реализующую алгоритм UPGMA. Матрица расстояний задается вами.

Neighbour joining, NJ (метод ближайших соседей)

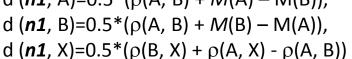
- Рисуем «звездное» дерево с ветвями условной длины, например, равными среднему расстоянию листа до всех остальных.
- Рассмотрим все m(m-1)/2 пар листьев и соединим ту пару, в которой листья близки друг к другу, но далеки ото всех остальных.

А и В — такая пара последовательностей, для которых минимальна величина

$$\rho(A, B) - M(A) - M(B),$$

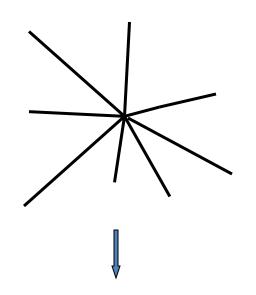
где ρ — расстояние из матрицы, а M — среднее расстояние от А или В до всех m-2 остальных последовательностей.

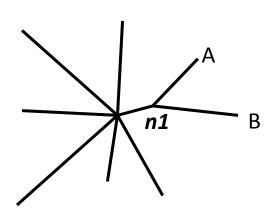
Объединим А и В в кластер АВ и строим первый узел, расстояние до которого зависит от среднего расстояния до других вершин: d (n1, A)=0.5*(ρ (A, B) + M(A) – M(B)),





5. Повторяем процедуру с п. 2 до тех пор, пока не останется 3 вершины





Задача 22. Написать программу, реализующую алгоритм Neighbour Joining. Матрица расстояний задается вами.

Переборные алгоритмы: выбор лучшего дерева из всех возможных

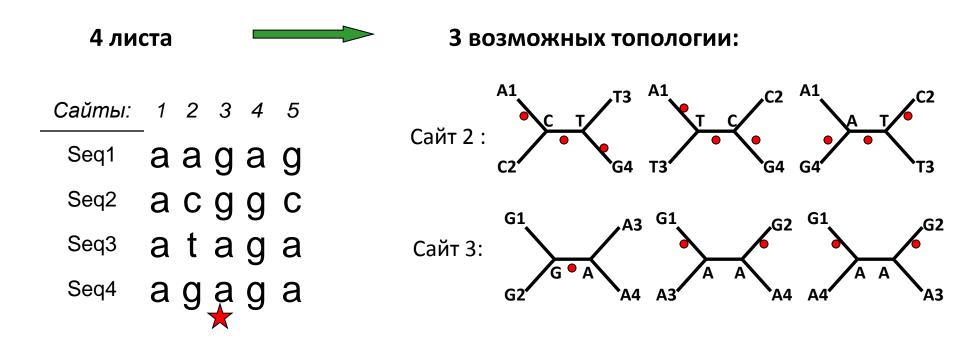
Основные критерии качества

- Максимальная экономия (maximum parsimony)
- Максимальное правдоподобие (maximum likelihood)
- Соответствие расстояний по дереву заданной матрице расстояний (Fitch Margoliash или minimal evolution)
- •

Лучшее дерево — не обязательно «истинное» дерево

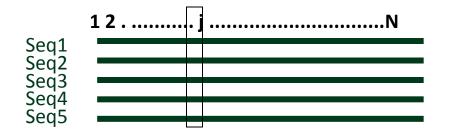
МР, метод максимальной экономии

Лучшее дерево — дерево, в котором различия в данных объясняются минимальным числом элементарных эволюционных событий.

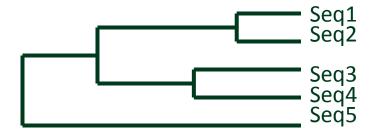


- 1. Найдем все информативные сайты.
 - Информативный сайт колонка выравнивания, символы в которой позволяют выбрать одну из возможных топологий, в данном примере сайт 3 информативный, а сайт 2 нет.
- 2. Для каждого из возможных деревьев определим минимальное число замен в каждом информативном сайте и сложим эти числа.
- 3. Выберем дерево с наименьшим числом замен; возможно, что окажется несколько деревьев с одинаковым числом замен.

Метод максимального правдоподобия



Рассмотрим все варианты деревьев, первый, например,



$$L = \max_{t \in T} \{ Pr(D|t) \},$$

где D – данные, Т –множество всех возможных деревьев, модель эволюции задана и фиксирована

Какова вероятность L(j) того, что в рамках принятой эволюционной модели (например, матрицы замен) и при данной эволюционной истории (т.е. при данной топологии дерева) получится исходное выравнивание в колонке j? Для полного выравнивания : L= L(1) x L(2) x L(N) Выбираем дерево, соответствующее максимальной вероятности или наиболее правдоподобное. Числа очень маленькие, поэтому

$$\ln L = \ln L(1) + \ln L(2) \dots + \ln L(N) = \sup_{j=1}^{N} \ln L(j)$$