

Филогенетические деревья

The time will come, I believe, though I shall not live to see it, when we shall have fairly true genealogical trees of each great kingdom of Nature.

Charles Darwin

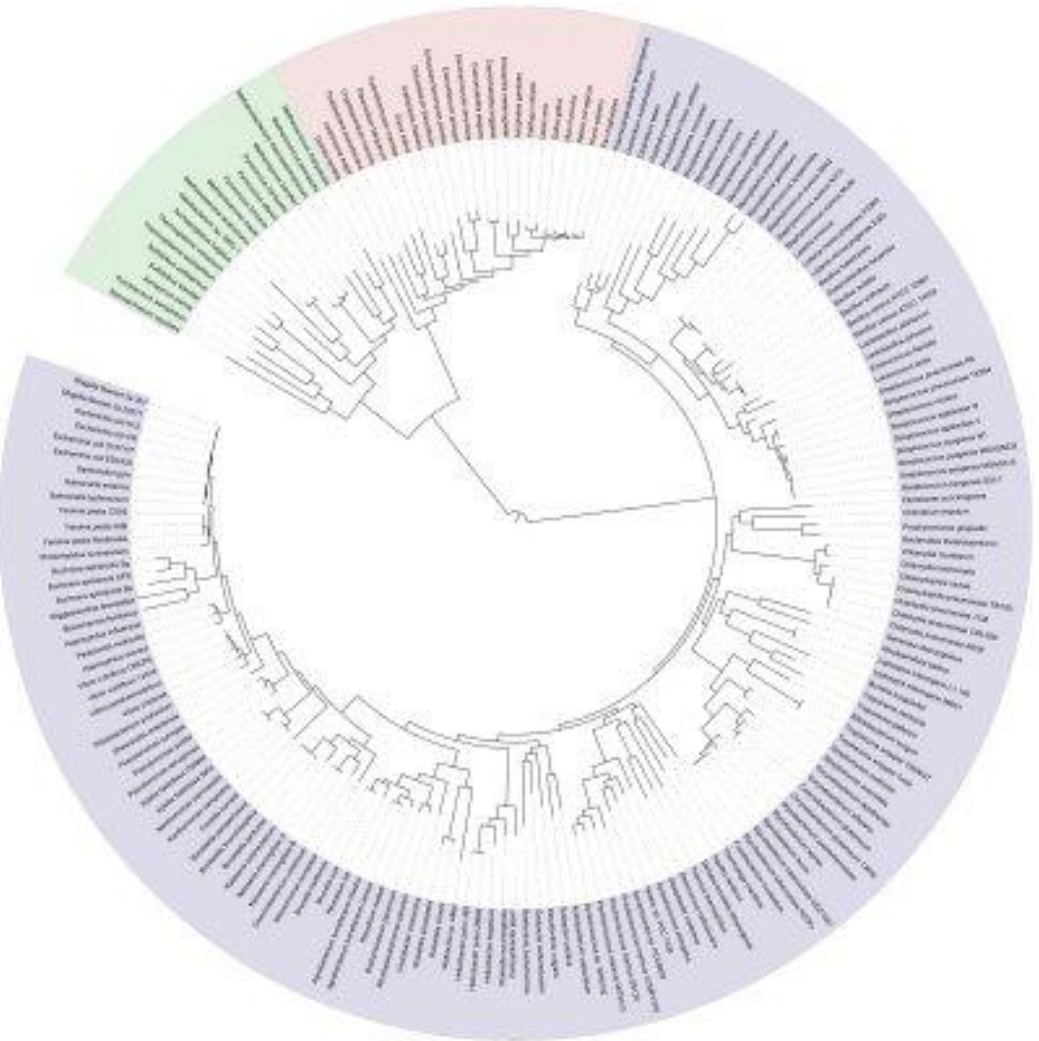
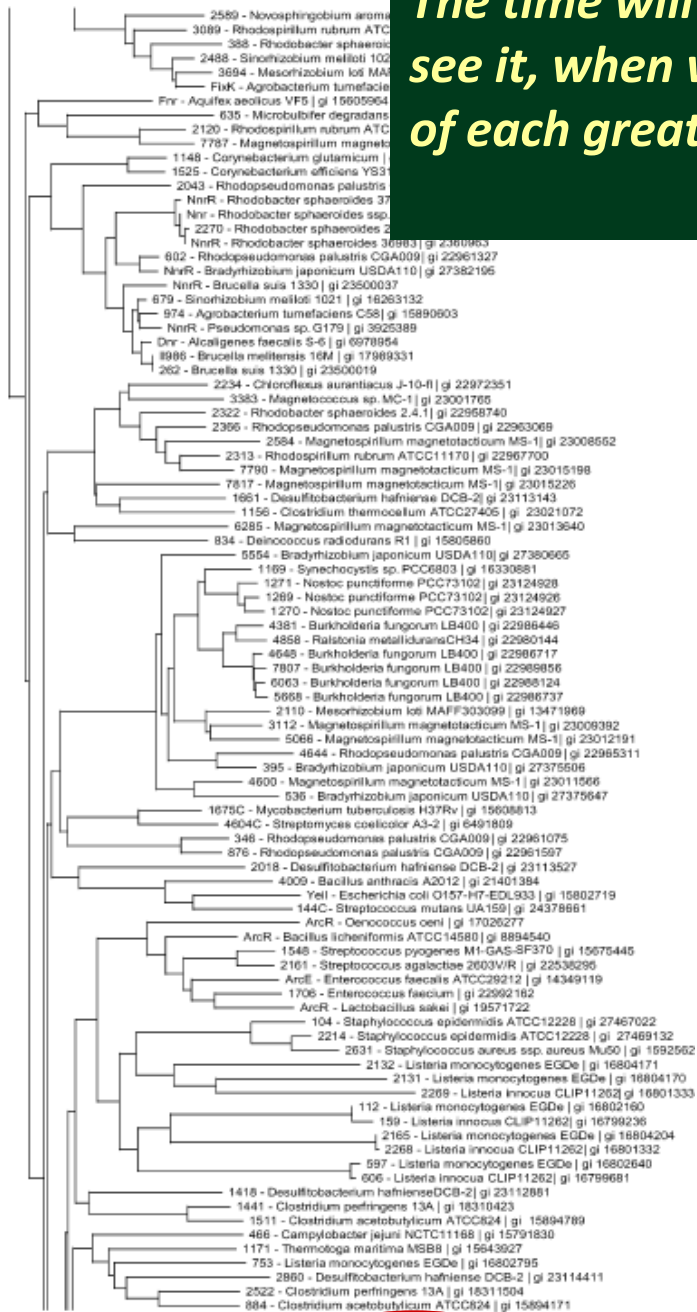
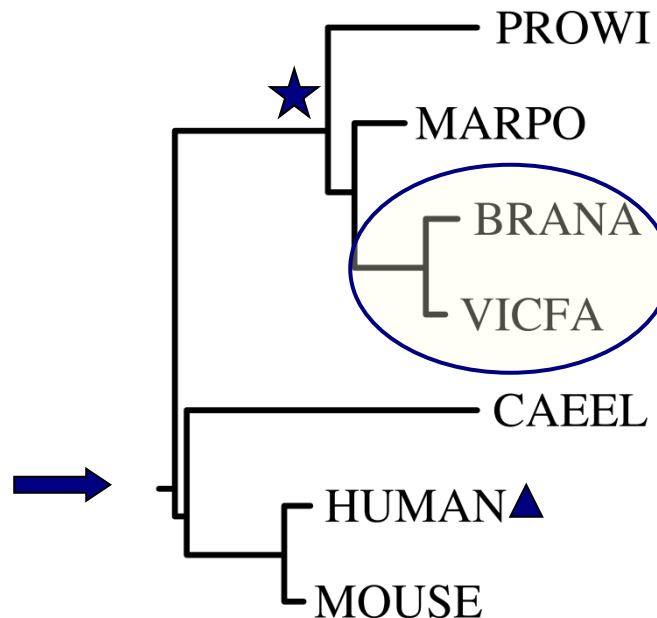


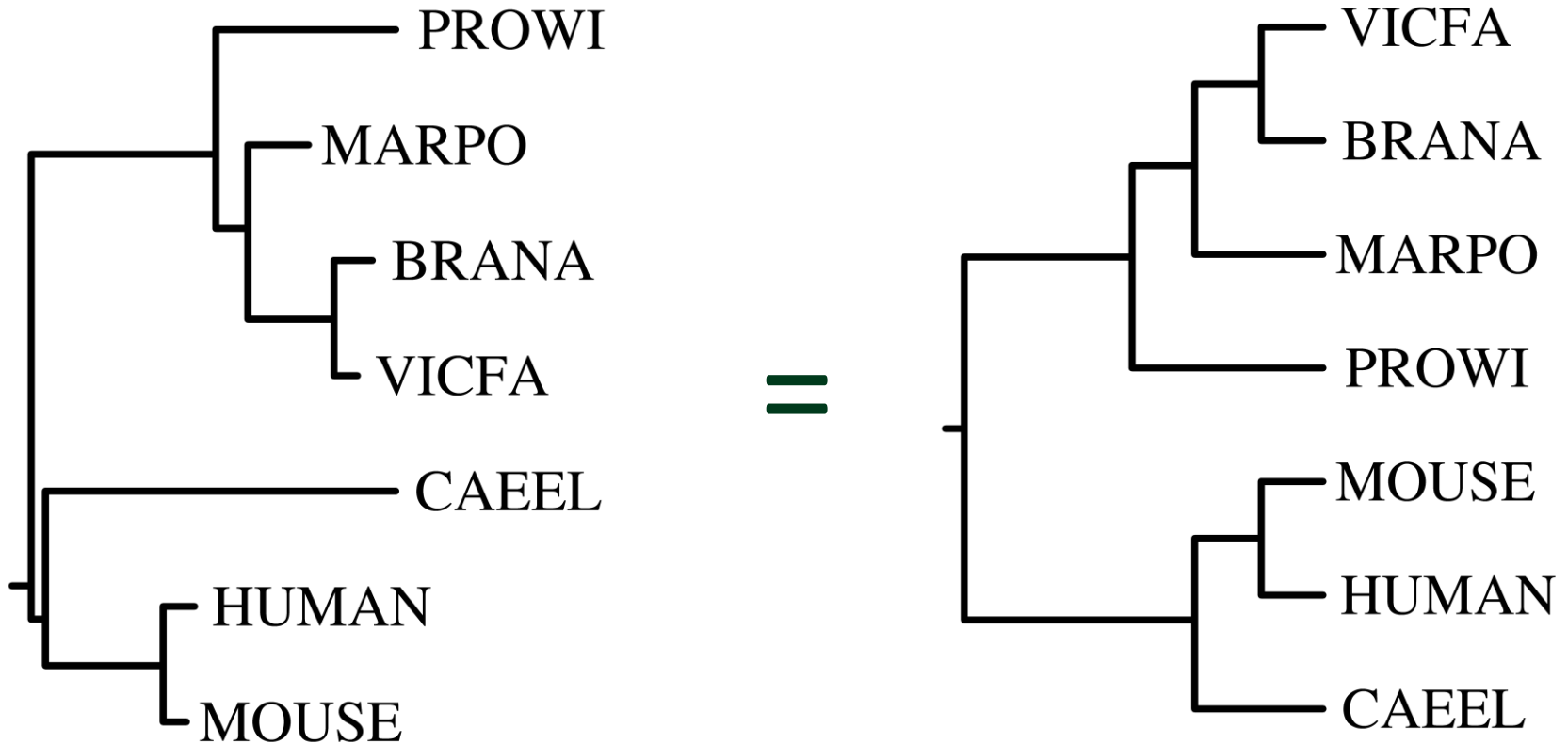
Fig. 1 (Continued)

Описание структуры дерева (терминология)

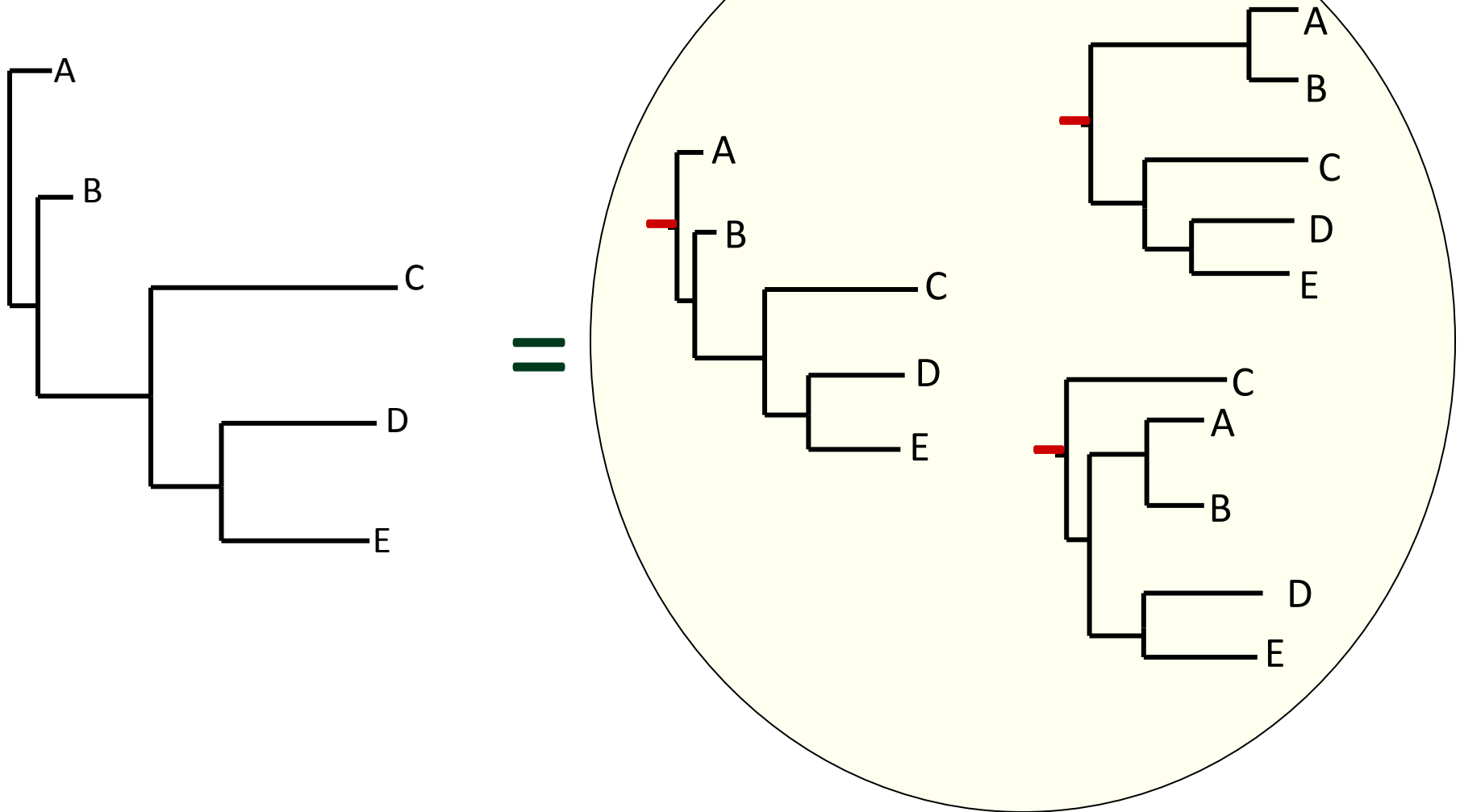
- Узел (node) — точка разделения предковой последовательности (вида, популяции) на две независимо эволюционирующие. Соответствует внутренней вершине графа, изображающего эволюцию.
- Лист — реальный (современный) объект; внешняя вершина графа.
OTU : Operational Taxonomic Unit.
- Ветвь (branch) — связь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



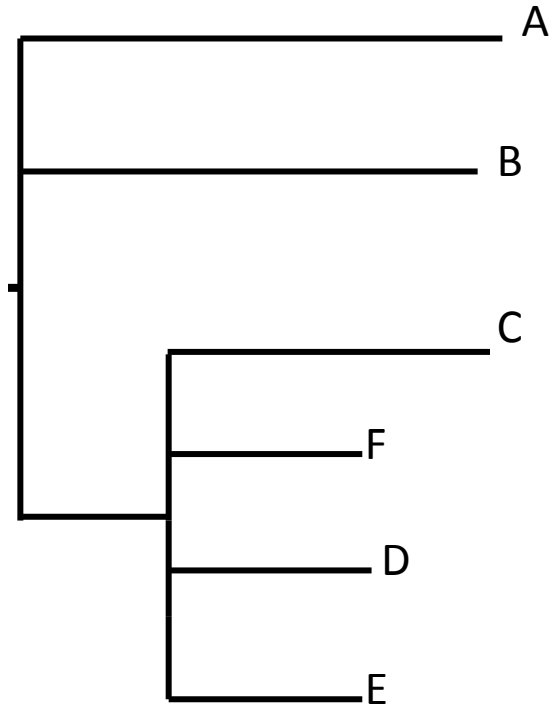
Топология дерева



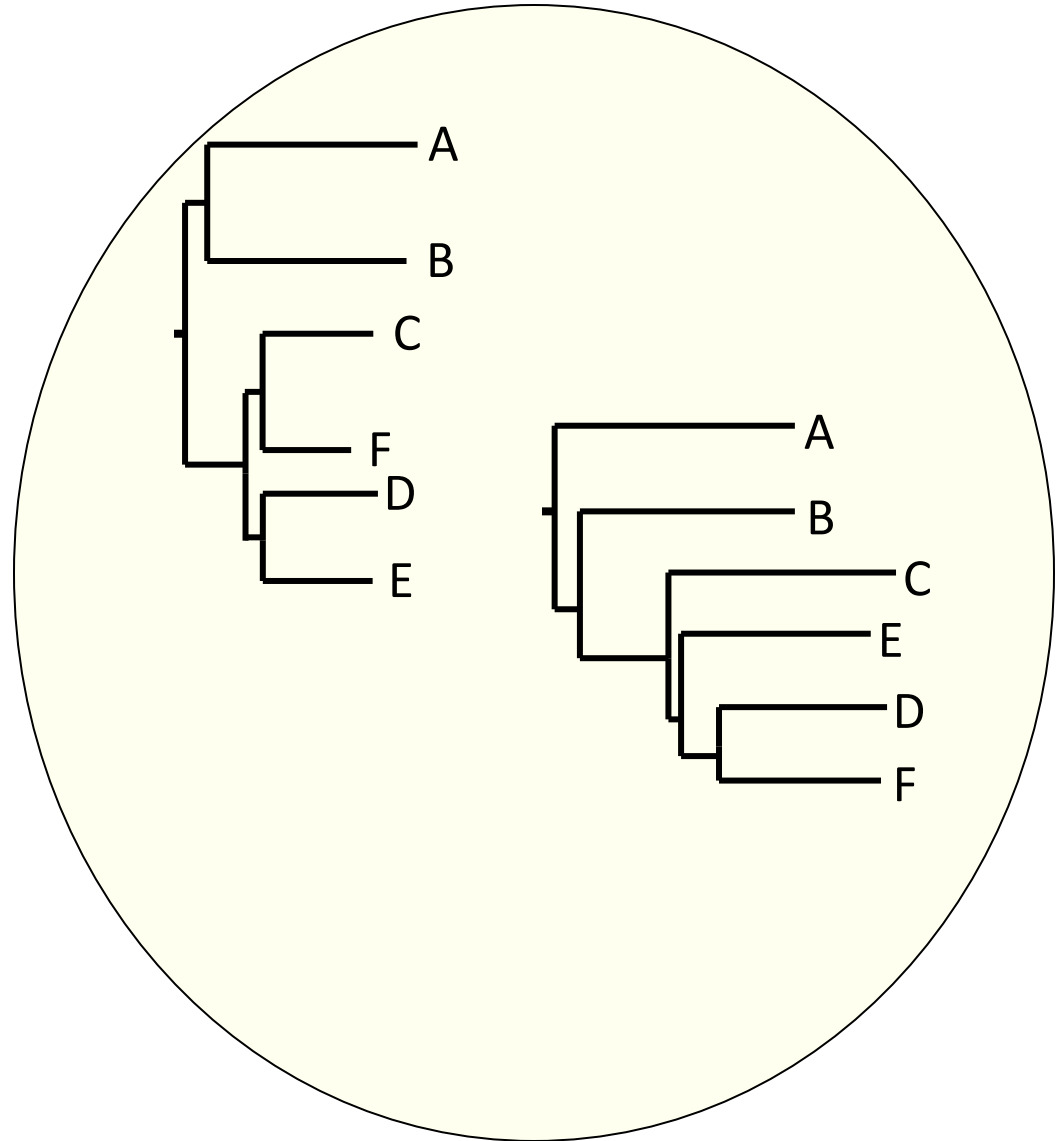
Неукоренённое дерево следует понимать как множество возможных укоренений



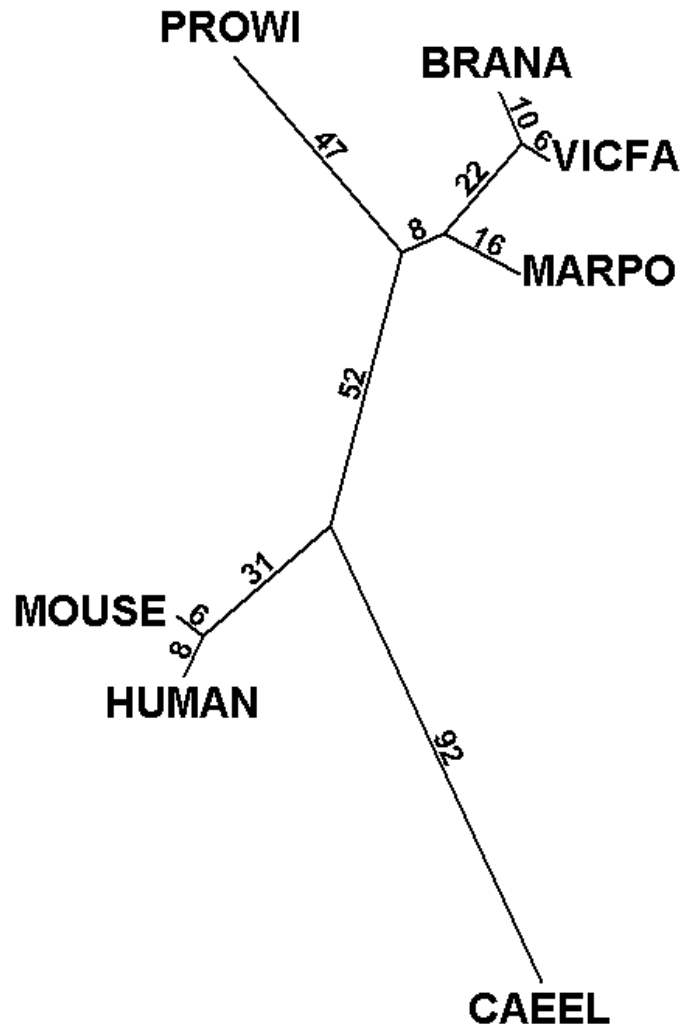
Небинарное дерево следует понимать как множество возможных «разрешений»



=



Расстояния по дереву между листьями



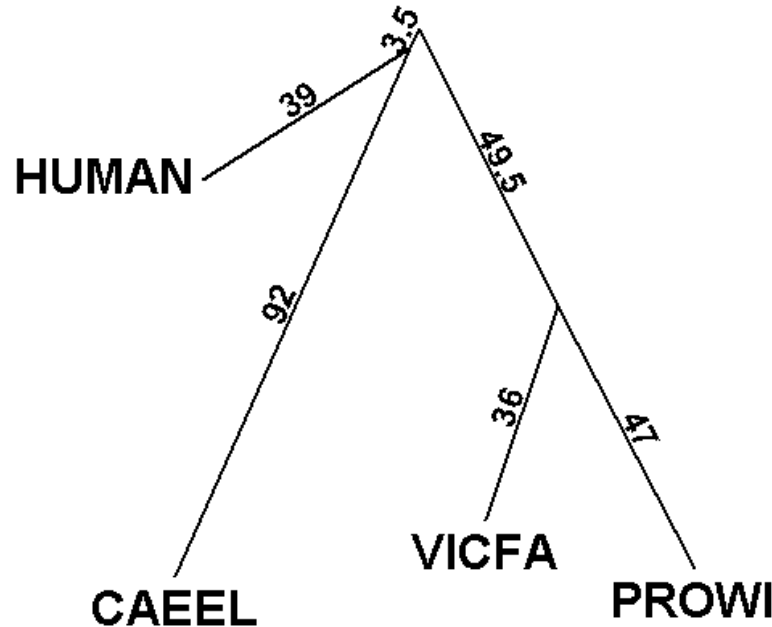
$$D(\text{MOUSE}, \text{CAEEL}) = 6 + 31 + 92 = 129$$

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.»

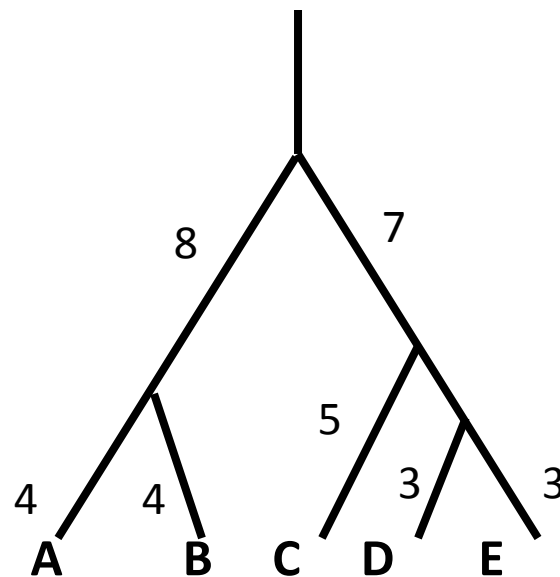
Joseph Felsenstein, <http://evolution.genetics.washington.edu/phylip/newicktree.html>

Ультраметрические деревья

Если на дереве можно найти точку такую, что расстояния от нее до всех листьев одинаковы, то 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*: $d(x, z) \leq \max(d(x, y), d(y, z))$.



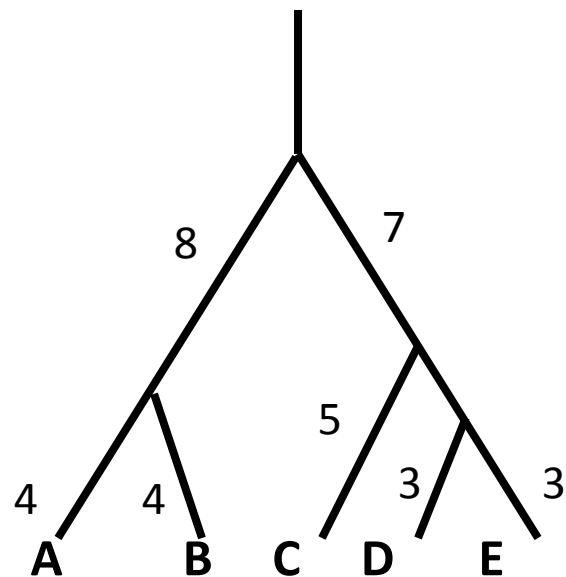
Ультраметрические деревья

Если на дереве можно найти точку такую, что расстояния от нее до всех листьев одинаковы, то 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*: $d(x, z) \leq \max(d(x, y), d(y, z))$.

Задача 19. Доказать равносильность предыдущих утверждений.



Ультраметрические деревья

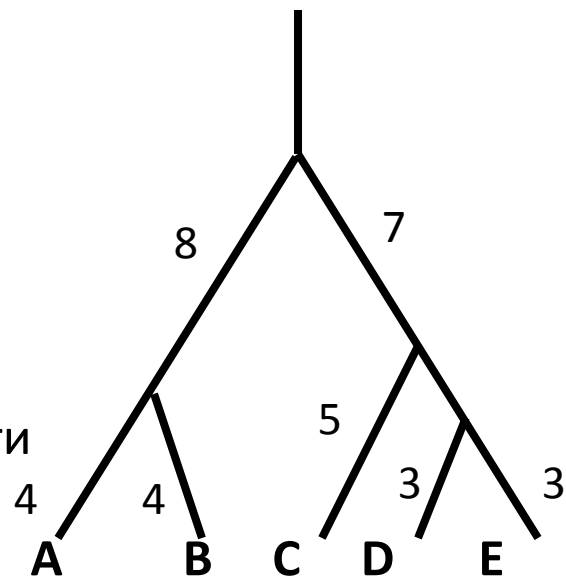
Если на дереве можно найти точку такую, что расстояния от нее до всех листьев одинаковы, то дерево называется “ультраметрическим”.

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

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

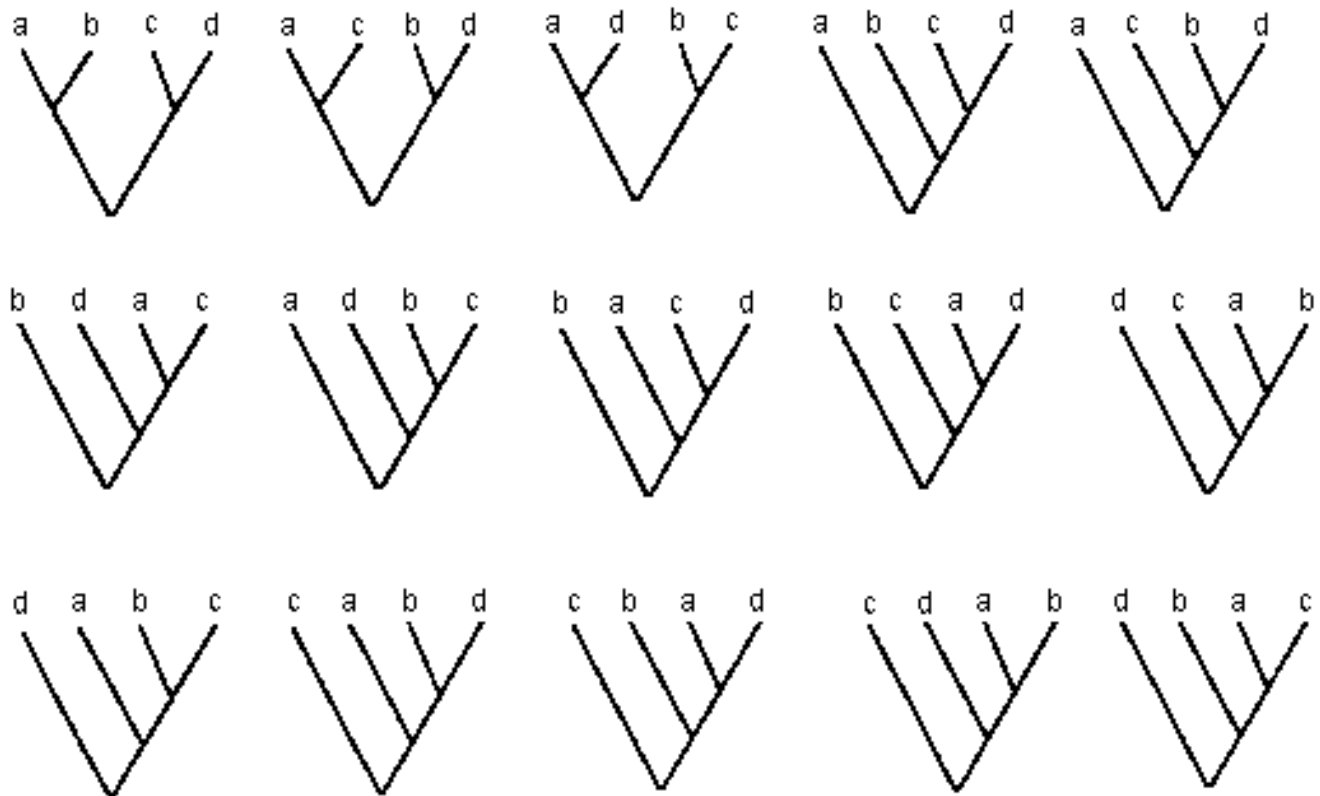
Задача 19. Доказать равносильность предыдущих утверждений.

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

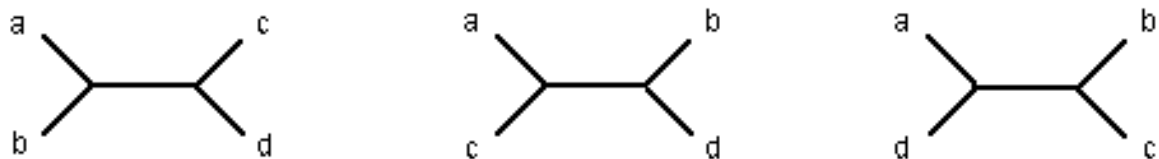


Задача 20. Сколько существует различных укорененных и неукорененных дихотомических деревьев для t листьев?

A



B



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				

	K	L	MN
K		16	16
L			8
MN			

	K	LMN
K		16
LMN		

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:

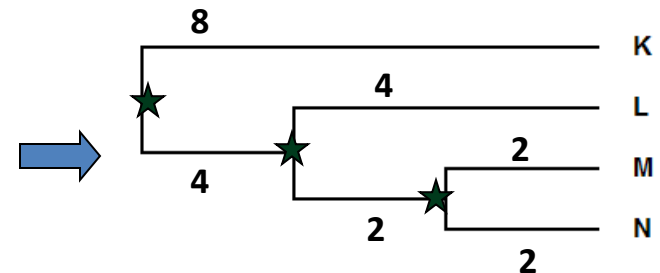
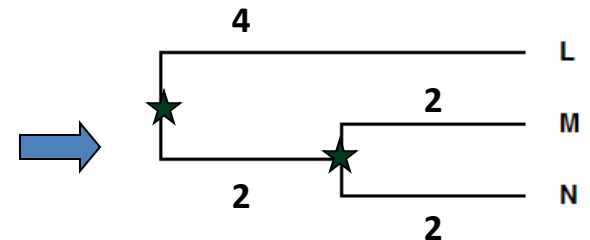
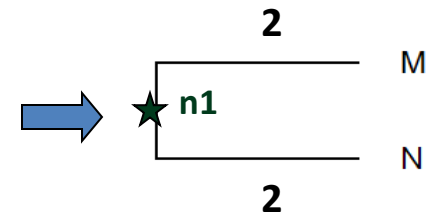
$$d(MN-K) = [d(M-K) + d(N-K)] / 2 = 16$$

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

Шаг 2.

см. шаг 1.

....



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:

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

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

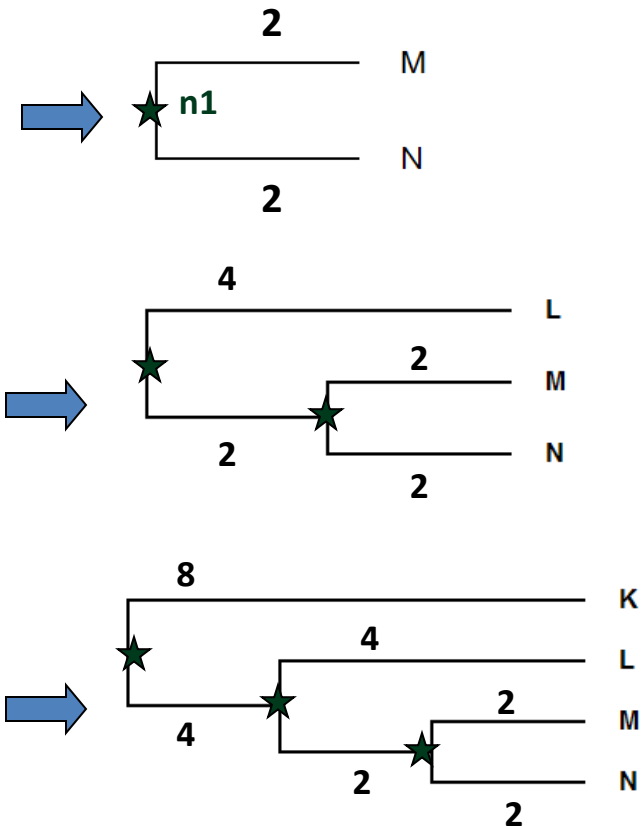
	K	L	MN
K		16	16
L			8
MN			

Шаг 2.

см. шаг 1.

	K	LMN
K		16
LMN		

....



Получаем единственное укорененное ультраметрическое дерево

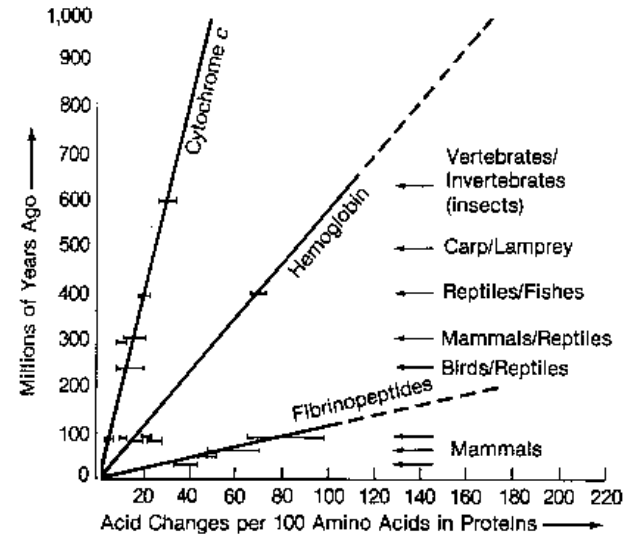
Гипотеза «молекулярных часов» (E.Zuckerkandl, L.Pauling, 1962)

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.

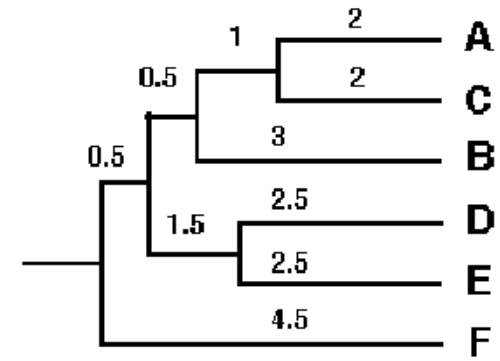
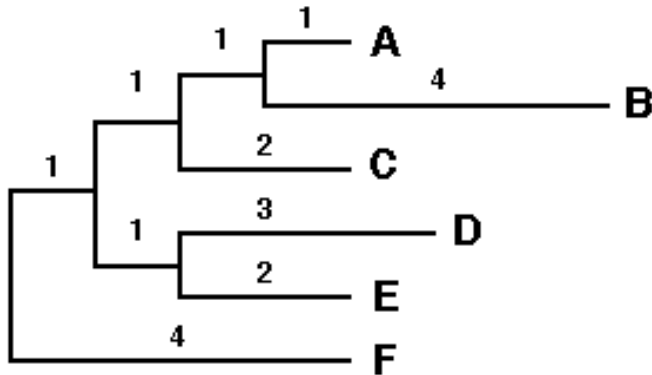
Отклонения от ультраметричности можно считать случайными. Эволюция реконструируется в виде **ультраметрического** дерева.

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



	A	B	C	D	E
B	5				
C	4	7			
D	7	10	7		
E	6	9	6	5	
F	8	11	8	9	8

«Реальное» дерево

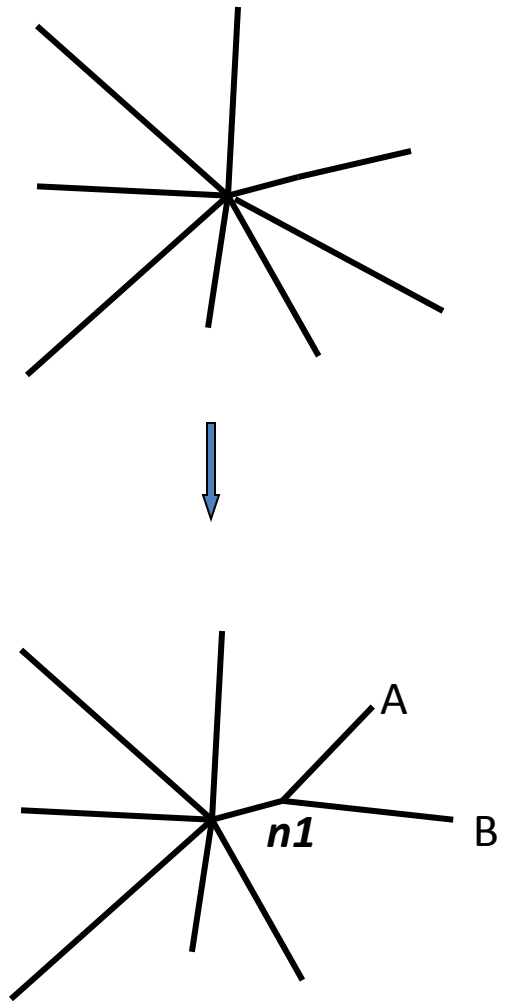
UPGMA

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

Neighbour joining, NJ (метод ближайших соседей)

1. Рисуем «звездное» дерево с ветвями условной длины, например, равными среднему расстоянию листа до всех остальных.
2. Рассмотрим все $m(m-1)/2$ пар листьев и соединим ту пару, в которой листья близки друг к другу, но далеки ото всех остальных.
A и B — такая пара последовательностей, для которых минимальна величина
 $\rho(A, B) - M(A) - M(B)$,
где ρ — расстояние из матрицы, а M — среднее расстояние от A или B до всех $m-2$ остальных последовательностей.
3. Объединим A и B в кластер AB и строим первый узел, расстояние до которого зависит от среднего расстояния до других вершин :
 $d(n1, A) = 0.5 * (\rho(A, B) + M(A) - M(B))$,
 $d(n1, B) = 0.5 * (\rho(A, B) + M(B) - M(A))$,
 $d(n1, X) = 0.5 * (\rho(B, X) + \rho(A, X) - \rho(A, B))$
4. Оставляем кластер AB и удаляем узлы A и B, после чего дерево стало меньше на 1 узел.
5. Повторяем процедуру с п. 2 до тех пор, пока не останется 3 вершины

Получаем единственное неукорененное дерево



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

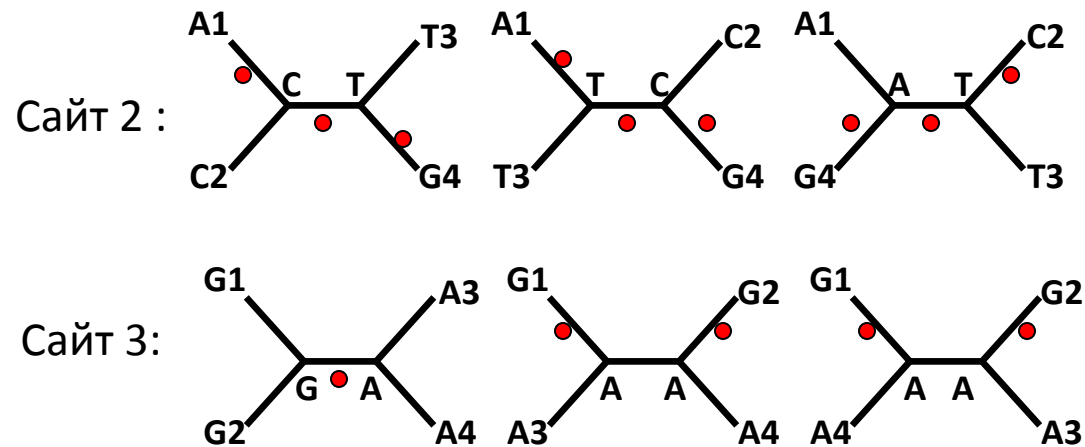
Лучшее дерево — дерево, в котором различия в данных объясняются минимальным числом элементарных эволюционных событий.

4 листа



3 возможных топологии:

Сайты:	1	2	3	4	5
Seq1	а	а	г	а	г
Seq2	а	с	г	г	с
Seq3	а	т	а	г	а
Seq4	а	г	а	г	а



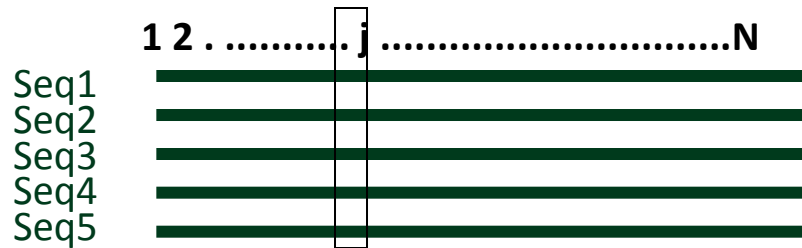
1. Найдем все информативные сайты.

Информативный сайт — колонка выравнивания, символы в которой позволяют выбрать одну из возможных топологий, в данном примере сайт 3 — информативный, а сайт 2 — нет.

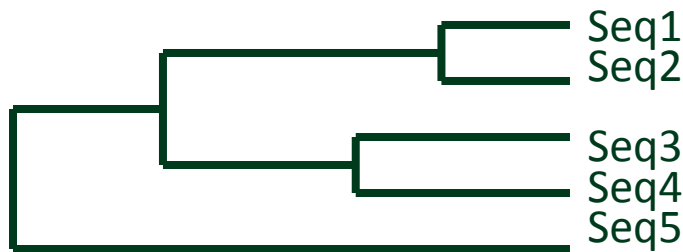
2. Для каждого из возможных деревьев определим минимальное число замен в каждом информативном сайте и сложим эти числа.

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,
 T – 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 : $L = L(1) \times L(2) \dots \times L(N)$
Выбираем дерево, соответствующее максимальной вероятности или наиболее правдоподобное. Числа очень маленькие, поэтому

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