

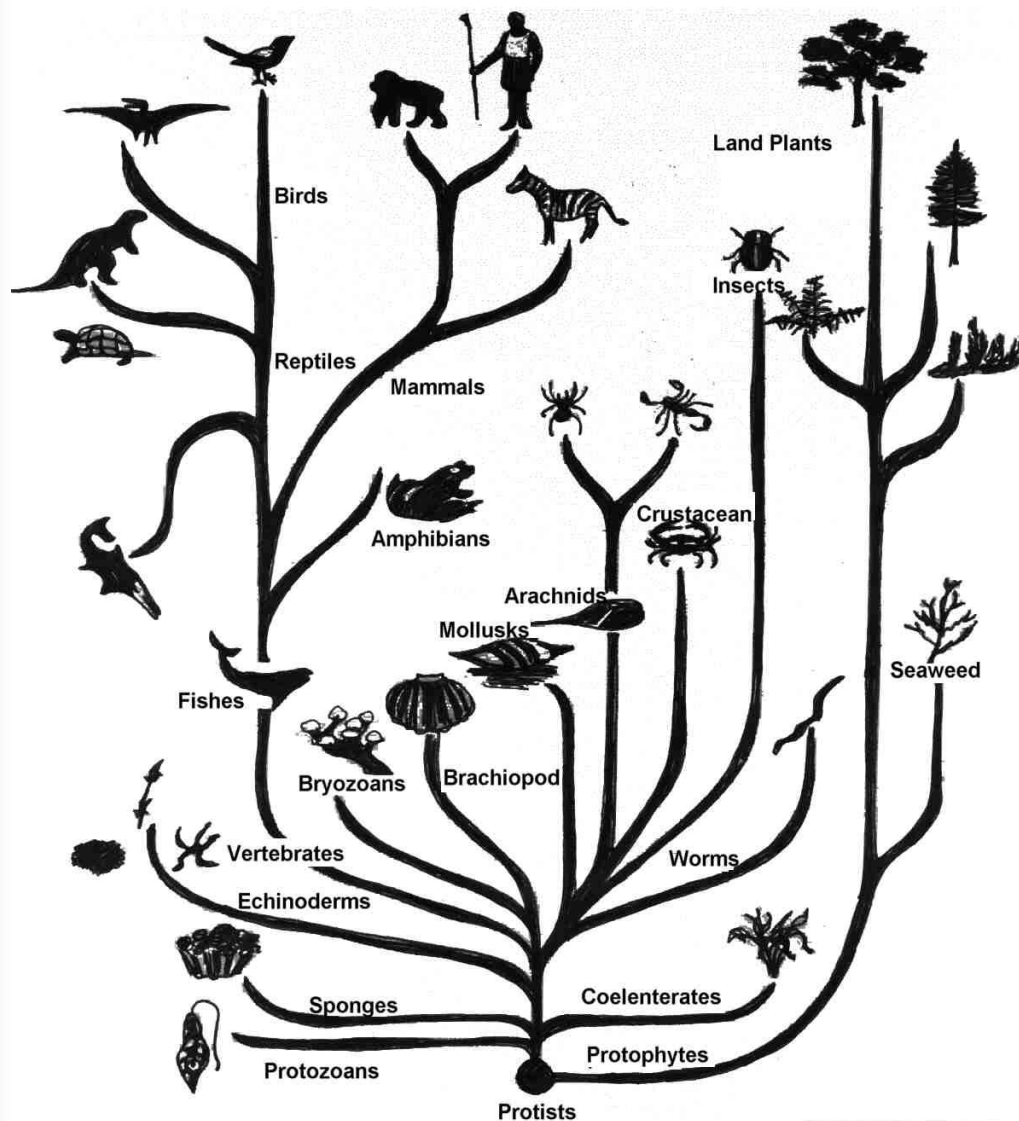


РЕКОНСТРУКЦИЯ ФИЛОГЕНИИ

С.А. Спирин

3 ноября 2016 г.

Древо жизни



В наше время
выяснять детали
происхождения
видов помогают
последовательности
ДНК и белков.

Эволюция белков

Мутации возникают случайно.

Конкретная мутация может быть:

- летальной;
- вредной;
- слабовредной;
- нейтральной;
- полезной.

Мутация порождает **полиморфизм** данного белка в популяции.

Доля каждого варианта подвержена случайным изменением (модель: «случайное блуждание с поглощением»).

За исторически короткое время один из вариантов (старый или новый) исчезает. Во втором случае говорят, что мутация **закрепилась**.

Эволюция видов и эволюция белков

Когда виды разделяются, то разделяются пути эволюции всех их белков...

В результате большинству белков одного вида соответствует ортолог в другом виде.

Но:

- бывают дубликации белков без разделения видов:

два родственных белка существуют в одном геноме и эволюционируют (почти) независимо – такие белки называются паралогами.

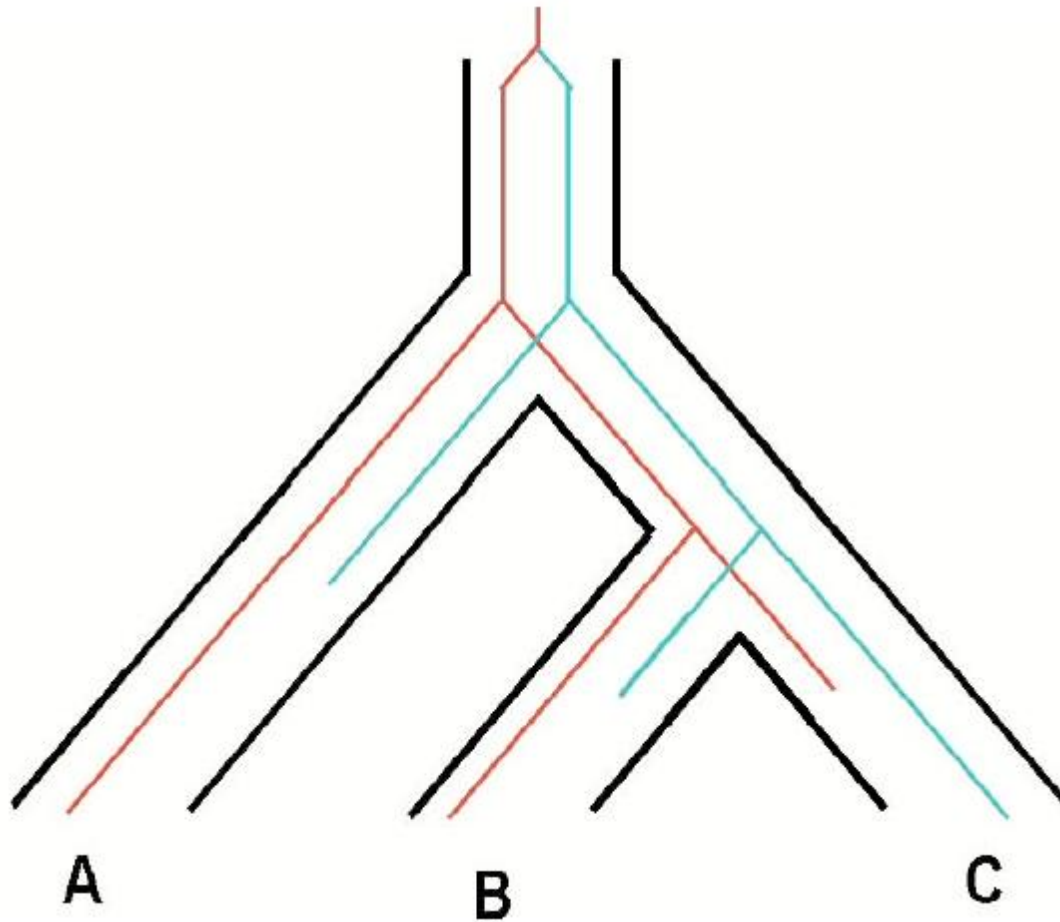
- бывают потери генов.

Если в двух видах потерялись по одному белку из пары паралогов, то получается, что общий предок белков, которые выглядят как ортологи, «жил» существенно раньше, чем общий предок видов.

- бывает, что два белка объединяются в один многодоменный, и наоборот.

Поэтому правильнее говорить об эволюции белковых доменов.

Дерево видов и дерево белков



Исходный материал для реконструкции филогении

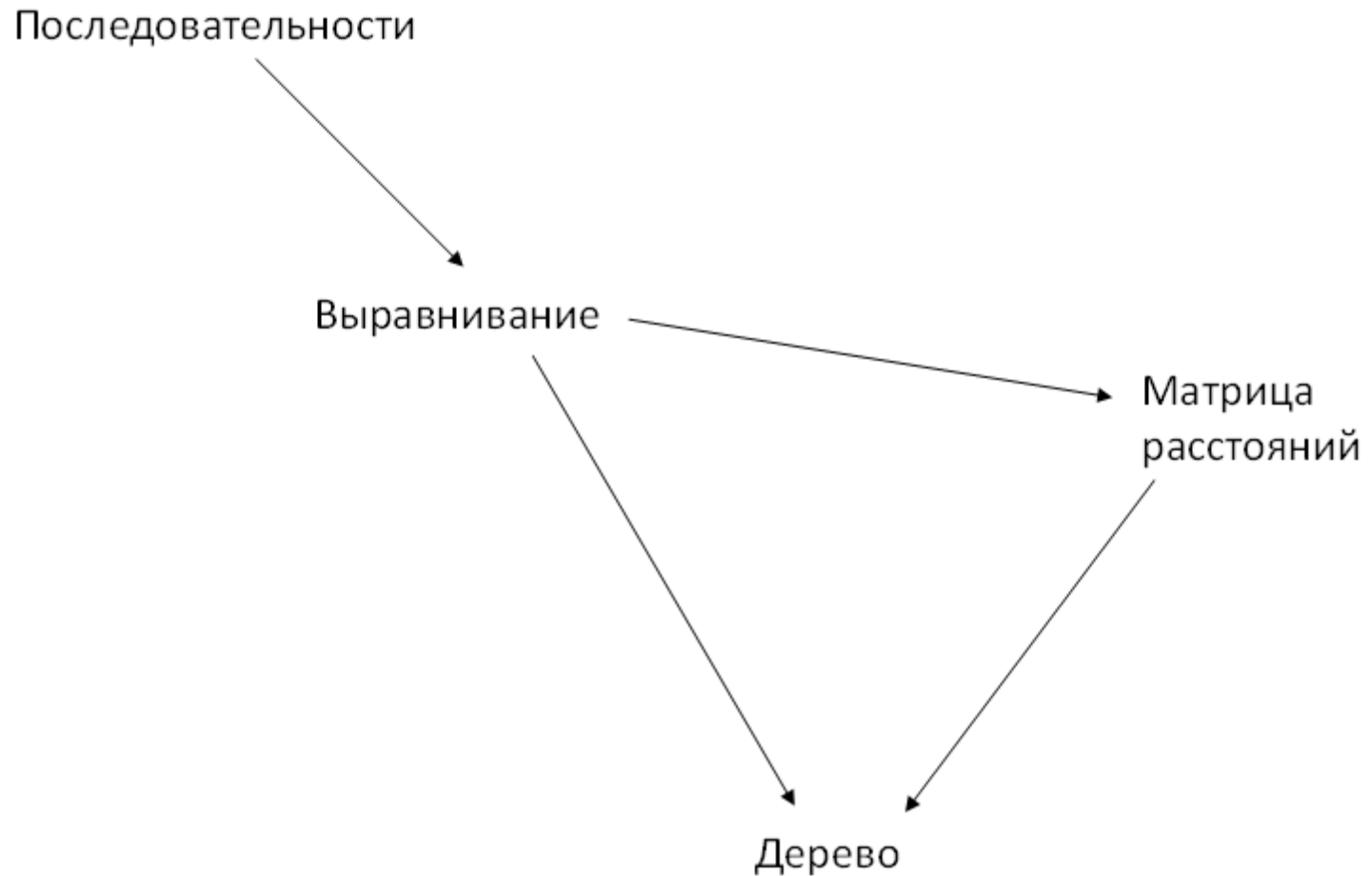
Множественное выравнивание биологических последовательностей

CYB5_CHICK	MVGSSEAGGEAWRGYYRLEEVQKHNNNSQSTWII VHHRIYDITKFLDEHPGGEEVLREQA
CYB5_HUMAN	---MAEQSDEAV-- KYITLEEIQKHNSKSTWLI LHNKVYDLTKFLEENPGGEEVLREQA
CYB5_HORSE	---MAEQSDKAV-- KYITLEEIKKHNSKSTWLI LHNKVYDLTKFLEDHPGGEEVLREQA
CYB5_MUSDO	-----MSSEDV-- KYFTRAEVAKNNTKDKNWF I IHNNVYDVTAFLNEHPGGEEVLIEQA
CYB5_DROME	-----MSSEET-- KTFTRAEVAKHNTNKTWLL IHNNIYDVTAFLNEHPGGEEVLIEQA

Методы реконструкции филогенетических деревьев:

- символно ориентированные:
 - * максимальной экономии (maximal parsimony)
 - * наибольшего правдоподобия
- использующие матрицу расстояний
 - * кластерные (UPGMA и др.)
 - * Neighbor-joining
 - * минимальной эволюции
 - * наименьших квадратов
 - * Фитча - Марголиаша
 - * ...

Схема реконструкции филогении по последовательностям



Классификация методов

Название	Переборный / эвристический	Предполагает молекулярные часы	Символьно ориентированный
UPGMA	Эвристический	Да	Нет
Neighbour-joining	Эвристический	Нет	Нет
Наименьших квадратов	Переборный	Может	Нет
Фитча - Марголиаша	Переборный	Может	Нет
Минимальной эволюции	Переборный	Может	Нет
Максимальной экономии	Переборный	Нет	Да
Наибольшего правдоподобия	Переборный	Может	Да

Деревья

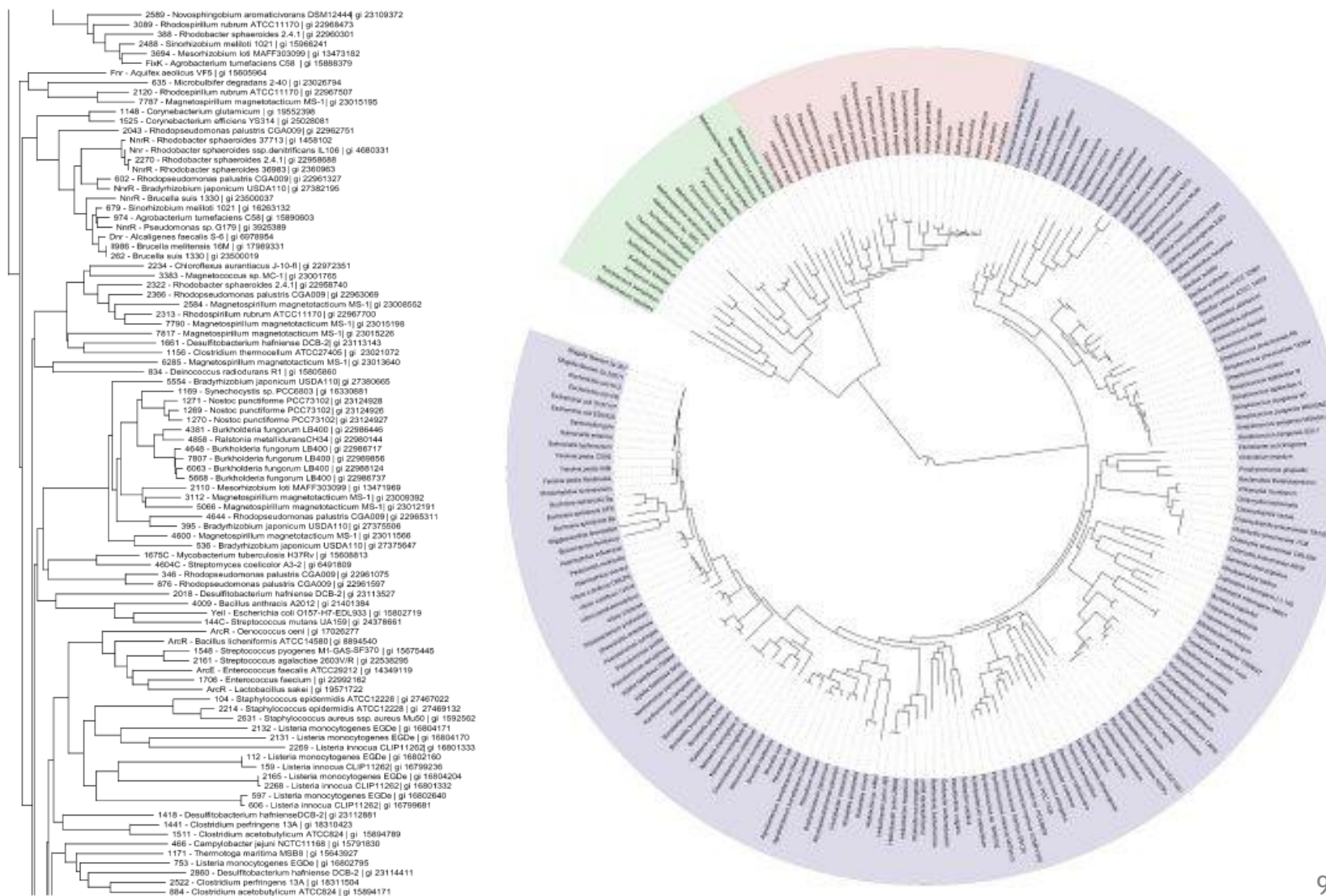


Fig. 1 (Continued).

The phylogeny of *The Canterbury Tales*

Geoffrey Chaucer's *The Canterbury Tales* survives in about 80 different manuscript versions¹. We have used the techniques of evolutionary biology to produce what is, in effect, a phylogenetic tree showing the relationships between 58 extant fifteenth-century manuscripts of "The Wife of Bath's Prologue" from *The Canterbury Tales*. We found that many of the manuscripts fall into separate groups sharing distinct ancestors.

Manuscripts such as these were created by copying, directly or indirectly, from the original material (written, in the case of *The Canterbury Tales*, in the late fourteenth century). In the process of copying, the scribes made (deliberately or otherwise) changes, which were themselves copied. Textual scholars have developed a system for reconstructing the relationships between textual traditions by analysing the distribution of these shared changes, and have constructed family trees (stemmata) on the basis of the results, with the ultimate aim of establishing precisely what the author actually wrote.

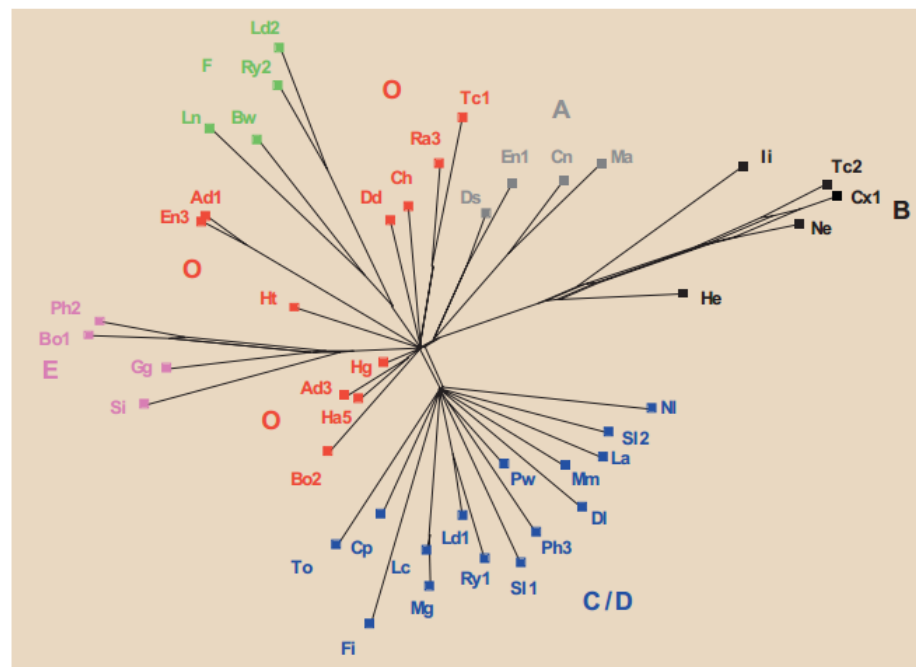


Figure 1 SplitsTree analysis of 44 manuscripts of "The Wife of Bath's Prologue" from Chaucer's *The Canterbury Tales*⁴. The two- or three-character codes indicate individual manuscripts, whereas the large capitals indicate groups of manuscripts, which are coloured the same.

the 58 manuscripts. Very similar results were given by PAUP (not shown). Several manuscripts form groups (A, B, C/D, E and F), each descended from a single and distinct common ancestor. The remaining 14 manuscripts were removed from the analysis shown in Fig. 1, as they were likely to

notes of passages to be deleted or added, and alternative drafts of sections. In time, this may lead editors to produce a radically different text of *The Canterbury Tales*. These results also demonstrate the power of applying phylogenetic techniques, and particularly split decomposition, to the study

Терминология

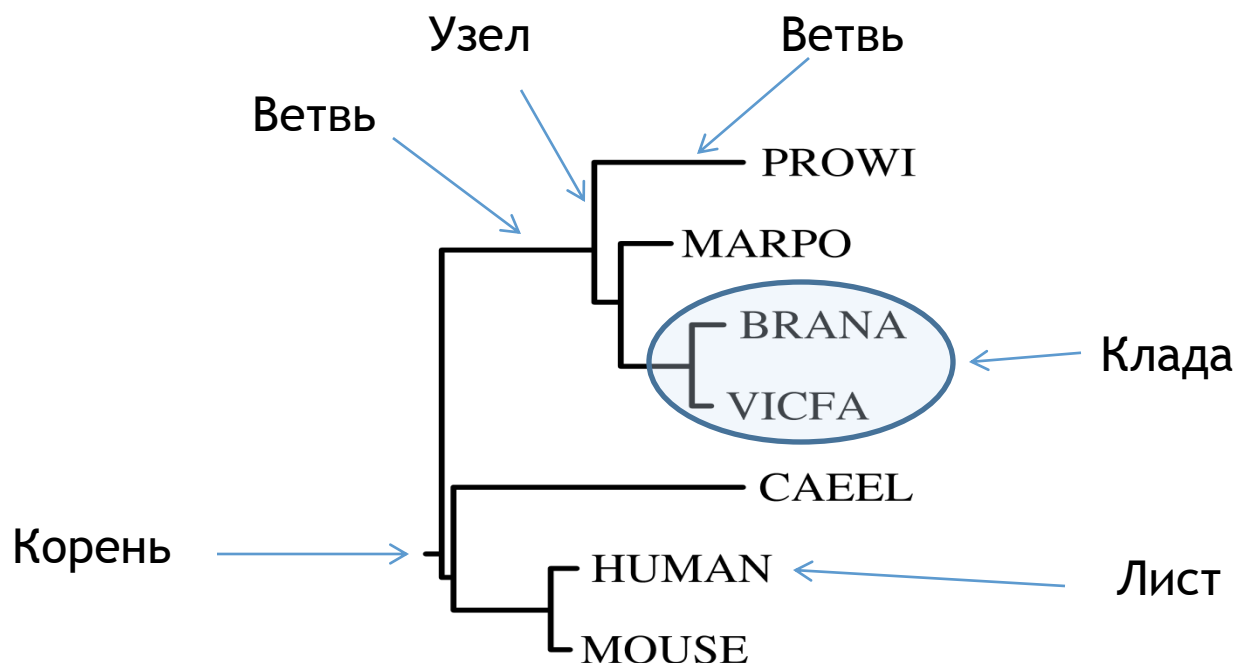
Узел (node) – точка разделения предковой последовательности, внутренняя вершина графа.

Лист (leaf) – реальный (современный) объект; внешняя вершина графа.

Ветвь (branch) — связь между узлами или между узлом и листом; ребро графа.

Корень (root) — общий предок всех рассматриваемых объектов.

да (clade) — группа всех потомков некоторого ранее существовавшего объекта.



Длины ветвей

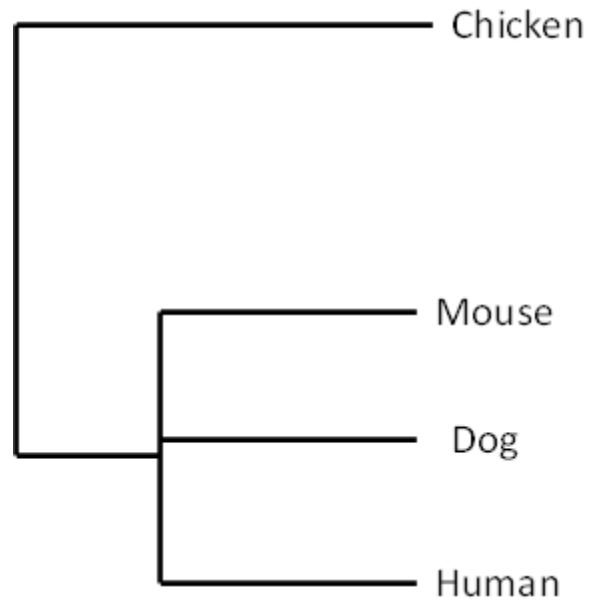
Каждая точка дерева – некоторая последовательность, существовавшая в некоторый момент времени (в прошлом, если эта точка – не лист).

Длины ветвей могут иметь двоякий смысл:

- 1) интервал времени между моментами существования двух последовательностей;
- 2) число мутаций, случившихся на пути от одной последовательности до другой.

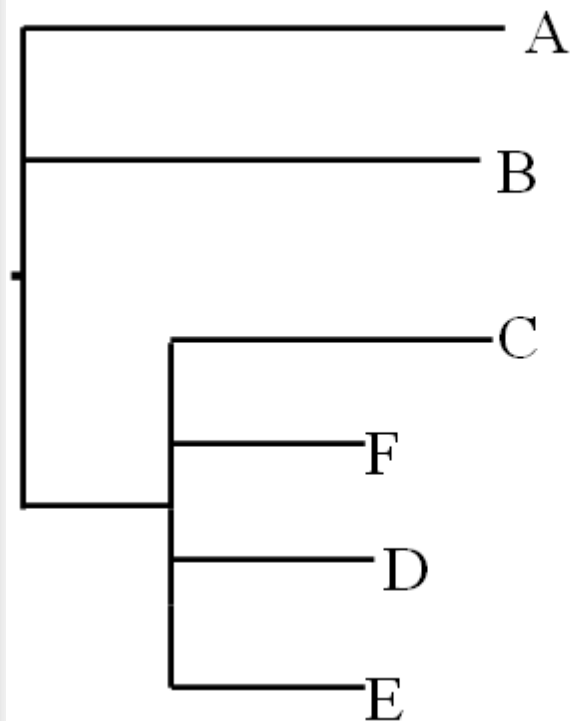
Дерево без длин ветвей часто называют кладограммой (cladogram), дерево, снабженное длинами ветвей – филограммой (phylogram) или фенограммой (phenogram).

Небинарное дерево

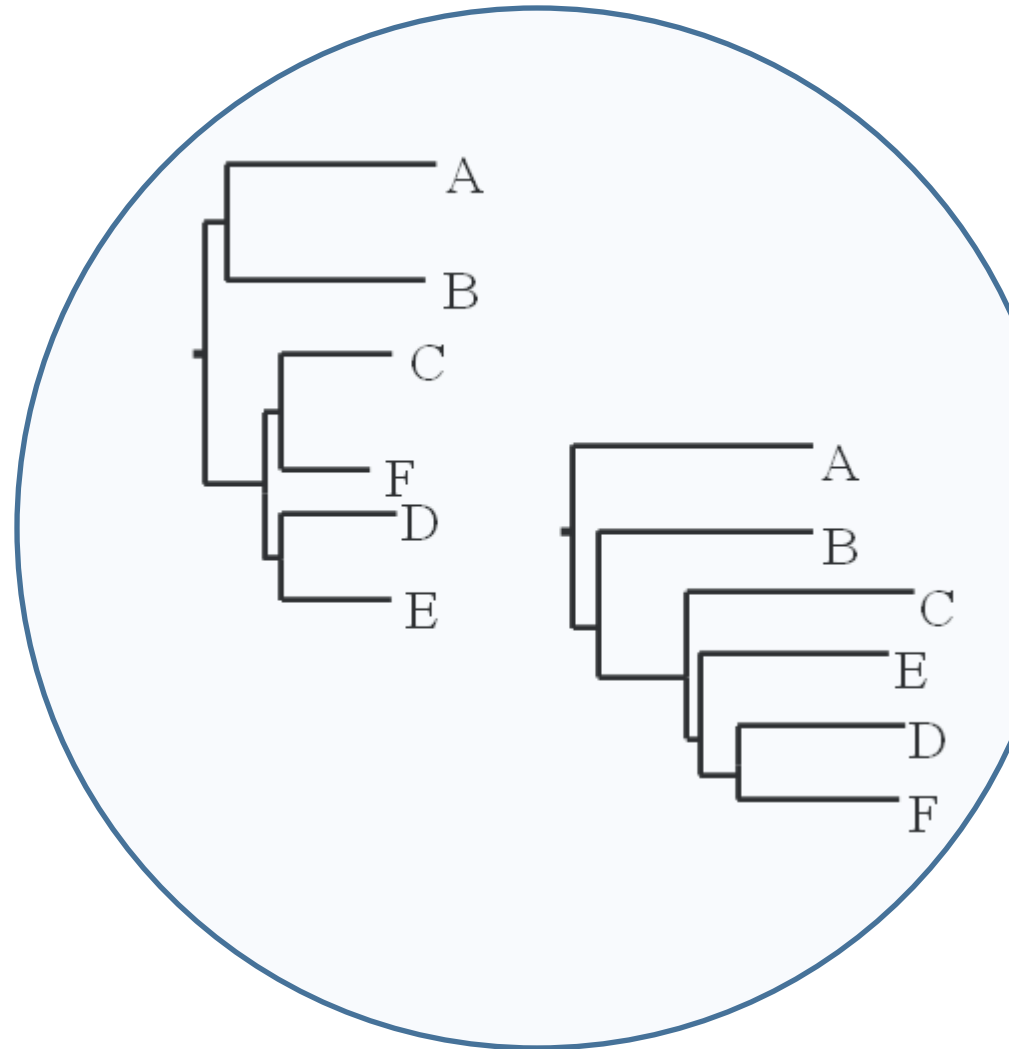


из узла выходят не два, а три или больше ветвей

Разрешения небинарного дерева

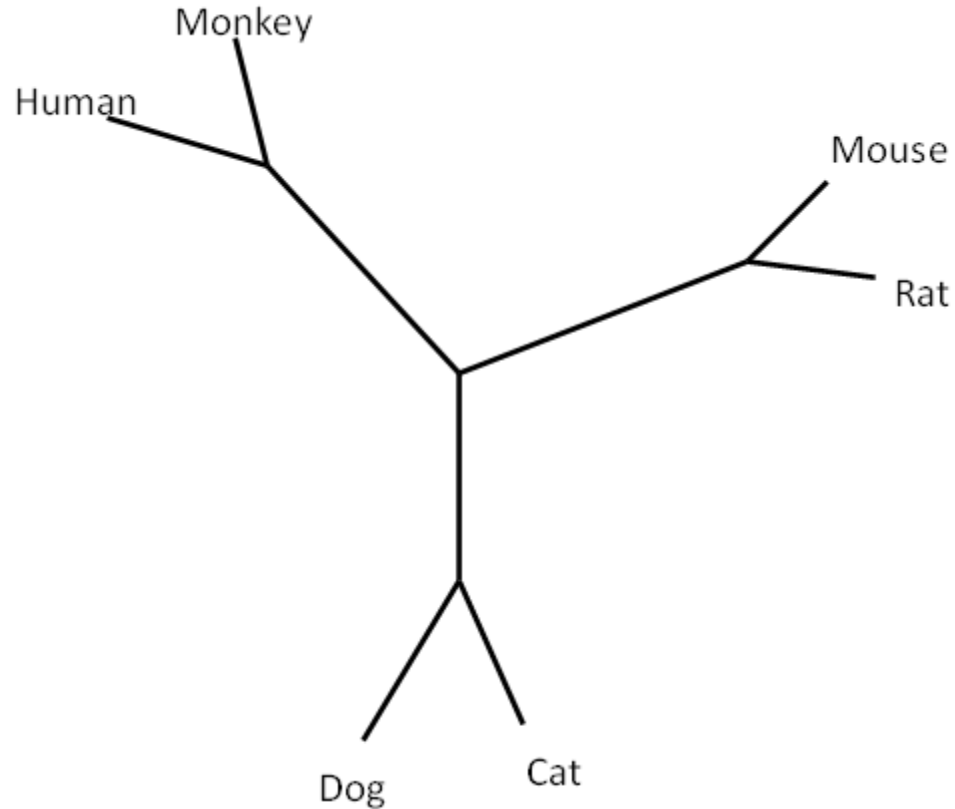


=



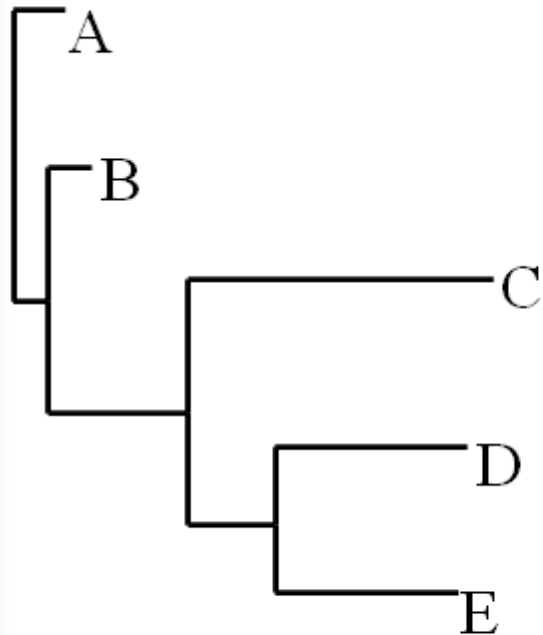
... и ещё 43 варианта

Неукоренённое дерево

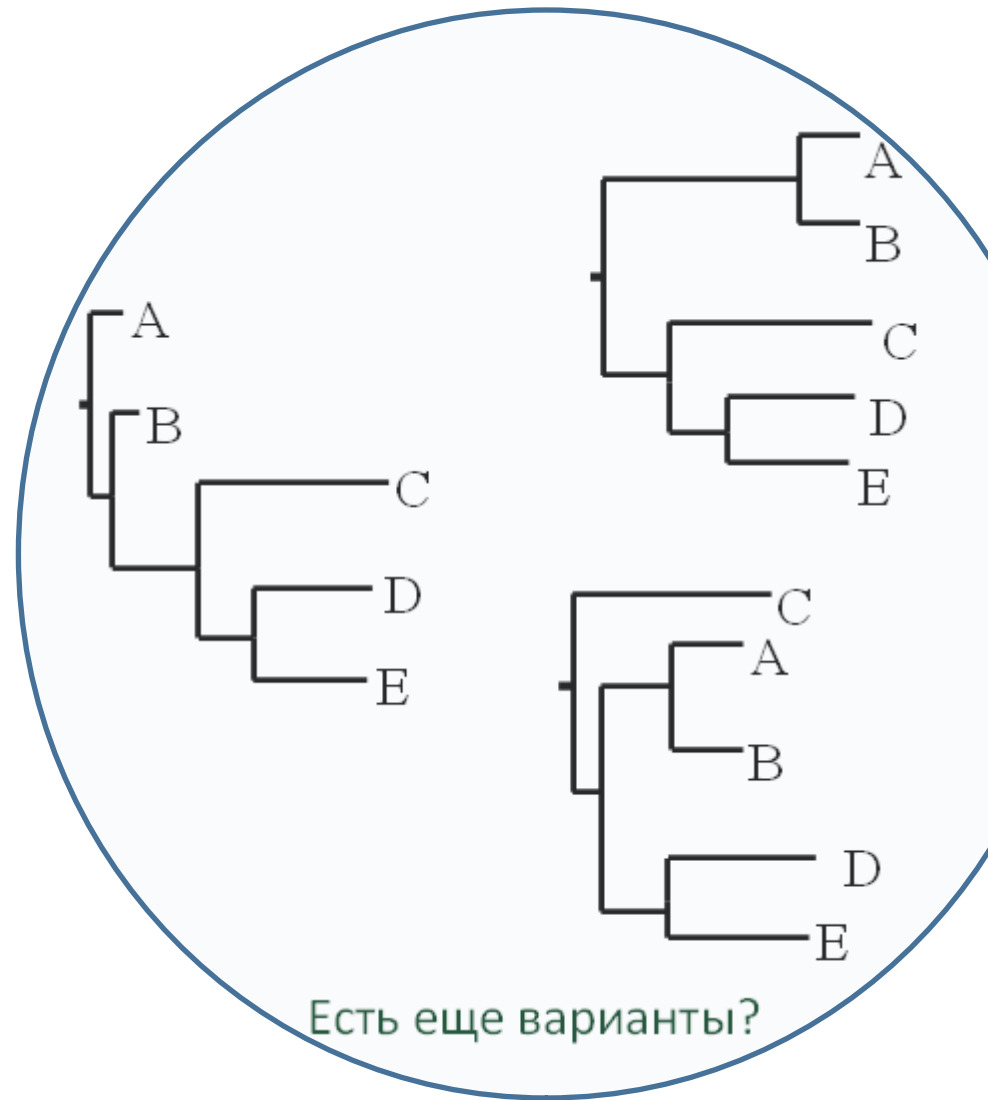


... кстати: в «радиальном» стиле узлы изображаются точками

Укоренения

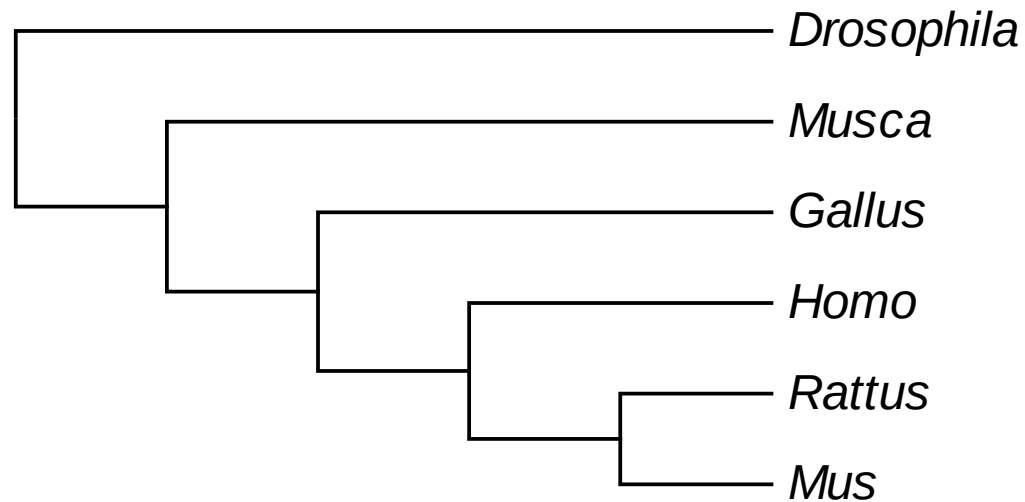


=

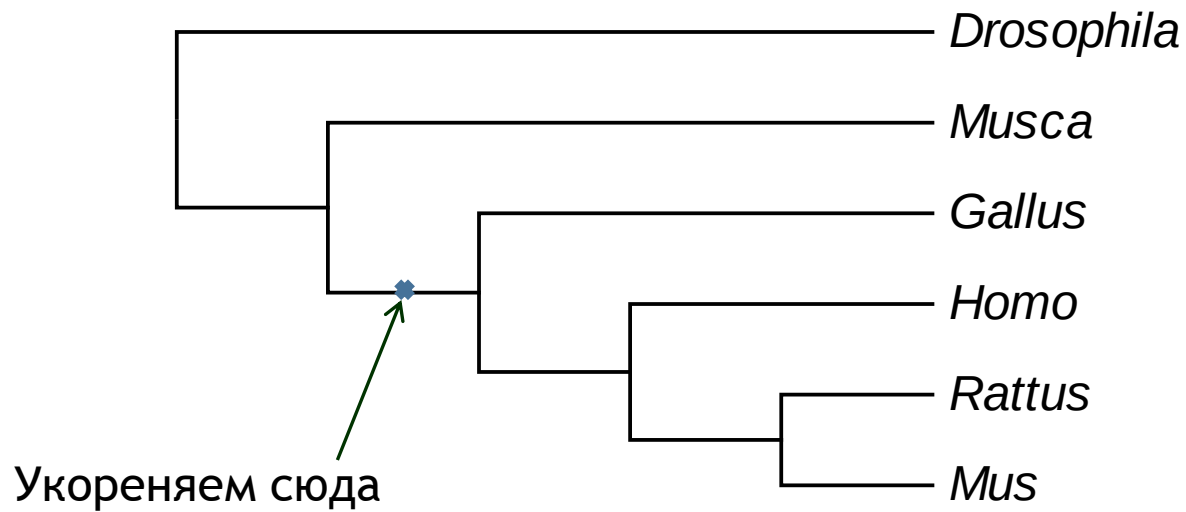


Неукоренённое дерево надо понимать как множество возможных укоренений. Укоренение возможно в любую ветвь.

Это - правильное дерево!



Это - правильное дерево!



Выдача программы пакета PHYLIP

Protein parsimony algorithm, version 3.69

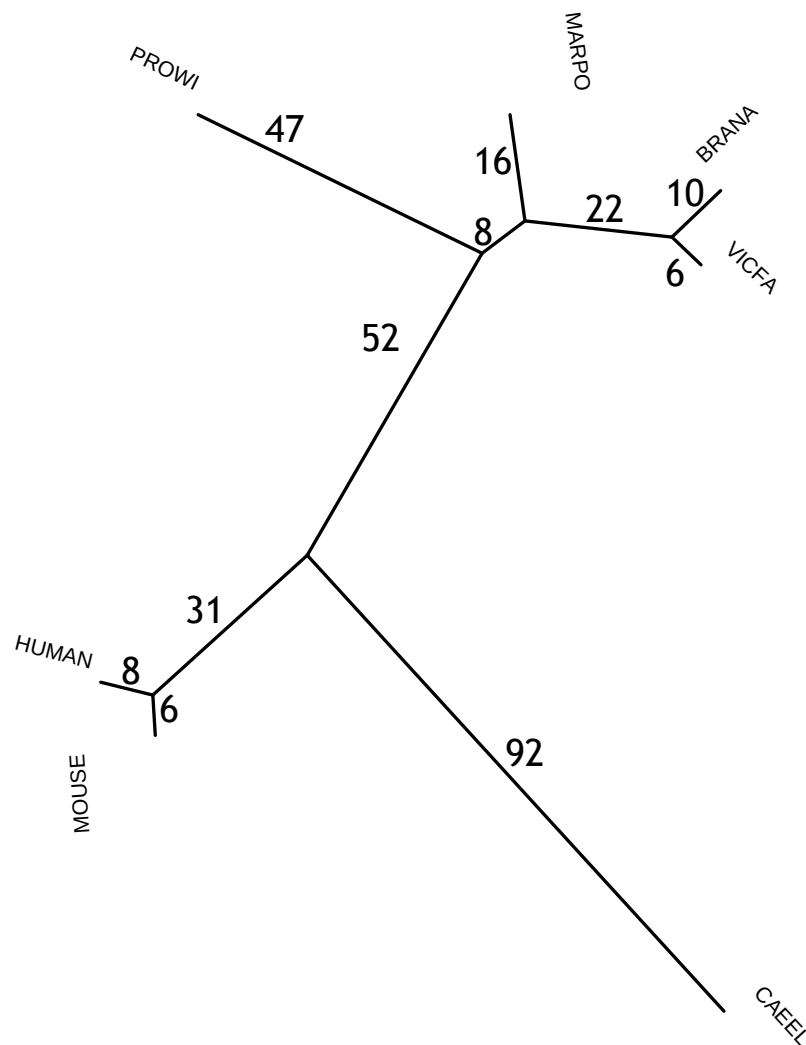
One most parsimonious tree found:

```

              +-----HUMAN
            +-----4
          !      !  +--RAT
          !      +--5
+--3          +--MOUSE
!  !
!  !          +--DROME
1  +-----2
!          +--MUSDO
!
+-----CHICK
```

remember: this is an unrooted tree!

Расстояния по дереву



Расстояние от CAEEL до
PROWI равно $92 + 52 + 47 = 191$

Ультраметрические деревья

Дерево называется ультраметрическим, если на нём есть точка, расстояния от которой до всех листьев одинаково.

Если такая точка есть, то её естественно считать корнем, поэтому можно считать, что ультраметрическое дерево всегда укоренённое (но не наоборот!)

Если все листья представляют **современные** последовательности, а длины ветвей имеют смысл **времени**, то дерево ультраметрическое (потому что в этом случае расстояние от корня до любого листа одинаковое – это время, прошедшее от первого разделения до настоящего времени).

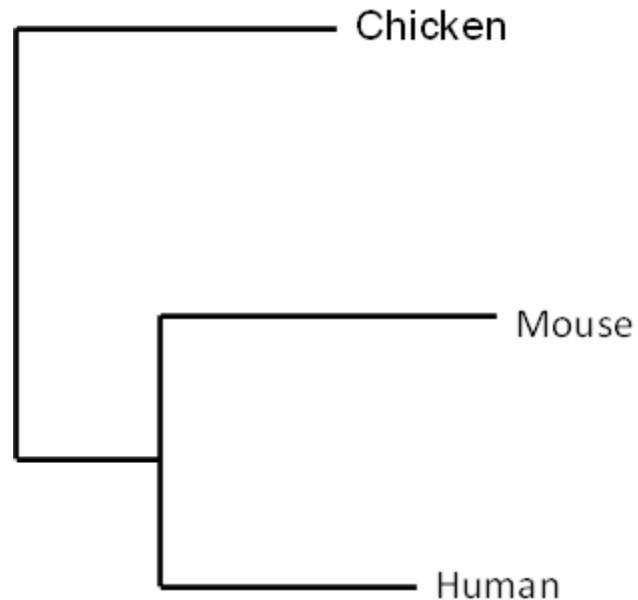
Молекулярные часы

Гипотеза молекулярных часов: за одинаковое время происходит в среднем одинаковое число мутаций

Если гипотеза верна, то можно оценивать **эволюционное время** между современными последовательностями и на основании этих оценок строить **укоренённое ультраметрическое дерево**.

Но гипотеза МЧ часто не выполняется.

Здесь длины ветвей - не время!



Как правило, длины ветвей отображают не время, а число мутаций, закрепившихся на данном отрезке эволюционной истории.

Свойство аддитивности расстояний по дереву

Если длины ветвей понимать как число мутаций, то дерево почти всегда не ультраметрично.

Однако расстояния по дереву между листьями – не произвольный набор чисел. Они удовлетворяют свойству, называемому «аддитивность»: для любых четырёх листьев A, B, C, D из трёх сумм

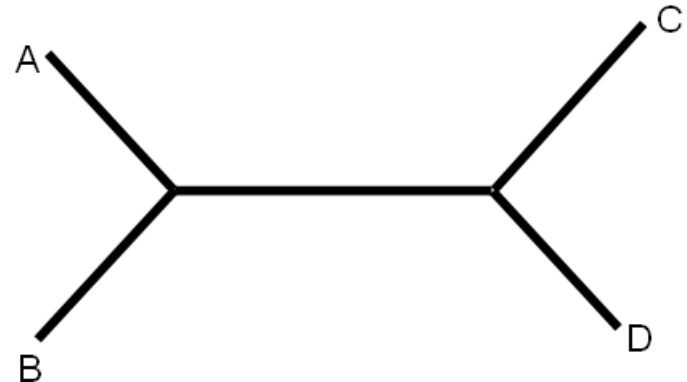
1) $d(A, B) + d(C, D)$

2) $d(A, C) + d(B, D)$

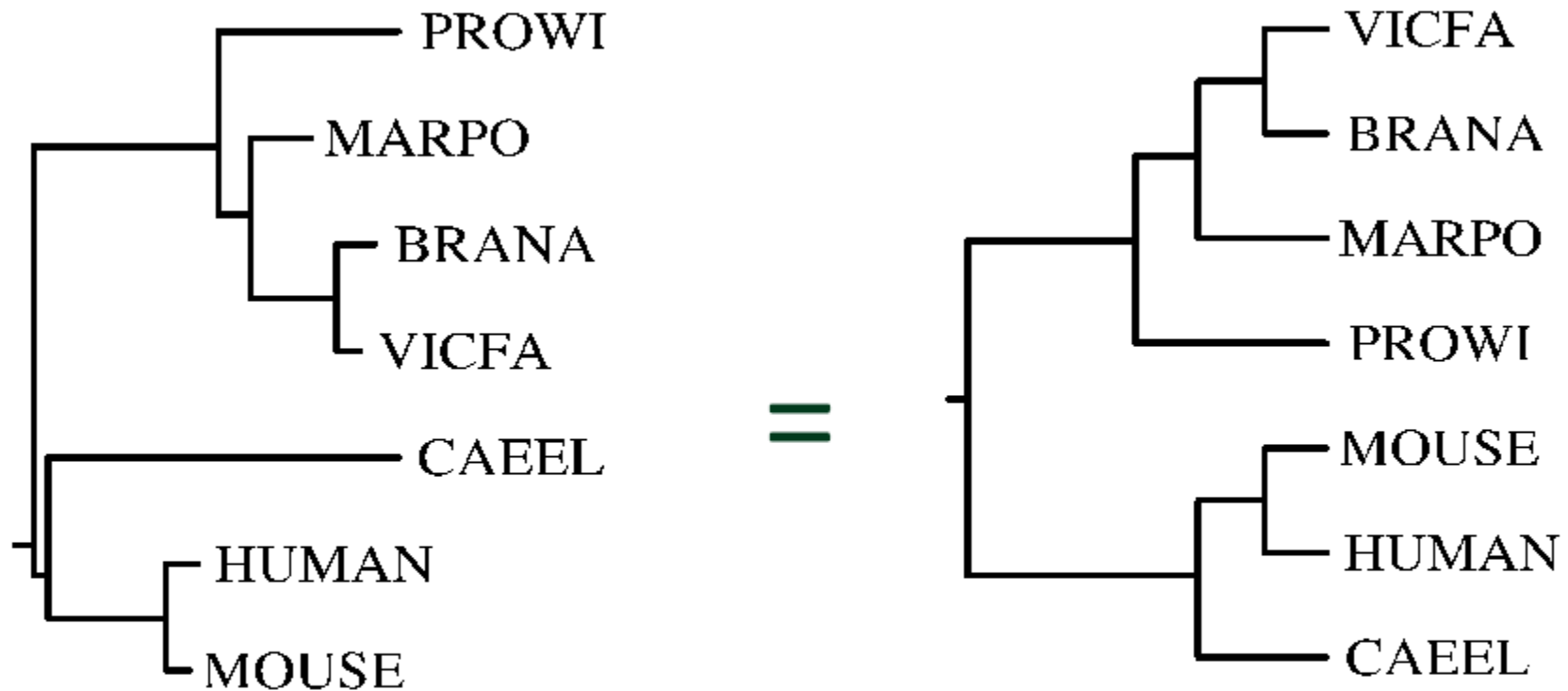
3) $d(A, D) + d(B, C)$

две равны между собой и больше третьей.

d - расстояние по дереву



Топология дерева

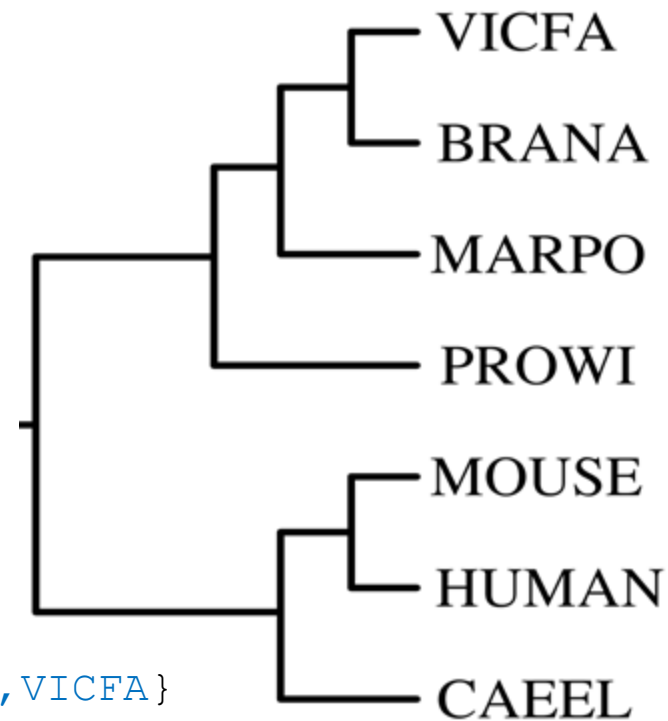


Топология дерева: разбиения

Каждая ветвь разбивает множество листьев на два подмножества.

В каждом дереве есть **тривиальные ветви** (отделяющие один лист от всех остальных), они не зависят от топологии.

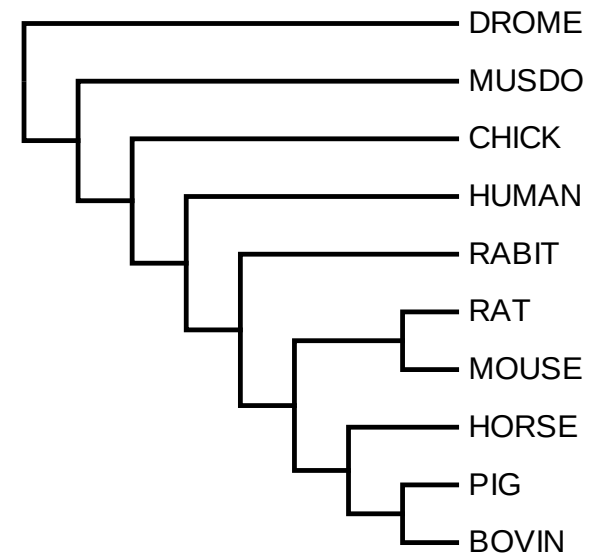
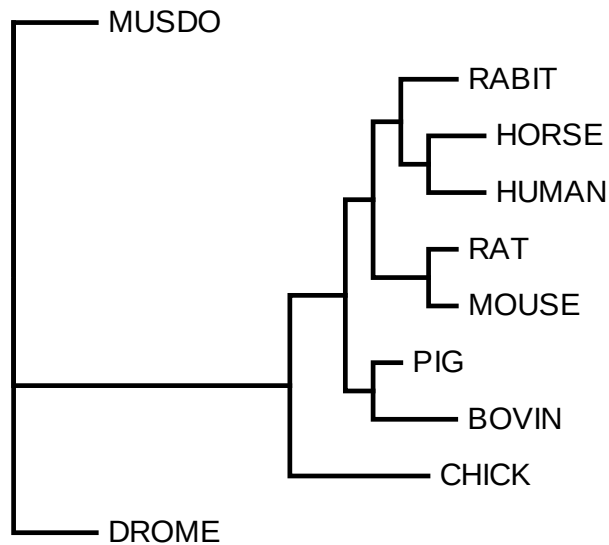
Топологию (неукоренённого) дерева можно однозначно записать набором **нетривиальных разбиений**. Например:



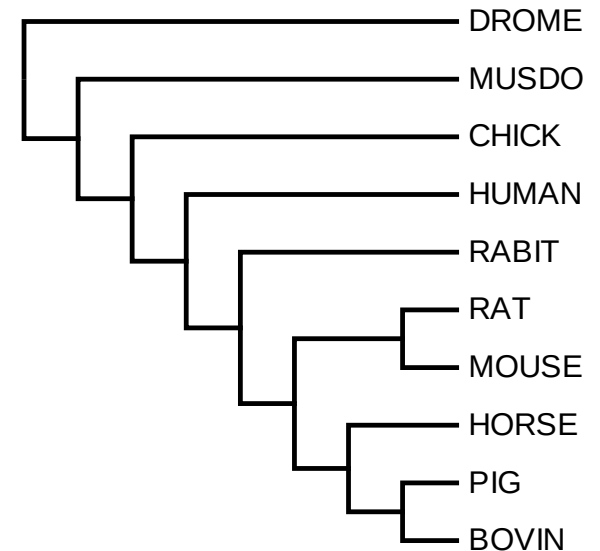
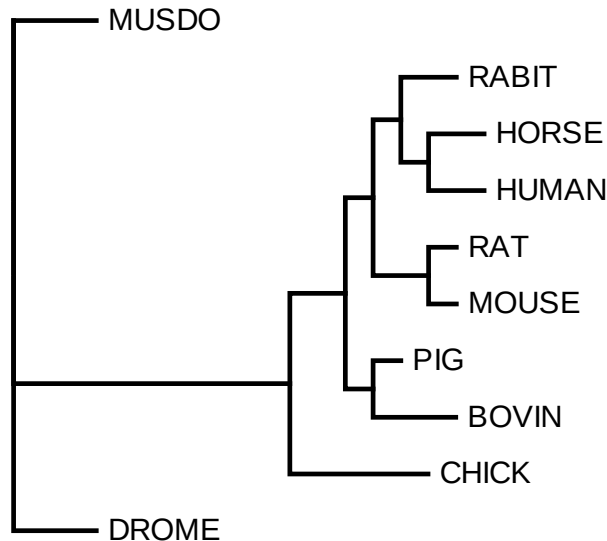
$\{\text{HUMAN, MOUSE}\} \text{ vs } \{\text{CAEEL, PROWI, MARPO, BRANA, VICFA}\}$
 $\{\text{HUMAN, MOUSE, CAEEL}\} \text{ vs } \{\text{PROWI, MARPO, BRANA, VICFA}\}$
 $\{\text{HUMAN, MOUSE, CAEEL, PROWI}\} \text{ vs } \{\text{MARPO, BRANA, VICFA}\}$
 $\{\text{HUMAN, MOUSE, CAEEL, PROWI, MARPO}\} \text{ vs } \{\text{BRANA, VICFA}\}$

HUMAN	MOUSE	CAEEL	VICFA	BRANA	MARPO	PROWI
+	+	-	-	-	-	-
+	+	+	-	-	-	-
+	+	+	-	-	-	+
+	+	+	-	-	+	+

Что общего у двух деревьев?



Что общего у двух деревьев?

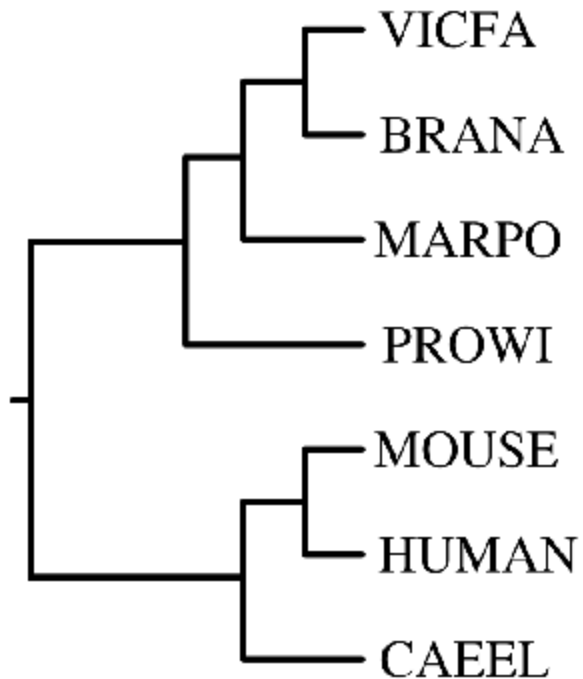


Ответ: четыре ветви:

- 1) отделяющая {DROME, MUSDO} от всех остальных;
- 2) отделяющая {RAT, MOUSE} от всех остальных;
- 3) отделяющая {PIG, BOVIN} от всех остальных;
- 4) отделяющая {DROME, MUSDO, CHICK} от всех остальных.

При этом на каждом дереве есть по три нетривиальных ветви, отсутствующих на другом.

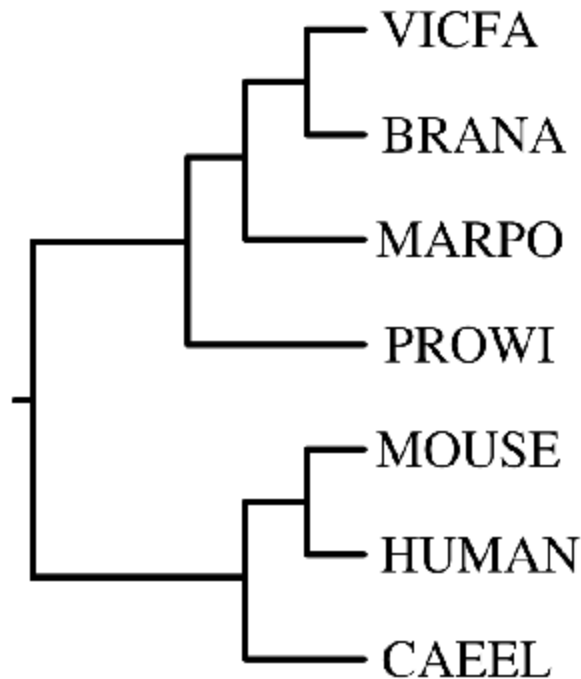
Формат Newick



```
((((VICFA, BRANA), MARPO), PROWI), ((MOUSE, HUMAN), CAEEL));
```

Формат Newick

(с длинами ветвей)



«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>

`((((VICFA:3, BRANA:3):3, MARPO:6):2, PROWI:8):7, ((MOUSE:3, HUMAN:3):3, CAEEL:6):15);`

Классификация методов

Название	Переборный / эвристический	Предполагает молекулярные часы	Символьно ориентированный
UPGMA	Эвристический	Да	Нет
Neighbour-joining	Эвристический	Нет	Нет
Наименьших квадратов	Переборный	Может	Нет
Фитча - Марголиаша	Переборный	Может	Нет
Минимальной эволюции	Переборный	Может	Нет
Максимальной экономии	Переборный	Нет	Да
Наибольшего правдоподобия	Переборный	Может	Да

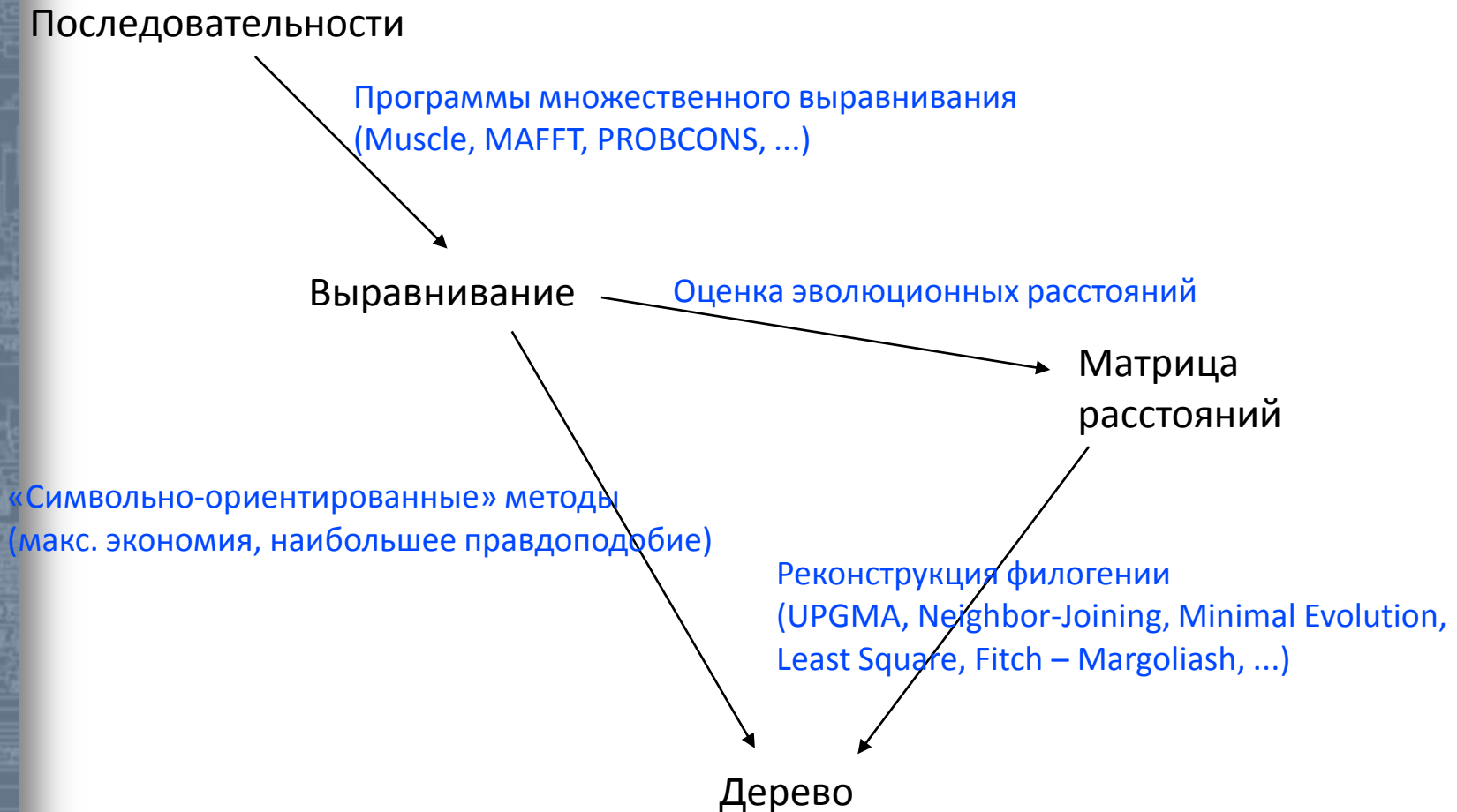
Исходный материал для реконструкции дерева: множественное выравнивание последовательностей белков

CYB5_CHICK	MVGSSEAGGEAWRGRYYRLEEVQKHNNNSQSTWIIIVHHRIYDITKFLDEHPGGEEVLREQA
CYB5_HUMAN	---MAEQSDEAV--KYYTLEEIQKHNSKSTWLILHHKVYDLTKFLEEHPGGEEVLREQA
CYB5_HORSE	---MAEQSDKAV--KYYTLEEIKKHNSKSTWLILHHKVYDLTKFLEDHPGGEEVLREQA
CYB5_MUSDO	-----MSSEDV--KYFTRAEVAKNNTKDKNWFI IHNNVYDVTAFLENEHPGGEEVLIEQA
CYB5_DROME	-----MSSEET--KTFTRAEVAKHNTNKDTWLLIHNNIYDVTAFLENEHPGGEEVLIEQA

Методы реконструкции филогенетических деревьев:

- символно ориентированные:
 - * максимальной экономии (maximal parsimony)
 - * наибольшего правдоподобия
- использующие матрицу расстояний
 - * кластерные (UPGMA и др.)
 - * Neighbor-joining
 - * минимальной эволюции
 - * наименьших квадратов
 - * Фитча – Марголиаша
 - * ...

Схема реконструкции филогении по последовательностям



Матрица расстояний

12						
YEAST	0.000000	1.226697	1.313711	1.187561	1.143864	1.198065
1.172131	1.171814	1.178217	1.168880	1.119266	1.136044	
TOBAC	1.226697	0.000000	1.130459	1.058881	1.185956	1.085361
1.168318	1.106069	1.156591	1.173863	1.193820	1.175669	
MUSDO	1.313711	1.130459	0.000000	0.281922	0.992650	0.859854
0.951601	0.863773	0.903579	0.874458	0.922674	0.976793	
DROME	1.187561	1.058881	0.281922	0.000000	0.968555	0.881378
0.935269	0.858443	0.913545	0.884540	0.939652	0.957328	
CHICK	1.143864	1.185956	0.992650	0.968555	0.000000	0.346056
0.281003	0.274432	0.288135	0.283946	0.272658	0.306560	
BOVIN	1.198065	1.085361	0.859854	0.881378	0.346056	0.000000
0.176332	0.107028	0.183987	0.165630	0.177340	0.166408	
HUMAN	1.172131	1.168318	0.951601	0.935269	0.281003	0.176332
0.000000	0.139340	0.133018	0.124695	0.102465	0.108339	
PIG	1.171814	1.106069	0.863773	0.858443	0.274432	0.107028
0.139340	0.000000	0.105960	0.082471	0.123761	0.090834	
MOUSE	1.178217	1.156591	0.903579	0.913545	0.288135	0.183987
0.133018	0.105960	0.000000	0.021980	0.102164	0.123996	
RAT	1.168880	1.173863	0.874458	0.884540	0.283946	0.165630
0.124695	0.082471	0.021980	0.000000	0.092132	0.099815	
RABIT	1.119266	1.193820	0.922674	0.939652	0.272658	0.177340
0.102465	0.123761	0.102164	0.092132	0.000000	0.090807	
HORSE	1.136044	1.175669	0.976793	0.957328	0.306560	0.166408
0.108339	0.090834	0.123996	0.099815	0.090807	0.000000	

Как оценить расстояние между последовательностями

По аддитивному набору расстояний дерево (с длинами ветвей) восстанавливается однозначно!

Но в реальности нам даны последовательности и требуется подсчитать расстояния, то есть оценить число произошедших мутаций.

Это не так просто, поскольку мутации могут происходить в одной и той же позиции.

Как оценить расстояние между последовательностями

По аддитивному набору расстояний дерево (с длинами веток) восстанавливается однозначно!

Но в реальности нам даны последовательности и требуется оценить число произошедших мутаций. Это не так просто, поскольку мутации могут происходить в одной и той же позиции.

Всё же простейшая оценка расстояния есть число различий, делённое на длину последовательности.

Более изощрённые методы учитывают тот факт, что чем больше наблюдаемое различие между последовательностями, тем больше можно ожидать повторных и возвратных мутаций в одинаковых позициях. Кроме того, учитывается, что разные замены аминокислотных остатков имеют в среднем разные шансы закрепиться.

То, что получается, как правило, не соответствует в точности расстояниям по какому-либо дереву!

Принцип наибольшего правдоподобия

Оцениваем причины по последствиям.

Принимаем как наиболее обоснованную гипотезу тот вариант причины, при котором вероятность наблюдаемых последствий наибольшая.

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

(моделей много, наиболее популярная называется JTT).

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

Переборные методы

Алгоритм, реализующий переборный метод, должен включать:

а) критерий сравнения деревьев (какая из двух топологий лучше соответствует исходным данным?)

б) алгоритм поиска лучшего по критерию дерева.

Пример критерия

(метод наименьших квадратов, OLS — ordinary least squares)

Пусть дана матрица расстояний и топология дерева;

i, j — две последовательности, тогда мы имеем расстояние $d(i, j)$ из матрицы. Приписав ветвям дерева длину, будем иметь расстояние $d'(i, j)$ «по дереву».

Подберём длины ветвей так, чтобы сумма величин $(d(i, j) - d'(i, j))^2$ (по всем парам листьев i, j) была наименьшей.

Это наименьшее значение и будет критерием качества: будем считать ту топологию лучшей, для которой это значение получится меньшим.

Поиск лучшего дерева

Имеется единственная топология (бинарного и неукоренённого) дерева с тремя листьями, три разных топологии деревьев с четырьмя листьями,

15 топологий деревьев с пятью листьями,

... ..

~ 2 млн. топологий деревьев с десятью листьями,

... ..

~ 8 трлн. топологий деревьев с 15 листьями,

... ..

Триллионы проверок компьютер будет делать слишком долго.

А ведь приходится строить деревья и с сотней листьев...

Поиск лучшего дерева

Имеется единственная топология (бинарного и неукоренённого) дерева с тремя листьями, три разных топологии деревьев с четырьмя листьями,

15 топологий деревьев с пятью листьями,

... ..

~ 2 млн. топологий деревьев с десятью листьями,

... ..

~ 8 трлн. топологий деревьев с 15 листьями,

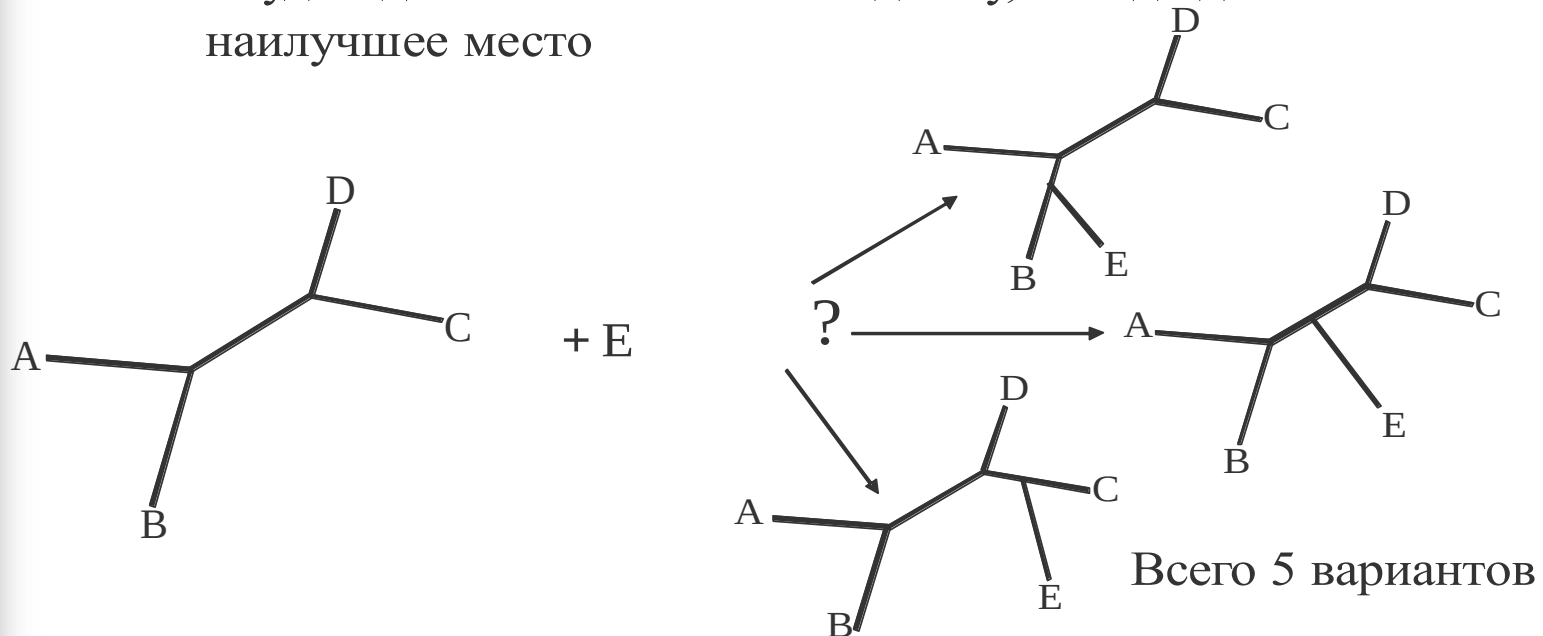
... ..

Триллионы проверок компьютер будет делать слишком долго. А ведь приходится строить деревья и с сотней листьев...

Поэтому программы, реализующие переборные методы, практически никогда не включают полный перебор всех возможных деревьев

Поиск лучшего дерева: «выращивание»

- Найдем лучшее дерево для части последовательностей
- Будем добавлять листья по одному, находя для них наилучшее место



Поиск лучшего дерева: «выращивание»

Дерево с N листьями всегда имеет $2N-3$ ветви.

Поэтому, чтобы “вырастить” дерево с N листьями, надо проанализировать

$3 + 5 + \dots + (2N - 5) = (N - 3)(N - 1)$ деревьев.

Уже для $N=10$ это число меньше числа всех возможных деревьев в 32175 раз!

Выращивание не гарантирует нахождение “лучшего” дерева, но при хороших данных не должно приводить к большим ошибкам.

Поиск лучшего дерева: просмотр соседних деревьев

Построим сначала «черновое» дерево, а затем попробуем его улучшить.

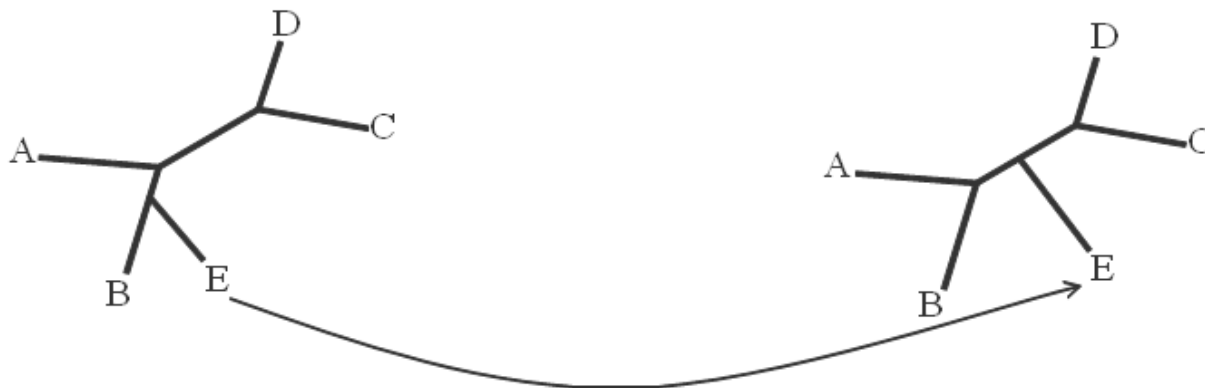
Черновое дерево можно построить одним из эвристических методов или «вырастить».

Улучшать будем, просматривая «соседние» деревья.

Поиск лучшего дерева: просмотр соседних деревьев

Что такое «соседние деревья»

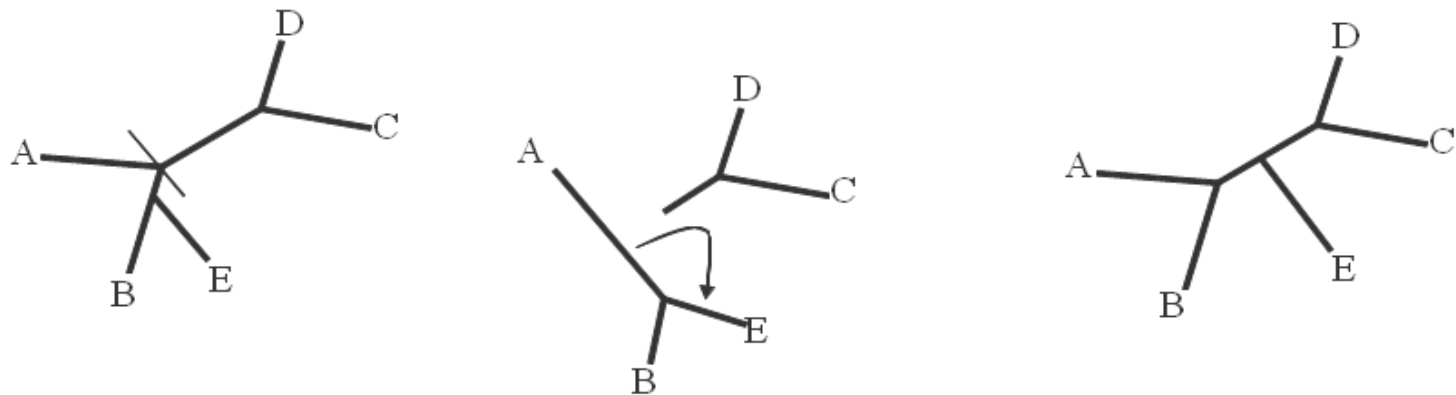
- Оторвём один лист и «привьём» его на другую ветвь



Поиск лучшего дерева: просмотр соседних деревьев

Что такое «соседние деревья»

- Можно проделать аналогичную операцию с целой кладой

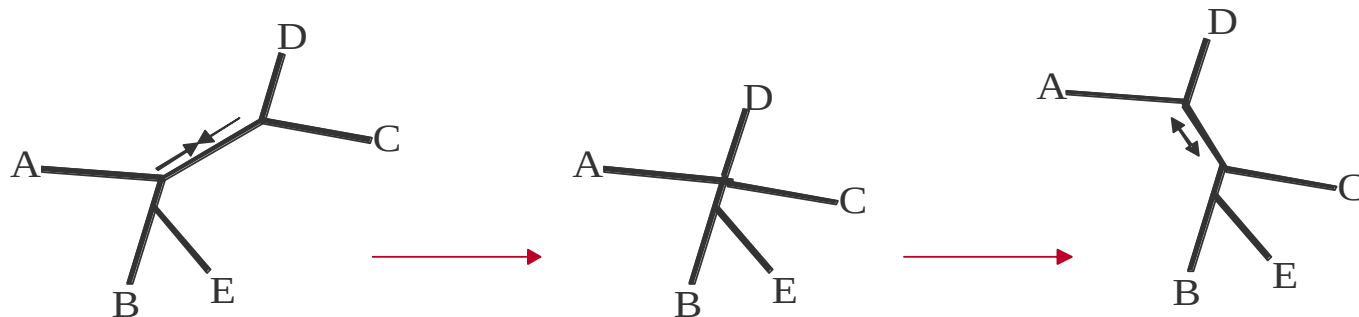


Такая операция обычно называется “SPR” : Subtree Pruning and Regrafting
В пакете PHYLIP она называется “Global rearrangement”.

Поиск лучшего дерева: просмотр соседних деревьев

Что такое «соседние деревья»

- Можно «схлопнуть» одну ветвь и заменить её другой



Такая операция обычно называется “NNI” : Nearest Neighbor Interchange.
В пакете PHYLIP она называется “Local rearrangement”.

Поиск лучшего дерева

Алгоритм поиска

- Строим черновое дерево
(два варианта: эвристический метод или «выращивание» с использованием критерия качества).
- Анализируем соседние деревья;
если находим среди соседей лучшее, берём за основу его.
- Повторяем предыдущий пункт, пока
текущее дерево не окажется лучше всех
своих соседей.

Переборные методы

Название метода всегда совпадает с названием критерия качества

- Максимальной экономии (или «бережливости», **maximum parsimony**, MP)
- Наибольшего правдоподобия (**maximum likelihood**, ML)
- Наименьших квадратов (**least squares**, LS)
- Фитча - Марголиаша (**Fitch - Margoliash**, FM)
- Минимальной эволюции (**minimal evolution**, ME)

Все методы, кроме максимальной экономии, допускают предположение о молекулярных часах (но чаще используются без этого предположения!) и оценивают длины ветвей.

Методы MP и ML — символно-ориентированные, LS, FM, ME и многие другие принимают на вход матрицу расстояний.

Эвристические методы

- UPGMA = «Unweighted pair group method with arithmetic mean»
 - Строит укоренённое ультраметрическое дерево
 - Видимо, реально лучший из методов, предполагающих молекулярные часы.
- Neighbor-Joining
 - **Строит неукоренённое дерево.** Если и уступает некоторым переборным алгоритмам, то не сильно.

Оба метода принимают на вход матрицу расстояний.

UPGMA - схема алгоритма

Укоренённое дерево строится «снизу вверх»

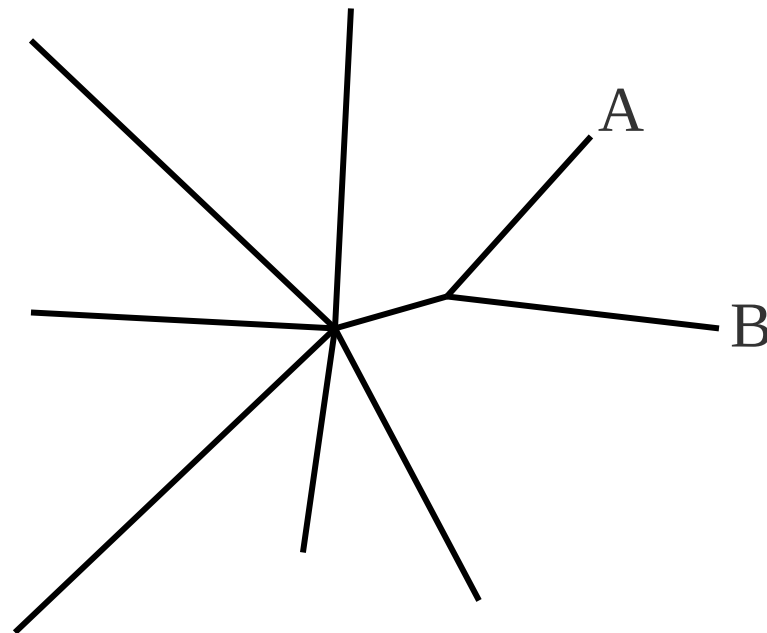
- Найдём в матрице расстояний наименьший элемент.
- Объединим два ближайших листа в кластер (это - узел дерева, соединённый ветвями с листьями, образовавшими его).
- Пересчитаем матрицу расстояний, рассматривая кластер как новый лист. Расстоянием до кластера будем считать **среднее арифметическое** расстояний 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»

Идея алгоритма neighbor-joining



A и B — такая пара последовательностей, для которых минимальна величина $(n - 2)\rho(A, B) - M(A, B)$, где n — число последовательностей, ρ — расстояние из матрицы, а $M(A, B) = \sum_C (\rho(A, C) + \rho(B, C))$

Такие «соседи» дальше рассматриваются как один лист. «Объединение соседей» продолжается, пока не останутся только три «листа».

Свойства алгоритма NJ

Время работы: $O(n^3)$

Предложен очень близкий алгоритм с временем работы $O(n^2)$

Если исходная матрица расстояний состояла из расстояний, вычисленных по некоторому дереву, то это дерево будет восстановлено.

Показывает хорошие результаты на практике.

Теорема (K. Atteson, 1999)

Пусть заданы расстояния между набором объектов, которые отличаются от расстояний между теми же объектами, посчитанными по некоторому дереву, самое большее на ε . Тогда если самая короткая ветвь дерева имеет длину больше 2ε , то алгоритм NJ восстановит дерево по заданному набору расстояний.

В этой науке данное утверждение формулируется так:

“Neighbor-joining has l_∞ - radius $1/2$ ”

В той же работе Аттесона была сформулирована следующая гипотеза.

Пусть расстояния отличаются от расстояний по некоторому дереву самое большее на ε . Тогда алгоритм NJ восстановит все ветви этого дерева, чья длина больше 4ε .

“Neighbor-joining has edge l_∞ - radius $1/4$ ”

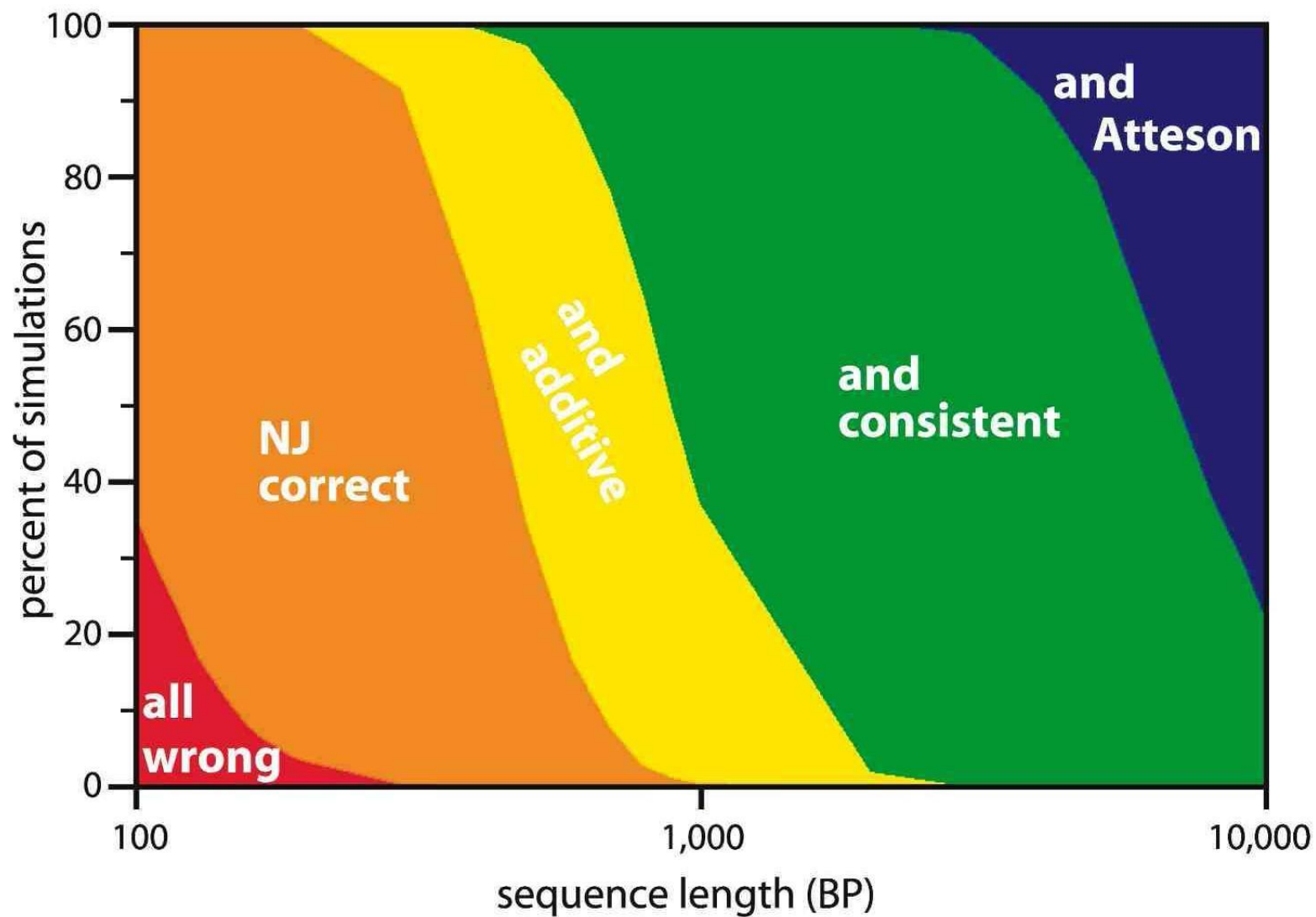
В работе Mihaescu et al. “Why neighbor-joining works?”(2006–2009)
доказано **усиление** гипотезы Аттесона

Пусть некоторая внутренняя ветвь дерева имеет длину L и пусть A и B – части, на которые эта ветвь делит множество листьев.

Предположим, что матрица расстояний между листьями такова, что:

- 1) расстояния между a и b для $a \in A$ и $b \in B$ отличаются от расстояний по дереву меньше, чем на $L / 4$;
- 2) для всех прочих пар (внутри A и внутри B) расстояния не превосходят истинные больше, чем на $L / 4$ (могут быть как угодно малы).

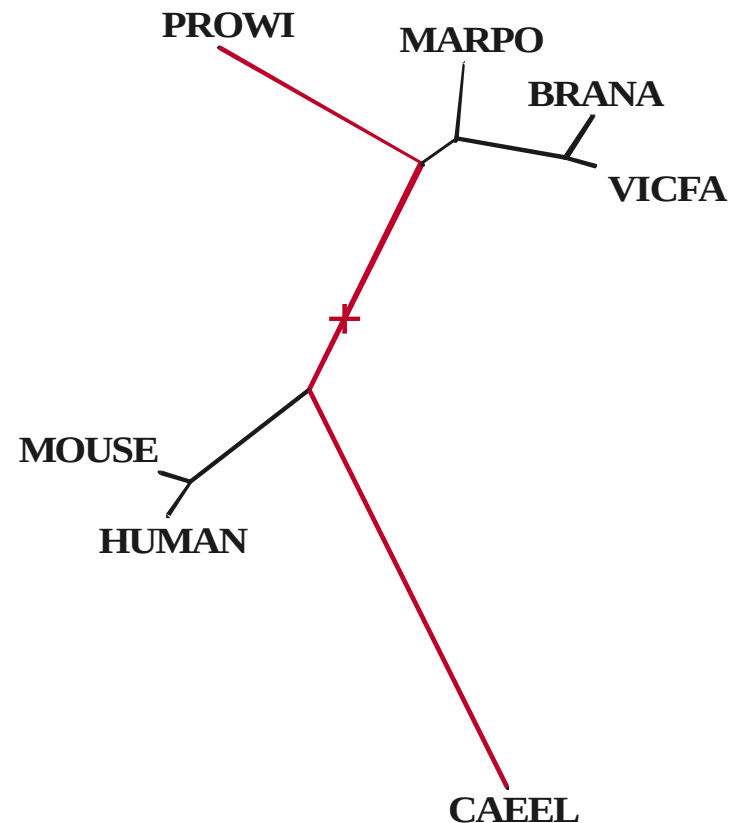
Тогда дерево, построенное алгоритмом NJ по такой матрице, будет содержать данную ветвь.



Укоренение

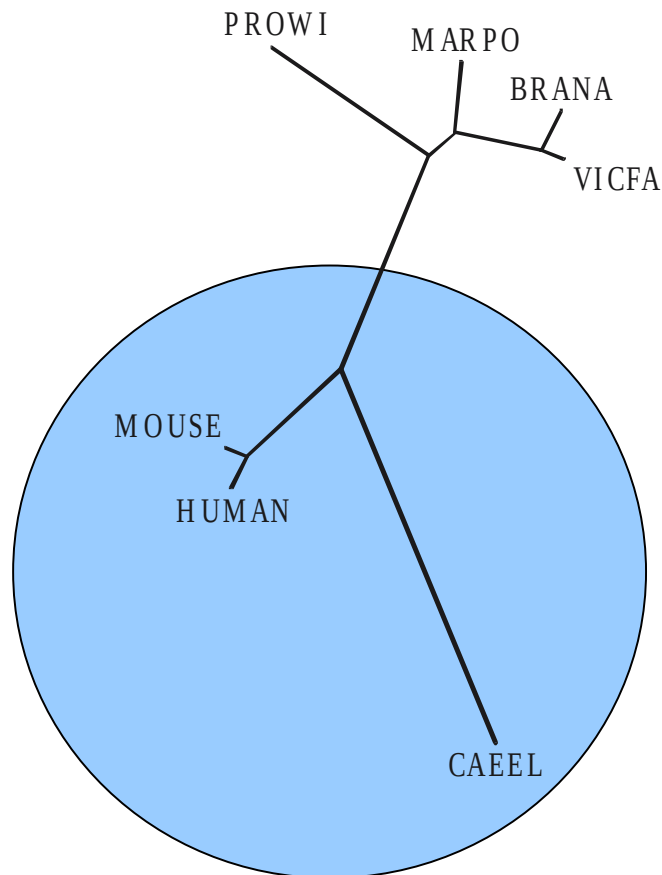
В среднюю точку:

Находим на дереве самый длинный путь от листа к листу и за корень принимаем середину этого пути



Укоренение

Используя внешнюю группу («аутгруппу», outgroup):

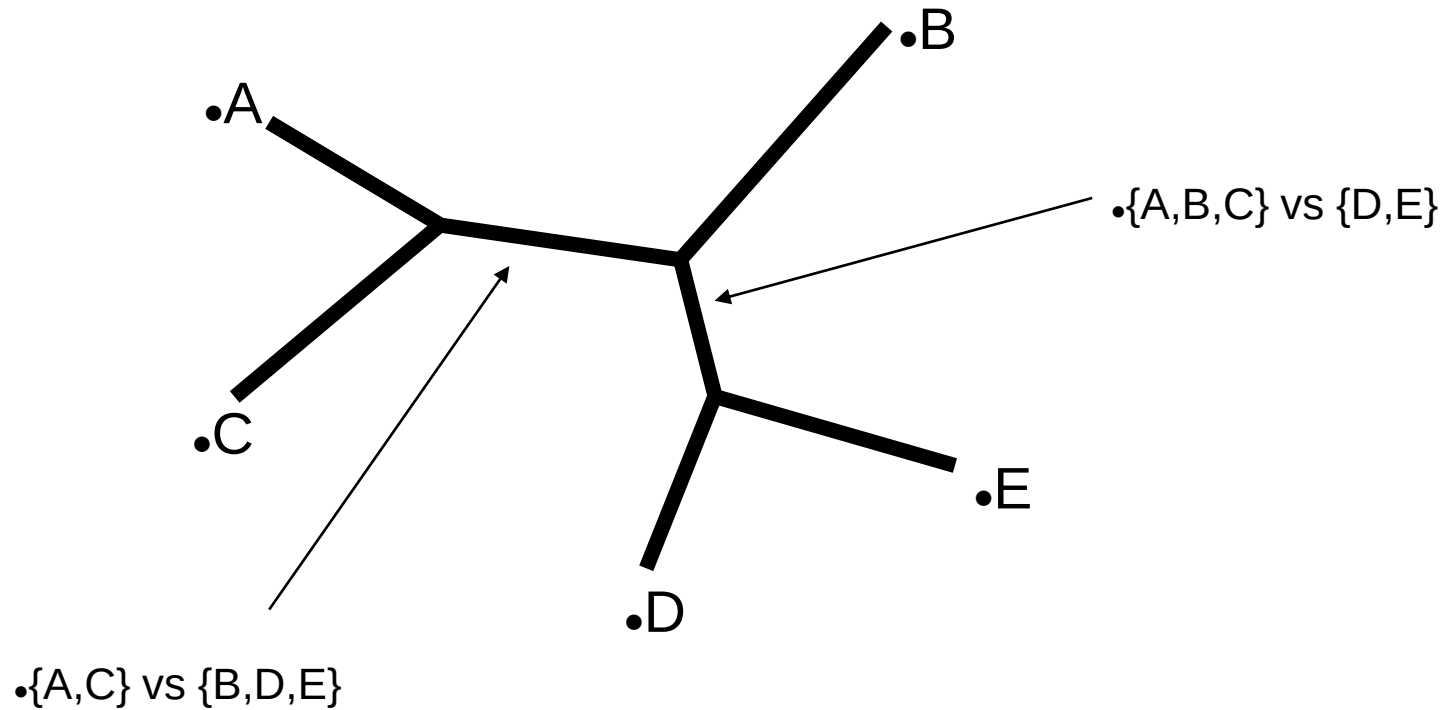


В данном случае укоренено дерево четырёх растений, для чего пришлось построить дерево с участием внешней группы — трёх животных (в синем круге)

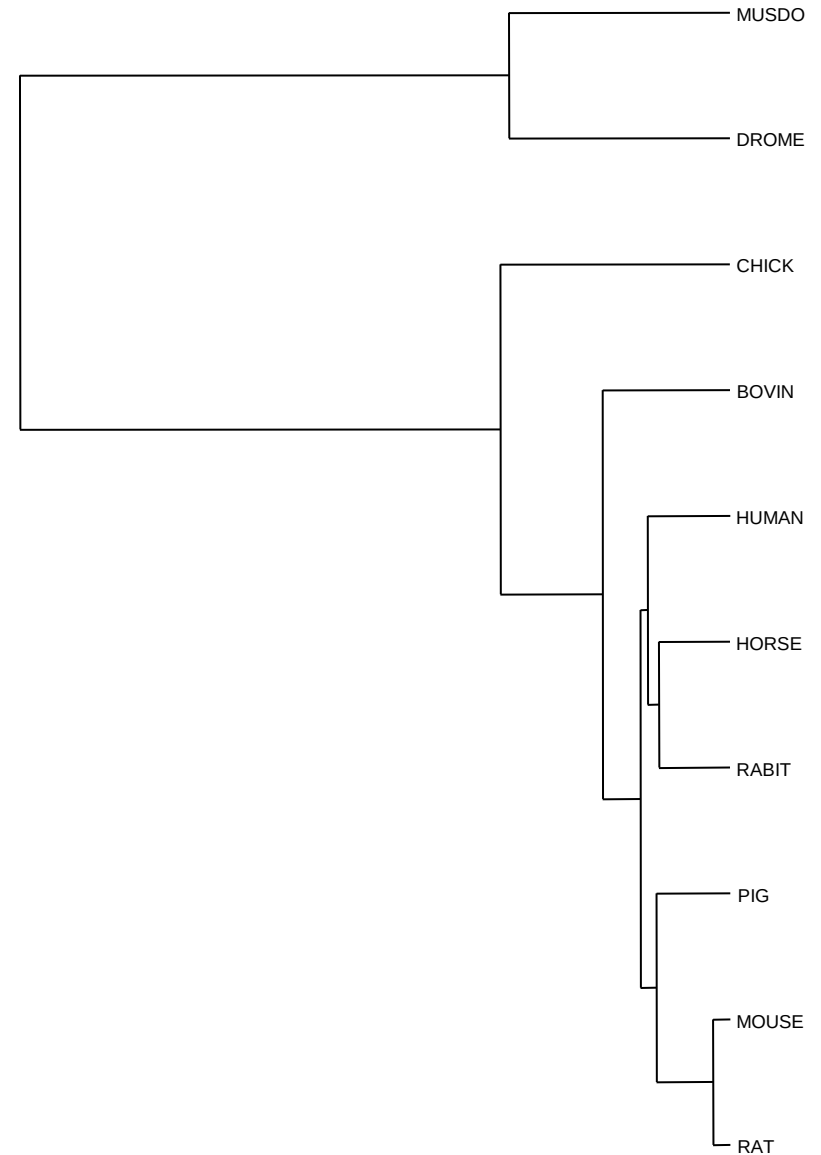
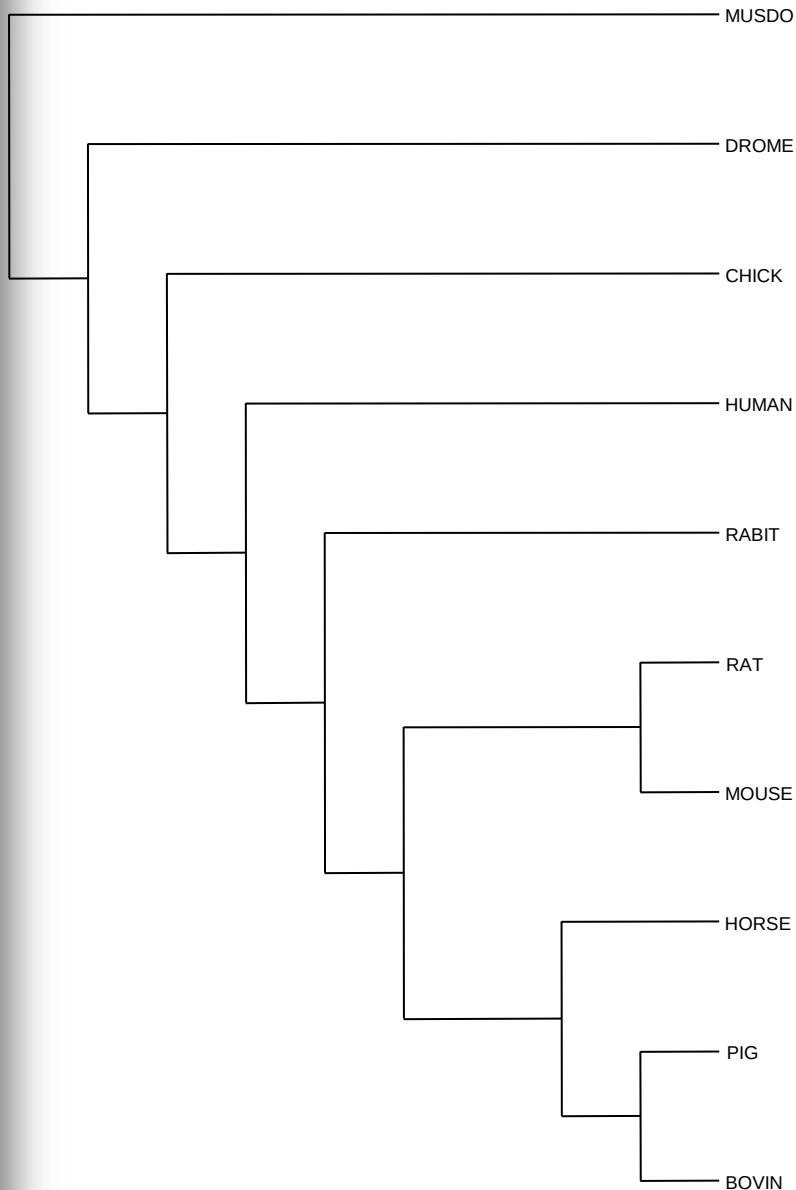
Сравнение деревьев

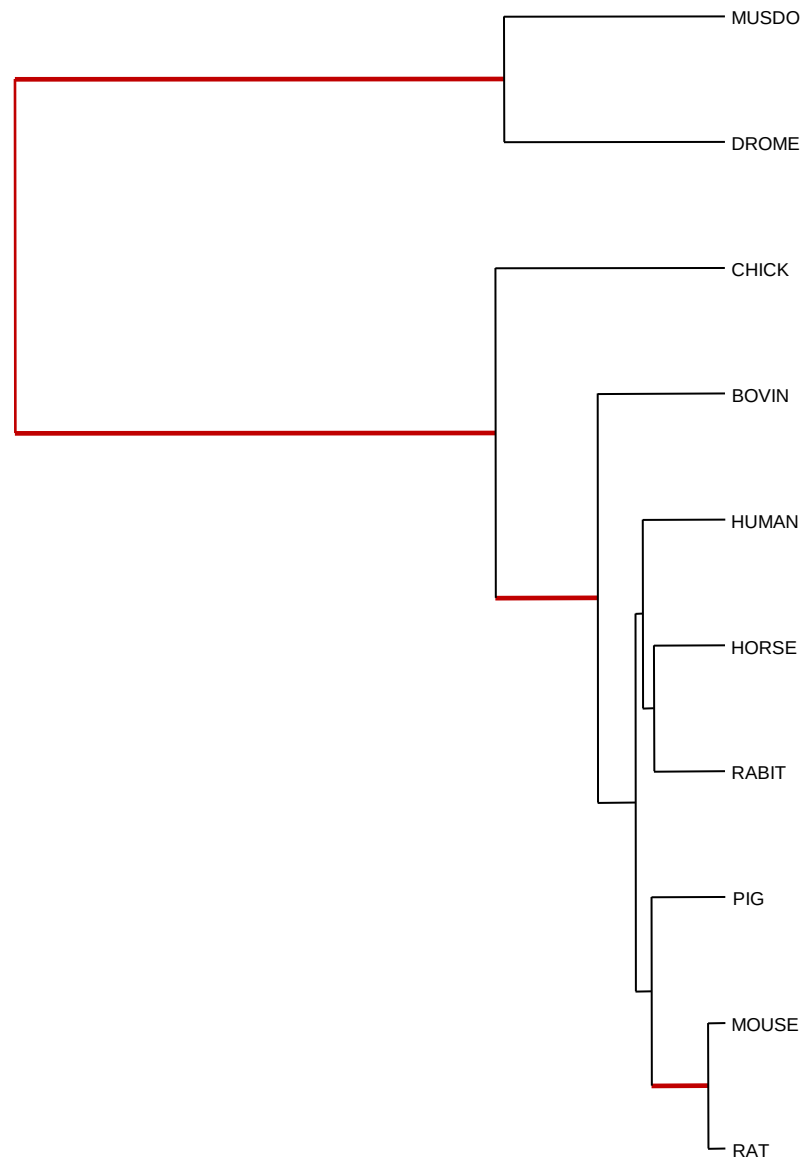
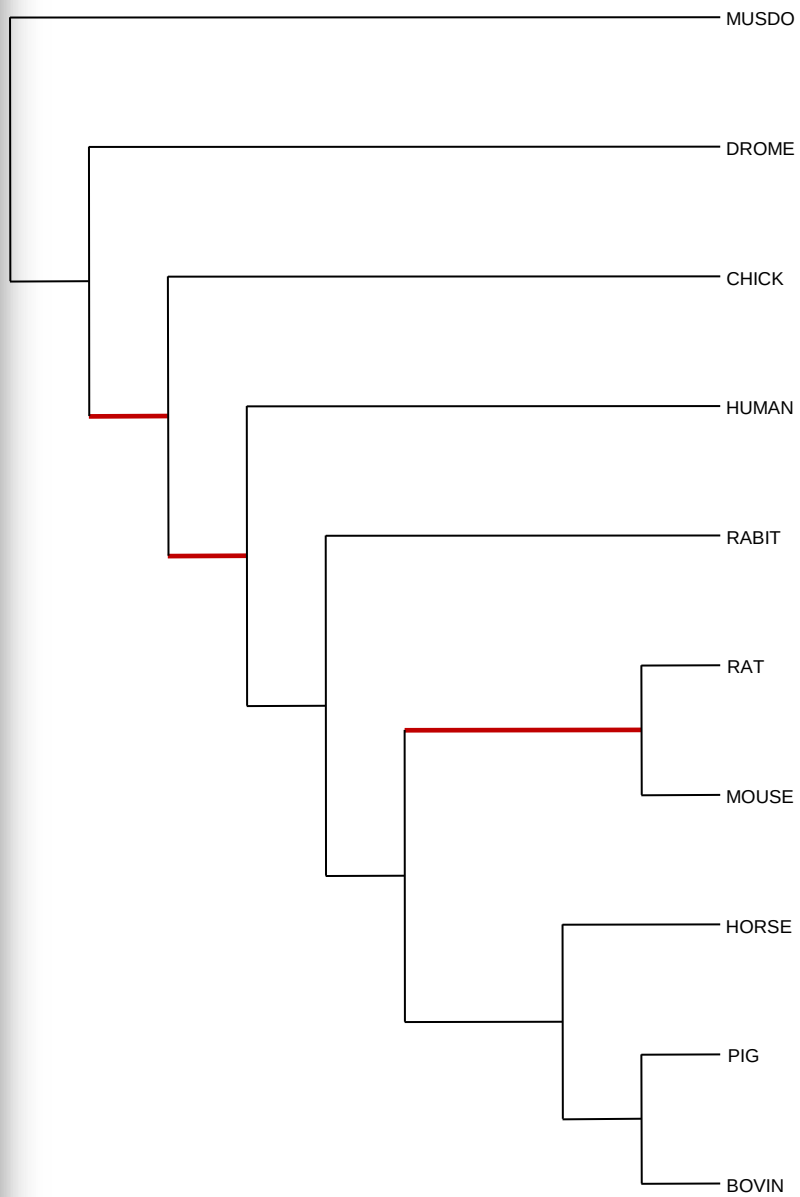
- Программы реконструкции филогении так же ненадёжны, как и любые другие компьютерные программы предсказания биологических фактов.
- Поэтому (в частности) возможны различные варианты реконструкции по одним и тем же данным.
- Встаёт задача сравнения различных деревьев с одним и тем же множеством листьев.
- Та же задача возникает, например, при изучении горизонтальных переносов генов, рекомбинации вирусных геномов и т.п.

Ветвь дерева как разбиение множества листьев

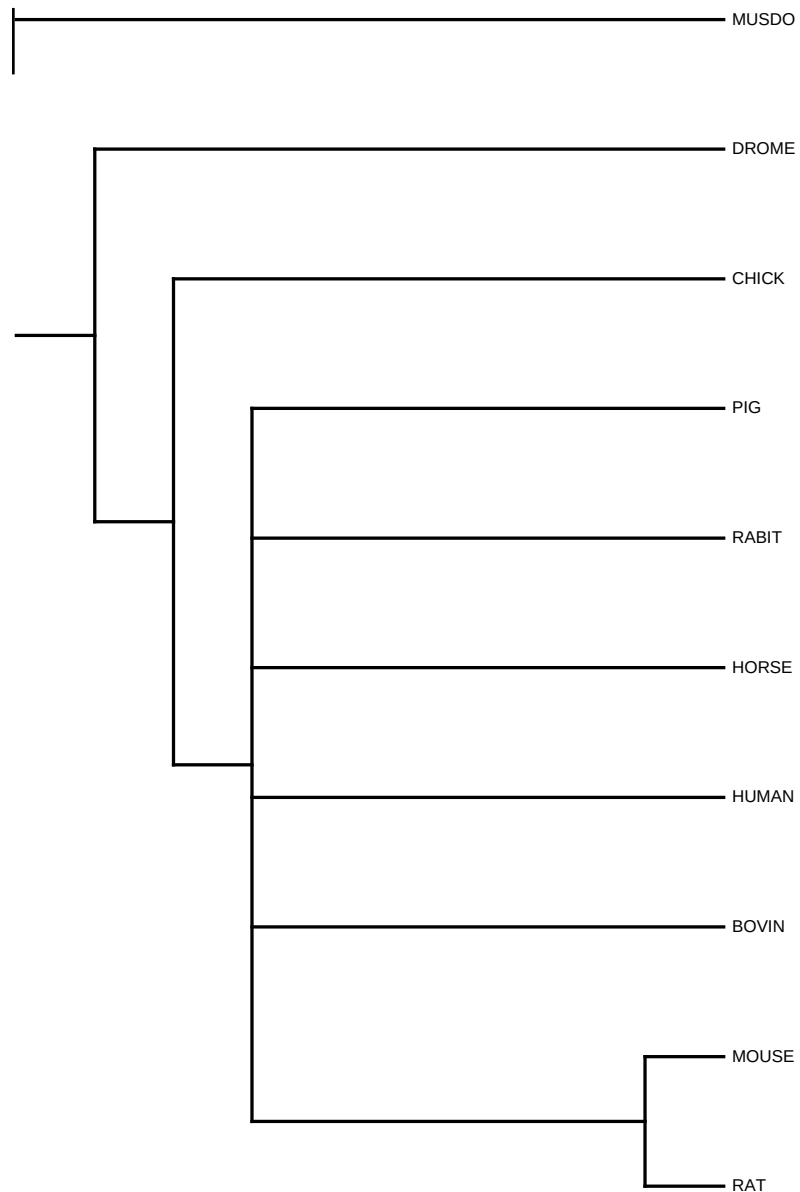


Что общего у этих двух деревьев?





Консенсусное дерево



Дерево большинства (Majority-rule tree)

- Строится по большому набору деревьев с одинаковым множеством листьев (например, деревья одного и того же набора бактерий, реконструированные по разным ортологическим рядам белков)
- Включает только те ветви, которые встретились в большинстве деревьев исходного набора.

Сравнение деревьев

- . Консенсусное (небинарное) дерево
- . Максимальное общее поддерево
- . Дерево из ветвей, поддержанных большинством (Majority-rule tree)

.Меры сходства между деревьями:

- . Доля общих ветвей
- . Размер максимального общего поддерева
- . Доля общих четвёрок
- . Длина пути в пространстве деревьев

Результат одного из сравнений методов

Программы пакета PHYLIP **neighbor** (метод Neighbor-Joining) и **protpars** (метод максимальной экономии) были применены к 1189 выравниваниям. Каждое выравнивание включало 26 последовательностей белковых доменов, по одному из 26 видов грибов. Последовательности были выбраны из Pfam так, чтобы в большинстве случаев представлять ортологический ряд. Поэтому их филогения должна в большинстве случаев совпадать с филогенией соответствующих видов. Для 431 случаев дерево, построенное программой **neighbor**, оказалось по всем мерам ближе к дереву видов, нежели дерево, построенное программой **protpars**. В 36 случаях наблюдалась обратная картина. Вероятность получить такое соотношение по случайным причинам – менее 10^{-86} .

А вот при сравнении программ **neighbor** и **fitch** (метод Фитча – Марголиаша) соотношение получилось 65:55. Такой результат не позволяет сделать никаких выводов: вероятность получить такое или большее соотношение при «нулевой гипотезе» – равенстве методов – около 0,4.

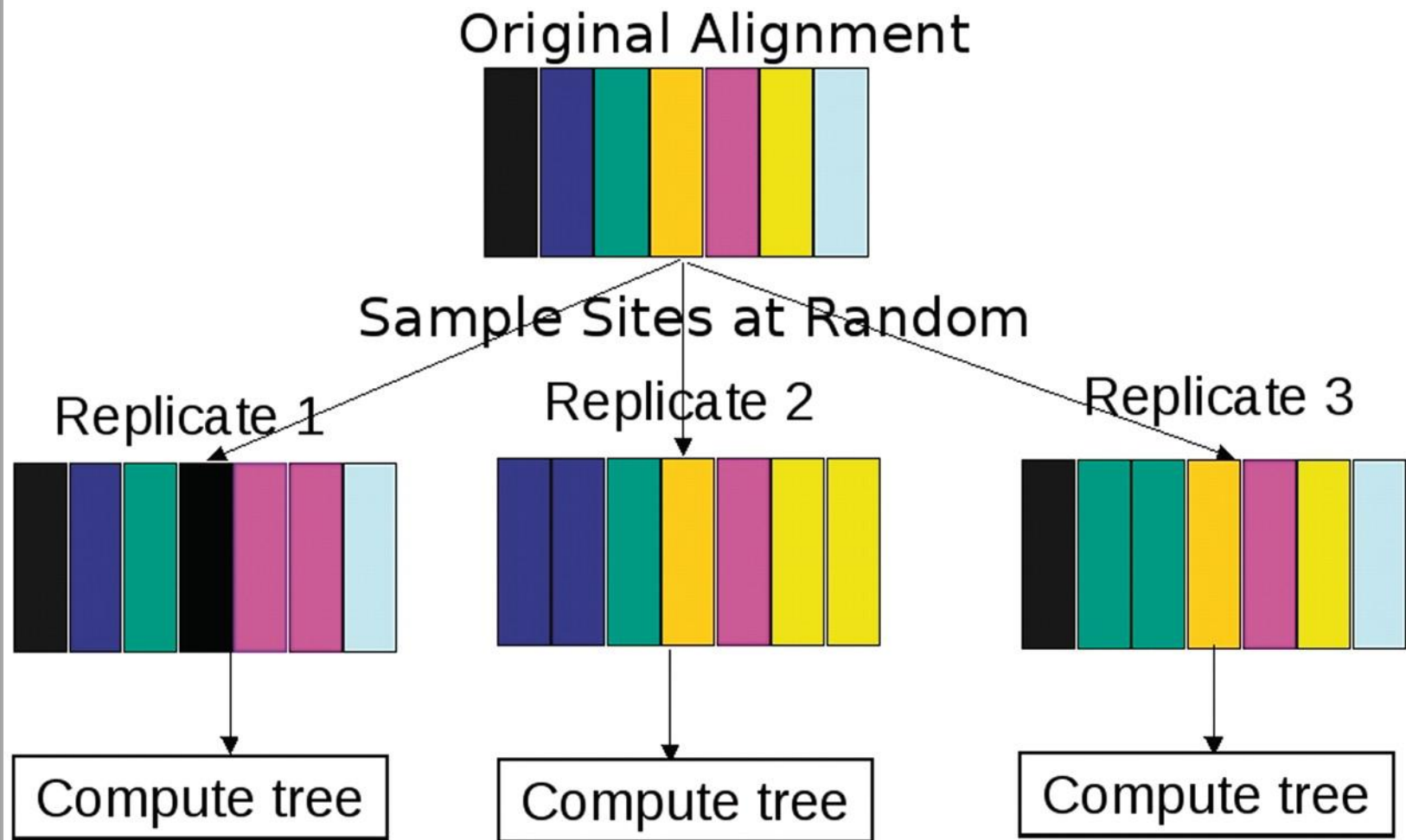
Бутстрэп-анализ

Из входного выравнивания делается много (например, 100) так наз. «бутстрэп-реплик».

Каждая бутстрэп-реплика получается в результате случайного удаления половины столбцов из выравнивания с заменой их копиями других (тоже случайно выбранных) столбцов.

Смысл в том, чтобы построить дерево по половине данных и затем сравнить результаты от по разному выбранных половин.

•Outline of the phylogenetic bootstrap procedure.



•Stamatakis A , Izquierdo-Carrasco F Brief Bioinform
2011;12:270-279

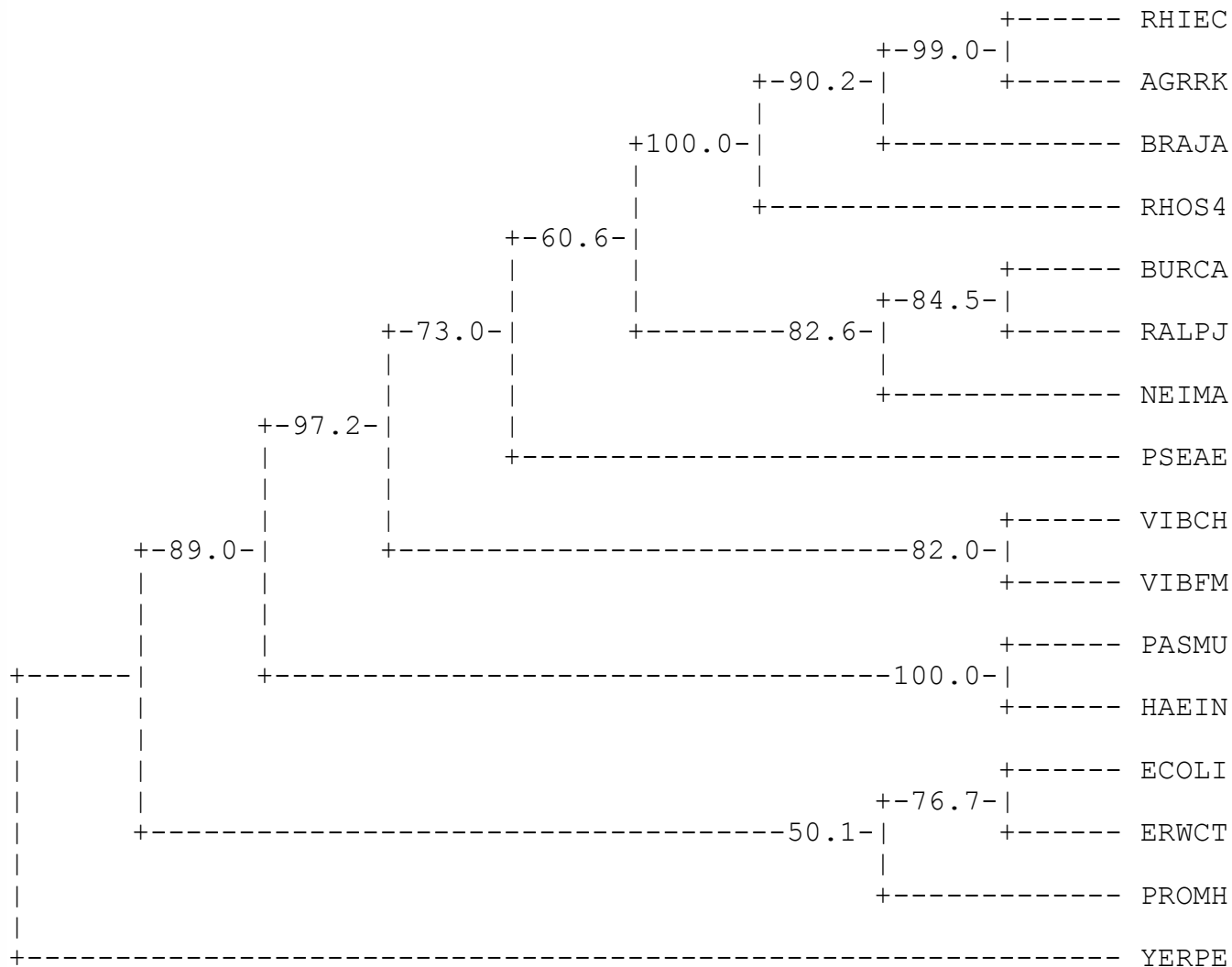
Бутстрэп-анализ

- создаём из входного выравнивания 100 бутстрэп-реплик;
- для каждой из реплик строим по дереву;
- из 100 деревьев строим дерево по методу расширенного большинства («Extended majority-rule tree»).

Помимо того, что (как правило) возрастает качество реконструкции, есть возможность оценить достоверность каждой ветви по т.н. «бутстрэп-поддержке», то есть проценту деревьев, в которых встретилась данная ветвь.

Бутстрэп-анализ

(пример результата)



Программы

- JalView

<http://www.jalview.org/>

Работа с выравниваниями.

По невыровненным последовательностям реконструировать эволюцию нельзя!

Может строить деревья (методы UPGMA и Neighbor-Joining)

- MEGA

<http://www.megasoftware.net/>

Визуализация и построение деревьев (оконный интерфейс)

- Пакет PHYLIP

<http://evolution.genetics.washington.edu/phylip.html>

Построение и визуализация деревьев (интерактивный интерфейс из командной строки). Имеется веб-интерфейс на портале <http://bioweb2.pasteur.fr/>

JalView

Основное назначение - работа с выравниваниями.

Включает реализацию двух примитивных способов оценки эволюционных расстояний (по числу совпадений и по весу сравнения с использованием BLOSUM62) и двух эвристических алгоритмов: UPGMA (переименованного в “Average Distance”) и Neighbor-Joining.

Очень часто этого вполне достаточно!

Пакет PHYLIP

- Реализация методов UPGMA и Neighbor-Joining (программа *neighbor*), наименьших квадратов и Фитча - Марголиаша (*fitch* и *kitsch*), максимальной экономии (*dnapars* и *protpars*), наибольшего правдоподобия (*dnaml*, *dnamlk*, *proml*, *promlk*)
- Оценка эволюционных расстояний: программы *dnadist* и *protdist*
- Сравнение деревьев: *consense*, *treedist*, *treedistpair*
- Редактура (включая укоренение в среднюю точку): *retree*
- Бутстрэп: *seqboot*
- Визуализация: *drawtree*, *drawgram*

Пакет PHYLIP

- Свободно распространяется, имеются версии для всех основных операционных систем. Доступен для скачивания на сайте
<http://evolution.genetics.washington.edu/phylip.html>
- Своеобразный интерактивный (не оконный) интерфейс
- В пакет EMBOSS в качестве дополнения включены варианты всех программ пакета PHYLIP, снабженные интерфейсом в стиле EMBOSS (отличаются буквой *f* в начале, например *fprotpars* вместо *protpars*)

MEGA

<http://www.megasoftware.net/>

Помимо хорошей визуализации деревьев, включает реализацию ряда алгоритмов: вычисление расстояний, UPGMA, NJ, ML, MP, один из вариантов ME (не лучший).



Другие бесплатные программы

В отличие от MEGA, без оконного интерфейса

- FastME - очень хороший вариант минимальной эволюции
- TNT - максимальная экономия
- PhyML и RaXML - два пакета, реализующих наибольшее правдоподобие, с большим количеством моделей
- MrBayes - т.н. байесов метод (вариант ML, очень медленный, но часто даёт лучшие результаты, чем обычный ML).
- PhyloBayes - ещё одна популярная реализация байесова метода

См. также

<http://evolution.genetics.washington.edu/phyip/software.html>

Что важно помнить при реконструкции филогении по последовательностям

1. Последовательности должны быть гомологичны по всей длине.
2. Если последовательности нуклеотидные, надо убедиться, что часть из них не представлена комплементарными вариантами.
3. Последовательности необходимо выравнивать.
Кстати, по виду выравнивания можно оценить, действительно ли последовательности такие, как надо: должно быть много консервативных колонок и мало хаотично расположенных гэпов. Помните: программа выравнивания выдаст результат даже для совершенно неродственных последовательностей, но смысла в этом результате не будет!
4. Большинство программ реконструируют неукоренённое дерево (даже если оно выглядит как укоренённое). Определение положения корня – отдельная задача.
5. Результат реконструкции – не абсолютная истина. Достоверность той или иной ветви можно оценить путём сравнения результатов разных программ и/или бутстрепа.
Как правило, чем короче выравнивание, тем хуже качество реконструкции.