Деревья (trees)

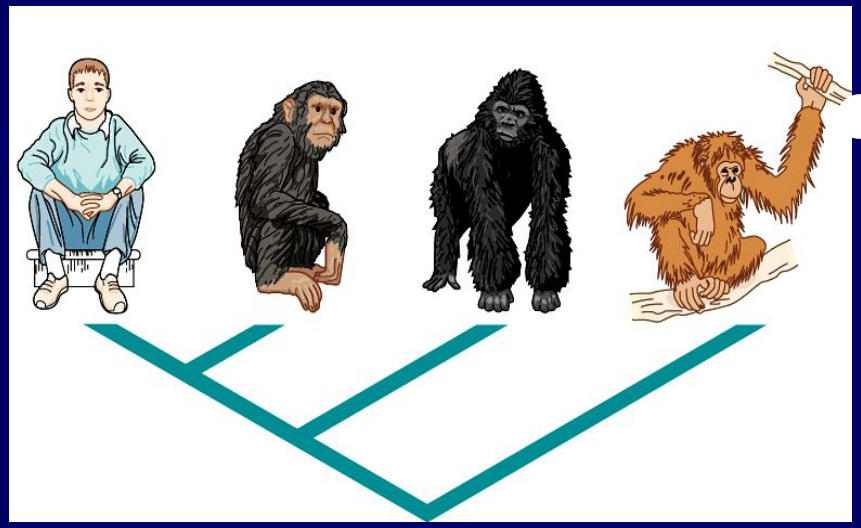
«…великое Дерево Жизни заполняет земную кору своими мертвыми и сломанными ветвями и покрывает поверхность вечно ветвящимися и прекрасными побегами»

Ч. Дарвин

Задача построения филогенетического дерева

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.

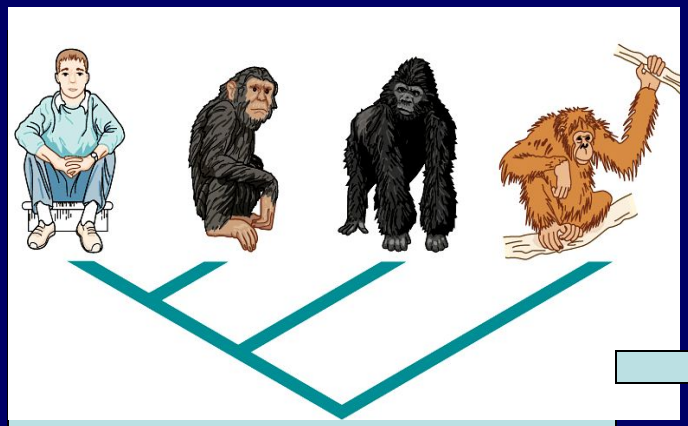
Charles Darwin

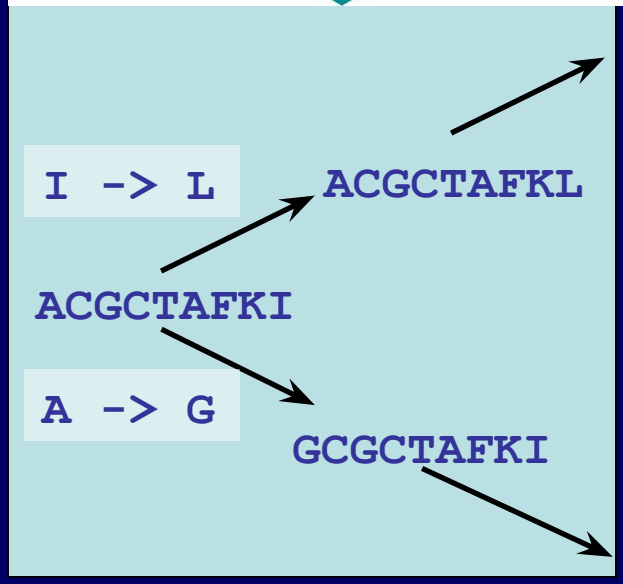


- **<u>Биологические задачи –</u>**
  - **сравнение 3-х и более объектов** (***кто на кого более похож ....)***
  - **реконструкция эволюции** (***кто от кого, как и когда произошел…)***
- **<u>Математическая задача –</u>** задача кластеризации, использование теории графов и комбинаторной оптимизации для того, чтобы на основе «грязных» биологических данных

Реальные события :

эволюция в природе или в лаборатории, компьютерная симуляция





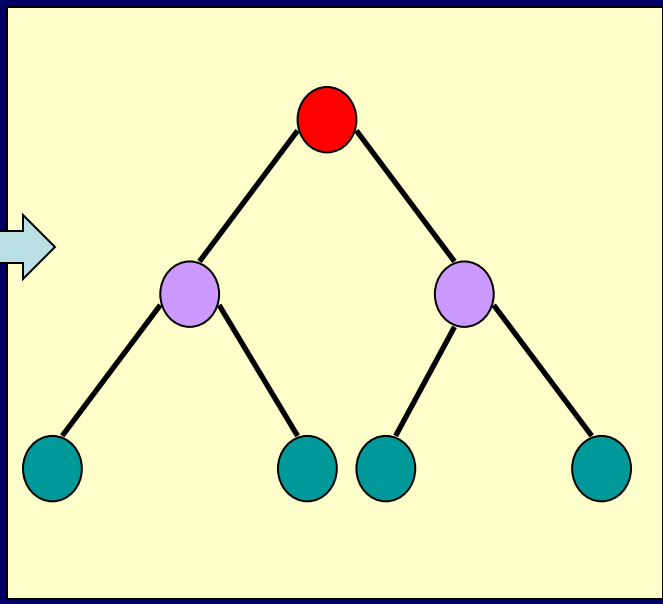
Данные:

например, а.к. последовательности или количество усиков

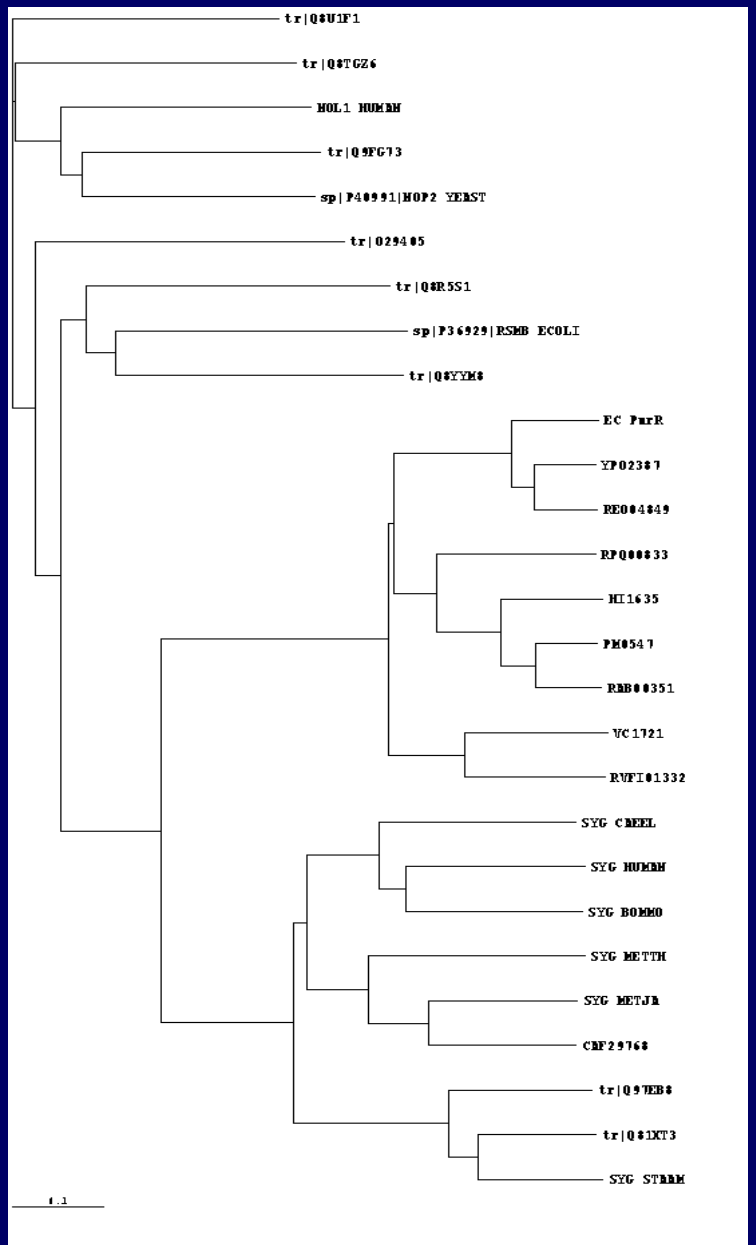
```
>Seq1
ASGCTAFKL
. . .
>Seq3
GCGCTLFKI
>Seq4
GCGCTGFKI
. . . . .
```

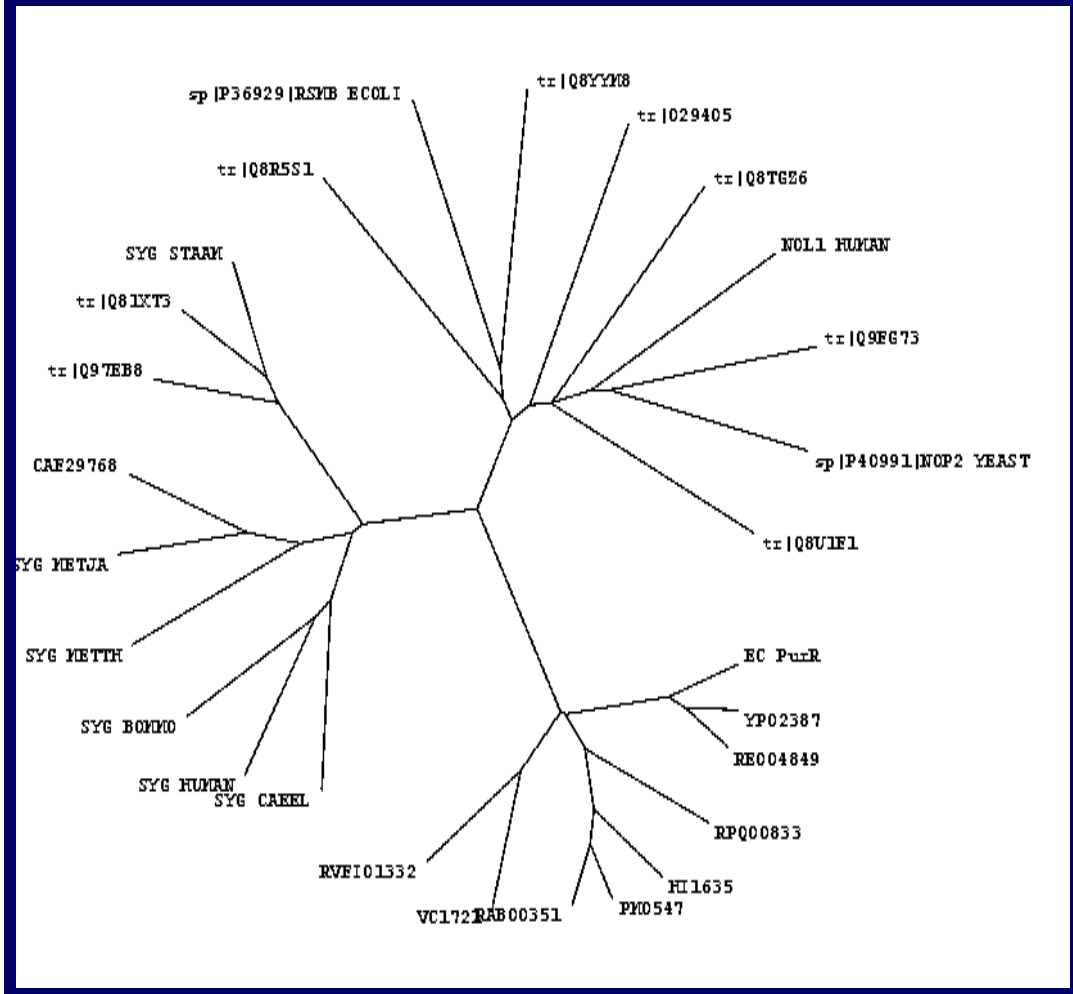
Построенное дерево

древовидный граф, ***вычисленный на основе данных***, может отражать или не отражать реальные события



Будни биоинформатика – деревья, деревья…





Рутинная процедура

Составление выборки последовательностей

Множественное выравнивание

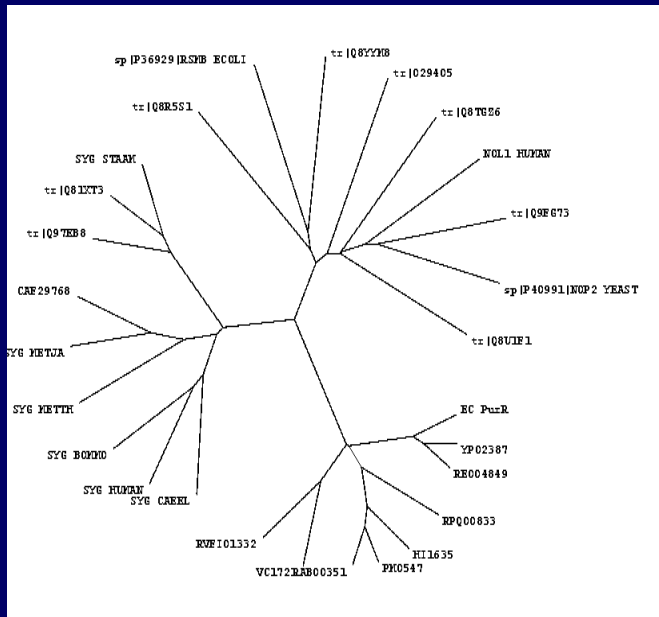
```
f53969          ALLRTPAEFDAYELNSSIKGAGTDEACLIEILSSRSNAEIKEINRIYKQEYKKTLEDAIK 313
con101          AMLKTPSQYDAYELKEAIKGAGTDEACLIEILASRSNAEIREINQVFKAENKKSLEDAIS 347
hum4            GMMTPTVLYDVQELRRAMKGAGTDEGCLIEILASRTPEEIRRISQTYQQQYGRSLEDDIR 140
ANX4_MOUSE      GLMTPTVLYDVQELRRAMKGAGTDEGCLIEILASRTPEEIRRINQTYQQQYGRSLEKDIC 140
max1            GLLMPAPVYDAYELKAAMKGAGTEEACLIDILASRSNSEMNAINEVYKKEYGKTLEDAVC 143
```

Построение дерева

фрагмент записи в виде правильной скобочной структуры:

(((((con101:38.51018,(f53969:28.26973,((f67220:8.39851,
max4:27.50591):4.92893,con92:30.19677):13.62315):9.53075):
25.83145,

Визуализация и редактура дерева



Основные термины

Узел (вершина, node) – таксономическая единица (taxonomic units – TU), может соответствовать видам, популяциям, нуклеотидным или аминокислотным последовательностям.

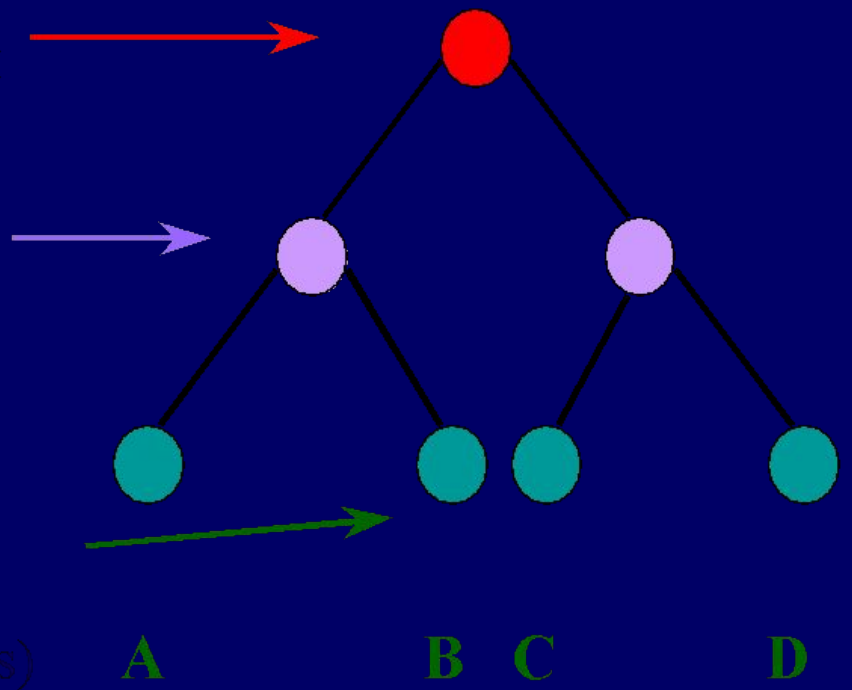
Ветвь (ребро, branch) – связь между узлами.

Топология дерева – порядок ветвления дерева.

Корень – гипотетический общий предок.

Внутренние узлы представляют ближайших *гипотетических* предков (HTUs).

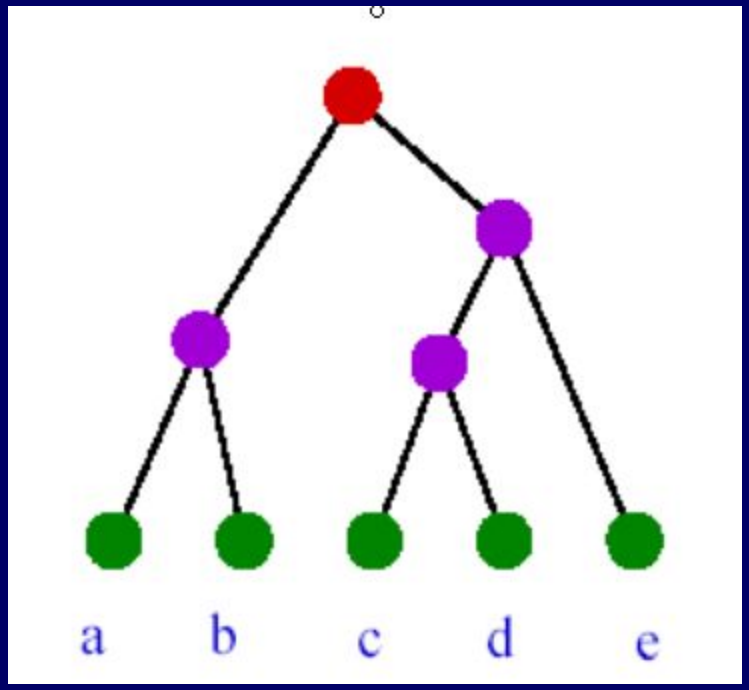
Листья или внешние узлы представляют реальные объекты (operational taxonomic units, OTUs)

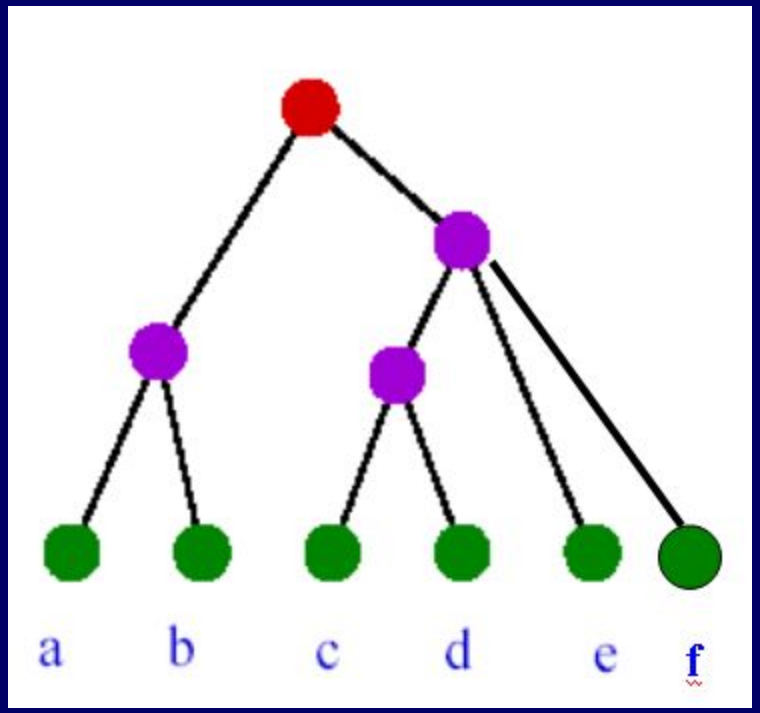


Какие бывают построенные деревья?

Бинарное разрешенное
(в один момент времени может произойти одно событие)

Бинарное неразрешенное
(может ли в один момент времени произойти два события?)

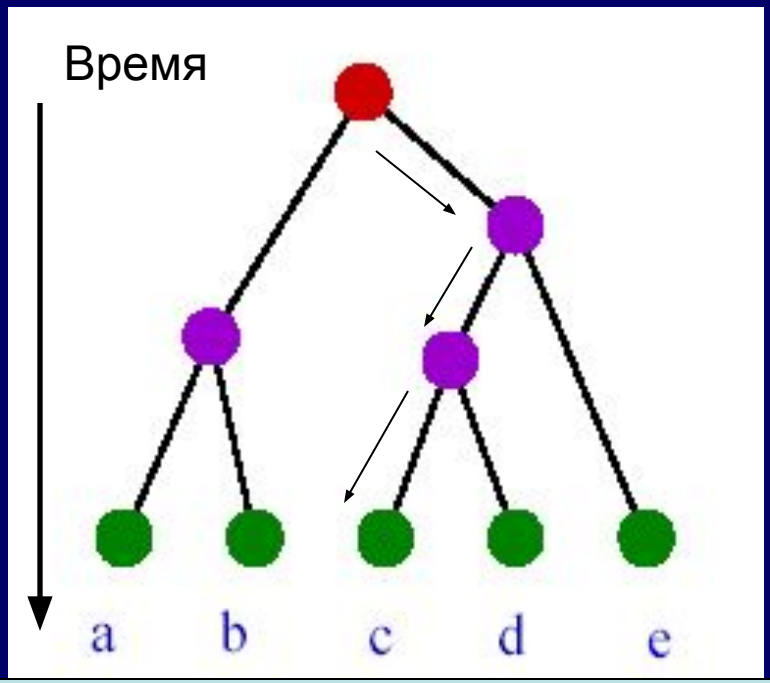




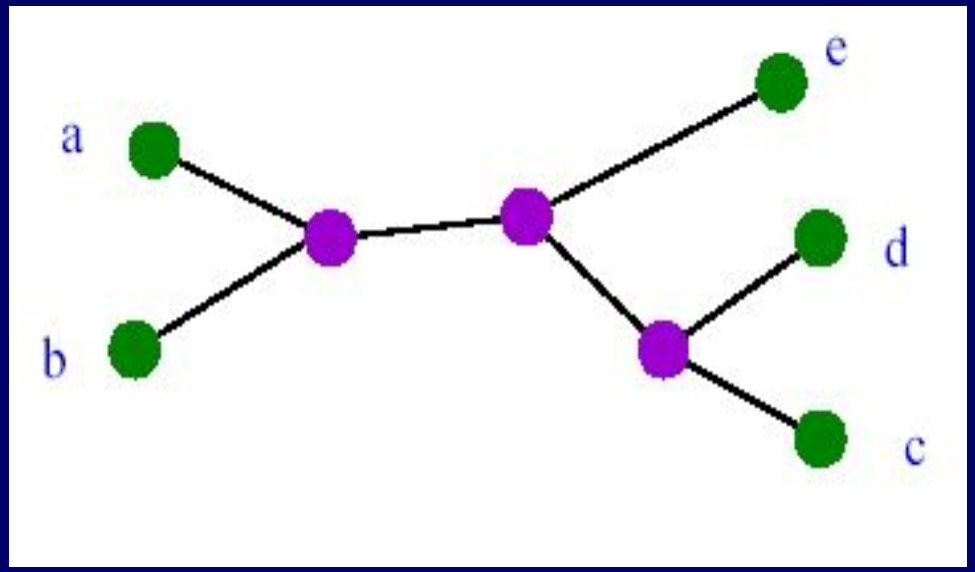
Время

Какие бывают построенные деревья?

Укорененное ориентированное дерево отражает направление эволюции



Неукорененное (бескорневое) неориентированное дерево показывает только связи между узлами



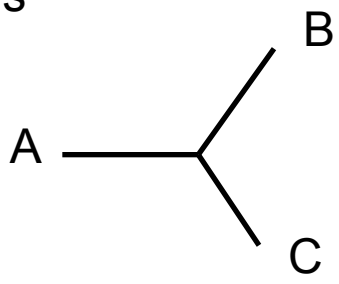
Если число листьев равно n, существует (2n-3)!! разных бинарных укоренных деревьев. (2n-3)!! – это нечто вроде факториала, но учитываются только четные числа.

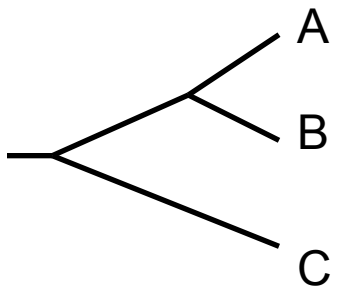
Существует (2n-5)!! разных бескорневых деревьев с n вершинами

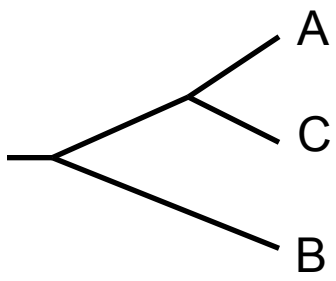
UNROOTED

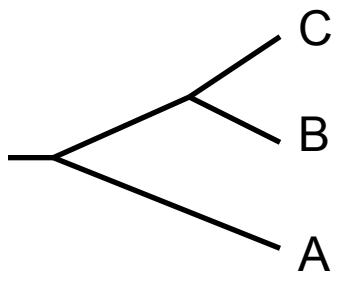
ROOTED

3 OTUs

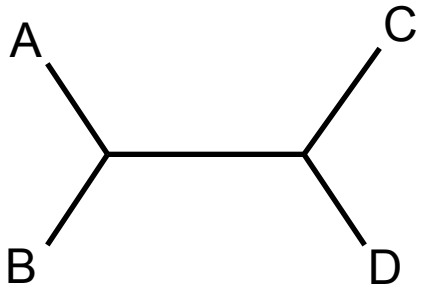


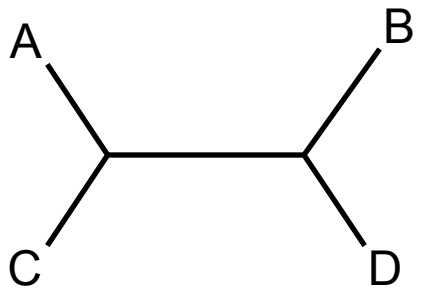


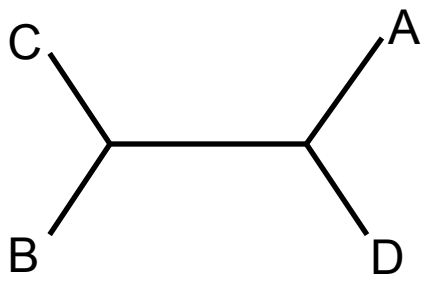


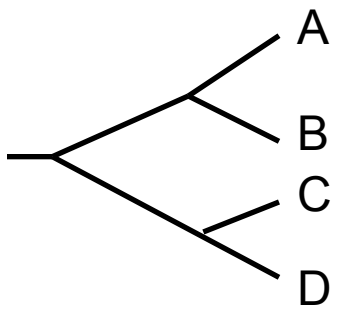


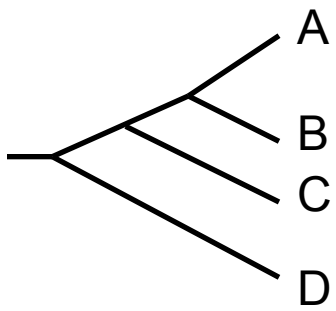
4 OTUs

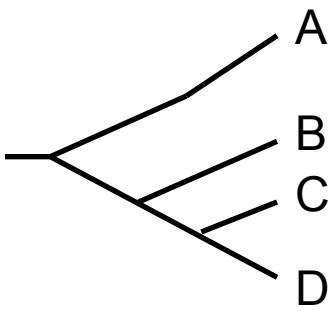


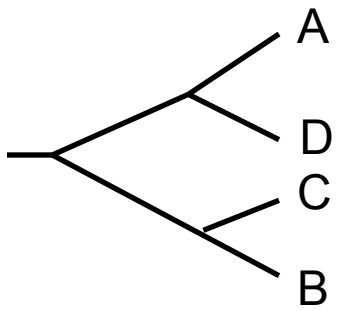










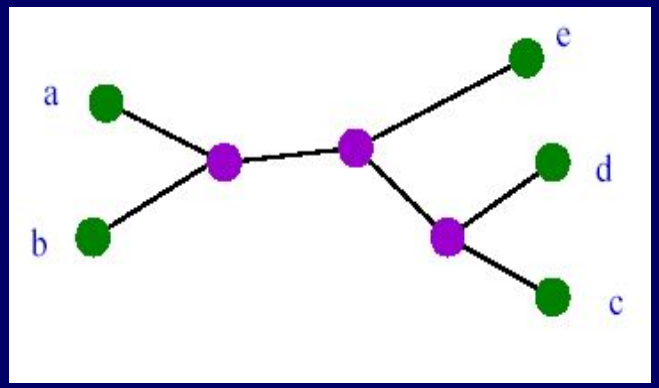


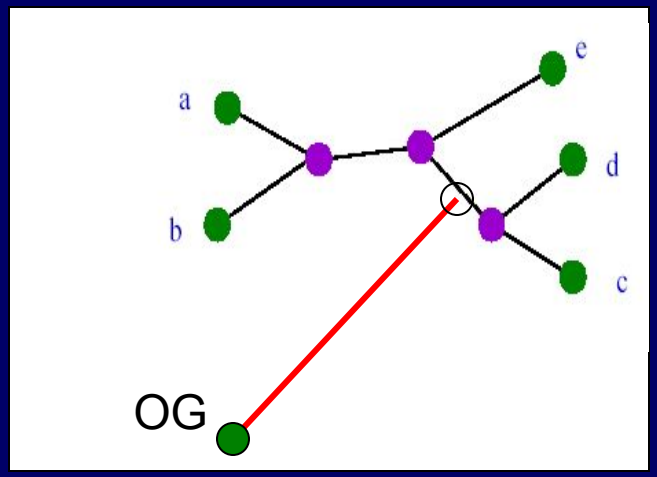
… 15 rooted trees of 4 OTUs

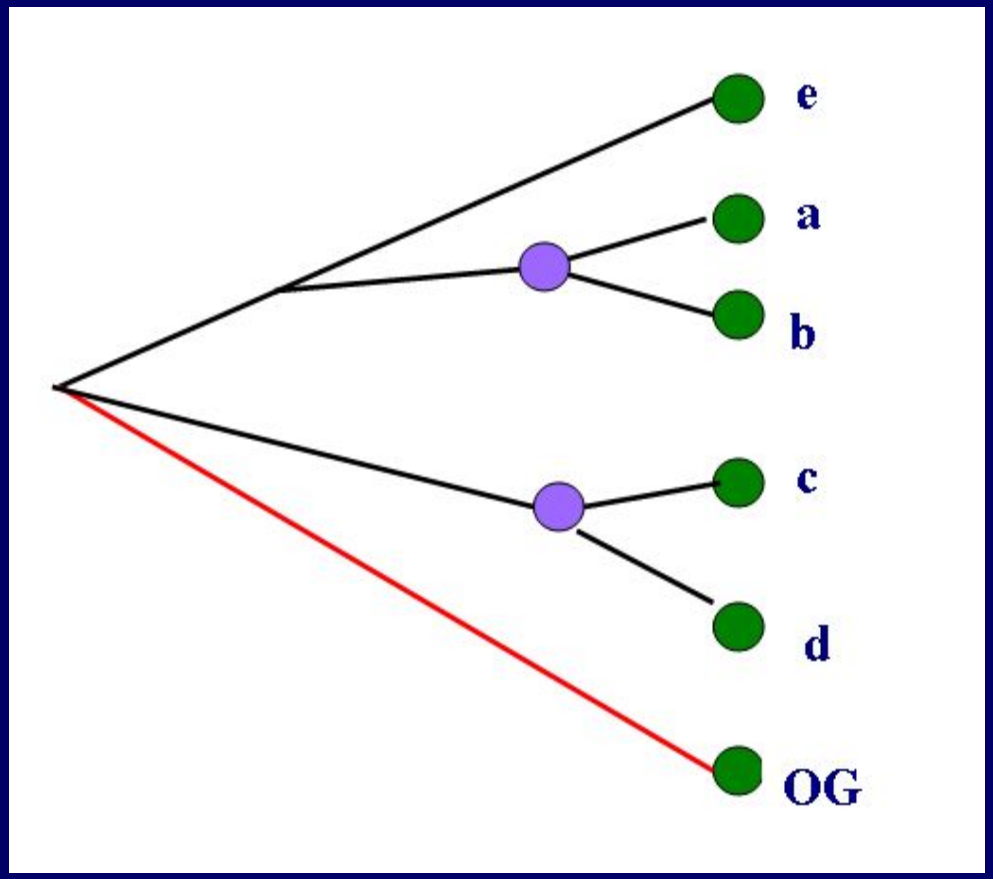
Искусственный способ укоренения деревьев

- **Бескорневое дерево можно «укоренить», если ввести внешнюю группу OTU (outgroup).**

  Внешния группа должна быть "старше", т.е. заведомо отделиться раньше, чем произошла дивергенция остальных OTU.







Какие бывают построенные деревья ?

Расстояние по дереву не то же самое, что эволюционное расстояние между данными

- **Ультраметрические деревья**

  Корневое дерево, в котором для любых листьев i и j расстояние D(i,j) – метка наименьшего общего предка i и j .

  В таком дереве все листья находятся на одинаковом от корня, что соответствует одинаковой скорости эволюции всех ветвей

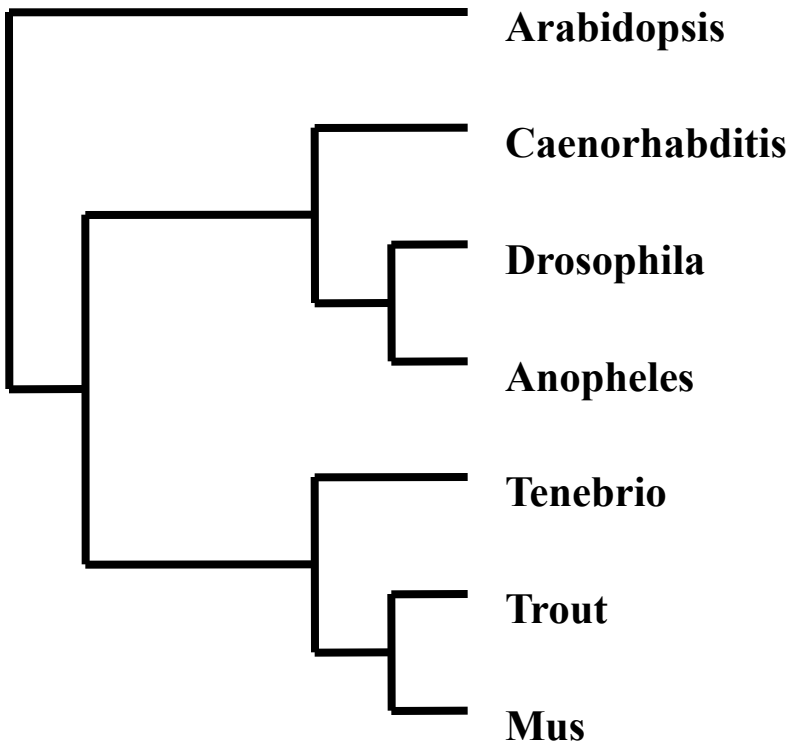
- **Аддитивные деревья**

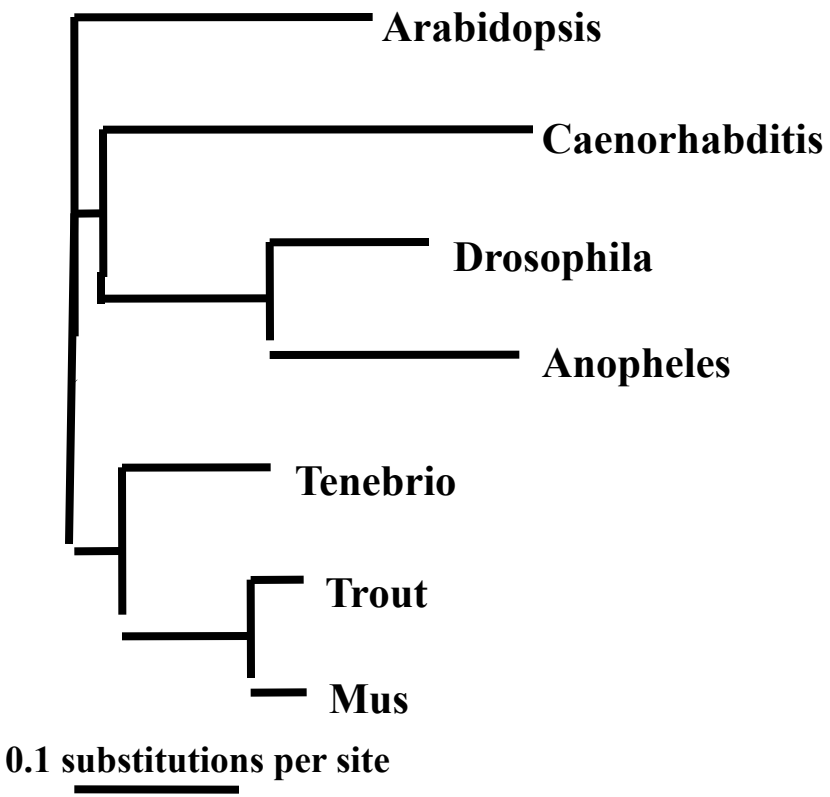
  Дерево, в котором для любых вершин i и j расстояние D(i,j) – это эволюционный путь от i к j . При этом расстояния от i и от j до их наименьшего общего предка могут сильно различаться.

- **Другие …**

Вообще говоря, строгое решение задачи построения аддитивного дерева невозможно
(следует из свойства задачи)

Как можно нарисовать построенное дерево?





Кладограмма:

представлена только топология, длина ребер игнорируется.

Филограмма:

Длина ребер пропорциональна эволюционному расстоянию между узлами.

Основные алгоритмы построения филогенетических деревьев

Методы, основанные на оценке расстояний (матричные методы):

Вычисляются эволюционные расстояния между всеми вершинами (OTUs) и строится дерево, в котором расстояния между вершинами наилучшим образом соответствуют матрице попарных расстояний.

- **UPGMA (Unweighted Pair Group with Arithmetic Mean)**
- **Ближайших соседей (Neighbor-joining, NJ)**

Символьно-ориентированные методы:

- **Наибольшего правдоподобия, Maximum likelihood, ML**

  Используется модель эволюции и строится дерево, которое наиболее правдоподобно при данной модели

- **Максимальной экономии (бережливости), maximum parsimony, MP**

  Выбирается дерево с минимальным количеством мутаций, необходимых для объяснения данных

Методы, основанные на оценке расстояний

- Дано:

  M – матрица *n* x *n*,

  где *Mij>=0* , *Mij* – эволюционное расстояние между листьями (OTU).

- Задача:

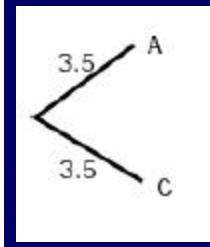
  Построить реберно взвешенное (an edge-weighted) дерево, где каждая вершина (лист) соответствует объекту из M , а расстояние, измеренное по дереву между вершинами (листьями) i and j соответствует Mij.

UPGMA

(алгоритм последовательной кластеризации)

- Выбираем 2 наиболее похожие вершины a, c.
- Строим новый узел k такой, что D(a,k)=D(b,k)=D(a,c)/2.
- Пересчитываем матрицу попарных расстояний :

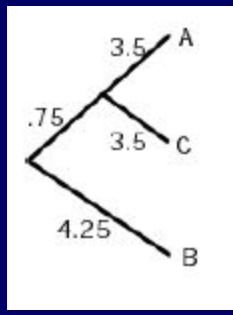
	A	B	C	D
A	0			
B	8	0		
C	7	9	0	
D	12	14	11	0



D(b, a or c) = [D(b,a) + D(b,c)] /2 = (8+9)/2=8.5
D(d, a or c) = [D(d,a) + D(d,c)] /2=(12+11)/2=11.5

- Повторяем процедуру….

	A or C	B	D
A or C	0		
B	8.5	0	
D	11.5	14	0

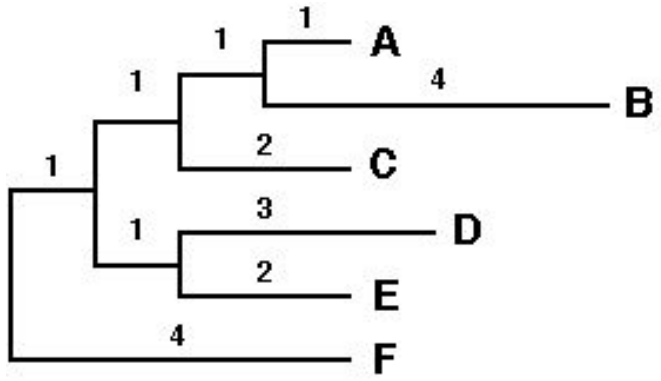


В конце концов получаем <u>единственное ультраметрическое укорененное</u> дерево

Не пользуйтесь UPGMA!

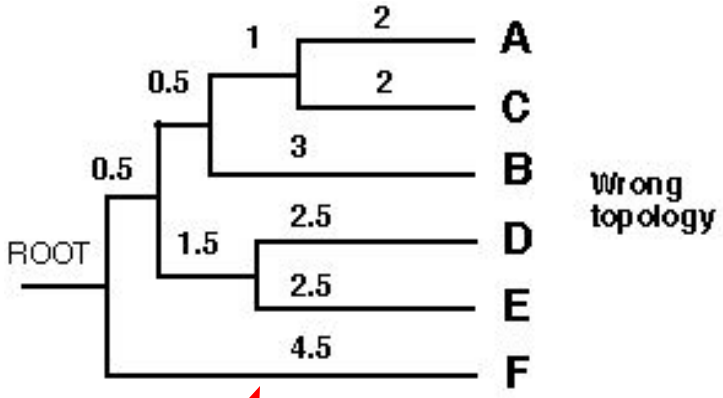
Алгоритм строит ультраметрическое дерево, а это означает, что скорость эволюции одинакова для всех ветвей дерева.

Использовать этот алгоритм имеет смысл только в случае ультраметрических данных (объектов эволюционирующих с одинаковой скоростью).



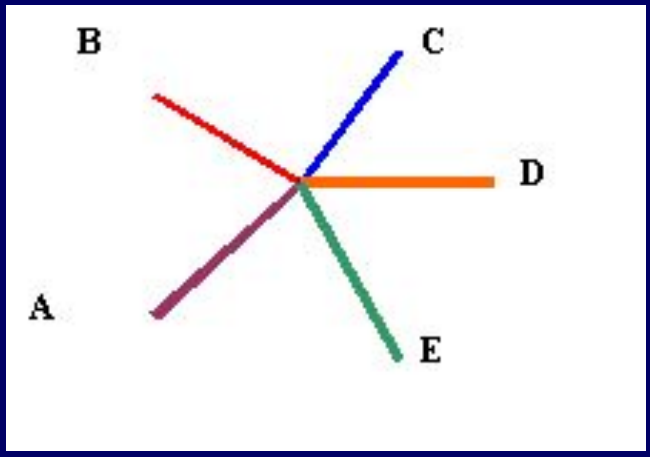
реальное с точки зрения эксперта дерево

	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



UPGMA

Метод ближайших соседей (Neighbor-joining, NJ)



1. Рисуем «звездное» дерево и будем "отщипывать" от него по паре вершин, рассмотрим все возможные пары вершины.

пусть $u_i = \sum_k \frac{M_{ik}}{n-2}$ - «среднее» расстояние до других вершин.

2. Выберем 2 вершины i и j с минимальным значением

Mij – ui –uj

т.е. выбираем 2 узла, которые близки друг к другу, но далеки ото всех остальных.

Метод ближайших соседей (Neighbor-joining, NJ)

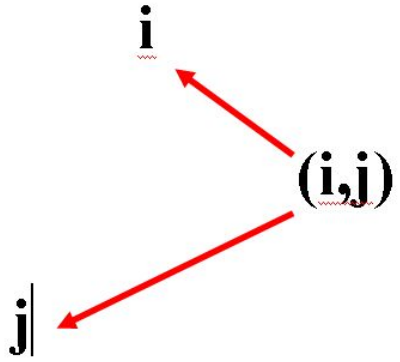
3. Кластер (i, j) – новый узел дерева

Расстояние от i или от j до узла (i,j):

$d_{i,(i,j)} = 0.5(M_{ij} + u_i - u_j)$

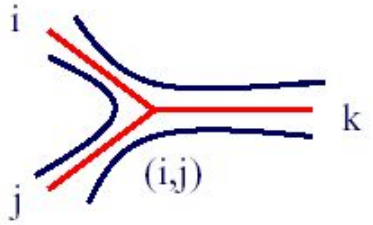
$d_{j,(i,j)} = 0.5(M_{ij} + u_j - u_i)$

т.е. длина ветви зависит от среднего расстояния до других вершин.



4. Вычисляем расстояние от нового кластера до всех других

$$M(ij)k = \frac{Mik + Mjk - Mij}{2}$$



5. В матрице M убираем i и j и добавляем (i, j).

Повторяем, пока не останутся 2 узла.....

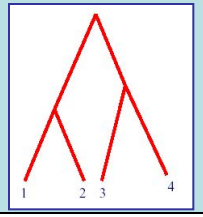
Метод ближайших соседей (Neighbor-joining, NJ)

- **Строит бескорневое аддитивное дерево**

- Может работать с большим количеством данных
- Достаточно быстрый алгоритм
- Хорошо зарекомендовал себя на практике: если есть недвусмысленное с точки зрения эксперта дерево, то оно будет построено.
- Используется при множественном выравнивании с помощью программы ClustalW
- Могут появиться ветви с длиной <0

Достоверность топологии. Bootstraps.

Есть множественное выравнивание и построенное по нему дерево.
Верим ли мы в топологию дерева?

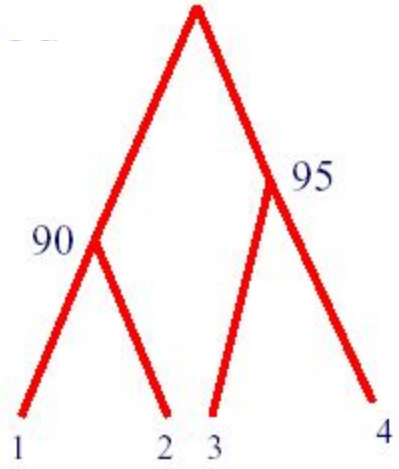


- Создадим псевдоданные:

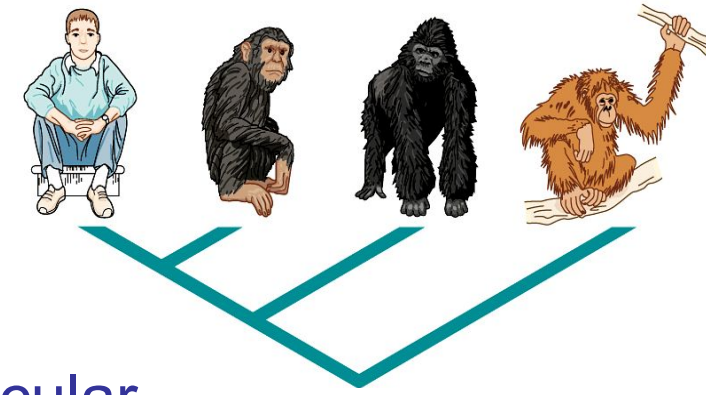
  N множественных выравниваний той же длины, что и исходное, каждое из псевдовыравниваний - случайный набор столбцов из исходного.

- Построим N деревьев:

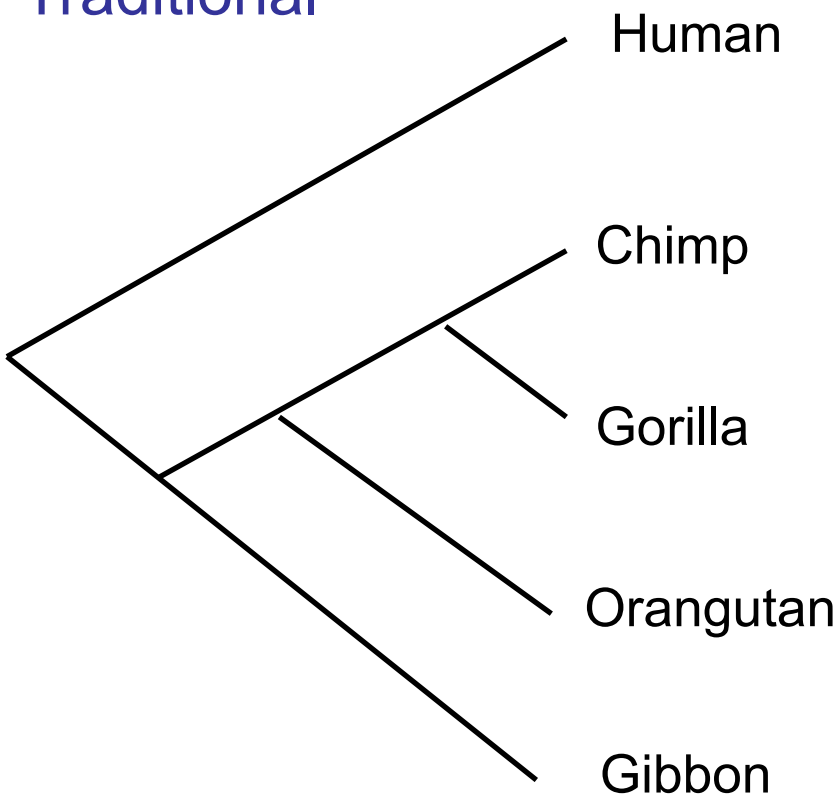
  на каждом внутреннем узле отметим доли случаев из N, в которых появлялся этот узел.



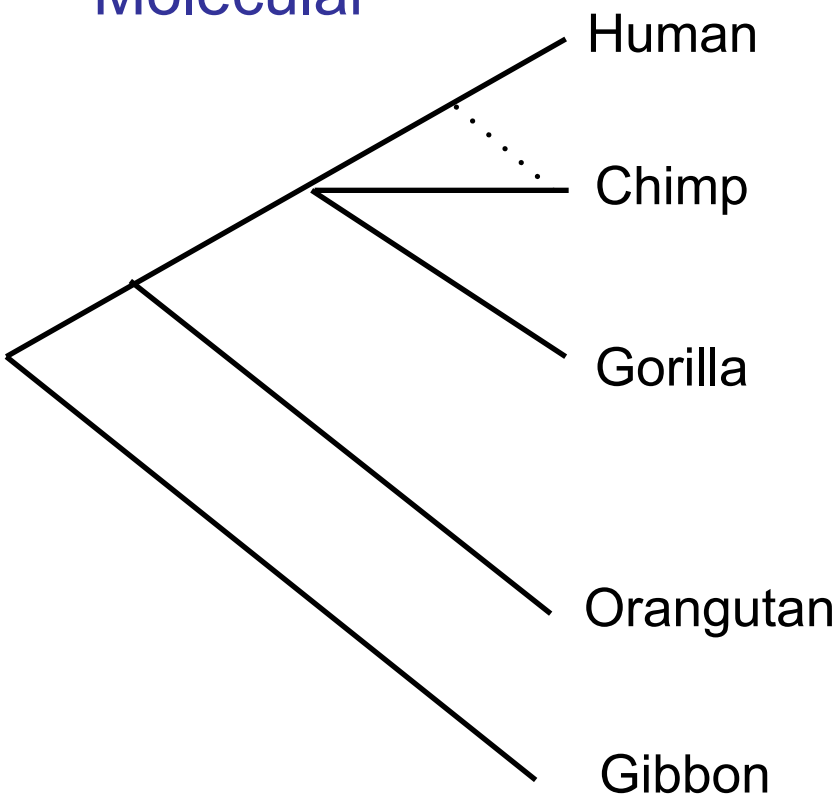
Обычно верят в топологию, если метки узлов на бутстрепном дереве больше 70-80% . Если меньше 30%, то не верим. В иных случаях – думаем…



Traditional



Molecular



Trees

plagiarized by Chuck Staben, 1998
Sergeant Joyce Kilmer, 1914

I think that I won't ever see
A really correct phylogeny.
A phylogeny whose root can rest
With truth and beauty ne'er stressed.
A tree that brings to our mind
What biology says that we should find.
A tree that even in this class
Puts bamboo correctly with other grass.
Upon this tree the truth may lie
Or, with exhaustive search, may die.
Phylogenies are made by fools like me,
But only God can make a tree.