

Филогенетические деревья

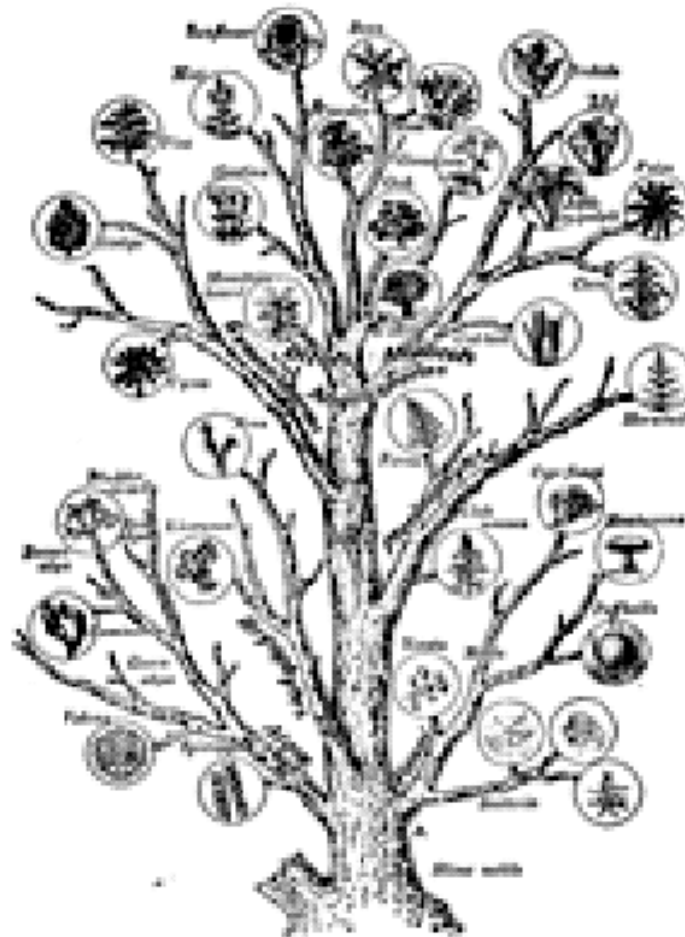


FIG. 121. Genealogical tree of plant life

This diagram is intended to suggest the common origin of all plant forms, with the constant progressive departure from elevated types, now in one direction and now in another, like the branching of a tree. Lower and higher mean nearer to or farther from the original types. The closer together two forms are on a given branch, the more closely related they are considered (cf. Fig. 122)

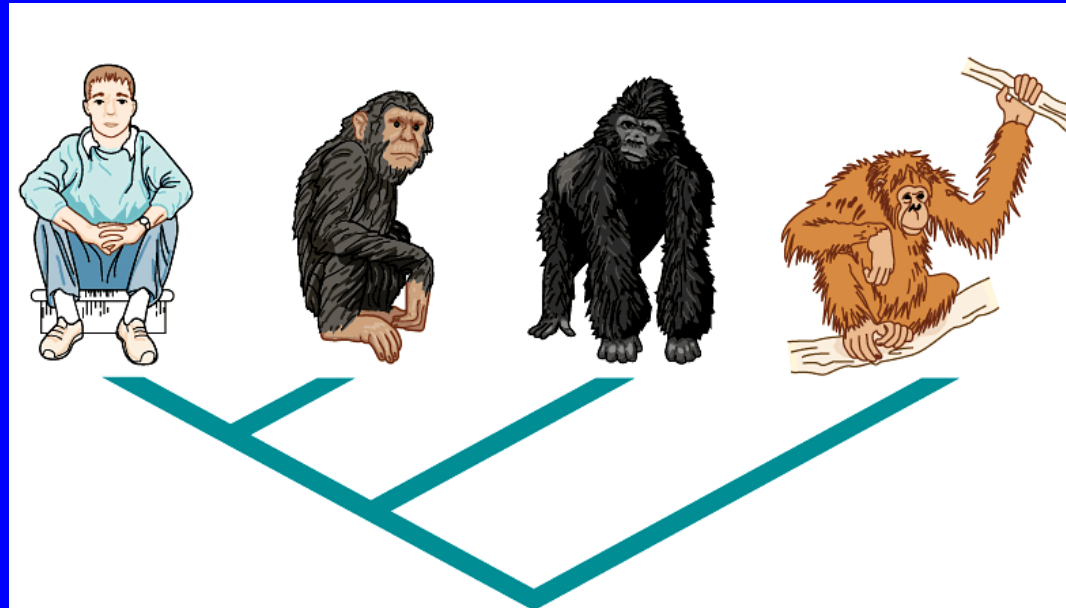
Что такое филогенетическое дерево?

- **Филогения** - раздел биологии, изучающий родственные взаимоотношения разных групп живых организмов. Филогению отображается обычно в виде "эволюционных древ" или систематических названий.
- **Филогенетика (=молекулярная филогенетика)** – те же взаимоотношения, но на уровне отдельных белковых (генных) семейств

Зачем нужны филогенетические деревья?

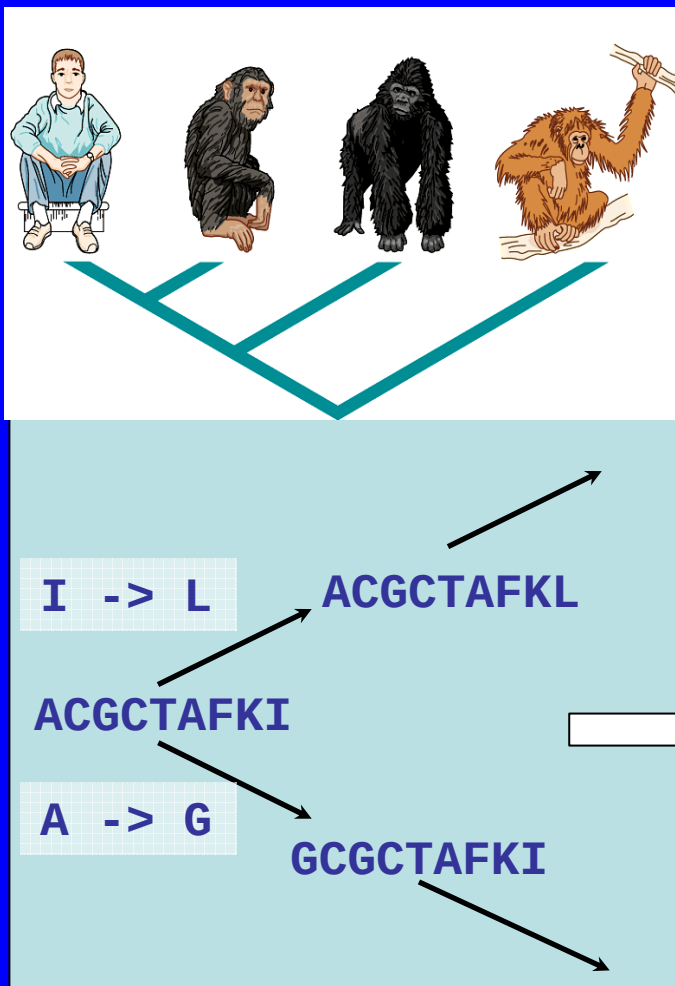
Биологические задачи:

- **сравнение 3-х и более объектов**
(кто на кого более похож)
- **реконструкция эволюции**
(кто от кого, как и когда произошел...)



Реальные события :

эволюция в природе или в лаборатории,
компьютерная симуляция



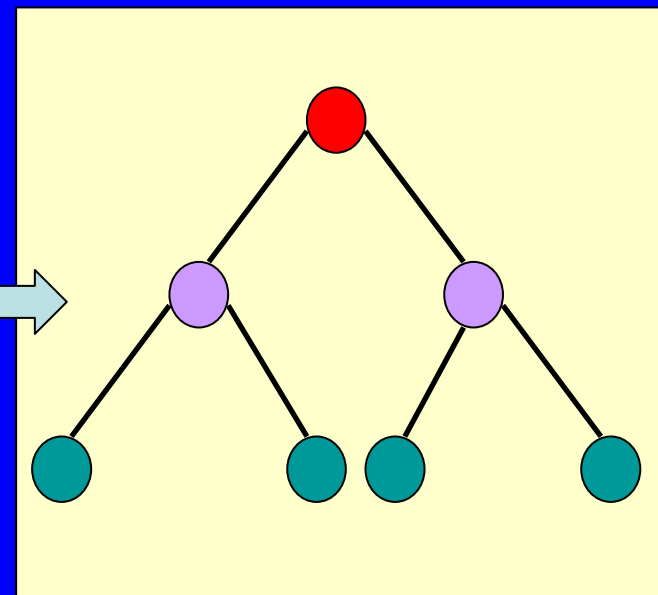
Данные:

например,
а.к. последовательности или
количество щетинок

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

Построенное дерево

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



Основные термины

Узел (node) — точка разделения предковой последовательности (вида, популяции) на две независимо эволюционирующие. Соответствует внутренней вершине графа, изображающего эволюцию.

Лист (leaf, OTU — оперативная таксономическая единица) — реальный (современный) объект; внешняя вершина графа.

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

Корень (root) — гипотетический общий предок.

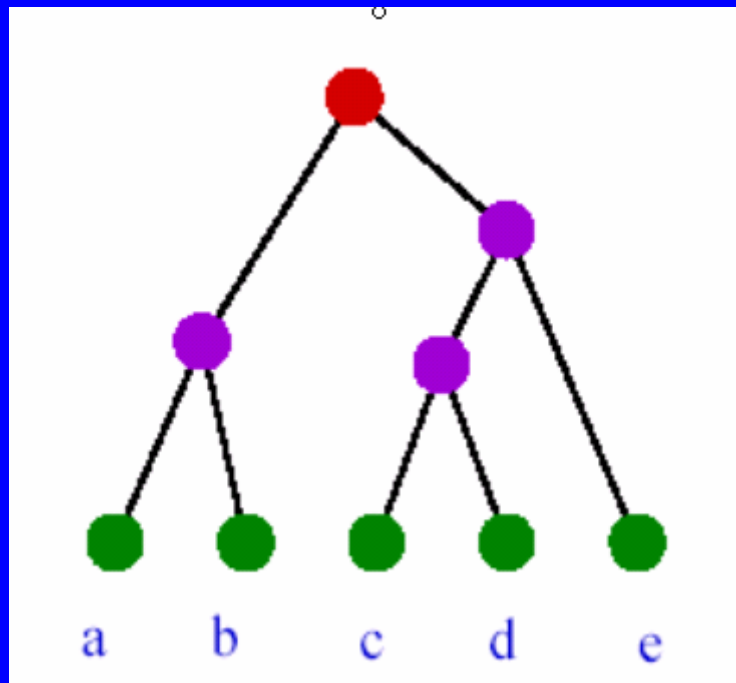
Клада (clade) - группа двух или более таксонов или последовательностей ДНК, которая включает как своего общего предка, так и всех его потомков.



Какие бывают деревья?

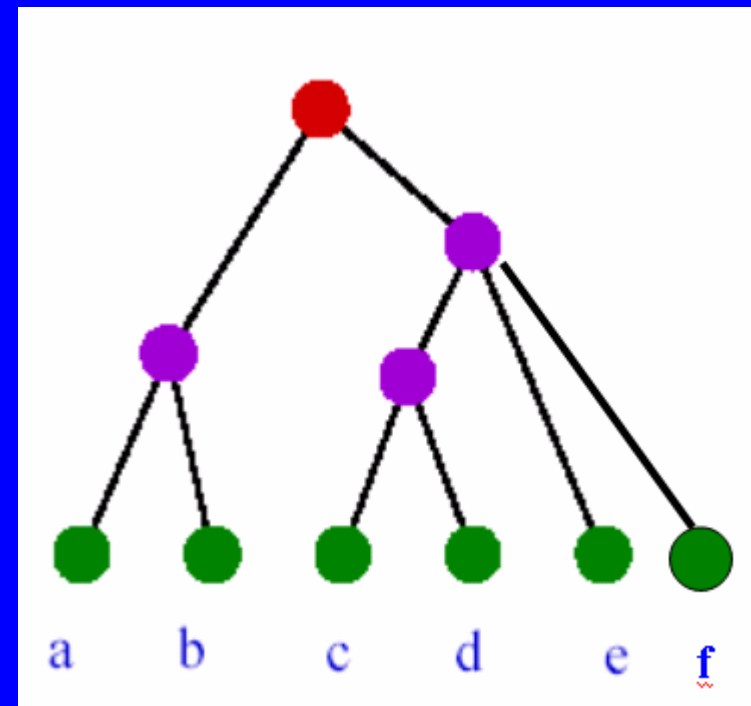
Бинарное (разрешённое)

(в один момент времени может произойти только одно событие)



Небинарное (неразрешённое)

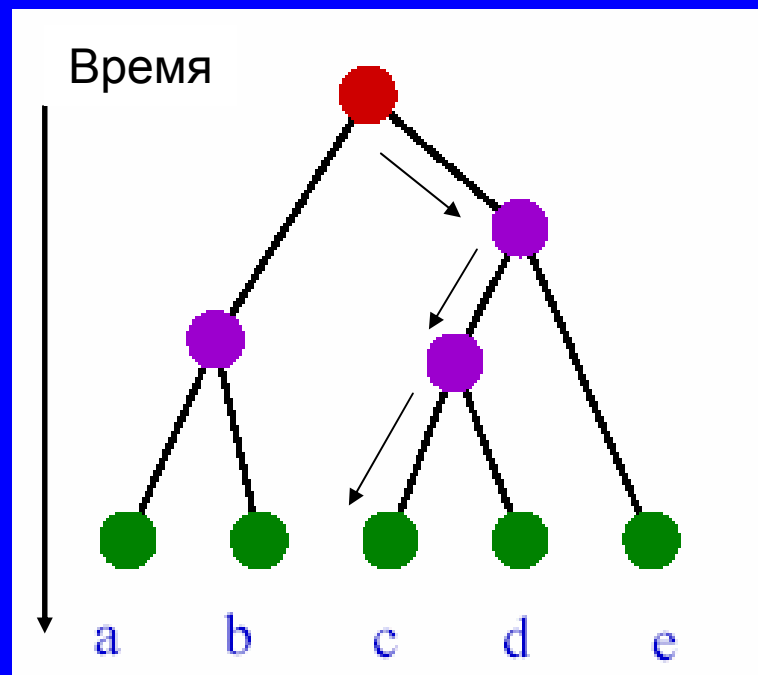
(может ли в один момент времени произойти два события?)



Время

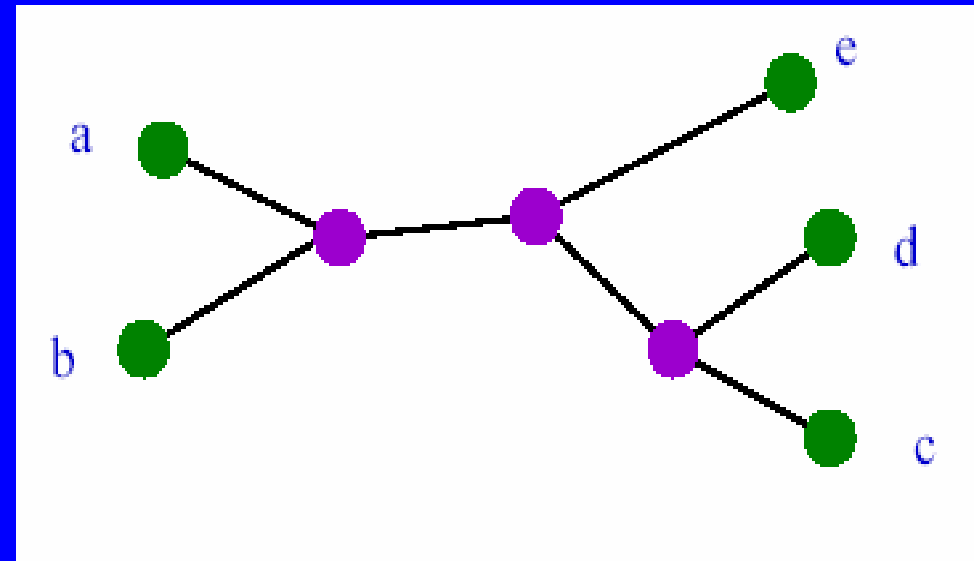
Какие бывают деревья?

Укорененное дерево (rooted tree)
отражает направление эволюции



Если число листьев равно n , существует $(2n-3)!!$ разных бинарных укоренных деревьев.
По определению, $(2n-3)!! = 1 \cdot 3 \cdot \dots \cdot (2n-3)$

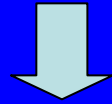
Неукорененное (бескорневое) дерево
(unrooted tree) показывает
только связи между узлами



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

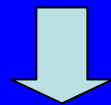
Рутинная процедура, или как строят деревья?

Составление выборки последовательностей



Множественное выравнивание

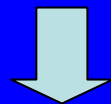
f53969	ALLKTPAEFDAYELNSSIKGAGTDEACLI EILSSRSNAEI
con101	AMLKTPSQYDAYELKEAIKGAGTDEACLI EILASRSNAEI
hum4	GMMTPTVLYDVQELRPAMKGAGTDEGCLIEILASRTPEEI



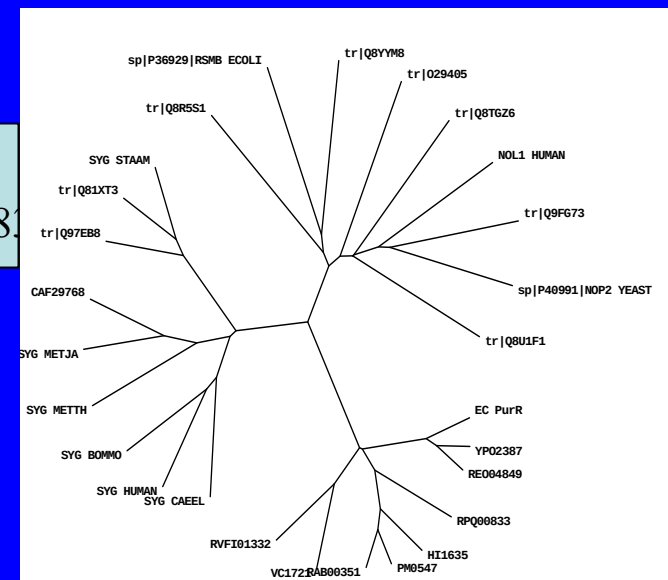
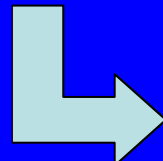
Построение дерева

фрагмент записи в виде скобочной формулы:

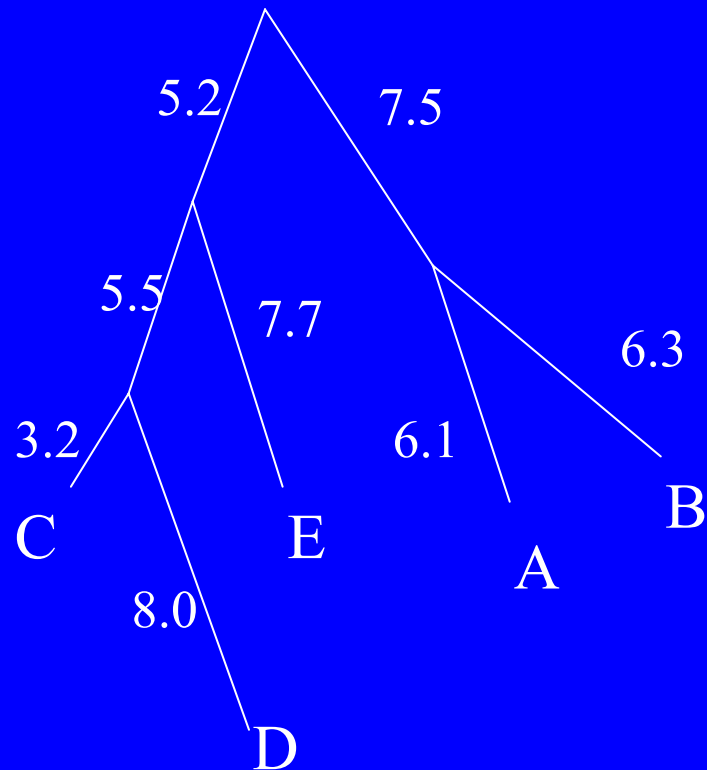
(((((con101:38.51018,(f53969:28.26973,((f67220:8.39851,
max4:27.50591):4.92893,con92:30.19677):13.62315):9.53075):25.8



Визуализация и редакция дерева



Скобочная формула (Newick format)



`(( (C,D),E), (A,B));` ТОЛЬКО ТОПОЛОГИЯ

`(( (C:3.2,D:8.0):5.5,E:7.7):5.2, (A:6.1,B:6.3):7.5);`

длины ветвей

Как выбирать последовательности для дерева?

- ✓ Кроме случаев очень близких последовательностей, проще работать с белками (а не с ДНК)
- ✓ Придерживайтесь небольшой выборки (< 50 последовательностей)
- ✓ Избегайте:
 - фрагментов;
 - ксенологов;
 - рекомбинантных последовательностей;
 - многодоменных белков и повторов
- ✓ Используйте outgroup (последовательность, ответвившаяся от общего предка заведомо (но минимально!) раньше разделения интересующих групп-клад)

Самое главное – хорошее выравнивание!

- ✓ Максимальный вклад в финальное дерево: нельзя построить хорошее дерево по плохому выравниванию
- ✓ Блоки, содержащие много гэпов, плохо выровненные N- и C- концы можно просто вырезать.

Основные алгоритмы построения филогенетических деревьев



Методы, основанные на оценке расстояний (матричные методы):

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

- **UPGMA**
- **Neighbor-joining**
- Минимальная эволюция
- Квартеты («топологический»)
- ...



Наибольшего правдоподобия, **Maximal likelihood, ML**

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



Максимальной экономии (бережливости), **maximal parsimony, MP**

Выбирается дерево с минимальным количеством мутаций, необходимых для объяснения данных

Пример матрицы расстояний

1	2	3	4	5	6	7	8	
0.00	10.53	9.77	12.78	12.03	16.54	13.53	25.00	HUMAN 1
	0.00	9.02	12.03	9.77	15.79	9.02	27.27	HORSE 2
		0.00	9.77	9.02	16.54	12.03	24.24	RABIT 3
			0.00	2.26	17.29	10.53	25.76	MOUSE 4
				0.00	15.79	8.27	25.76	RAT 5
					0.00	10.53	29.55	BOVIN 6
						0.00	25.00	PIG 7
							0.00	CHICK 8

Расстояние (уровень дивергенции) между
соответствующими последовательностями из
геномов мыши и свиньи

Как понимать расстояние между объектами?

- Как время, в течение которого они эволюционировали
- Как число «эволюционных событий» (мутаций)

В первом случае объекты образуют

ультраметрическое пространство

(если все объекты наблюдаются в одно время, что, как правило, верно)

Но время непосредственно измерить невозможно

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

За равное время во всех ветвях эволюции 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.

Укоренённое дерево называется ультраметрическим, если расстояние от корня до любого из листьев одинаково.

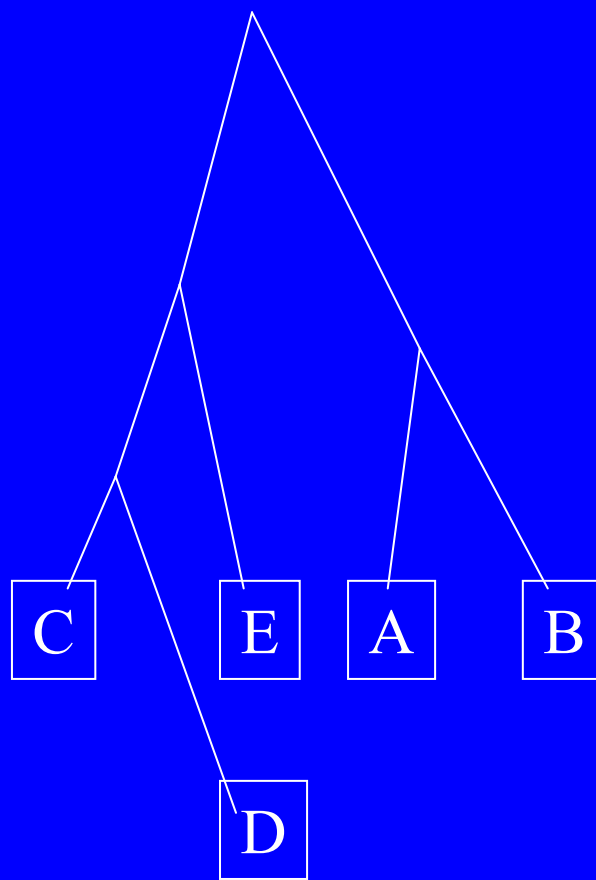
UPGMA

Unweighted Pair Group Method with Arithmetic Mean

разновидность кластерного метода

Расстояние между кластерами вычисляется как среднее арифметическое всевозможных расстояний между последовательностями из кластеров

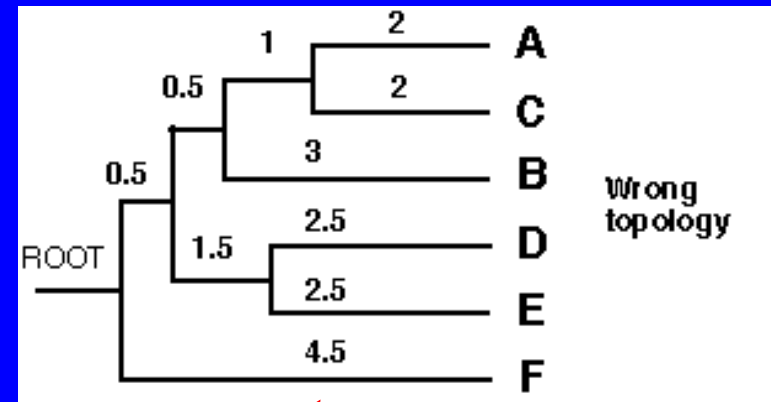
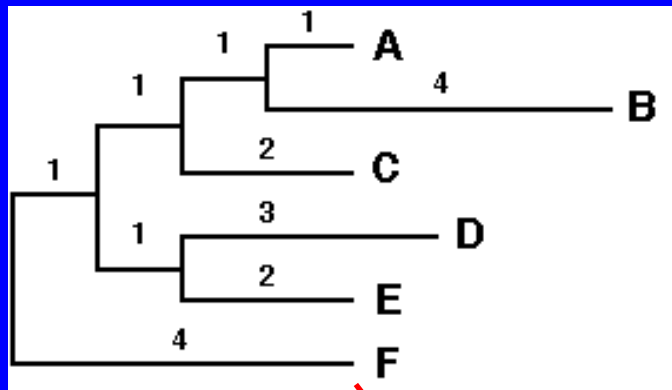
Гипотеза молекулярных часов не всегда справедлива



(длина ветвей пропорциональна числу мутаций)

Недостатки UPGMA

Алгоритм строит ультраметрическое дерево, а это означает, что скорость эволюции предполагается одинаковой для всех ветвей дерева. Использовать этот алгоритм имеет смысл только в случае ультраметрических данных (справедливости «молекулярных часов»).



Реальное дерево

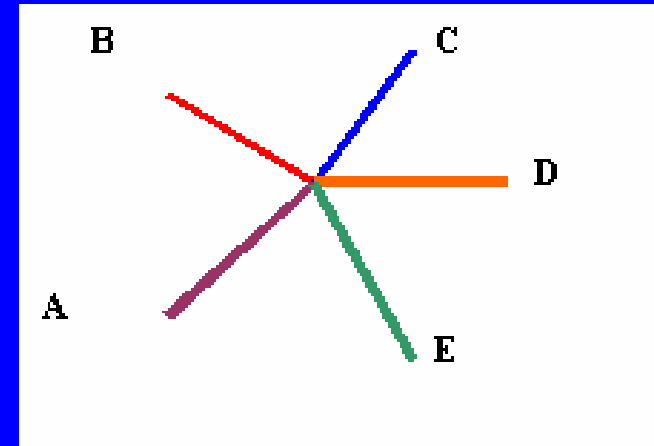
	A	B	C	D	E
B	5				
C	4	7			
D	7	10	7		
E	6	9	6	5	
F	8	11	8	9	8

UPGMA

Метод ближайших соседей (Neighbor-joining, NJ)

- ✓ Строит неукоренённое дерево
- ✓ Может работать с большим количеством данных
- ✓ Достаточно быстрый
- ✓ Хорошо зарекомендовал себя на практике: если есть недвусмысленное с точки зрения эксперта дерево, то оно будет построено.
- ✓ Могут появиться ветви с длиной <0

Метод Neighbor-joining



Рисуем «звездное» дерево и будем «отщипывать» от него по паре листьев

Пусть $u_i = \sum_k M_{ik}/(n-2)$ — среднее расстояние от листа i до других листьев

1. Рассмотрим все возможные пары листьев. Выберем 2 листа i и j с минимальным значением величины

$$M_{ij} - u_i - u_j$$

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

Метод ближайших соседей (Neighbor-joining, NJ)

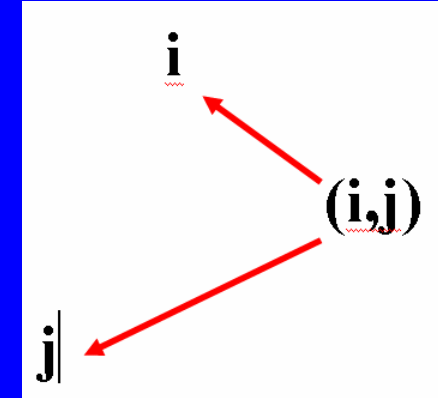
2. Кластер (i, j) – новый узел дерева

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

$$D(i, (i,j)) = 0,5 \cdot (M_{ij} + u_i - u_j)$$

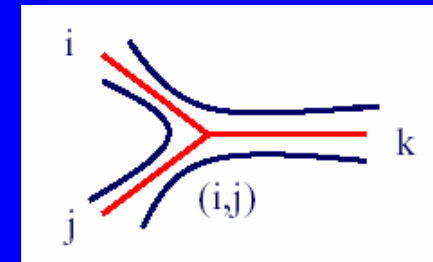
$$D(j, (i,j)) = 0,5 \cdot (M_{ij} + u_j - u_i)$$

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



3. Вычисляем расстояние от нового кластера до всех других

$$M(ij)k = \frac{M_{ik} + M_{jk} - M_{ij}}{2}$$



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

Повторяем, пока не останутся 3 узла ...

Стандартная ситуация

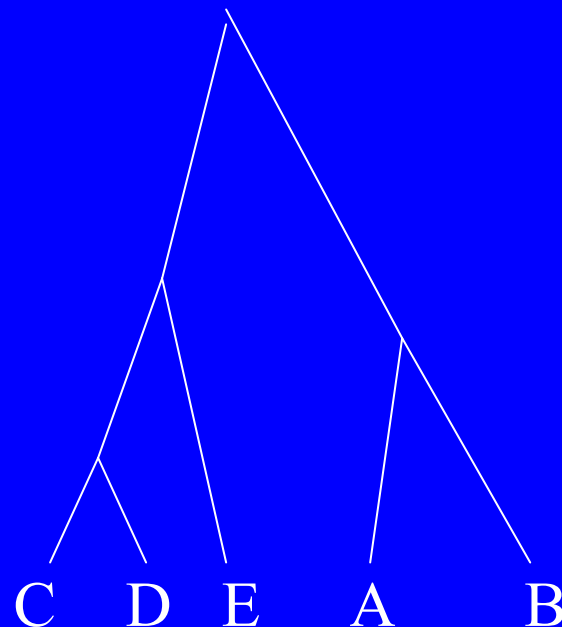
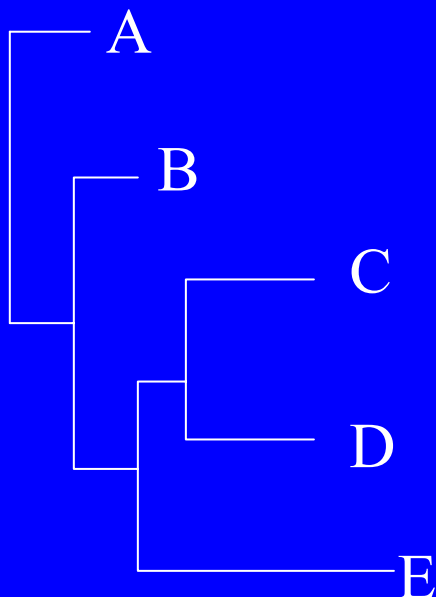
- ✓ Понимаем расстояние как число мутаций
- ✓ Реальное (неизвестное нам) дерево — укоренённое, но не ультраметрическое
- ✓ Мы реконструируем неукоренённое дерево (топологию и длины ветвей). Его надо понимать как **множество** всех возможных укоренений.

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

Как изобразить дерево?

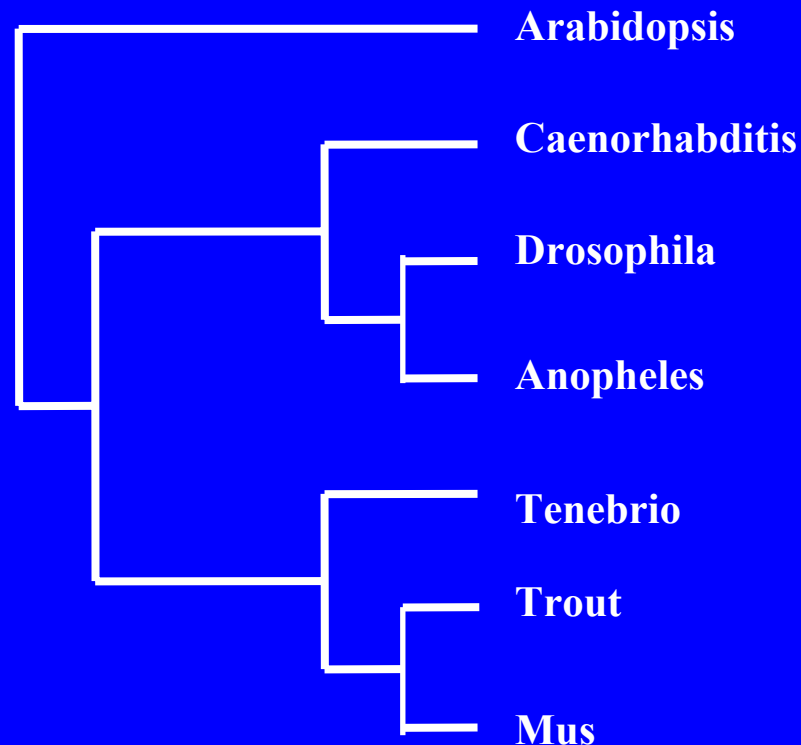
Топология дерева

Топология дерева — только листья, узлы, (корень)
и связывающие их ветви
(топология не зависит от способа изображения дерева)



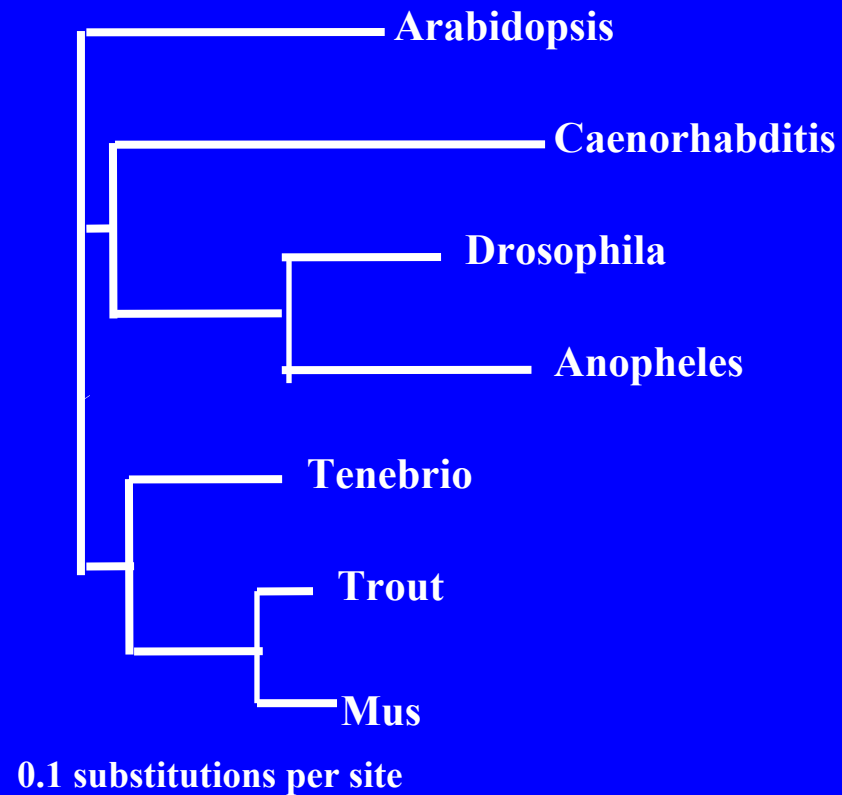
Два изображения одной и той же топологии

Как можно нарисовать построенное дерево?



Кладограмма:

представлена только топология,
длина ребер игнорируется.



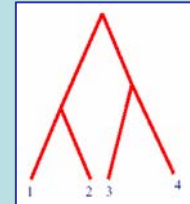
Филограмма:

Длина ребер пропорциональна
эволюционному расстоянию
между узлами.

Достоверность топологии. Bootstraps

Есть множественное выравнивание и построенное по нему дерево.

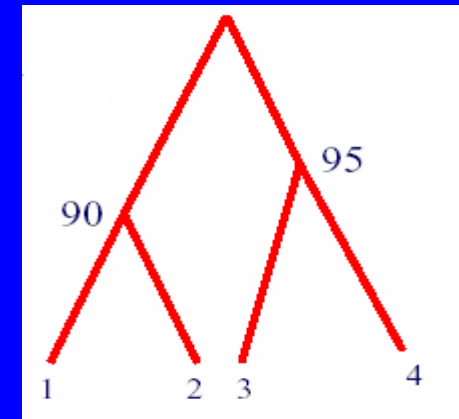
Верим ли мы в топологию дерева?



- Создадим псевдоданные:

N множественных выравниваний той же длины, что и исходное, каждое из псевдовыравниваний - случайный набор столбцов из исходного (выборка с возвращением!)

- Построим N деревьев:
на каждой внутренней ветви отметим долю случаев из N , в которых появлялся этот узел.



Обычно верят в топологию, если метки ветвей на бутстрепном дереве больше 70-80% . Если меньше 50%, то не верим. В иных случаях — думаем...

Какие on-line программы строят деревья?

- ✓ ClustalW. “Tree type” – nj, phylip: строит только методом NJ, но результат – в разных форматах, no bootstraps
- ✓ Phylip (Felsenstein, 1993) – пакет программ для построения филогенетических деревьев (stand-alone)
On-line (partly): например,
<http://bioweb.pasteur.fr/seqanal/phylogeny/phylip-uk.html>
- ✓ PAUP (Phylogenetic Analysis Using Parsimony)