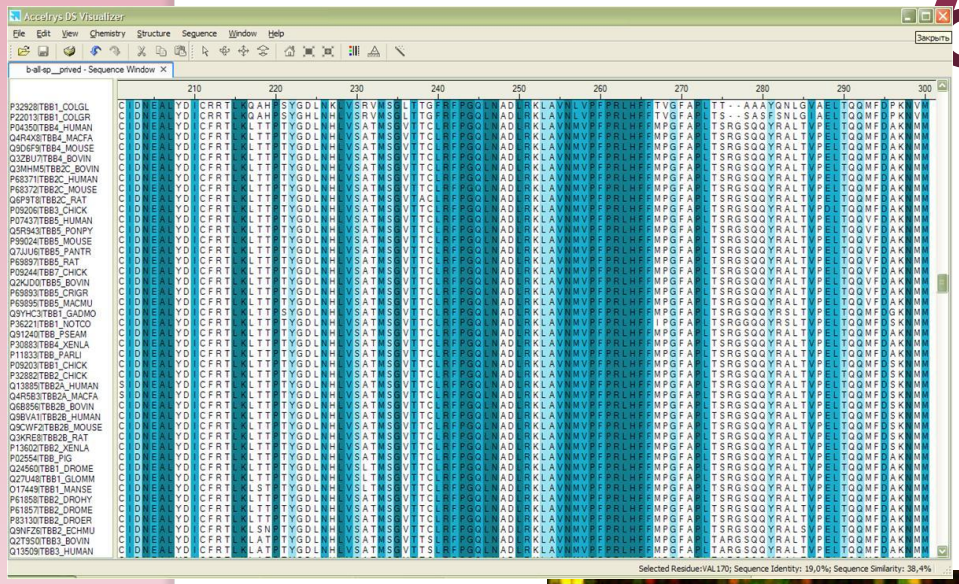
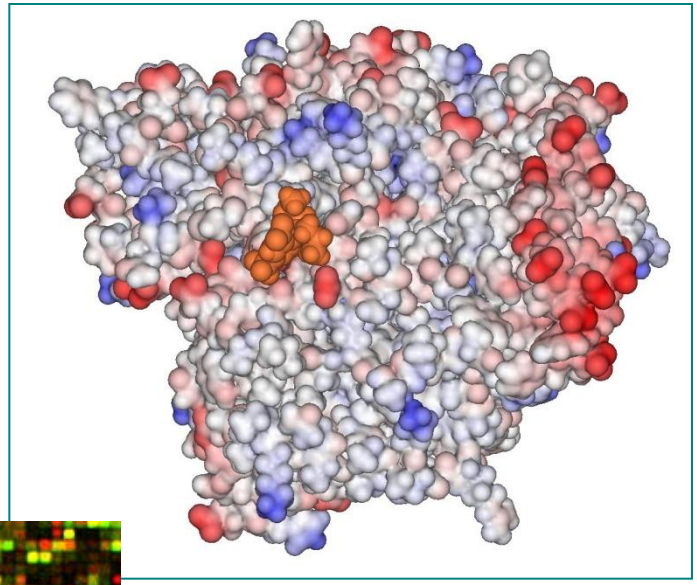
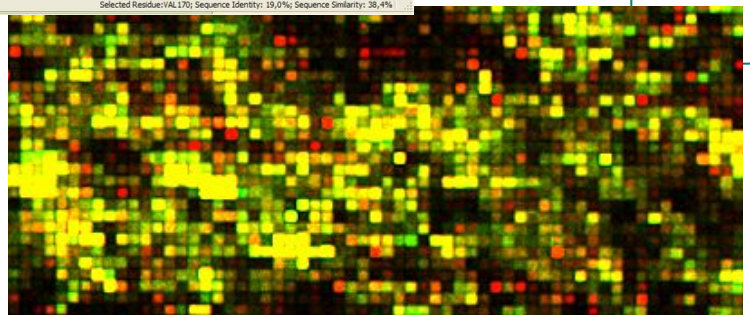
Київський національний університет ім. Тараса Шевченка
Інститут високих технологій

БІОІНФОРМАТИК

к.б.н. Нилорко О.А.







Дендрограммы

Молекулярная филогения

Графы и деревья

Граф — это простая диаграмма (абстрактная структура), применяемая для представления отношений между элементами например чисел, объектов или мест. Сами элементы изображают в виде узлов, а отношения между ними показывают в виде связей, или ребер (соединительных линий).

В теории графов деревом называют граф особого вида. Граф представляет собой структуру, состоящую из узлов (абстрактных точек) и соединяющих их ребер (линий между точками). Путь от одного узла к другому складывается из множества последовательных ребер, первое из которых выходит из начальной точки (узла), а последнее входит в конечную точку (узел). Граф называют связным, если в нем между любыми двумя узлами можно провести по крайней мере один путь.

Деревом называют связный ациклический граф, между каждыми двумя точками которого имеется строго один путь.

Терминология

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.

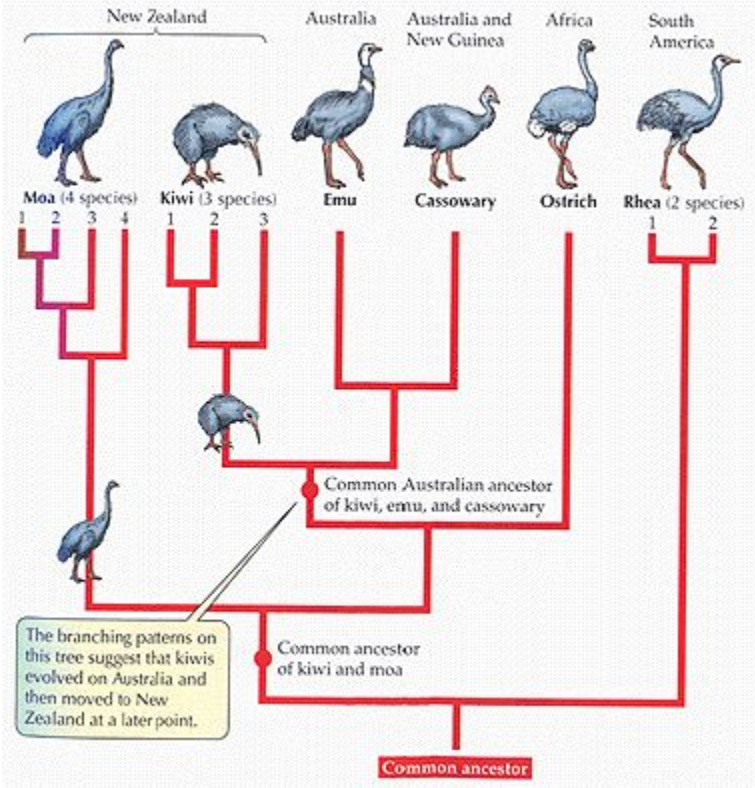
Корень (root) — общий предок.

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

Зачем нужны деревья?

<u>Биологические задачи</u>:

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



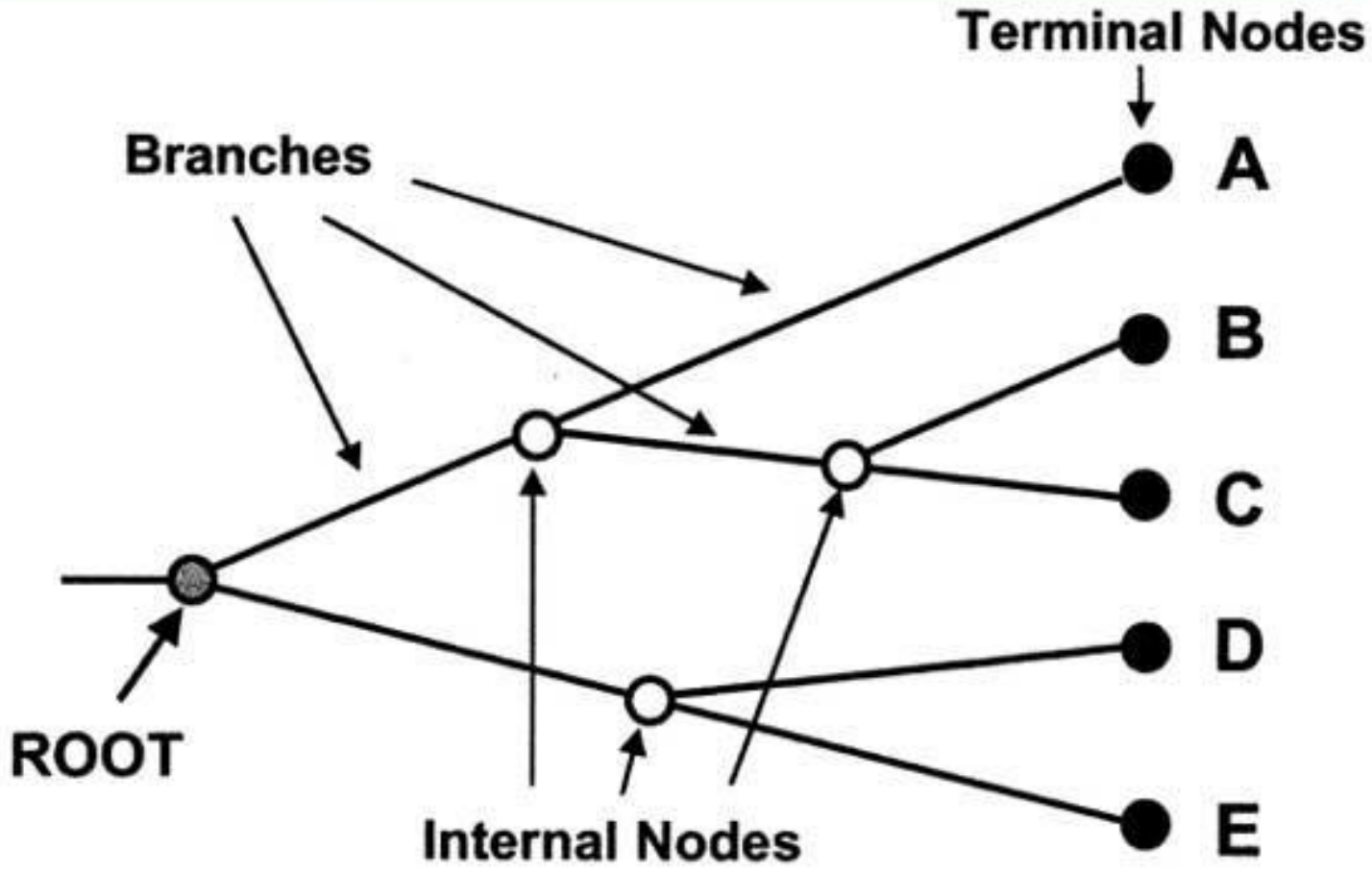
Филогенетическое дерево (древо)

Филогения - раздел биологии, изучающий родственные взаимоотношения разных групп живых организмов.

Молекулярная филогения -

Древо сходства и филогенетическое древо – не одно и то же!!!

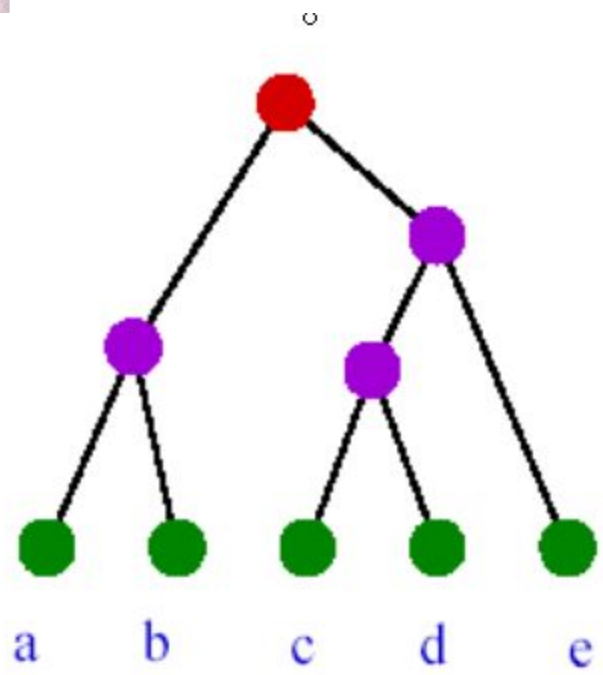
Phylogenetic Tree Terminology



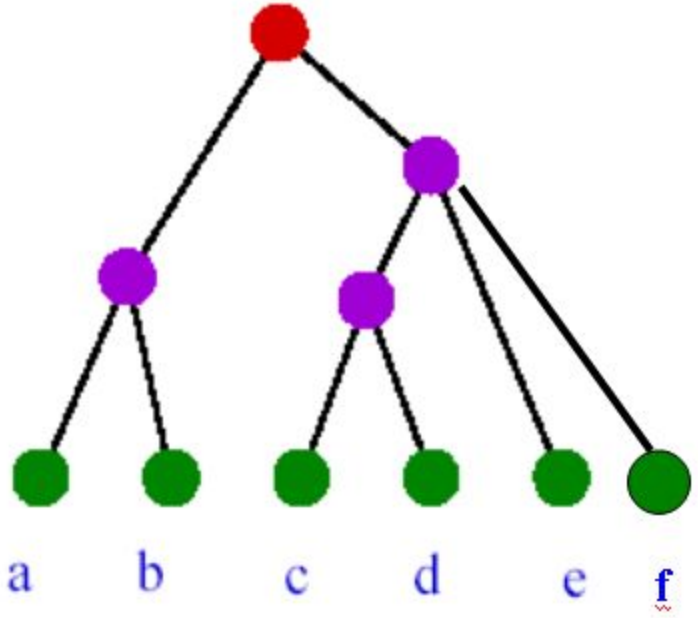
HTU (hypothetical taxonomic unit)

Какие бывают деревья?

Бинарное (разрешённое)
(в один момент времени может произойти только одно событие)



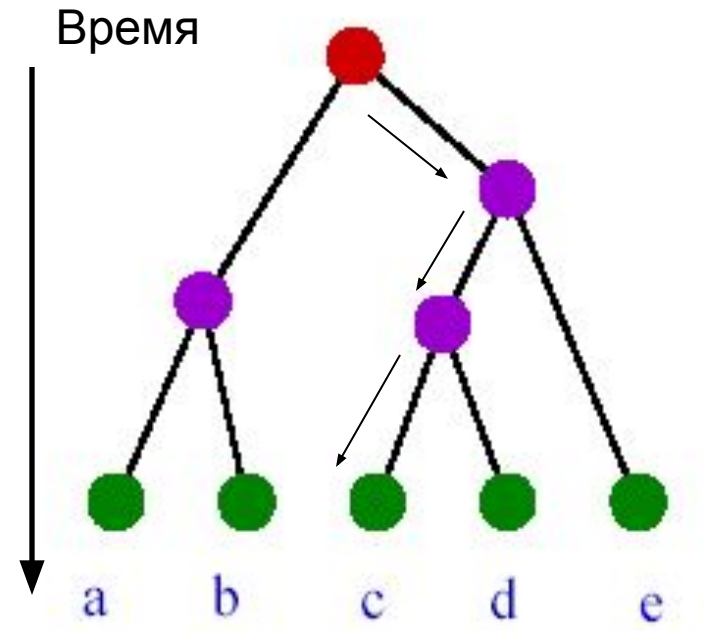
Небинарное (неразрешённое)
(может ли в один момент времени произойти два события?)



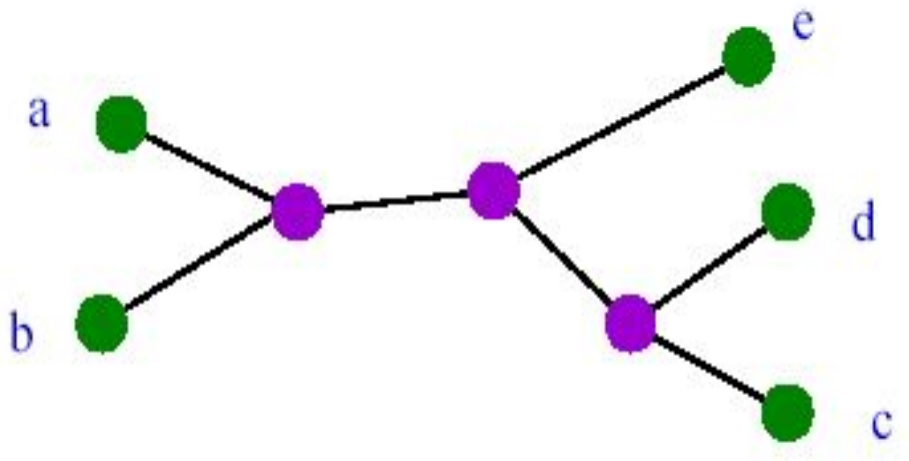
Вре

Какие бывают деревья?

Укорененное дерево (rooted tree)
отражает направление эволюции



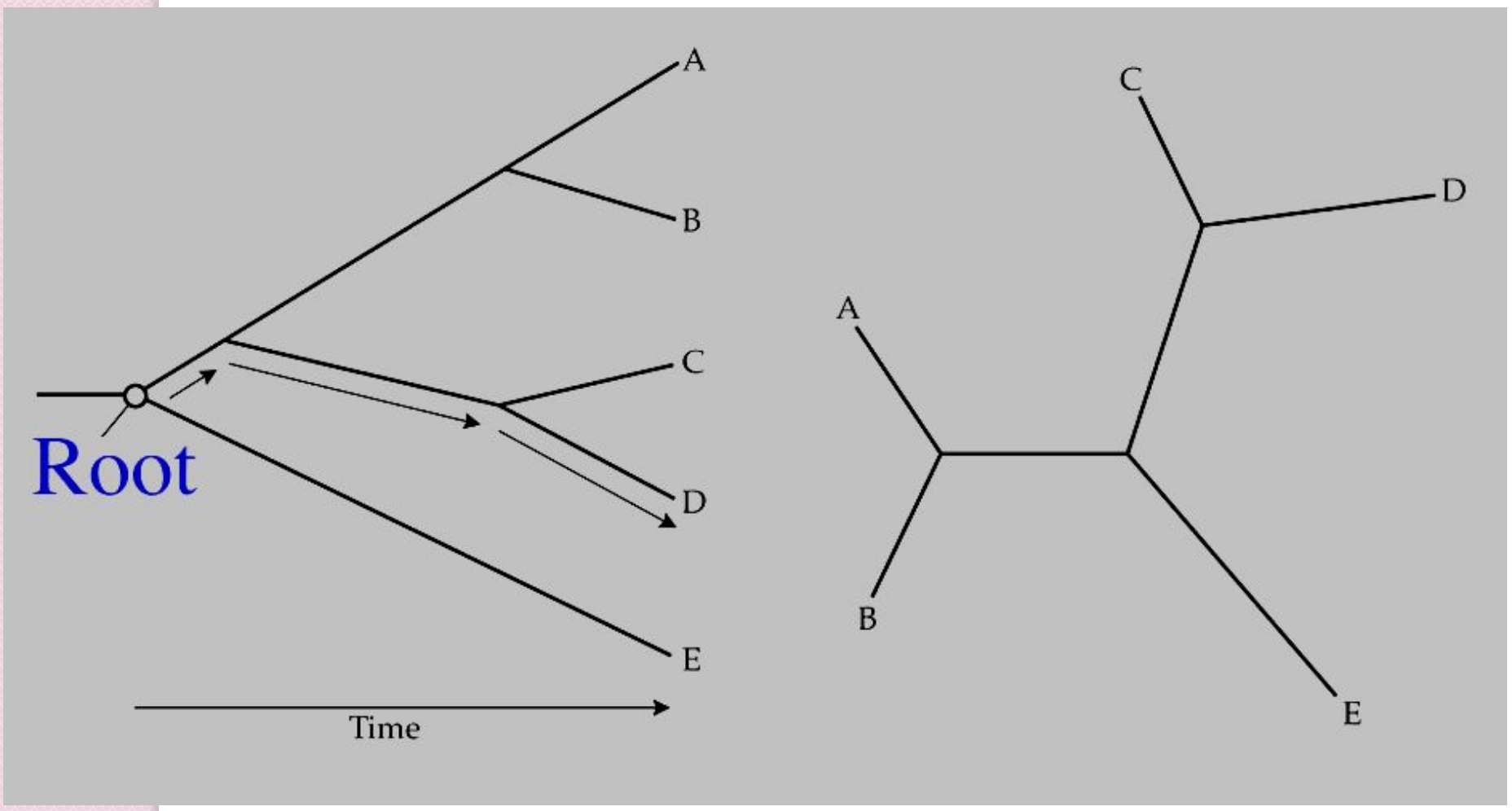
Неукорененное (бескорневое) дерево (unrooted tree) показывает только связи между узлами



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

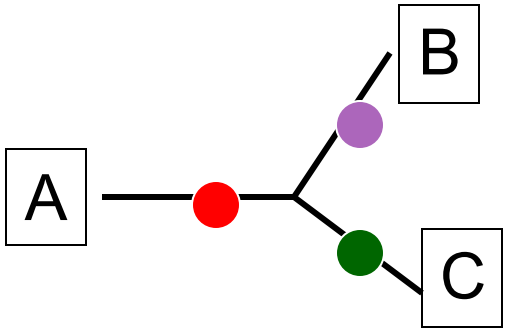
Существует (2n-5)!! разных бескорневых деревьев с n листьями

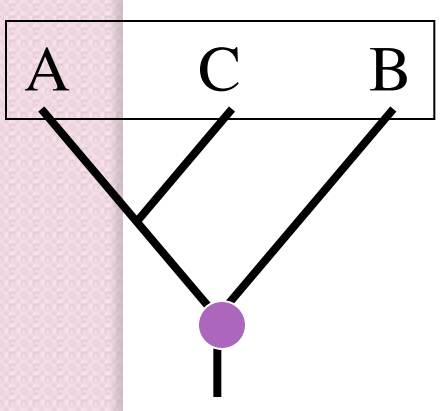
Rooting

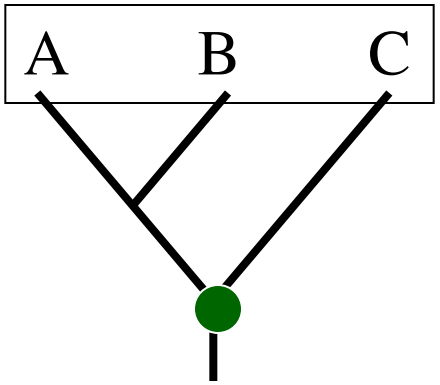


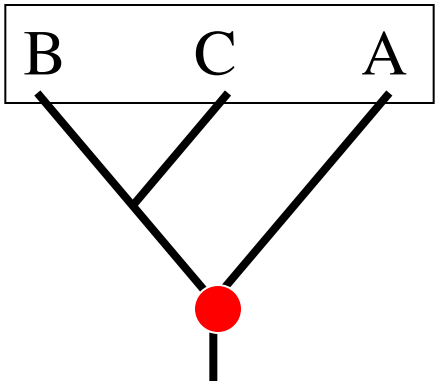
3 OTUs ⇒ 1 неукорененное дерево

3 укорененных деревьев

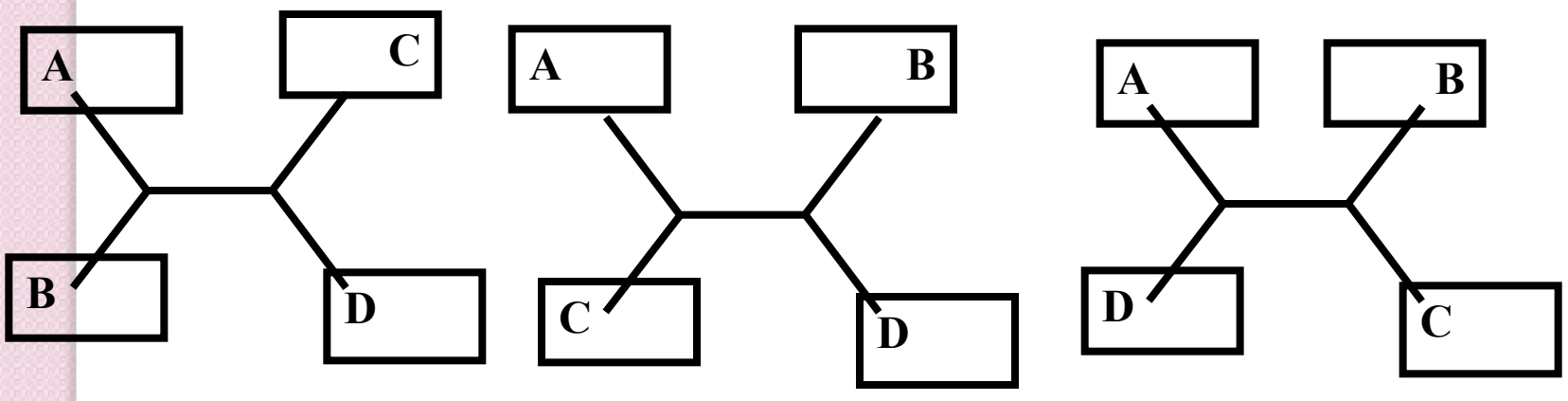


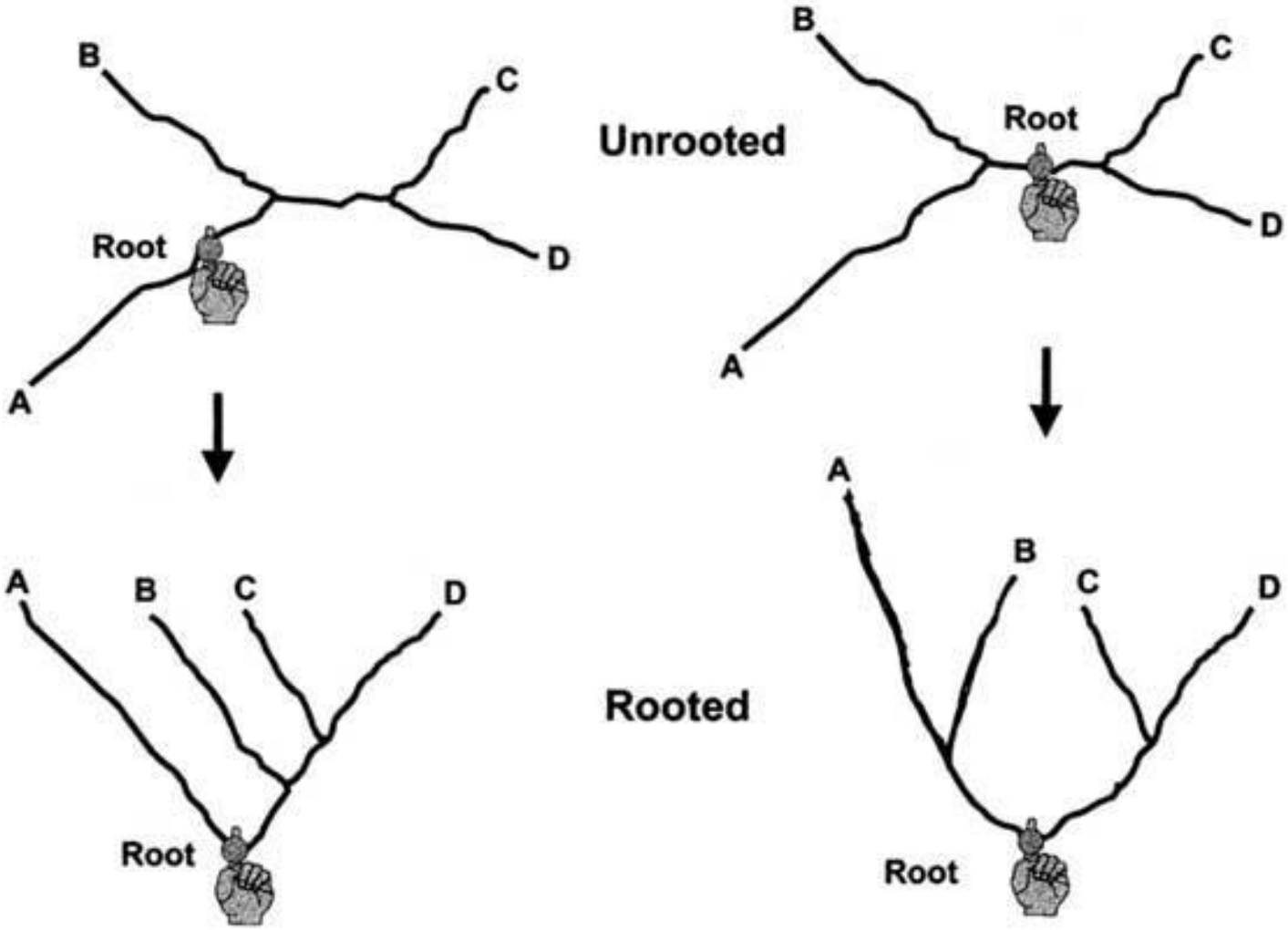






4 OTUs ⇒ 3 неукорененных филогенетических деревьев



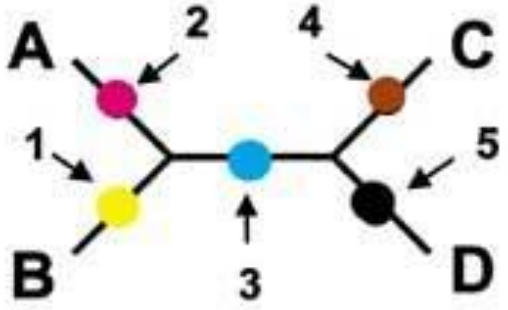


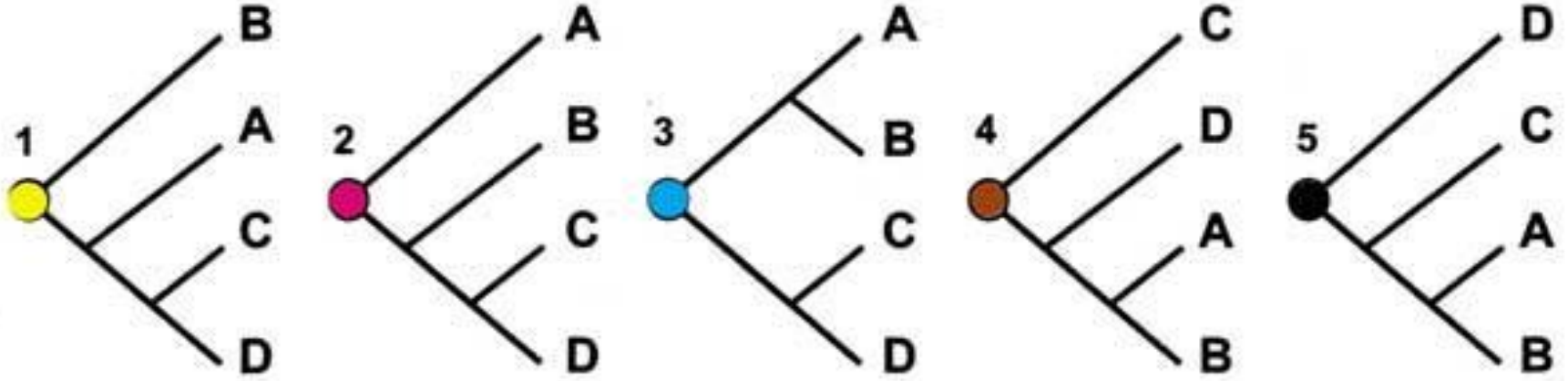
Note that in this rooted tree, taxon A is no more closely related to taxon B than it is to C or D.

Note that in this rooted tree, taxon A is most closely related to taxon B, and together they are equally distantly related to taxa C and D.

An unrooted, four-taxon tree can be rooted in five different places to produce five different rooted trees

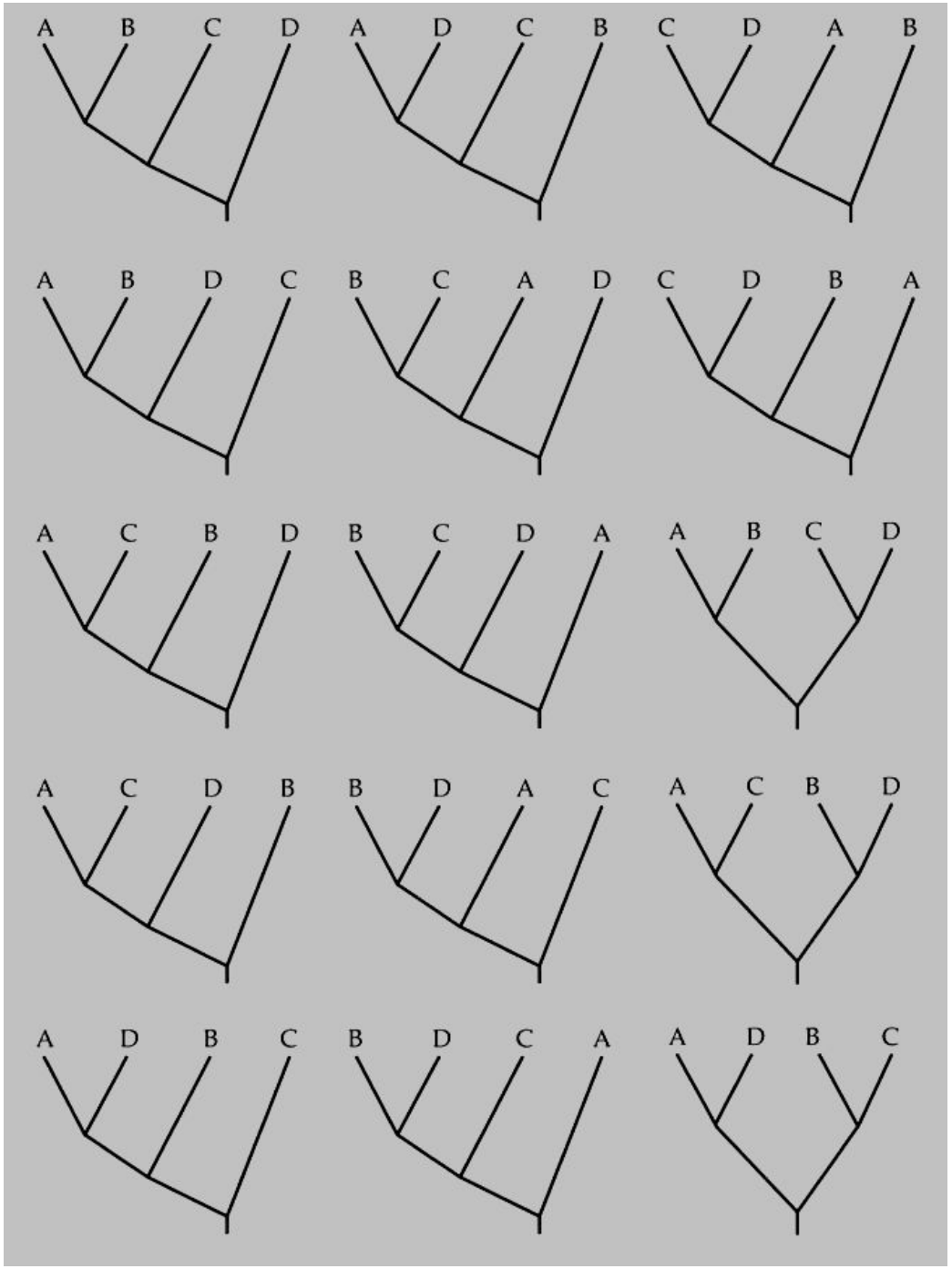
The unrooted tree:





4 OTUs ⇒

15 укорененных деревьев



Количество возможных деревьев

OTU	Количество укорененных	Количество неукорененных
2	1	1
3	3	1
4	15	3
5	105	15
6	954	105
7	10,395	954
8	135,135	10,395
9	2,027,025	135,135
10	34,459,425	2,027,025
11	654,729,075	34,459,425
12	13,749,310,575	654,729,075

Рутинная процедура, или как строят деревья?

Составление выборки последовательностей

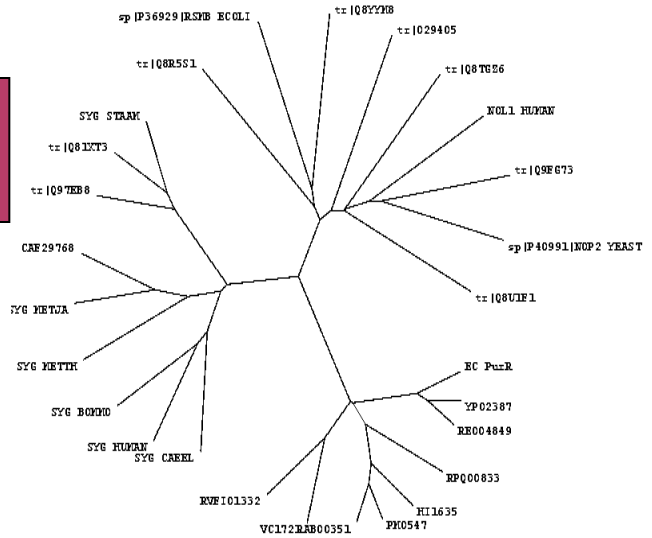
Множественное выравнивание

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

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):25.83145,



Визуализация и редактура дерева

Рутинная процедура, или как строят деревья?

Составление выборки последовательностей

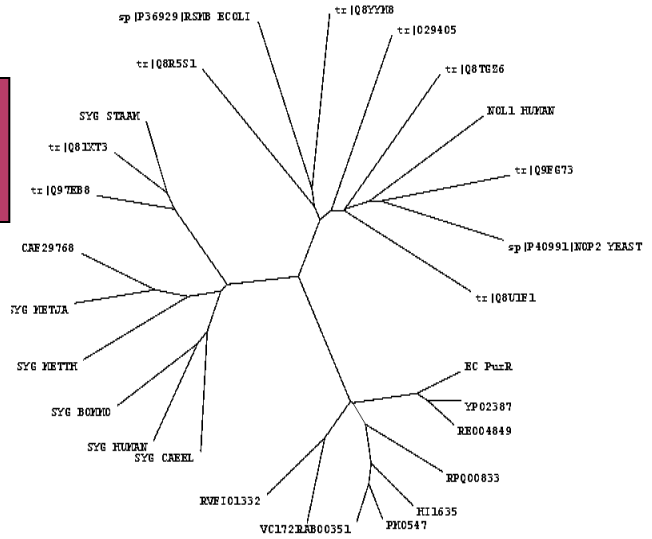
Множественное выравнивание (или всё-таки попарное)

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

Построение дерева

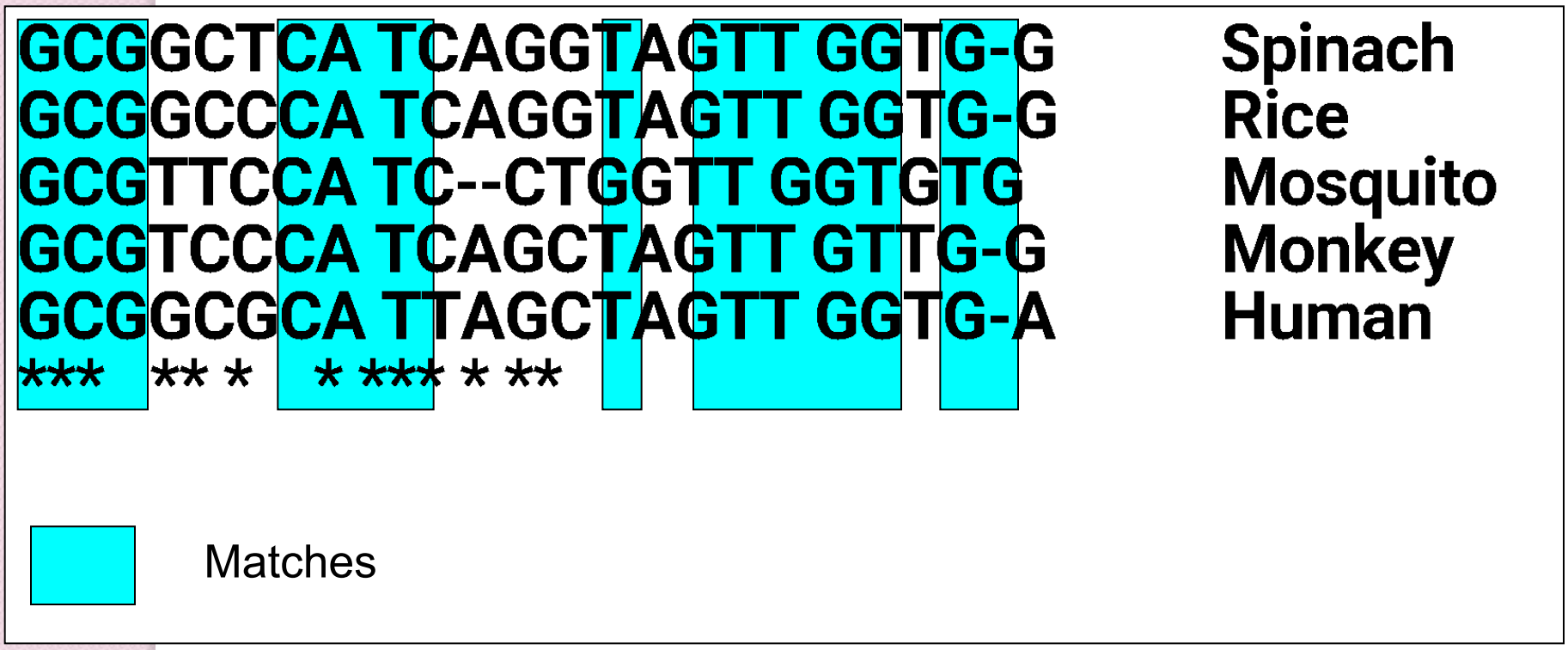
фрагмент записи в виде скобочной формулы:

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

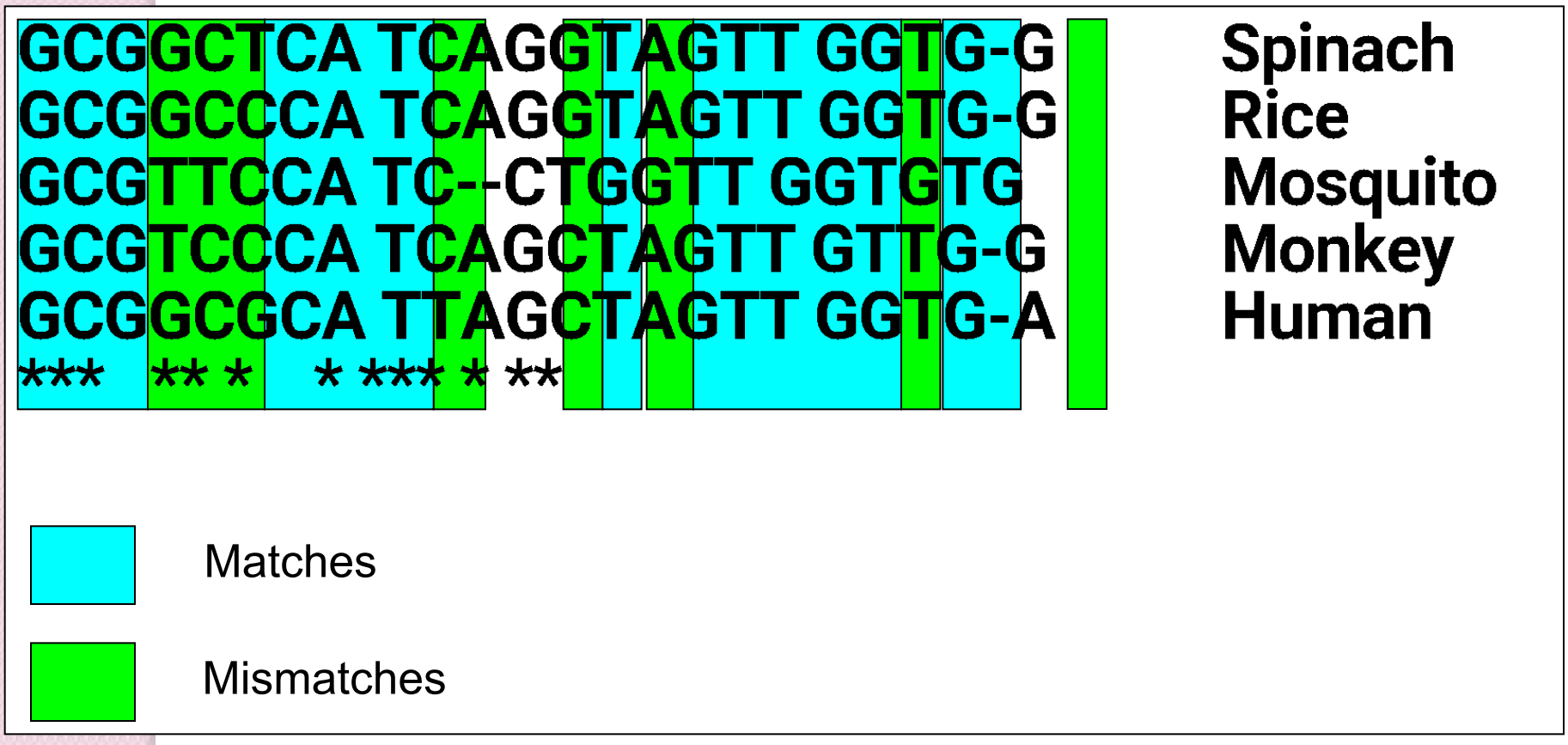


Визуализация и редактура дерева

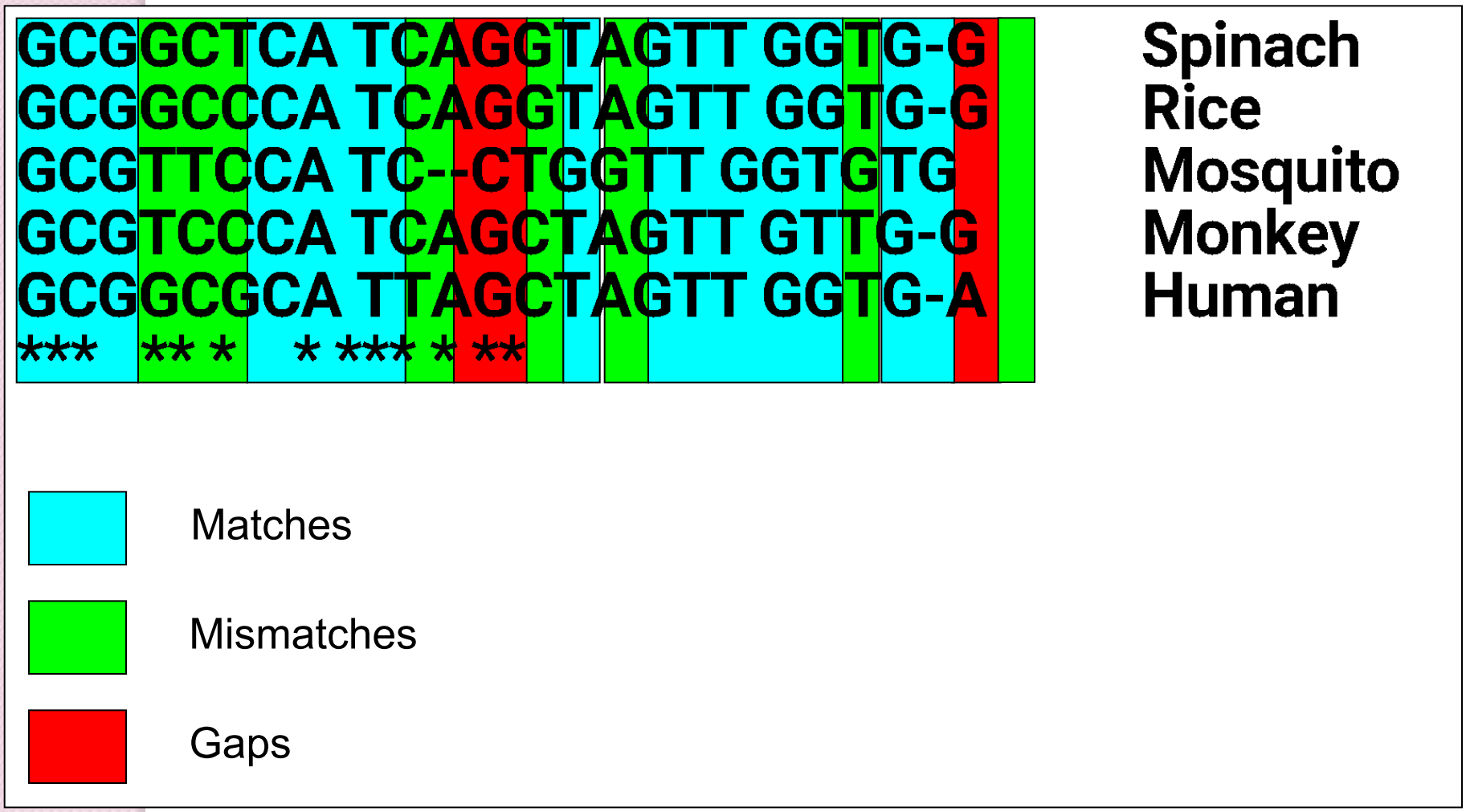
Множественное выравнивание



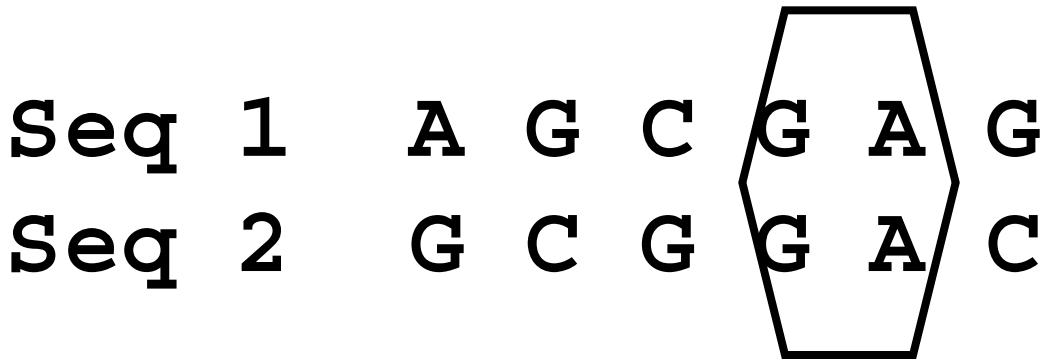
Multiple Alignment



Multiple Alignment



Шаг 3. Перевод количества расхождений в индексы замен

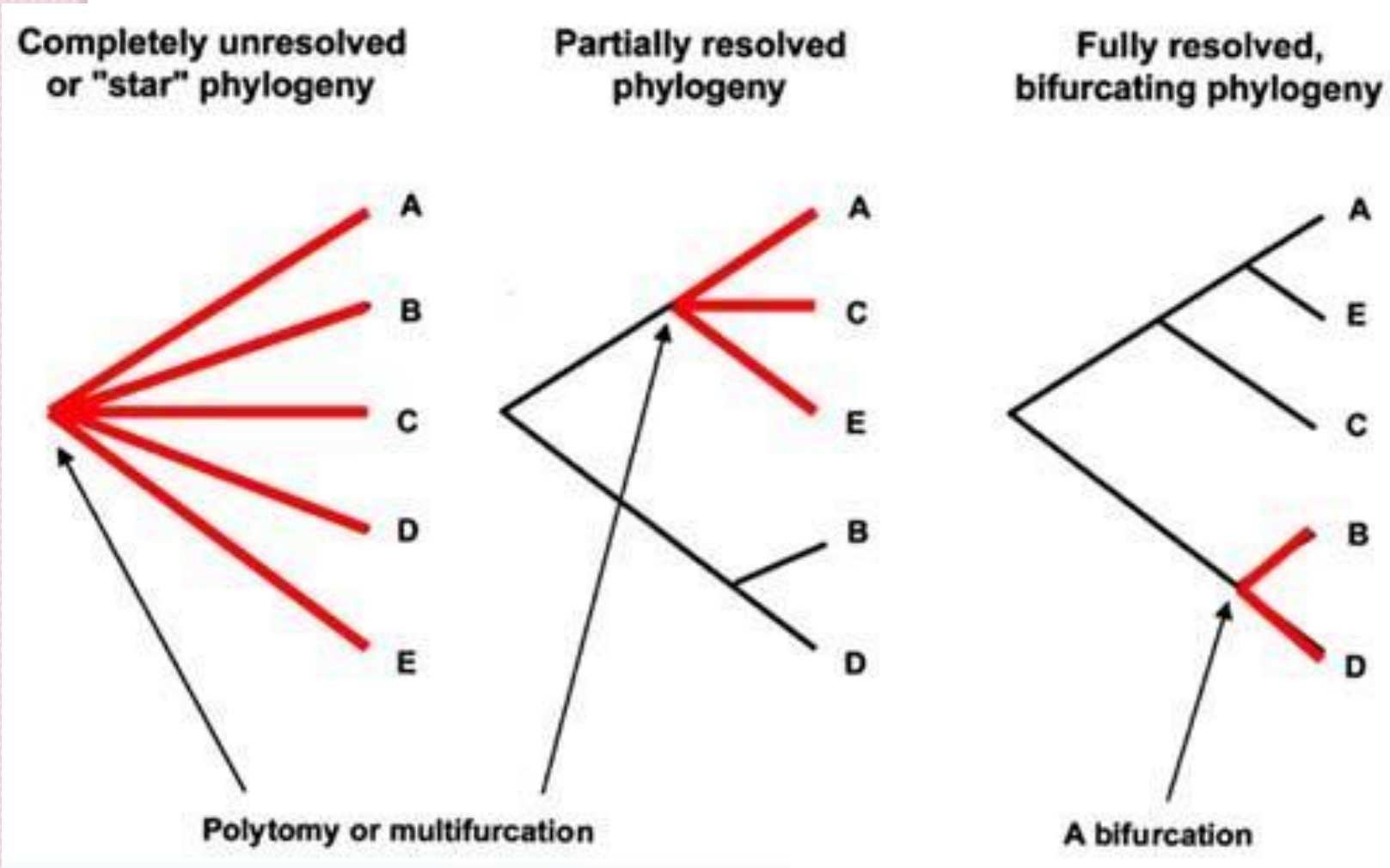


Distance Matrix*

	Spinach	Rice	Mosquito	Monkey	Human
Spinach		9	106	91	86
Rice			118	122	122
Mosquito				55	51
Monkey					3
Human					

* Units

Шаг 4: построение филогенетического дерева



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 (горизонтальный перенос генов);
  - 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):

- **UPGMA (кластеризация)**
- **Neighbor-joining**

Наибольшего правдоподобия, Maximal likelihood, ML

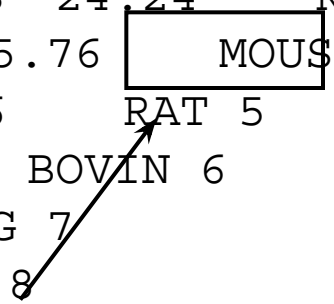
Используется модель эволюции и строится дерево, которое наиболее правдоподобно при данной модели

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
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

TABLE 7-1 Rates of Amino Acid Substitutions per Amino Acid Site per 10^9 Years ($\lambda \times 10^9$) in Various Proteins

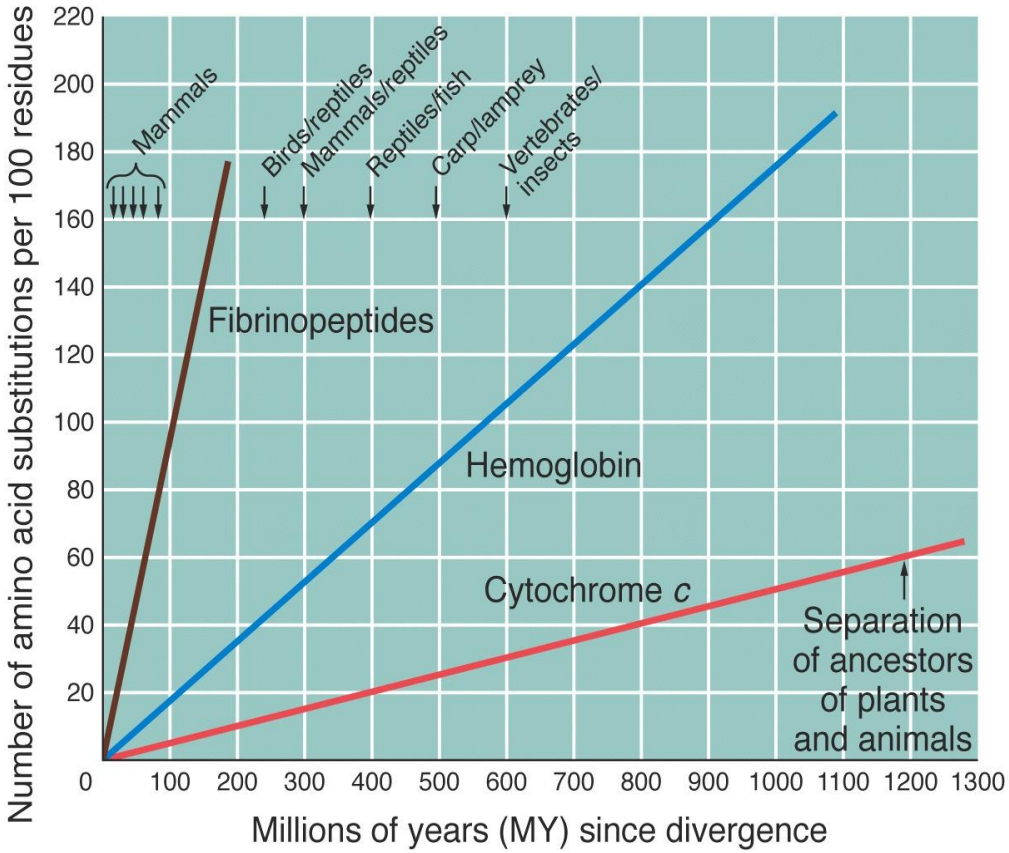
Protein	Rate	Protein	Rate
Fibrinopeptides	9.0	Thyrotropin beta chain	0.74
Growth hormone	3.7	Parathyrin	0.73
Immunoglobulin (Ig) kappa chain C region	3.7	Parvalbumin	0.70
Kappa casein	3.3	Trypsin	0.59
Ig gamma chain C region	3.1	Melanotropin beta	0.56
Lutropin beta chain	3.0	Alpha crystallin A chain	0.50
Ig lambda chain C region	2.7	Endorphin	0.48
Lactalbumin	2.7	Cytochrome b_5	0.45
Epidermal growth factor	2.6	Insulin (except guinea pig and coypu)	0.44
Somatotropin	2.5	Calcitonin	0.43
Pancreatic ribonuclease	2.1	Neurophysin 2	0.36
Serum albumin	1.9	Plastocyanin	0.35
Phospholipase A2	1.9	Lactate dehydrogenase	0.34
Prolactin	1.7	Adenylate kinase	0.32
Carbonic anhydrase C	1.6	Cytochrome c	0.22
Hemoglobin alpha chain	1.2	Troponin C, skeletal muscle	0.15
Hemoglobin beta chain	1.2	Alpha crystallin B chain	0.15
Gastrin	0.98	Glucagon	0.12
Lysozyme	0.98	Glutamate dehydrogenase	0.09
Myoglobin	0.89	Histone H2B	0.09
Amyloid AA	0.87	Histone H2A	0.05
Nerve growth factor	0.85	Histone H3	0.014
Acid proteases	0.84	Ubiquitin	0.010
Myelin basic protein	0.74	Histone H4	0.010

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

За равное время во всех ветвях эволюции данного гена\белка 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

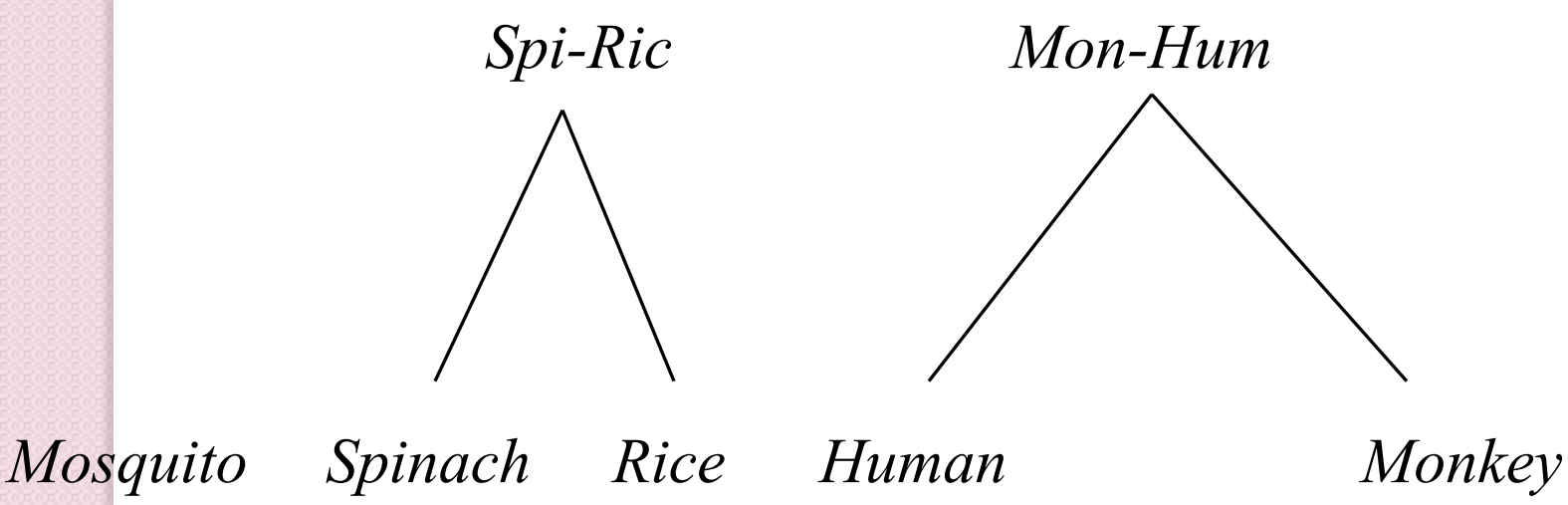
	Spinach	Rice	Mosquito	Monkey	Human
Spinach	0.0	9	106	91	86
Rice		0.0	118	122	122
Mosquito			0.0	55	51
Monkey				0.0	3
Human					0.0

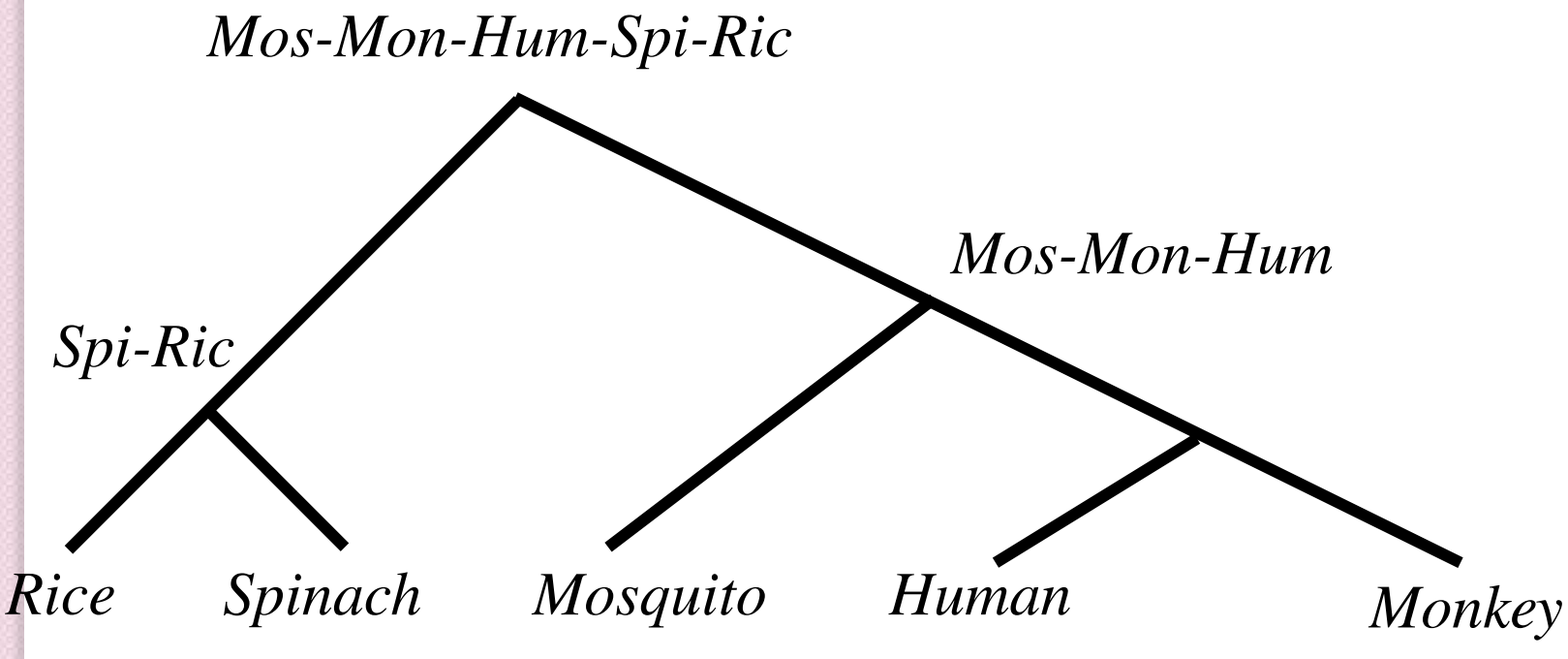
Дистанция между человеком и обезьяной минимальна. Эти группы объединяются в Monkey-Human, а все остальные дистанции пересчитываются

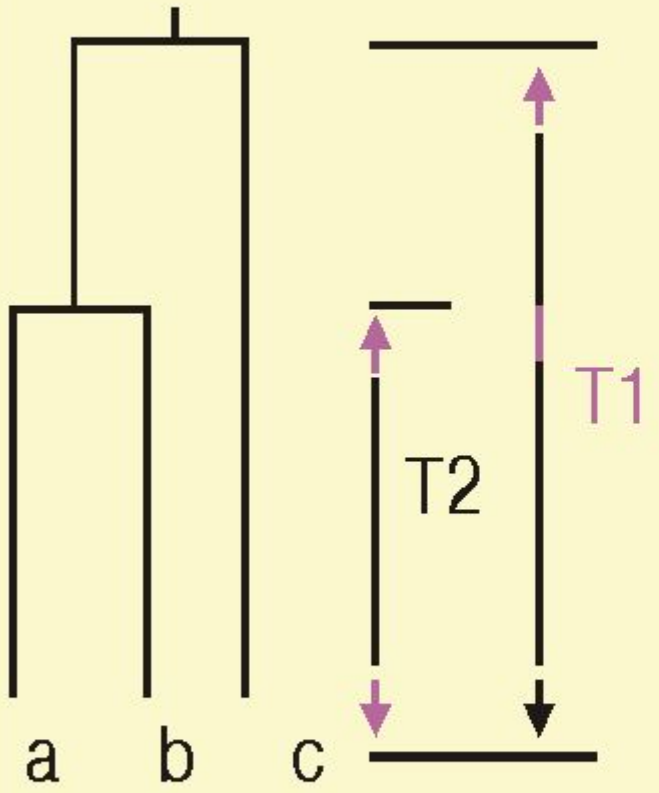
Dist[Spinach, MonHum] = (Dist[Spinach, Monkey] +
Dist[Spinach, Human])/2 = (91 + 86)/2 = 88.5

Редуцированная матрица дистанций

	Spinach	Rice	Mosquito	Mon-Hum
Spinach	0.0	9	106	88.5
Rice		0.0	118	122
Mosquito			0.0	53
Mon-Hum				0.0


Spi-Ric
Mon-Hum
Mosquito
Spinach
Rice
Human
Monkey


Mos-Mon-Hum-Spi-Ric
Mos-Mon-Hum
Spi-Ric
Rice
Spinach
Mosquito
Human
Monkey



В случае если известно время T1 и не известно T2, время T2 вычисляется по формуле:

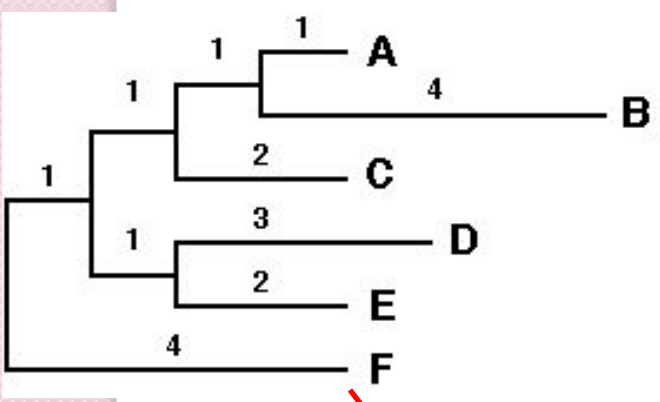
$$T2 = \frac{2*K(ab)*T1}{K(ac)+K(bc)}$$

Если же наоборот, то:

$$T1 = \frac{(K(ac)+K(bc))*T2}{2*K(ab)}$$

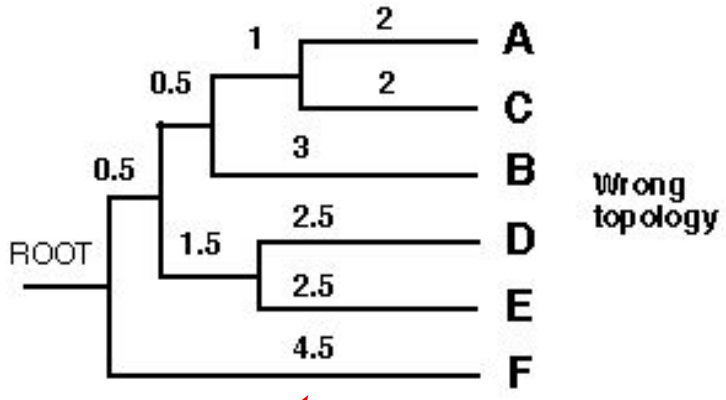
Недостатки UPGMA

Алгоритм строит ультраметрическое дерево – 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

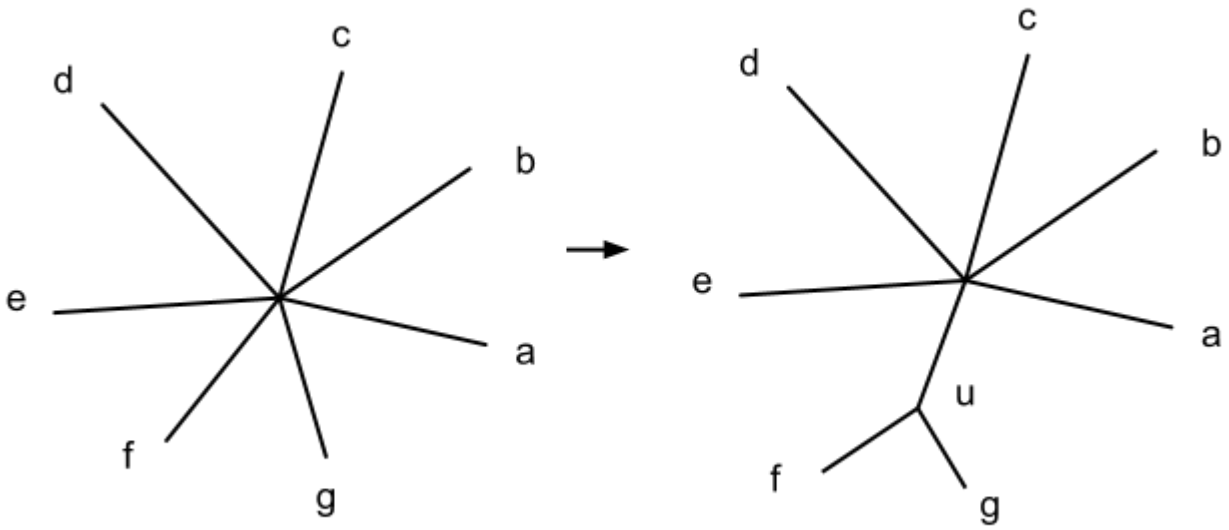


UPGMA

Метод ближайших соседей (Neighbor-joining, NJ)

- ✔ Строит неукоренённое дерево
- ✔ Может работать с большим количеством данных
- ✔ Достаточно быстрый
- ✔ Если есть недвусмысленное с точки зрения эксперта дерево, то оно будет построено.
- ✔ !!! Только древо сходства – не филогенетическое

Метод Neighbor-joining



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

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

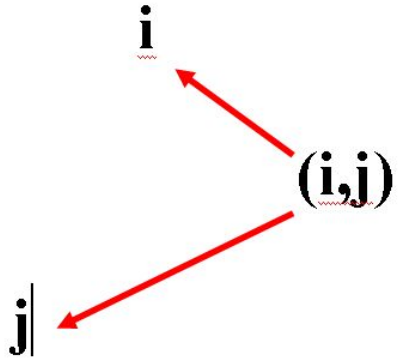
1. Рассмотрим все возможные пары листьев. Выберем 2 листа *i* и *j* с минимальным значением величины

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

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):

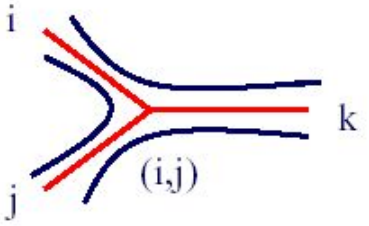
$D(i, (i,j)) = 0,5 \cdot (Mij + ui - uj)$

$D(j, (i,j)) = 0,5 \cdot (Mij + uj - ui)$

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

3. Вычисляем расстояние от нового кластера до всех других

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

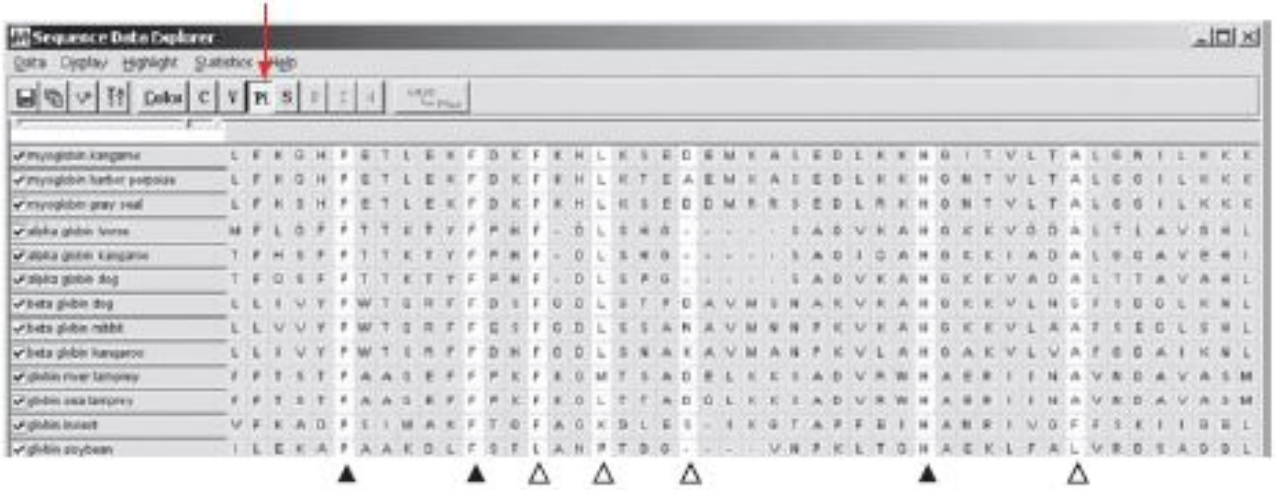


5. В матрице M убираем i и j и добавляем (i, j).

Повторяем, пока не останутся 3 узла ...

Maximum Parsimony (MP)

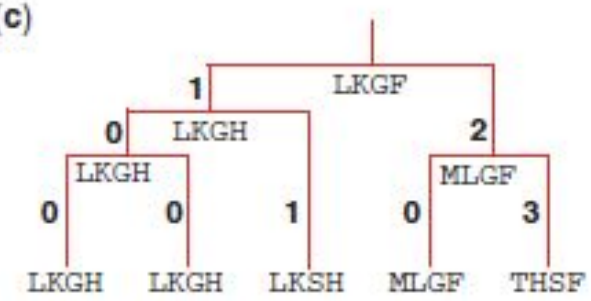
(a)



(b)

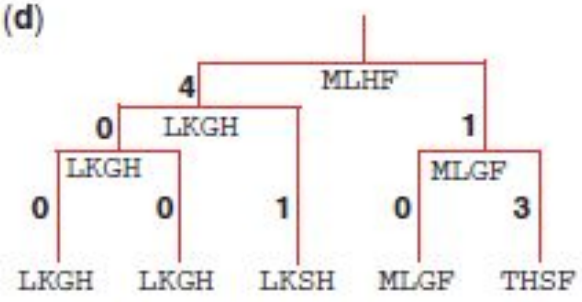
```
kangaroo    LKGH
porpoise    LKGH
gray seal   LKSH
horse α     MLGF
kangaroo α  THSF
```

(c)



Total cost: 7

(d)



Total cost: 9

Методы, основанные на последовательностях:
Maximum Likelihood (ML), Maximum Parsimony (MP)

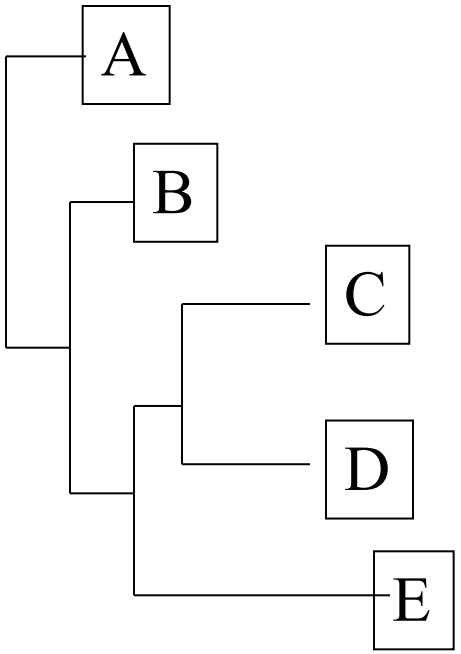
Input:

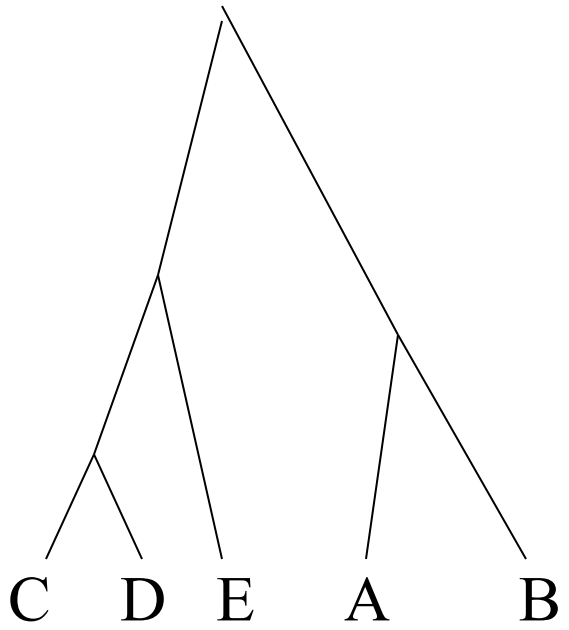
MSA для n последовательностей, одна последовательность для каждого вида.

Как изобразить дерево?
Топология дерева

Топология дерева — только листья, узлы, (корень) и связывающие их ветви

(топология не зависит от способа изображения дерева)

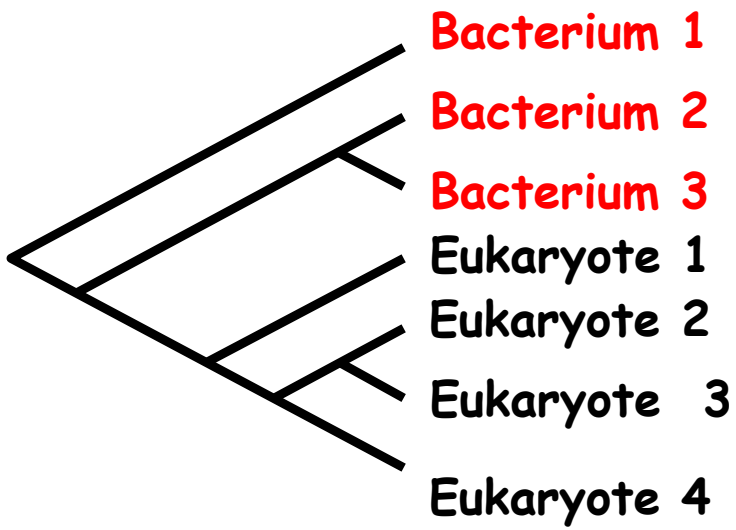




Два изображения одной и той же топологии

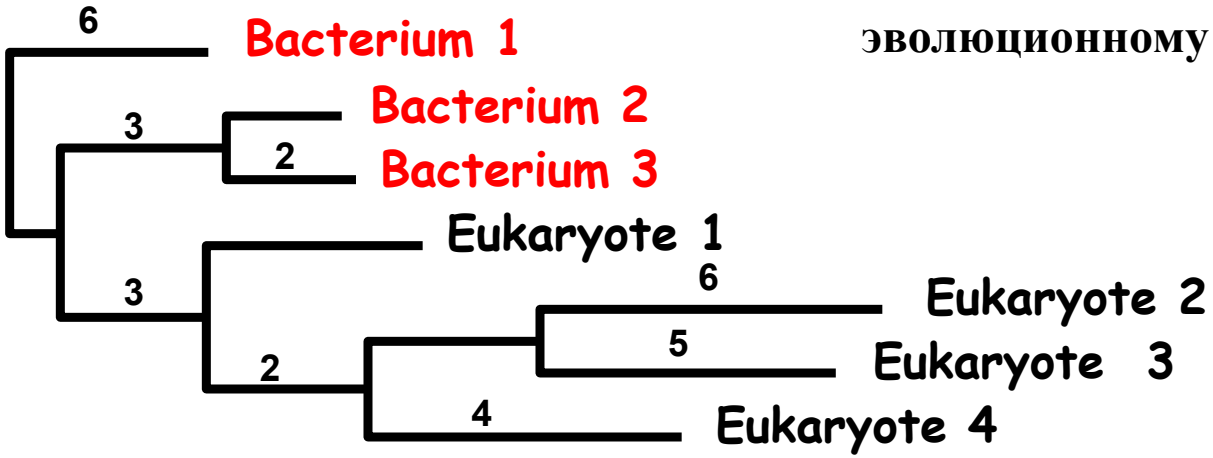
Как можно нарисовать построенное дерево?

Кладограммы и филограммы



Кладограммы – только топологя. Длины ветвей не учитываются

Филограммы – длины ветвей пропорциональны эволюционному расстоянию.



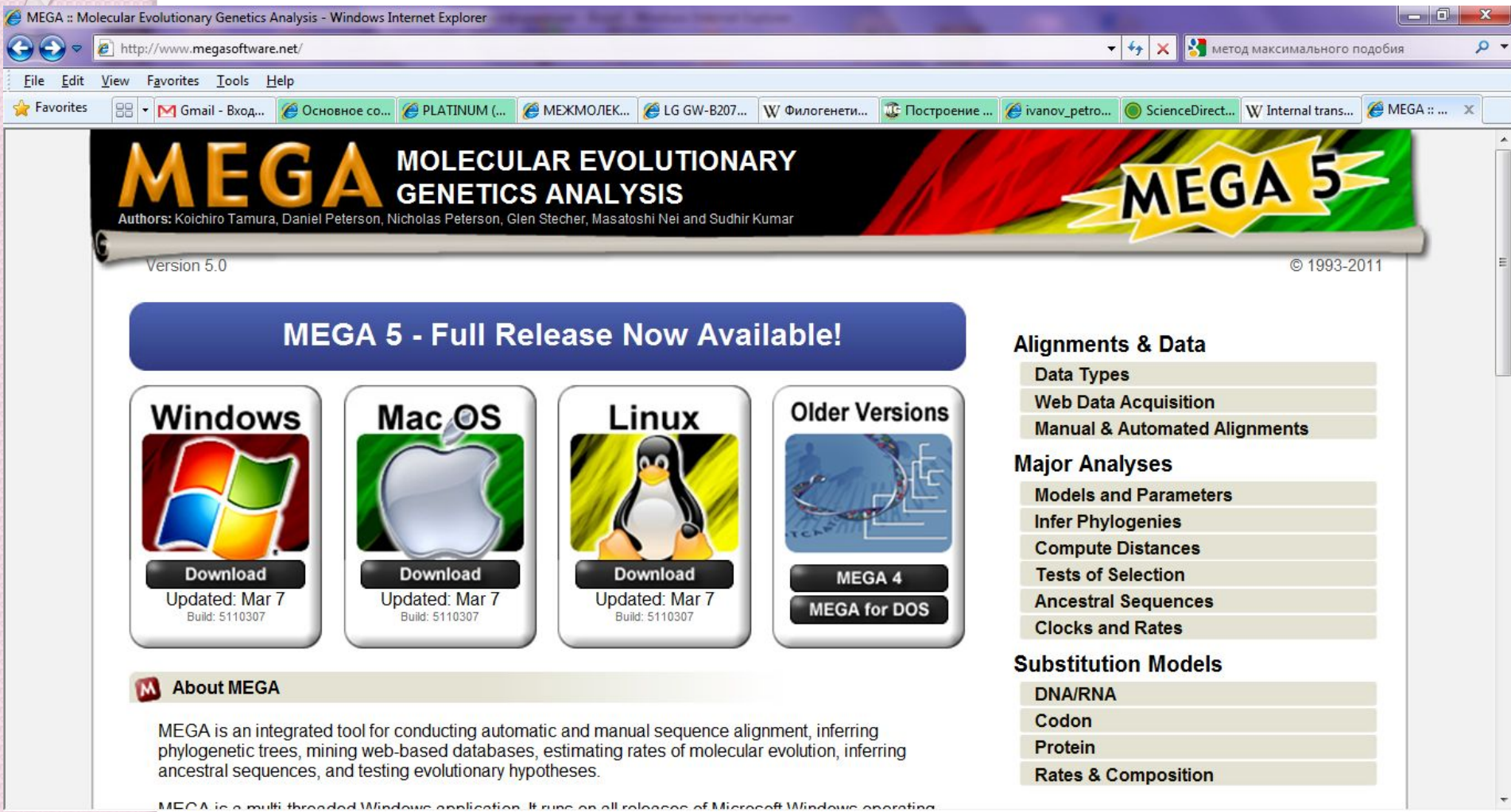
Какие on-line программы строят деревья?

- 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)

PAUP (Phylogenetic Analysis Using Parsimony)

MEGA: филогенетический анализ последовательностей

http://www.megasoftware.net/



Эволюция – исторический процесс.

Из 8,200,794,532,637,891,559,375 деревьев для 20 OTUs, 1 является верным и 8,200,794,532,637,891,559,374 неверны.

Truth is one, falsehoods are many.

Дякую за увагу
Благодарю за внимание
Thank you for your attention