ВВЕДЕНИЕ В БИОИНФОРМАТИКУ

Лекция №8

Методы кластеризации.
Филогенетические деревья.
Множественные выравнивания.
Скрытые марковские модели

Новоселецкий Валерий Николаевич
к.ф.-м.н., доц. каф. биоинженерии
valery.novoseletsky@yandex.ru

Сайт курса http://intbio.org/bioinf2018

Множественное выравнивание последовательностей

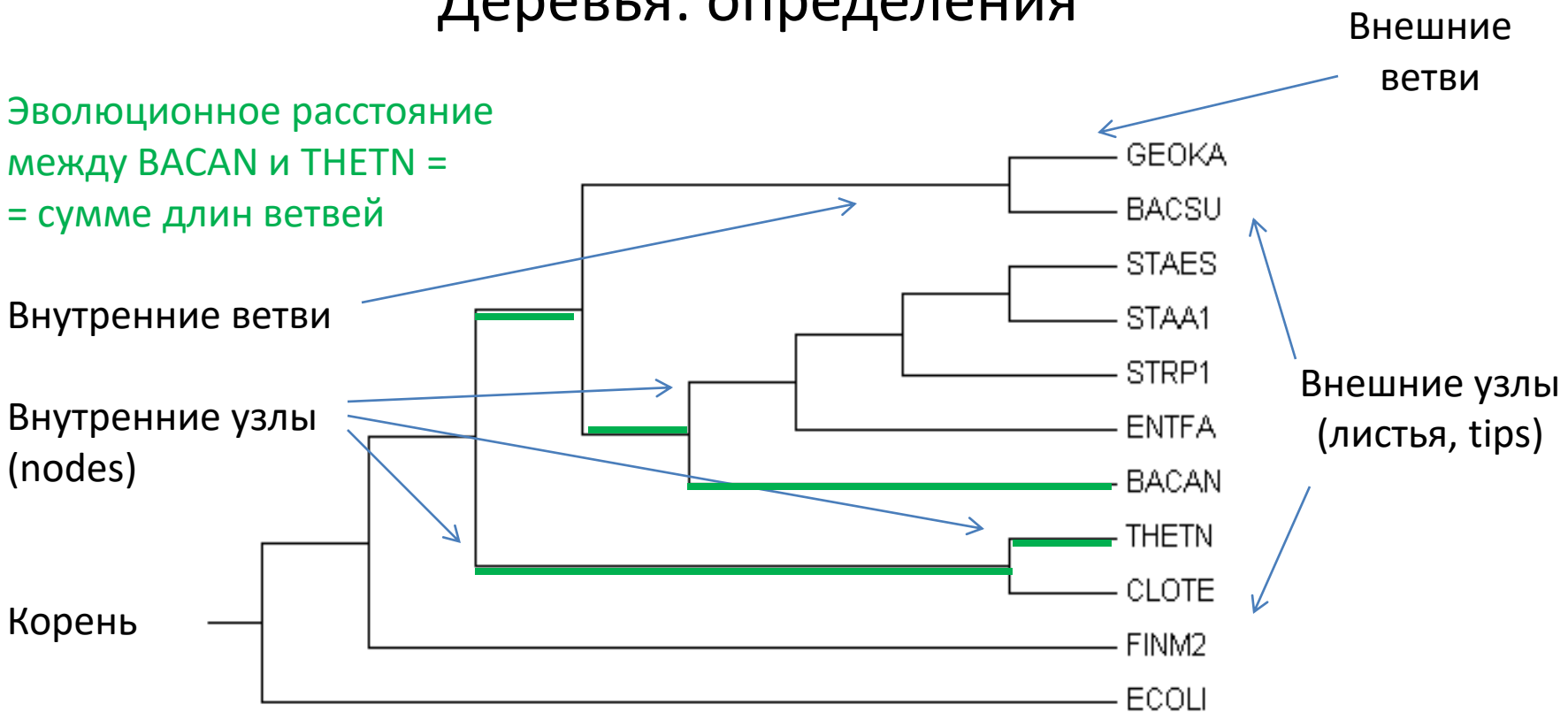
Цели:

- **Построение филогенетических деревьев**
- **Выявление консервативных остатков и мотивов**
- **Построение профилей (визуализация)**
- **Итеративное выявление удаленной гомологии**
- …

Алгоритмы:

- Динамическое программирование – не годится
- Прогрессивное выравнивание
- Скрытые марковские модели
- Квантовые компьютеры?

Деревья: определения



Кладограмма — филогенетическое дерево, не содержащее информации о длинах ветвей.

Филограмма — филогенетическое дерево, содержащее информацию о длинах ветвей; эти длины представляют изменение некой характеристики.

Деревья: свойства

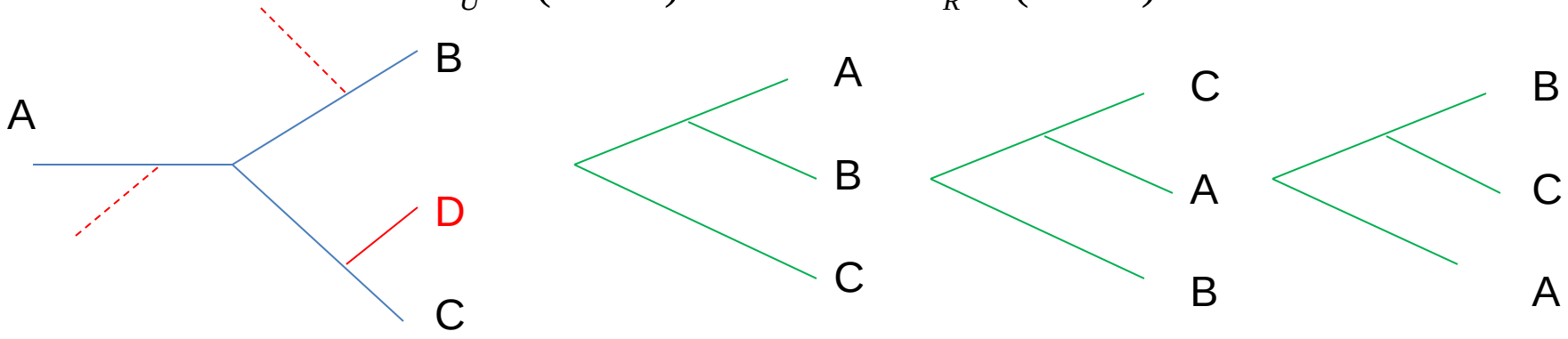
Число деревьев

неукорененных

$$N_U = (2n-5)!!$$

укорененных

$$N_R = (2n-3)!!$$

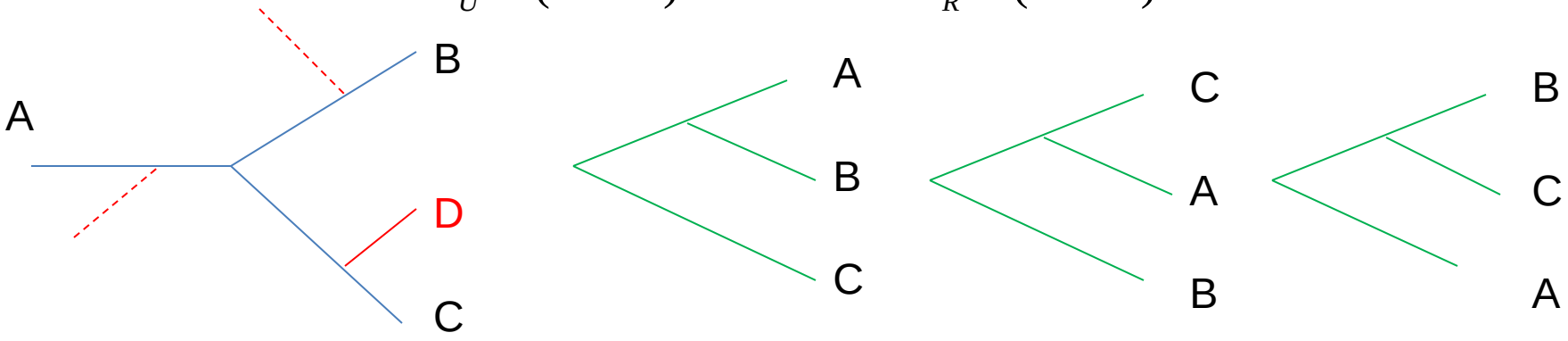


Деревья: свойства

Число деревьев

неукорененных $N_U = (2n-5)!!$

укорененных $N_R = (2n-3)!!$

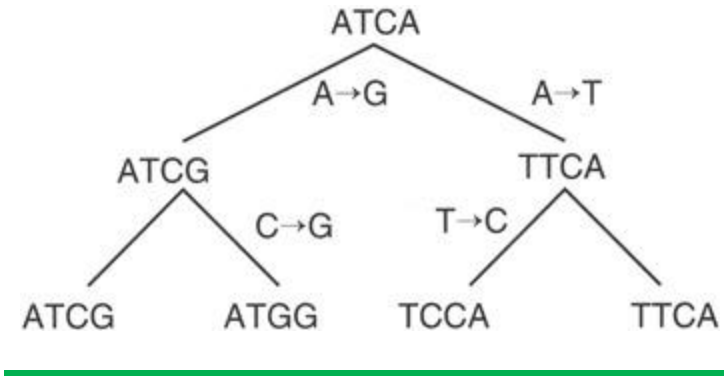


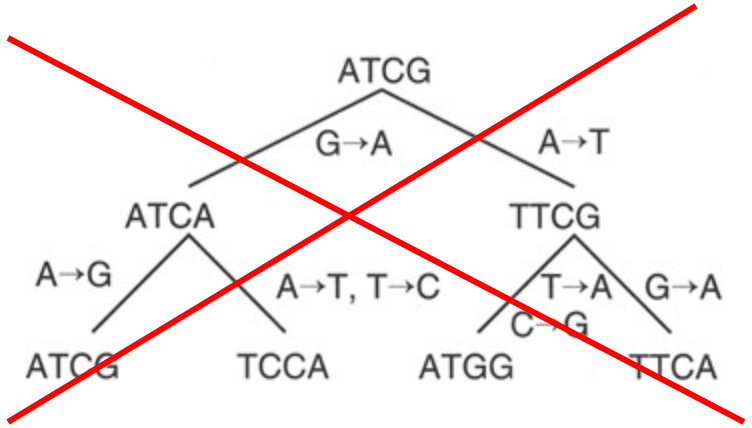
Число листьев n	Число неукорененных деревьев Nu	Число укорененных деревьев Nr
3	1	3
4	3	15
10	2027025	34459425
20	2,21643E+20	8,20079E+21
	Возраст Земли	1,4E+17 секунд

С точки зрения филогении, правильное только одно! Как его отыскать?

Филогенетические деревья – методы построения

Максимальная экономия (Fitch, 1971) (**метод оценки!**) – критерий оптимальности, согласно которому предпочтительнее деревья с меньшим суммарным числом мутаций. Однако алгоритма быстрого построения такого дерева не существует.





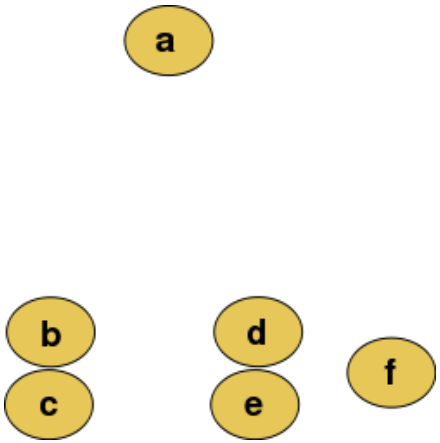
Метод максимального правдоподобия учитывает не просто число мутаций, но и их вероятность.

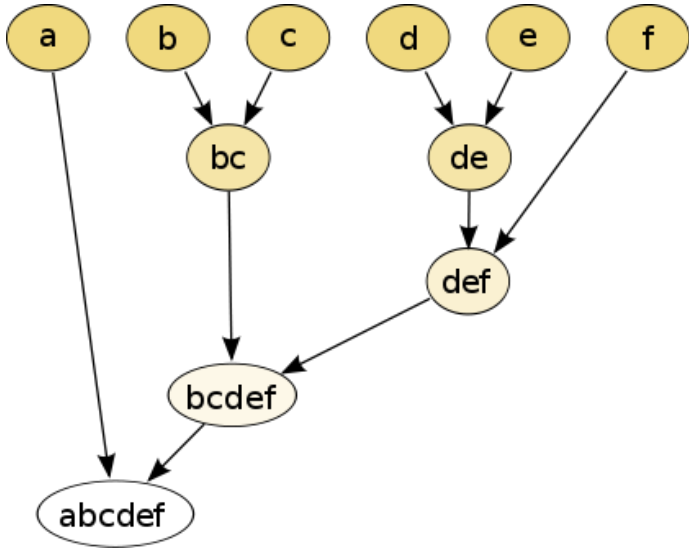
Методы иерархической кластеризации. UPGMA

Unweighted Pair Group Method with Arithmetic mean (Sokal, Michener, 1958)

– метод невзвешенной группировки с арифметическим средним

– пример алгоритма иерархической кластеризации





Расстояние между элементами

$$\|a-b\|_2 = \sqrt{\sum_i (a_i - b_i)^2}$$

Расстояние между кластерами

$$\frac{1}{|\mathcal{A}| \cdot |\mathcal{B}|} \sum_{x \in \mathcal{A}} \sum_{y \in \mathcal{B}} d(x,y).$$

Методы иерархической кластеризации. UPGMA

Дан набор объектов *Sk*, где для каждой пары (*Si, Sj*) установлена мера сходства *L(Si, Sj)*. Для построения дерева выбирают два наиболее близких объекта *(Sm, Sn)* и добавляют вершину, изображающую их общего «предка» *(Smn)*. Затем замещают эти два объекта группой, содержащий обоих, и присваивают расстояниям от этой пары до остальных объектов *Sk* средние значения от каждого из элементов этой группы до *Sk*:

$$L(S_{mn}, S_k) = \frac{L(S_m, S_k) + L(S_n, S_k)}{2}$$

В случае объединения кластеров *Ci* и *Cj* с образованием кластера *Ck*, содержащего *ni + nj = nk* элементов, расстояние от кластера *Ck* до остальных кластеров *Cm* вычисляется как

$$L(C_k, C_m) = \frac{n_i L(C_i, C_m) + n_j L(C_j, C_m)}{n_i + n_j}$$

Методы иерархической кластеризации. UPGMA

Дано 6 последовательностей – ATTTG, AGCGT, ACCGT, CGCGA, GGCGA, CGGGG.

Используя расстояние по Хэммингу, получаем матрицу расстояний *D0*:

D0	ATTTG	AGCGT	ACCGT	CGCGA	GGCGA	CGGGG
ATTTG	0	4	4	5	5	4
AGCGT		0	***1***	2	2	3
ACCGT			0	3	3	4
CGCGA				0	***1***	2
GGCGA					0	3
CGGGG						0

Объединяем первую пару элементов

Методы иерархической кластеризации. UPGMA

Дано 6 последовательностей – ATTTG, AGCGT, ACCGT, CGCGA, GGCGA, CGGGG.

D0	ATTTG	AGCGT	ACCGT	CGCGA	GGCGA	CGGGG
ATTTG	0	4	4	5	5	4
AGCGT		0	***1***	2	2	3
ACCGT			0	3	3	4
CGCGA				0	***1***	2
GGCGA					0	3
CGGGG						0

Матрицу D0 преобразуем в матрицу D1:

D1	ATTTG	AGCGT, ACCGT	CGCGA	GGCGA	CGGGG
ATTTG	0	(4+4)/2=4	5	5	4
AGCGT, ACCGT		0	(2+3)/2=2,5	(2+3)/2=2,5	(3+4)/2=3,5
CGCGA			0	**1**	2
GGCGA				0	3
CGGGG					0

Методы иерархической кластеризации. UPGMA

D1	ATTTG	AGCGT, ACCGT	CGCGA	GGCGA	CGGGG
ATTTG	0	(4+4)/2=4	5	5	4
AGCGT, ACCGT		0	(2+3)/2=2,5	(2+3)/2=2,5	(3+4)/2=3,5
CGCGA			0	**1**	2
GGCGA				0	3
CGGGG					0

Объединяем вторую пару элементов

Методы иерархической кластеризации. UPGMA

D1	ATTTG	AGCGT, ACCGT	CGCGA	GGCGA	CGGGG
ATTTG	0	(4+4)/2=4	5	5	4
AGCGT, ACCGT		0	(2+3)/2=2,5	(2+3)/2=2,5	(3+4)/2=3,5
CGCGA			0	**1**	2
GGCGA				0	3
CGGGG					0

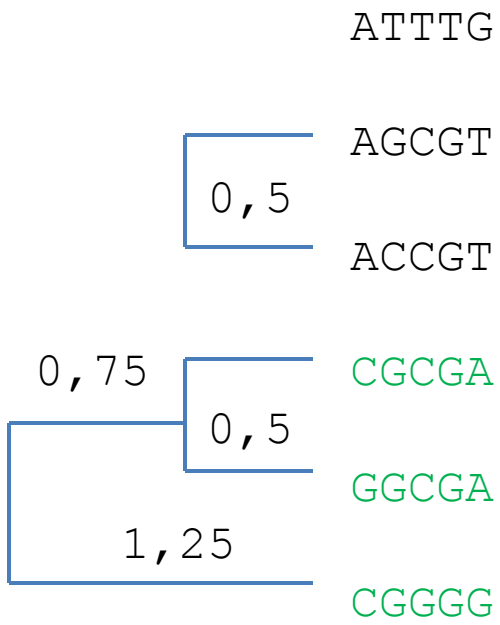
Матрицу D1 преобразуем в матрицу D2:

D2	ATTTG	AGCGT, ACCGT	CGCGA, GGCGA	CGGGG
ATTTG	0	4	(5+5)/2=5	4
AGCGT, ACCGT		0	(2,5+2,5)/2=2,5	3,5
CGCGA, GGCGA			0	(2+3)/2=**2,5**
CGGGG				0

Методы иерархической кластеризации. UPGMA

D2	ATTTG	AGCGT, ACCGT	CGCGA, GGCGA	CGGGG
ATTTG	0	4	(5 + 5)/2 = 5	4
AGCGT, ACCGT		0	(2,5 + 2,5)/2 = 2,5	3,5
CGCGA, GGCGA			0	(2 + 3)/2 = **2,5**
CGGGG				0

Добавляем третий элемент
во второй кластер



Методы иерархической кластеризации. UPGMA

D2	ATTTG	AGCGT, ACCGT	CGCGA, GGCGA	CGGGG
ATTTG	0	4	(5 + 5)/2 = 5	4
AGCGT, ACCGT		0	(2,5 + 2,5)/2 = 2,5	3,5
CGCGA, GGCGA			0	(2 + 3)/2 = **2,5**
CGGGG				0

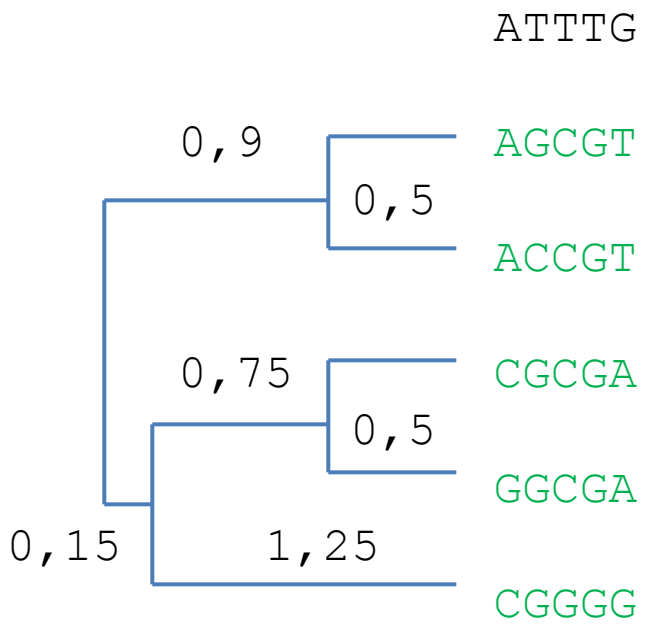
Матрицу D2 преобразуем в матрицу D3:

D3	ATTTG	AGCGT, ACCGT	(CGCGA, GGCGA), CGGGG
ATTTG	0	4	(2*5 + 4)/3 = 4,7
AGCGT, ACCGT		0	(2*2,5 + 3,5)/3 = **2,8**
(CGCGA, GGCGA), CGGGG			0

Методы иерархической кластеризации. UPGMA

D3	ATTTG	AGCGT, ACCGT	(CGCGA, GGCGA), CGGGG
ATTTG	0	4	(2*5 + 4)/3 = 4,7
AGCGT, ACCGT		0	(2*2,5 + 3,5)/3 = **2,8**
(CGCGA, GGCGA), CGGGG			0

Объединяем кластеры

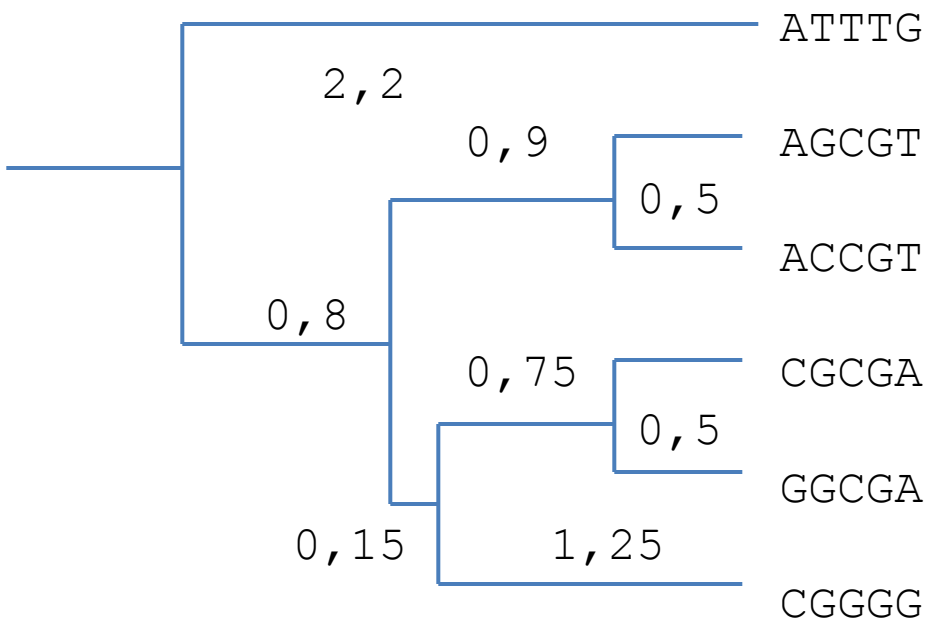


Методы иерархической кластеризации. UPGMA

Рассчитываем последнюю матрицу D4:

D4	ATTTG	((CGCGA, GGCGA), CGGGG), (AGCGT, ACCGT)
ATTTG	0	(4*2 + 4,7*3)/5 = 4,4
((CGCGA, GGCGA), CGGGG), (AGCGT, ACCGT)		0

И добавляем последний элемент

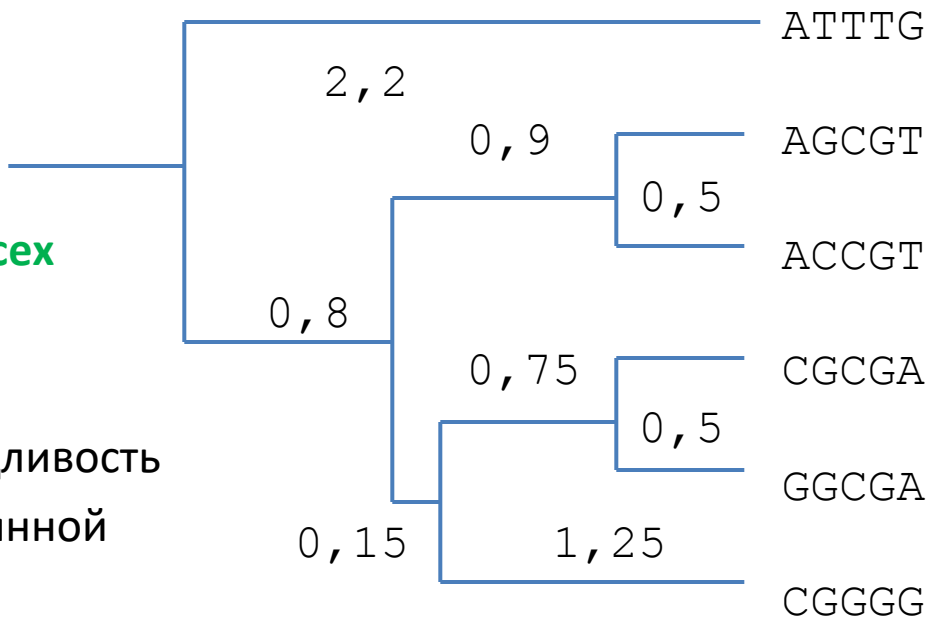


Методы иерархической кластеризации. UPGMA

D4	ATTTG	((CGCGA, GGCGA), CGGGG), (AGCGT, ACCGT)
ATTTG	0	(4*2 + 4,7*3)/5 = 4,4
((CGCGA, GGCGA), CGGGG), (AGCGT, ACCGT)		0

Длины ветвей установлены так, что **расстояние от корня одинаково для всех листьев - ультраметричность**.

Метод UPGMA подразумевает справедливость **гипотезы молекулярных часов** (постоянной скорости эволюции).



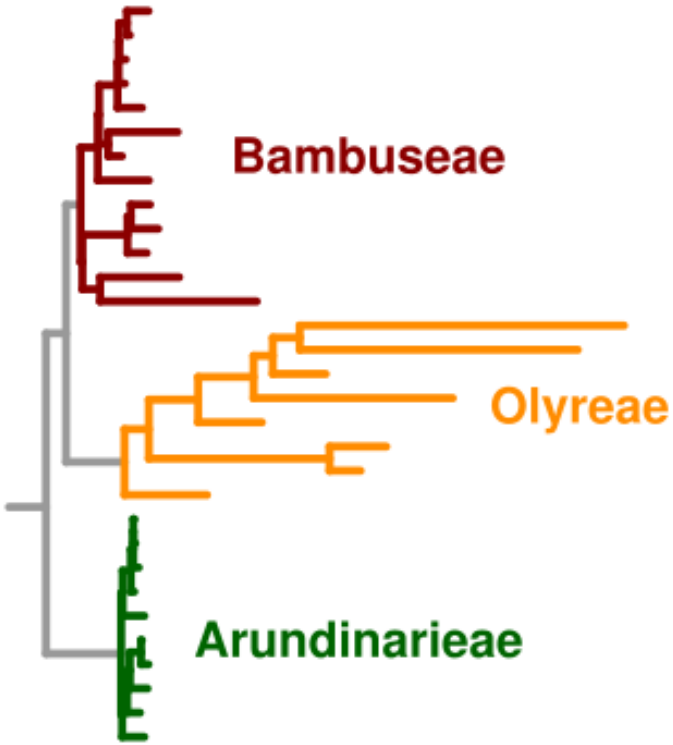
Гипотеза молекулярных часов

Сопоставления числа различий в аминокислотных последовательностях **гемоглобинов** млекопитающих и срока дивергенции (Э. Цукеркандль, Л. Полинг, 1962).

Сопоставления идентичности последовательностей **цитохрома *c*** у рыб, птиц и млекопитающих (Э. Марголиаш, 1963).

Гипотеза широко распространена, хотя и имеется довольно много примеров, ей противоречащих.

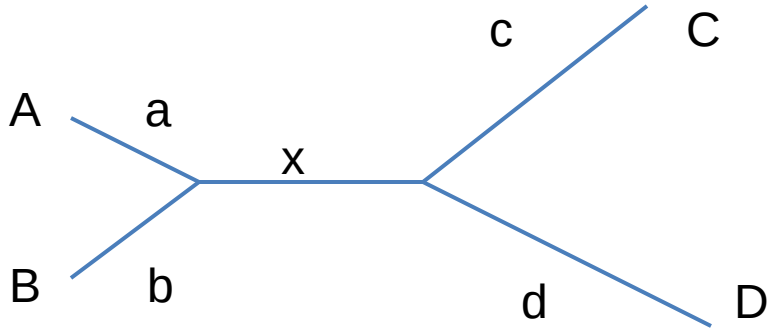
Трудности калибровки



Методы иерархической кластеризации. Метод связи между соседями

Соседи – последовательности, расположенные в дереве через один узел (A и B, C и D).

Считая длины ветвей известными, видим, что:



$$L_{AC} + L_{BD} = L_{AD} + L_{BC} = a + b + c + d + 2x = L_{AB} + L_{CD} + 2x$$

Очевидно, что

$$L_{AB} + L_{CD} < L_{AC} + L_{BD}$$

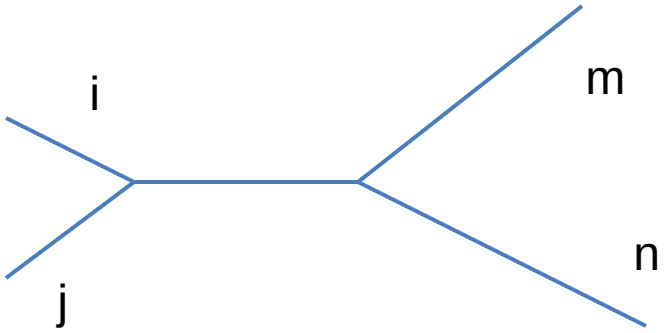
$$L_{AB} + L_{CD} < L_{AD} + L_{BC}$$

Условие четырех точек (1974)

Если известны расстояния между последовательностями, но неизвестны эволюционные отношения, то **метод позволяет установить топологию филогенетического дерева**, т.е. как раз отношения.

Методы иерархической кластеризации. Метод связи между соседями

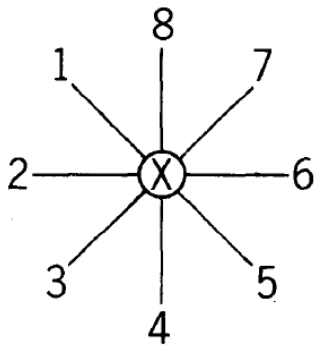
На большее число последовательностей обобщается путем рассмотрения всех четверок и определением тех из них, для которых суммы расстояний минимальны.



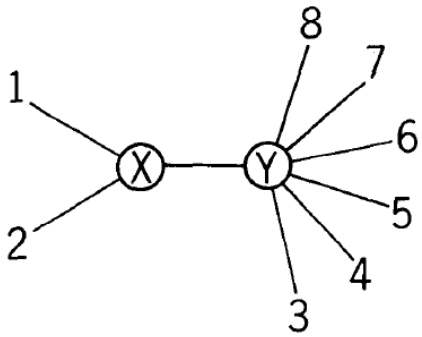
$$Если\ \min(L_{ij}+L_{mn},\ \ L_{im}+L_{jn},\ \ L_{in}+L_{jm}) = L_{ij}+L_{mn},\ \ то$$

N	…	*i*	*j*	…	*m*	*n*
…						
i			**+1**		+0	+0
j		**+1**			+0	+0
…						
m		+0	+0			**+1**
n		+0	+0		**+1**	

Методы иерархической кластеризации. NJ



(a)



(b)

Neighbor joining – метод присоединения соседей (Saitou, Nei, 1987)

- еще один алгоритм иерархической кластеризации.

Цель: построить **дерево с минимальной суммой длин ветвей.**

Пусть

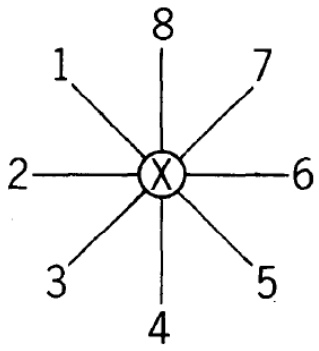
D_{ij} – *расстояние между таксонами* i *и* j, — известно из матрицы расстояний

L_{ab} – *длина ветви между узлами* a *и* b — нужно найти

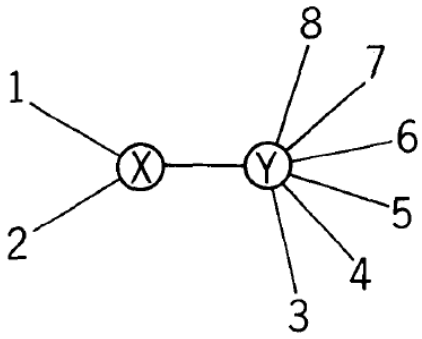
Тогда для суммы длин ветвей дерева на рис. (a) *выполняется*

$$S_0 = \sum_{i=1}^{N} L_{iX} = \frac{1}{N-1} \sum_{i<j} D_{ij} \quad (\text{поскольку } D_{ij} = L_{iX} + L_{Xj}) \qquad (1)$$

Методы иерархической кластеризации. NJ



(a)



(b)

Выделив пару таксонов 1 и 2 и добавив узел Y, для суммы расстояний от таксонов 1 и 2 до всех остальных имеем (рис. (b)):

$$\sum_{k=3}^{N}(D_{1k}+D_{2k}) = (N-2)L_{1X}+(N-2)L_{2X}+2(N-2)L_{XY}+2\sum_{i=3}^{N}L_{iY} \quad (2)$$

А для суммы всех ветвей дерева (рис. (b))

$$S_{12} = L_{XY}+L_{1X}+L_{2X}+\sum_{i=3}^{N}L_{iY} \quad (3)$$

Теперь надо так выбрать таксоны 1 и 2, чтобы **сумма всех ветвей дерева была минимальна**

Методы иерархической кластеризации. NJ

Итак, имеем

$$\sum_{k=3}^{N}(D_{1k}+D_{2k}) = (N-2)L_{1X} + (N-2)L_{2X} + 2(N-2)L_{XY} + 2\sum_{i=3}^{N} L_{iY} \qquad (2)$$

$$S_{12} = L_{XY} + L_{1X} + L_{2X} + \sum_{i=3}^{N} L_{iY} \qquad (3)$$

Выражаем L_XY из (2) и подставляем в (3)

$$L_{XY} = \frac{1}{2(N-2)}\left(\sum_{k=3}^{N}(D_{1k}+D_{2k}) - (N-2)(L_{1X}+L_{2X}) - 2\sum_{i=3}^{N} L_{iY}\right)$$

$$S_{12} = \frac{1}{2(N-2)}\left(\sum_{k=3}^{N}(D_{1k}+D_{2k}) - (N-2)(L_{1X}+L_{2X}) - 2\sum_{i=3}^{N} L_{iY}\right) +$$

$$+ (L_{1X}+L_{2X}) + \sum_{i=3}^{N} L_{iY} = \frac{\sum_{k=3}^{N}(D_{1k}+D_{2k})}{2(N-2)} + \frac{L_{1X}+L_{2X}}{2} + \frac{N-3}{N-2}\sum_{i=3}^{N} L_{iY}$$

Методы иерархической кластеризации. NJ

$$Заметим, что\ L_{1X} + L_{2X} = D_{12},\ \ а\ также,\ по\ аналогии\ с\ (1),$$

$$\sum_{i=3}^{N} L_{iY} = \frac{1}{N-3} \sum_{3 \le i < j} D_{ij}$$

$$Теперь S_{12}\ принимает\ вид$$

$$S_{12} = \frac{\sum_{i=3}^{N} (D_{1i} + D_{2i})}{2(N-2)} + \frac{D_{12}}{2} + \frac{\sum_{3 \le i < j} D_{ij}}{N-2}$$

$$Для\ произвольных\ соседей\ k\ и\ l\ получим$$

$$S_{kl} = \frac{\sum_{i \ne k,l}^{N} (D_{ki} + D_{li})}{2(N-2)} + \frac{D_{kl}}{2} + \frac{\sum_{i < j, i \ne k,l} D_{ij}}{N-2}$$

Здесь все слагаемые так или иначе зависят от *k* и *l*, что неудобно.

Попробуем упростить.

Методы иерархической кластеризации. NJ

$$S_{kl} = \frac{\sum_{i \neq k,l}^{N} (D_{ki} + D_{li})}{2(N-2)} + \frac{D_{kl}}{2} + \frac{\sum_{i<j, i \neq k,l} D_{ij}}{N-2} =$$

$$= \frac{\sum_{i}^{N} D_{ki} - D_{kk} + \sum_{i}^{N} D_{li} - D_{ll} + 2\left(\sum_{i<j} D_{ij} - \sum_{i}^{N} D_{ki} - \sum_{i}^{N} D_{li} \right)}{2(N-2)} + \frac{D_{kl}}{2} =$$

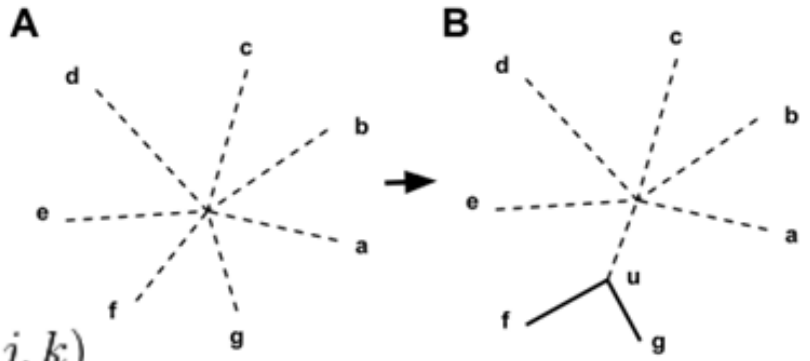
$$= \frac{2\sum_{i<j} D_{ij} - \sum_{i}^{N} D_{ki} - \sum_{i}^{N} D_{li}}{2(N-2)} + \frac{D_{kl}}{2}$$

Заметим, что $\sum_{i<j} D_{ij}$ *– не зависит от* k *и* l.

Поэтому домножив на $2(N-2)$ *и вычтя эту постоянную, получим* $Q_{kl} = D_{kl}(N-2) - \sum_{i}^{N} D_{ki} - \sum_{i}^{N} D_{li}$

Методы иерархической кластеризации. NJ

Пошаговая реализация:



1. По текущей матрице расстояний $d(i, j)$ рассчитывается Q-матрица

$$Q(i,j) = (n-2)d(i,j) - \sum_{k=1}^{n} d(i,k) - \sum_{k=1}^{n} d(j,k)$$

2. Ищется пара различных таксонов i и j, для которых значение $Q(i,j)$ наименьшее. Эти таксоны присоединяются к новому узлу u, который, в свою очередь, соединяется с центральным <виртуальным> узлом.

3. Рассчитывается расстояние от каждого из присоединенных таксонов к новому узлу

$$\delta(f,u) = \frac{1}{2}d(f,g) + \frac{1}{2(n-2)}\left[\sum_{k=1}^{n} d(f,k) - \sum_{k=1}^{n} d(g,k)\right]$$

Заметим, что при таком определении

$$\delta(f,u) + \delta(g,u) = d(f,g)$$

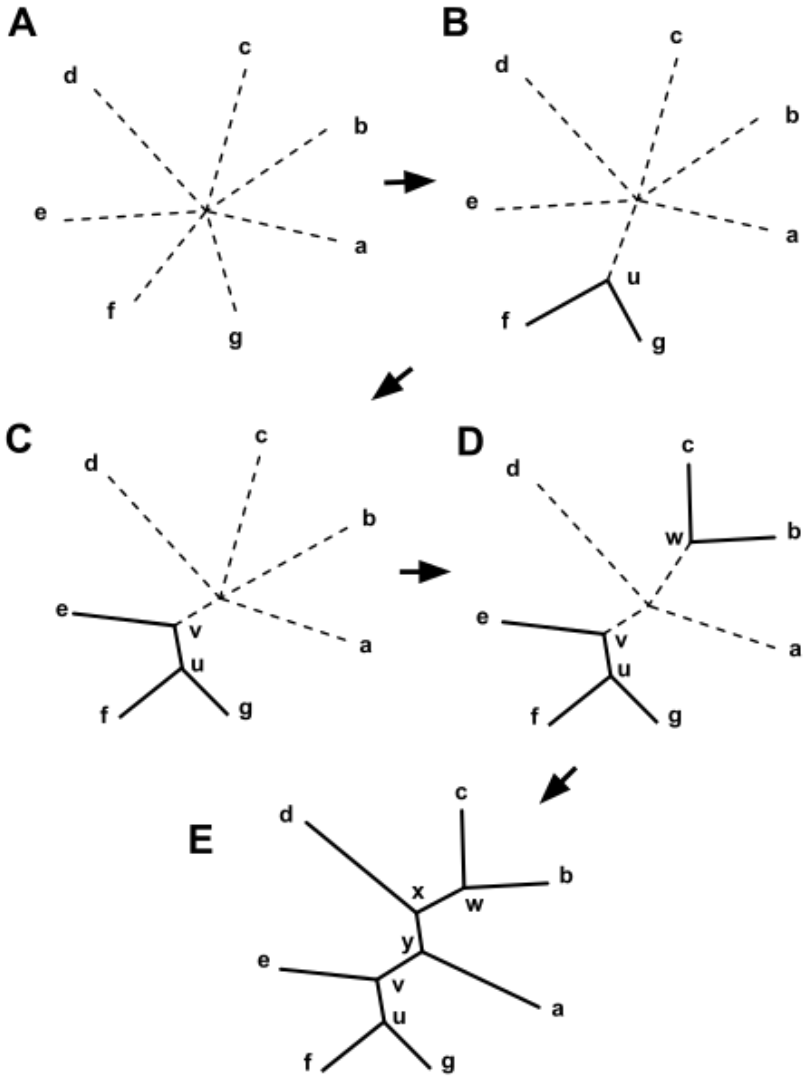
Методы иерархической кластеризации. NJ

4. Рассчитывается расстояние от каждого из оставшихся таксонов до нового узла

$$d(u,k) = \frac{1}{2}[d(f,k) + d(g,k) - d(f,g)]$$

5. Алгоритм запускается снова, заменяя пару соединенных таксонов на новый узел и используя расстояния, посчитанные на предыдущих шагах.

Длины ветвей, выходящих из одного узла в общем случае неравны.



Методы иерархической кластеризации. NJ

Те же 6 последовательностей, та же матрица расстояний:

D0	ATTTG	AGCGT	ACCGT	CGCGA	GGCGA	CGGGG
ATTTG	0	4	4	5	5	4
AGCGT		0	***1***	2	2	3
ACCGT			0	3	3	4
CGCGA				0	***1***	2
GGCGA					0	3
CGGGG						0

$$Q(i,j) = (n-2)d(i,j) - \sum_{k=1}^{n} d(i,k) - \sum_{k=1}^{n} d(j,k)$$

n = 6

Q0	ATTTG	AGCGT	ACCGT	CGCGA	GGCGA	CGGGG
ATTTG		-18	-21	-15	-16	-22
AGCGT			-23	-17	-18	-16
ACCGT				-16	-17	-15
CGCGA					-23	-21
GGCGA						-18
CGGGG						

Методы иерархической кластеризации. NJ

D0	ATTTG	AGCGT	ACCGT	CGCGA	GGCGA	CGGGG
ATTTG	0	4	4	5	5	4
AGCGT		0	***1***	2	2	3
ACCGT			0	3	3	4
CGCGA				0	***1***	2
GGCGA					0	3
CGGGG						0

$$\delta(f,u) = \frac{1}{2}d(f,g) + \frac{1}{2(n-2)}\left[\sum_{k=1}^{n} d(f,k) - \sum_{k=1}^{n} d(g,k)\right]$$

δ1	AGCGT, ACCGT
AGCGT	0, 125
ACCGT	0, 875

Методы иерархической кластеризации. NJ

D0	ATTTG	AGCGT	ACCGT	CGCGA	GGCGA	CGGGG
ATTTG	0	4	4	5	5	4
AGCGT		0	***1***	2	2	3
ACCGT			0	3	3	4
CGCGA				0	***1***	2
GGCGA					0	3
CGGGG						0

$$d(u,k) = \frac{1}{2}[d(f,k) + d(g,k) - d(f,g)]$$

D1	ATTTG	AGCGT, ACCGT	CGCGA	GGCGA	CGGGG
ATTTG	0	3,5	5	5	4
AGCGT, ACCGT		0	2	2	3
CGCGA			0	1	2
GGCGA				0	3
CGGGG					0

Методы иерархической кластеризации. NJ

D1	ATTTG	AGCGT, ACCGT	CGCGA	GGCGA	CGGGG
ATTTG	0	3,5	5	5	4
AGCGT, ACCGT		0	2	2	3
CGCGA			0	1	2
GGCGA				0	3
CGGGG					0

$$Q(i,j) = (n-2)d(i,j) - \sum_{k=1}^{n} d(i,k) - \sum_{k=1}^{n} d(j,k)$$

n = 5

Q1	ATTTG	AGCGT, ACCGT	CGCGA	GGCGA	CGGGG
ATTTG		-17,5	-12,5	-13,5	-17,5
AGCGT, ACCGT			-14,5	-15,5	-13,5
CGCGA				**-18**	-16
GGCGA					-14
CGGGG					

Методы иерархической кластеризации. NJ

D1	ATTTG	AGCGT, ACCGT	CGCGA	GGCGA	CGGGG
ATTTG	0	3,5	5	5	4
AGCGT, ACCGT		0	2	2	3
CGCGA			0	1	2
GGCGA				0	3
CGGGG					0

δ2	CGCGA, GGCGA
CGCGA	0,333
GGCGA	0,667

$$\delta(f,u) = \frac{1}{2}d(f,g) + \frac{1}{2(n-2)}\left[\sum_{k=1}^{n} d(f,k) - \sum_{k=1}^{n} d(g,k)\right]$$

$$d(u,k) = \frac{1}{2}[d(f,k) + d(g,k) - d(f,g)]$$

D2	ATTTG	AGCGT, ACCGT	CGCGA, GGCGA	CGGGG
ATTTG	0	3,5	4,5	4
AGCGT, ACCGT		0	1,5	3
CGCGA, GGCGA			0	2
CGGGG				0

Методы иерархической кластеризации. NJ

D2	ATTTG	AGCGT, ACCGT	CGCGA, GGCGA	CGGGG
ATTTG	0	3,5	4,5	4
AGCGT, ACCGT		0	1,5	3
CGCGA, GGCGA			0	2
CGGGG				0

Q2	ATTTG	AGCGT, ACCGT	CGCGA, GGCGA	CGGGG
ATTTG		-13	-11	-13
AGCGT, ACCGT			-13	-11
CGCGA, GGCGA				**-13**
CGGGG				

δ3	(CGCGA, GGCGA), CGGGG
CGCGA, GGCGA	0,75
CGGGG	1,25

Методы иерархической кластеризации. NJ

D2	ATTTG	AGCGT, ACCGT	CGCGA, GGCGA	CGGGG
ATTTG	0	3,5	4,5	4
AGCGT, ACCGT		0	1,5	3
CGCGA, GGCGA			0	2
CGGGG				0

Q2	ATTTG	AGCGT, ACCGT	CGCGA, GGCGA	CGGGG
ATTTG		-13	-11	-13
AGCGT, ACCGT			-13	-11
CGCGA, GGCGA				**-13**
CGGGG				

D3	ATTTG	AGCGT, ACCGT	(CGCGA, GGCGA), CGGGG
ATTTG	0	3,5	3,25
AGCGT, ACCGT		0	1,25
(CGCGA, GGCGA), CGGGG			0

Методы иерархической кластеризации. NJ

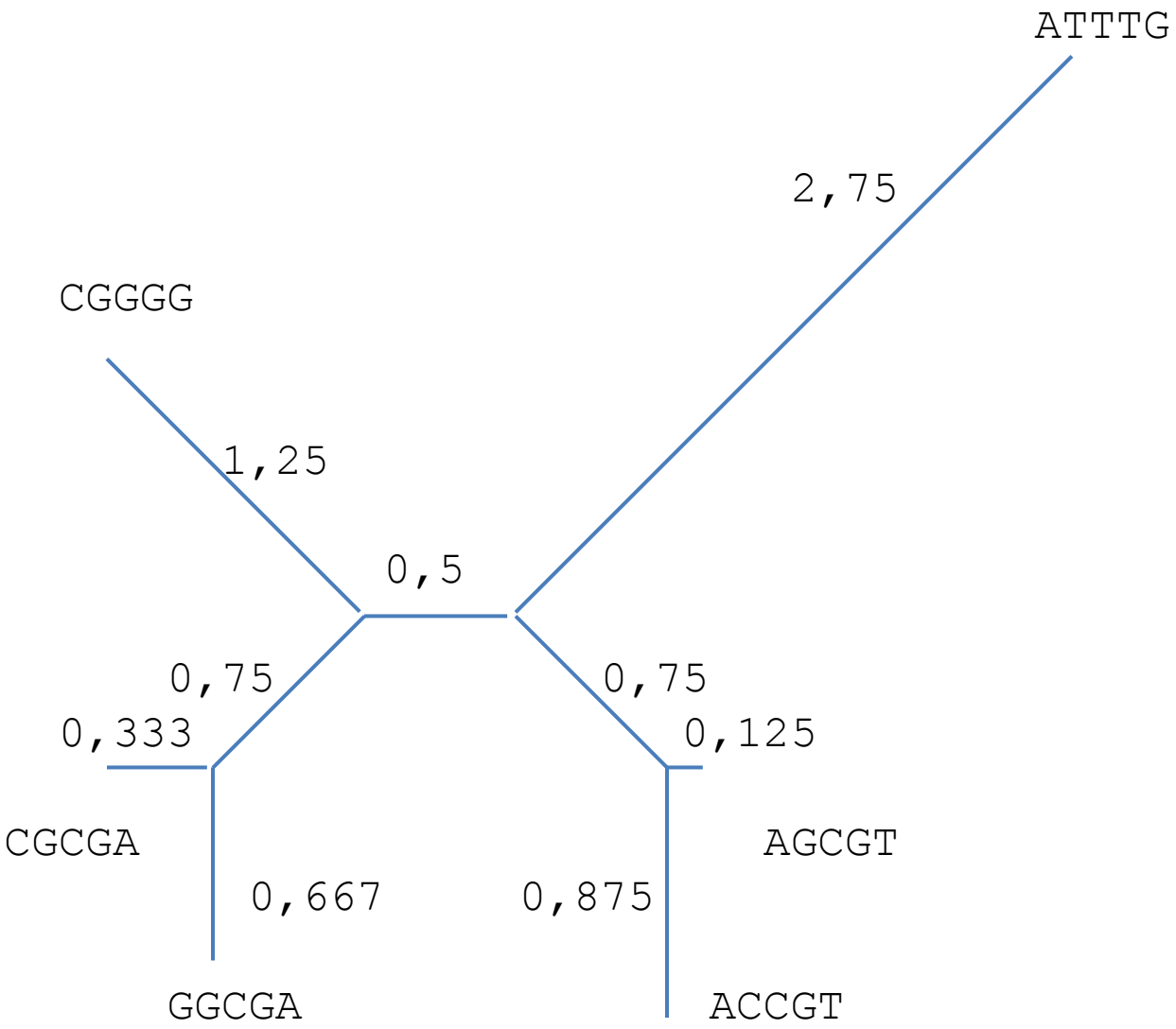
D3	ATTTG	AGCGT, ACCGT	(CGCGA, GGCGA), CGGGG
ATTTG	0	3,5	3,25
AGCGT, ACCGT		0	1,25
(CGCGA, GGCGA), CGGGG			0

Q3	ATTTG	AGCGT, ACCGT	(CGCGA, GGCGA), CGGGG
ATTTG		-8	-8
AGCGT, ACCGT			-8
(CGCGA, GGCGA), CGGGG			

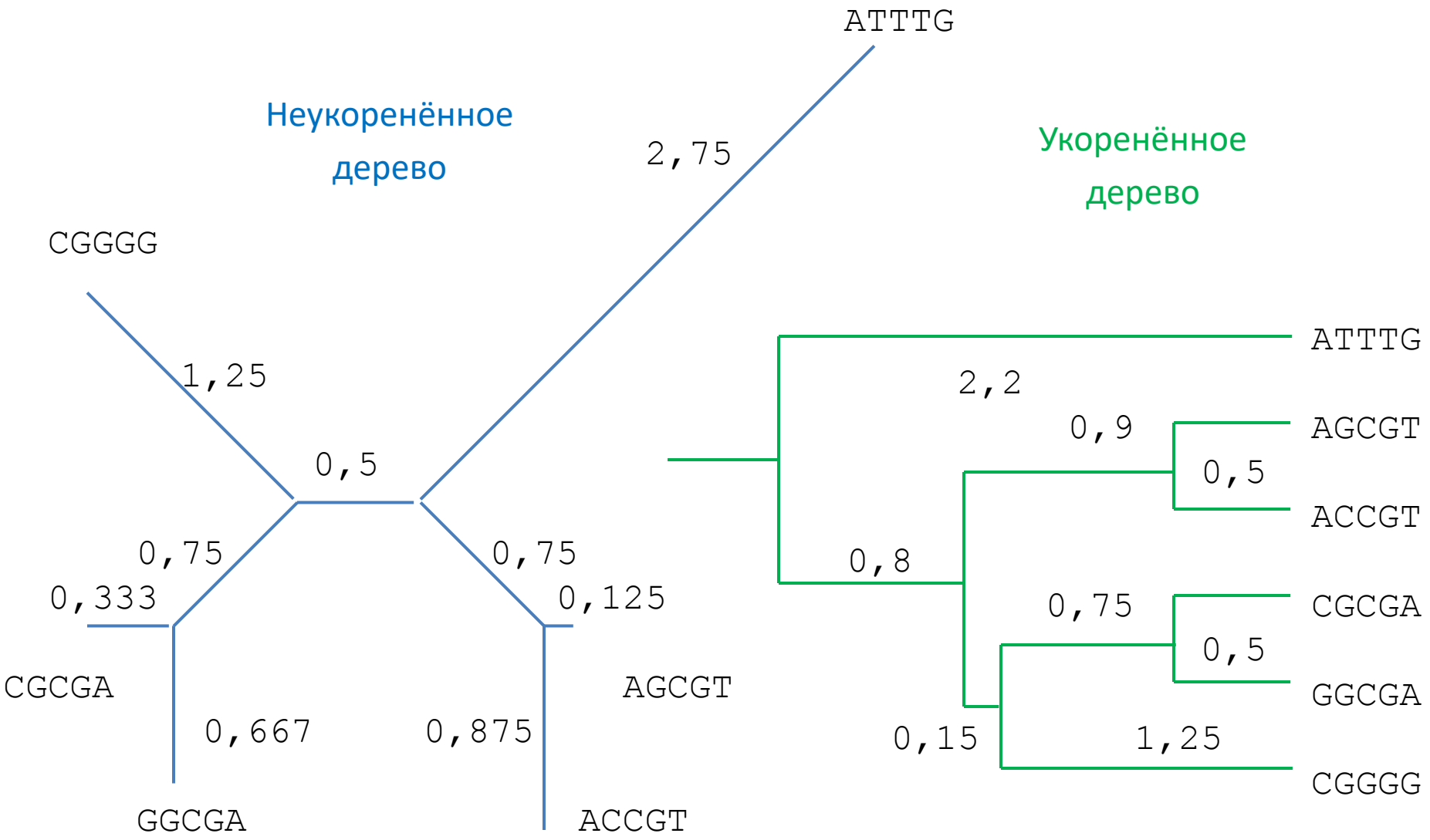
δ4	ATTTG, AGCGT, ACCGT
ATTTG	0,75
AGCGT, ACCGT	1,25

D4	AGCGT, ACCGT, ATTTG	CGCGA, GGCGA, CGGGG
AGCGT, ACCGT, ATTTG	0	0,5
CGCGA, GGCGA, CGGGG	0,5	0

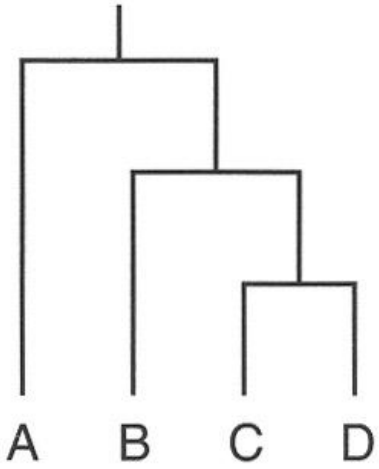
Методы иерархической кластеризации. NJ



Методы иерархической кластеризации. NJ и UPGMA

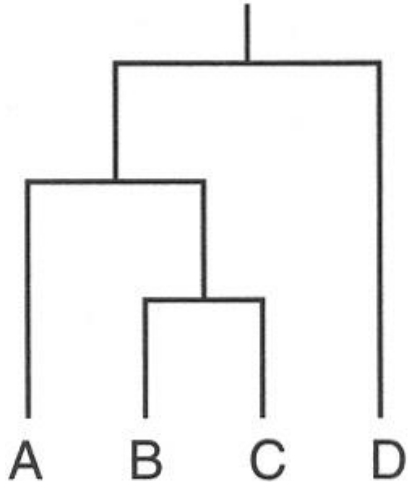


Филогенетические деревья – проблема переменной скорости эволюции

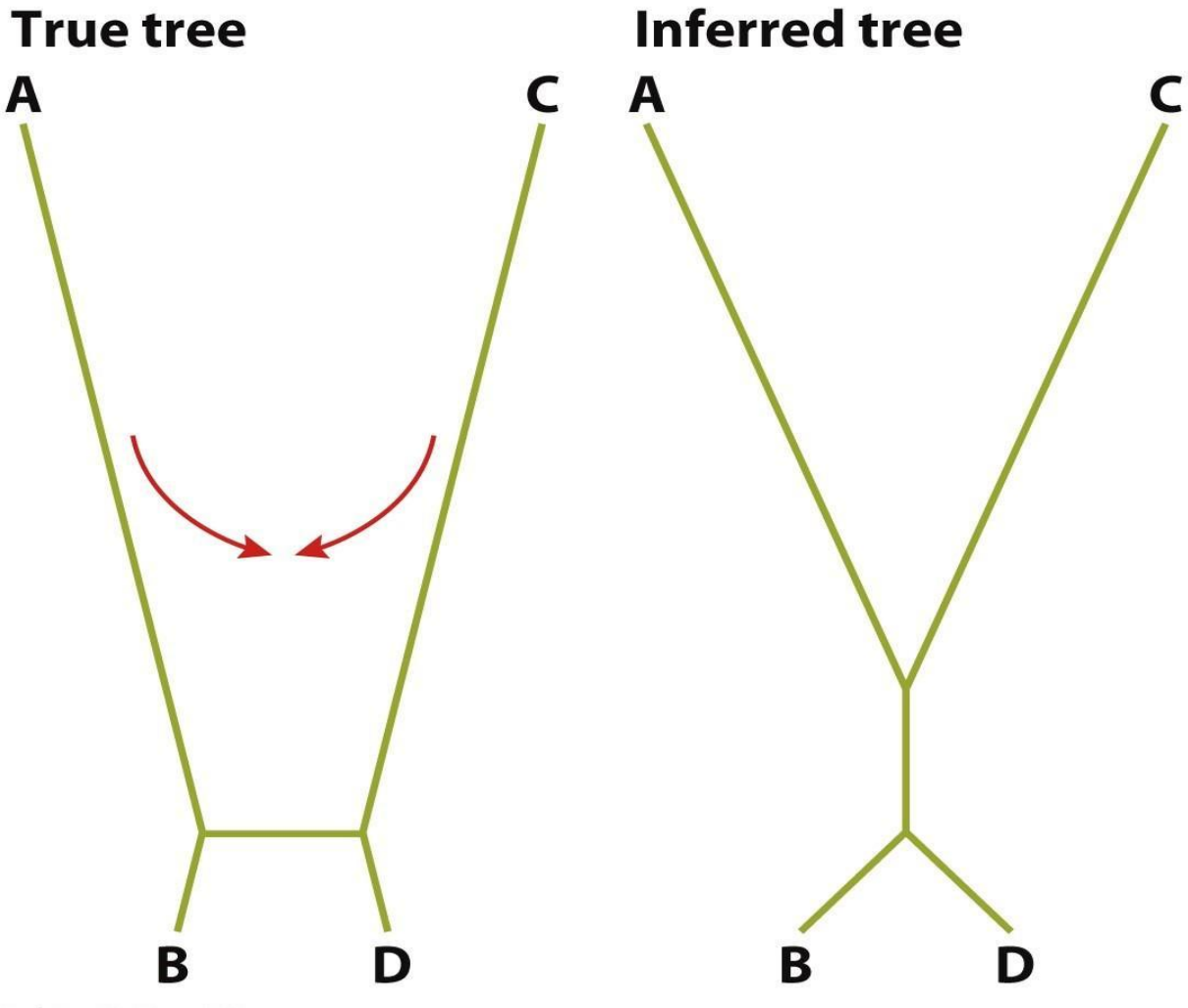


	A	B	C	D
A	0	3	3	3
B		0	2	2
C			0	1
D				0

	A	B	C	D
A	0	3	3	20
B		0	2	20
C			0	20
D				0



Филогенетические деревья – «притяжение длинных ветвей»

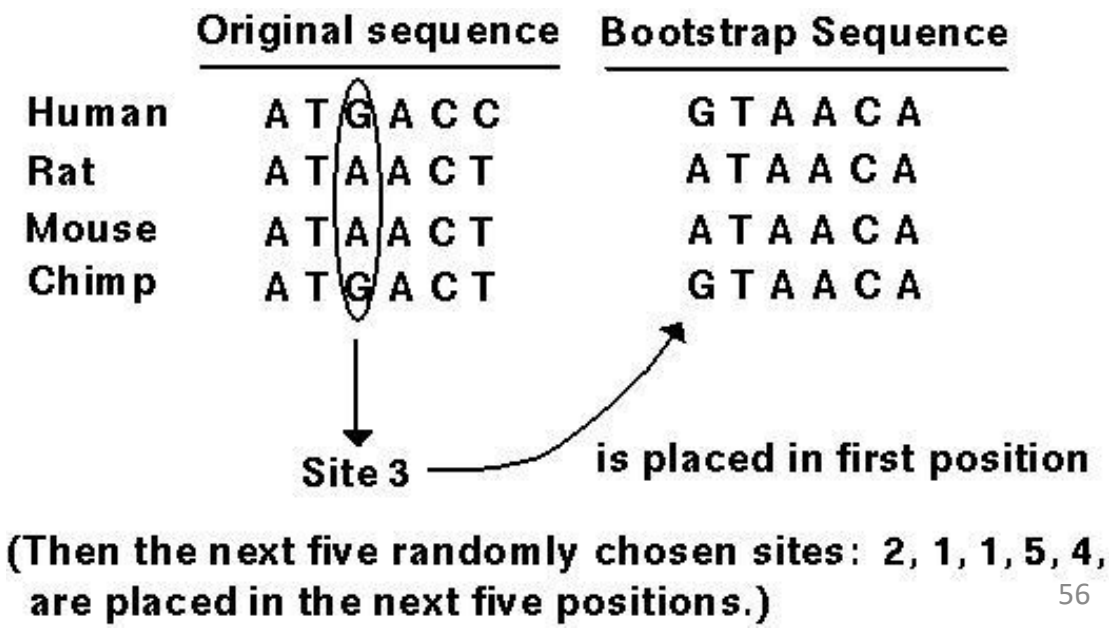


Филогенетические деревья – методы проверки

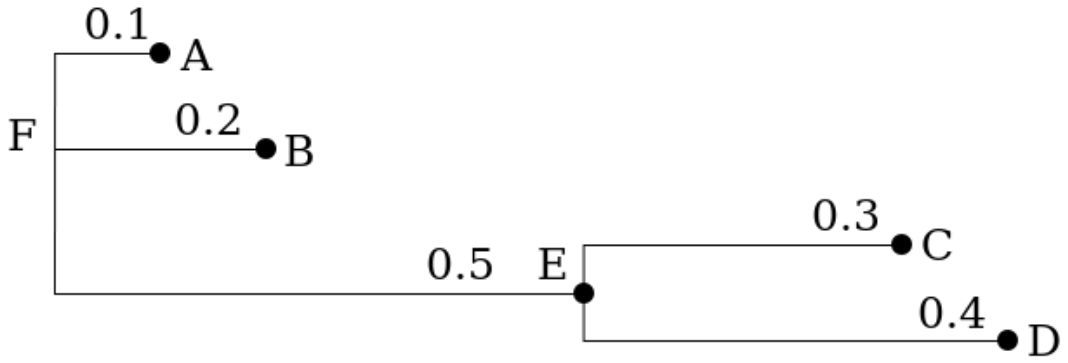
1) **Использование внешней группы**, т.е. видов, которые заведомо более удалены ото всех видов, для которых строится дерево (приматы и корова);
2) **Сравнение деревьев**, полученных на основе разных характеристик. Очевидно, они должны быть согласованными;

Оценка результата с помощью формальных статистических тестов:

3) Например, **построение дерева для подмножества последовательностей** из исходного множественного выравнивания должно дать поддерево дерева, полученного для этого выравнивания (jack-knife);
4) **Бутстреп** (boot strap).



Филогенетические деревья.
Скобочная формула (Newick format) (1986)



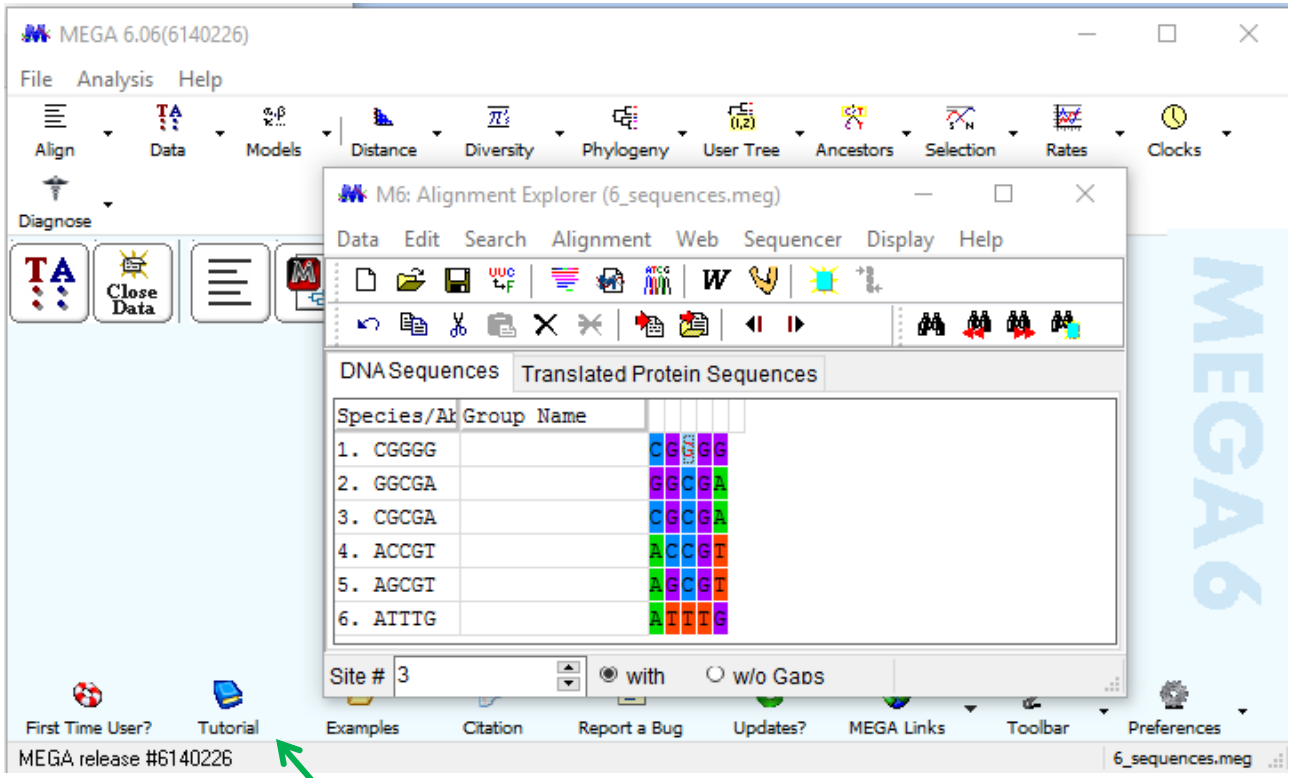
`(,,(,));`	*имена узлов не указаны*
`(A,B,(C,D));`	*указаны только имена листьев*
`(A,B,(C,D)E)F;`	*указаны имена всех узлов*
`(:0.1,:0.2,(:0.3,:0.4):0.5);`	*для всех узлов кроме корня указано расстояние до родительского узла*
`(:0.1,:0.2,(:0.3,:0.4):0.5):0.0;`	*для всех узлов указано расстояние до родительского узла*
`(A:0.1,B:0.2,(C:0.3,D:0.4):0.5);`	*указаны имена листьев и расстояния*
`(A:0.1,B:0.2,(C:0.3,D:0.4)E:0.5)F;`	*указаны все имена и расстояния*

Molecular Evolutionary Genetics Analysis (1993)

MEGA Molecular Evolutionary Genetics Analysis

Масатоси Неи

EN = 4

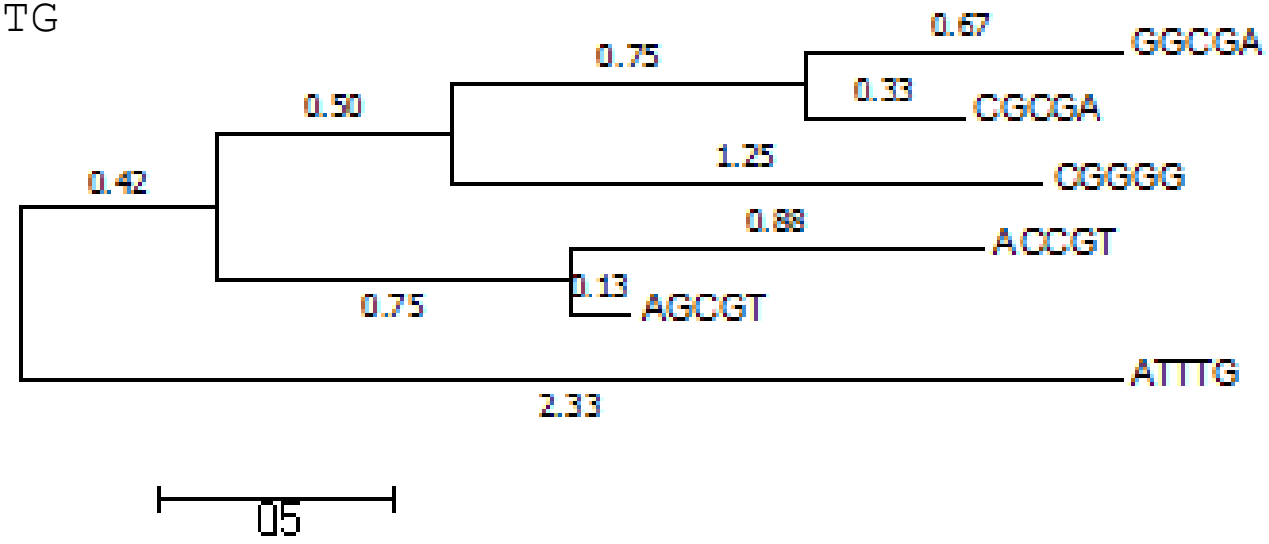


Tutorial (in English)

Molecular Evolutionary Genetics Analysis (1993)

CGGGG CGGGG
GGCGA GGCGA
CGCGA CGCGA
ACCGT ACCGT
AGCGT AGCGT
ATTTG ATTTG

This is a Neighbour-joining tree.

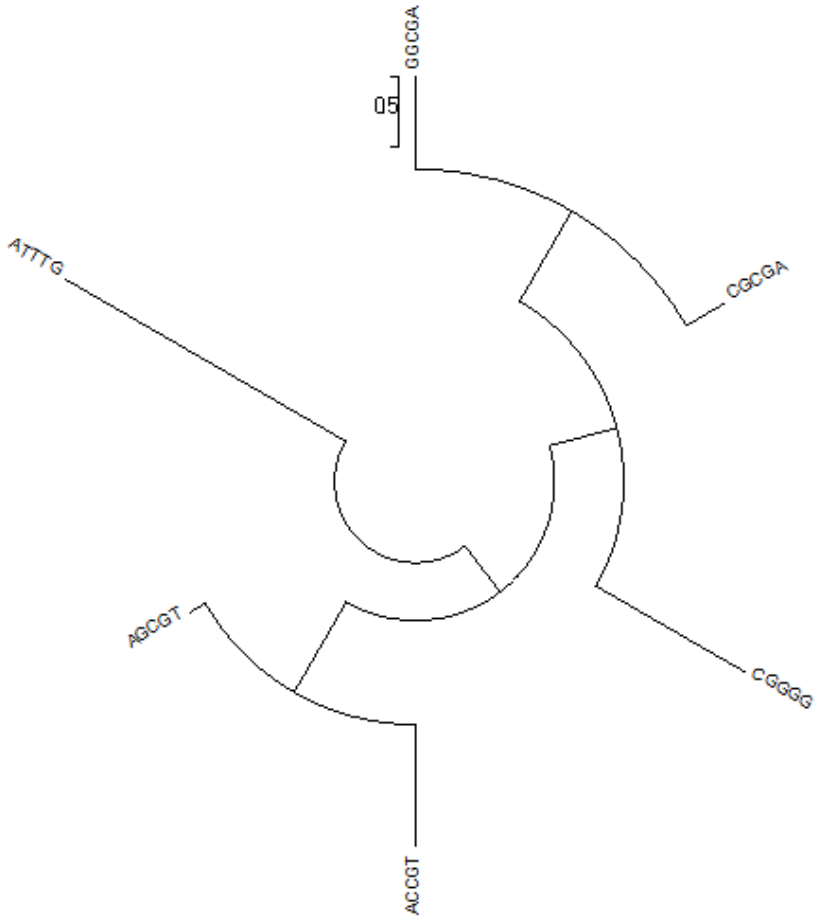


Не корень! ⟶

(((GGCGA:0.667,CGCGA:0.333):0.750,CGGGG:1.250):0.500,
(ACCGT:0.875,AGCGT:0.125):0.750):0.417,ATTTG:2.333)

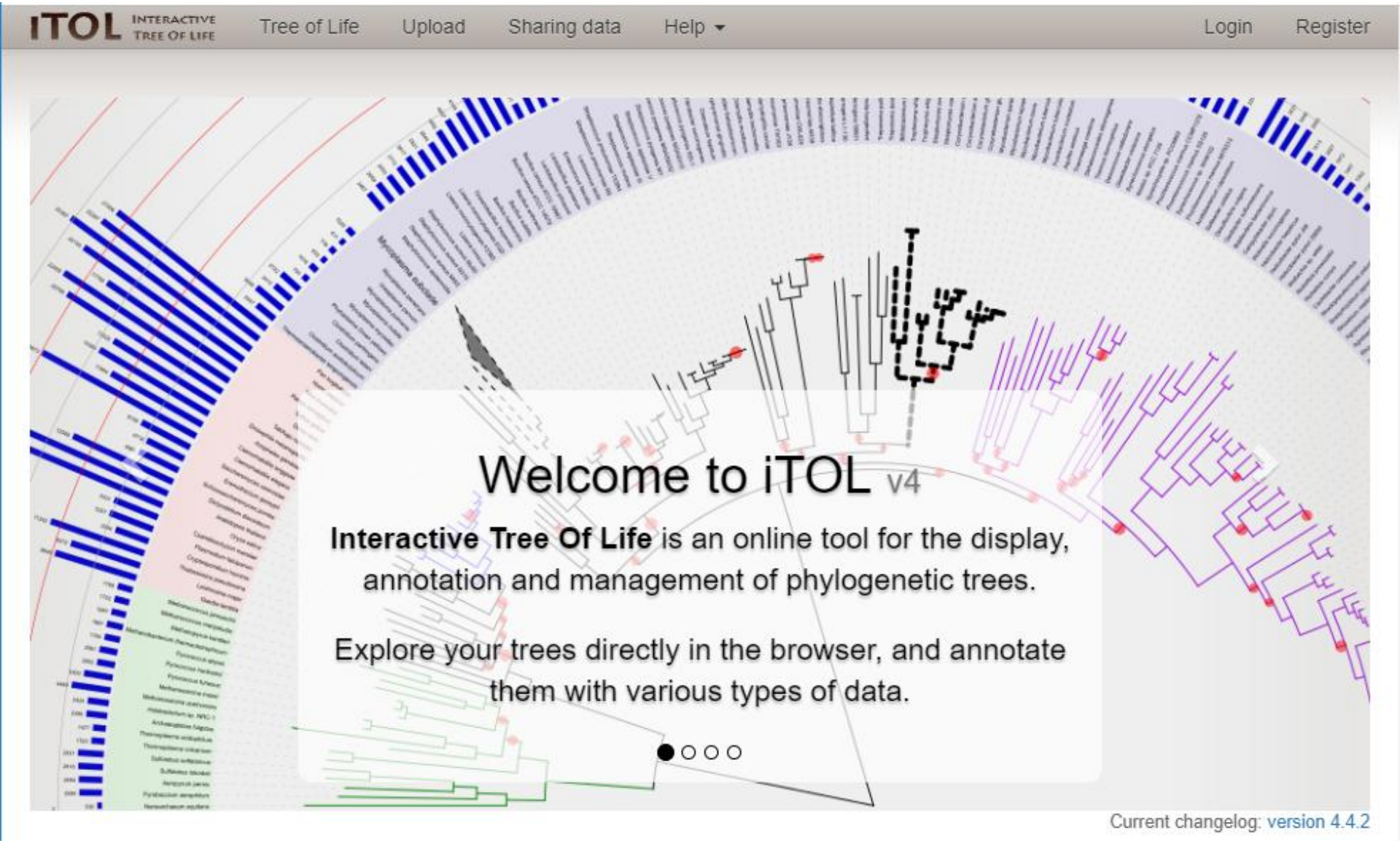
Molecular Evolutionary Genetics Analysis (1993)

CGGGG	CGGGG
GGCGA	GGCGA
CGCGA	CGCGA
ACCGT	ACCGT
AGCGT	AGCGT
ATTTG	ATTTG



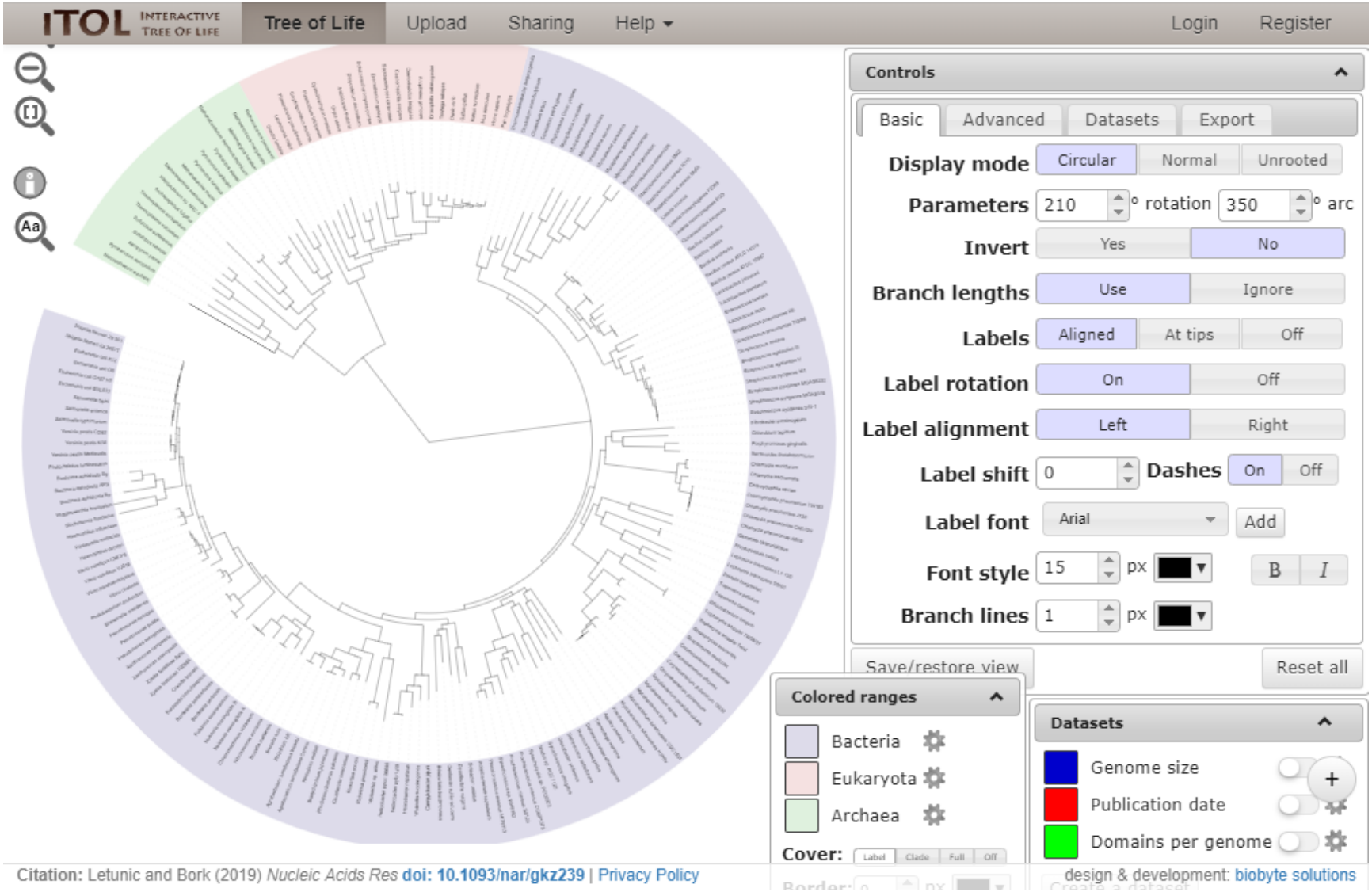
(((GGCGA:0.667,CGCGA:0.333):0.750,CGGGG:1.250):0.500,
(ACCGT:0.875,AGCGT:0.125):0.750):0.417,ATTTG:2.333)

Interactive Tree of Life



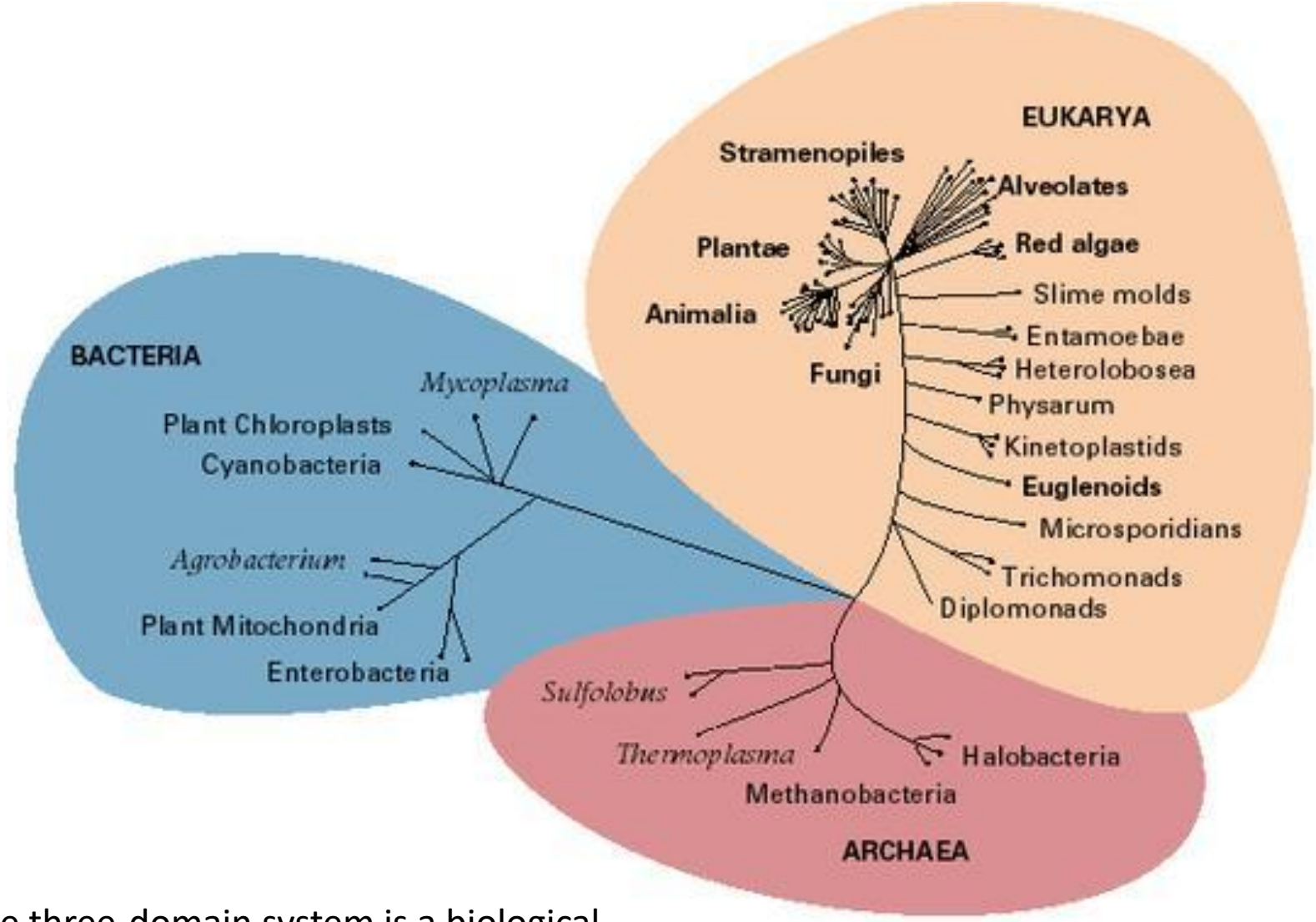
http://itol.embl.de

Interactive Tree of Life



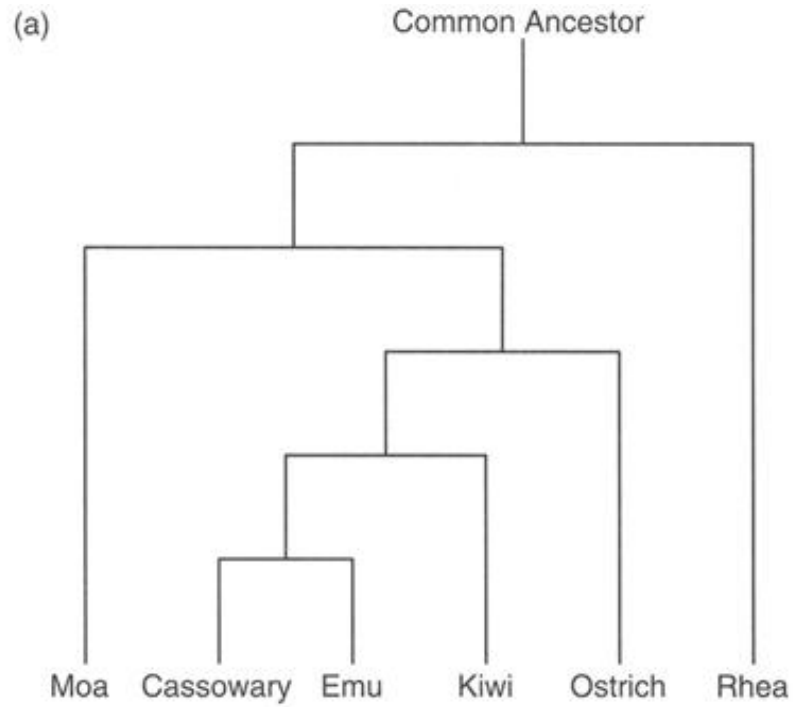
http://itol.embl.de

Phylogenetic trees: so different

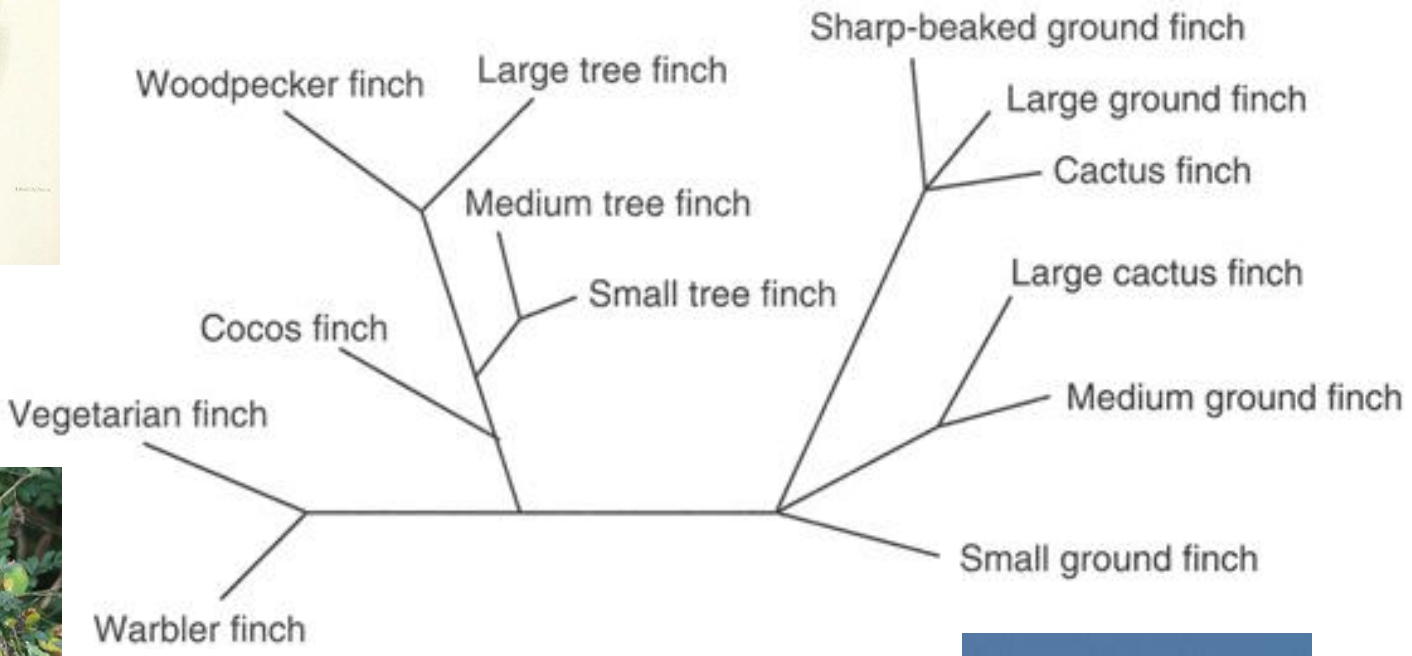


The three-domain system is a biological classification introduced by Carl Woese et al. in 1990

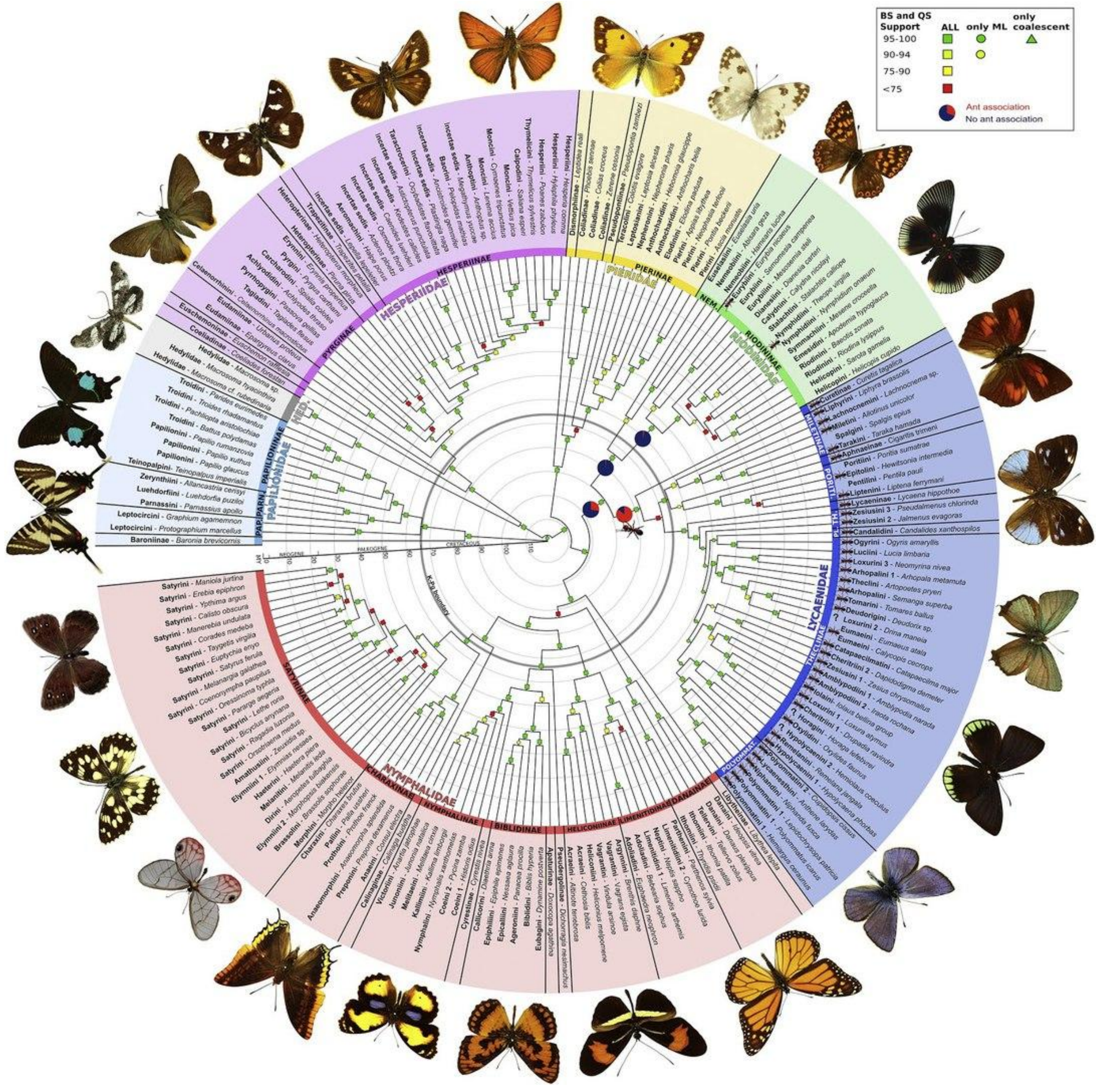
Филогенетические деревья – примеры



Филогенетические деревья – примеры



Филогенетические деревья – такие разные



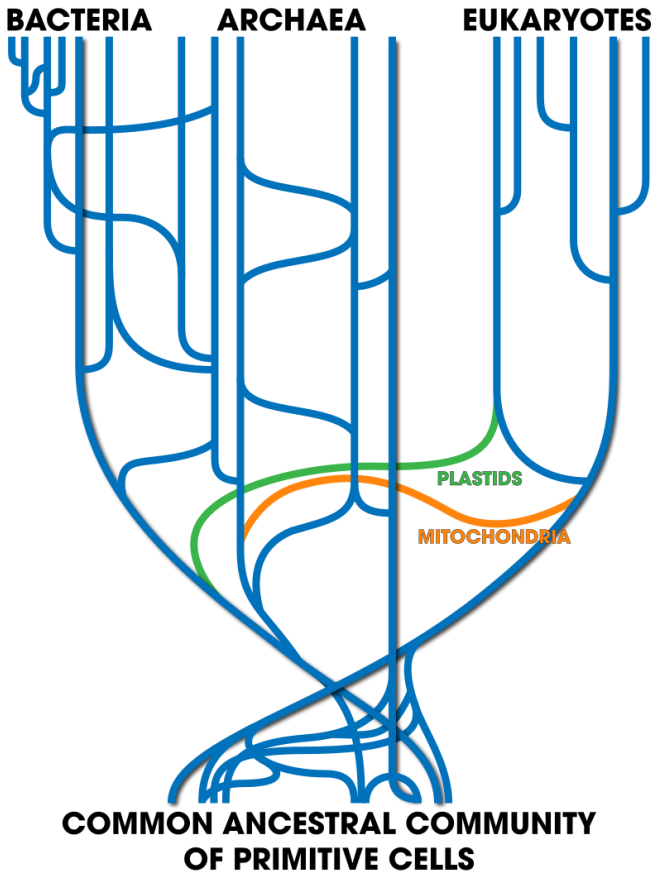
Филогенетические деревья. LUCA

Last Universal Common Ancestor (Darwin, 1859).

Современное представление о существовании такого организма основано на **сопоставлении геномов** (2000 – …).

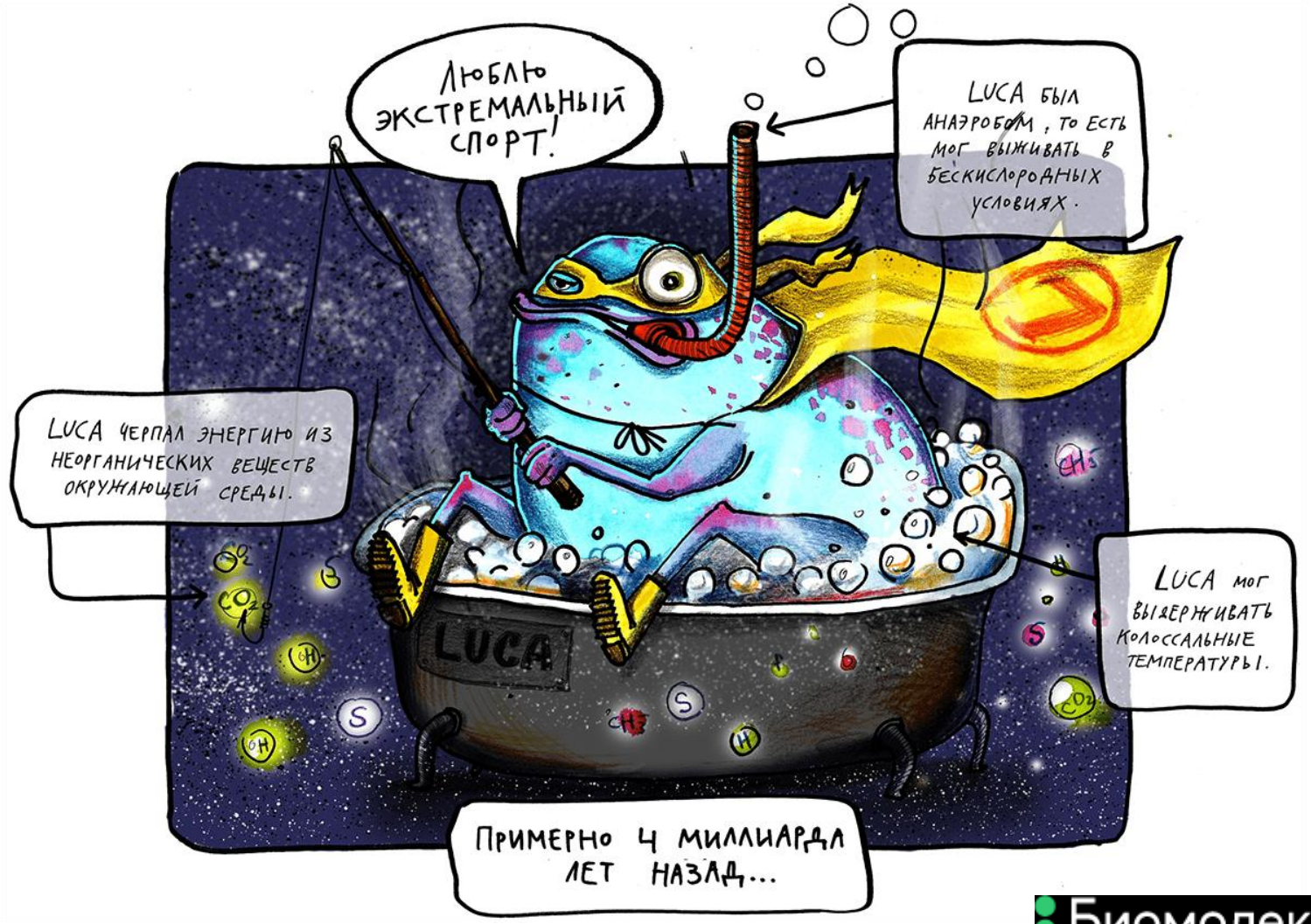
Показано, что наличие общих генов является скорее следствием их общего происхождения, а не горизонтального переноса (Theobald, 2010).

На основании анализа 6,1 млн белок-кодирующих генов можно предположить, что LUCA обладал 355 генами, схожими с генами современных организмов, что даёт возможность детально представить его метаболизм (2016).



Comparative genomics, minimal gene-sets and the last universal common ancestor (**Koonin**, 2003)

Last Universal Common Ancestor



Биомолекула

[В диких условиях: как жил последний всеобщий предок LUCA.]

Множественное выравнивание последовательностей

Цели:

- **Построение филогенетических деревьев**
- **Выявление консервативных остатков и мотивов**
- **Построение профилей (визуализация)**
- **Итеративное выявление удаленной гомологии**
- …

Алгоритмы:

- Динамическое программирование – не годится
- **Прогрессивное выравнивание**
- Скрытые марковские модели
- Квантовые компьютеры?

Прогрессивное выравнивание. Clustal

Clustal (Higgins and Sharp, 1988) выполняет постепенное выравнивание все новых последовательностей, начиная с наиболее <эволюционно> близких, ориентируясь на предварительно построенное на основании парных выравниваний (**сравнений**) филогенетическое дерево.

Алгоритм:

1) Экспресс-оценка сходства двух последовательностей вычисляется как число совпадающих остатков в словах длины K (K = 1-2 для белковых последовательностей и 2-4 для нуклеотидных) за вычетом штрафа за сделанные вставки.
2) Методом UPGMA (позже NJ) рассчитывается направляющее дерево, по которому затем рассчитываются веса последовательностей, причем более близкие последовательности получают меньшие веса.
3) Согласно направляющего дерева выбираются наиболее близкие последовательности и выполняется их выравнивание методом динамического программирования с использованием матрицы замен и штрафов за открытие/расширение вставок, с полученным выравниванием сопоставляются всё новые последовательности.

Прогрессивное выравнивание. Clustal

Особое внимание уделено значениям штрафов за вставки. Введена их зависимость от:

А) типа сопоставляемого и предшествующего остатков;

Б) степени близости последовательностей;

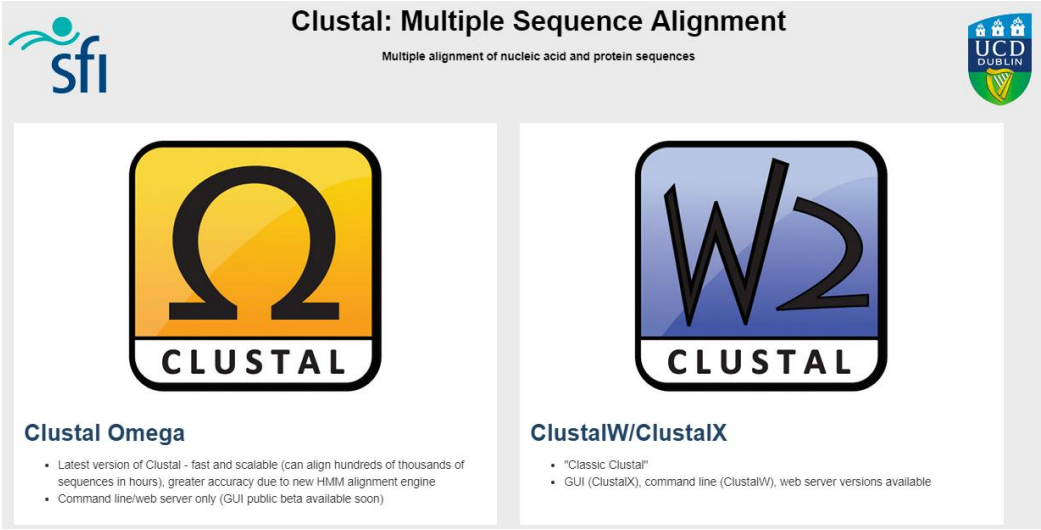
В) длин рассматриваемых последовательностей;

Г) наличия вставок в уже имеющемся выравнивании;

Д) характера аминокислотной последовательности.

One gap, always gap

Текущие версии: **ClustalW2** и **Clustal Omega (использует СММ)**



Итеративное выравнивание. MUSCLE

MUSCLE (MUltiple Sequence Comparison by Log-Expectation) (Edgar, 2004)

Три стадии:

1) **быстрое** «черновое» **множественное выравнивание** (попарные глобальные выравнивания; оценка сходства как доля совпадающих позиций; построение направляющего дерева методом UPGMA или NJ; прогрессивное выравнивание)
2) **улучшенное множественное выравнивание** (оценка сходства как доля совпадающих позиций в текущем множественном выравнивании; построение направляющего дерева через построение матрицы расстояний по Кимуре и ее кластеризацию; сравнение текущего дерева с построенным ранее; пересчет выравнивания для отличающихся узлов; повторение до сходимости)
3) **уточнение выравнивания** (удаление произвольного узла для разбиения дерева на два; построение профилей для каждого поддерева и их выравнивание; расчет суммы парных оценок в получающемся множественном выравнивании; перебор всех узлов от листьев к корню и выбор выравнивания с максимальной суммой)

http://www.drive5.com/muscle/ 25000+ цитирований (2018)

Множественное выравнивание последовательностей

Цели:

- **Построение филогенетических деревьев**
- **Выявление консервативных остатков и мотивов**
- **Построение профилей (визуализация)**
- **Итеративное выявление удаленной гомологии**
- …

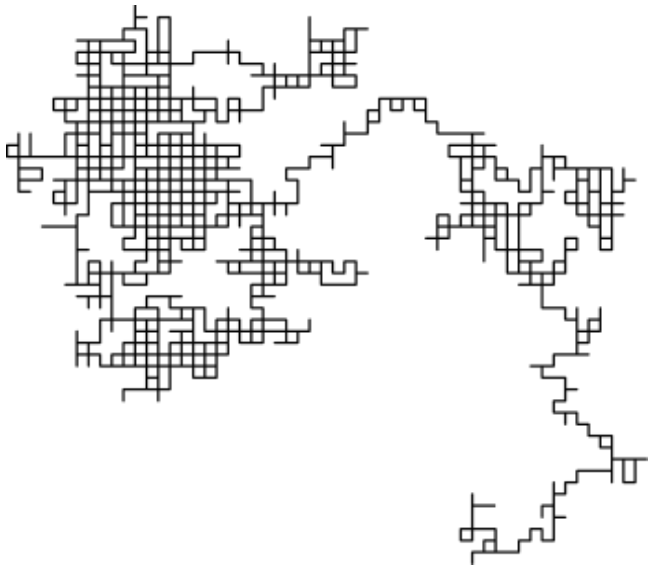
Алгоритмы:

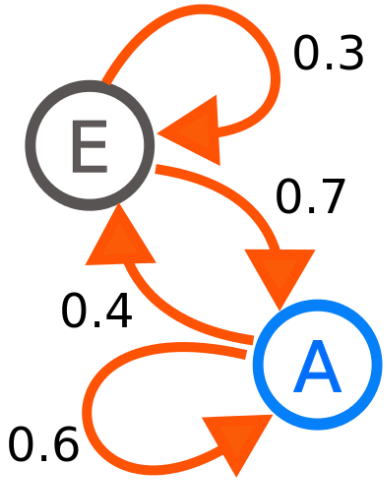
- Динамическое программирование – не годится
- **Прогрессивное выравнивание**
- **Скрытые марковские модели**
- Квантовые компьютеры?

Скрытые марковские модели

Марковский процесс — случайный процесс, эволюция которого после любого заданного значения временно́го параметра *t* не зависит от эволюции, предшествовавшей *t*, при условии, что значение процесса в этот момент фиксировано («будущее» процесса не зависит от «прошлого» при известном «настоящем») (бросание/перекатывание игрального кубика, случайное блуждание,…).

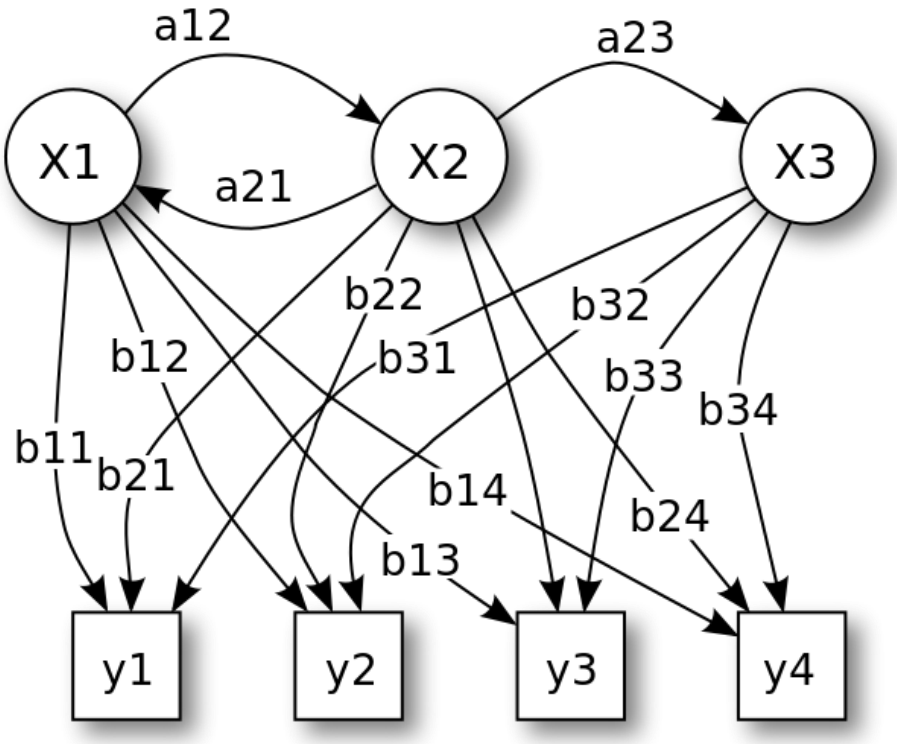
А.А. Марков (ст.)
(1856 -1922)





Скрытые марковские модели

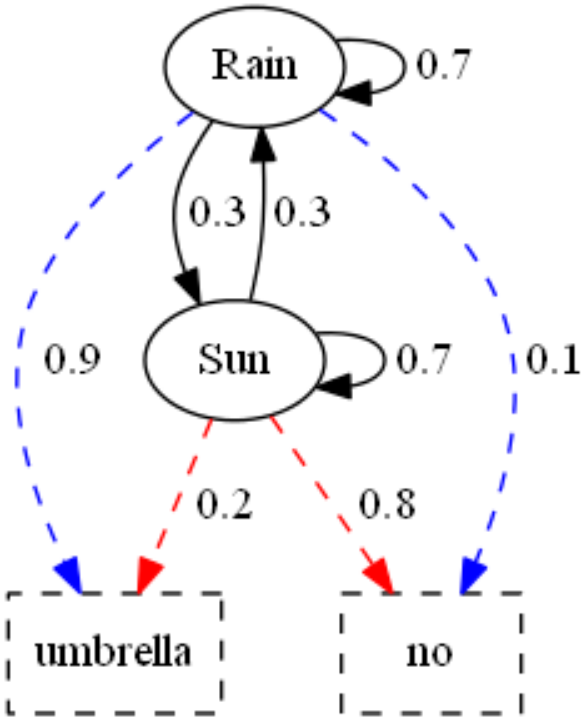
Скрытая марковская модель (СММ, hidden Markov model, HMM) — статистическая модель, имитирующая работу процесса, похожего на марковский процесс с неизвестными параметрами. Задачей ставится разгадывание неизвестных параметров на основе наблюдаемых.



Скрытые марковские модели. Пример

За день погода может поменяться с вероятностью 0,3, и если на улице идет дождь, то некий человек приносит зонтик с вероятностью 0,9, а если солнечно — то с вероятностью 0,2.

За рабочую неделю вы заметили, что он не принес зонтик лишь в среду. С какой вероятностью во вторник шел дождь?



СММ. Казино

Вы играете с игроком от казино:

1. Делаются ставки;
2. Вы бросаете кубик;
3. Соперник бросает кубик;
4. Обе ставки забирает тот, у кого выпало больше.

Вы бросаете «честный» кубик, для которого P(1) = P(2) = … = P(6) = 1/6.

Соперник может бросить как «честный» кубик, так и «нечестный» кубик (со смещенным центром тяжести), для которого, например, P(1) = … = P(5) = 1/10, P(6) = 1/2.

Вероятность смены кубика перед очередным броском **0,4**.

При ряде бросков у Вашего соперника выпала следующая последовательность:

12156216241461461361366616646616366163661636165

Какие вопросы могут у Вас возникнуть?

СММ. Казино

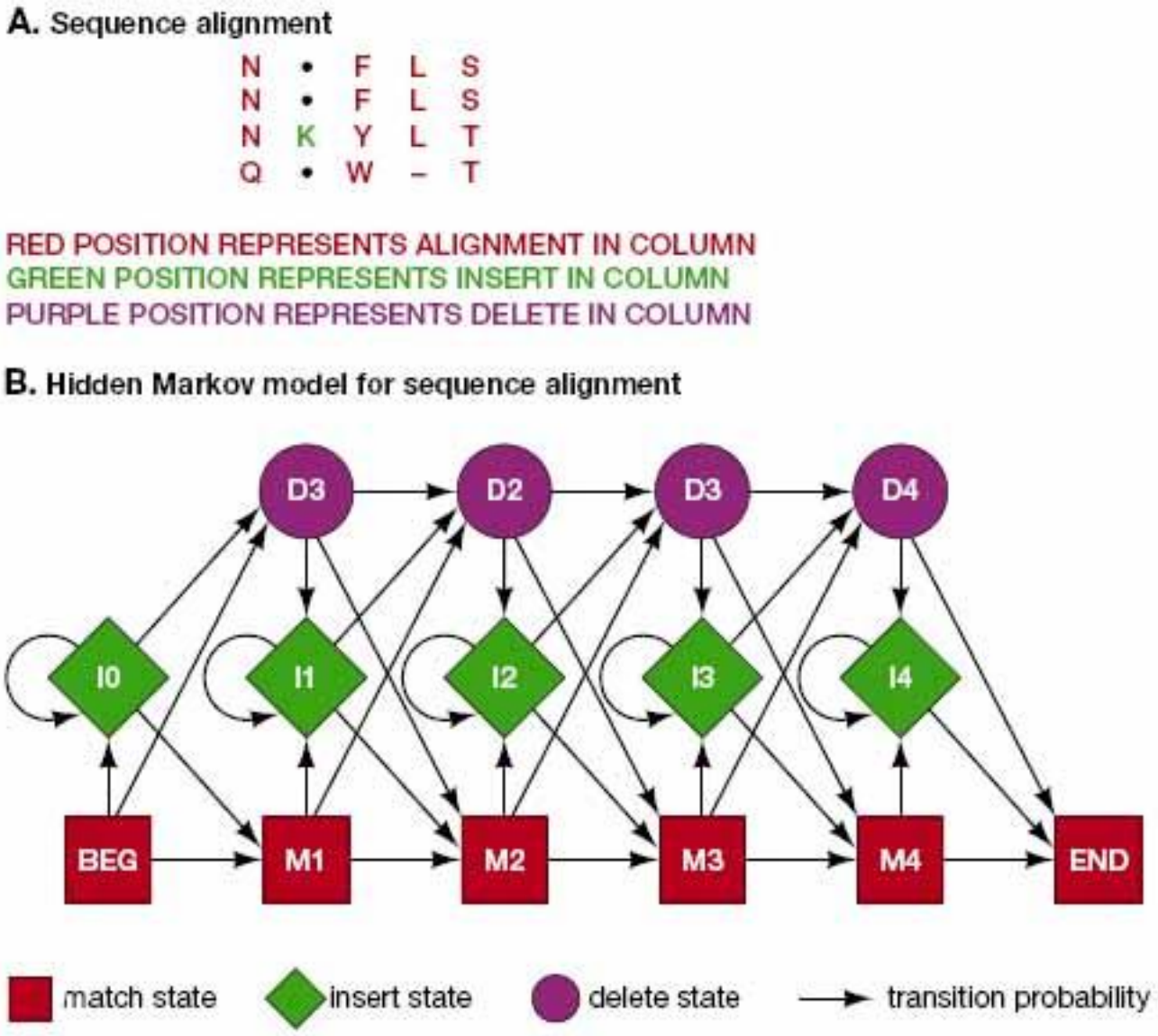
При ряде бросков у Вашего соперника выпала следующая последовательность:

12156216241461461361366616646616366163661636165

Какие вопросы могут у Вас возникнуть?

1. Насколько вероятно выпадение такой последовательности в рамках известной модели казино?
   В терминах СММ это **задача ОЦЕНКИ**
2. Считая модель верной, какие фрагменты последовательности могли быть сгенерированы «честным» кубиком, а какие «нечестным»?
   В терминах СММ это **задача ДЕШИФРОВКИ**
3. Насколько смещён центр тяжести в «нечестном» кубике? Насколько идеален «честный» кубик? Как часто соперник меняет кубики?
   Это **задача** определения параметров или **ОБУЧЕНИЯ** и она самая сложная.

CMM. Построение множественного выравнивания



CMM. Построение множественного выравнивания

Алгоритм:

- **Обучение.** Имея ряд невыровненных последовательностей, можно выровнять их и подогнать вероятности переходов и порождения остатков, чтобы определить модель, описывающую данный набор последовательностей.
- **Поиск гомологов.** Имея модель и исследуемую последовательность, можно посчитать вероятность того, то модель могла бы сгенерировать эту последовательность. Если вероятность достаточно высока, то рассматриваемая последовательность принадлежит тому же семейству, что и обучающие.

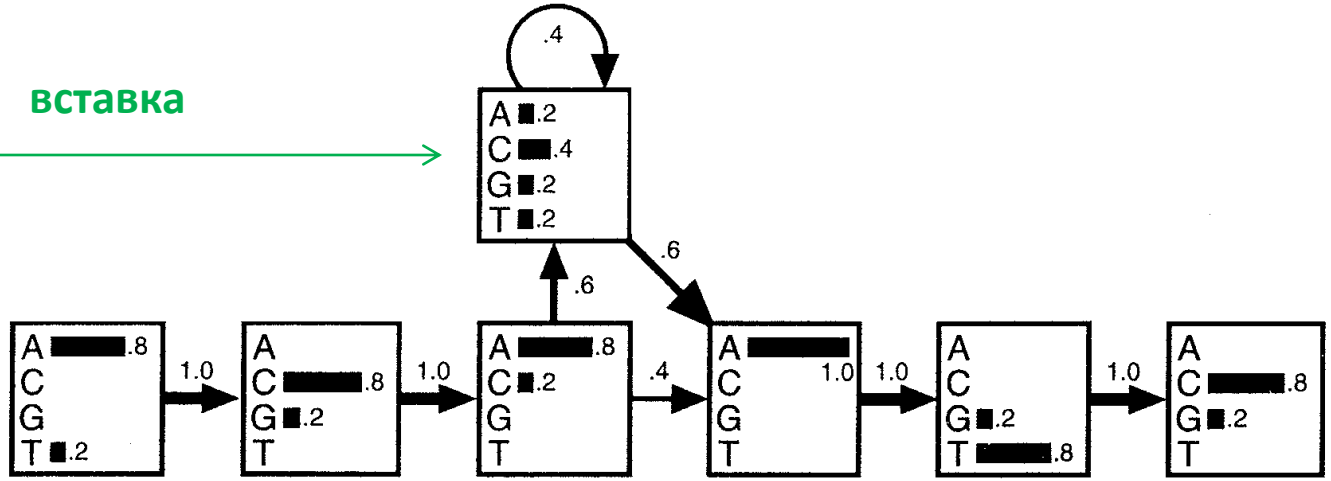
```
ACA---ATC
TCAACTATC
ACAC--AGC
AGA---ATC
ACCG--ATG
```

Построим?

СММ. Построение множественного выравнивания

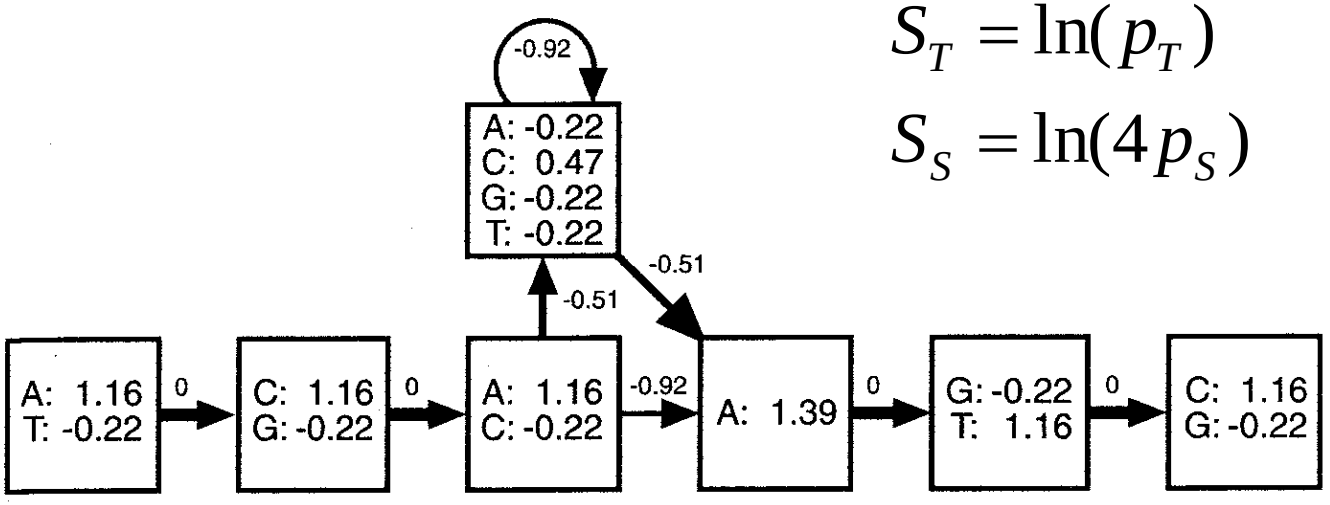
ACA---ATC
TCAACTATC
ACAC--AGC
AGA---ATC
ACCG--ATG

вставка



$$S_T = \ln(p_T)$$

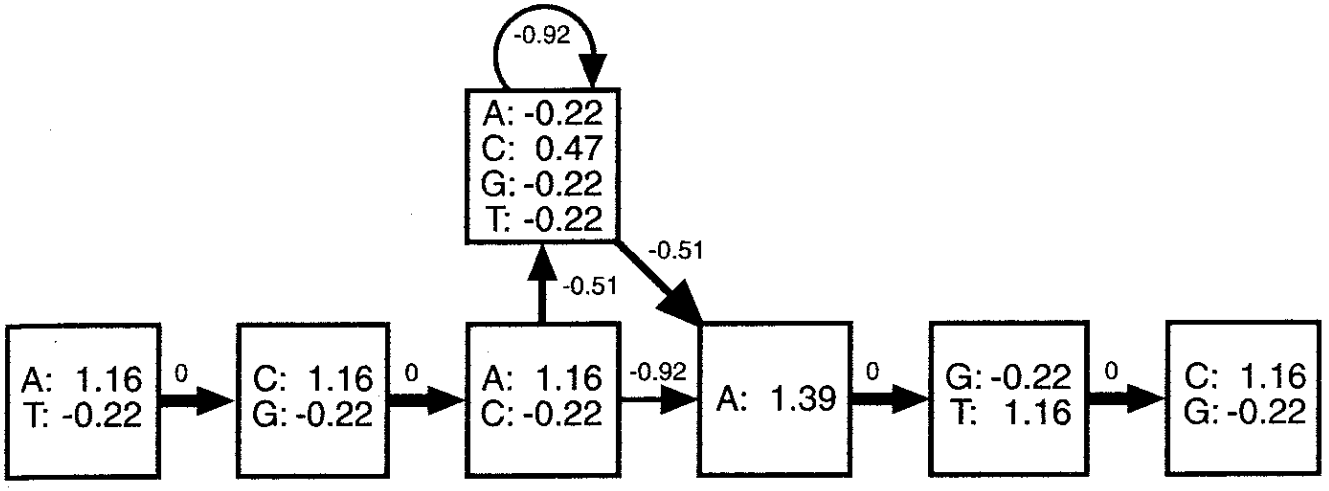
$$S_S = \ln(4p_S)$$



СММ. Построение множественного выравнивания

```
ACA---ATC
TCAACTATC
ACAC--AGC
AGA---ATC
ACCG--ATG
CGCGT-CGG
```

Посчитаем: описывает ли построенная модель новую последовательность?



СММ. Построение множественного выравнивания

ACA---ATC S = 1.16 + 0 + 1.16 + 0 + 1.16 - 0.92 + 1.39 + 0 + 1.16 + 0 + 1.16 = 6.29

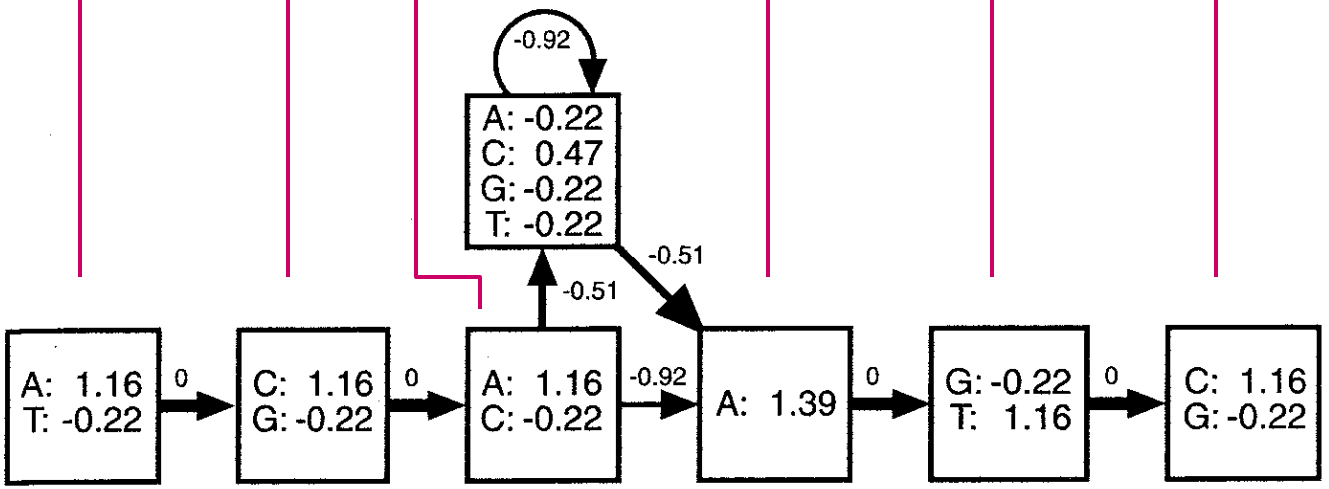
TCAACTATC ?

ACAC--AGC

AGA---ATC

ACCG--ATG

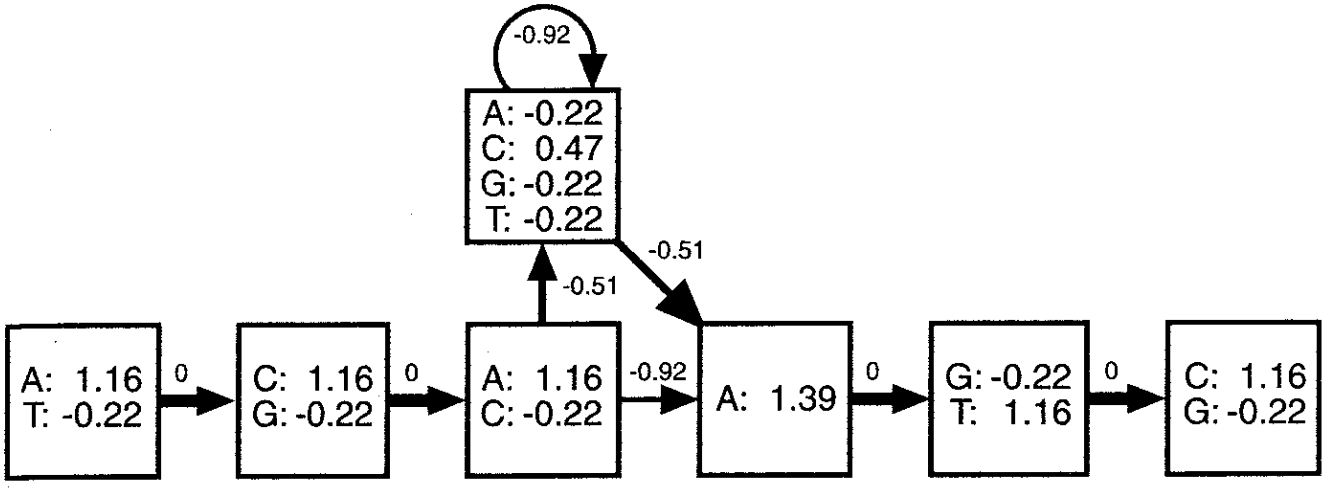
CGCGT-CGG ?



C: 0

СММ. Построение множественного выравнивания

ACA---ATC	6,29
TCAACTATC	2,99
ACAC--AGC	5,26
AGA---ATC	4,90
ACCG--ATG	3,18
CGCGT-CGG	-3,28

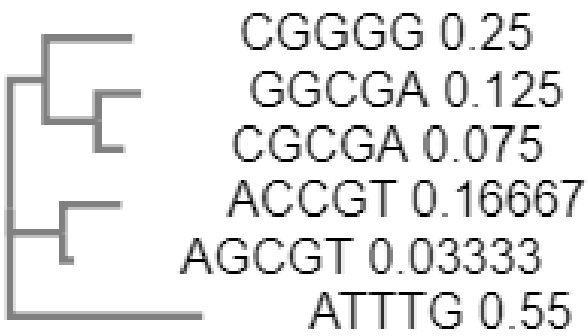


CMM. Clustal Omega

```
CLUSTAL O(1.2.1) multiple sequence alignment
```

http://www.ebi.ac.uk/Tools/msa/clustalo/

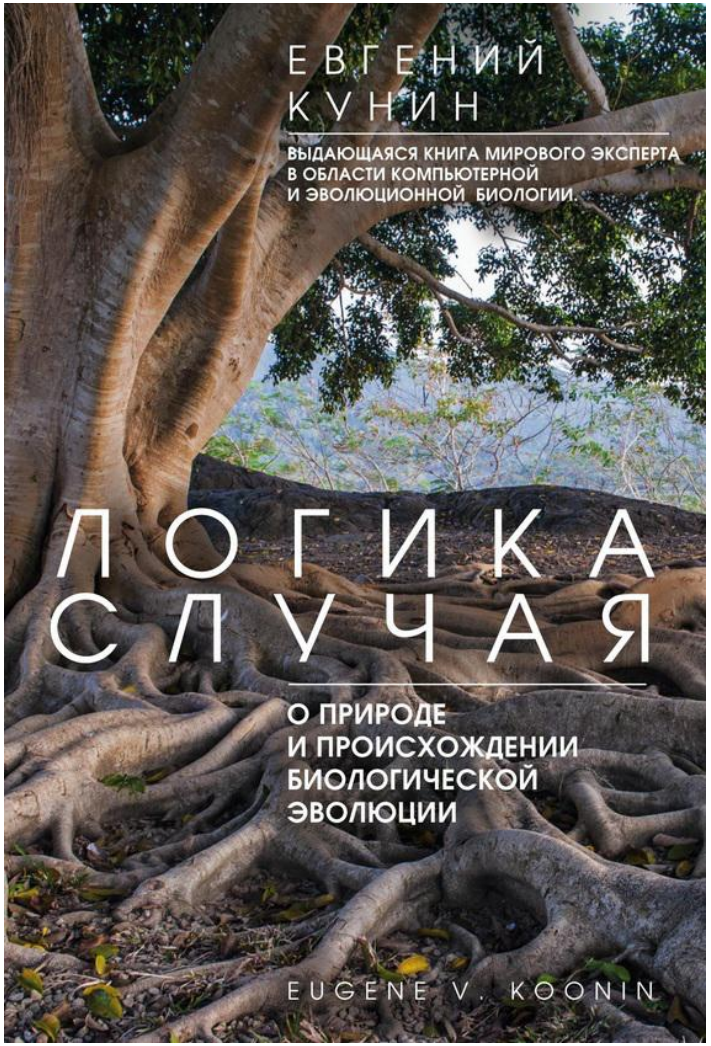

```
CGGGG           CGGGG
GGCGA           GGCGA
CGCGA           CGCGA
ACCGT           ACCGT
AGCGT           AGCGT
ATTTG           ATTTG
```

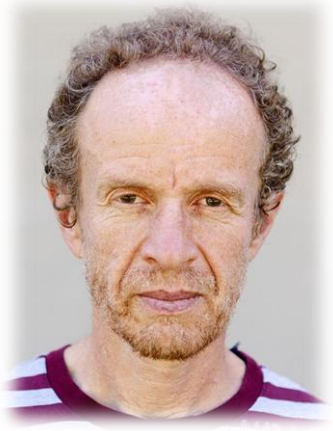


This is a Neighbour-joining tree without distance corrections.

Что почитать?

Что почитать?





Евгений Кунин
EN = 2

«**В XXI веке** вопрос о необходимости перевода научной литературы с английского на какие-либо другие языки, мягко говоря, неоднозначен. Научные тексты теперь публикуются по- английски, и **умение** их **читать на этом языке – элементарное требование профессиональной пригодности**.»

Благодарю за внимание!