# ВВЕДЕНИЕ В БИОИНФОРМАТИКУ

Лекция №7

Методы кластеризации. Множественные выравнивания. Модели эволюции. Филогенетические деревья

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

Сайт курса http://intbio.org/bioinf2018

# Множественное выравнивание последовательностей

Что полезного?

- Выявление удаленной гомологии
- Выявление консервативных остатков и мотивов
- Построение филогенетических деревьев
- …

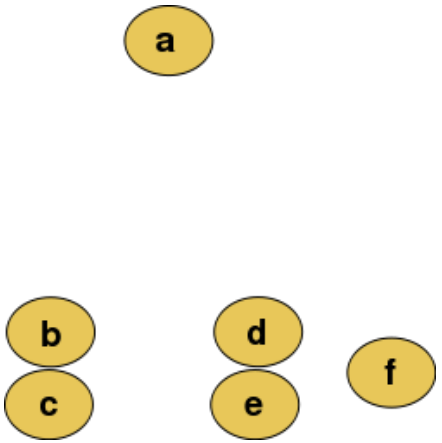
Алгоритмы:

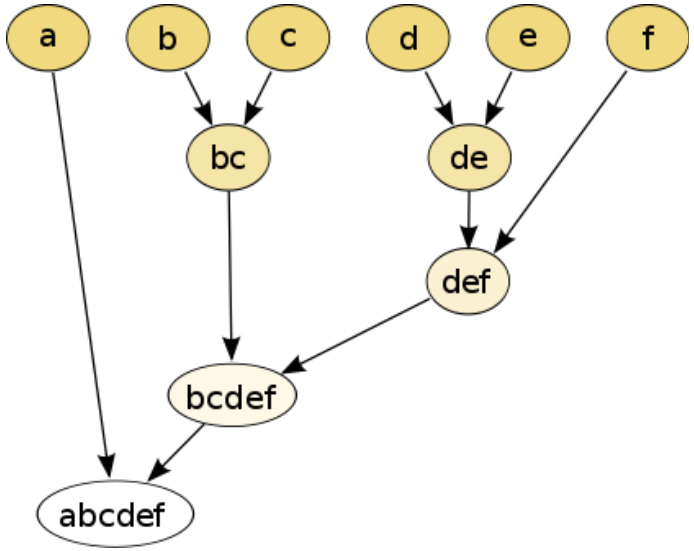
- Динамическое программирование – не годится
- Прогрессивное выравнивание
- Итеративное выравнивание
- Скрытые марковские модели
- Квантовые компьютеры?! (2017)

Визуализация: Построение профилей

# Методы иерархической кластеризации. UPGMA

**UPGMA – Unweighted Pair Group Method with Arithmetic mean** (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 C_i и C_j с образованием кластера C_k, содержащего *n_i + n_j = n_k* элементов, расстояние от кластера C_k до остальных кластеров C_m 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, CGGGC.

Используя расстояние по Хэммингу, получаем матрицу расстояний:

| D0 | ATTTG | AGCGT | ACCGT | CGCGA | GGCGA | CGGGC |
|---|---|---|---|---|---|---|
| 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 |
| CGGGC | | | | | | 0 |

# Методы иерархической кластеризации. UPGMA

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

Используя расстояние по Хэммингу, получаем матрицу расстояний:

| D0 | ATTTG | AGCGT | ACCGT | CGCGA | GGCGA | CGGGC |
|---|---|---|---|---|---|---|
| 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 |
| CGGGC | | | | | | 0 |

| D1 | ATTTG | AGCGT, ACCGT | CGCGA | GGCGA | CGGGC |
|---|---|---|---|---|---|
| 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 |
| CGGGC | | | | | 0 |

# Методы иерархической кластеризации. UPGMA

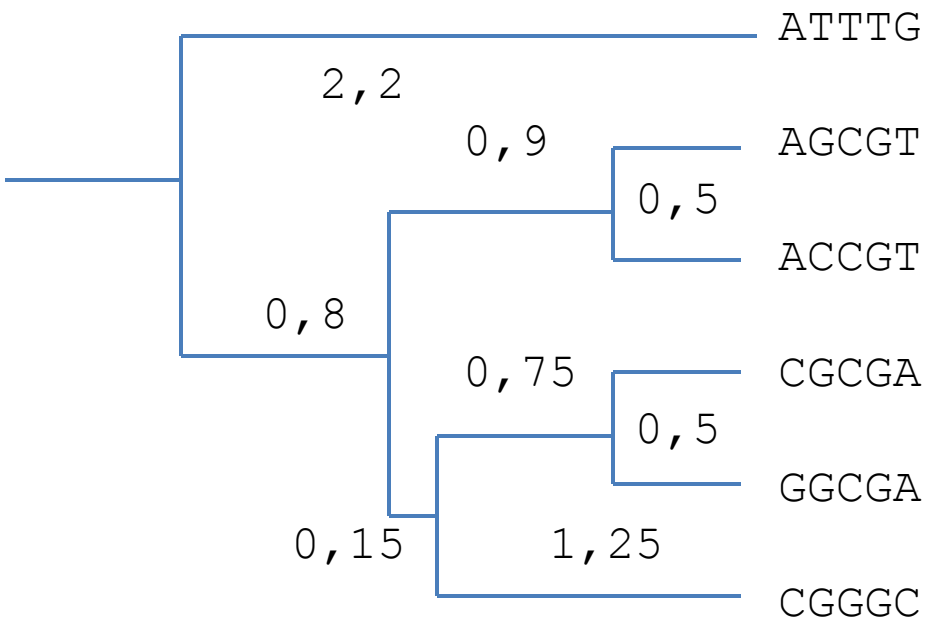
| D2 | ATTTG | AGCGT, ACCGT | CGCGA, GGCGA | CGGGC |
|---|---|---|---|---|
| ATTTG | 0 | 4 | (5+5)/2=5 | 4 |
| AGCGT, ACCGT | | 0 | (2,5+2,5)/2=2,5 | 3,5 |
| CGCGA, GGCGA | | | 0 | (2+3)/2=**2,5** |
| CGGGC | | | | 0 |

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

| D4 | ATTTG | ((CGCGA, GGCGA), GGGC), (AGCGT, ACCGT) |
|---|---|---|
| ATTTG | 0 | (4*2+4,7*3)/5=4,4 |
| ((CGCGA, GGCGA), CGGGC), (AGCGT, ACCGT) | | 0 |

# Методы иерархической кластеризации. UPGMA

Объединяя теперь кластеры, получим дерево :




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

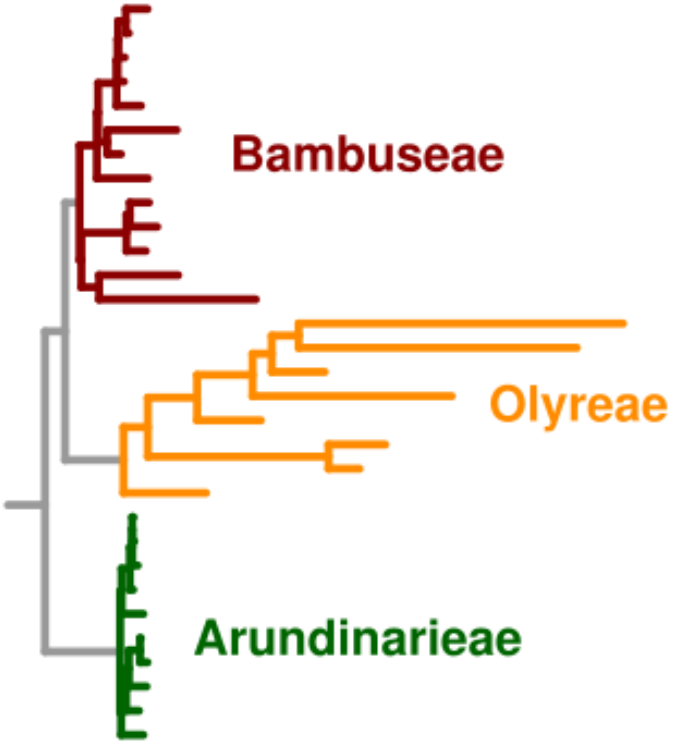
Метод UPGMA подразумевает справедливость гипотезы молекулярных часов (постоянной скорости эволюции).

# Гипотеза молекулярных часов

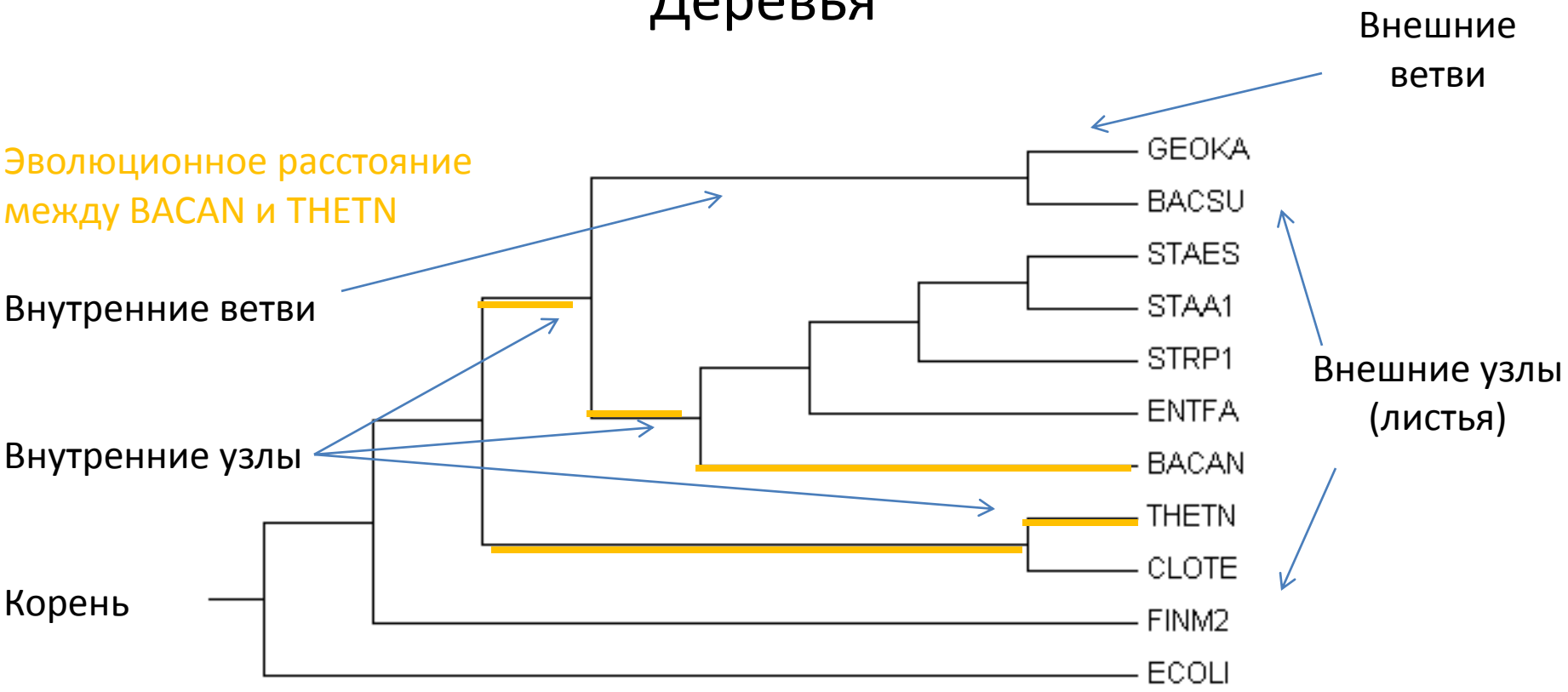
Сопоставления числа различий в аминокислотных последовательностях гемоглобинов млекопитающих и срока дивергенции (Э. Цукеркандль, Л. Полинг, 1962).

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

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



# Деревья

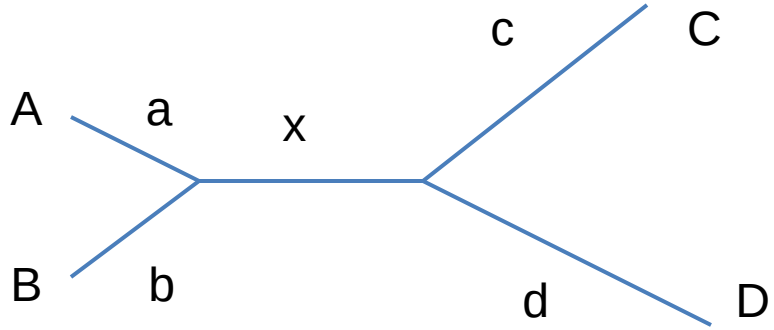



Кладограмма — филогенетическое дерево, не содержащее информации о длинах ветвей.

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

# Методы иерархической кластеризации. Метод связи между соседями

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




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

*Очевидно, что*

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

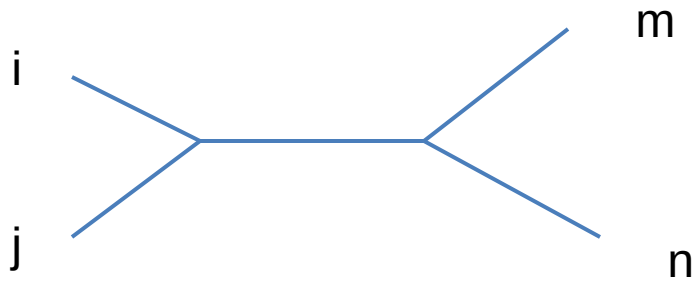
$$L_{AB} + L_{CD} < L_{AD} + L_{BC}$$

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

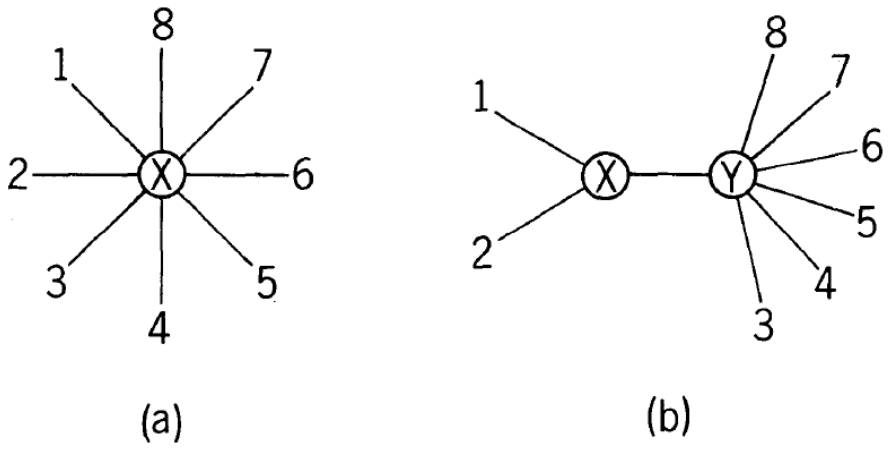
На большее число последовательностей обобщается путем рассмотрения всех четверок и определением тех из них, для которых суммы расстояний минимальны




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

| 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



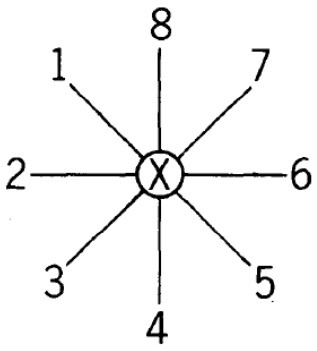

**Neighbor joining** – метод присоединения соседей (Saitou, Nei, 1987)

Еще один алгоритм иерархической кластеризации.

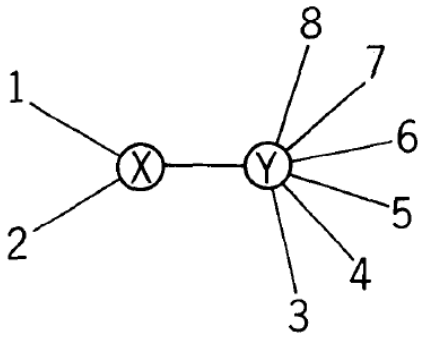
$$\textit{Пусть}$$
$$D_{ij} - \textit{расстояние между таксонами } i \textit{ и } j,$$
$$L_{ab} - \textit{длина ветви между узлами } a \textit{ и } b$$

$$\textit{Тогда для суммы длин ветвей дерева на рис.}(a) \textit{ выполняется}$$
$$S_0 = \sum_{i=1}^{N} L_{iX} = \frac{1}{N-1} \sum_{i<j} D_{ij} \; (\textit{поскольку } D_{ij} = L_{iX} + L_{Xj}) \qquad (1)$$

# 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))

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

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

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

Цель: найти такие таксоны 1 и 2, чтобы **сумма всех ветвей дерева была минимальна**

# Методы иерархической кластеризации. NJ

*Итак, имеем*

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

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

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

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

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

$$Теперь\ S_{12}\ принимает\ вид$$

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

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

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

# Методы иерархической кластеризации. NJ

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

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

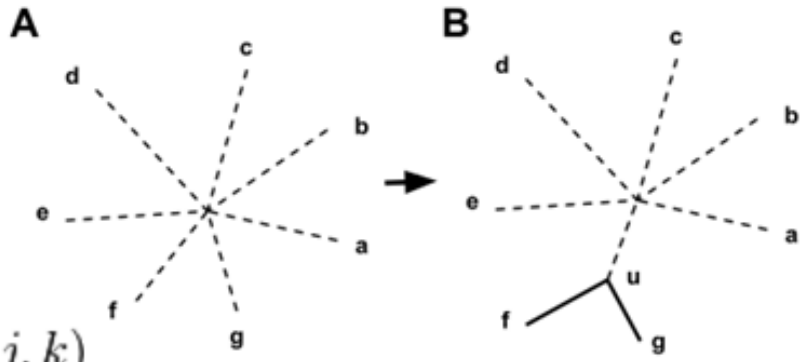
$$= \frac{2\sum_{i<j} D_{ij} - \sum_{i}^{N} D_{ki} - \sum_{i}^{N} D_{li}}{2(N-2)} + \frac{D_{kl}}{2}$$

*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 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

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

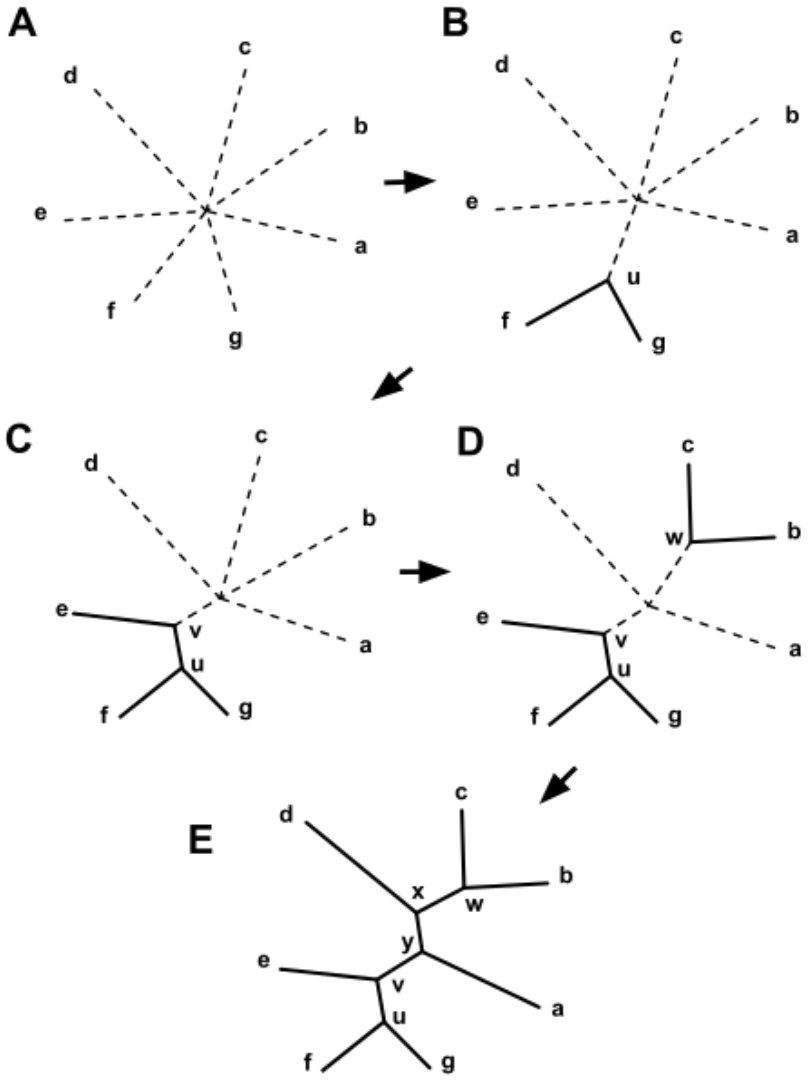
# Методы иерархической кластеризации. NJ

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 присоединенных соседей 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 | CGGGC |
|---|---|---|---|---|---|---|
| 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 |
| CGGGC | | | | | | 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 | CGGGC |
|---|---|---|---|---|---|---|
| ATTTG | | -18 | -21 | -15 | -16 | -22 |
| AGCGT | | | -23 | -17 | -18 | -16 |
| ACCGT | | | | -16 | -17 | -15 |
| CGCGA | | | | | -23 | -21 |
| GGCGA | | | | | | -18 |
| CGGGC | | | | | | |

# Методы иерархической кластеризации. NJ

| **D0** | ATTTG | AGCGT | ACCGT | CGCGA | GGCGA | CGGGC |
|---|---|---|---|---|---|---|
| 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 |
| CGGGC | | | | | | 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 | 0,875 |
| ACCGT | 0,125 |

# Методы иерархической кластеризации. NJ

| **D0** | ATTTG | AGCGT | ACCGT | CGCGA | GGCGA | CGGGC |
|---|---|---|---|---|---|---|
| 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 |
| CGGGC | | | | | | 0 |

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

| **D1** | ATTTG | AGCGT, ACCGT | CGCGA | GGCGA | CGGGC |
|---|---|---|---|---|---|
| ATTTG | 0 | 3,5 | 5 | 5 | 4 |
| AGCGT, ACCGT | | 0 | 2 | 2 | 3 |
| CGCGA | | | 0 | 1 | 2 |
| GGCGA | | | | 0 | 3 |
| CGGGC | | | | | 0 |

# Методы иерархической кластеризации. NJ

| **D1** | ATTTG | AGCGT, ACCGT | CGCGA | GGCGA | CGGGC |
|---|---|---|---|---|---|
| ATTTG | 0 | 3,5 | 5 | 5 | 4 |
| AGCGT, ACCGT | | 0 | 2 | 2 | 3 |
| CGCGA | | | 0 | 1 | 2 |
| GGCGA | | | | 0 | 3 |
| CGGGC | | | | | 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 | CGGGC |
|---|---|---|---|---|---|
| ATTTG | | -17,5 | -12,5 | -13,5 | -17,5 |
| AGCGT, ACCGT | | | -14,5 | -15,5 | -13,5 |
| CGCGA | | | | **-18** | -16 |
| GGCGA | | | | | -14 |
| CGGGC | | | | | |

# Методы иерархической кластеризации. NJ

| D1 | ATTTG | AGCGT, ACCGT | CGCGA | GGCGA | CGGGC |
|---|---|---|---|---|---|
| ATTTG | 0 | 3,5 | 5 | 5 | 4 |
| AGCGT, ACCGT | | 0 | 2 | 2 | 3 |
| CGCGA | | | 0 | 1 | 2 |
| GGCGA | | | | 0 | 3 |
| CGGGC | | | | | 0 |

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

| δ2 | CGCGA, GGCGA |
|---|---|
| CGCGA | 0,333 |
| GGCGA | 0,667 |

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

| D2 | ATTTG | AGCGT, ACCGT | CGCGA, GGCGA | CGGGC |
|---|---|---|---|---|
| ATTTG | 0 | 3,5 | 4,5 | 4 |
| AGCGT, ACCGT | | 0 | 1,5 | 3 |
| CGCGA, GGCGA | | | 0 | 2 |
| CGGGC | | | | 0 |

# Методы иерархической кластеризации. NJ

| **D2** | ATTTG | AGCGT, ACCGT | CGCGA, GGCGA | CGGGC |
|---|---|---|---|---|
| ATTTG | 0 | 3,5 | 4,5 | 4 |
| AGCGT, ACCGT | | 0 | 1,5 | 3 |
| CGCGA, GGCGA | | | 0 | 2 |
| CGGGC | | | | 0 |

| **Q2** | ATTTG | AGCGT, ACCGT | CGCGA, GGCGA | CGGGC |
|---|---|---|---|---|
| ATTTG | | -13 | -11 | -13 |
| AGCGT, ACCGT | | | -13 | -11 |
| CGCGA, GGCGA | | | | **-13** |
| CGGGC | | | | |

| **δ3** | (CGCGA, GGCGA), CGGGC |
|---|---|
| CGCGA, GGCGA | 0,75 |
| CGGGC | 1,25 |

# Методы иерархической кластеризации. NJ

| **D2** | ATTTG | AGCGT, ACCGT | CGCGA, GGCGA | CGGGC |
|---|---|---|---|---|
| ATTTG | 0 | 3,5 | 4,5 | 4 |
| AGCGT, ACCGT | | 0 | 1,5 | 3 |
| CGCGA, GGCGA | | | 0 | 2 |
| CGGGC | | | | 0 |

| **Q2** | ATTTG | AGCGT, ACCGT | CGCGA, GGCGA | CGGGC |
|---|---|---|---|---|
| ATTTG | | -13 | -11 | -13 |
| AGCGT, ACCGT | | | -13 | -11 |
| CGCGA, GGCGA | | | | **-13** |
| CGGGC | | | | |

| **D3** | ATTTG | AGCGT, ACCGT | (CGCGA, GGCGA), CGGGC |
|---|---|---|---|
| ATTTG | 0 | 3,5 | 3,25 |
| AGCGT, ACCGT | | 0 | 1,25 |
| (CGCGA, GGCGA), CGGGC | | | 0 |

# Методы иерархической кластеризации. NJ

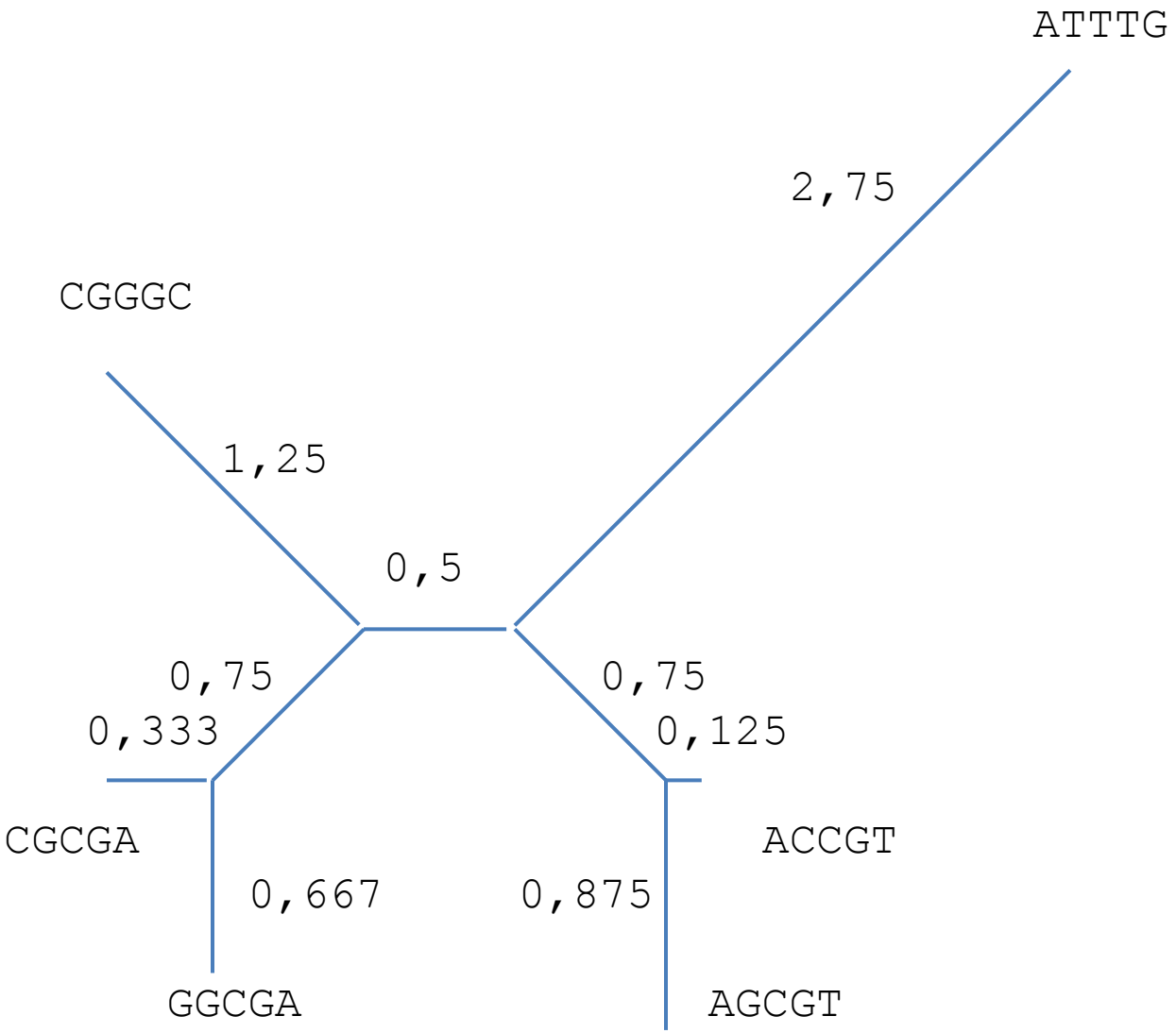
| **D3** | ATTTG | AGCGT, ACCGT | (CGCGA, GGCGA), CGGGC |
|---|---|---|---|
| ATTTG | 0 | 3,5 | 3,25 |
| AGCGT, ACCGT | | 0 | 1,25 |
| (CGCGA, GGCGA), CGGGC | | | 0 |

| **Q3** | ATTTG | AGCGT, ACCGT | (CGCGA, GGCGA), CGGGC |
|---|---|---|---|
| ATTTG | | -8 | -8 |
| AGCGT, ACCGT | | | -8 |
| (CGCGA, GGCGA), CGGGC | | | |

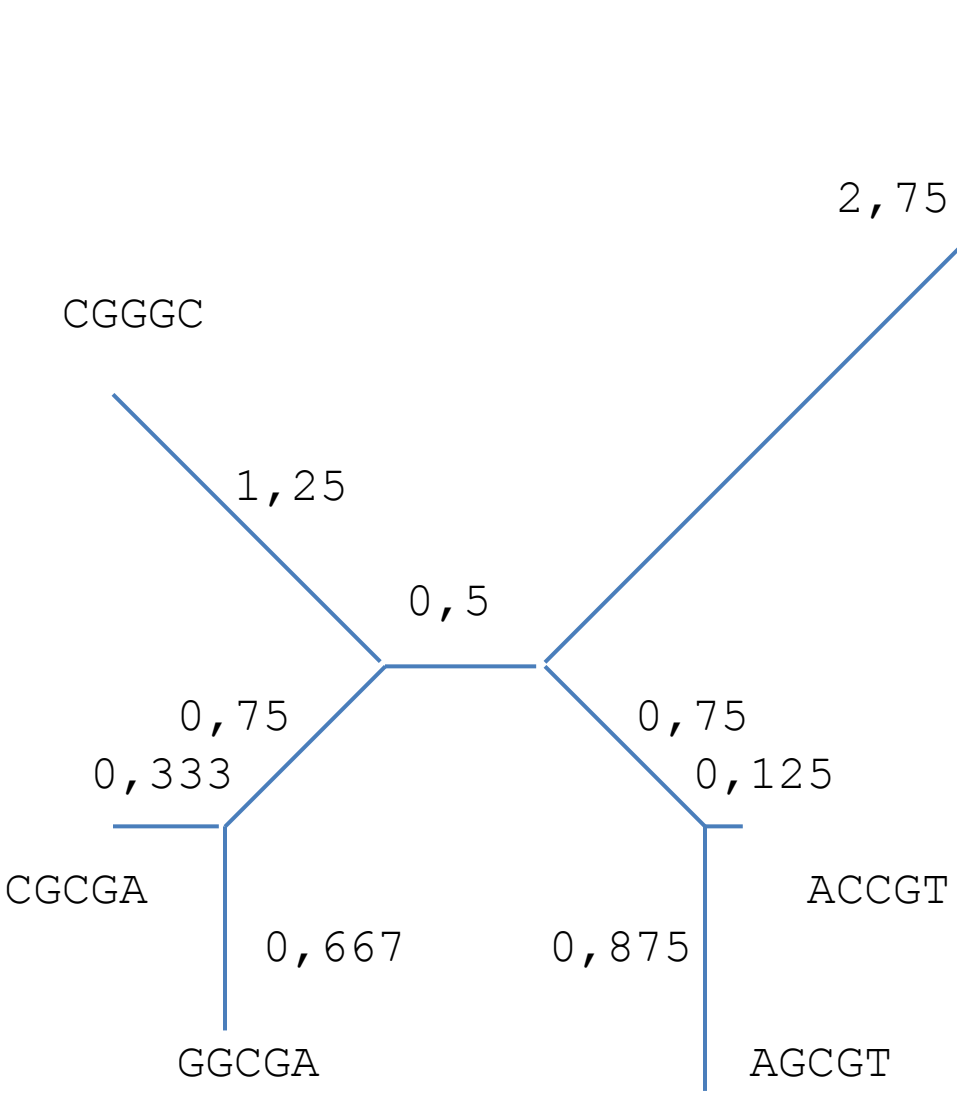
| **δ4** | ATTTG, AGCGT, ACCGT |
|---|---|
| ATTTG | 0,75 |
| AGCGT, ACCGT | 1,25 |

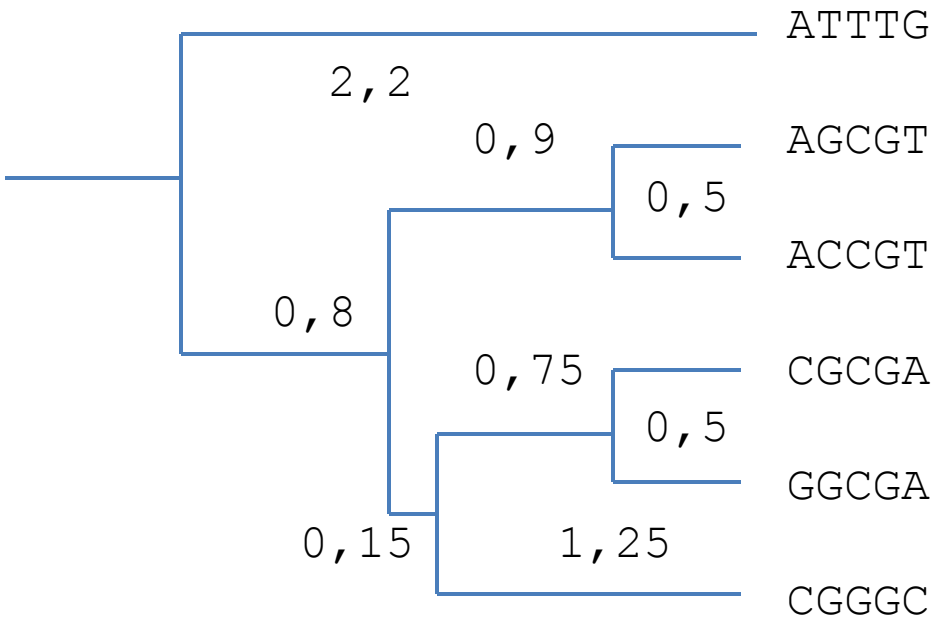
| **D4** | AGCGT, ACCGT, ATTTG | CGCGA, GGCGA, CGGGC |
|---|---|---|
| AGCGT, ACCGT, ATTTG | 0 | 0,5 |
| CGCGA, GGCGA, CGGGC | 0,5 | 0 |

# Методы иерархической кластеризации. NJ



# Методы иерархической кластеризации. NJ и UPGMA



# Прогрессивное выравнивание. Clustal

**Clustal** (1988) выполняет постепенное выравнивание все новых последовательностей, начиная с наиболее <эволюционно> близких, ориентируясь на предварительно построенное на основании парных выравниваний 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

Текущая версия - Clustal Omega - использует СММ

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

# Итеративное выравнивание. MUSCLE

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)

# Модели эволюции. Модель Джукса-Кантора (1969)

Вероятность $\alpha$ замены любого нуклеотида на любой другой в единицу времени одинакова.

$P\_A(0) = 1$ – в некой позиции в момент времени 0 содержится нуклеотид A

$P\_A(1) = 1 - 3\alpha$ – вероятность сохранения нуклеотида A в этой позиции в момент 1

$P\_A(2) = (1 - 3\alpha)\ P\_A(1) + \alpha\ (1 - P\_A(1))$ – вероятность сохранения нуклеотида A в этой позиции в момент 2.

$P\_A(t+1) = (1 - 3\alpha)\ P\_A(t) + \alpha\ (1 - P\_A(t))$

$\Delta\ P\_A(t) = P\_A(t+1) - P\_A(t) = - 3\alpha P\_A(t) + \alpha\ (1 - P\_A(t)) = - 4\ \alpha\ P\_A(t) + \alpha$

# Модели эволюции. Модель Джукса-Кантора (1969)

$\Delta P_A(t) = P_A(t+1) - P_A(t) = -3\alpha P_A(t) + \alpha (1 - P_A(t)) = -4\alpha P_A(t) + \alpha$

Переходя к непрерывному времени, получаем

$$d P_A(t)/dt = -4\alpha P_A(t) + \alpha$$

- дифференциальное уравнение, решение которого можно искать в виде

$$P_A(t) = U \exp(-4\alpha t) + V.$$

Подставляя, находим $V = ¼$, что с учетом известного $P_A(0)$ даёт

$$P_A(t) = (P_A(0) – ¼) \exp(-4\alpha t) + ¼.$$

Если в момент 0 нуклеотид A присутствовал в позиции, то $P_A(0) = 1$ и

$$P_A(t) = ¾ \exp(-4\alpha t) + ¼.$$

Иначе $P_A(0) = 0$ и

$$P_A(t) = -¼ \exp(-4\alpha t) + ¼.$$

# Модели эволюции. Модель Джукса-Кантора (1969)

Если в момент 0 нуклеотид A присутствовал в позиции, то $P_A(0) = 1$ и

$$P_A(t) = ¾ \exp(-4 \alpha t) + ¼.$$

Иначе $P_A(0) = 0$ и

$$P_A(t) = -¼ \exp(-4 \alpha t) + ¼.$$

Таким образом, для любого нуклеотида при любом начальном состоянии со временем вероятность обнаружения этого нуклеотида в данной позиции стремиться к ¼, а значит доля нуклеотида в последовательности к 25% - в общем случае это неверно

# Модели эволюции. Модель Джукса-Кантора (1969)

Рассмотрим теперь две гомологичные последовательности x и y, имеющие неизвестного общего предка.

Если в некой позиции этих последовательностей наблюдается нуклеотид A, то он мог остаться от общего предка с вероятностью P_AA(t)*P_AA(t), либо в обеих последовательностях произошли одинаковые замены, вероятность чего P_CA(t)*P_CA(t) + P_GA(t)*P_GA(t) + P_TA(t)*P_TA(t).

Итого нуклеотид A будет наблюдаться в обеих последовательностях с вероятностью

P_AA(t)*P_AA(t) + P_CA(t)*P_CA(t) + P_GA(t)*P_GA(t) + P_TA(t)*P_TA(t).

Это же верно и для всех позиций в целом, поэтому доля I(t) идентичных нуклеотидов в последовательностях x и y будет

I(t) = P_AA(t)*P_AA(t) + P_CA(t)*P_CA(t) + P_GA(t)*P_GA(t) + P_TA(t)*P_TA(t) = ¾ exp (- 8 α t) + ¼

# Модели эволюции. Модель Джукса-Кантора (1969)

Рассмотрим теперь две гомологичные последовательности x и y, имеющие неизвестного общего предка.

Доля I(t) идентичных нуклеотидов в последовательностях x и y будет

$$I(t) = ¾ \exp(-8 \alpha t) + ¼$$

При этом доступная для наблюдения доля p различий между последовательностями x и y будет

$$p = 1 - I(t) = ¾ (1 - \exp(-8 \alpha t))$$

что дает для α

$$8 \alpha t = - \ln(1 - 4/3p).$$

В то же время расчетное число замен за это время, согласно модели, будет D = 2(3 α t). Таким образом,

$$D = - ¾ \ln(1 - 4/3p).$$

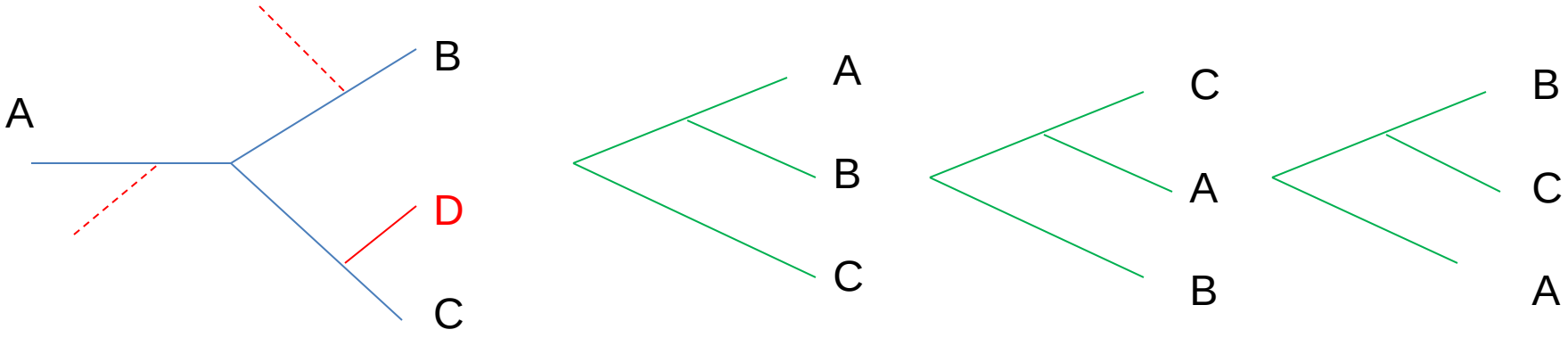
# Деревья: свойства

Число деревьев

неукорененных $N_U = (2n-5)!!$

укорененных $N_R = (2n-3)!!$

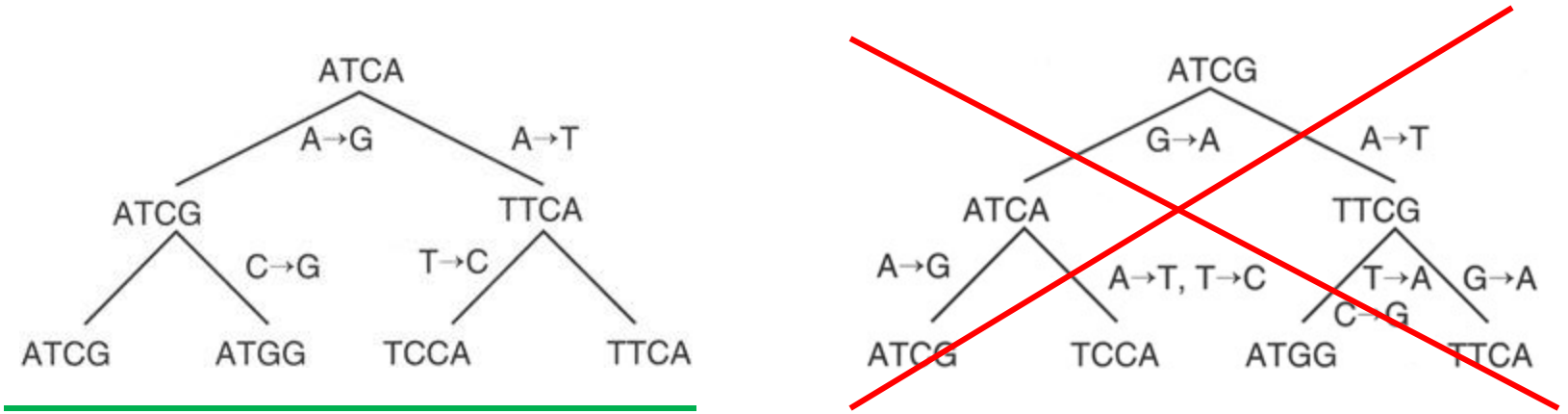
С точки зрения филогении, правильное только одно!




| **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 |

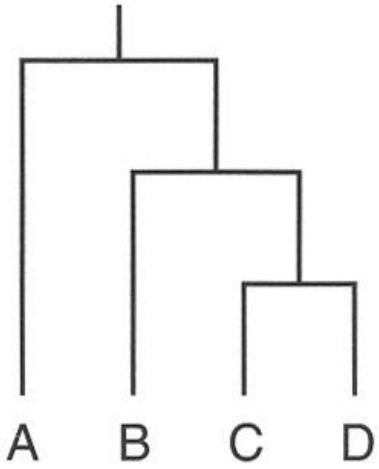
# Филогенетические деревья – методы

**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.



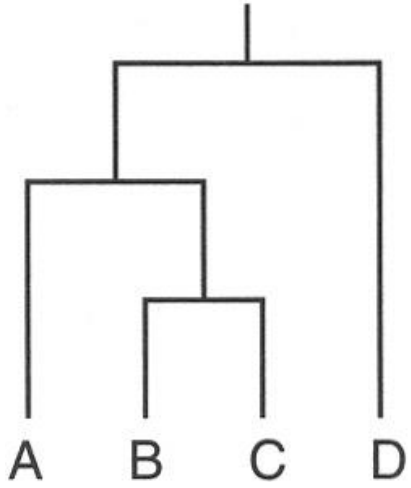

**Метод максимального правдоподобия** учитывает не просто число мутаций, но и их вероятность.

# Филогенетические деревья – проблема переменной скорости эволюции

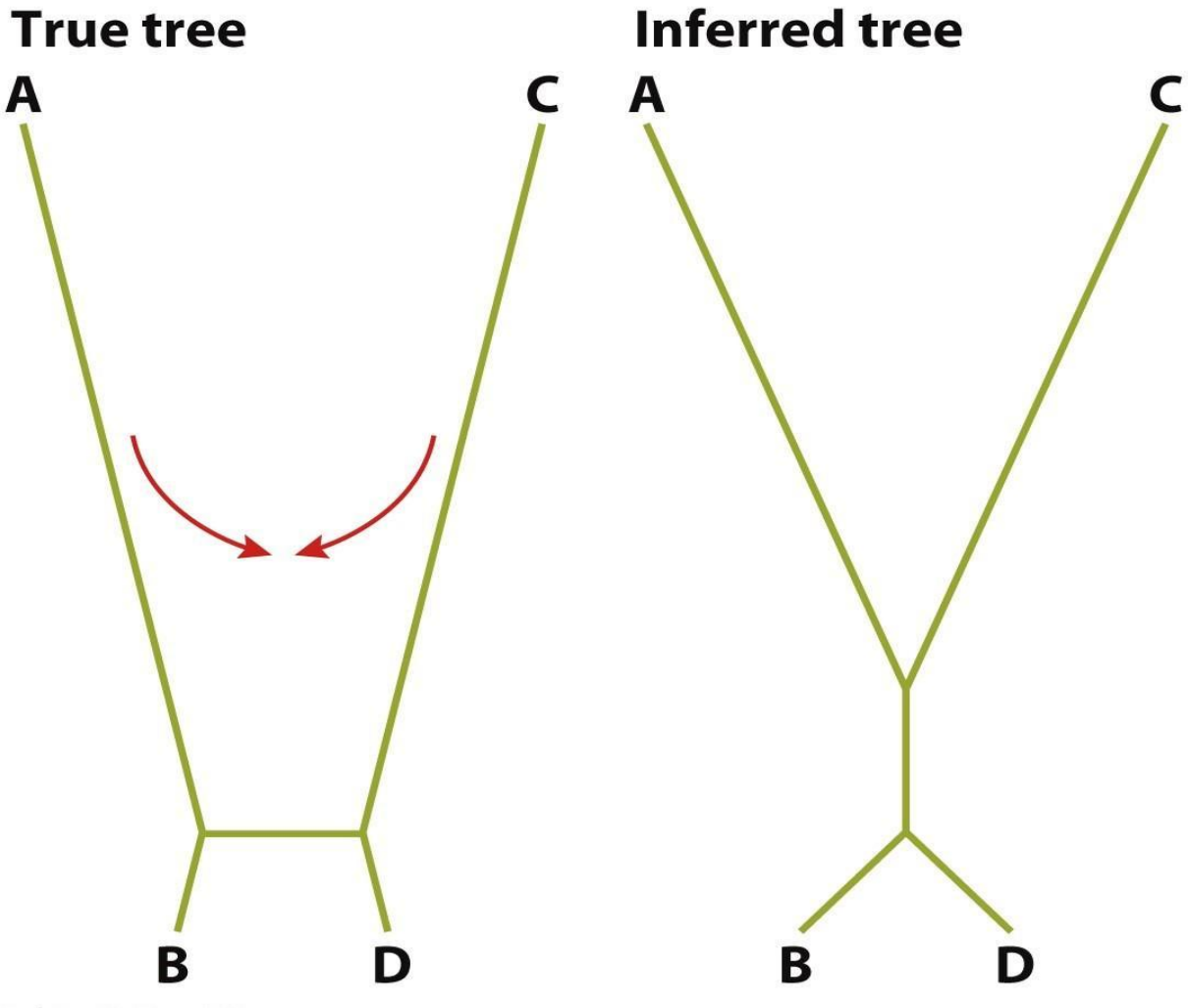



| | A | B | C | D |
|---|---|---|---|---|
| **A** | 0 | 3 | 3 | 3 |
| **B** | | 0 | 2 | 2 |
| **C** | | | 0 | 1 |
| **D** | | | | 0 |

| | A | B | C | D |
|---|---|---|---|---|
| **A** | 0 | 3 | 3 | 20 |
| **B** | | 0 | 2 | 20 |
| **C** | | | 0 | 20 |
| **D** | | | | 0 |

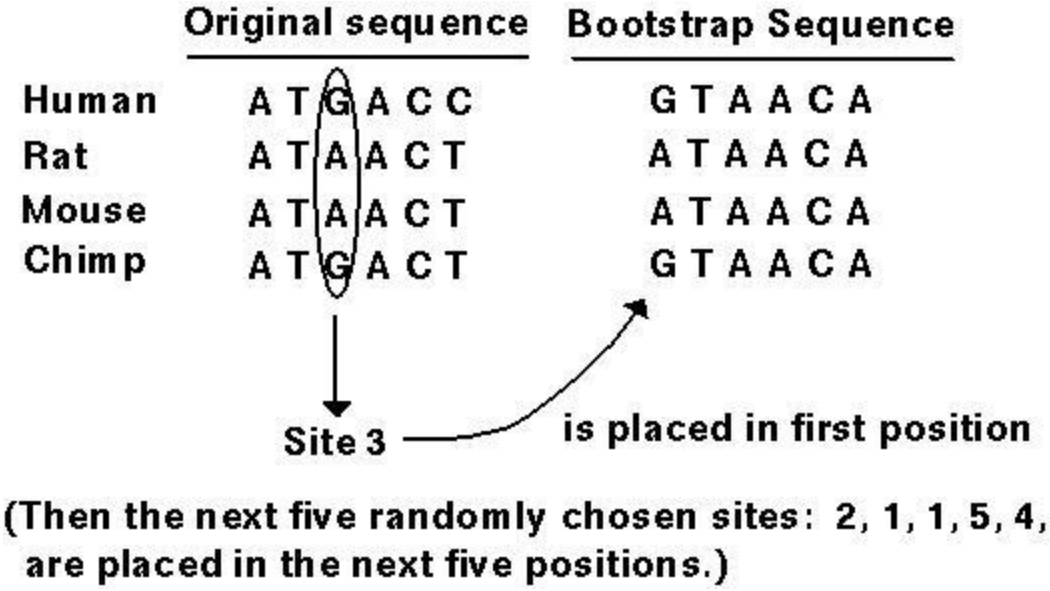


# Филогенетические деревья – «притяжение длинных ветвей»

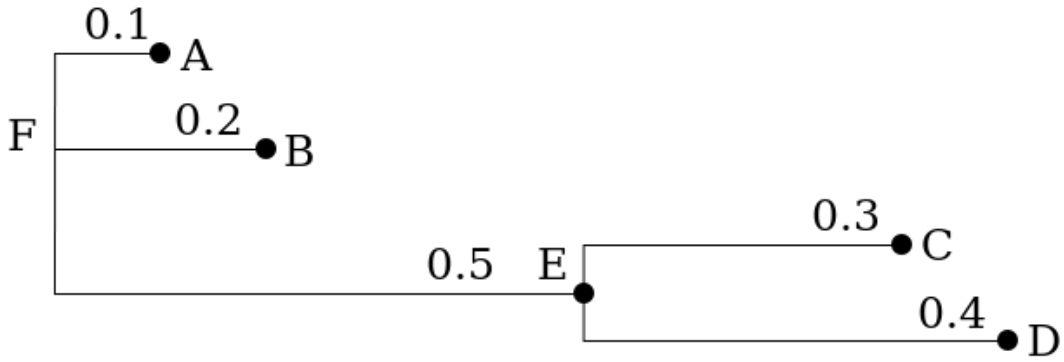


# Филогенетические деревья – методы проверки

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;
3) Оценка результата с помощью формальных статистических тестов. Например, построение дерева для подмножества последовательностей из исходного множественного выравнивания должно дать поддерево дерева, полученного для этого выравнивания;
4) Бутстреп (boot strap)



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

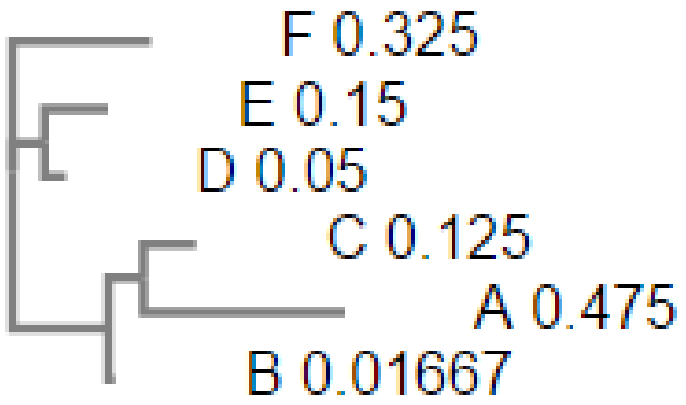



| | |
|---|---|
| `(,,(,));` | *имена узлов не указаны* |
| `(A,B,(C,D));` | *указаны только имена листьев* |
| `(A,B,(C,D)E)F;` | *указаны имена всех узлов* |
| `(:0.1,:0.2,(:0.3,:0.4):0.5);` | *для всех узлов кроме корня указано расстояние до родительского узла* |
| `(:0.1,:0.2,(:0.3,:0.4):0.5):0.0;`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

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

```
CLUSTAL O(1.2.1) multiple sequence alignment

F CGGGC
C ACCGT
B AGCGT
E GGCGA
A ATTTG
D CGCGA
```

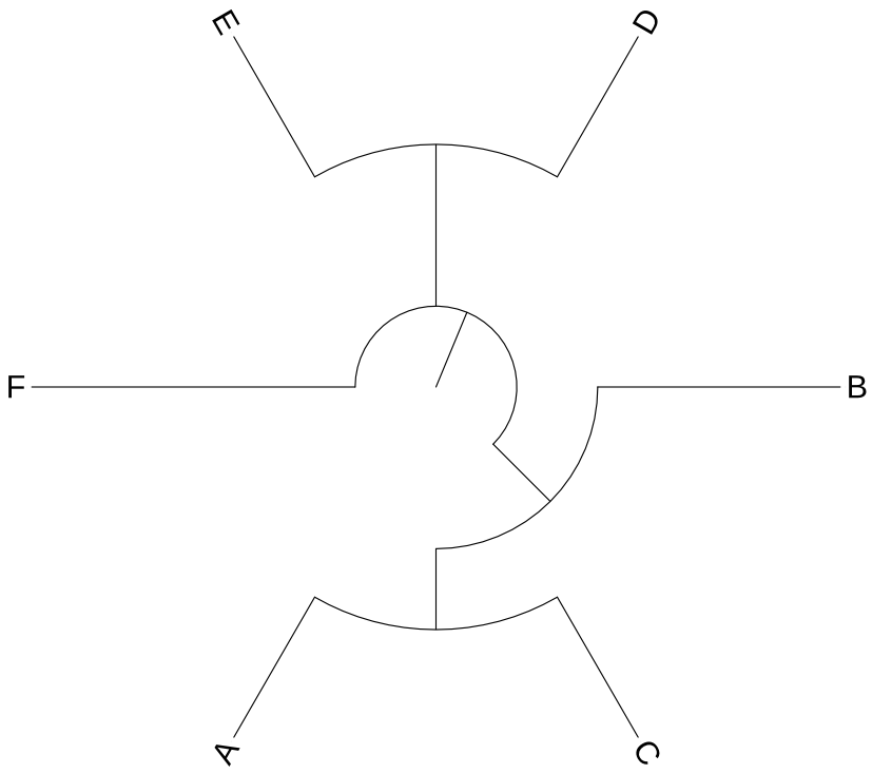
```
(
F:0.32500,
(
E:0.15000,
D:0.05000)
:0.07500,
(
(
C:0.12500,
A:0.47500)
:0.08333,
B:0.01667)
:0.22500);
```




*This is a Neighbour-joining tree without distance corrections.*

# Филогенетические деревья. Круговое представление

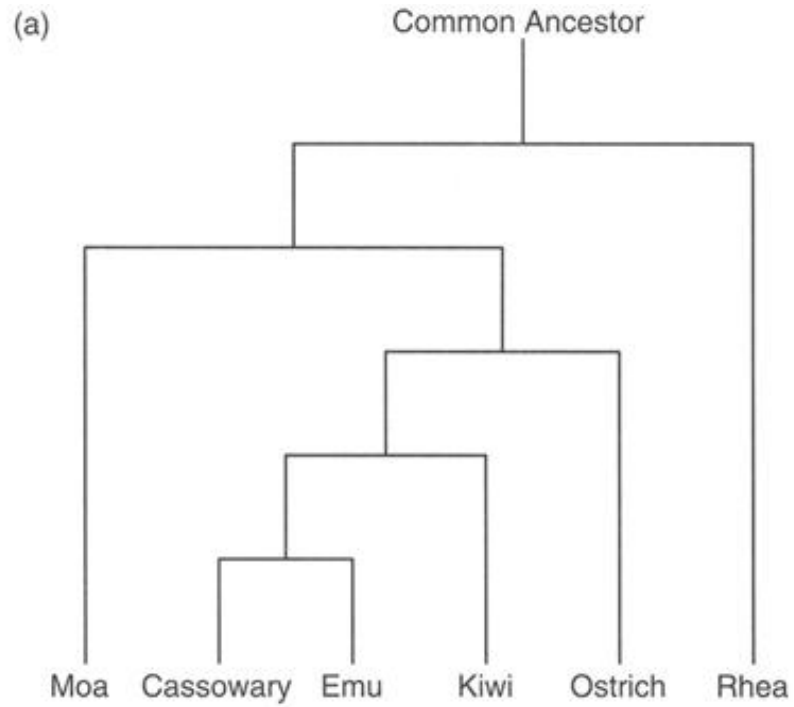
```
CLUSTAL O(1.2.1) multiple sequence alignment

F CGGGC
C ACCGT
B AGCGT
E GGCGA
A ATTTG
D CGCGA
```



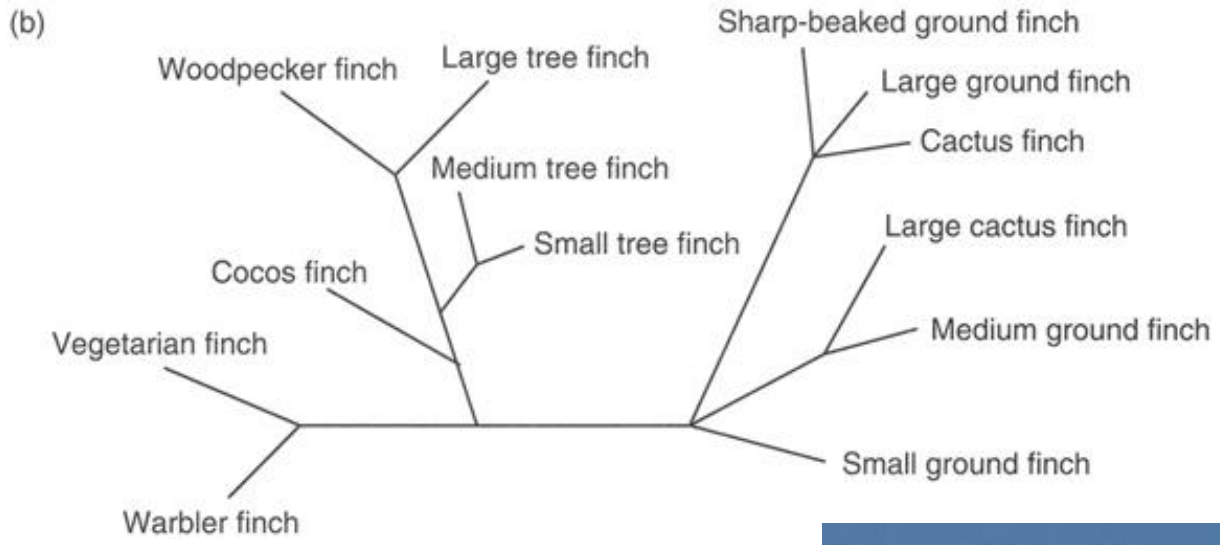

```
(
F:0.32500,
(
E:0.15000,
D:0.05000)
:0.07500,
(
(
C:0.12500,
A:0.47500)
:0.08333,
B:0.01667)
:0.22500);
```

http://itol.embl.de

# Филогенетические деревья – примеры




# Филогенетические деревья – примери




# Филогенетические деревья – такие разные