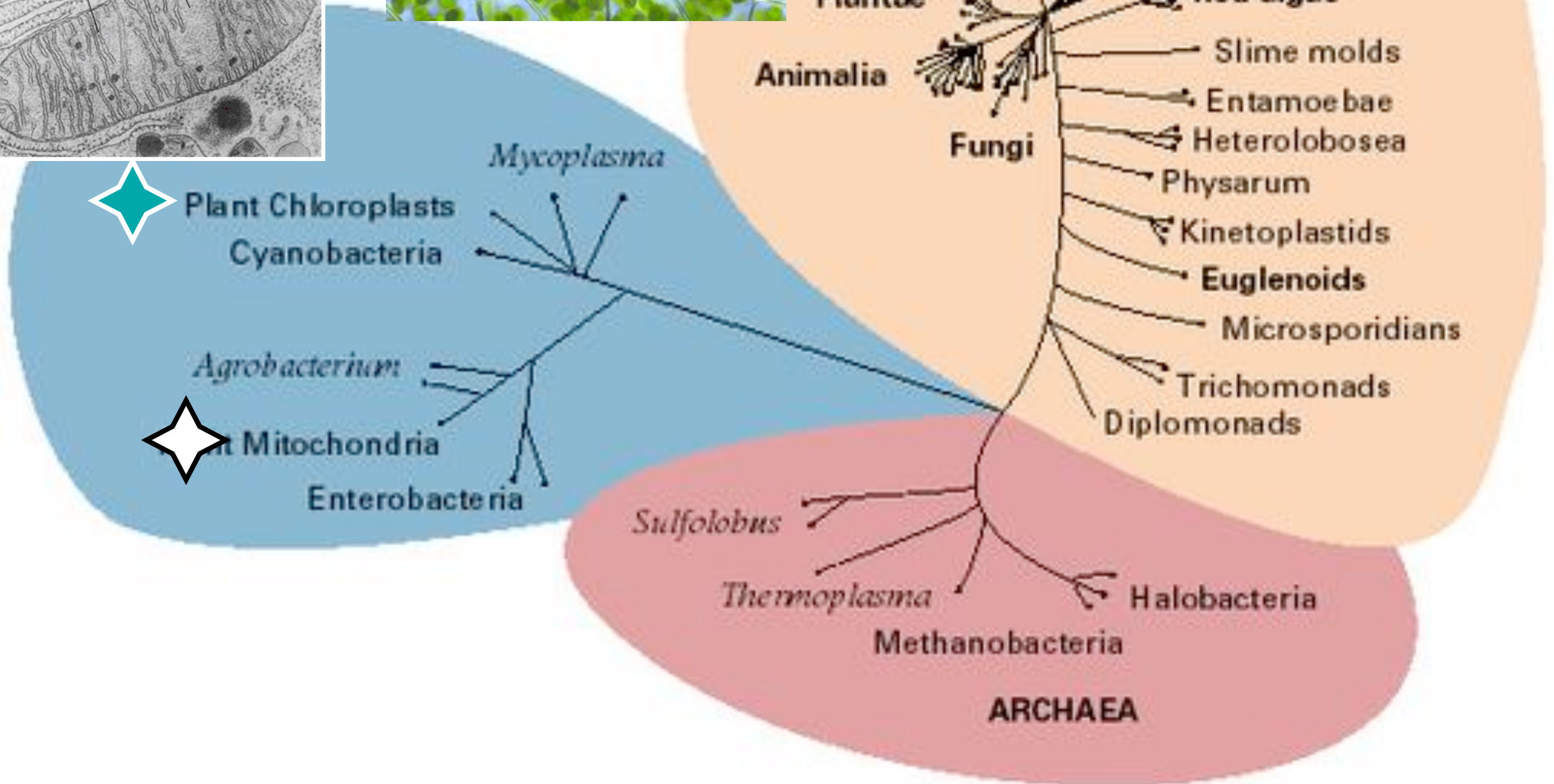
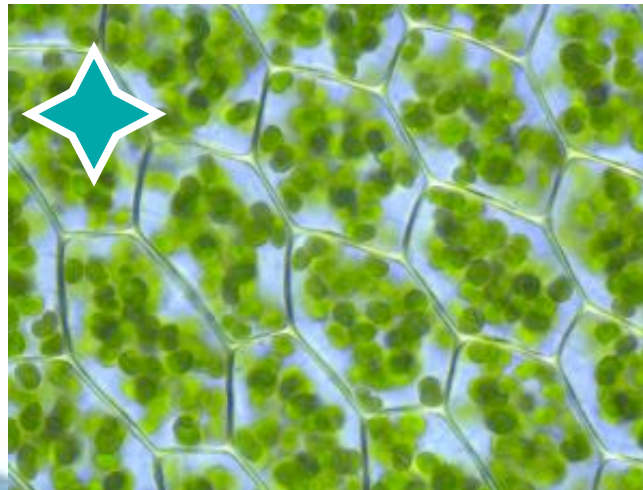
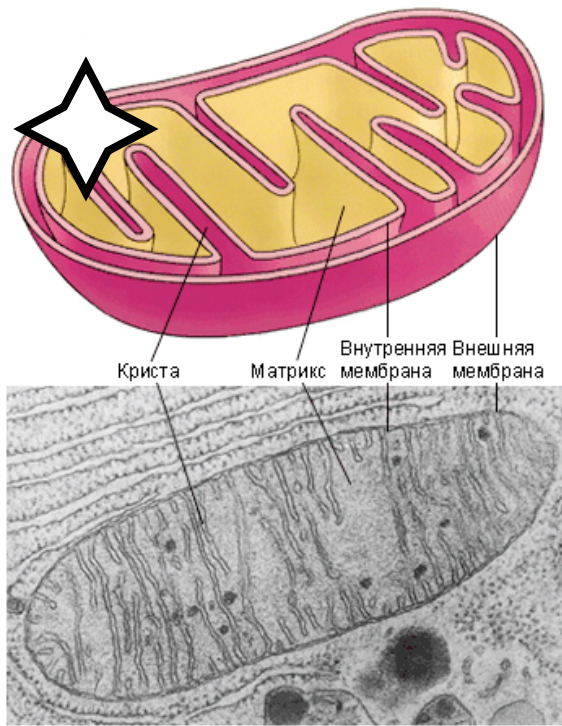


Филогенетические деревья

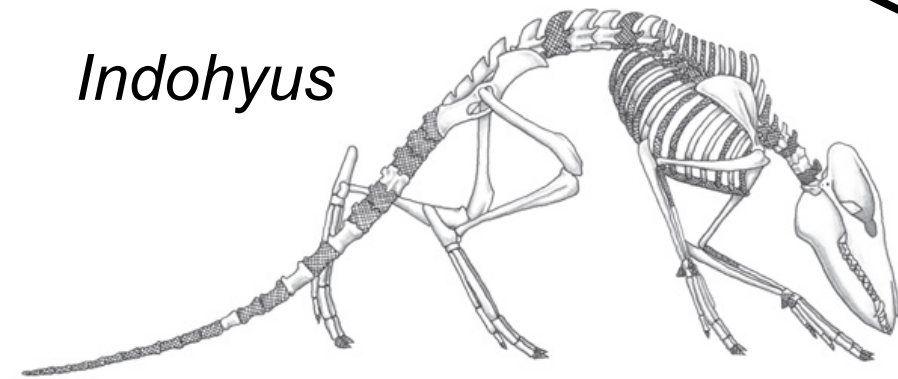
Три домена: бактерии, археи, эукариоты



Неожиданные родственные отношения между млекопитающими



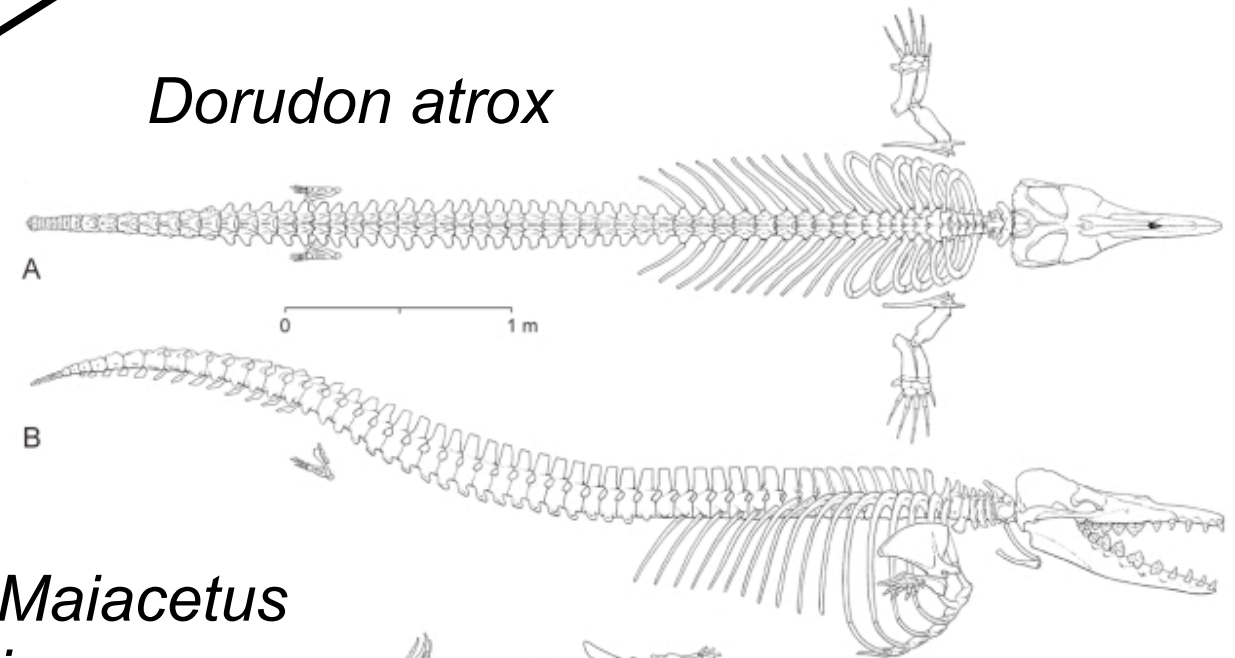
Indohyus



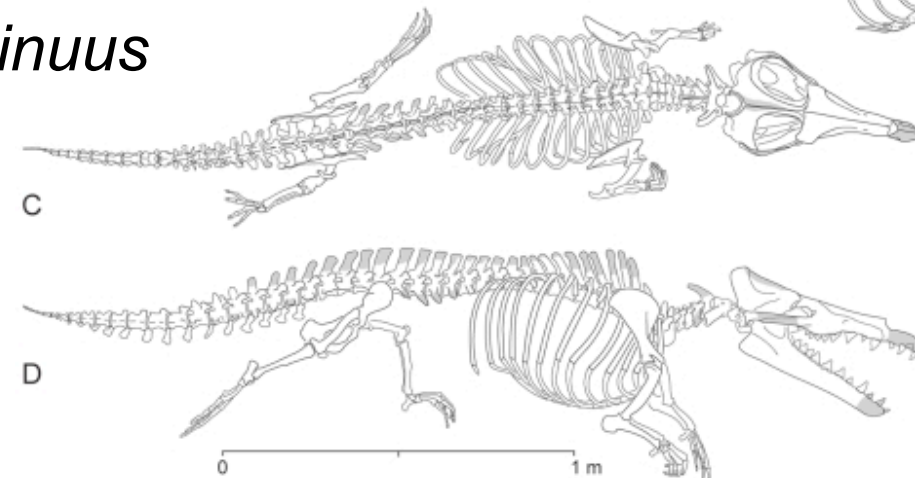
Rodhocetus



Dorudon atrox

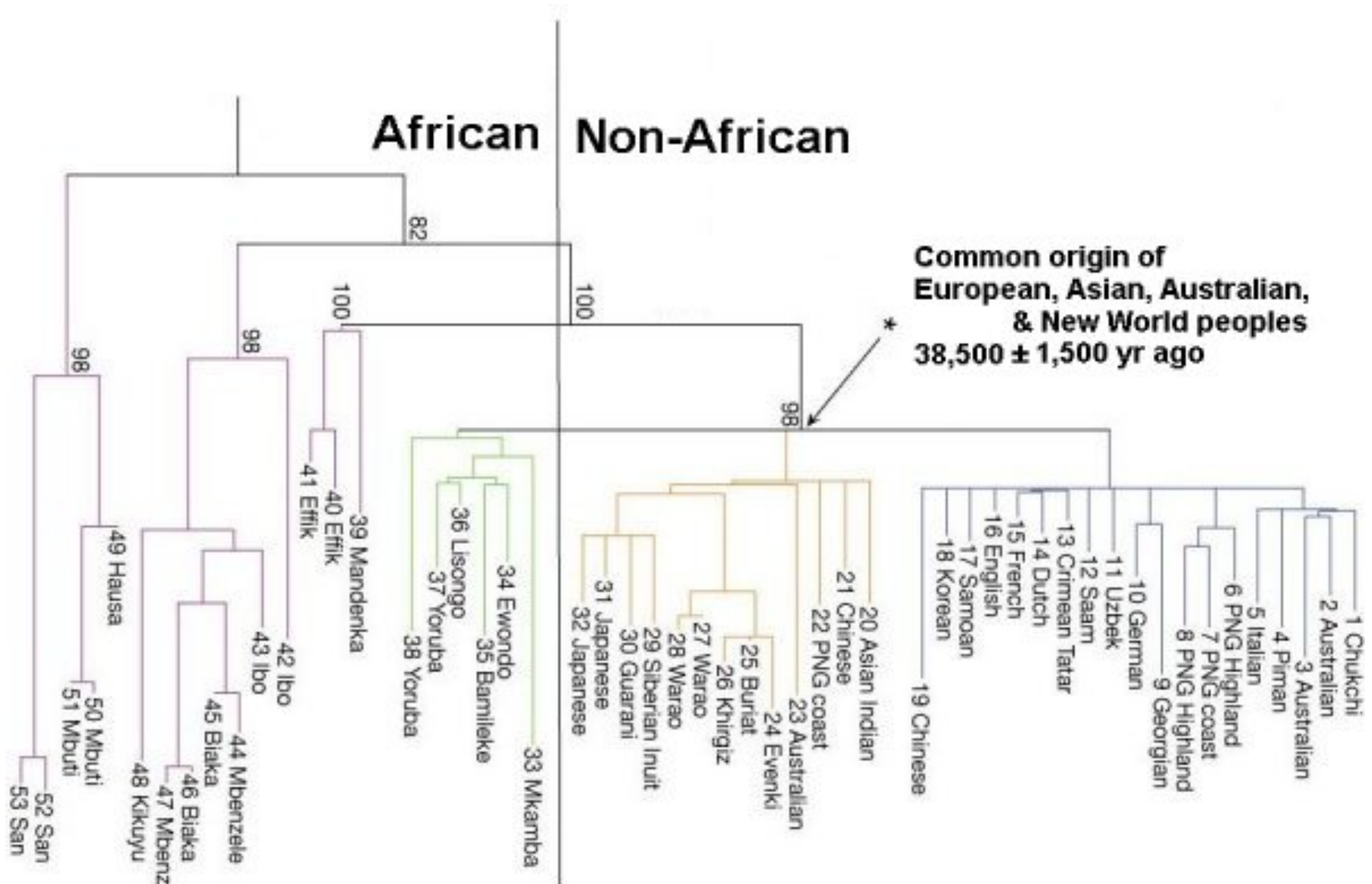


*Maiacetus
inuus*



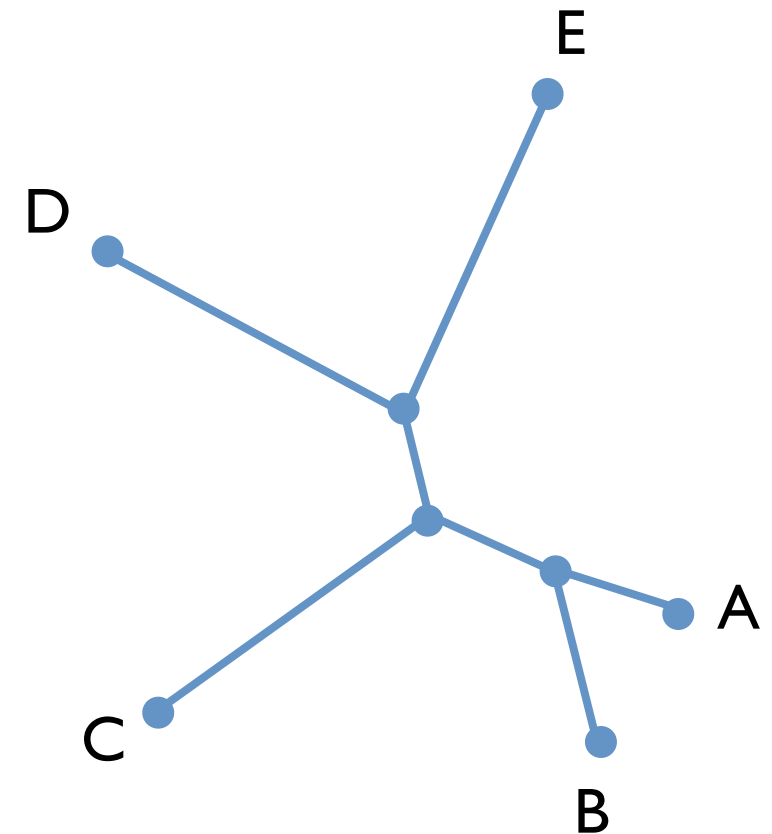
Филогения на основе молекулярных данных.

Пример: гипотеза африканского происхождения человека



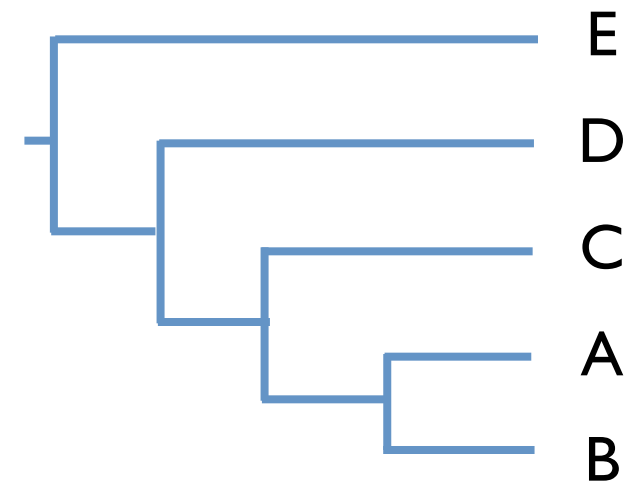
Терминология

- *Дерево* - связный граф, не имеющий циклов
- *Дерево: вершины (узлы), ребра (ветви)*
- *Листья (внешние узлы)* соответствуют существующим в наше время видам, в то как *внутренние узлы* - предковым, последовательности которых неизвестны.
- Последовательность, являющаяся предковой для всех узлов - *корень* дерева.

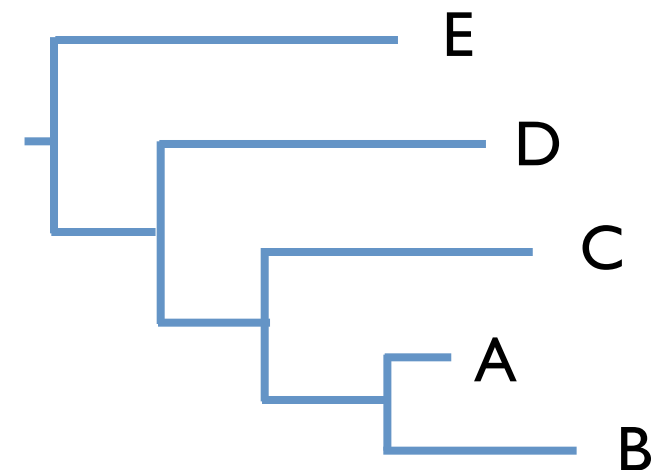


Топология дерева

- *Топология дерева* - уникальная конфигурация узлов и ветвей.
- Дерево, отражающее только топологию и не содержащее информации о длине ветвей, называется *кладограммой*.
- *Филограмма*, в отличие от кладограммы, несет информацию о длине ветвей.
- *Степень узла* - число связанных с ним ветвей.
- Дерево, в котором степень всех узлов не превышает 3-х, а степень корня не превышает 2-х называется *бинарным*.



Кладограмма



Филограмма

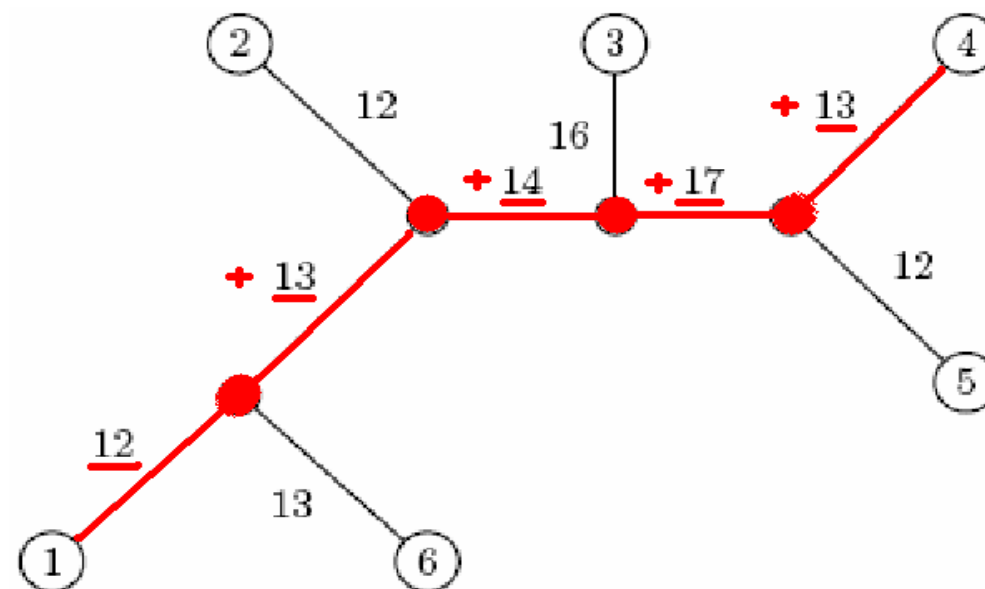
Методы построения филогенетических деревьев

Тип методов	Название метода
Методы, рассматривающие расстояния между последовательностями (Distance-based methods)	Кластеризация (UPGMA, ближайшего соседа)
Методы, рассматривающие признаки (Character-based methods)	Метод максимальной бережливости
	Метод наибольшего правдоподобия

Методы, рассматривающие расстояния между последовательностями (distance-based methods): матрица расстояний

- Имея n последовательностей мы можем составить матрицу расстояний D_{ij} размерности $n \times n$.
- Расстояние между двумя последовательностями можно определить различными способами. Простейший из них - доля позиций в которых остатки двух последовательностей различаются.
- Задача - реконструировать дерево, в котором расстояния между узлами d_{ij} (определение d_{ij} на рисунке) наилучшим образом соответствовали расстояниям между последовательностями: d_{ij} .

$$d_{14} = 12 + 13 + 14 + 17 + 13 = 69$$



UPGMA: метод невзвешенной попарной группировки с усреднением (unweighted pair group method using arithmetic averages)

- UPGMA - алгоритм кластеризации, вычисляющий расстояния между кластерами как усредненные попарные узловые расстояния:

$$d_{ij} = \frac{1}{|C_i| |C_j|} \sum_{p \in C_i, q \in C_j} d_{pq}$$

Алгоритм UPGMA

Инициализация:

Присвоим каждой последовательности x_i ее собственный кластер C_i

Определим по одному листу для каждой последовательности, поместим их на высоту 0

Итерации:

Выберем два кластера C_i, C_j , для которых d_{ij} минимально

Определим $C_k = C_i \cup C_j$, пересчитаем матрицу расстояний

Определим новый узел, соединяющий C_i и C_j , и поместим его на высоту $d_{ij}/2$

Удалим кластеры C_i, C_j

Завершение:

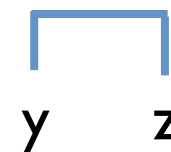
Когда останется только два кластера i и j , поместим корень дерева на высоте $d_{ij}/2$

Алгоритм UPGMA: пример

	v	w	x	y	z
v	0	6	8	8	8
w		0	8	8	8
x			0	4	4
y				0	2
z					0

Алгоритм UPGMA: пример

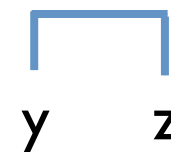
	v	w	x	y	z
v	0	6	8	8	8
w		0	8	8	8
x			0	4	4
y				0	2
z					0



Алгоритм UPGMA: пример

	v	w	x	y	z
v	0	6	8	8	8
w		0	8	8	8
x			0	4	4
y				0	2
z					0

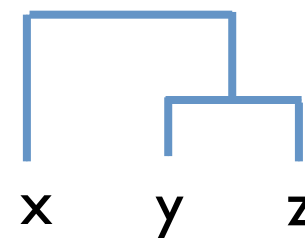
	v	w	x	yz
v	0	6	8	8
w		0	8	8
x			0	4
yz				0



Алгоритм UPGMA: пример

	v	w	x	y	z
v	0	6	8	8	8
w		0	8	8	8
x			0	4	4
y				0	2
z					0

	v	w	x	yz
v	0	6	8	8
w		0	8	8
x			0	4
yz				0

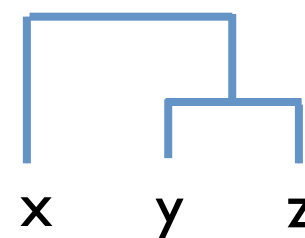


Алгоритм UPGMA: пример

	v	w	x	y	z
v	0	6	8	8	8
w		0	8	8	8
x			0	4	4
y				0	2
z					0

	v	w	xyz
v	0	6	8
w		0	8
xyz			0

	v	w	x	yz
v	0	6	8	8
w		0	8	8
x			0	4
yz				0

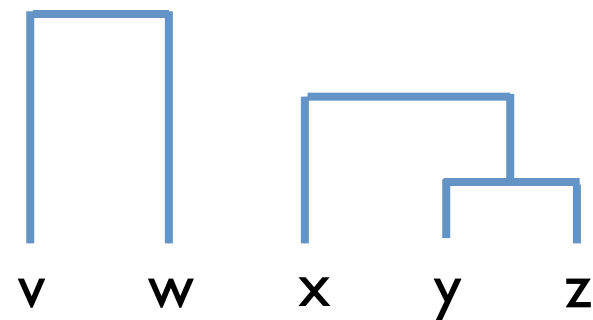


Алгоритм UPGMA: пример

	v	w	x	y	z
v	0	6	8	8	8
w		0	8	8	8
x			0	4	4
y				0	2
z					0

	v	w	xyz
v	0	6	8
w		0	8
xyz			0

	v	w	x	yz
v	0	6	8	8
w		0	8	8
x			0	4
yz				0



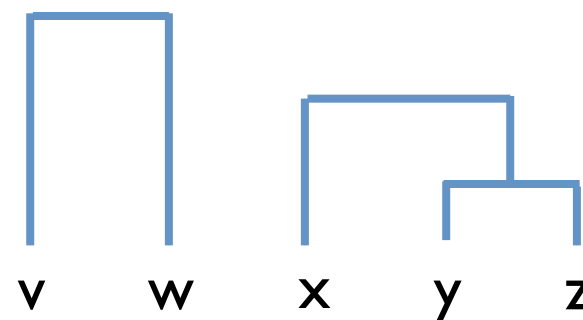
Алгоритм UPGMA: пример

	v	w	x	y	z
v	0	6	8	8	8
w		0	8	8	8
x			0	4	4
y				0	2
z					0

	v	w	xyz
v	0	6	8
w		0	8
xyz			0

	vw	xyz
vw	0	8
xyz		0

	v	w	x	yz
v	0	6	8	8
w		0	8	8
x			0	4
yz				0



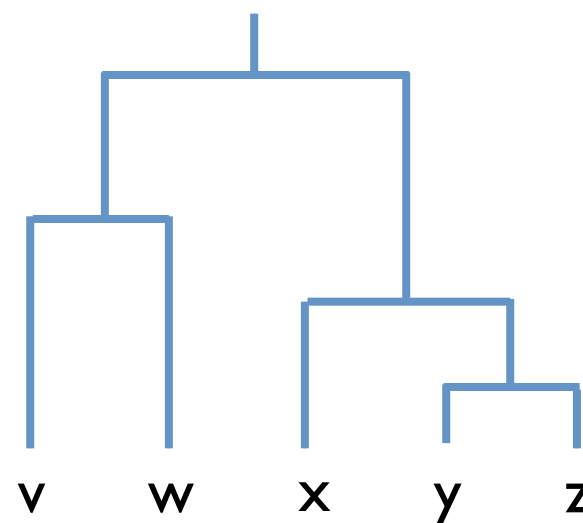
Алгоритм UPGMA: пример

	v	w	x	y	z
v	0	6	8	8	8
w		0	8	8	8
x			0	4	4
y				0	2
z					0

	v	w	xyz
v	0	6	8
w		0	8
xyz			0

	vw	xyz
vw	0	8
xyz		0

	v	w	x	yz
v	0	6	8	8
w		0	8	8
x			0	4
yz				0

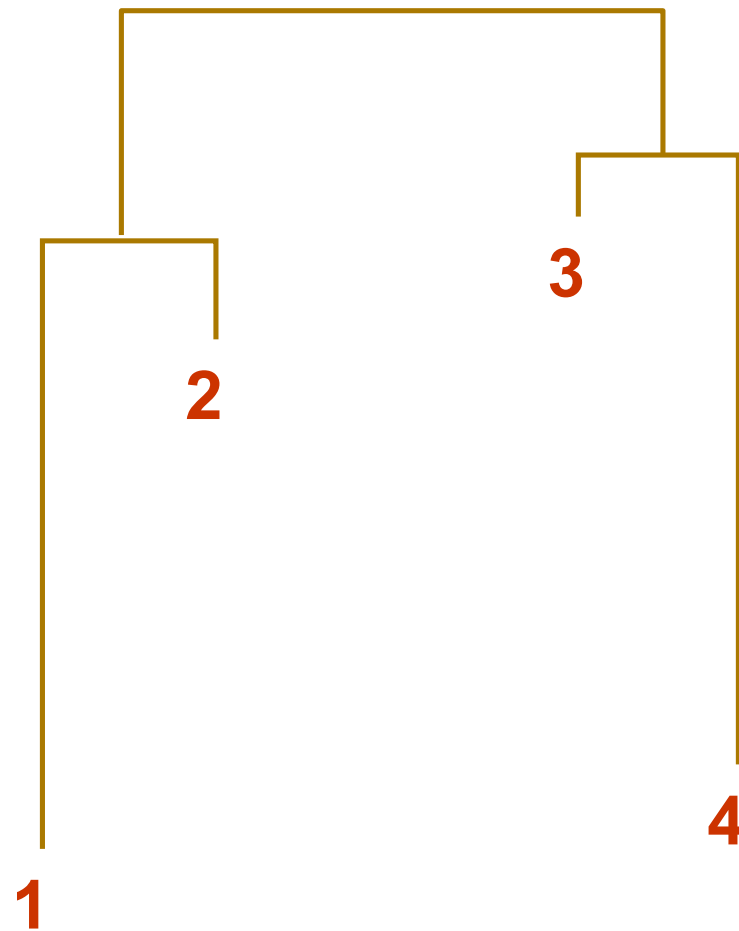


Недостатки UPGMA

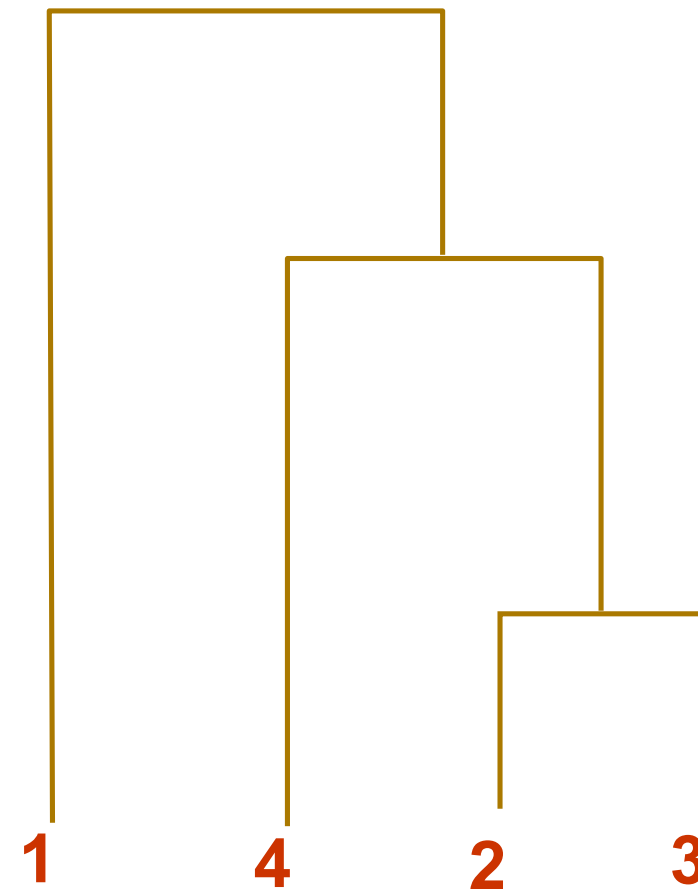
- Алгоритм строит *ультраметрические* деревья - деревья, в которых расстояния от корня до всех листьев одинаково. Это подразумевает постоянство скорости *молекулярных часов* - процесса накопления мутаций (эволюции) последовательностями.
- Известно, что некоторые виды или белки могут эволюционировать быстрее чем другие, т.е. молекулярные часы не всегда постоянны.

Недостатки UPGMA: пример

Исходное дерево



Реконструкция методом UPGMA

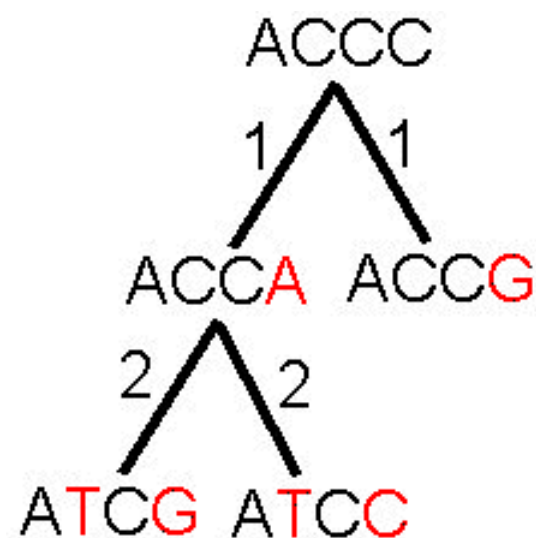


Методы рассматривающие признаки (character-based methods): метод максимальной бережливости (maximum parsimony)

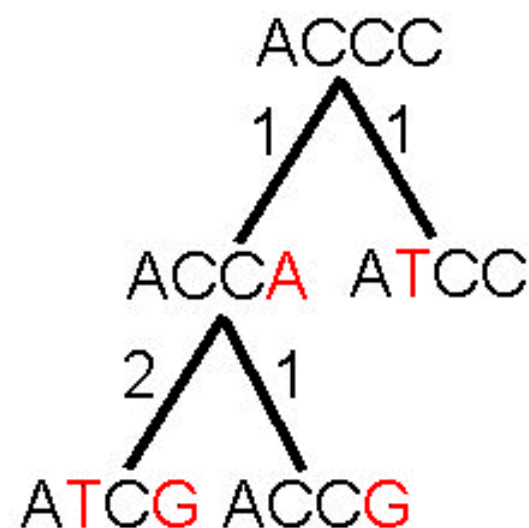
- Методы рассматривающие признаки реконструируют дерево используя информацию из каждой позиции выравнивания, а не последовательности в целом (как методы на расстояниях). Признаны более точными методами.
- Эти методы могут применяться к любым признакам, не только к последовательностям.
- Метод максимальной бережливости реконструирует дерево, которое может объяснить возникновение наблюдаемых последовательностей путем наименьшего количества мутаций (принцип минимальной эволюции).

Метод максимальной бережливости (maximum parsimony)

- В отличие от методов на расстояниях, данный метод не строит дерево, а дает оценку максимальной бережливости для дерева с конкретной топологией.
- Для того, чтобы найти оптимальное дерево, с точки зрения максимальной бережливости, нужно рассмотреть всевозможные топологии.
- Таким образом метод можно поделить на две составляющие:
 - оценка максимальной бережливости для конкретного дерева;
 - поиск оптимальной топологии по всем деревьям;



оценка бережливости = 6



оценка бережливости = 5

Метод максимальной бережливости (maximum parsimony): алгоритмы оценки

- Взвешенная максимальная бережливость [Sankoff 1983]:
подсчитывается не количество замен, а сумму стоимостей каждой замены. Пример стоимостей замен:

	A	T	G	C
A	0	3	4	9
T	3	0	2	4
G	4	2	0	4
C	9	4	4	0

- Традиционная максимальная бережливость [Fitch 1971] -
частный случай взвешенной бережливости $S(a,a) = 0$, $S(a,b) = 1$

Алгоритмы поиска оптимальной топологии

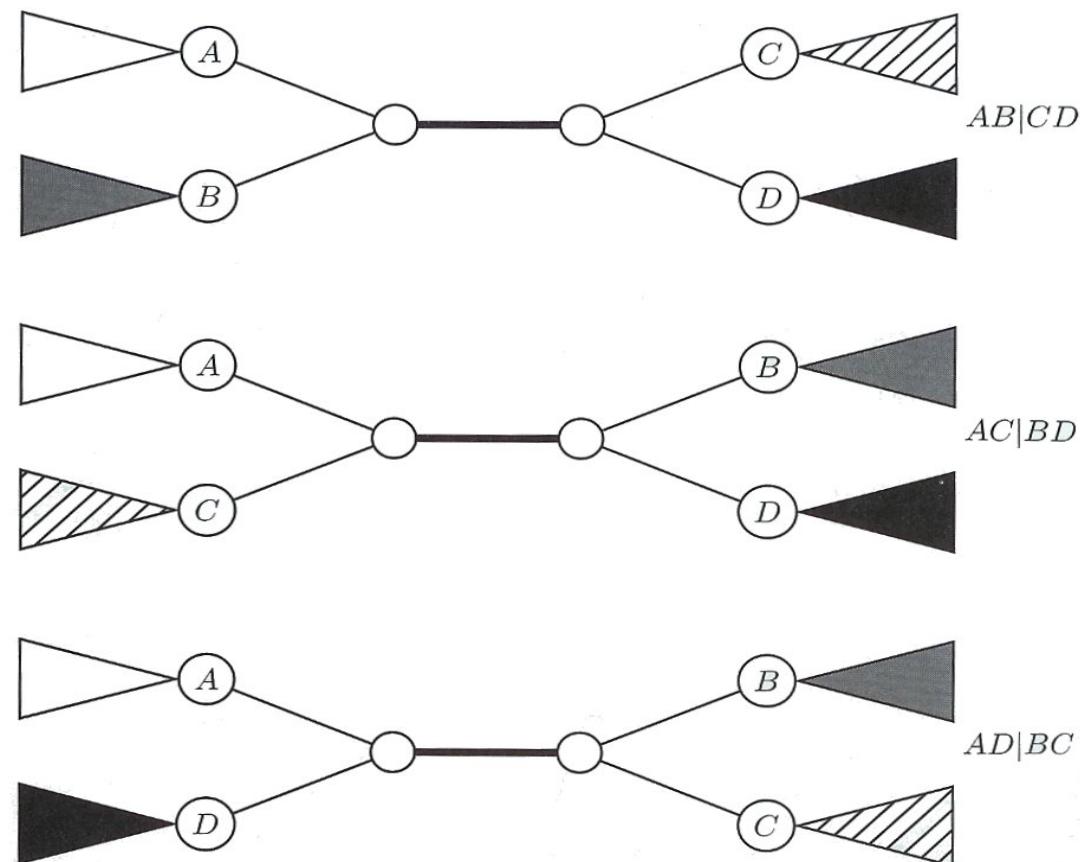
- Полный перебор - $(2n-5)!!$ вариантов, где n - число листьев. Осуществимо только для небольших n ($n < 10$).
- Для метода максимальной бережливости существует оптимизированный метод перебора топологий - метод *ветвей и границ*.
- Эвристические методы - алгоритмы объединения/разделения, перестановок.

Метод ветвей и границ (branch-and-bound algorithm)

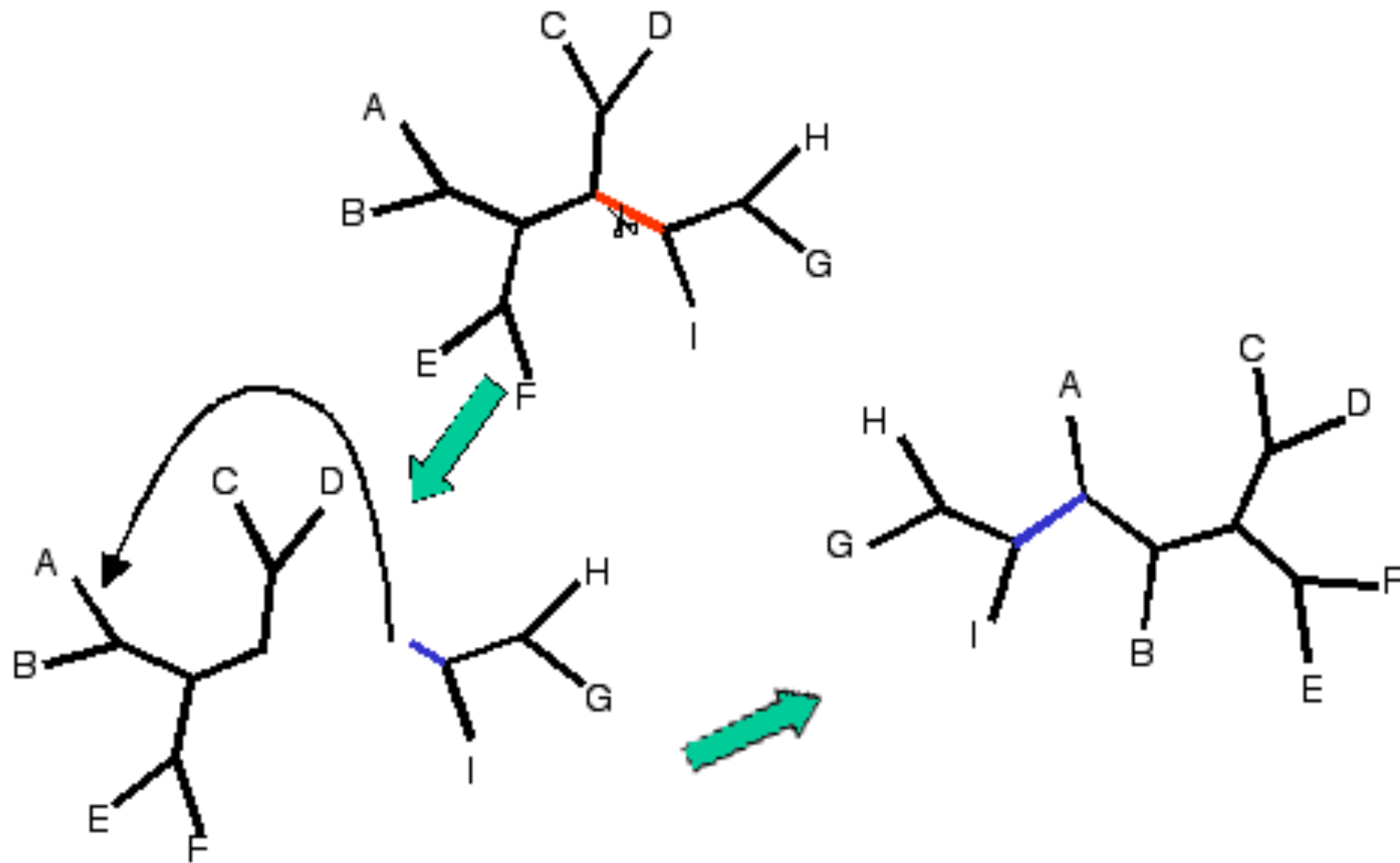
- Используется для подхода максимальной бережливости
- Использует тот факт, что при добавлении новой ветви к существующему дереву оценка максимальной бережливости может только увеличиваться
- Систематически просматривает топологии в порядке увеличения количества ветвей, и останавливает текущий путь построения, если стоимость текущего дерева превышает наименьшую стоимость полного дерева, полученного на предыдущих шагах.

Методы перестановок: обмен ближайшими соседями (nearest-neighbour interchange)

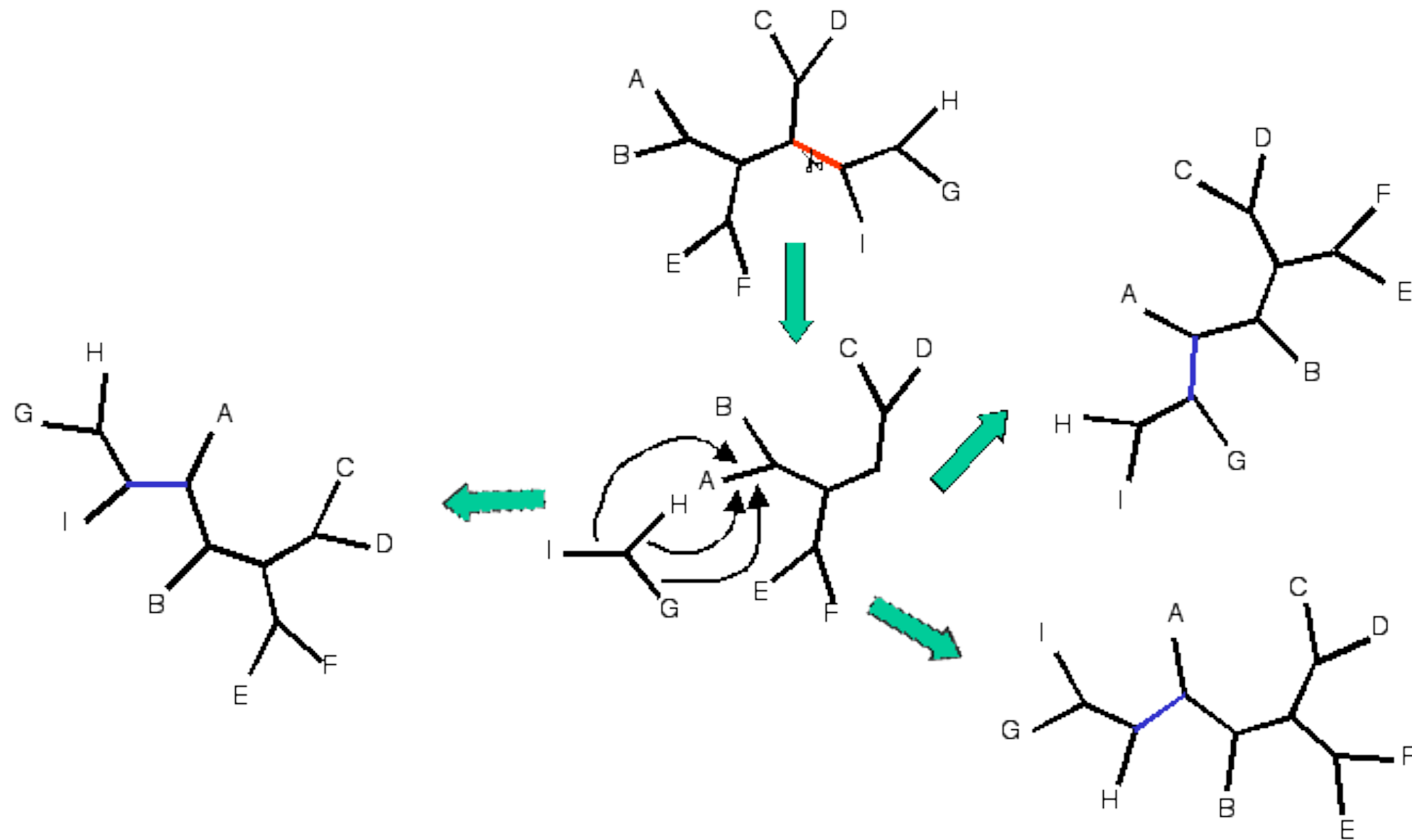
- Метод начинает работу с произвольной топологии.
- Для каждого ребра выполняется перестановка ближайших соседей.
- Выполняется оценка новых топологий.
- Процесс продолжается для наилучшей из новых топологий.



Методы перестановок: метод отсечения и пересадки (subtree pruning and regrafting)



Методы перестановок: метод деления пополам и повторного соединения (tree bisection and reconnection)



Вероятностные модели эволюции

- Модель Джукса-Кантора: каждый нуклеотид имеет одинаковую скорость замены λ в другой нуклеотид.

$$Q = \{q_{ij}\} = \begin{bmatrix} -3\lambda & \lambda & \lambda & \lambda \\ \lambda & -3\lambda & \lambda & \lambda \\ \lambda & \lambda & -3\lambda & \lambda \\ \lambda & \lambda & \lambda & -3\lambda \end{bmatrix}$$

$$P(t) = e^{Qt}$$

$$P(t) = \begin{bmatrix} p_0(t) & p_1(t) & p_1(t) & p_1(t) \\ p_1(t) & p_0(t) & p_1(t) & p_1(t) \\ p_1(t) & p_1(t) & p_0(t) & p_1(t) \\ p_1(t) & p_1(t) & p_1(t) & p_0(t) \end{bmatrix}$$

$$\begin{cases} p_0(t) = \frac{1}{4} + \frac{3}{4}e^{-4\lambda t} \\ p_1(t) = \frac{1}{4} - \frac{1}{4}e^{-4\lambda t} \end{cases}$$

Вероятностные модели эволюции

- Модель Кимуры: транзиции: $T \Leftrightarrow C, A \Leftrightarrow G$
 трансверсии: $\{T, C\} \Leftrightarrow \{A, G\}$.
- Транзиции (скорость замен - α) более вероятны, чем трансверсии (скорость замен - β)

$$Q = \begin{bmatrix} -(\alpha + 2\beta) & \alpha & \beta & \beta \\ \alpha & -(\alpha + 2\beta) & \beta & \beta \\ \beta & \beta & -(\alpha + 2\beta) & \alpha \\ \beta & \beta & \alpha & -(\alpha + 2\beta) \end{bmatrix}$$

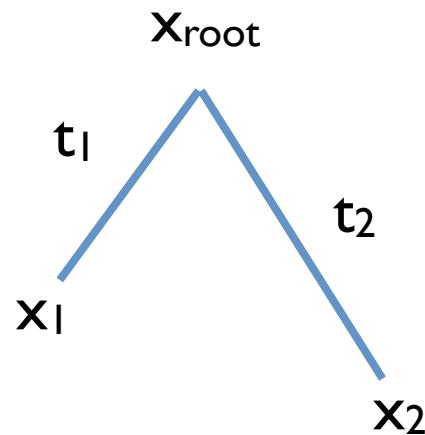
$$P(t) = e^{Qt} = \begin{bmatrix} p_0(t) & p_1(t) & p_2(t) & p_2(t) \\ p_1(t) & p_0(t) & p_2(t) & p_2(t) \\ p_2(t) & p_2(t) & p_0(t) & p_1(t) \\ p_2(t) & p_2(t) & p_1(t) & p_0(t) \end{bmatrix}$$

$$p_0(t) = \frac{1}{4} + \frac{1}{4}e^{-4\beta t} + \frac{1}{2}e^{-2(\alpha+\beta)t}$$

$$p_1(t) = \frac{1}{4} + \frac{1}{4}e^{-4\beta t} - \frac{1}{2}e^{-2(\alpha+\beta)t}$$

$$p_2(t) = \frac{1}{4} - \frac{1}{4}e^{-4\beta t}$$

Вероятностные методы реконструкции деревьев

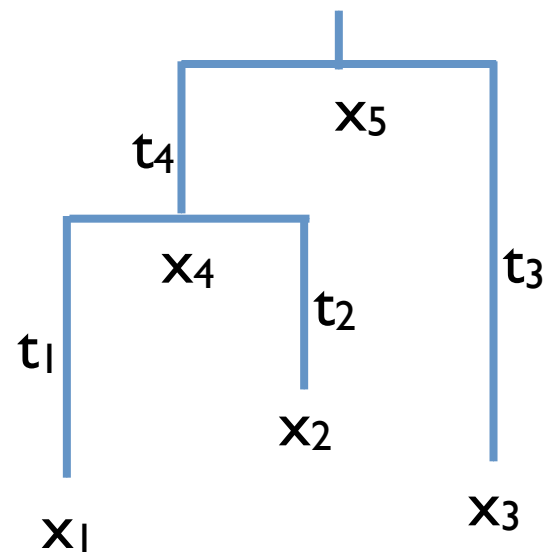


$$P(x_1, x_2, x_{\text{root}} | T, t_1, t_2) = P(x_{\text{root}}) P(x_1 | t_1, x_{\text{root}}) P(x_2 | t_2, x_{\text{root}})$$

Используя модель Джукса-Кантора, если $x_1 = x_{\text{root}} = A$, $x_2 = C$, $t_1 = t_2 = 1$

$$P(x_1, x_2, x_{\text{root}} | T, t_1, t_2) = p_A \frac{1}{4} (1 + 3e^{-4\alpha}) \frac{1}{4} (1 - e^{-4\alpha}) = \left(\frac{1}{4}\right)^3 (1 + 3e^{-4\alpha})(1 - e^{-4\alpha})$$

Вероятностные методы реконструкции деревьев



$$P(x_1, \dots, x_5 | T, t) = P(x_1 | x_4, t_1) P(x_2 | x_4, t_2) P(x_3 | x_5, t_3) P(x_4 | x_5, t_4) P(x_5)$$

т.е., если мы знаем предковые последовательности и во всех узлах и длины ветвей, то вероятность дерева:

$$P(x_1, x_2, \dots, x_N, x_{N+1}, \dots, x_{2N-1} | T, t) = P(x_{\text{root}}) \prod_{j \neq \text{root}} P(x_j | x_{\text{parent}(j)}, t_{j, \text{parent}(j)})$$

но, обычно мы не знаем предковые последовательности, поэтому для вычисления вероятности необходимо просуммировать вероятности по всем возможным значениям внутренних узлов:

$$P(x_1, x_2, \dots, x_N | T, t) = \sum_{x_{N+1}} \sum_{x_{N+2}} \dots \sum_{x_{2N-1}} P(x_1, x_2, \dots, x_{2N-1} | T, t)$$

Метод максимального правдоподобия

Имеется множественное выравнивание из M -колонок и N последовательностей.

- Правдоподобие дерева:

$$L(T, \mathbf{t}) = P(\text{данные} \mid T, \mathbf{t}) = \prod_{m=1 \dots M} P(x_{1m}, \dots, x_{nm} \mid T, \mathbf{t})$$

- Метод максимального правдоподобия:

Имея данные $X = (x_{ij})$, найти топологию **T** и вектор длин ребер **$\mathbf{t}$** , максимизирующие правдоподобие $L(\mathbf{T}, \mathbf{t})$

Благодарности

- При подготовке слайдов использовались материалы лекций:
 - Михаила Гельфанда (ИППИ)
 - Андрея Миронова (МГУ)
 - Serafim Batzoglou (Stanford)
 - Manolis Kellis (MIT)
 - Pavel Pevzner (UCSD)