

Лекция 5

Метод максимальной парсимонии (продолжение)

Метод максимального правдоподобия

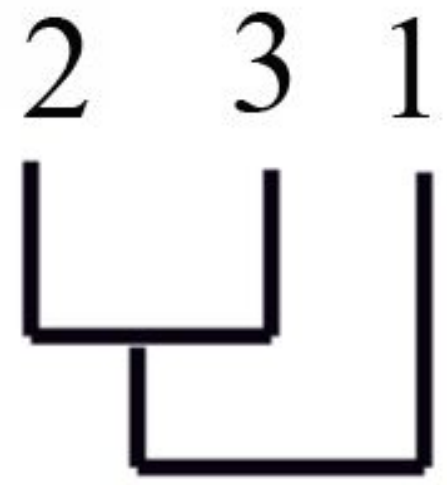
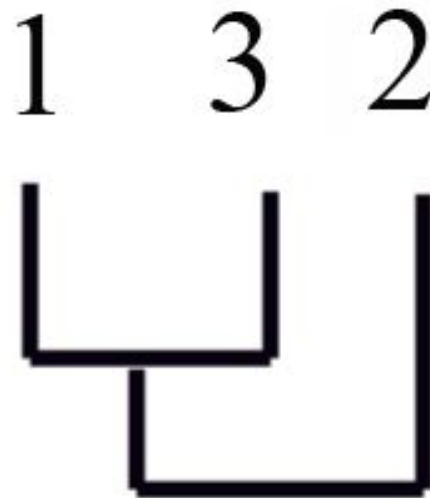
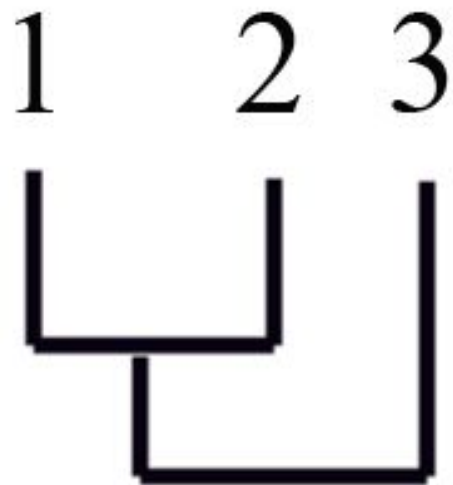
ДНК:

1 5 10
tagcaaaatg

Метод максимальной парсимонии (наибольшей экономии)

Критерий оптимальности:
лучшее дерево - самое простое дерево (самое
короткое)

Ищем все
ВОЗМОЖНЫЕ ТОПОЛОГИИ



Варианты топологий в случае трех таксонов

Для 5 таксонов возможны 15 неукорененных деревьев и 105 укорененных деревьев

Table 1.1: A simple data set with 0/1 characters.

Species	Characters					
	1	2	3	4	5	6
Alpha	1	0	0	1	1	0
Beta	0	0	1	0	0	0
Gamma	1	1	0	0	0	0
Delta	1	1	0	1	1	1
Epsilon	0	0	1	1	1	0

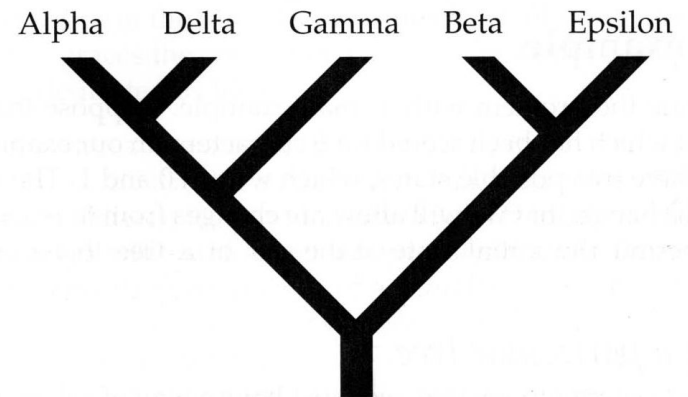


Figure 1.1: A phylogeny that we want to evaluate using parsimony.

Один из вариантов топологии

Существует $(2n-5)!!$ разных неукорененных деревьев с n вершинами

Number of sequences	Number of unrooted trees	Number of rooted trees
2	1	1
3	1	3
4	3	15
5	15	105
6	105	945
7	945	10395
8	10395	135135
9	135135	2027025
10	2027025	34459425

Numbers of unrooted and rooted trees for 2–10 sequences.

Вначале ищем все возможные ТОПОЛОГИИ

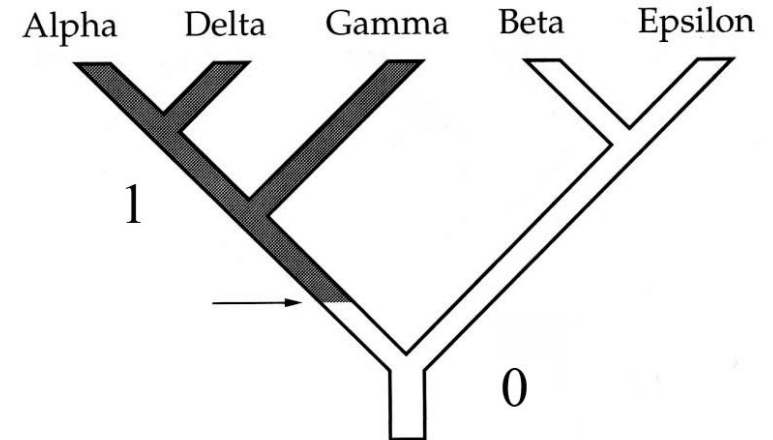
Если число таксонов равно n , существует $(2n-3)!!$ разных бинарных укорененных деревьев. $(2n-3)!!$ – это нечто вроде факториала, но

Для каждой топологии рассматриваем все возможные варианты эволюции каждого признака

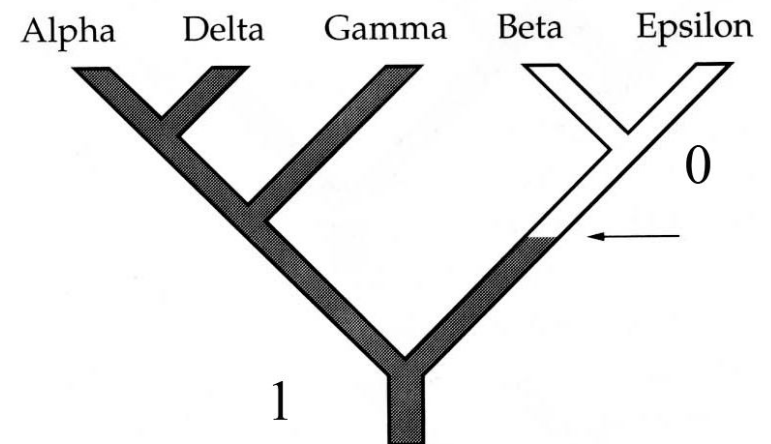
Признак 1

Table 1.1: A simple data set with 0/1 characters.

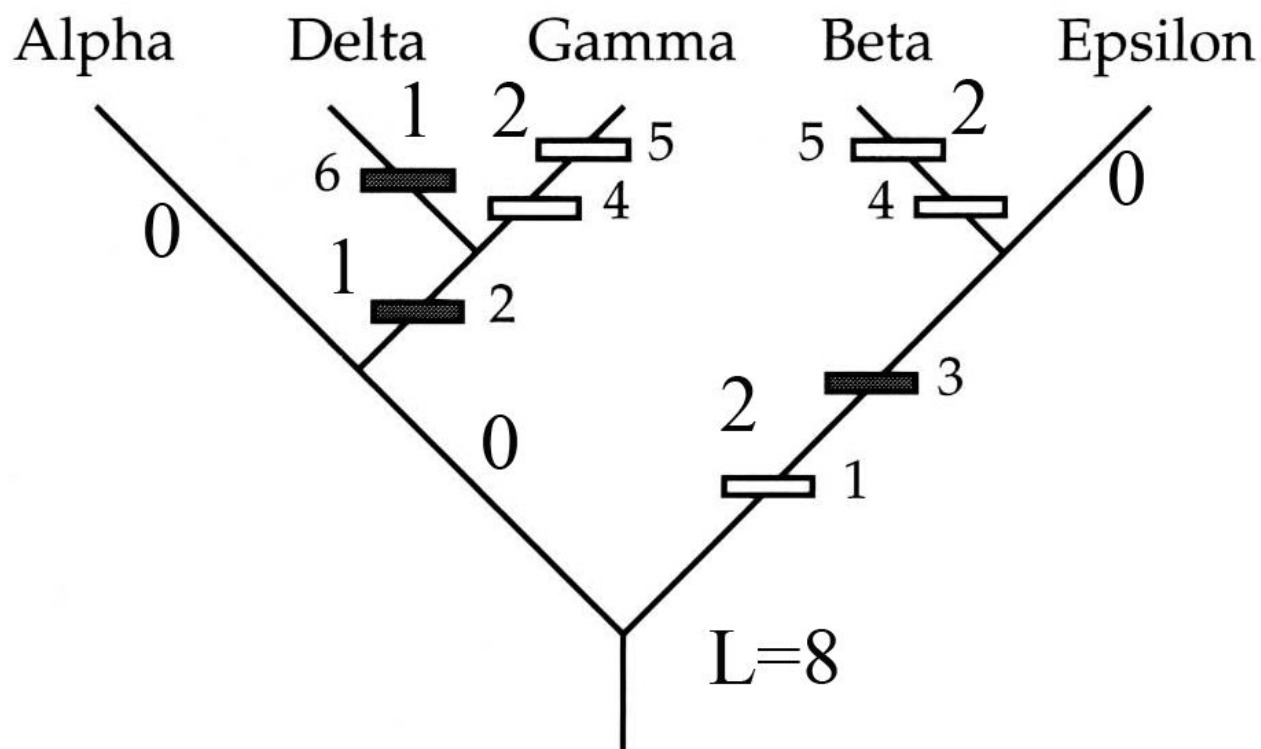
Species	Characters					
	1	2	3	4	5	6
Alpha	1	0	0	1	1	0
Beta	0	0	1	0	0	0
Gamma	1	1	0	0	0	0
Delta	1	1	0	1	1	1
Epsilon	0	0	1	1	1	0



or



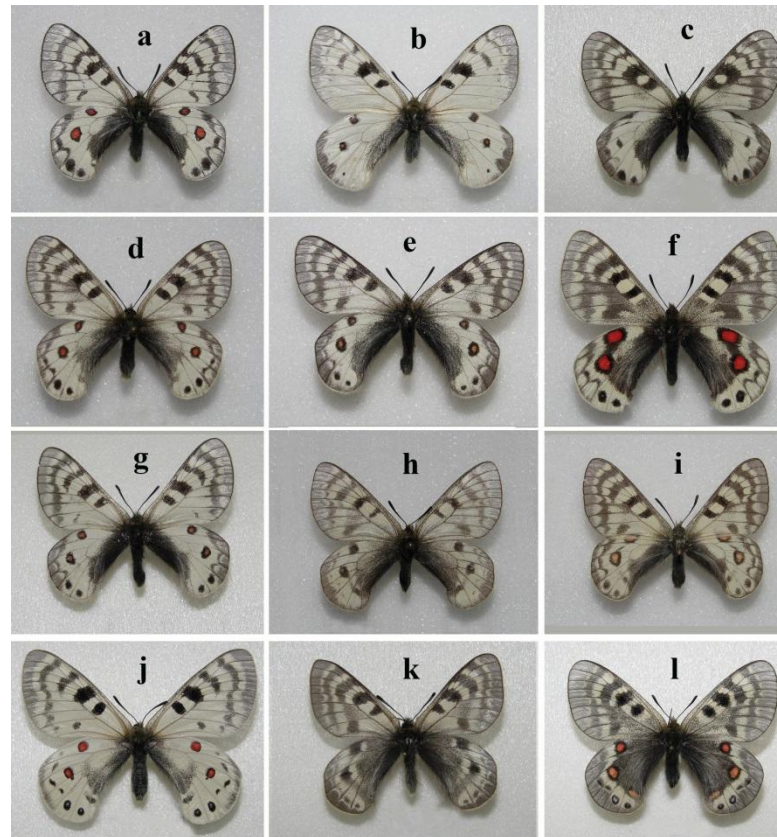
Считаем число изменений признаков в каждом из эволюционных сценариев



Анализ парсимониальных деревьев

- Выявление равнопарсимониальных деревьев
- Построение консенсуса

Пример
Филогения бабочек рода *Parnassius*,
основанная на анализе гена COI с
использованием метода максимальной
парсимонии



Проверка устойчивости филогенетической реконструкции

Нужна статистика:
среднее значение и уровень изменчивости

Варианты

реальная статистика и
bootstrapping



Проверка устойчивости филогенетической реконструкции

Jackknife (метод вырезания)

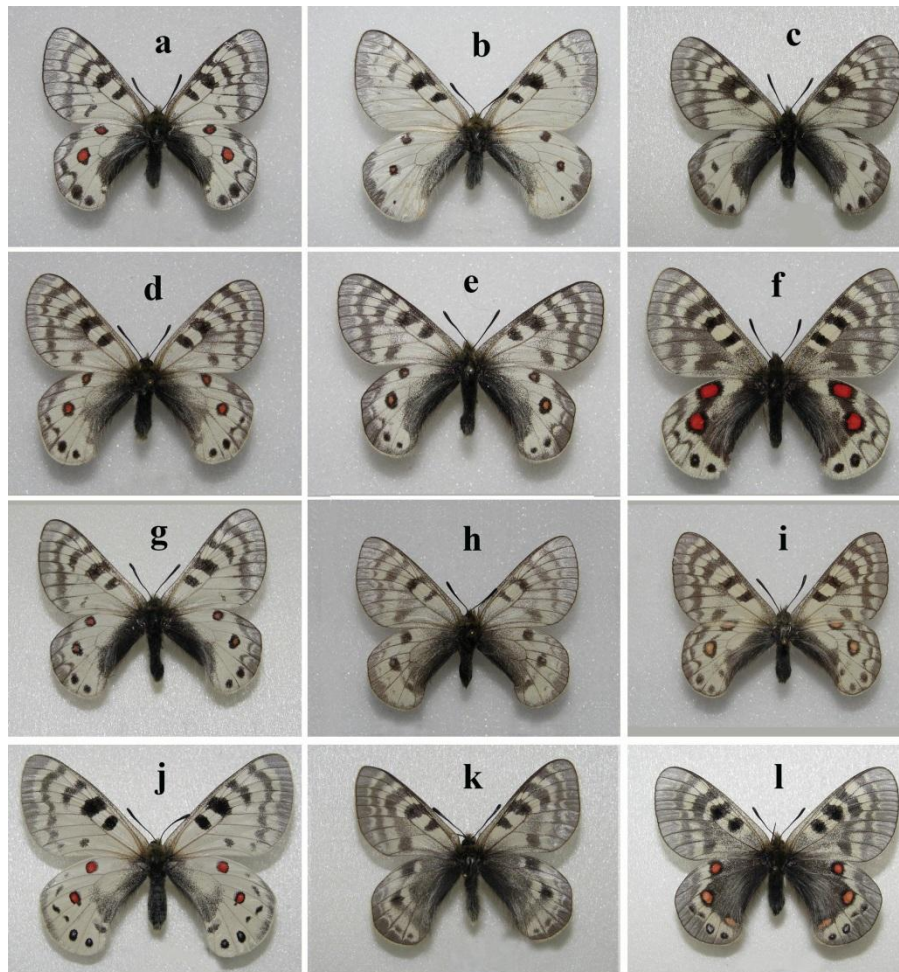
1	2	3	4	5	6	7	8	9	10	11
A	T	A	A	C	A	T	A	A	G	A
C	T	A	T	T	A	T	A	A	G	A
C	T	A	T	C	A	T	A	A	G	A
A	C	A	A	C	G	T	A	A	G	A
A	T	A	A	C	A	T	A	T	G	C
A	T	G	A	C	A	A	A	T	G	A
A	T	G	A	C	A	T	A	A	G	A
A	T	G	A	C	A	T	A	A	G	A

Проверка устойчивости филогенетической реконструкции

- Бутстреп (bootstrap)
 - Что это такое?

1	2	3	4	5	6	7	8	9	10	11
A	T	A	A	C	A	T	A	A	G	A
C	T	A	T	T	A	T	A	A	G	A
C	T	A	T	C	A	T	A	A	G	A
A	C	A	A	C	G	T	A	A	G	A
A	T	A	A	C	A	T	A	T	G	C
A	T	G	A	C	A	A	A	T	G	A
A	T	G	A	C	A	T	A	A	G	A
A	T	G	A	C	A	T	A	A	G	A

Бутстреп-анализ
филогении бабочек рода *Parnassius*
(ген *COI*, метод максимальной парсимонии)



- Бутстреп - это не вероятность данной клады!!!!
- Это скорее мера ее устойчивости при искусственной манипуляции с данными

Проверка устойчивости филогенетической реконструкции

- Бутстреп
 - Что это такое?
 - Сколько псевдореплик нужно получать?
 - Какой смысл имеют разные проценты бутстреп-поддержки?
 - Ограничение в применении метода бутстрепа (малое число признаков)

Bremer support (поддержка Бремера)

- Мы выбрали наиболее парсимониальное дерево, в этом случае на дереве имеется определенная клада
- А что будет если мы возьмем менее парсимониальное (т.е. более длинное дерево)?
Сохранится ли эта клада?
- Да, если есть запас прочности в виде набора синапоморфий

Bremer support

- $BS=0$

Удлинение дерева на один шаг приводит к тому, что клада исчезает

- $BS=1$

При удлинении дерева на один шаг данная клада сохраняется.

Взвешивание признаков и сайтов – способ задать более сложные модели эволюции в рамках метода максимальной парсимонии

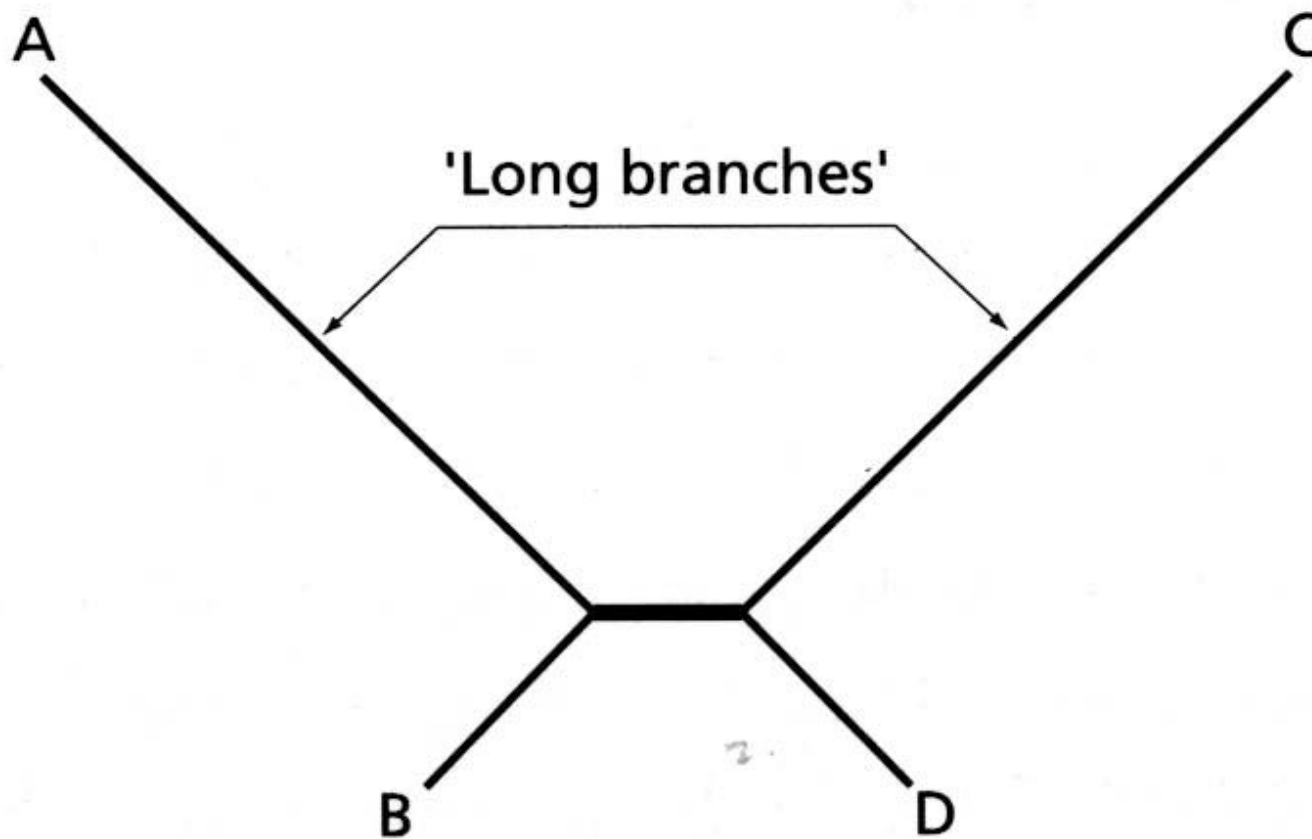
Возможности и ограничения метода максимальной парсимонии

Парсимония как философский принцип и парсимония как математическая модель

Чем реже встречается признак (чем реже его изменения), тем более
адекватно применение принципа парсимонии

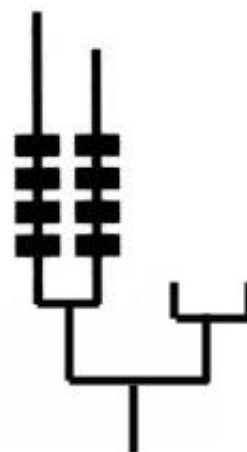
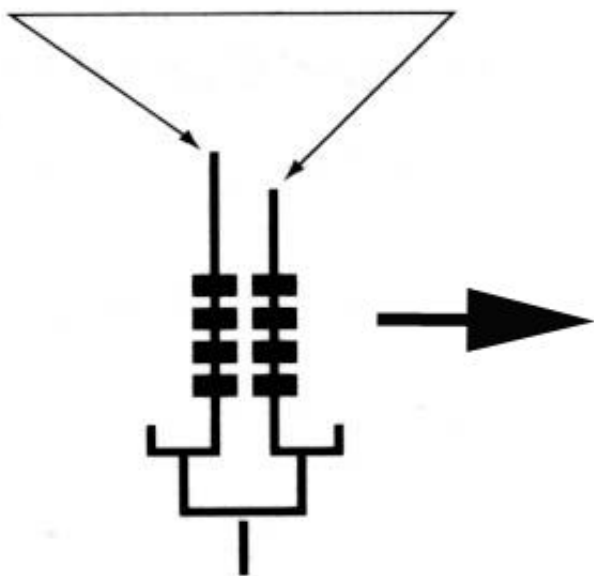
- критерий парсимонии имеет некоторое теоретическое обоснование. Однако в общем виде он является несостоятельным, и при ряде условий его использование приводит к ошибочным реконструкциям (Felsenstein, 1978, 2004)

Проблема длинных ветвей

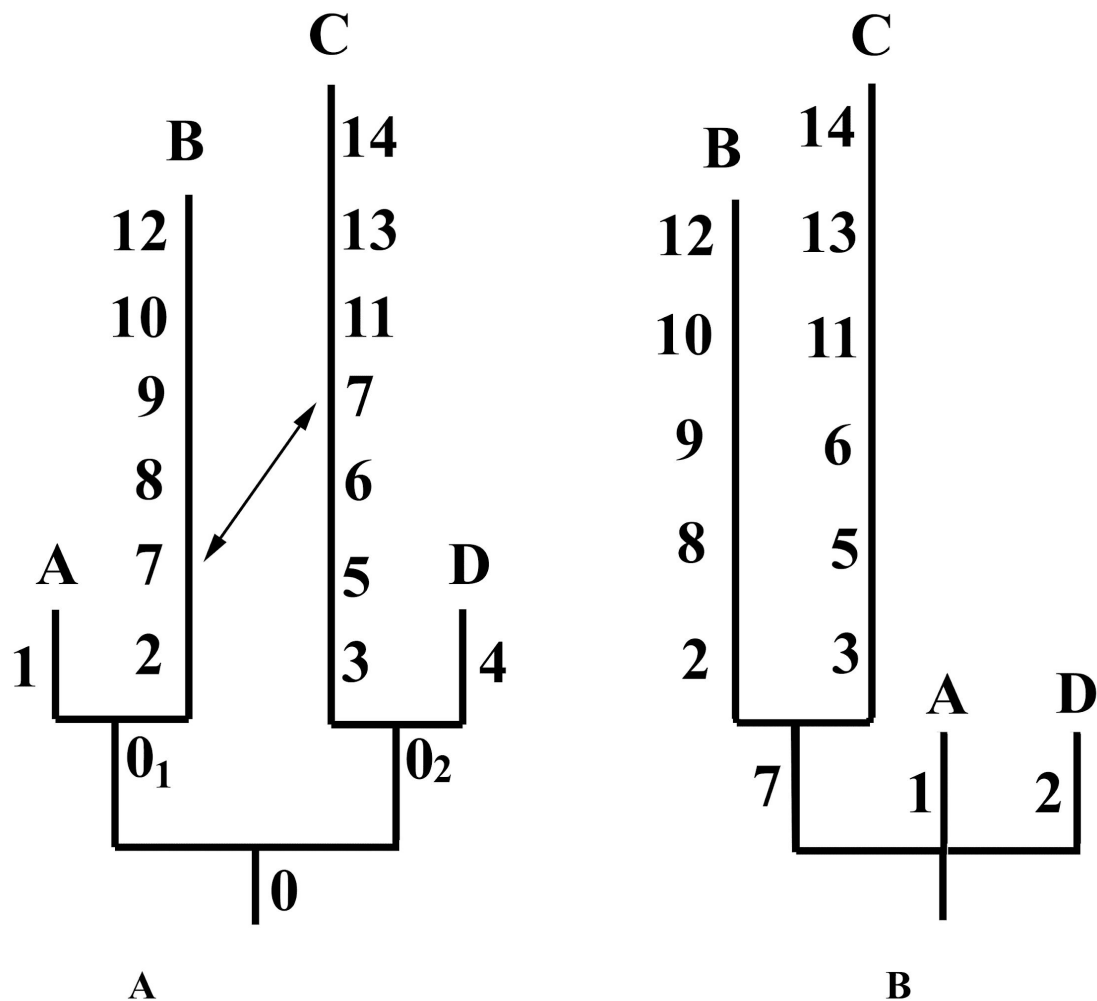


Small tree

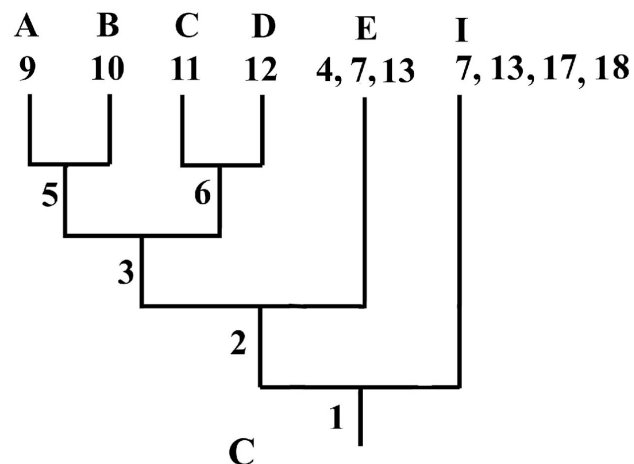
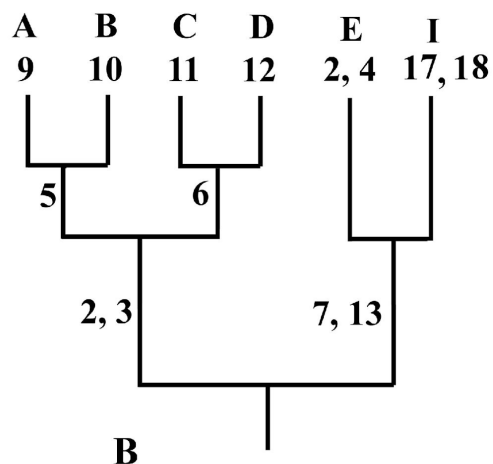
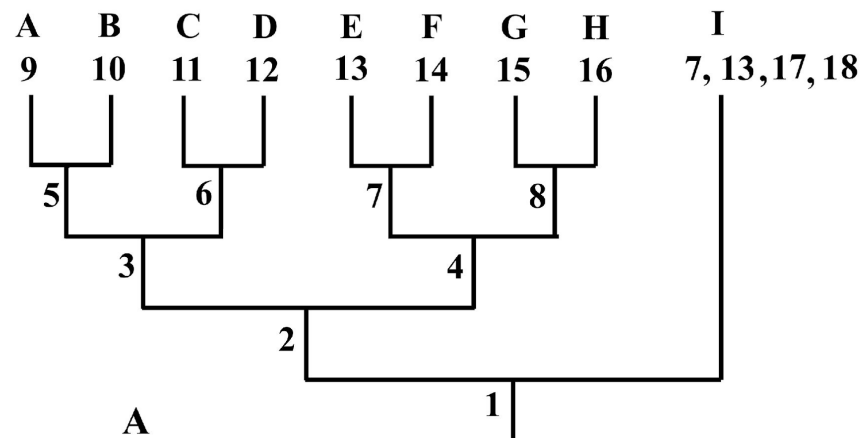
Длинные ветви имеют
большую вероятность, что в них независимо появятся
одинаковые признаки (гомоплазии)



ошибочная интерпретация их
в качестве синапоморфий
ведет к ложной реконструкции



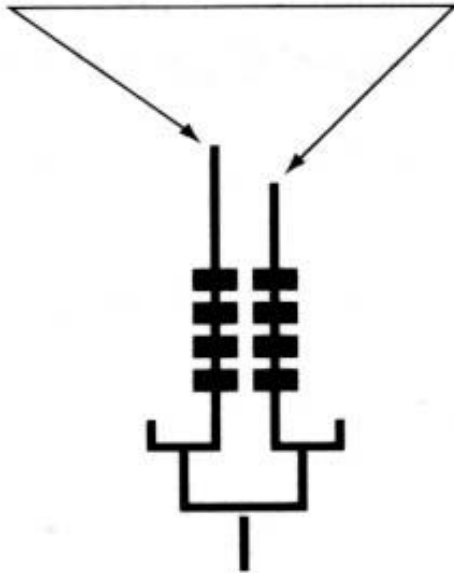
Влияние эффекта притяжения длинных ветвей на результаты парсимониального филогенетического анализа таксонов A, B, C и D. 0 - плезиоморфный признак, 1-14 - апоморфные признаки. A - реальная (истинная) филогения и распределение на ней признаков. B - ложная реконструкция филогении A, получаемая при проведении кладистического анализа с использованием метода максимальной парсимонии



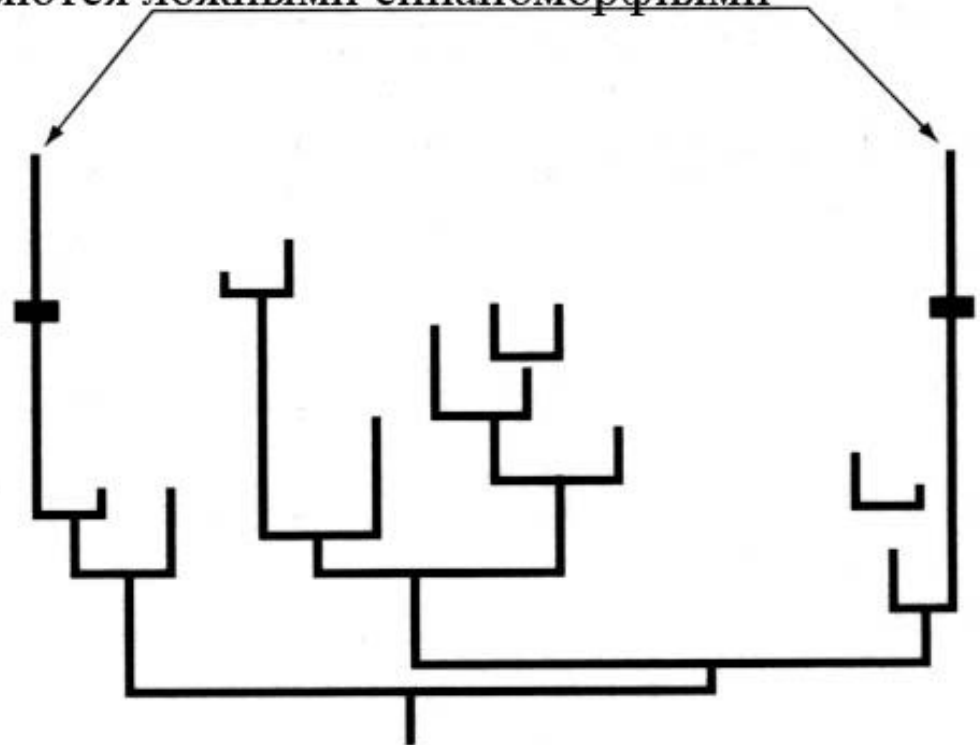
Влияние неполноты выборки таксонов на результаты парсимониального кладистического анализа

Small tree

разбиение длинных ветвей: промежуточные ветки несут
филогенетический сигнал, позволяющий понять,
какие признаки являются ложными синапоморфиями



Large tree



■ = Covarying sites

Критерии оценки методов построения деревьев

- скорость (быстродействие)
- трудоемкость получения исходных данных
- соответствуют ли реконструкции действительности
- помехоустойчивость (чувствительность к отклонениям в модели, в данных)
- проверяемость получаемых выводов

- Правильную ли филогению мы получили?
- Возможные источники ошибок
- Как проверить правильность реконструкции

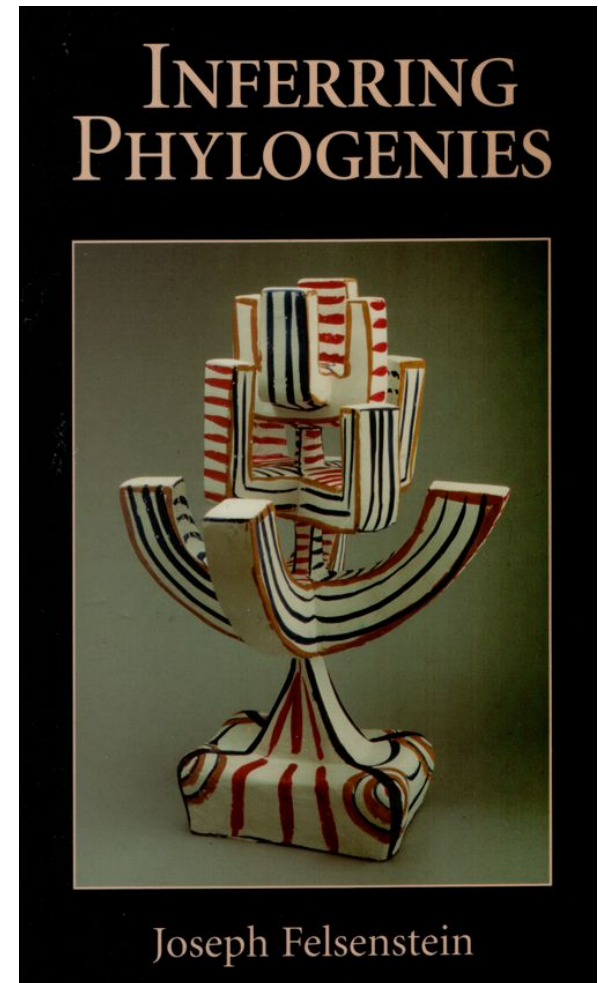
Источники ошибок в филогенетических реконструкциях

- 1) не правильный и/или недостаточный выбор признаков
- 2) неправильный sampling
- 3) неправильный выбор внешней группы (для укорененного дерева)
- 4) выбор неправильной модели или метода
- 5) объективные трудности - сложность структуры самого дерева

Метод максимального правдоподобия



Joseph Felsenstein



Принципы работы метода максимального правдоподобия

- если имеется информация о закономерностях эволюционных преобразований признаков (иными словами, если есть модель эволюции признака),

Принципы работы метода максимального правдоподобия

- если имеется информация о закономерностях эволюционных преобразований признаков (иными словами, если есть модель эволюции признака),
- и известно распределение состояний признаков у изучаемых организмов,

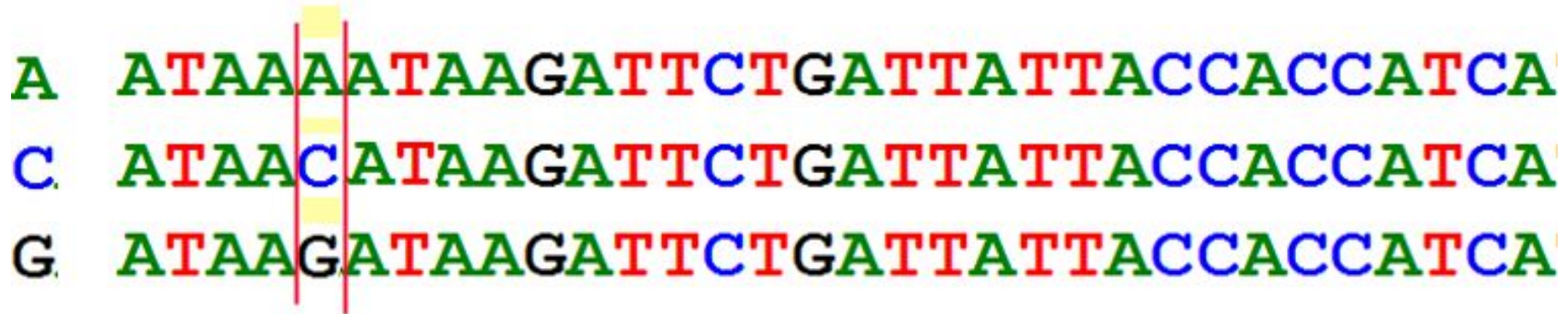
Принципы работы метода максимального правдоподобия

- если имеется информация о закономерностях эволюционных преобразований признаков (иными словами, если есть модель эволюции признака),
- и известно распределение состояний признаков у изучаемых организмов,
- то можно рассчитать **вероятности** различных эволюционных траекторий, которые могли привести к современным формам

Принципы работы метода максимального правдоподобия

- А затем к качеству оптимального дерева выбрать ту траекторию, которая имеет наибольшую вероятность

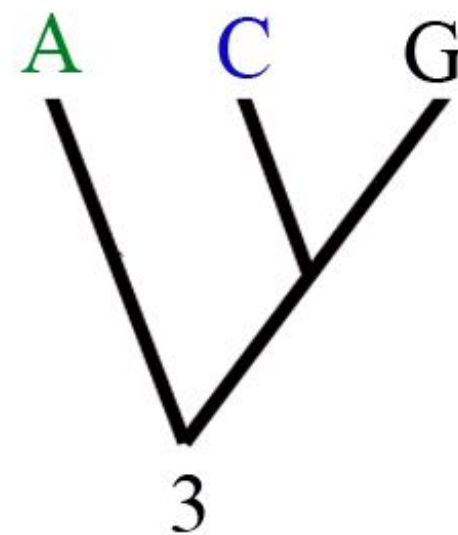
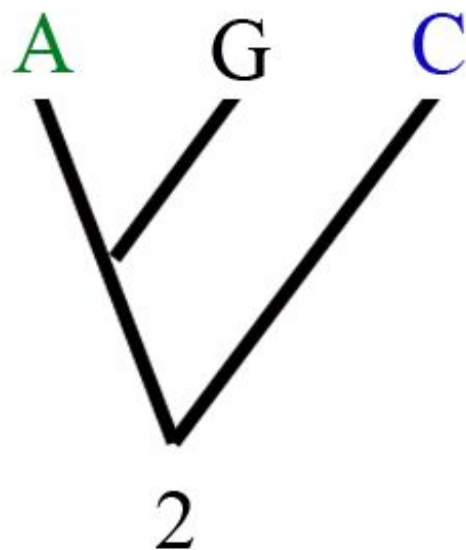
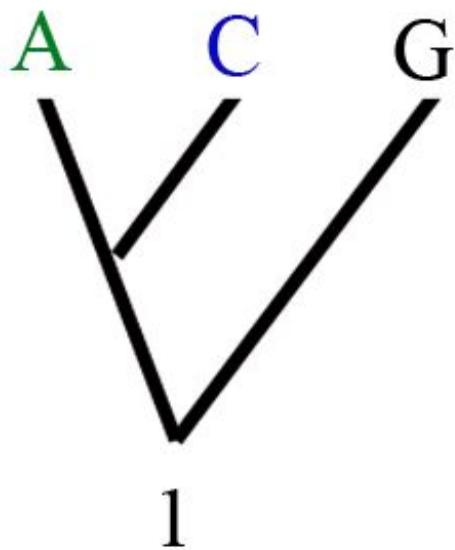
Построение дерева, состоящего из 3 таксонов, с использованием метода максимального правдоподобия



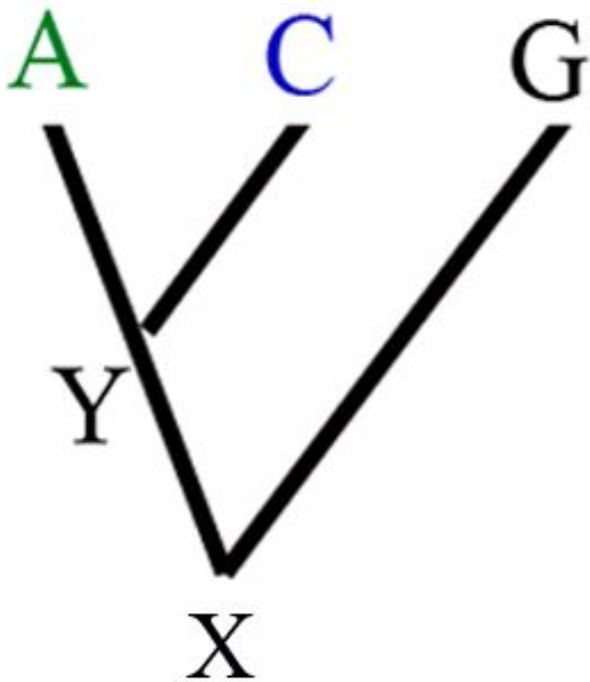
The image displays a sequence alignment of three taxa (A, C, G) across 20 positions. The sequences are color-coded: A (green), T (red), A (green), A (green), A (green), T (red), A (green), G (black), A (green), T (red), T (red), C (blue), T (red), G (black), A (green), T (red), T (red), A (green), C (blue), C (blue), A (green), T (red), C (blue), A (green). A phylogenetic tree is overlaid on the alignment, with two vertical red lines indicating the branching points. The tree structure shows a root at the top, branching into two main lineages. The first lineage leads to a clade containing taxa A and C, while the second lineage leads to taxon G. The branching points are marked by vertical red lines, and the nodes are highlighted with yellow squares.

Taxon	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20					
A	A	T	A	A	A	T	A	G	A	T	T	C	T	G	A	T	T	A	C	C	A	T	C	A	
C	A	T	A	A	C	A	T	A	G	A	T	T	C	T	G	A	T	T	A	C	C	A	T	C	A
G	A	T	A	A	G	A	T	A	G	A	T	T	C	T	G	A	T	T	A	C	C	A	T	C	A

Три возможных дерева



Рассмотрим дерево 1

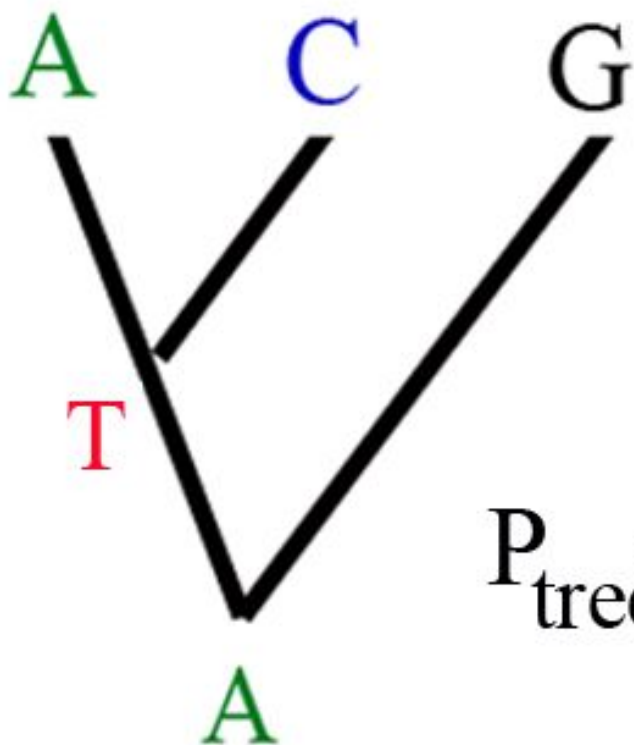


$X = A, \text{ or } C, \text{ or } G, \text{ or } T$

$Y = A, \text{ or } C, \text{ or } G, \text{ or } T$

Возможны 16 вариантов нуклеотидных переходов

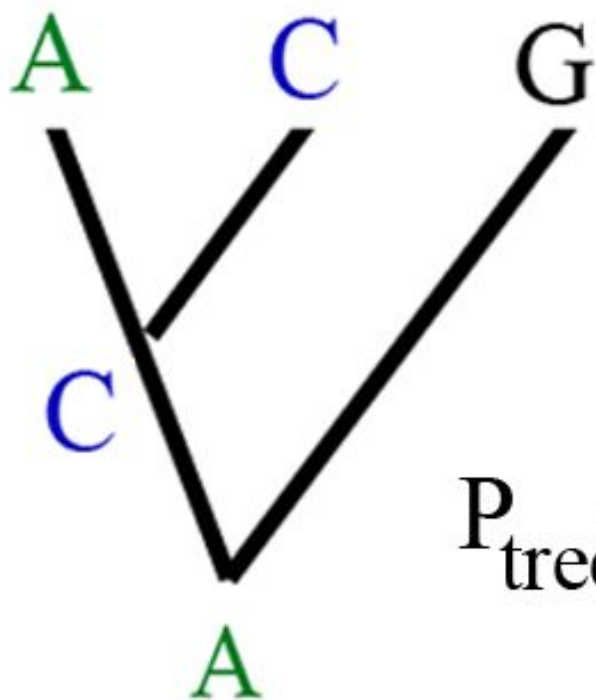
Дерево 1 из 3



$$P_{\text{tree}} = P_A \cdot P_{AT} \cdot P_{TA} \cdot P_{TC} \cdot P_{AG}$$

Вариант 1 из 16

Дерево 1 из 3



$$P_{\text{tree}} = P_A \cdot P_{AC} \cdot P_{CA} \cdot P_{CC} \cdot P_{AG}$$

Вариант 2 из 16

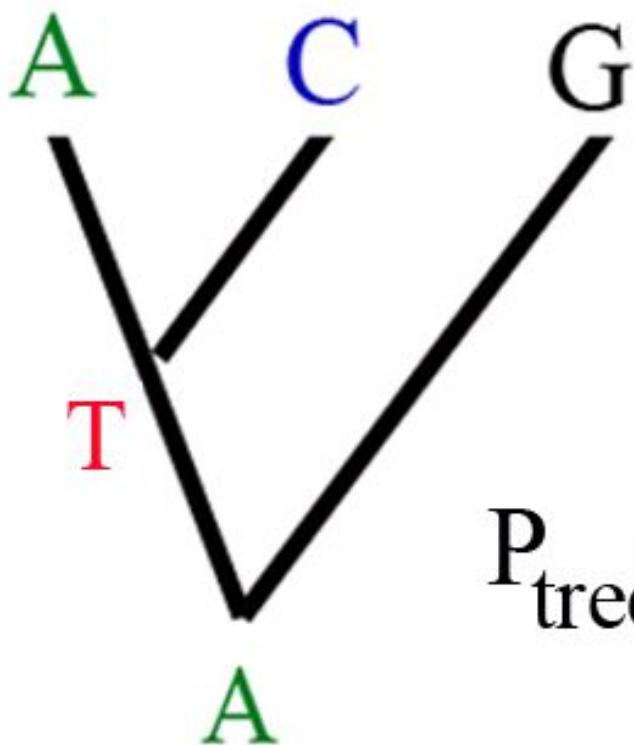
Модель Фитча-Вагнера (Fitch-Wagner parsimony) для нуклеотидных замен

$A \longleftrightarrow C$ $A \longleftrightarrow G$ $A \longleftrightarrow T$

$C \longleftrightarrow G$ $C \longleftrightarrow T$

$G \longleftrightarrow T$

Дерево 1 из 3



Вероятности всех замен одинаковы,
т.е. $P(AC)=P(AG)=P(AT)=P(CA)=P(CG)=P(CT)=$
 $P(GA)=P(GC)=P(GT)=P(TA)=P(TC)=P(TG)=\alpha$

частоты нуклеотидов равны,
т.е. $f(A)=f(C)=f(G)=f(T)=0.25$
 $P_{xy} = \alpha = 1/12=0,083$

$$P_{\text{tree}} = 0.25 \times \alpha \times \alpha \times \alpha \times \alpha =$$
$$= 0.25 \times 0.083 \times 0.083 \times 0.083 \times 0.083$$
$$= 0.00001186$$

$$P_{\text{tree}} = P_A \cdot P_{AT} \cdot P_{TA} \cdot P_{TC} \cdot P_{AG}$$

- Это вероятность конкретного сценария в контексте вероятностей отдельных событий.
 - Поэтому для этой величины используют понятие *правдоподобие*
 - Правдоподобие гипотезы = 0.00001186
 - Сумма правдоподобий не равна единице!
 $0.00001186 \times 48 = 0.00056928$
- Но это не тоже самое что вероятность дерева как гипотезы.
 - $P(\text{Вероятность гипотезы}) = 1/48 = 0.0208$
 - Сумма вероятностей = 1!

- Вопрос: какую модель мы использовали?

JC model

Вероятности всех замен одинаковы,
т.е. $P(AC)=P(AG)=P(AT)=P(CG)=P(CT)=P(GT)=\alpha$

частоты нуклеотидов равны, т.е. $f(A)=f(C)=f(G)=f(T)=0.25$

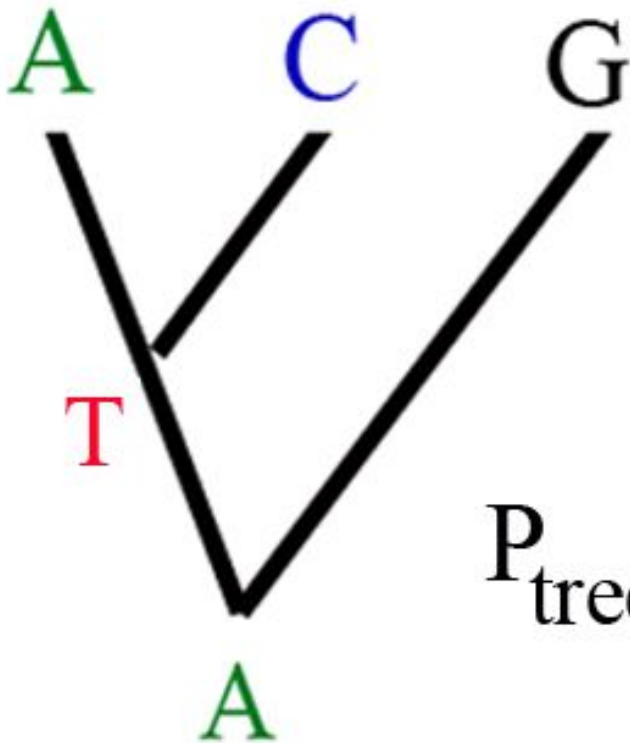
Jukes–Cantor (JC)

$$\mathbf{P}_t = \begin{bmatrix} . & \alpha & \alpha & \alpha \\ \alpha & . & \alpha & \alpha \\ \alpha & \alpha & . & \alpha \\ \alpha & \alpha & \alpha & . \end{bmatrix}, \quad \mathbf{f} = \left[\frac{1}{4} \frac{1}{4} \frac{1}{4} \frac{1}{4} \right]$$

Дерево 1 из 3

А если более сложная модель?

Рассчитываем параметры,
исходя из матрицы данных



$$P_{\text{tree}} = P_A \cdot P_{AT} \cdot P_{TA} \cdot P_{TC} \cdot P_{AG}$$

Как рассчитать эти вероятности (а вернее правдоподобия)?

Обращаемся к моделям нуклеотидных замен

$$\mathbf{P}_t = \begin{bmatrix} p_{AA} & p_{AC} & p_{AG} & p_{AT} \\ p_{CA} & p_{CC} & p_{CG} & p_{CT} \\ p_{GA} & p_{GC} & p_{GG} & p_{GT} \\ p_{TA} & p_{TC} & p_{TG} & p_{TT} \end{bmatrix} \quad \mathbf{f} = [f_A \ f_C \ f_G \ f_T]$$

Где t - это время, P_{AC} -

$$P_{AC} = P_{CA}$$

Используются те же модели, что и для расчета генетических дистанций

JC model

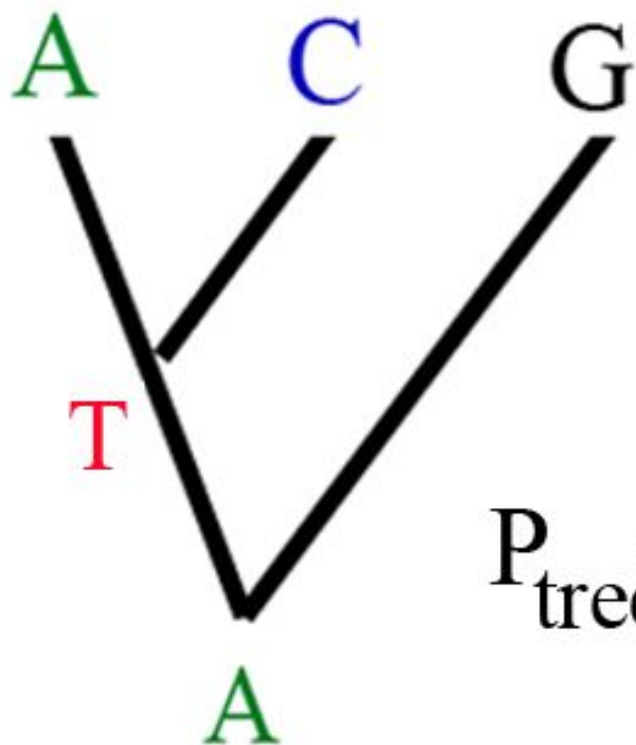
Вероятности всех замен одинаковы,
т.е. $P(AC)=P(AG)=P(AT)=P(CG)=P(CT)=P(GT)=\alpha$

частоты нуклеотидов равны, т.е. $f(A)=f(C)=f(G)=f(T)=0.25$

Jukes–Cantor (JC)

$$\mathbf{P}_t = \begin{bmatrix} . & \alpha & \alpha & \alpha \\ \alpha & . & \alpha & \alpha \\ \alpha & \alpha & . & \alpha \\ \alpha & \alpha & \alpha & . \end{bmatrix}, \quad \mathbf{f} = \left[\frac{1}{4} \ \frac{1}{4} \ \frac{1}{4} \ \frac{1}{4} \right]$$

Дерево 1 из 3



Вероятности всех замен одинаковы,
т.е. $P(AC)=P(AG)=P(AT)=P(CA)=P(CG)=P(CT)=$
 $P(GA)=P(GC)=P(GT)=P(TA)=P(TC)=P(TG)=\alpha$

частоты нуклеотидов равны,
т.е. $f(A)=f(C)=f(G)=f(T)=0.25$
 $P_{xy} = \alpha = 1/12=0,083$

$$P_{\text{tree}} = 0.25 \times \alpha \times \alpha \times \alpha \times \alpha =$$
$$= 0.25 \times 0.083 \times 0.083 \times 0.083 \times 0.083$$
$$= 0.00001186$$

$$P_{\text{tree}} = P_A \cdot P_{AT} \cdot P_{TA} \cdot P_{TC} \cdot P_{AG}$$

K2P

Вероятности транзиций и трансверсий разные,
частоты нуклеотидов равны, т.е. $f(A)=f(C)=f(G)=f(T)=0.25$

Kimura's 2 parameter model (K2P)

α – транзиция

β – трансверсия

Параметры α и β
(т.е. вероятность
транзиций и
трансверсий)
можно оценить,
исходя из данных

$$P_t = \begin{bmatrix} . & \beta & \alpha & \beta \\ \beta & . & \beta & \alpha \\ \alpha & \beta & . & \beta \\ \beta & \alpha & \beta & . \end{bmatrix}, \quad \mathbf{f} = \left[\frac{1}{4} \frac{1}{4} \frac{1}{4} \frac{1}{4} \right].$$

Type of sequences	Transition/transversion ratio (κ)
mtDNA	9.0
12S rRNA	1.75
α - and β -globins	0.66
Pseudo η -globin	2.70

F81

Вероятности всех замен одинаковы, но частоты нуклеотидов разные

Felsenstein (1981)

$$\mathbf{P}_t = \begin{bmatrix} . & \pi_C \alpha & \pi_G \alpha & \pi_T \alpha \\ \pi_A \alpha & . & \pi_G \alpha & \pi_T \alpha \\ \pi_A \alpha & \pi_C \alpha & . & \pi_T \alpha \\ \pi_A \alpha & \pi_C \alpha & \pi_G \alpha & . \end{bmatrix}, \quad \mathbf{f} = [\pi_A \ \pi_C \ \pi_G \ \pi_T]$$

K2P

Вероятности транзиций и трансверсий разные,
частоты нуклеотидов разные

Hasegawa, Kishino and Yano (1985)

$$\mathbf{P}_t = \begin{bmatrix} . & \pi_C \beta & \pi_G \alpha & \pi_T \beta \\ \pi_A \beta & . & \pi_G \beta & \pi_T \alpha \\ \pi_A \alpha & \pi_C \beta & . & \pi_T \beta \\ \pi_A \beta & \pi_C \alpha & \pi_G \beta & . \end{bmatrix}, \quad \mathbf{f} = [\pi_A \ \pi_C \ \pi_G \ \pi_T]$$

General Reversible Model

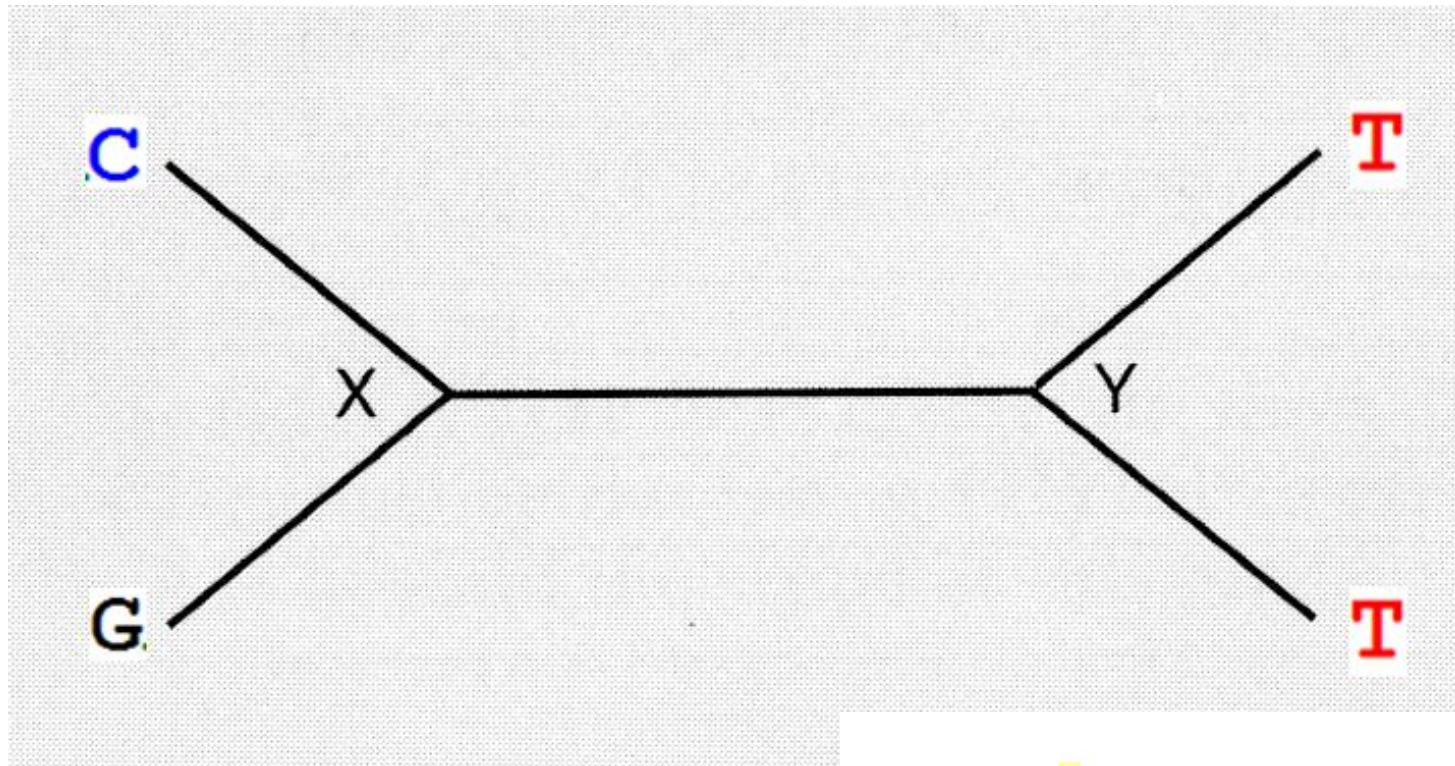
Вероятности ВСЕХ ЗАМЕН разные, т.е. $P(AC)=a$,
 $P(AG)=b$, $P(AT)=c$, $P(CG)=d$, $P(CT)=e$, $P(GT)=f$

частоты нуклеотидов разные
т.е. $f(A)=\pi_1$, $f(C)=\pi_2$, $f(G)=\pi_3$, $f(T)=\pi_4$

General reversible model (REV)

$$\mathbf{P}_t = \begin{bmatrix} . & \pi_C a & \pi_G b & \pi_T c \\ \pi_A a & . & \pi_G d & \pi_T e \\ \pi_A b & \pi_C d & . & \pi_T f \\ \pi_A c & \pi_C e & \pi_G f & . \end{bmatrix}, \quad \mathbf{f} = [\pi_A \ \pi_C \ \pi_G \ \pi_T]$$

Для 4 таксонов возможны 3 варианта неукорененного дерева и
15 вариантов укорененного дерева



Один из них

- 1 ATAACATAAGATTCTGATTATTACCACCATCA
- 2 ATAATATAAGATTCTGATTATTACCACCATCA
- 3 ATAATATAAGATTCTGATTATTACCACCATCA
- 4 ATAAGATAAGATTCTGATTATTACCACCATCA

Возможность использования метода максимального правдоподобия опирается в первую очередь на наличие реалистичных моделей эволюции признаков

- Для **морфологических** признаков, как правило, имеются только **вербальные** (словесные) **модели эволюции**, прописанные в виде эволюционных сценариев, а не формул.
- Количественные параметры этих моделей трудно, фактически невозможно разработать, исходя из имеющихся эмпирических данных
- Но даже если мы создадим модель для одного признака, она не пригодна для других, так как признаки очень разнородны

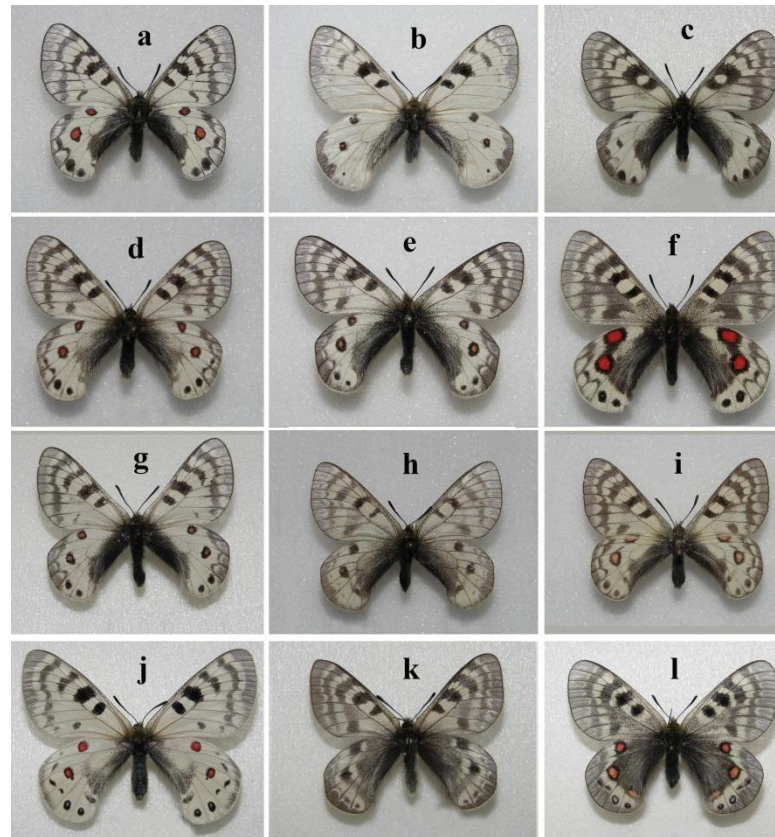
Модели молекулярной эволюции

- легко формализуются в виде формул, так как признаки стереотипны, а из изменения стандартны
- например, модели, описывающие эволюцию нуклеотидных последовательностей:
 - JC (Jukes-Cantor model)
 - K2P (Kimura 2 parameter model)
 - F81 (Felsenstein 1981 model)
 - HKY85 (Hasegawa et al. 1985 model)
 - REV (general reversible model)
 - HKY85 + Γ (Hasegawa et al. 1985 +gamma distribution model)

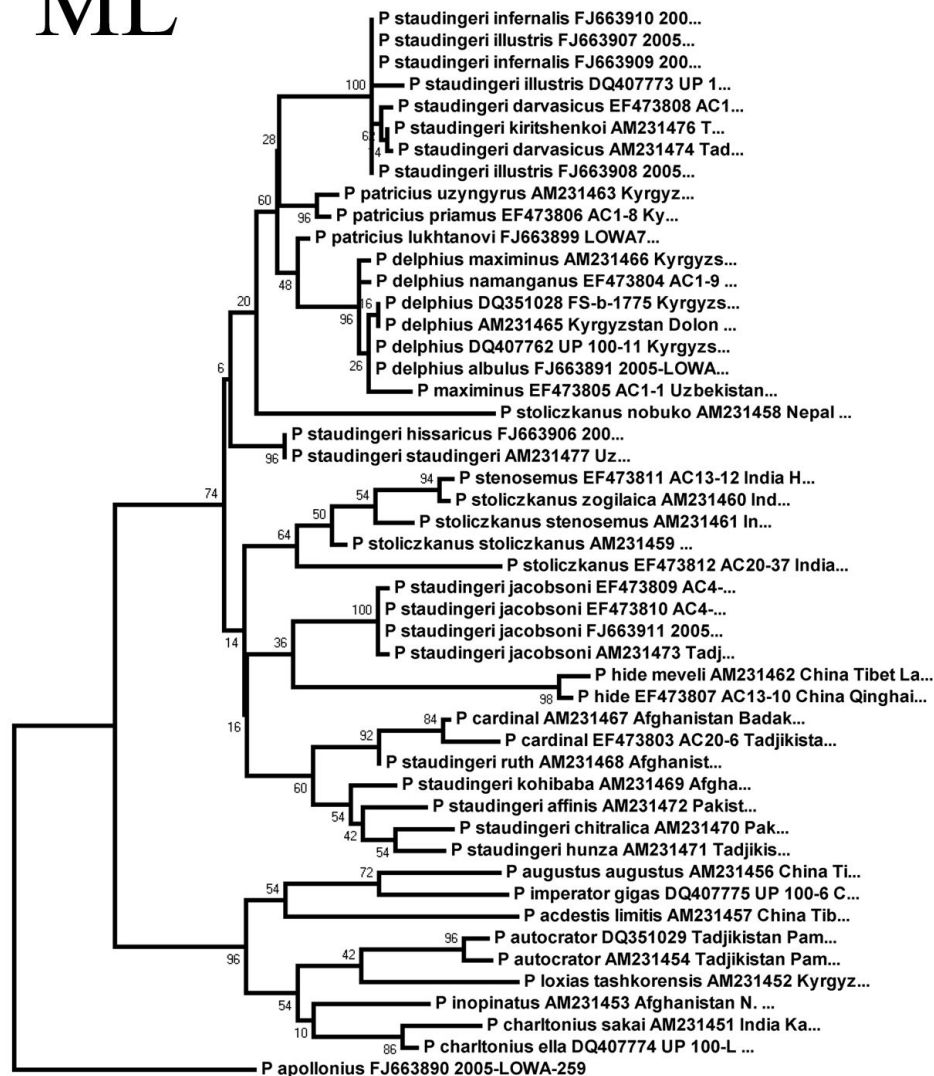
- Аналитический и эвристические методы построения дерева максимального правдоподобия
- Бутстреп

Пример

Филогения бабочек рода *Parnassius*, основанная на анализе гена COI с использованием метода максимального правдоподобия

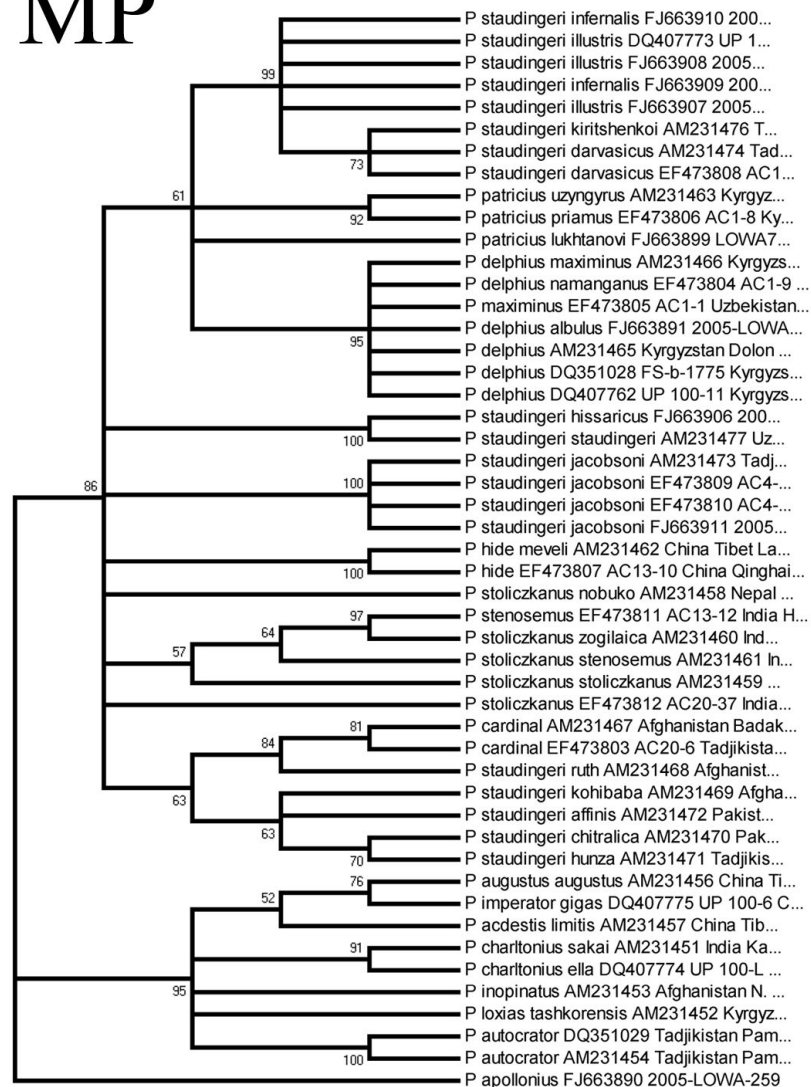


ML



0.01

MP



Соотношение парсимонии и максимального правдоподобия

Преимущества метода максимального правдоподобия:

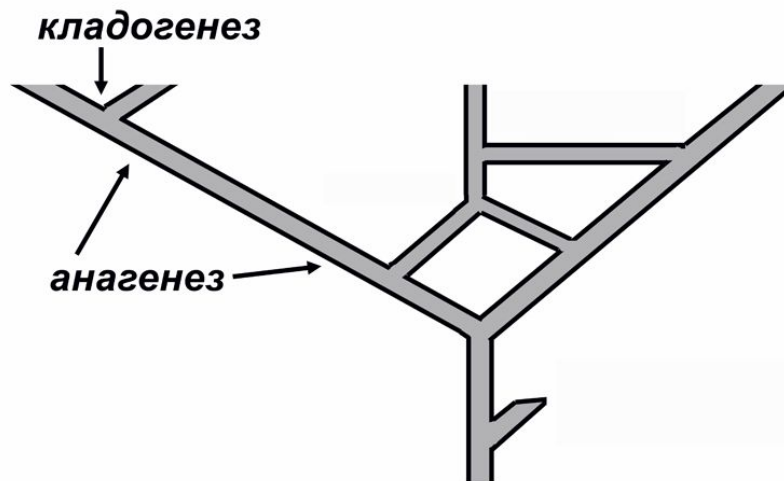
- Теоретически более состоятелен, так как не ограничен в выборе модели эволюции
 - 1) не нуждается в теоретически и практически несостоятельном принципе парсимонии в качестве критерия правильной реконструкции
 -

Преимущество метода максимального правдоподобия:

- 2) возможность использования гораздо большего числа признаков
 - не только синапоморфий, но и аутапоморфий (на самом деле еще и плезиоморфий [роль инвариантных сайтов] ! – эволюционные филогенетики должны возрадоваться -
- что дает принципиальную возможность разрешения большего числа узлов ветвления филогенетического дерева

Преимущества метода максимального правдоподобия:

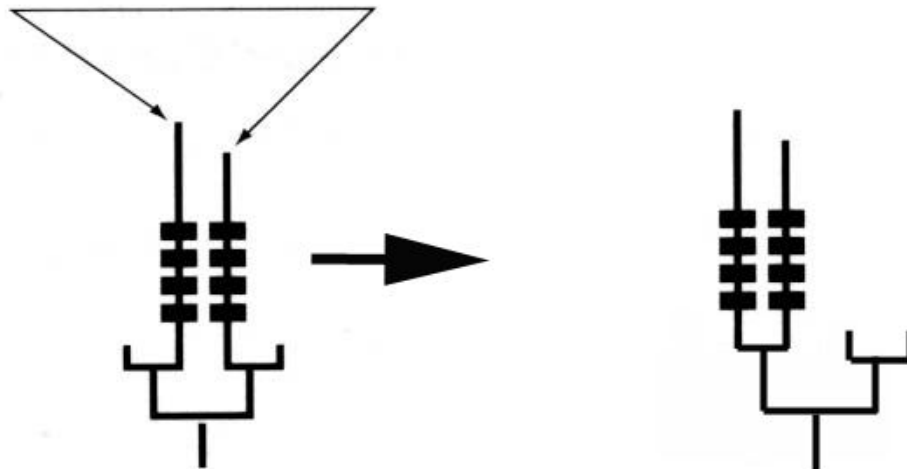
- 3) дает более адекватное представление об анагенетической составляющей эволюции



4) Менее чувствителен к эффекту длинных ветвей

Small tree

Длинные ветви имеют
большую вероятность, что в них независимо появятся
одинаковые признаки (гомоплазии)



ошибочная интерпретация их
в качестве синапоморфий
ведет к ложной реконструкции

Недостатки

- Ошибка в выборе модели может быть фатальна, т.е. иногда лучше упрощенная модель, чем более совершенная, но явно ошибочная

Методы укоренения деревьев

- По внешней группе
 - Принципы выбора внешней группы
- По средней точке – чтобы расстояние от общего предка до конца ветвей было одинаковым (основан на принципе молекулярных часов)

- По внешней группе
 - Принципы выбора внешней группы
 - Внешняя точка должна быть заведомо внешней

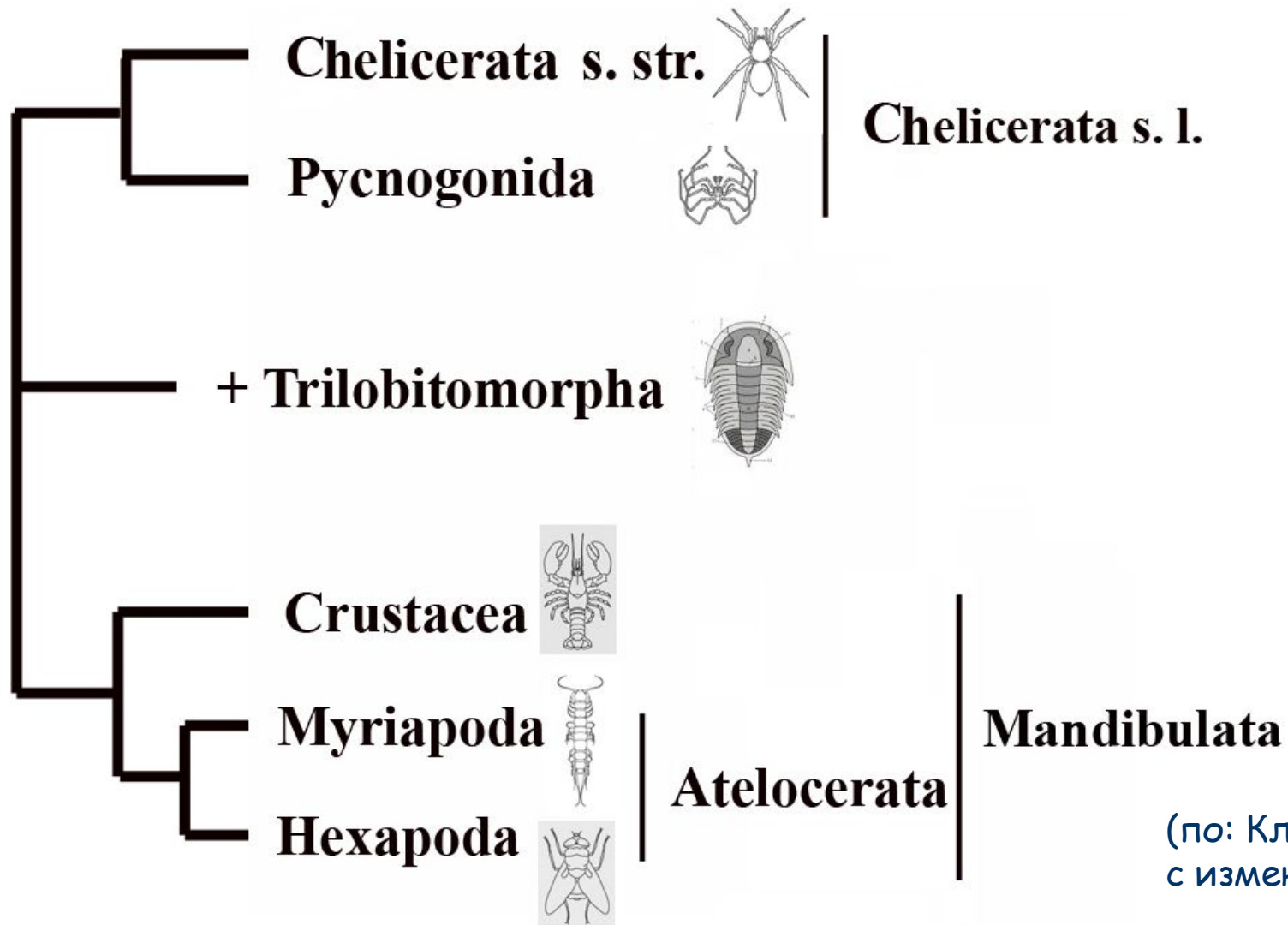
- По внешней группе
 - Принципы выбора внешней группы
 - Внешняя точка должна быть заведомо внешней
 - Но желательно не слишком далекой (т.е. максимально близкая, но заведомо внешняя)

- По внешней группе
 - Принципы выбора внешней группы
 - Внешняя группа должна быть заведомо внешней
 - Но желательно не слишком далекой (т.е. максимально близкая, но заведомо внешняя)
 - Внешняя группа желательно должна быть множественной

- По внешней группе
 - Принципы выбора внешней группы
 - Внешняя группа должна быть заведомо внешней
 - Но желательно не слишком далекой (т.е. максимально близкая, но заведомо внешняя)
 - Внешняя группа желательно должна быть множественной
 - Внешняя группа не должна быть полифилетической

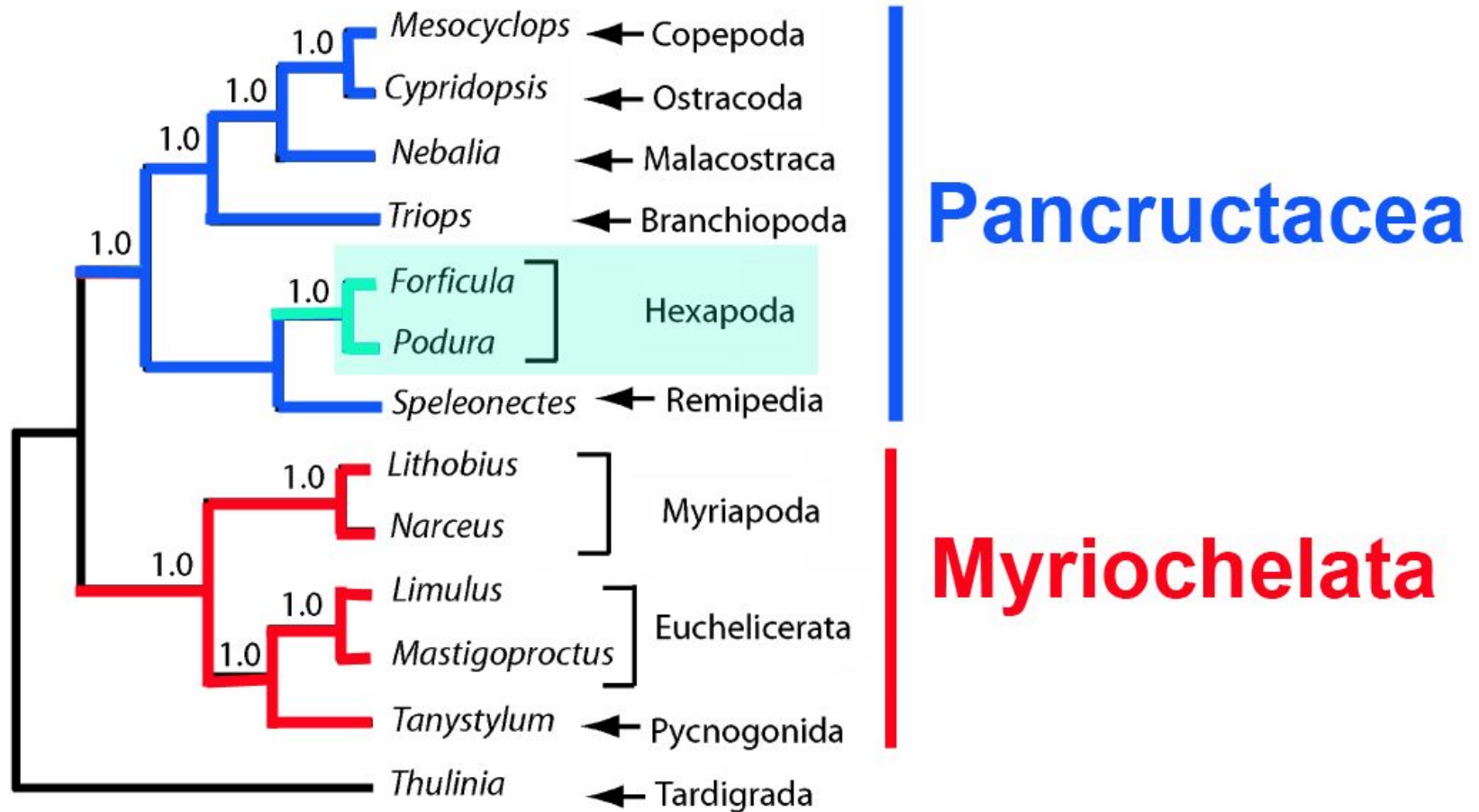
Представление о филогении членистоногих, которое недавно считалось классическим:

насекомые (Hexapoda) и многоножки (Myriapoda) – сестринские группы



(по: Ключе, 2000, с изменениями)

формальный анализ всей совокупности молекулярных признаков (62 гена, 41000 пар нуклеотидов) поддерживает как Pancrustacea, так и Myriochelata



Regier et al., 2008. Resolving Arthropod phylogeny: Exploring phylogenetic signal within 41 kb of protein-coding nuclear gene sequence. *Syst.biol.* 57:920-938

Методы укоренения деревьев

- По средней точке – чтобы расстояние от общего предка до конца ветвей было одинаковым (основан на принципе молекулярных часов)

- Метод ML основан на оптимизации соответствия выбранной модели и наблюдаемых данных, НО
- Пример с гномами