

Лекция 7

Байесова филогенетика

Проблема конкурирующих гипотез и метод проб и ошибок

- Примеры гипотез:
 - Встречу ли я динозавра, выйдя на улицу?
 - Гипотезы: встречу – не встречу
 - 50% и 50% ???????

Проблема конкурирующих гипотез и метод проб и ошибок

- Примеры гипотез:
 - Встречу ли я динозавра, выйдя на улицу?
 - Гипотезы: встречу – не встречу
 - 50% и 50% ???????

Проверка гипотез при помощи эмпирических испытаний
позволяет изменить первичную оценку вероятности гипотез

гипотеза H_1

10 черных шаров
30 белых шаров

гипотеза H_2

20 черных шаров
20 белых шаров

Если ящики одинаковы и закрыты,
то вероятность угадать, где какой $P = 0.5$

Если мы вынули из ящика 21 белый шар, то это точно гипотеза H_1

**Но не обязательно вынимать 21 белый шар
и тем более все шары:**

**Можно вынимать по одному, и сам факт преобладания
белых шаров постепенно повышает вероятность H_1**

Проблема конкурирующих гипотез и метод проб и ошибок

- Примеры гипотез:
 - Филогенетическая реконструкция: топология 1, топология 2
... топология_n
 - Каждый вариант – гипотеза. Какую выбрать?

Проблема конкурирующих гипотез

- Решения:
 - MP: выбираем наиболее простую гипотезу
 - ML: выбираем наиболее правдоподобную гипотезу
 - NO:
 - 1) за бортом остаются все другие гипотезы (слишком упрощенное решение)
 - Хорошо бы оценивать вероятность “лучшей” гипотезы в процентах
 - А еще лучше иметь совокупность всех гипотез с оценками их вероятностей

Проблема конкурирующих гипотез

- А еще лучше иметь совокупность всех гипотез с прямыми оценками их вероятностей
- Есть ли такой метод? Да! Байесова статистика!
- Она основана на выдвижении предварительных (априорных) гипотез и их испытании методом взятия проб. После взятия пробы можно рассчитать вероятность постериорной гипотезы

Метод Байеса (Bayes Inference)



Thomas Bayes
1702-1761 England

Байесова статистика .

Обычная статистика рассматривает вероятности (частоты статистических распределений) как константные величины.

Байесова статистика рассматривает вероятности (частоты статистических распределений) как предварительные гипотезы (priors), которые могут быть уточнены в ходе анализа.

Метод Байеса (Bayes Inference)

ОСНОВНЫЕ ПОНЯТИЯ:

- Априорная вероятность гипотезы
- Постериорная вероятность гипотезы
- правдоподобие гипотезы (вероятность наблюдения данных при условии, что гипотеза верна)

- Априорные и постериорные гипотезы

- Схема анализа:
 - 1) выбираются (задаются) априорные гипотезы (вероятности)
 - 2) получение данных (эмпирическое тестирование)
 - 3) на основании проведенных испытаний рассчитываются постериорные гипотезы (вероятности)

Тестирование двух гипотез - H_1 и H_2

$$P(H_1|E) = \frac{P(E|H_1) P(H_1)}{P(E|H_1) P(H_1) + P(E|H_2) P(H_2)}$$

H_1 – гипотеза 1

H_2 – гипотеза 2

E - испытание

$P(H_1|E)$ – постериорная вероятность гипотезы H_1

(после получения данных E , т.е. после проведенного испытания E)

$P(H_1)$ – априорная вероятность гипотезы H_1

$P(E|H_1)$ – вероятность наблюдения данных при условии, что гипотеза H_1 верна (=правдоподобие гипотезы)

В числителе $P(E|H_1) P(H_1)$ – произведение вероятности наблюдения данных на априорную вероятность данной гипотезы

В знаменателе – сумма произведений $P(E|H_1) P(H_1)$ для каждой из альтернативных гипотез (H_1, H_2 и т.д.)

гипотеза H_1

10 черных шаров
30 белых шаров

гипотеза H_2

20 черных шаров
20 белых шаров

Если ящики одинаковы и закрыты,
то вероятность угадать, где какой $P = 0.5$

P априорное для $H_1 = 0.5$

P априорное для $H_2 = 0.5$

гипотеза H_1

10 черных шаров 30 белых шаров

гипотеза H_2

20 черных шаров 20 белых шаров

Если ящики одинаковы и закрыты,
то вероятность угадать, где какой $P = 0.5$

Правдоподобие для $H_1 = 0.75$ (вероятность, что первый вынутый шар будет белым)

Правдоподобие для $H_2 = 0.5$ (вероятность, что первый вынутый шар будет белым)

$$\begin{aligned}
 P(H_1|E) &= \frac{P(E|H_1) P(H_1)}{P(E|H_1) P(H_1) + P(E|H_2) P(H_2)} \\
 &= \frac{0.75 \times 0.5}{0.75 \times 0.5 + 0.5 \times 0.5} \\
 &= 0.6
 \end{aligned}$$

$P(H_1)$ - априорная вероятность гипотезы H_1

$P(H_1/E)$ - постериорная вероятность гипотезы H_1

$P(E/H_1)$ - вероятность наблюдения данных при условии, что гипотеза H_1 верна (=правдоподобие гипотезы)

В числителе $P(E/H_1) P(H_1)$ - произведение вероятности наблюдения данных на априорную вероятность данной гипотезы

В знаменателе - сумма произведений $P(E/H_1) P(H_1)$ для каждой из альтернативных гипотез (H_1, H_2 и т.д.)

- Итеративная процедура – многократное возвращение к тестированию исходной гипотезы, но каждый раз с учетом уже измененной априорной вероятности

Вторая итерация

априорные вероятности гипотез уже другие

$$P(H_1)=0.6; P(H_2)=0.4$$

$$P(H_1|E) = \frac{P(E|H_1) P(H_1)}{P(E|H_1) P(H_1) + P(E|H_2) P(H_2)}$$

- $P = (0.6 \times 0.75)/(0.6 \times 0.75 + 0.4 \times 0.5) = 0.45/(0.45 + 0.2) = 0.69$

Третья итерация:
априорные вероятности снова изменились

$$P(H_1)=0.69; P(H_2)=0.31$$

$$P(H_1|E) = \frac{P(E|H_1) P(H_1)}{P(E|H_1) P(H_1) + P(E|H_2) P(H_2)}$$

- $P = (0.69 \times 0.75)/(0.69 \times 0.75 + 0.31 \times 0.5) = 0.5175/(0.5175 + 0.155) = 0.5175/0.6725 = 0.77$

Продолжаем процесс до тех пор пока вероятность одной из гипотез не достигнет 100% [$P(H_1)=1$], т.е. гипотеза доказана
(или до стационарного уровня, когда вероятность гипотез стабилизируется)

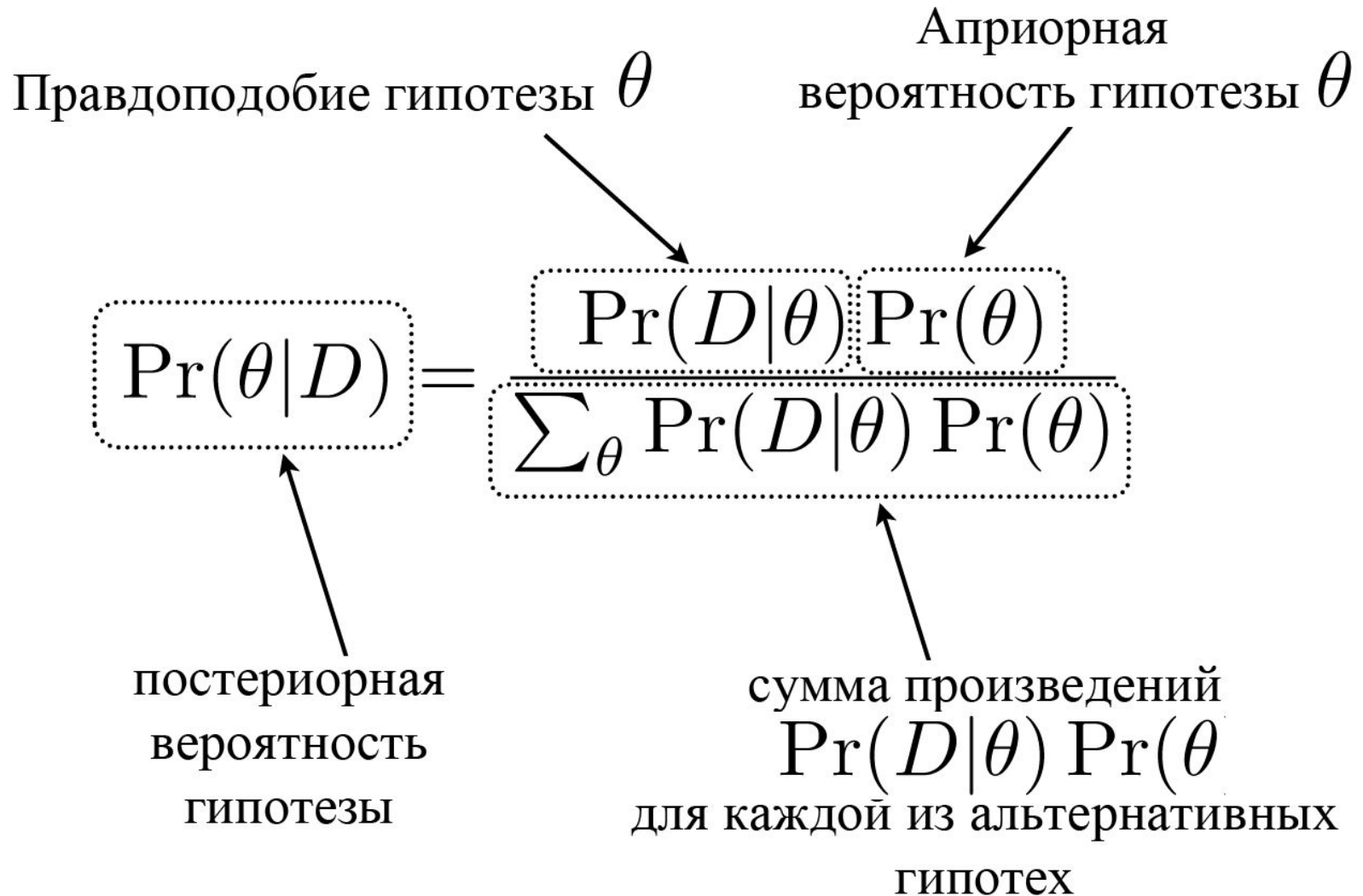
$$P(H_1|E) = \frac{P(E|H_1) P(H_1)}{P(E|H_1) P(H_1) + P(E|H_2) P(H_2)}$$

Photo # NH 70305 USS Scorpion comes alongside USS Tallahatchie County, April 1968



- Лодка затонула 21 мая [1968 года](#) Лодка затонула 21 мая 1968 года в 740 км (400 миль) к юго-западу от [Азорских островов](#) Лодка затонула 21 мая 1968 года в 740 км (400 миль) к юго-западу от Азорских островов [\[1\]](#) Лодка затонула 21 мая 1968 года в 740 км (400 миль) к юго-западу от Азорских островов [\[1\]](#) на глубине в 3000 м (9800 футов), за 5 дней до возвращения на базу в [Норфолк](#). Официально о потере USS Scorpion (SSN-589) было объявлено 5 июня 1968

D данные



- Как все это перенести на реконструкцию филогении?
- - нужны предварительные гипотезы
- - нужны значения правдоподобий

Метод максимального правдоподобия

JC model

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

частоты нуклеотидов равны,

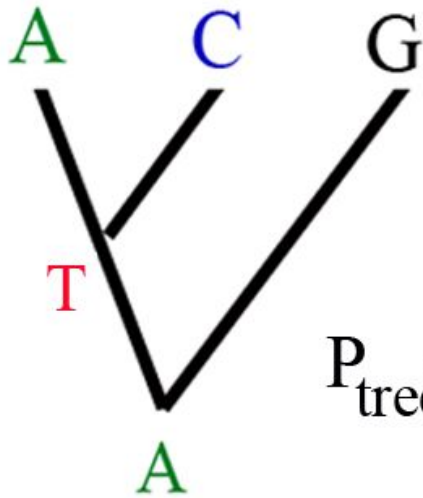
т.е. $f(A)=f(C)=f(G)=f(T)=0.25$

$P_{xy} = \alpha = 1/16=0,0625$

$P_{tree} = 0.25 \times \alpha \times \alpha \times \alpha =$

$= 0.25 \times 0.0625 \times 0.0625 \times 0.0625 \times 0.0625$

$= 0.00000381$



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

Правдоподобие гипотезы θ

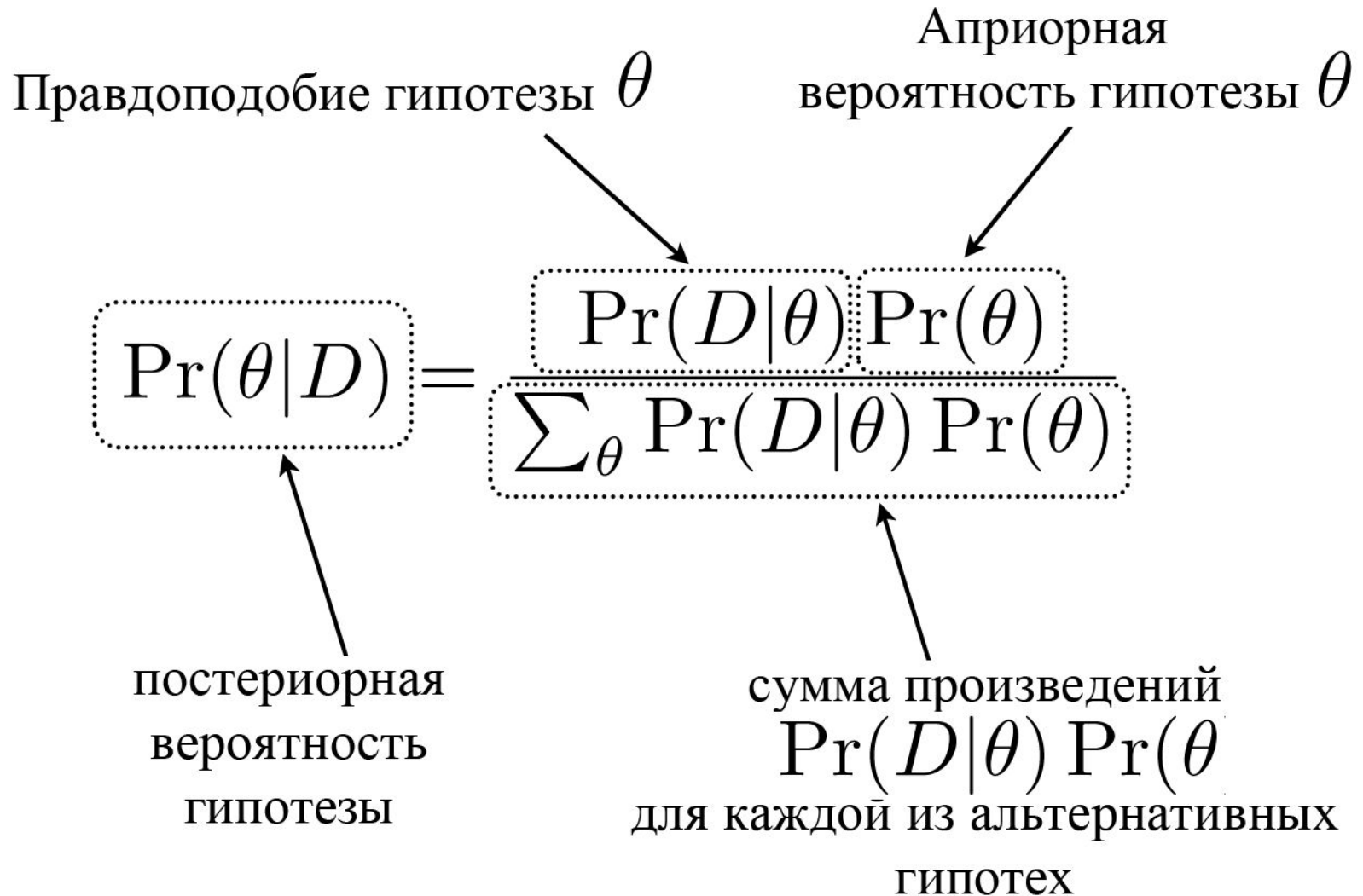
$$\Pr(D|\theta)$$

это вероятность

получения наблюдений D

при условии выполнения гипотезы θ

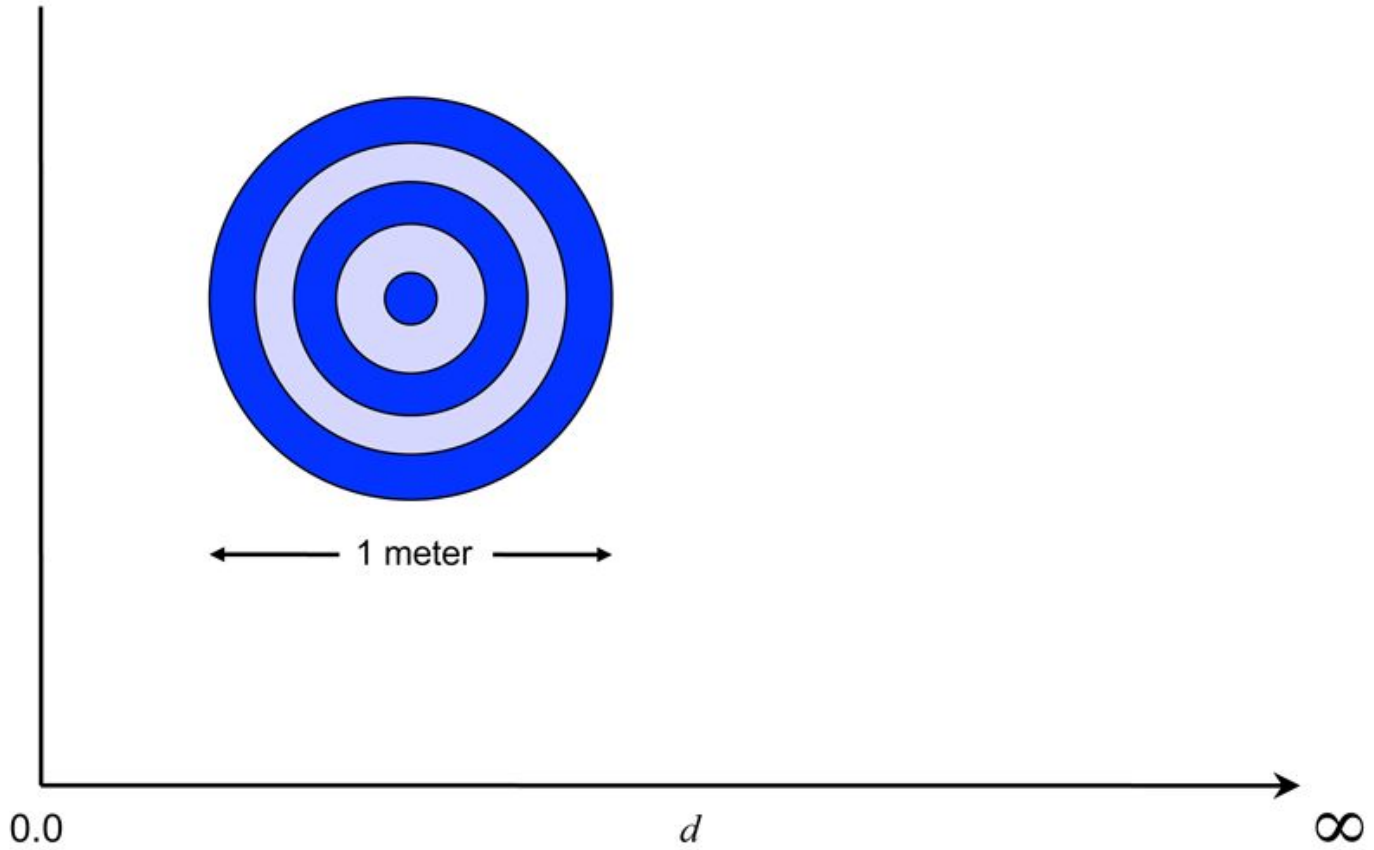
D данные



- Теперь вопрос, как перейти к филогенетическим гипотезам, т.е. деревьям

- В филогенетике эволюционные модели составляют очень большое число гипотез: (каждая уникальная комбинация дерева [топологии] и параметров может быть представлена в виде отдельной гипотезы
- Как использовать Байесову статистику, когда гипотезы составляют непрерывный ряд (континуум)?

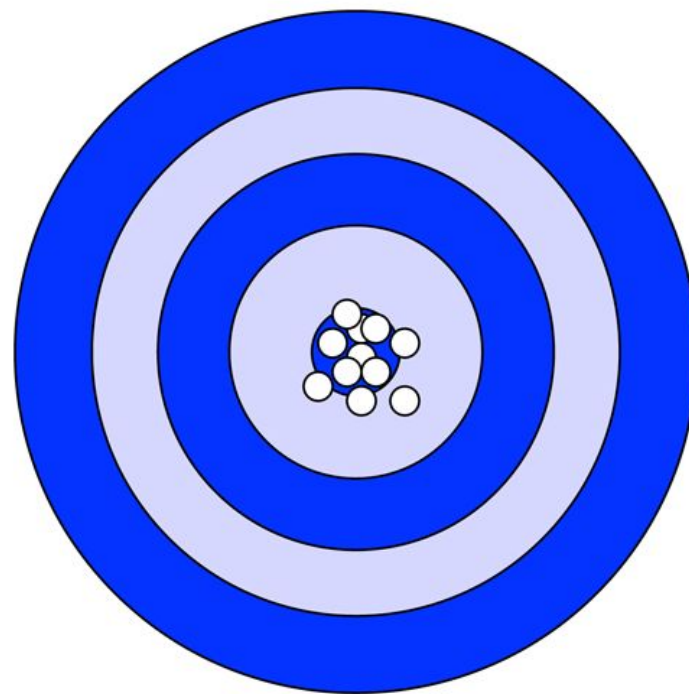
- анализировать не отдельные гипотезы (их может быть неограниченное множество), а статистические распределения этих гипотез



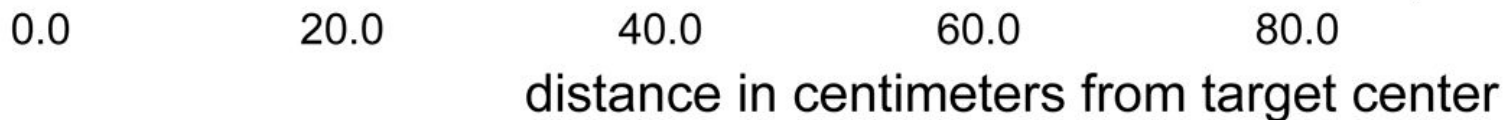
Априорная гипотеза 1

Я - меткий стрелок

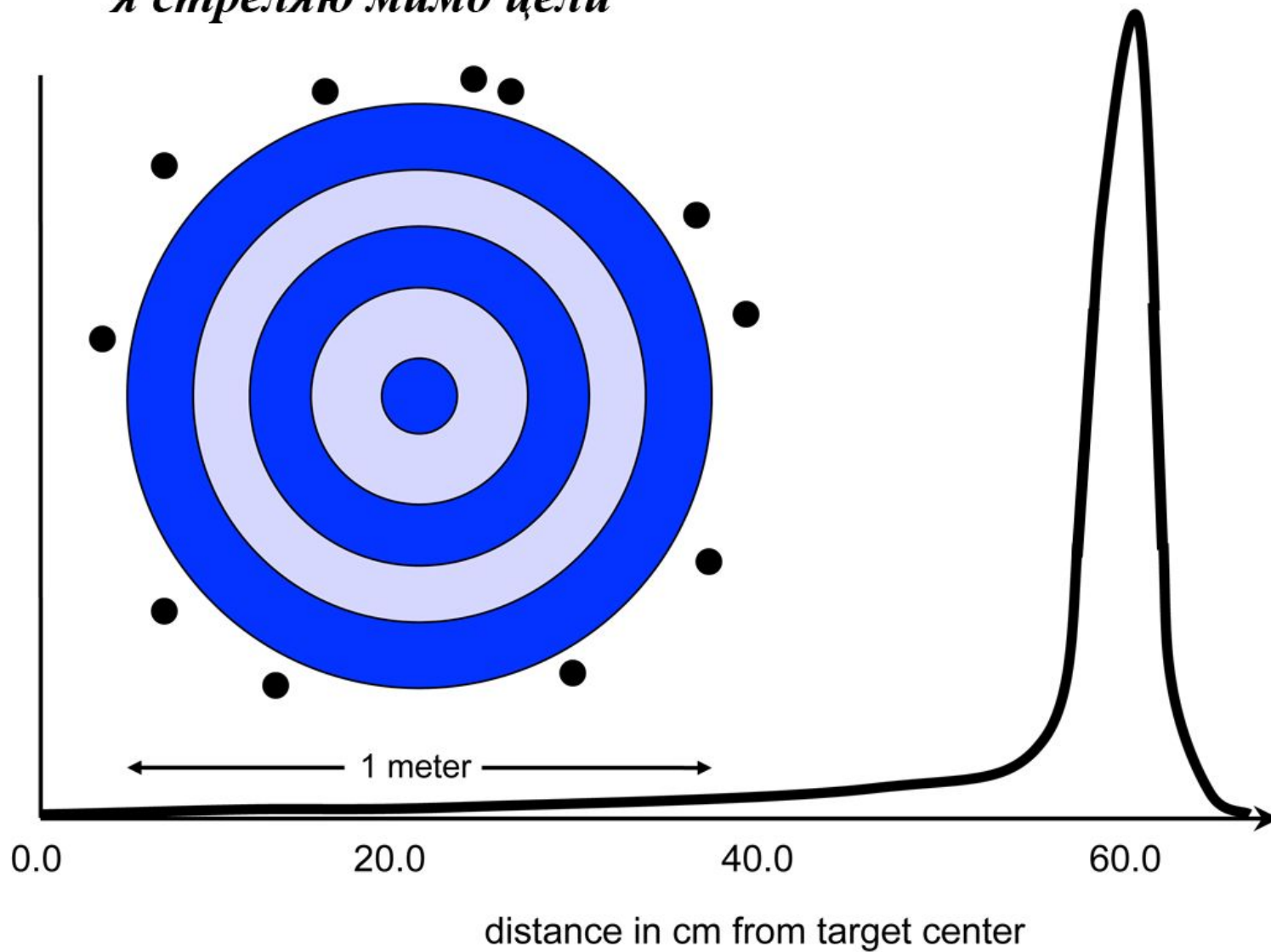
An *informative* prior
(low dispersion =
low variance) that
says most of my
arrows will fall within
10 cm of the center
(thanks for your
confidence!)



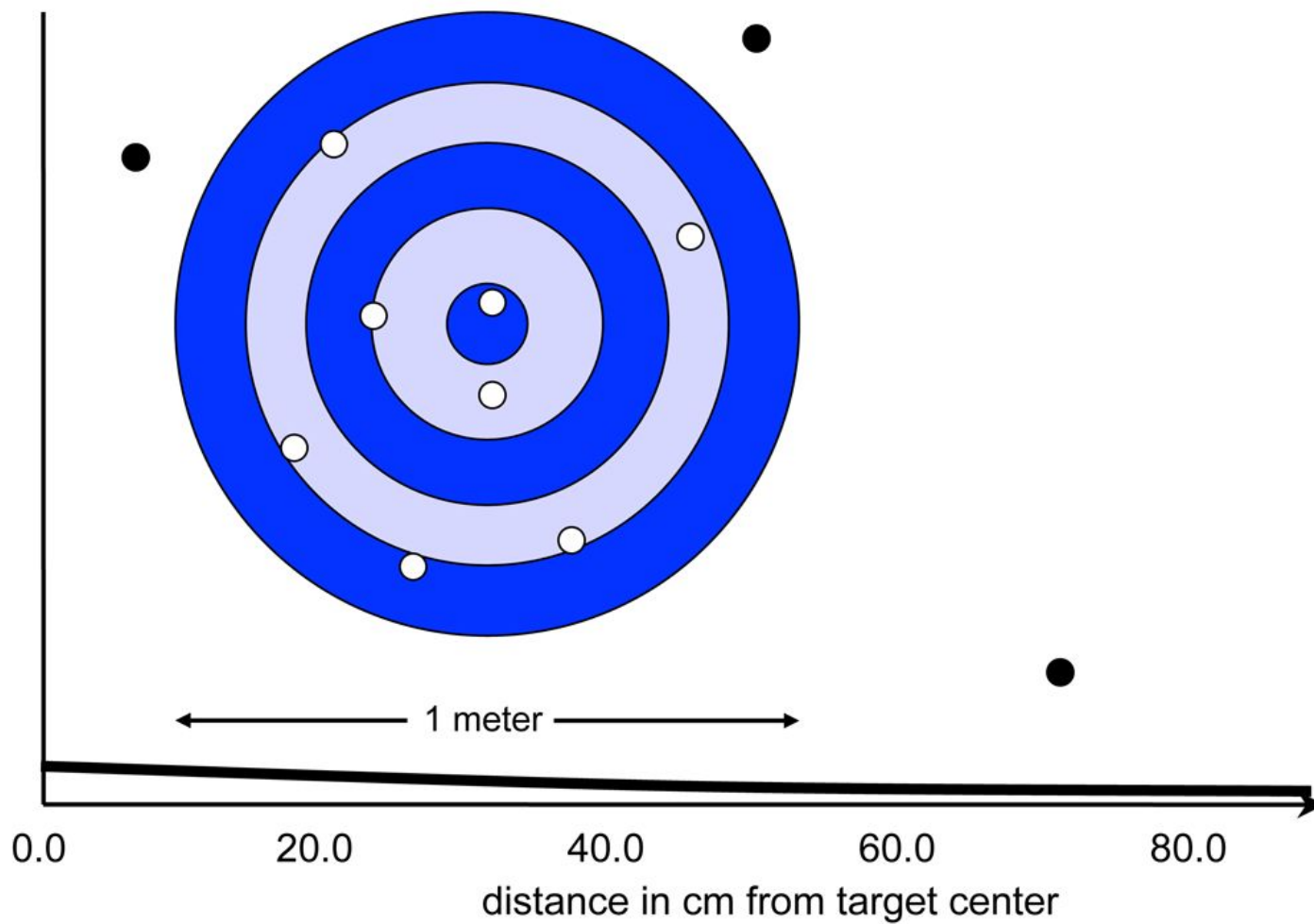
← 1 meter →



*Априорная гипотеза 2:
я стреляю мимо цели*

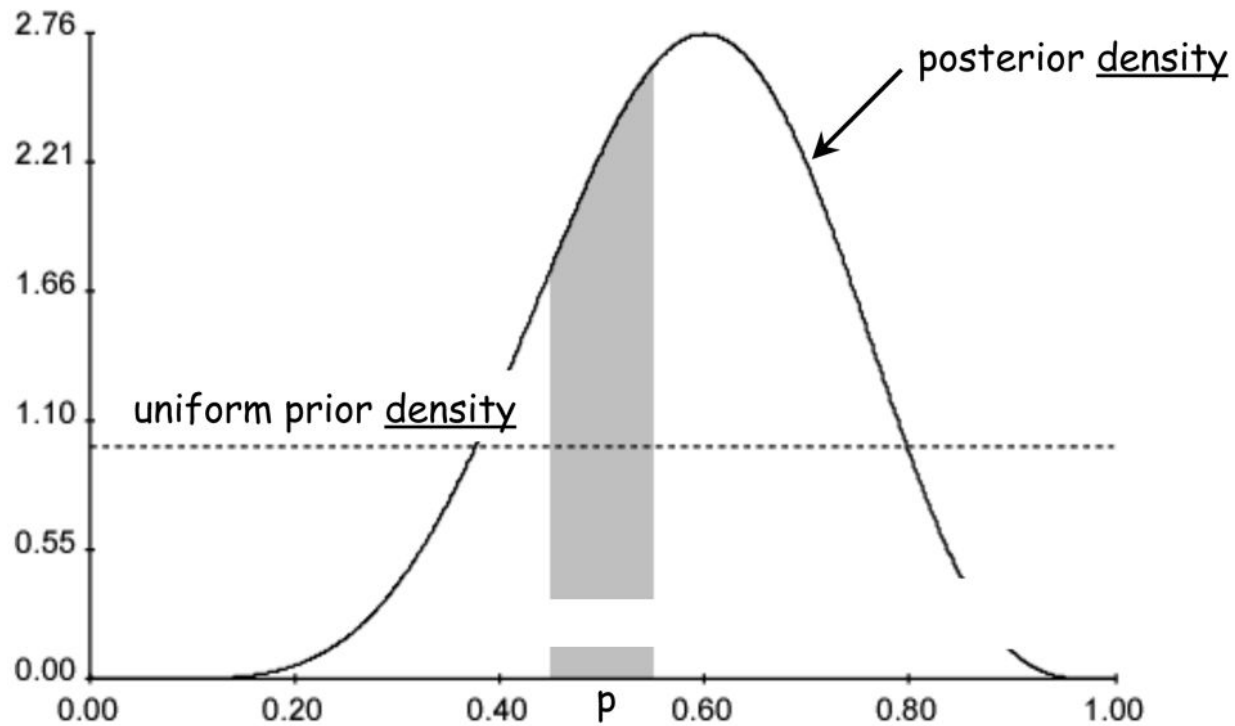


*Априорная гипотеза 3
нет никаких особых идей*



- Униформный (неспецифический прайор), казалось бы, какая от него польза.
- Но вспомним про итеративность... Итерации постепенно сдвигают распределение к более информативному

*Постериорные распределения вероятностей, как правило,
более информативны*



- Еще один прием: расчленить тестируемую гипотезу: представить ее в виде совокупности более простых гипотез

- В случае филогенетической гипотезы вместо дерева можно дать совокупность:
 - 1) топология
 - 2) информация о длине ветвей
 - 3) частоты нуклеотидов
 - 4) вероятности нуклеотидных замен разного типа
 - 5) распределение вероятности замен по длине нуклеотидного выравнивания (параметр гамма)
 - 6) доля инвариантных сайтов

(1) и (2) - параметры самого дерева

(3-6) - параметры ассоциированные с деревом

То есть априорную гипотезу о распределении деревьев можно представить в виде совокупности 6 более простых априорных гипотез (прайоров):

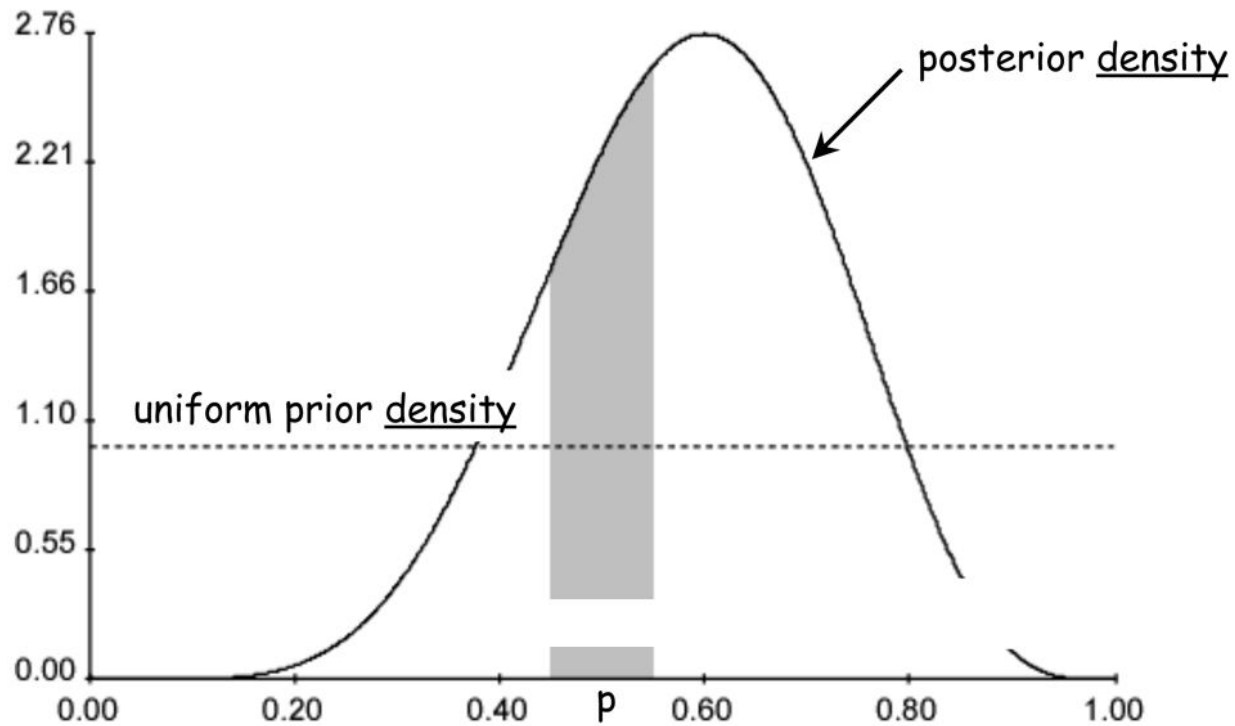
- 1) прайор о топологии
- 2) прайор о длине ветвей
- 3) прайор о частотах нуклеотидов
- 4) прайор о вероятности нуклеотидных замен разного типа
- 5) прайор о распределении вероятности замен по длине нуклеотидного выравнивания (параметр γ)
- 6) прайор о доле инвариантных сайтов

Как рассчитать эти прайоры?

(3-6) мы можем взять прямо из матрицы данных

Для (1) и (2) можно использовать равномерные (неспецифические прайоры)

*Постериорные распределения вероятностей, как правило,
более информативны*

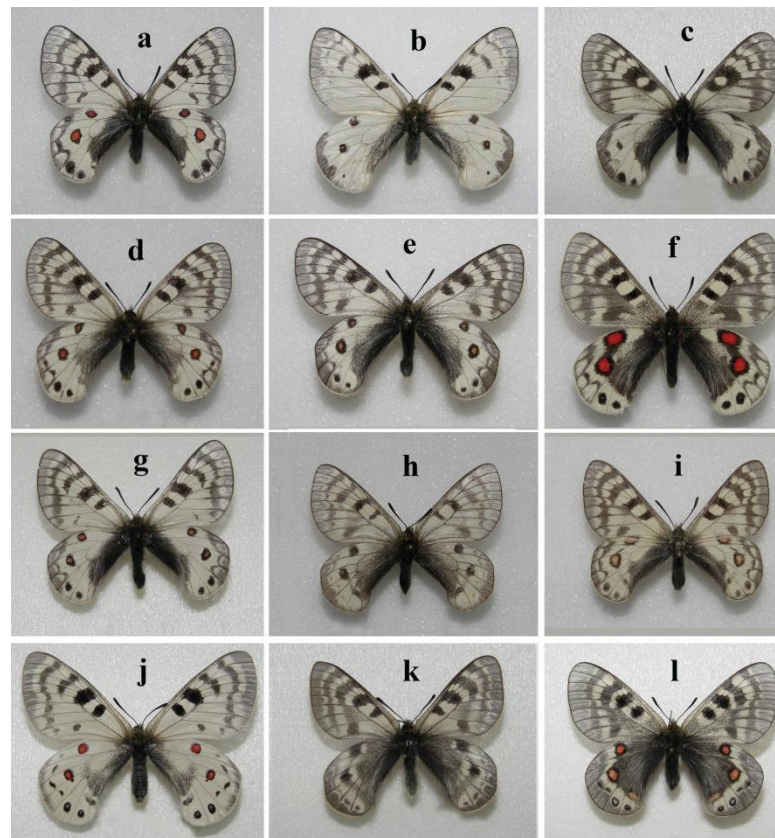


- При проведении анализа запускается несколько цепей (обычно 4), каждая из которых ищет оптимальные деревья
- Цепи могут обмениваться информацией, что позволяет "проскакивать" локальные оптимумы
- Получаемые деревья сравниваются и рассчитываются стандартные отклонения в положении ветвей. Анализ заканчивают, когда уровень этих отклонений стабилизируется.

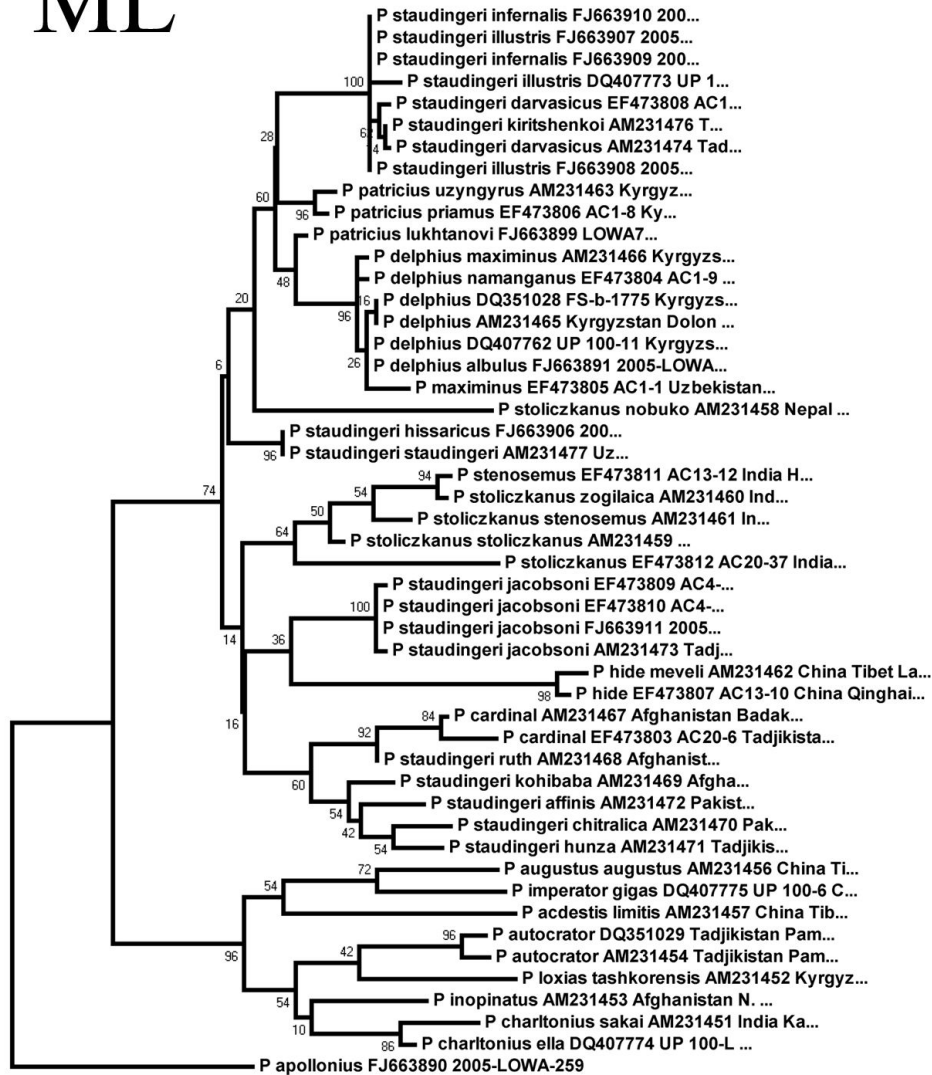
- Как задать прайоры в Байесовом анализе?
- Как выбрать модель эволюции в Байесовом анализе?
- GTR+I+G

Пример

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

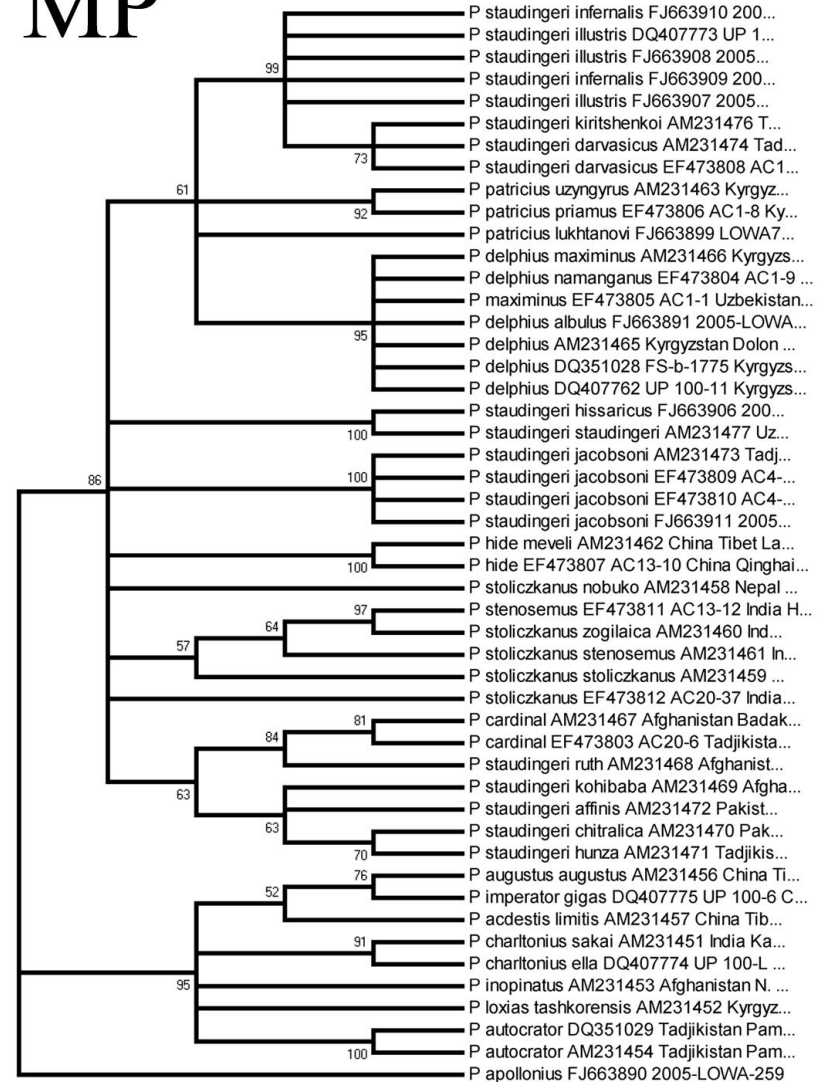


ML

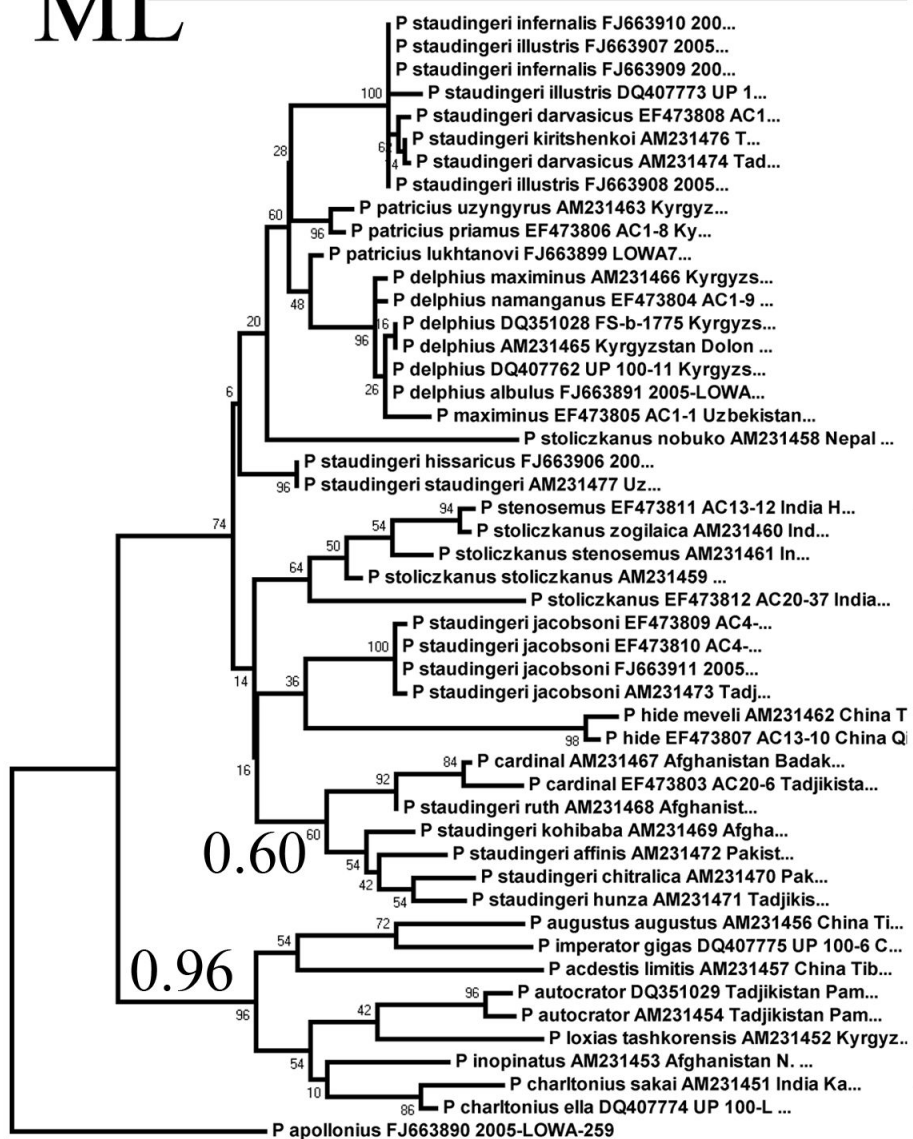


0.01

MP

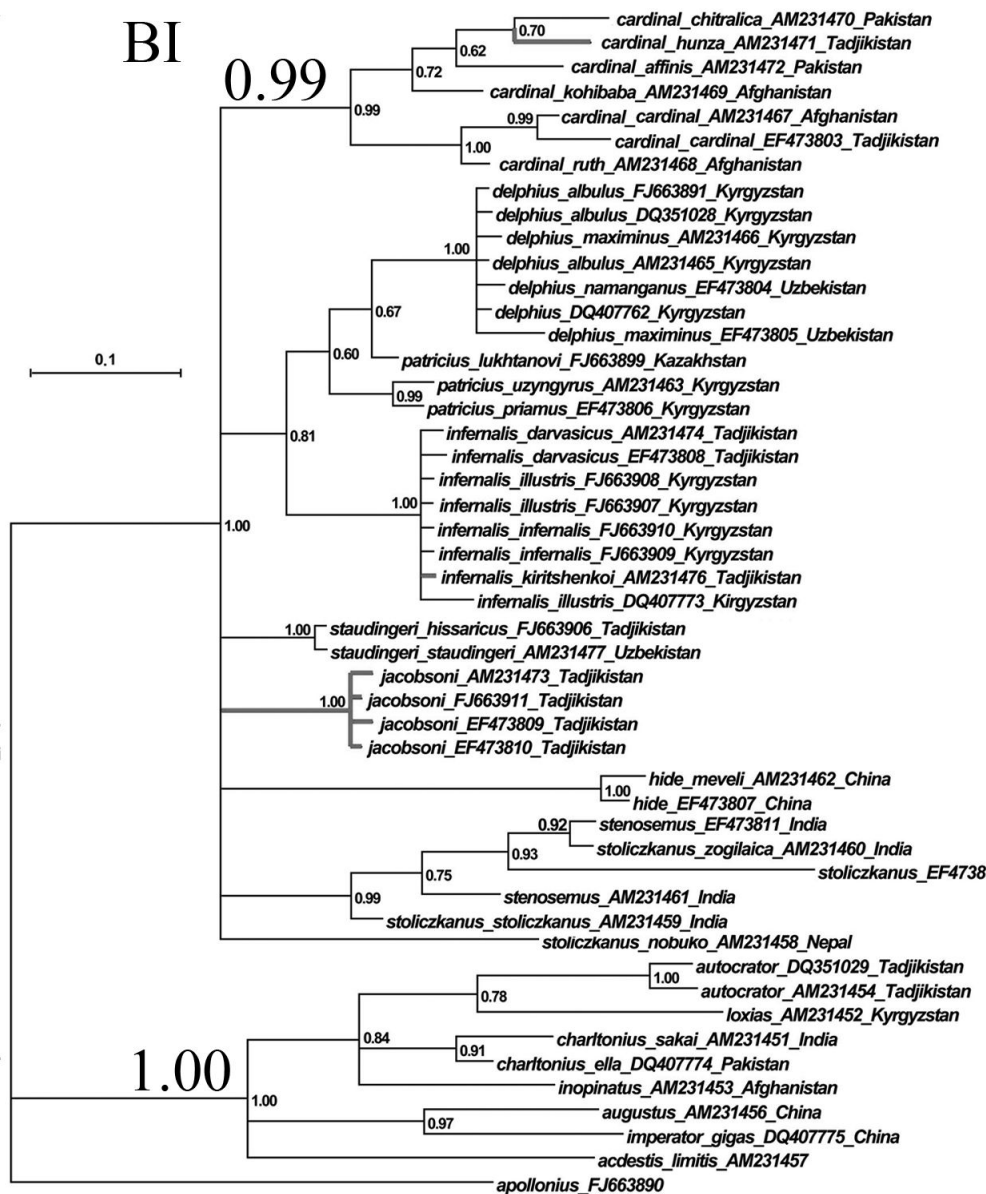


ML



0.01

BI



0.1

1.00

НО

- Основан на другой статистике, которая позволяет, получив вероятность дерева, пересчитать ее с учетом той топологии, которая исходно была неизвестна
- Дает множество деревьев, а не одно

Получаемые в ходе Байесова анализа деревья образуют распределение, которое позволяет рассчитать так называемую постериорную вероятность отдельных деревьев и клад (posterior probability)

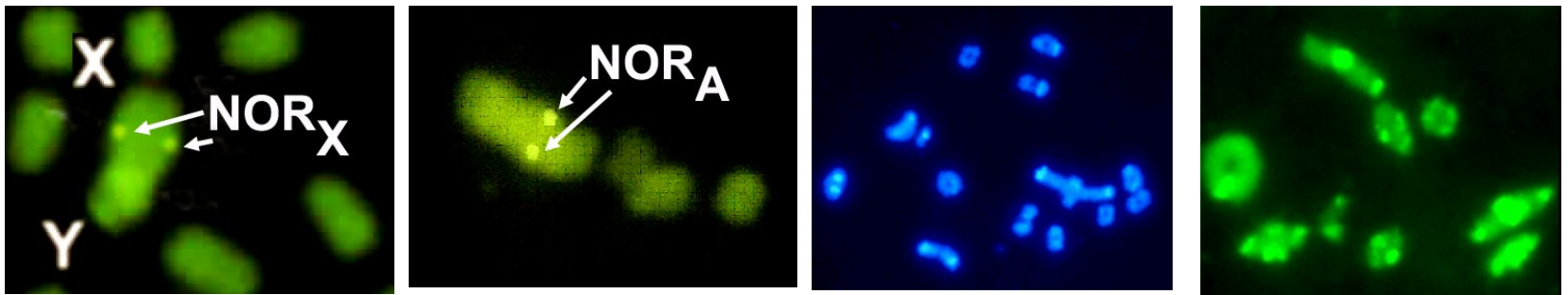
- Распределение этих деревьев позволяет рассчитать так называемую апостериорную вероятность (РВ), которая является прямой оценкой вероятности филогенетической реконструкции - поэтому не нужен бутстреп!

Методы максимального правдоподобия и Байеса: сходство и различия, плюсы и минусы

ML говорит лишь о степени соответствия данных и модели, но не говорит о достоверности тестируемой гипотезы (пример с гномами)

MB пытается заглянуть внутрь черного ящика.
Оценка вероятности самой гипотезы

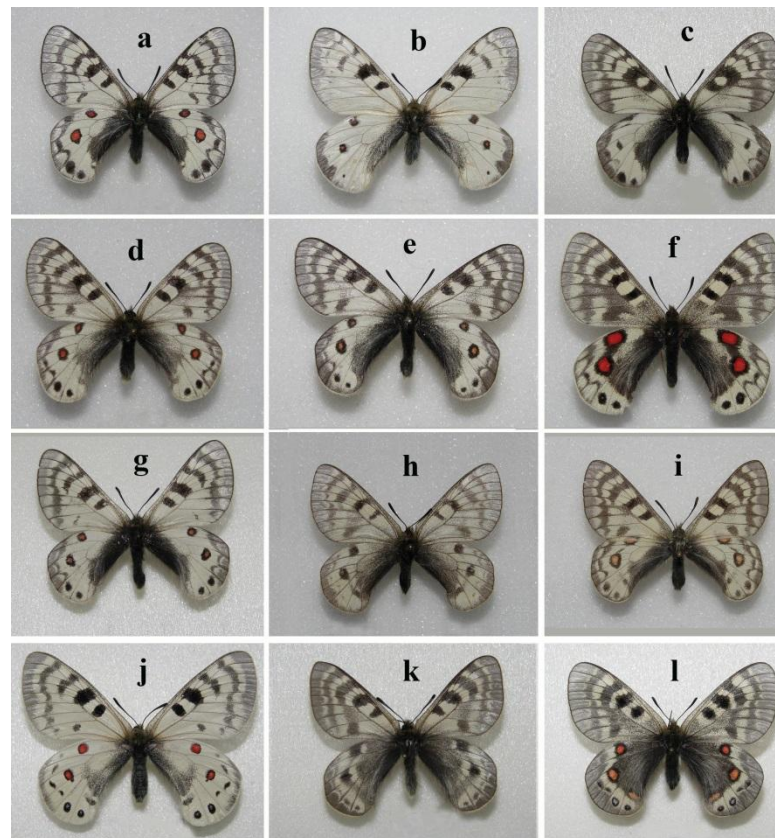
- Методы максимального правдоподобия и Байеса применимы для анализа любых структур, закономерности эволюции которых могут быть формализованы в виде параметрической модели
- Например, для филогенетического анализа хромосомных перестроек



- Не существует никакого теоретического запрета на использование морфологических признаков в рамках метода максимального правдоподобия и Байесова метода
- Однако здесь возникает проблема отсутствия приемлемых моделей морфологической эволюции

Пример

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



Методы реконструкции филогенезов , основанные на анализе генетических дистанций

ДНК:

1 5 10
tagcaaaatg

- Суть метода
- Откуда берутся генетические дистанции?
 - ДНК-ДНК гибридизация, иммунологические реакции, анализ анонимных маркеров - все, что исходно дает информацию в виде % сходства
 - Превращение дискретных данных в генетические дистанции

Преобразование матрицы дискретных данных в матрицу дистанций

sequences

sites

	1	2	3	4	5	6	7
1	T	T	A	T	T	A	A
2	A	A	T	T	T	A	A
3	A	A	A	A	A	T	A
4	A	A	A	A	A	A	T

sequences

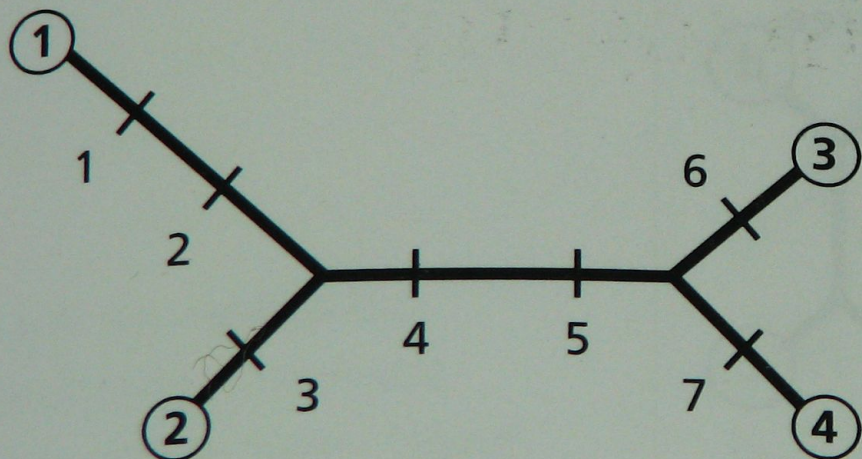
2	3		
3	5	4	
4	5	4	2
	1	2	3

sequences

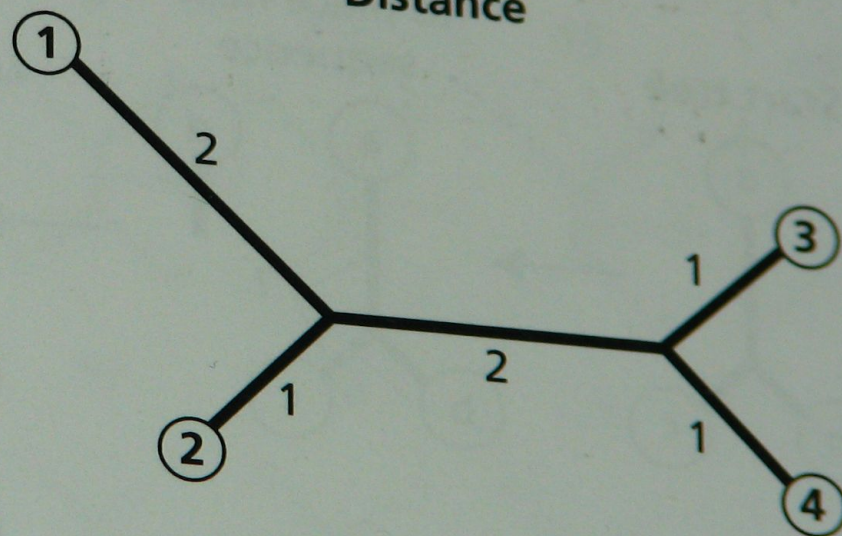
sequences		sites						
		1	2	3	4	5	6	7
1		T	T	A	T	T	A	A
2		A	A	T	T	T	A	A
3		A	A	A	A	A	T	A
4		A	A	A	A	A	A	T

sequences		sequences		
		1	2	3
2		3		
3		5	4	
4		5	4	2

Parsimony



Distance



Построение дерева на основании матрицы дискретных данных и на основании в матрицы дистанций

- Чем генетические дистанции отличаются от фенетических?
- Понятия сырой "р" дистанции и скорректированной дистанции
- модели эволюции

- Методы коррекции генетических дистанций

Если вероятности нуклеотидных замен (p) и частоты нуклеотидов (f) константны во времени, то суммарная эволюционная дистанция (доля измененных нуклеотидов) =

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

Где t это время, P_{AC} -

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

JC

Вероятности всех замен одинаковы, частоты нуклеотидов равны

Jukes–Cantor (JC)

$$\mathbf{P}_t = \begin{bmatrix} . & \alpha & \alpha & \alpha \\ \alpha & . & \alpha & \alpha \\ \alpha & \alpha & . & \alpha \\ \alpha & \alpha & \alpha & . \end{bmatrix},$$

$$\mathbf{f} = \left[\frac{1}{4} \frac{1}{4} \frac{1}{4} \frac{1}{4} \right]$$

K2P

Вероятности транзиций и трансверсий разные,
частоты нуклеотидов равны

Kimura's 2 parameter model (K2P)

α - транзиция

β - трансверсия

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

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)

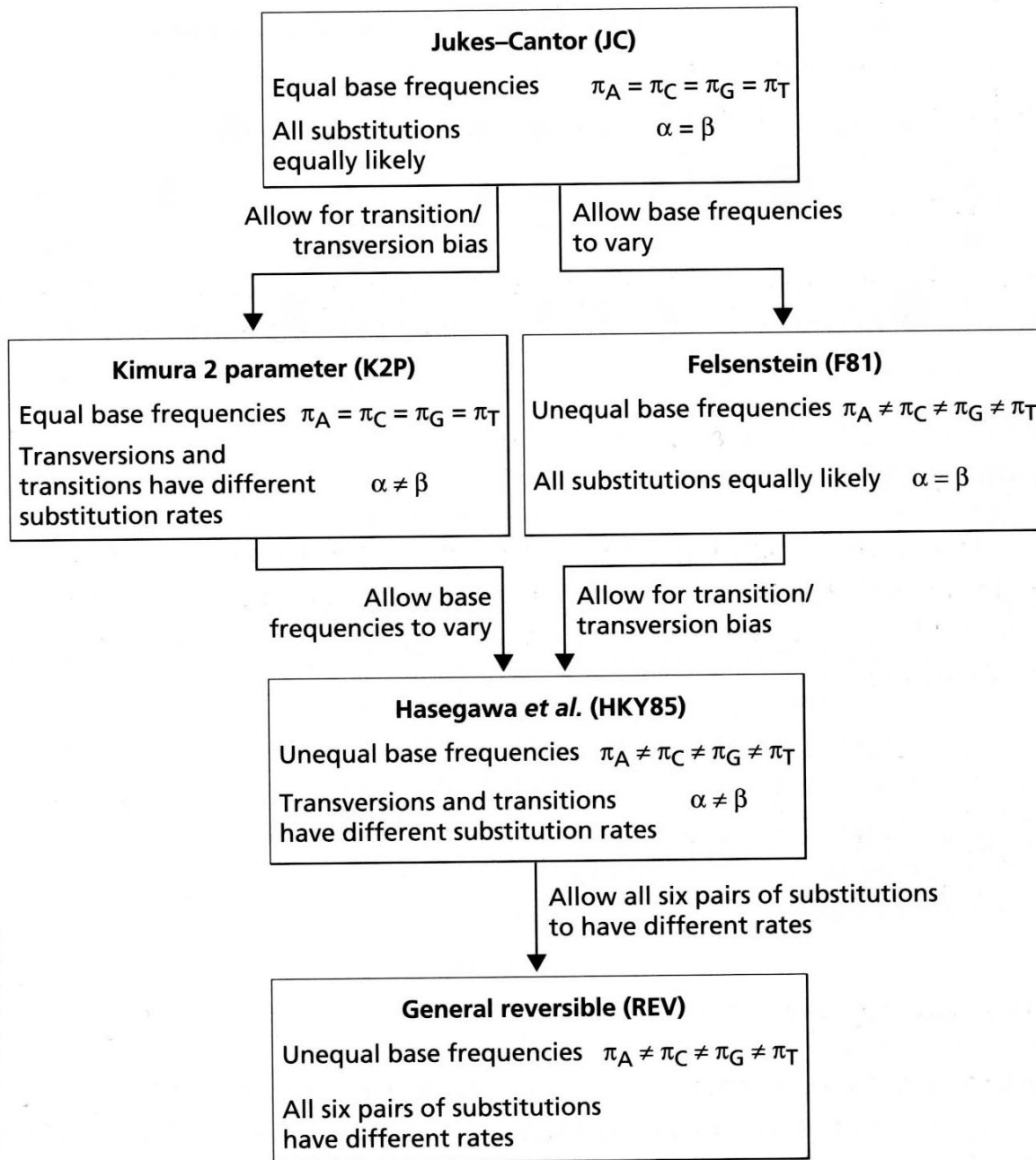
$$\mathbf{P}_t = \begin{bmatrix} . & \pi_C \beta & \pi_G \alpha & \pi_T \beta \\ \pi_A \beta & . & \pi_G \beta & \pi_T \alpha \\ \pi_A \alpha & \pi_C \beta & . & \pi_T \beta \\ \pi_A \beta & \pi_C \alpha & \pi_G \beta & . \end{bmatrix}, \quad \mathbf{f} = [\pi_A \ \pi_C \ \pi_G \ \pi_T]$$

REV

Вероятности ВСЕХ ЗАМЕН разные,
частоты нуклеотидов разные

General reversible model (REV)

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



Методы построения “дистантных” деревьев

- Методы основанные на использовании критериев оптимальности
- Методы, основанные на алгоритмах кластеризации

- Методы основанные на использовании критериев оптимальности
 - Метод наименьших квадратов
 - Оптимальным деревом признается то, при котором сумма квадратов генетических дистанций минимальна
 - Метод минимальной эволюции
 - Оптимальным деревом признается то, которое имеет наименьшую эволюционную длину (близко к идее максимальной парсимонии)

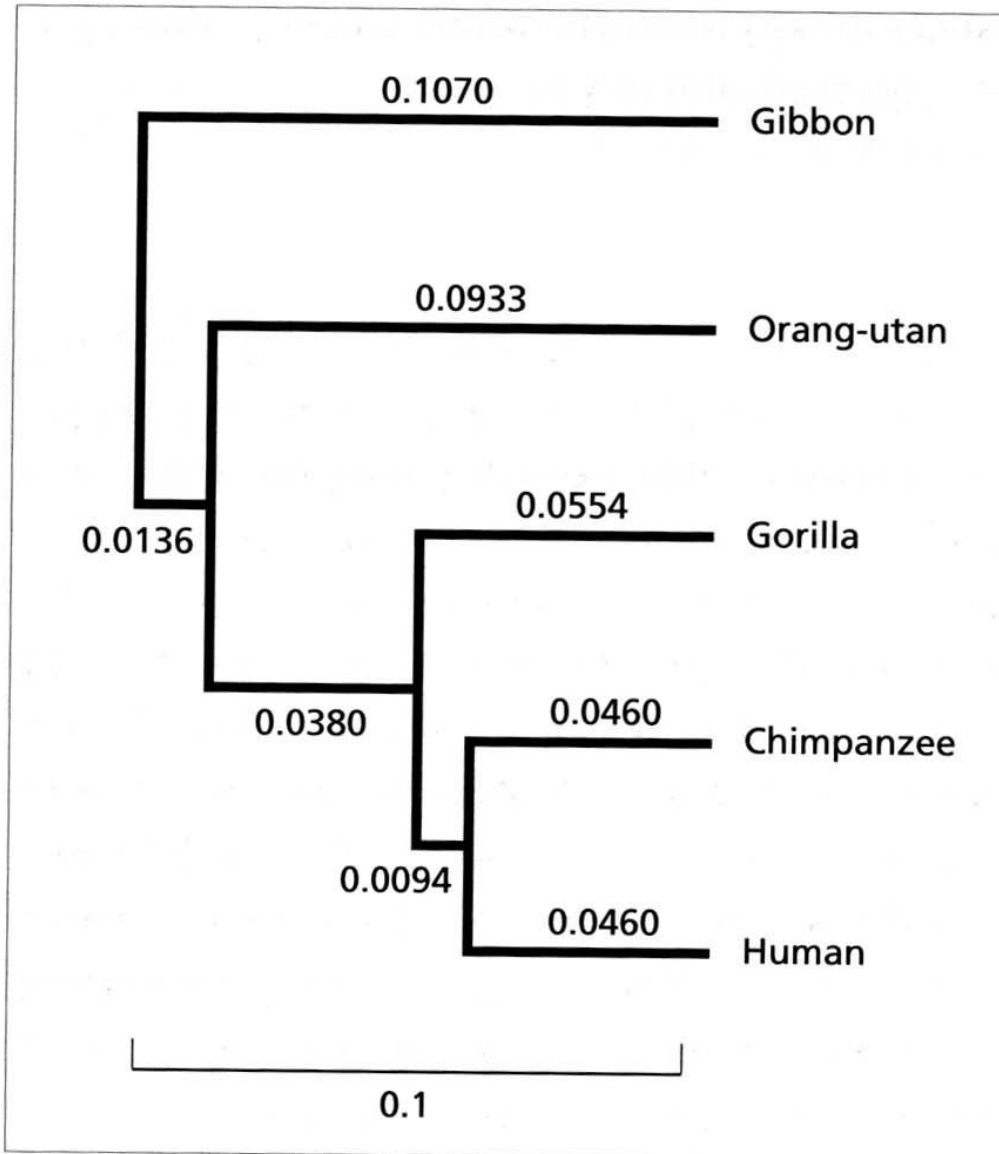


Fig. 6.9 Ultrametric tree for hominoid mtDNA with least squares branch lengths computed from the Kimura 2-parameter distances shown in the upper right triangle of Table 6.1. Compare with the additive tree in Fig. 6.7.

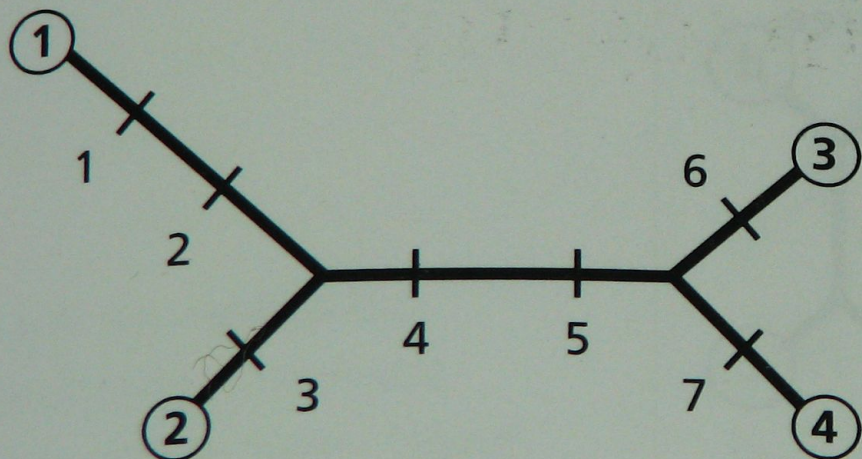
- **Методы основанные на использовании критериев оптимальности**
 - **Метод наименьших квадратов**
 - Оптимальным деревом признается то, при котором сумма квадратов генетических дистанций минимальна
 - **Метод минимальной эволюции**
 - Оптимальным деревом признается то, которое имеет наименьшую эволюционную длину (близко к идее максимальной парсимонии)

- Методы, основанные на алгоритмах кластеризации
 - Метод ближайшего соседа (Neighbour Joining)
 - Метод UPGMA (unweighted pair group method with arithmetic means)

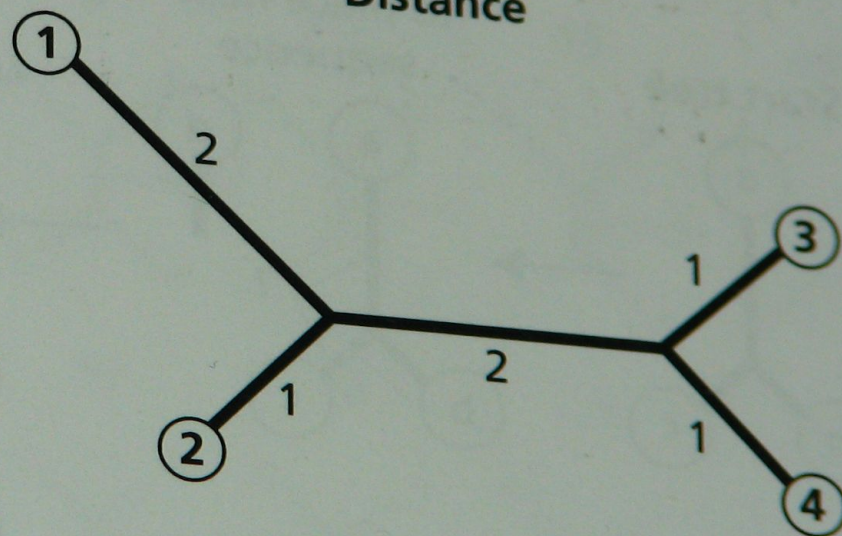
sequences		sites						
		1	2	3	4	5	6	7
1		T	T	A	T	T	A	A
2		A	A	T	T	T	A	A
3		A	A	A	A	A	T	A
4		A	A	A	A	A	A	T

sequences		sequences		
		1	2	3
2		3		
3		5	4	
4		5	4	2

Parsimony



Distance



Построение дерева на основании матрицы дискретных данных и на основании в матрицы дистанций

- Методы, основанные на алгоритмах кластеризации
 - Метод ближайшего соседа (Neighbour Joining)
 - Метод UPGMA (unweighted pair group method with arithmetic means)

Fig. 6.26 The condition required for UPGMA to successfully reconstruct the true tree is that the sum of the edges leading to sequence 1 ($a + b$) must be greater than the larger of c and d . After Mooers *et al.* (1994).

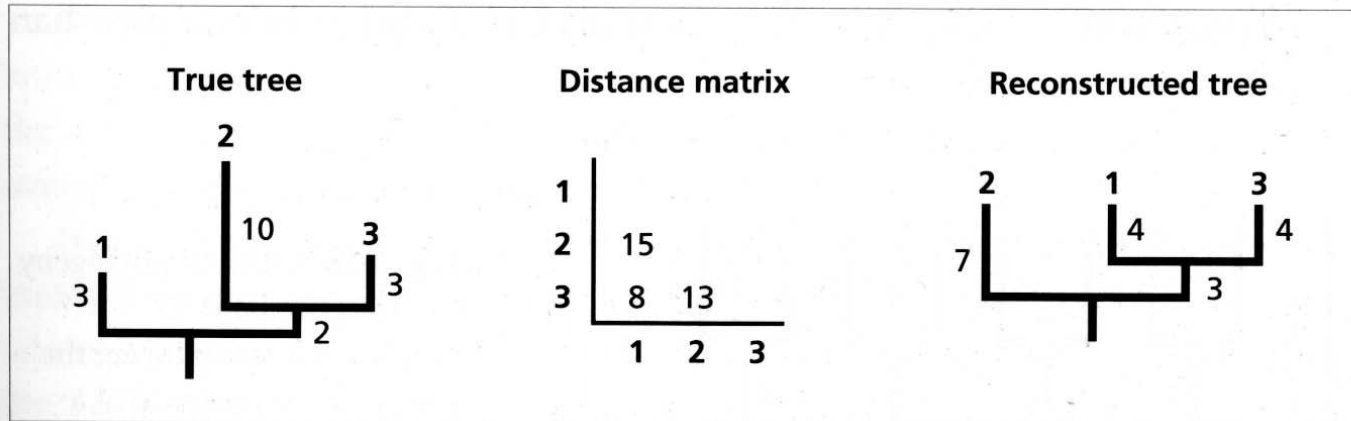
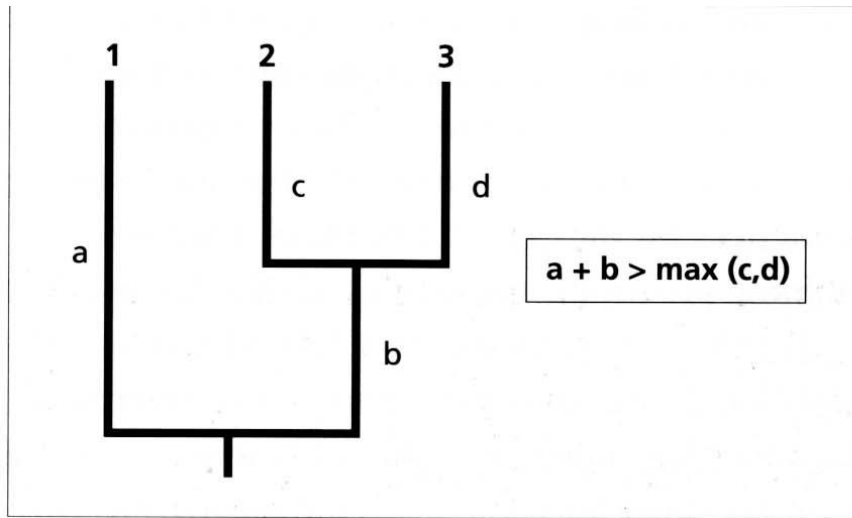


Fig. 6.27 An example where UPGMA will reconstruct the wrong tree. The edge lengths on the true tree violate the condition shown in Fig. 6.26, as $a + b = 3 + 2 < \max(c, d) = 10$. Sequence 2 has evolved more rapidly than the other two sequences, so that sequence 1 and 3 are more similar to each other than either is to 2.

Пример
Филогения бабочек рода *Parnassius*,
основанная на анализе гена COI с
использованием метода ближайшего соседа

