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	*гипотеза H_2*
10 черных шаров 30 белых шаров	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 H1

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 H1

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

  - НО:
  - 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 – H1 и H2

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

H1 –гипотеза 1
H2 – гипотеза 2
E - испытание
(P(H1/E) – постериорная вероятность гипотезы H1
(после получения данных E, т.е. после проведенного испытания E)
P(H1) – априорная вероятность гипотезы H1
P(E/H1) – вероятность наблюдения данных при условии,
что гипотеза H1 верна (=правдоподобие гипотезы)

В числителе P(E/H1) P(H1) – произведение вероятности наблюдения данных на априорную вероятность данной гипотезы

В знаменателе – сумма произведений P(E/H1) P(H1) для каждой из альтернативных гипотез (H1, H2 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 H1 = 0.5
P априорное для H2 = 0.5

гипотеза H1

10 черных шаров
30 белых шаров

гипотеза H2

20 черных шаров
20 белых шаров

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

Правдоподобие для H1 = 0.75 (вероятность, что первый вынутый шар будет белым)

Правдоподобие для H2 = 0.5 (вероятность, что первый вынутый шар будет белым)

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

$$= \frac{0.75 \times 0.5}{0.75 \times 0.5 + 0.5 \times 0.5}$$

$$= 0.6$$

P(H1) – априорная вероятность гипотезы H1
(P(H1/E) – постериорная вероятность гипотезы H1
P(E/H1) – вероятность наблюдения данных при условии, что гипотеза H1 верна (=правдоподобие гипотезы)
В числителе P(E/H1) P(H1) – произведение вероятности наблюдения данных на априорную вероятность данной гипотезы
В знаменателе – сумма произведений P(E/H1) P(H1) для каждой из альтернативных гипотез (H1, H2 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(H1)=0.6; P(H2)=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 x 0.75)/(0.6 x 0.75 + 0.4 x 0.5) = 0.45/(0.45 + 0.2) = 0.69

Третья итерация:
априорные вероятности снова изменились

P(H1)=0.69; P(H2)=0.31

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

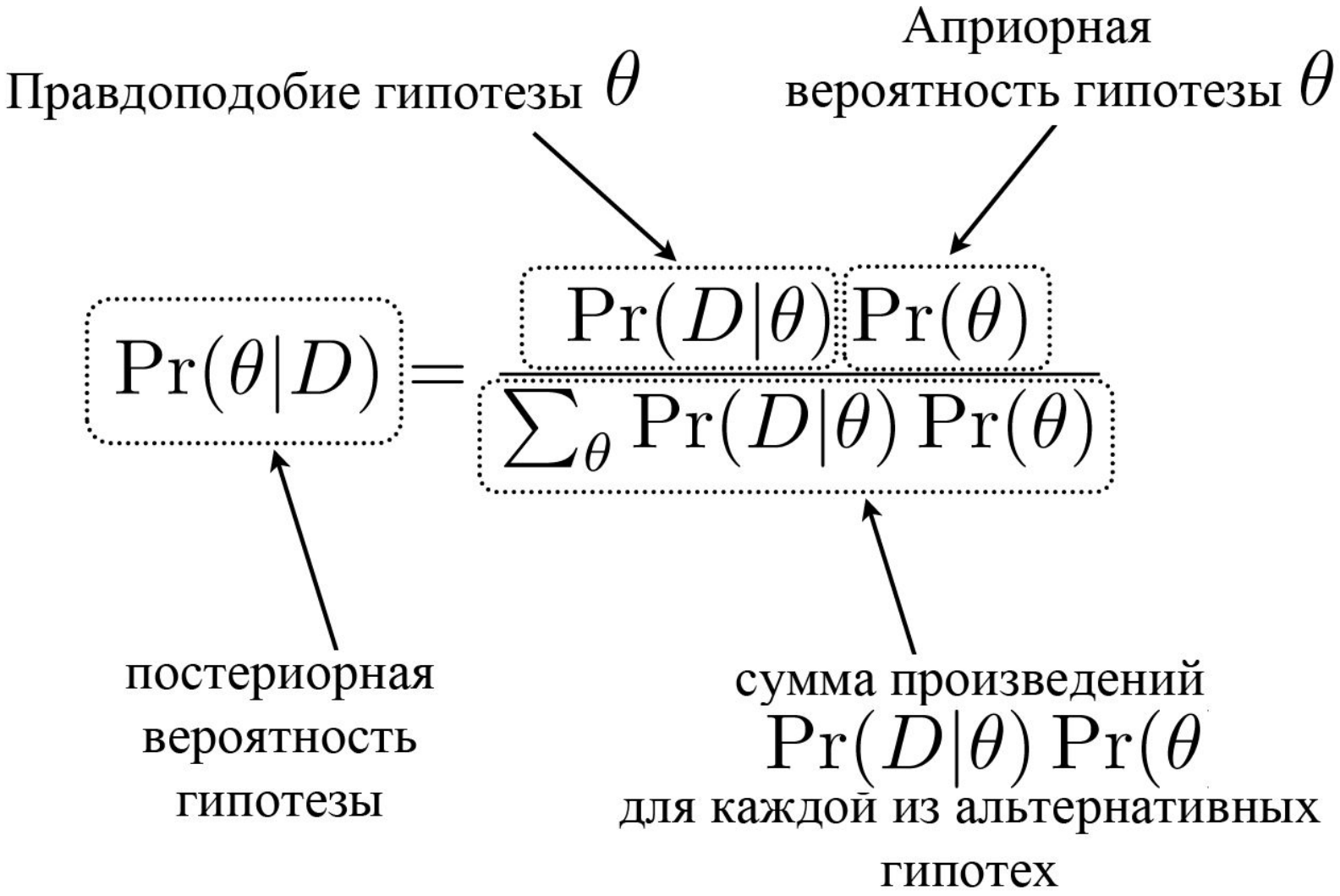
- P = (0.69 x 0.75)/(0.69 x 0.75 + 0.31 x 0.5) = 0.5175/(0.5175 + 0.155) = 0.5175/0.6725 = 0.77

Продолжаем процесс до тех пор пока вероятность одной из гипотез не достигнет 100% [P(H1)=1], т.е. гипотеза доказана
(или до стационарного уровня, когда вероятность гипотез стабилизируется)

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

- Лодка затонула 21 мая 1968 годаЛодка затонула 21 мая 1968 года в 740 км (400 миль) к юго-западу от Азорских острововЛодка 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 вНорфолк. Официально о потере USS Scorpion (SSN-589) было объявлено 5 июня 1968

D данные


Правдоподобие гипотезы θ
Априорная вероятность гипотезы θ
$\Pr(\theta|D) = \frac{\Pr(D|\theta)\Pr(\theta)}{\sum_\theta \Pr(D|\theta)\Pr(\theta)}$
постериорная вероятность гипотезы
сумма произведений $\Pr(D|\theta)\Pr(\theta$ для каждой из альтернативных гипотех

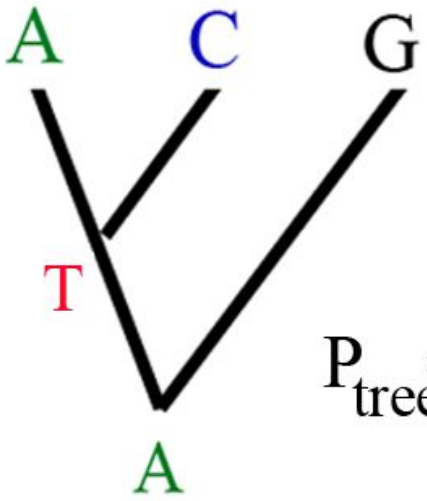
- Как все это перенести на реконструкцию филогении?

- - 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)=a

частоты нуклеотидов равны,
т.е. f(A)=f(C)=f(G)=f(T)=0.25
Pxy = a = 1/16=0,0625



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

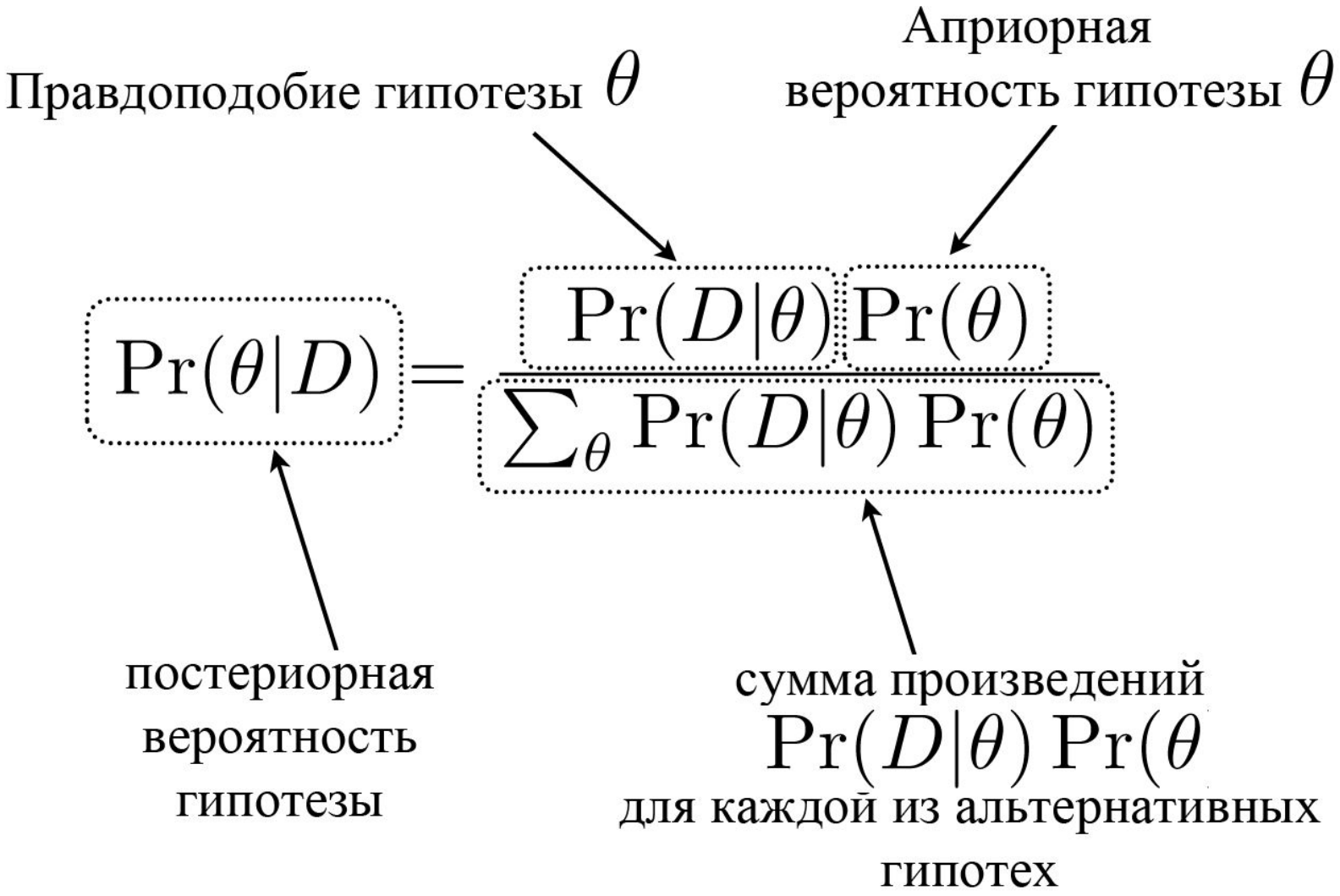
P_{tree} = 0.25xaxaxaxa =
= 0.25x0.0625x0.0625x0.0625x0.0625
=0.00000381

Правдоподобие гипотезы θ

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

это вероятность
получения наблюдений D
при условии выполнения гипотезы θ

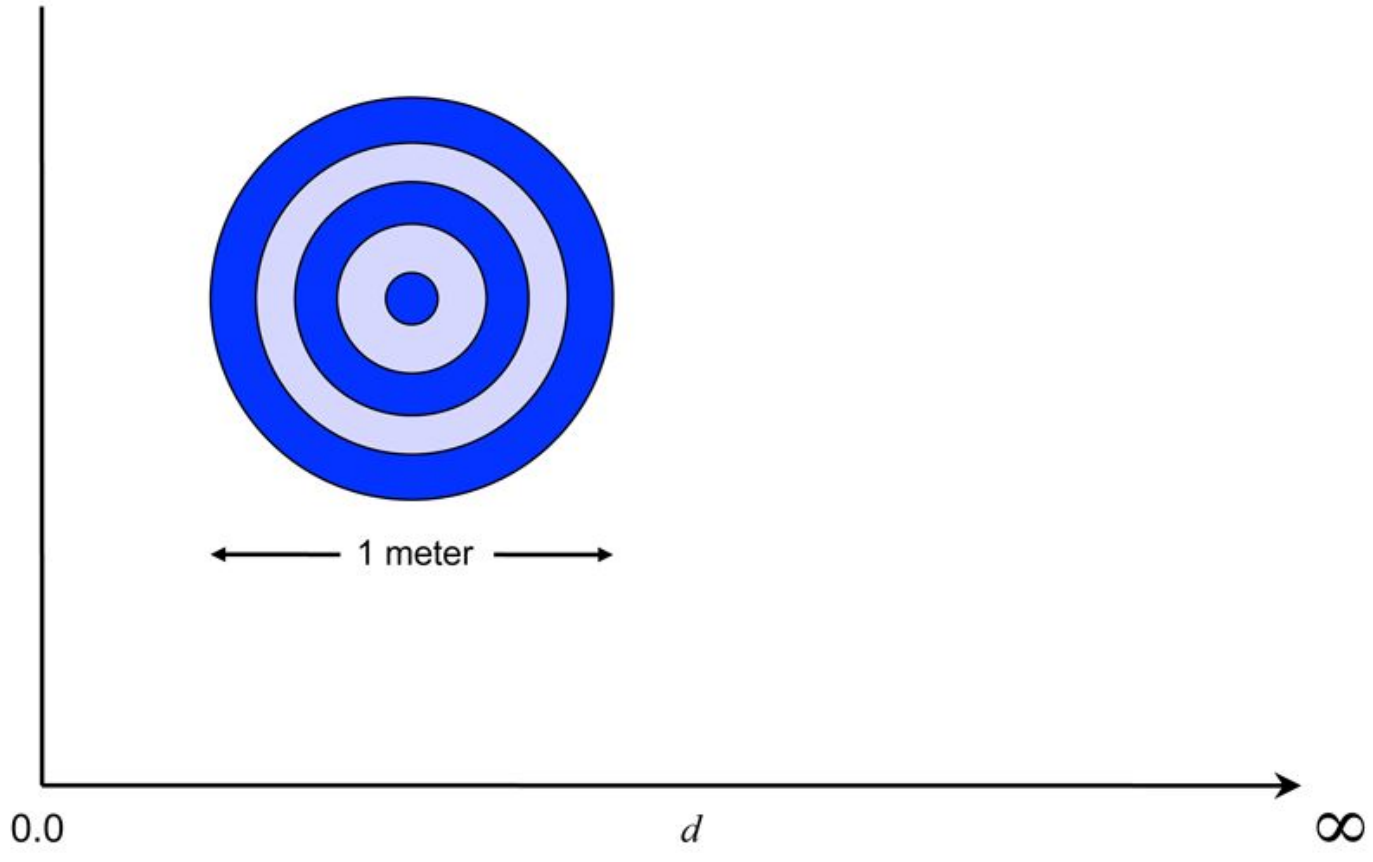
D данные


Правдоподобие гипотезы θ
Априорная вероятность гипотезы θ
$\Pr(\theta|D) = \frac{\Pr(D|\theta)\Pr(\theta)}{\sum_\theta \Pr(D|\theta)\Pr(\theta)}$
постериорная вероятность гипотезы
сумма произведений $\Pr(D|\theta)\Pr(\theta$ для каждой из альтернативных гипотех

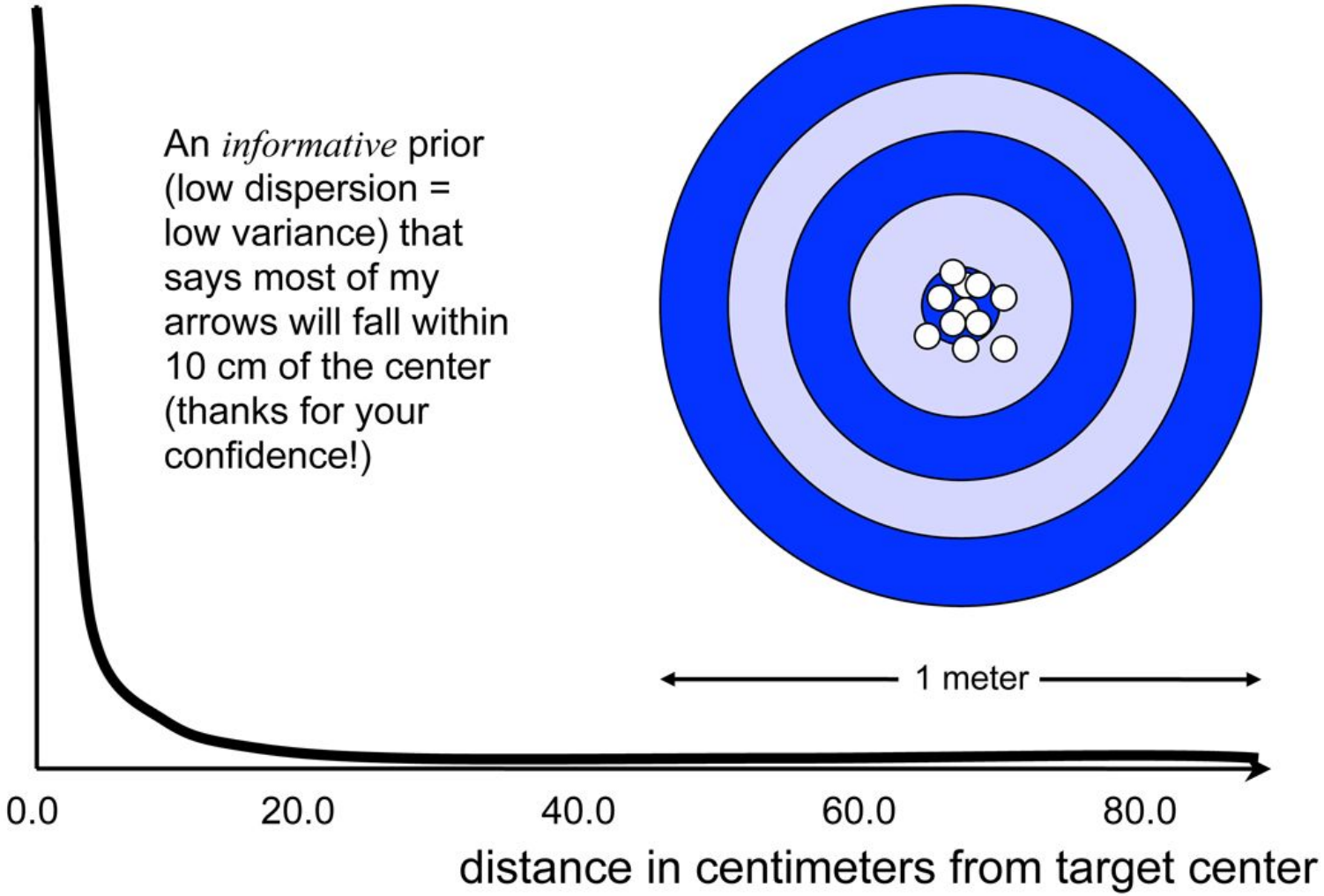
- Теперь вопрос, как перейти к филогенетическим гипотезам, т.е. деревьям

- 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)?

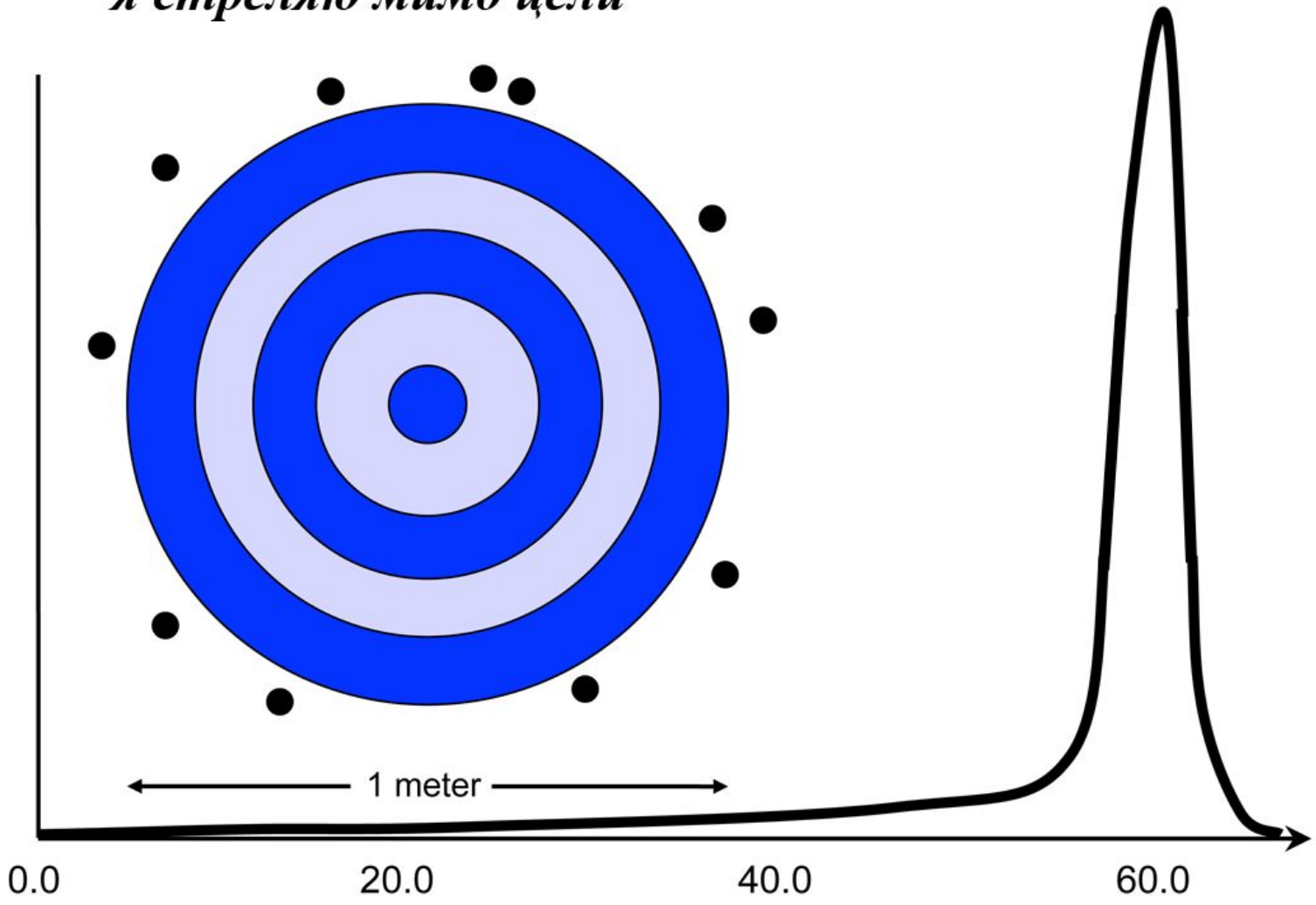
- анализировать не отдельные гипотезы (их может быть неограниченное множество), а статистические распределения этих гипотез


1 meter
0.0
d
∞

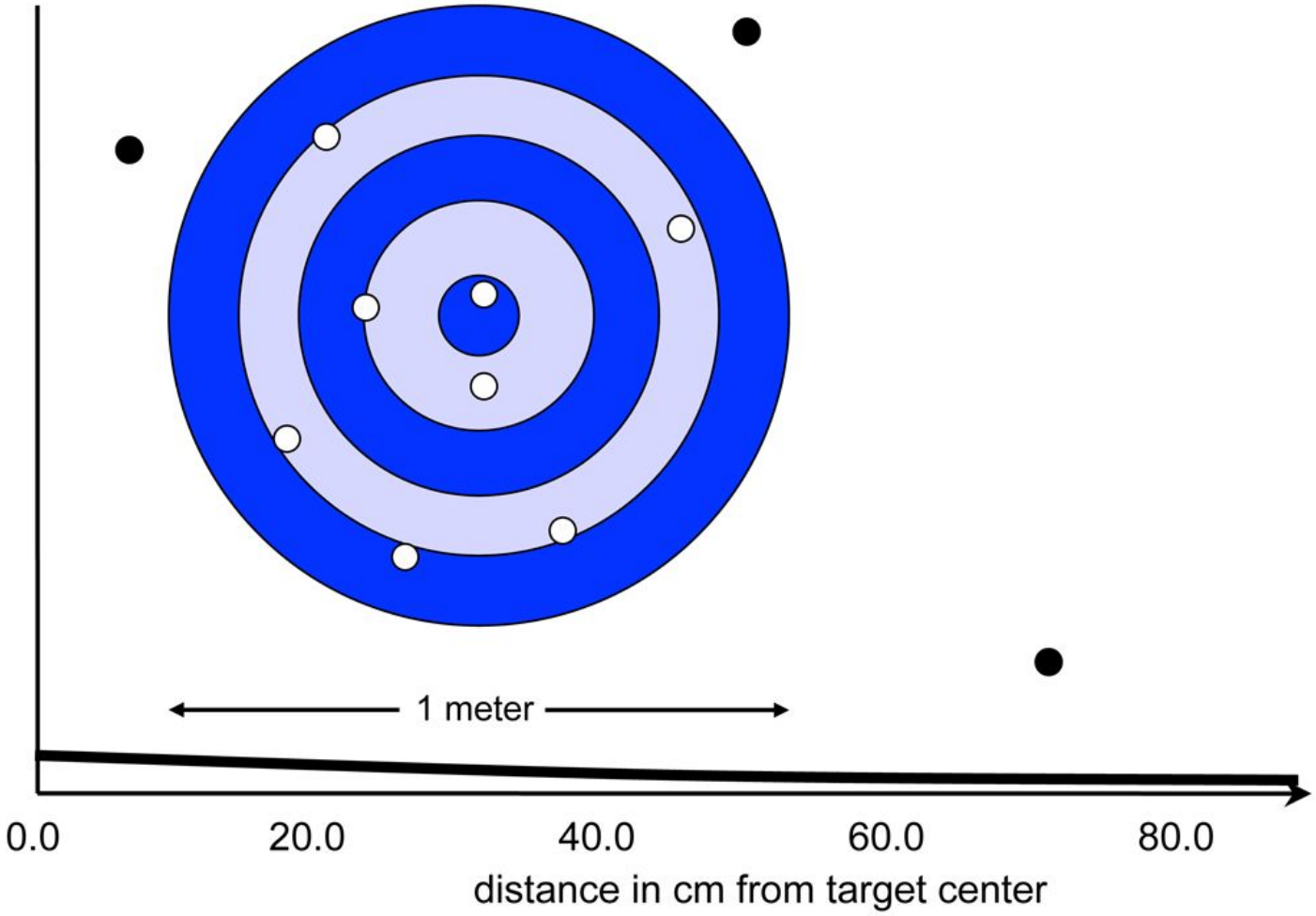
Априорная гипотеза 1
Я - меткий стрелок



Априорная гипотеза 2: я стреляю мимо цели



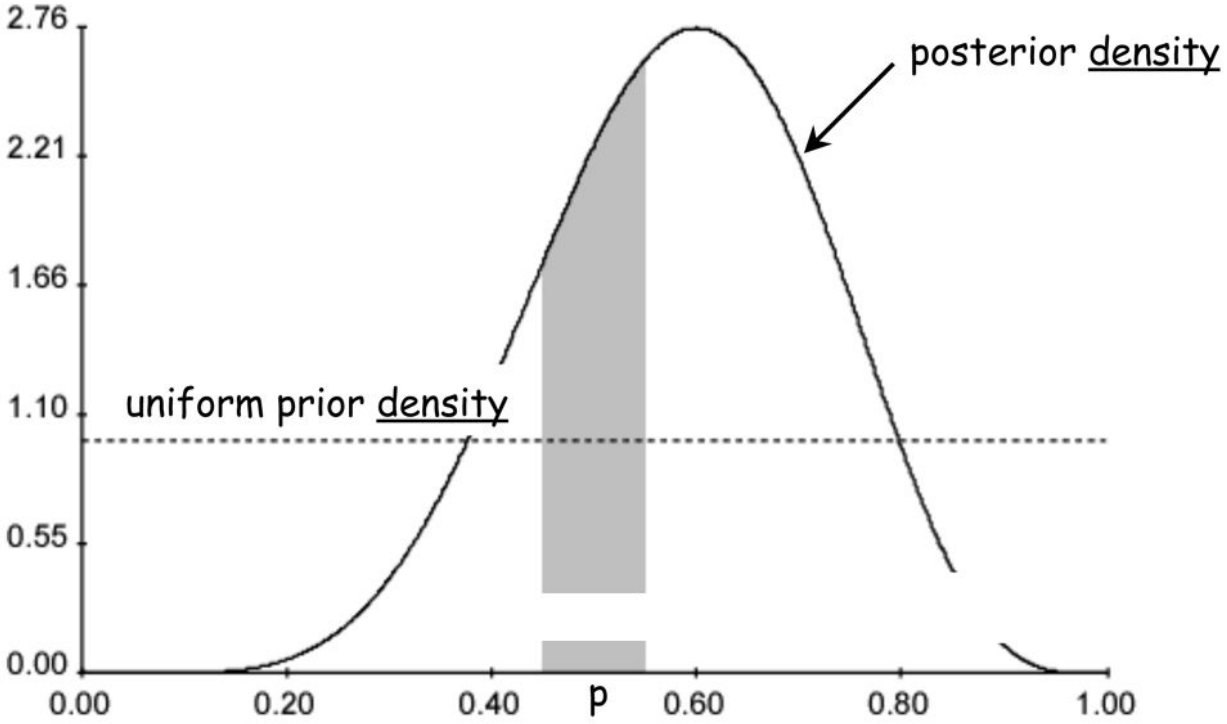
Априорная гипотеза 3
нет никаких особых идей



- Униформный (неспецифический прайор), казалось бы, какая от него польза.

- Но вспомним про итеративность… Итерации постепенно сдвигают распределение к более информативному

Постериорные рапределения 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):

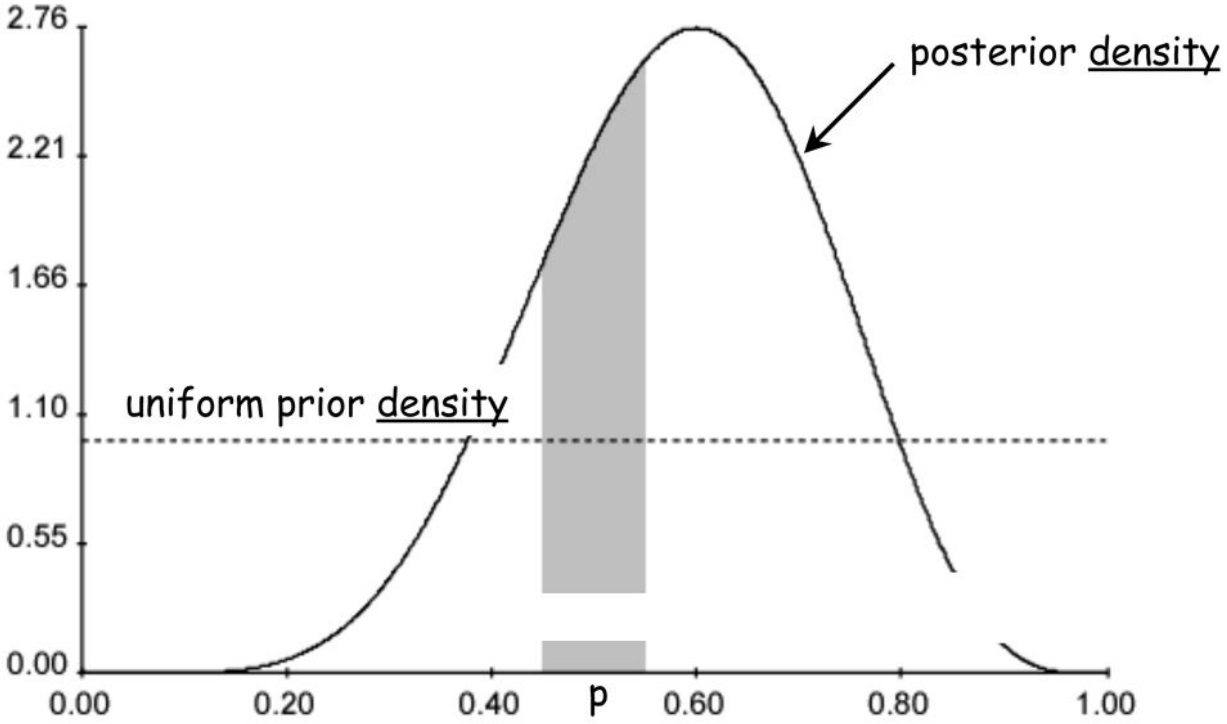
- 1) прайор о топологии
- 2) прайор о длине ветвей
- 3) прайор о частотах нуклеотидов
- 4) прайор о вероятности нуклеотидных замен разного типа
- 5) прайор о распределение вероятности замен по длине нуклеотидного выравнивания (параметр гамма)
- 6) прайор о доле инвариантных сайтов

Как рассчитать эти прайоры?

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

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

Постериорные рапределения вероятностей, как правило, более информативны

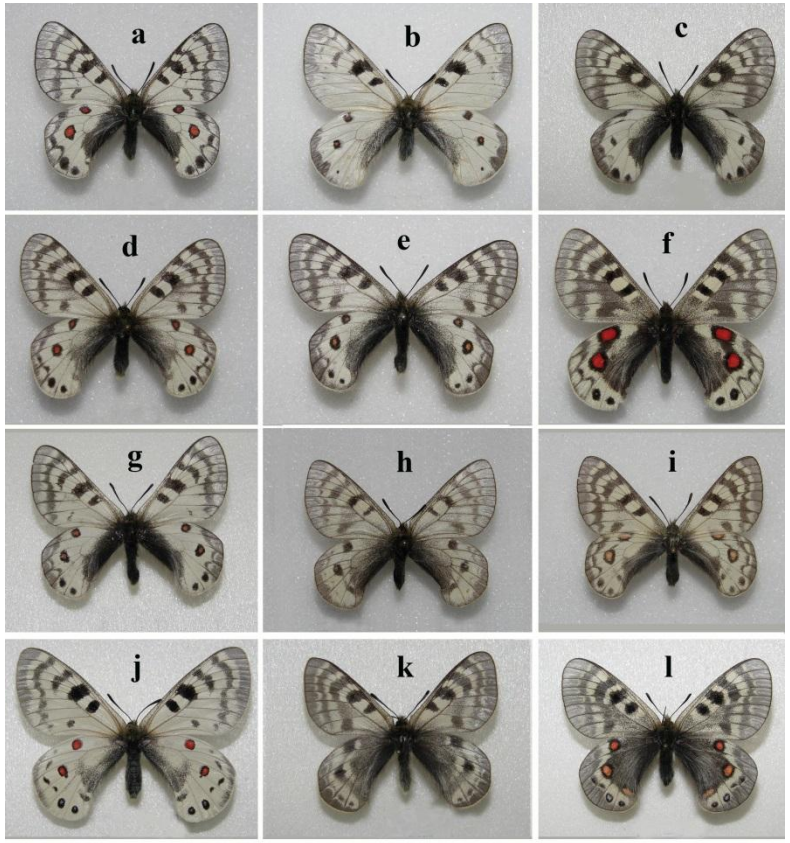


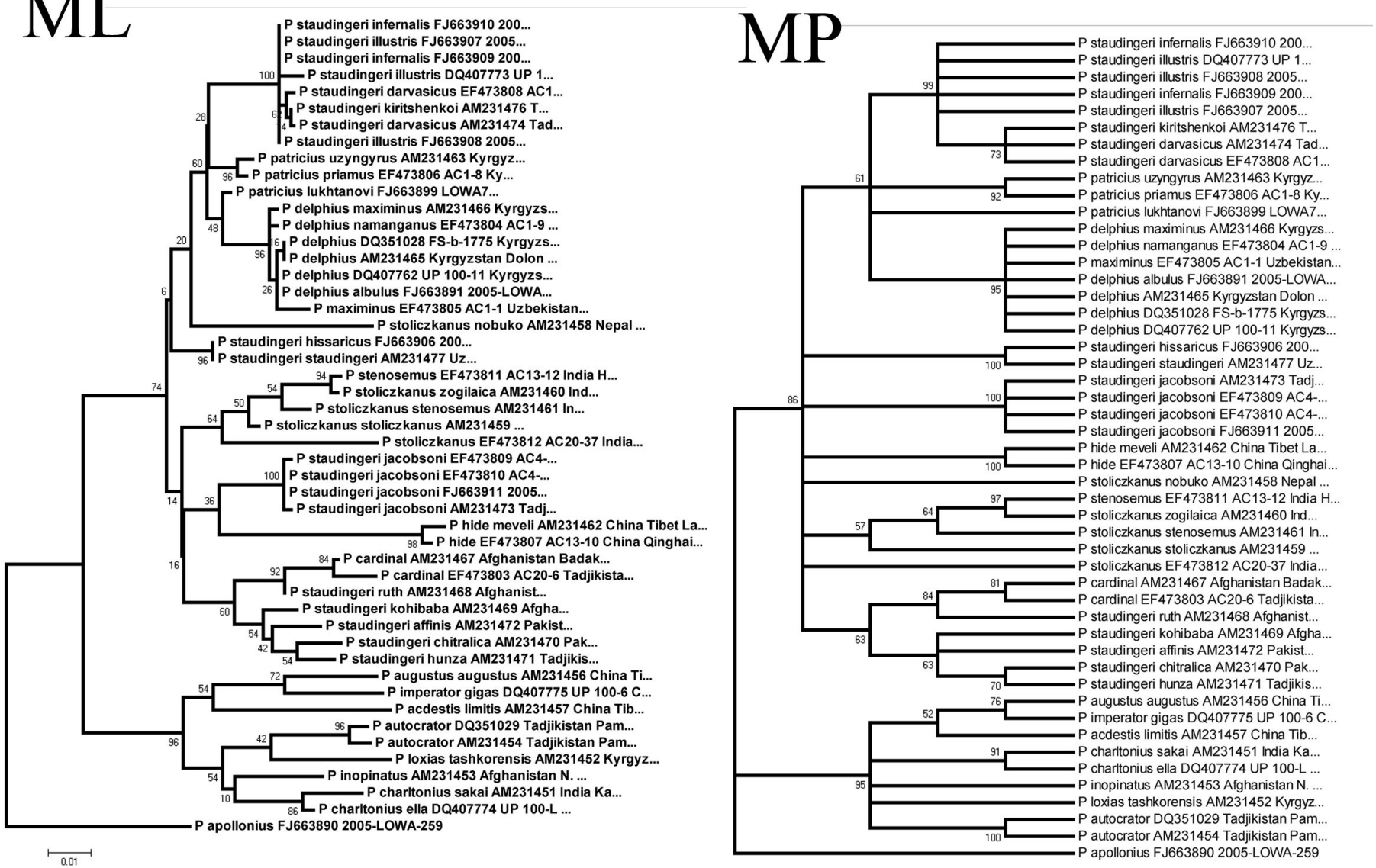
- При проведении анализа запускается несколько цепей (обычно 4), каждая из которых ищет оптимальные деревья
- Цепи могут обмениваться информацией, что позволяет “проскакивать” локальные оптимумы
- Получаемые деревья сравниваются и рассчитываются стандартные отклонения в положении ветвей. Анализ заканчивают,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

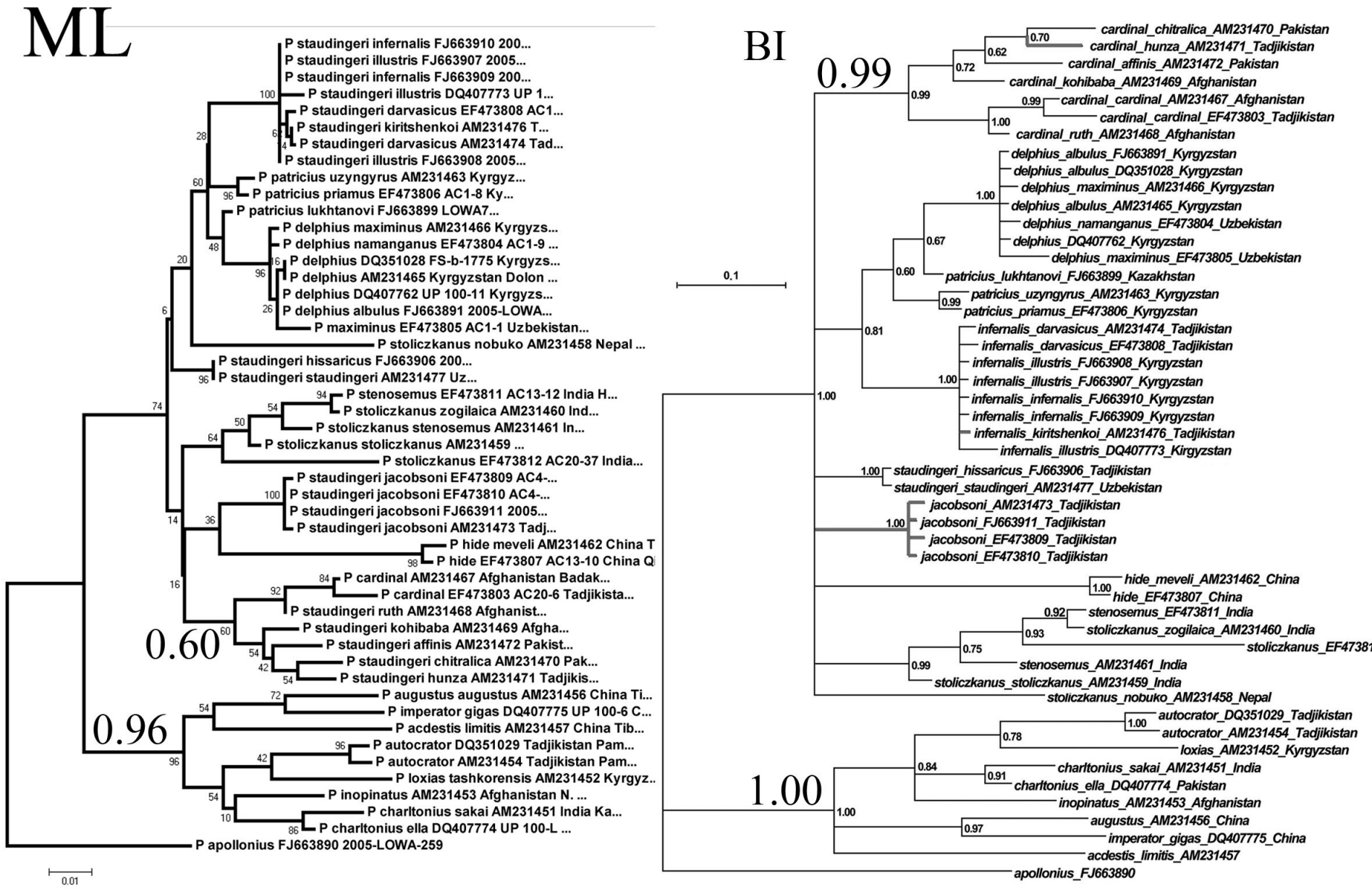
- Как задать прайоры в Байесовом анализе?
- Как выбрать модель эволюции в Байесовом анализе?

- GTR+I+G

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




ML
P staudingeri infernalis FJ663910 200...
P staudingeri illustris FJ663907 2005...
P staudingeri infernalis FJ663909 200...
P staudingeri illustris DQ407773 UP 1...
P staudingeri darvasicus EF473808 AC1...
P staudingeri kiritshenkoi AM231476 T...
P staudingeri darvasicus AM231474 Tad...
P staudingeri illustris FJ663908 2005...
P patricius uzyngyrus AM231463 Kyrgyz...
P patricius priamus EF473806 AC1-8 Ky...
P patricius lukhtanovi FJ663899 LOWA7...
P delphius maximinus AM231466 Kyrgyzs...
P delphius namanganus EF473804 AC1-9 ...
P delphius DQ351028 FS-b-1775 Kyrgyzs...
P delphius AM231465 Kyrgyzstan Dolon ...
P delphius DQ407762 UP 100-11 Kyrgyzs...
P delphius albulus FJ663891 2005-LOWA...
P maximinus EF473805 AC1-1 Uzbekistan...
P stoliczkanus nobuko AM231458 Nepal ...
P staudingeri hissaricus FJ663906 200...
P staudingeri staudingeri AM231477 Uz...
P stenosemus EF473811 AC13-12 India H...
P stoliczkanus zogilaica AM231460 Ind...
P stoliczkanus stenosemus AM231461 In...
P stoliczkanus stoliczkanus AM231459 ...
P stoliczkanus EF473812 AC20-37 India...
P staudingeri jacobsoni EF473809 AC4-...
P staudingeri jacobsoni EF473810 AC4-...
P staudingeri jacobsoni FJ663911 2005...
P staudingeri jacobsoni AM231473 Tadj...
P hide meveli AM231462 China Tibet La...
P hide EF473807 AC13-10 China Qinghai...
P cardinal AM231467 Afghanistan Badak...
P cardinal EF473803 AC20-6 Tadjikista...
P staudingeri ruth AM231468 Afghanist...
P staudingeri kohibaba AM231469 Afgha...
P staudingeri affinis AM231472 Pakist...
P staudingeri chitralica AM231470 Pak...
P staudingeri hunza AM231471 Tadjikis...
P augustus augustus AM231456 China Ti...
P imperator gigas DQ407775 UP 100-6 C...
P acdestis limitis AM231457 China Tib...
P autocrator DQ351029 Tadjikistan Pam...
P autocrator AM231454 Tadjikistan Pam...
P loxias tashkorensis AM231452 Kyrgyz...
P inopinatus AM231453 Afghanistan N. ...
P charltonius sakai AM231451 India Ka...
P charltonius ella DQ407774 UP 100-L ...
P apollonius FJ663890 2005-LOWA-259
0.01
MP
P staudingeri infernalis FJ663910 200...
P staudingeri illustris DQ407773 UP 1...
P staudingeri illustris FJ663908 2005...
P staudingeri infernalis FJ663909 200...
P staudingeri illustris FJ663907 2005...
P staudingeri kiritshenkoi AM231476 T...
P staudingeri darvasicus AM231474 Tad...
P staudingeri darvasicus EF473808 AC1...
P patricius uzyngyrus AM231463 Kyrgyz...
P patricius priamus EF473806 AC1-8 Ky...
P patricius lukhtanovi FJ663899 LOWA7...
P delphius maximinus AM231466 Kyrgyzs...
P delphius namanganus EF473804 AC1-9 ...
P maximinus EF473805 AC1-1 Uzbekistan...
P delphius albulus FJ663891 2005-LOWA...
P delphius AM231465 Kyrgyzstan Dolon ...
P delphius DQ351028 FS-b-1775 Kyrgyzs...
P delphius DQ407762 UP 100-11 Kyrgyzs...
P staudingeri hissaricus FJ663906 200...
P staudingeri staudingeri AM231477 Uz...
P staudingeri jacobsoni AM231473 Tadj...
P staudingeri jacobsoni EF473809 AC4-...
P staudingeri jacobsoni EF473810 AC4-...
P staudingeri jacobsoni FJ663911 2005...
P hide meveli AM231462 China Tibet La...
P hide EF473807 AC13-10 China Qinghai...
P stoliczkanus nobuko AM231458 Nepal ...
P stenosemus EF473811 AC13-12 India H...
P stoliczkanus zogilaica AM231460 Ind...
P stoliczkanus stenosemus AM231461 In...
P stoliczkanus stoliczkanus AM231459 ...
P stoliczkanus EF473812 AC20-37 India...
P cardinal AM231467 Afghanistan Badak...
P cardinal EF473803 AC20-6 Tadjikista...
P staudingeri ruth AM231468 Afghanist...
P staudingeri kohibaba AM231469 Afgha...
P staudingeri affinis AM231472 Pakist...
P staudingeri chitralica AM231470 Pak...
P staudingeri hunza AM231471 Tadjikis...
P augustus augustus AM231456 China Ti...
P imperator gigas DQ407775 UP 100-6 C...
P acdestis limitis AM231457 China Tib...
P charltonius sakai AM231451 India Ka...
P charltonius ella DQ407774 UP 100-L ...
P inopinatus AM231453 Afghanistan N. ...
P loxias tashkorensis AM231452 Kyrgyz...
P autocrator DQ351029 Tadjikistan Pam...
P autocrator AM231454 Tadjikistan Pam...
P apollonius FJ663890 2005-LOWA-259


ML
P staudingeri infernalis FJ663910 200...
P staudingeri illustris FJ663907 2005...
P staudingeri infernalis FJ663909 200...
P staudingeri illustris DQ407773 UP 1...
P staudingeri darvasicus EF473808 AC1...
P staudingeri kiritshenkoi AM231476 T...
P staudingeri darvasicus AM231474 Tad...
P staudingeri illustris FJ663908 2005...
P patricius uzyngyrus AM231463 Kyrgyz...
P patricius priamus EF473806 AC1-8 Ky...
P patricius lukhtanovi FJ663899 LOWA7...
P delphius maximinus AM231466 Kyrgyzs...
P delphius namanganus EF473804 AC1-9 ...
P delphius DQ351028 FS-b-1775 Kyrgyzs...
P delphius AM231465 Kyrgyzstan Dolon ...
P delphius DQ407762 UP 100-11 Kyrgyzs...
P delphius albulus FJ663891 2005-LOWA...
P maximinus EF473805 AC1-1 Uzbekistan...
P stoliczkanus nobuko AM231458 Nepal ...
P staudingeri hissaricus FJ663906 200...
P staudingeri staudingeri AM231477 Uz...
P stenosemus EF473811 AC13-12 India H...
P stoliczkanus zogilaica AM231460 Ind...
P stoliczkanus stenosemus AM231461 In...
P stoliczkanus stoliczkanus AM231459 ...
P stoliczkanus EF473812 AC20-37 India...
P staudingeri jacobsoni EF473809 AC4-...
P staudingeri jacobsoni EF473810 AC4-...
P staudingeri jacobsoni FJ663911 2005...
P staudingeri jacobsoni AM231473 Tadj...
P hide meveli AM231462 China T
P hide EF473807 AC13-10 China Q
P cardinal AM231467 Afghanistan Badak...
P cardinal EF473803 AC20-6 Tadjikista...
P staudingeri ruth AM231468 Afghanist...
P staudingeri kohibaba AM231469 Afgha...
P staudingeri affinis AM231472 Pakist...
P staudingeri chitralica AM231470 Pak...
P staudingeri hunza AM231471 Tadjikis...
P augustus augustus AM231456 China Ti...
P imperator gigas DQ407775 UP 100-6 C...
P acdestis limitis AM231457 China Tib...
P autocrator DQ351029 Tadjikistan Pam...
P autocrator AM231454 Tadjikistan Pam...
P loxias tashkorensis AM231452 Kyrgyz...
P inopinatus AM231453 Afghanistan N. ...
P charltonius sakai AM231451 India Ka...
P charltonius ella DQ407774 UP 100-L ...
P apollonius FJ663890 2005-LOWA-259
0.60
0.96
0.01
BI
cardinal_chitralica_AM231470_Pakistan
cardinal_hunza_AM231471_Tadjikistan
cardinal_affinis_AM231472_Pakistan
cardinal_kohibaba_AM231469_Afghanistan
cardinal_cardinal_AM231467_Afghanistan
cardinal_cardinal_EF473803_Tadjikistan
cardinal_ruth_AM231468_Afghanistan
delphius_albulus_FJ663891_Kyrgyzstan
delphius_albulus_DQ351028_Kyrgyzstan
delphius_maximinus_AM231466_Kyrgyzstan
delphius_albulus_AM231465_Kyrgyzstan
delphius_namanganus_EF473804_Uzbekistan
delphius_DQ407762_Kyrgyzstan
delphius_maximinus_EF473805_Uzbekistan
patricius_lukhtanovi_FJ663899_Kazakhstan
patricius_uzyngyrus_AM231463_Kyrgyzstan
patricius_priamus_EF473806_Kyrgyzstan
infernalis_darvasicus_AM231474_Tadjikistan
infernalis_darvasicus_EF473808_Tadjikistan
infernalis_illustris_FJ663908_Kyrgyzstan
infernalis_illustris_FJ663907_Kyrgyzstan
infernalis_infernalis_FJ663910_Kyrgyzstan
infernalis_infernalis_FJ663909_Kyrgyzstan
infernalis_kiritshenkoi_AM231476_Tadjikistan
infernalis_illustris_DQ407773_Kirgyzstan
staudingeri_hissaricus_FJ663906_Tadjikistan
staudingeri_staudingeri_AM231477_Uzbekistan
jacobsoni_AM231473_Tadjikistan
jacobsoni_FJ663911_Tadjikistan
jacobsoni_EF473809_Tadjikistan
jacobsoni_EF473810_Tadjikistan
hide_meveli_AM231462_China
hide_EF473807_China
stenosemus_EF473811_India
stoliczkanus_zogilaica_AM231460_India
stoliczkanus_EF47381
stenosemus_AM231461_India
stoliczkanus_stoliczkanus_AM231459_India
stoliczkanus_nobuko_AM231458_Nepal
autocrator_DQ351029_Tadjikistan
autocrator_AM231454_Tadjikistan
loxias_AM231452_Kyrgyzstan
charltonius_sakai_AM231451_India
charltonius_ella_DQ407774_Pakistan
inopinatus_AM231453_Afghanistan
augustus_AM231456_China
imperator_gigas_DQ407775_China
acdestis_limitis_AM231457
apollonius_FJ663890
0.99
1.00
0.1

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 (PB), которая является прямой оценкой вероятности филогенетической реконструкции – поэтому не нужен бутстреп!

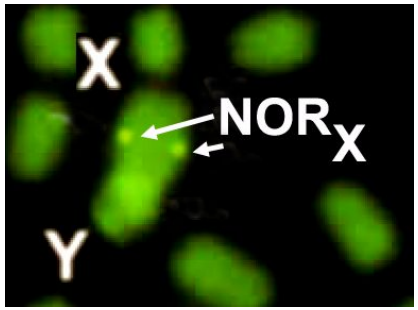
Методы максимального правдоподобия и Байеса: сходство и различия, плюсы и минусы

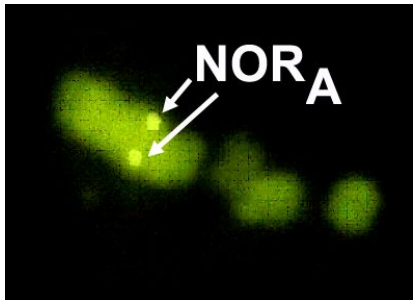
ML говорит лишь о степени соответствии данных и модели, но не говорит о достоверности тестируемой гипотезы (пример с гномами)

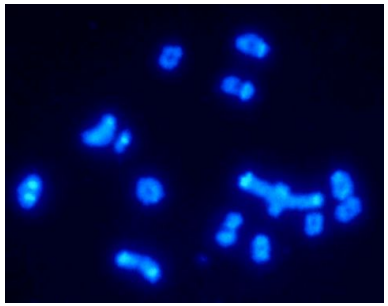
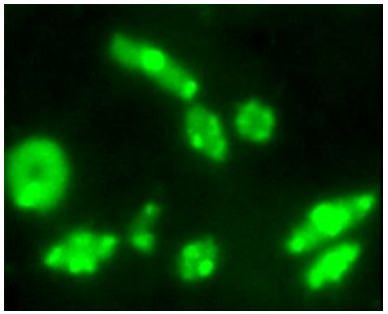
MB пытается заглянуть внутрь черного ящика. Оценка вероятности самой гипотезы

- Методы максимального правдоподобия и Байеса **применимы для анализа любых структур**, закономерности эволюции которых могут быть формализованы в виде параметрический моделей

- Например, для филогенетического анализа хромосомных перестроек



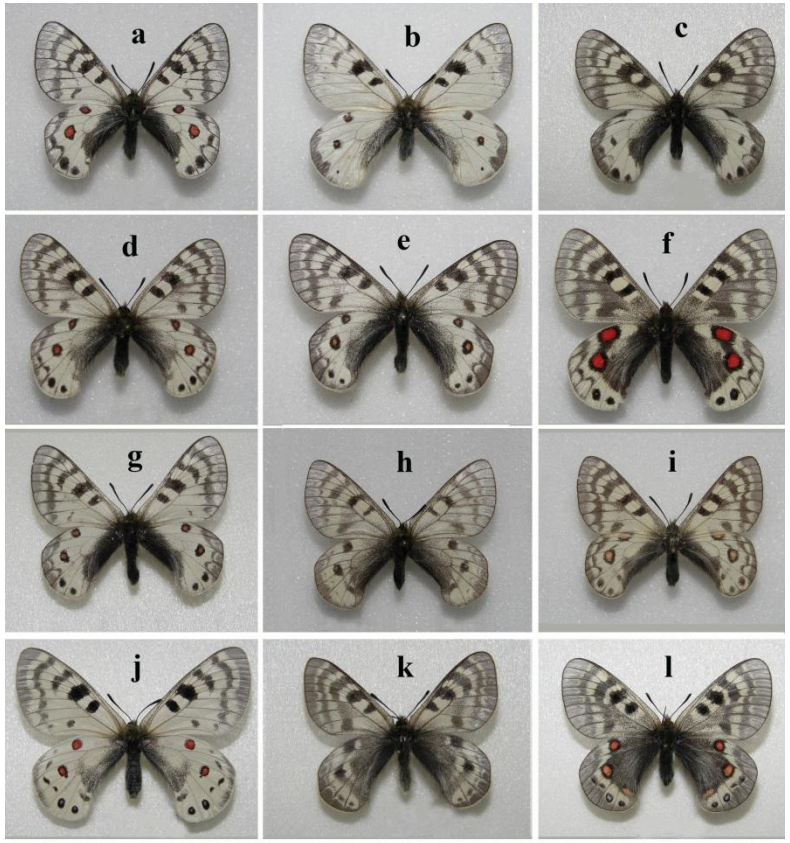


- **Не существует** никакого **теоретического запрета на использование морфологических признаков** в рамках метода максимального правдоподобия и Байесова метода

- 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

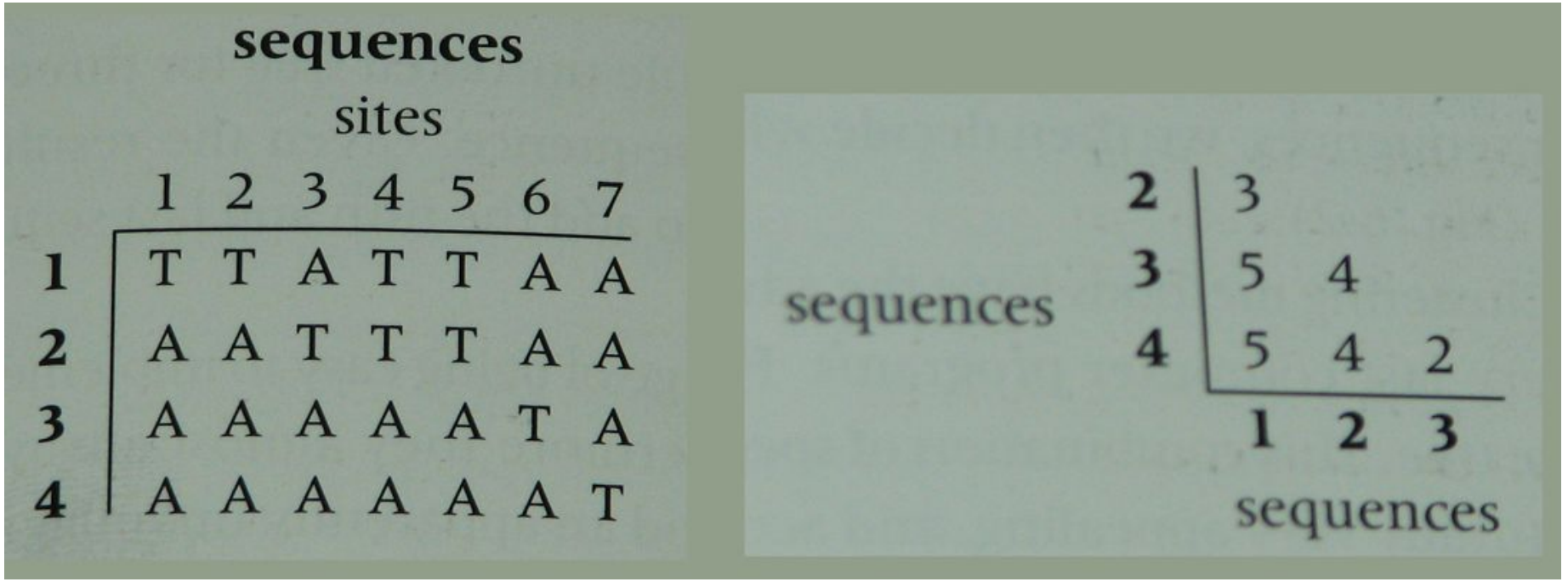
ДНК:

```
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

Превращение матрицы дискретных данных в матрицу дистанций



sequences

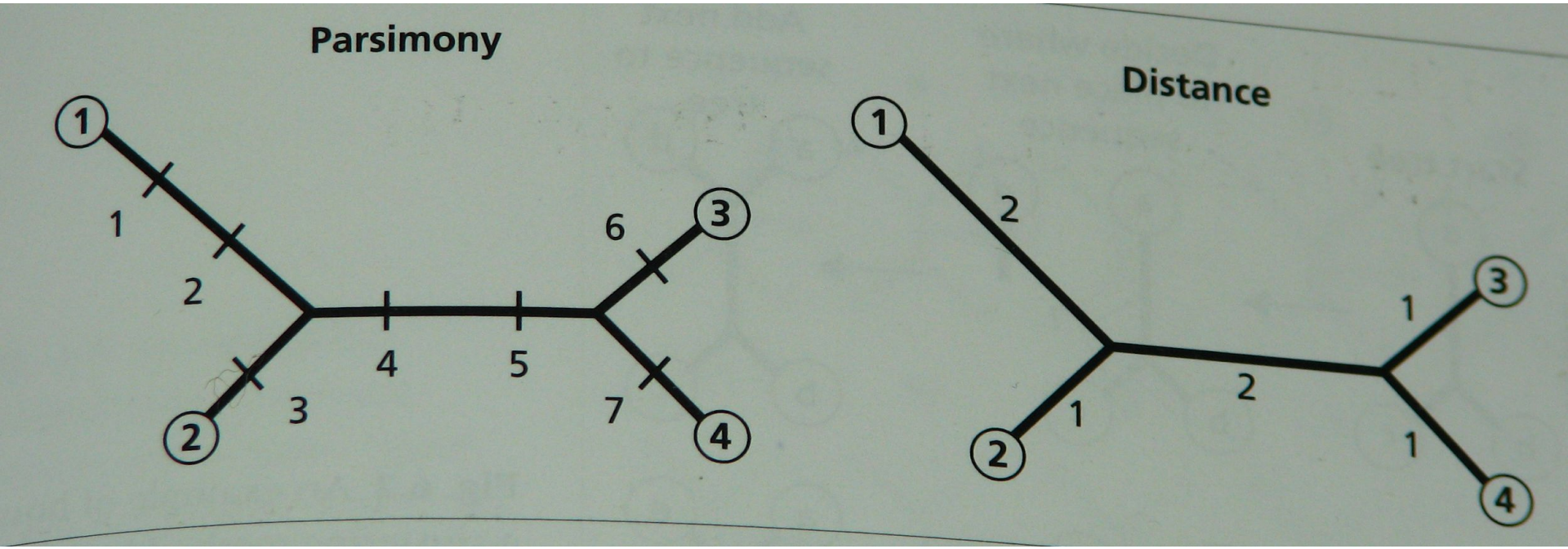
sites

	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



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

- Чем генетические дистанции отличаются от фенетических?
- Понятия сырой “p” 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) =

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

Где t это время, P_{AC} –

$P_{AC} = P_{CA}$

JC

Вероятности всех замен одинаковы, частоты нуклеотидов равны

Jukes–Cantor (JC)

$$\mathbf{P}_t = \begin{bmatrix} . & \alpha & \alpha & \alpha \\ \alpha & . & \alpha & \alpha \\ \alpha & \alpha & . & \alpha \\ \alpha & \alpha & \alpha & . \end{bmatrix}, \qquad \mathbf{f} = [\tfrac{1}{4}\ \tfrac{1}{4}\ \tfrac{1}{4}\ \tfrac{1}{4}]$$

K2P
Вероятности транзиций и трансверсий разные, частоты нуклеотидов равны

Kimura's 2 parameter model (K2P)

α – транзиция

β – трансверсия

$$\mathbf{P}_t = \begin{bmatrix} . & \beta & \alpha & \beta \\ \beta & . & \beta & \alpha \\ \alpha & \beta & . & \beta \\ \beta & \alpha & \beta & . \end{bmatrix}, \qquad \mathbf{f} = [\tfrac{1}{4}\ \tfrac{1}{4}\ \tfrac{1}{4}\ \tfrac{1}{4}].$$

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)

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

K2P
Вероятности транзиций и трансверсий разные, частоты нуклеотидов разные

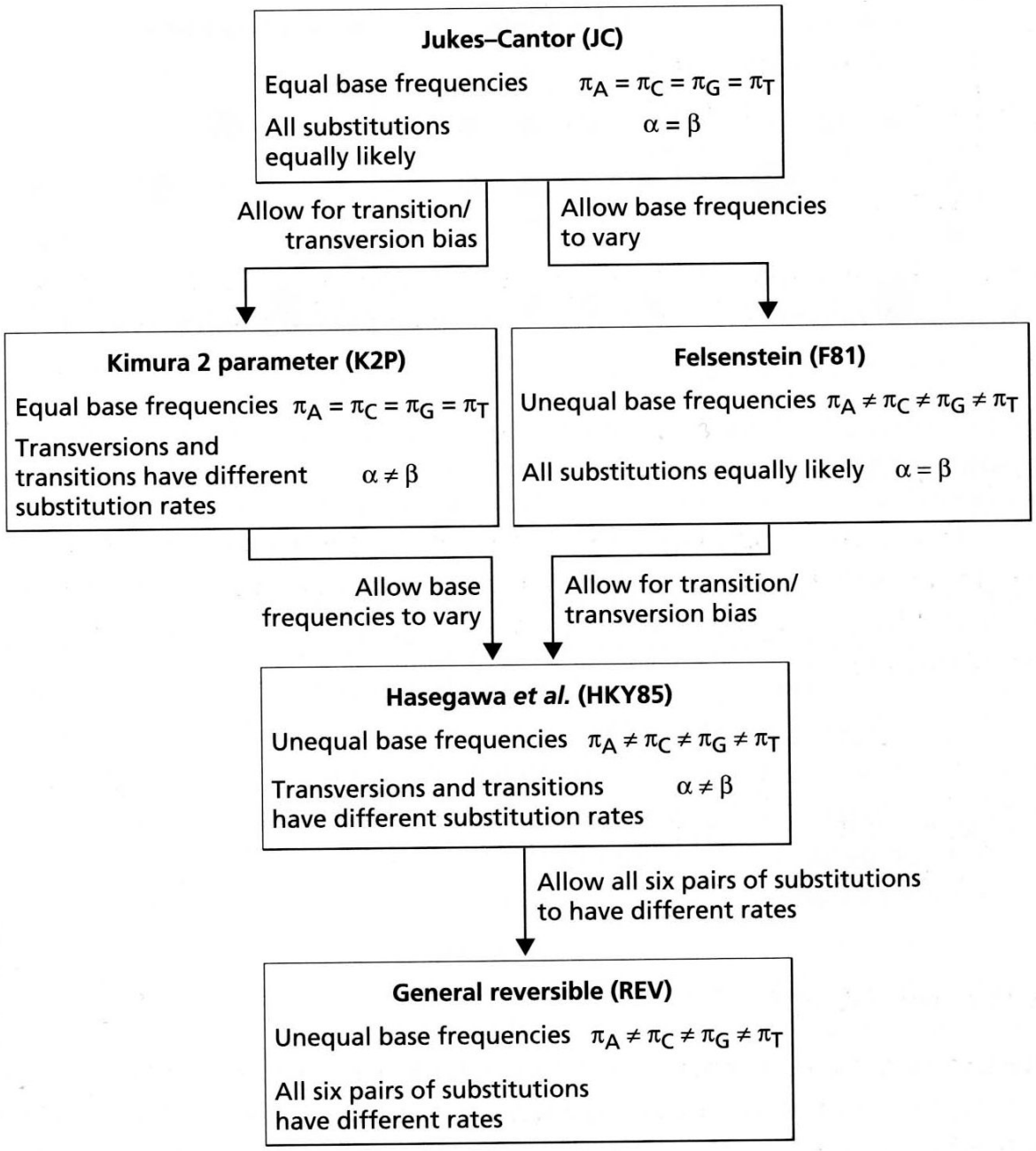
Hasegawa, Kishino and Yano (1985)

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

REV
Вероятности ВСЕХ ЗАМЕН разные,
частоты нуклеотидов разные

General reversible model (REV)

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


Jukes–Cantor (JC)
Equal base frequencies $\pi_A = \pi_C = \pi_G = \pi_T$
All substitutions equally likely $\alpha = \beta$
Allow for transition/ transversion bias
Allow base frequencies to vary
Kimura 2 parameter (K2P)
Equal base frequencies $\pi_A = \pi_C = \pi_G = \pi_T$
Transversions and transitions have different substitution rates $\alpha \neq \beta$
Felsenstein (F81)
Unequal base frequencies $\pi_A \neq \pi_C \neq \pi_G \neq \pi_T$
All substitutions equally likely $\alpha = \beta$
Allow base frequencies to vary
Allow for transition/ transversion bias
Hasegawa *et al.* (HKY85)
Unequal base frequencies $\pi_A \neq \pi_C \neq \pi_G \neq \pi_T$
Transversions and transitions have different substitution rates $\alpha \neq \beta$
Allow all six pairs of substitutions to have different rates
General reversible (REV)
Unequal base frequencies $\pi_A \neq \pi_C \neq \pi_G \neq \pi_T$
All six pairs of substitutions have different rates

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)

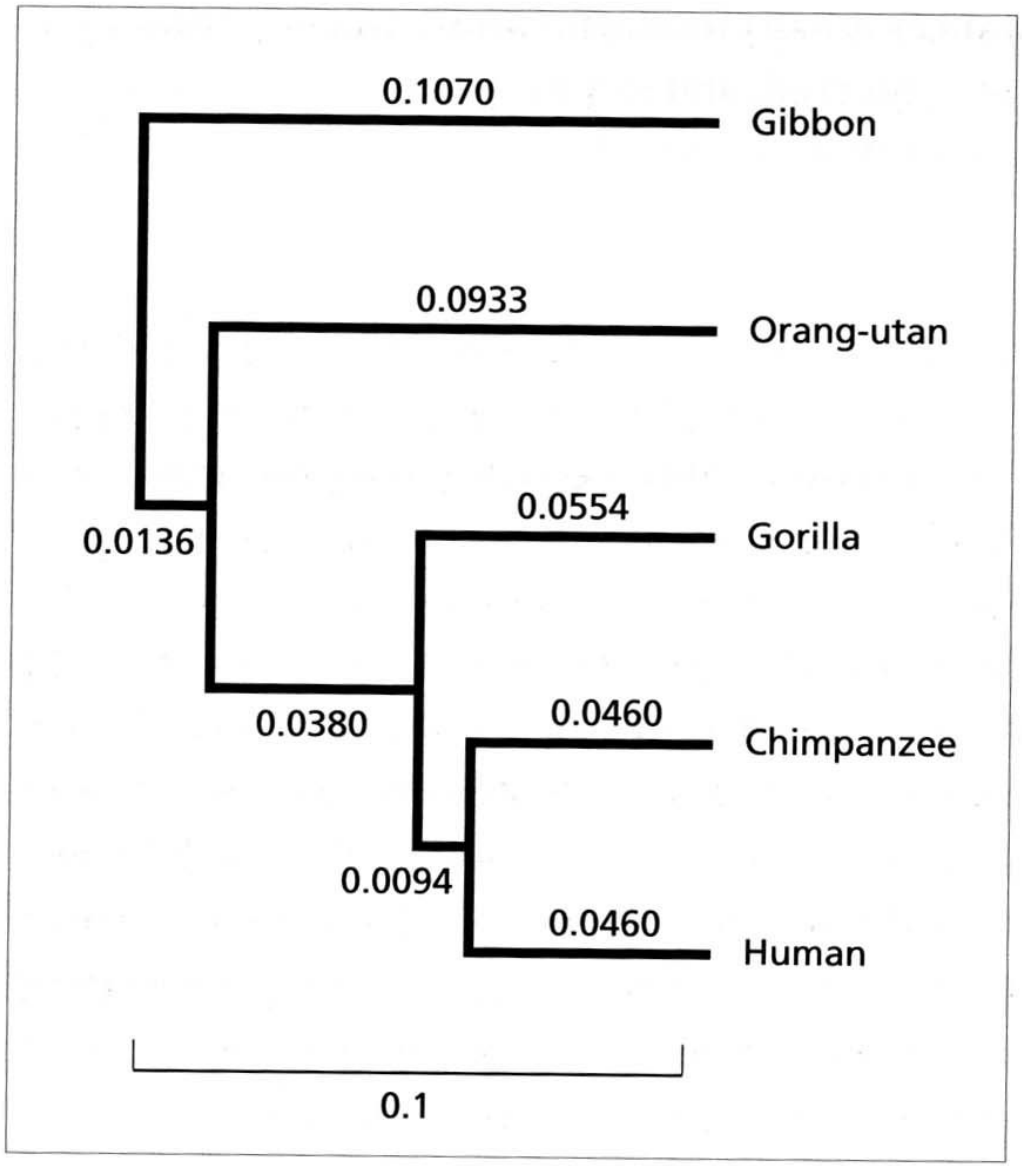


Fig. 6.9 Ultrametric tree for hominoid mtDNA with least squares branch lengths computed from the Kimura 2-parameter distances shown in the upper right triangle of Table 6.1. Compare with the additive tree in Fig. 6.7.

- 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

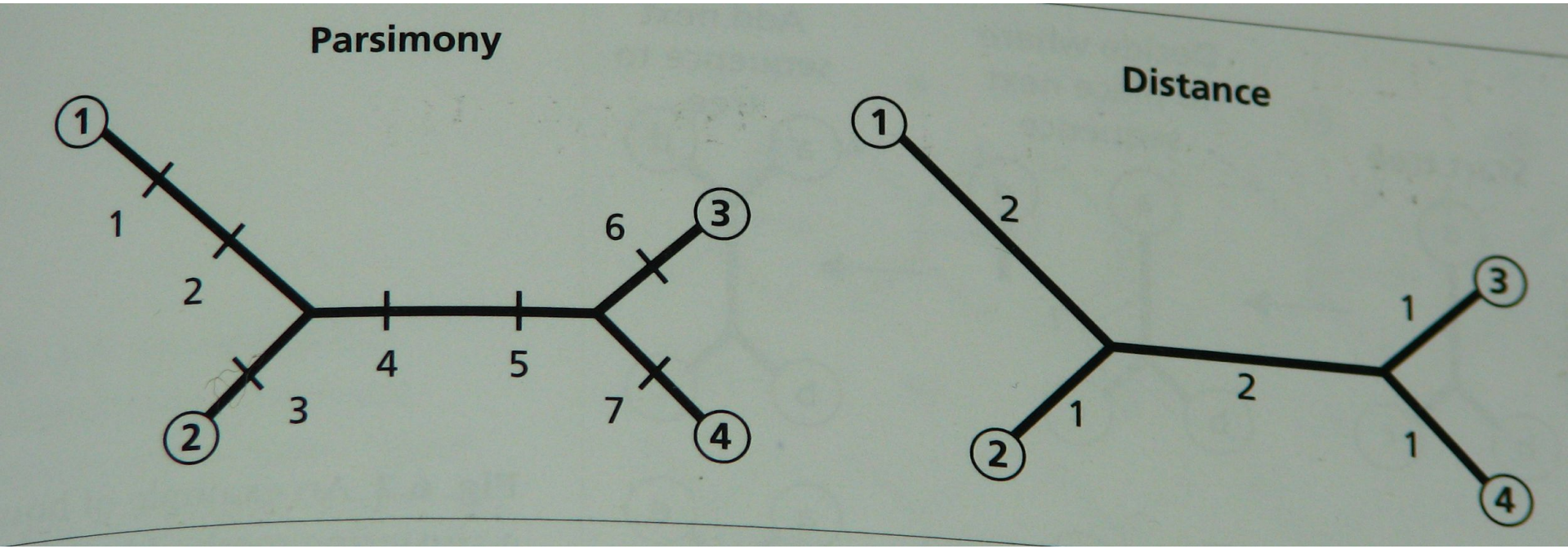
sites

	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

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

sequences



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)

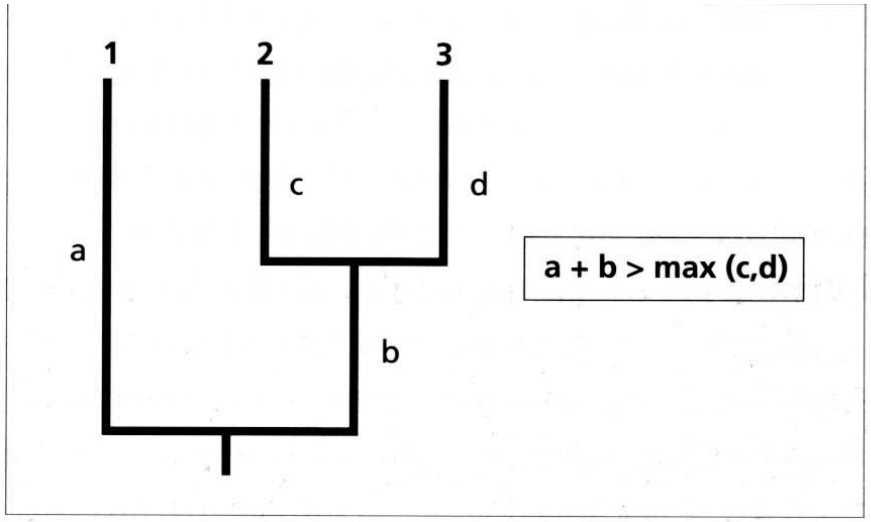


Fig. 6.26 The condition required for UPGMA to successfully reconstruct the true tree is that the sum of the edges leading to sequence 1 (a + b) must be greater than the larger of c and d. After Mooers *et al.* (1994).

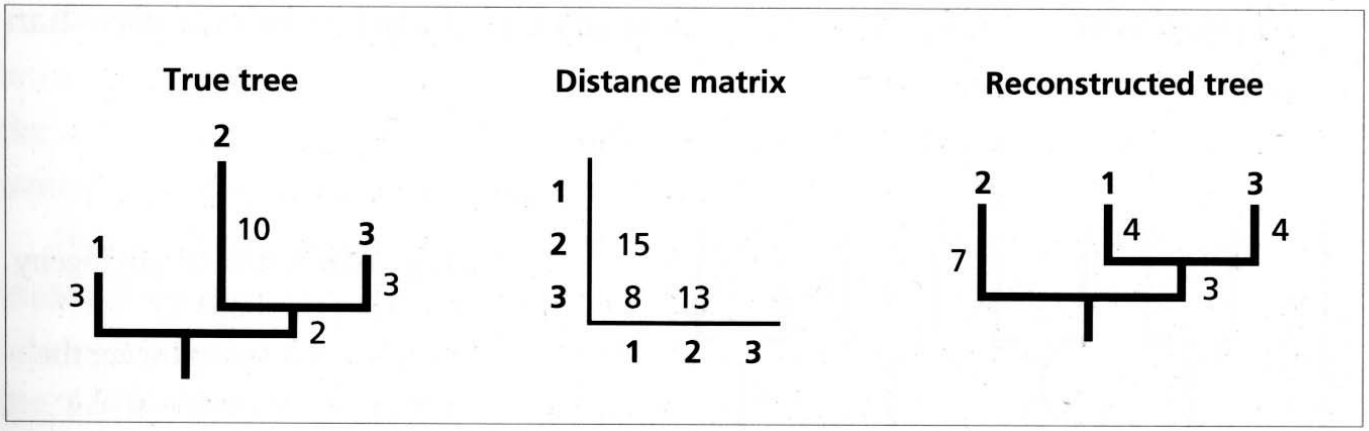


Fig. 6.27 An example where UPGMA will reconstruct the wrong tree. The edge lengths on the true tree violate the condition shown in Fig. 6.26, as a + b = 3 + 2 < max (c, d) = 10. Sequence 2 has evolved more rapidly than the other two sequences, so that sequence 1 and 3 are more similar to each other than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

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