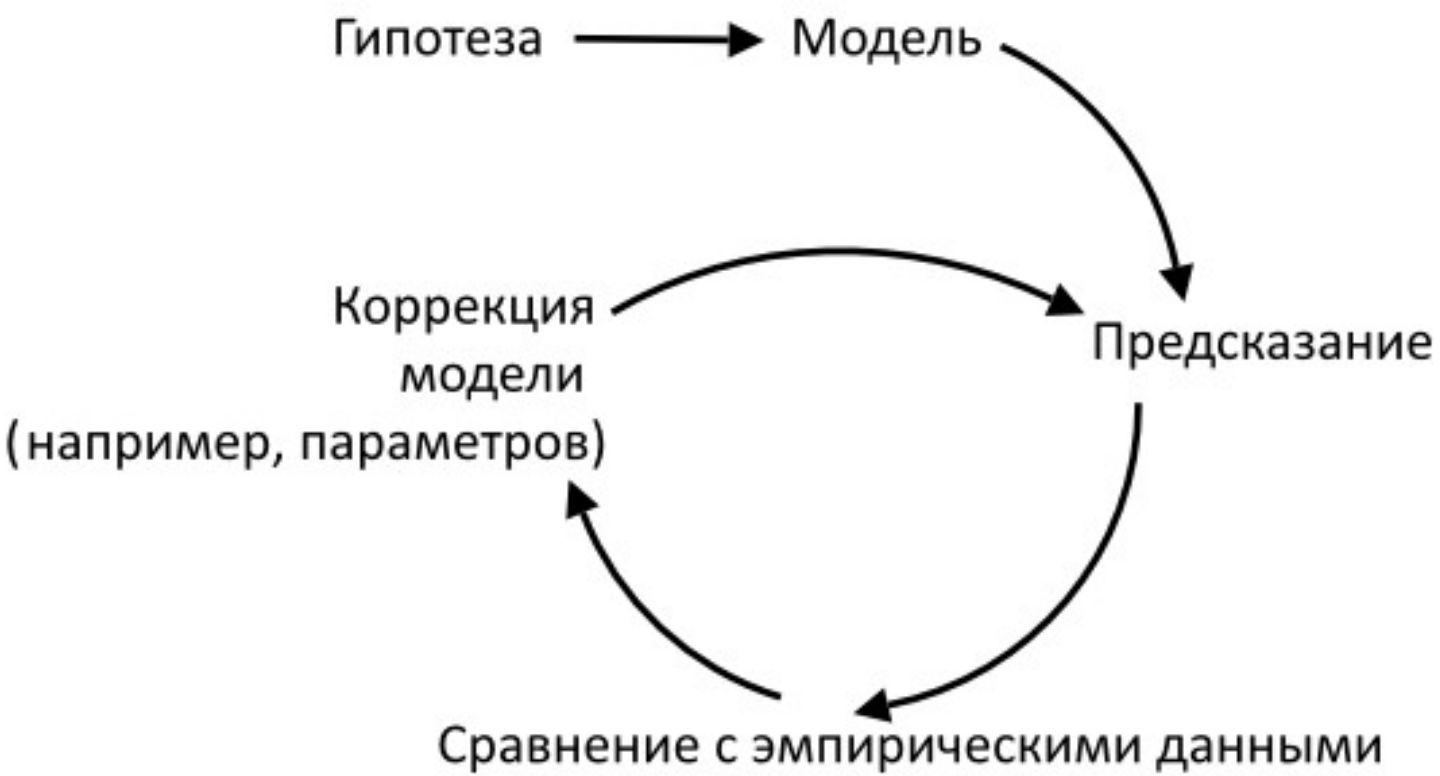
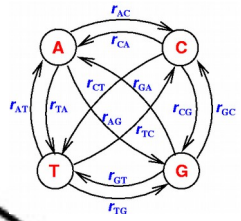
Метод максимального правдоподобия для филогенетики

Молекулярная филогенетика.
Лекция 6

Гипотеза, модель, данные



Данные, гипотеза и модель в контексте филогенетического дерева

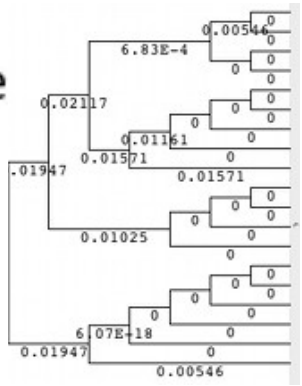


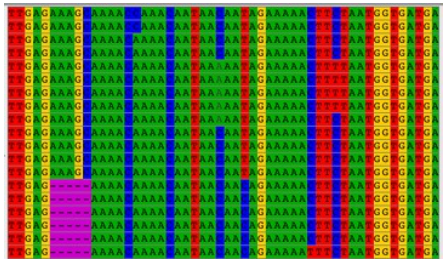
Гипотеза → Модель

Предсказание

Коррекция модели (например, параметров)

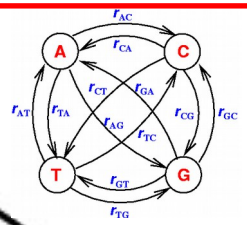
Сравнение с эмпирическими данными





Данные, гипотеза и модель в контексте филогенетического дерева

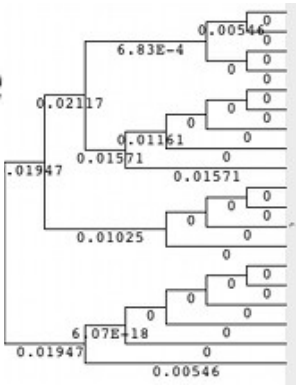
JC, HKY, GTR…



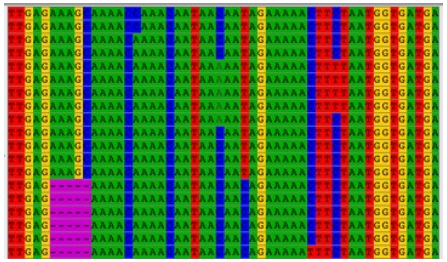
Гипотеза → Модель

Предсказание

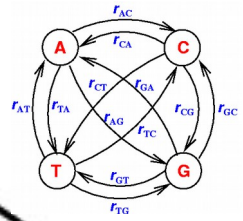
Коррекция модели (например, параметров)



Сравнение с эмпирическими данными



Данные, гипотеза и модель в контексте филогенетического дерева

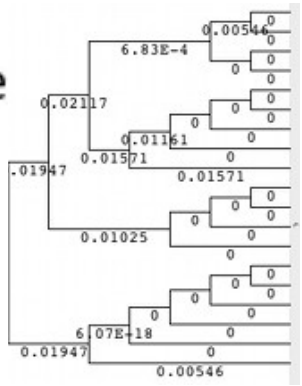


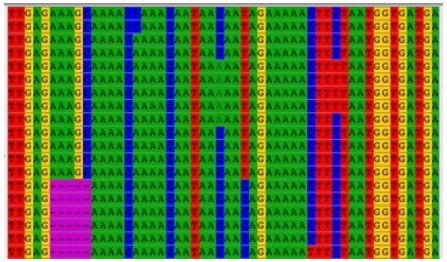
Гипотеза → Модель

Предсказание

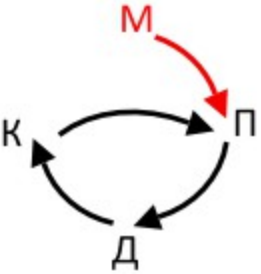
Сравнение с эмпирическими данными

Коррекция модели (например, параметров)



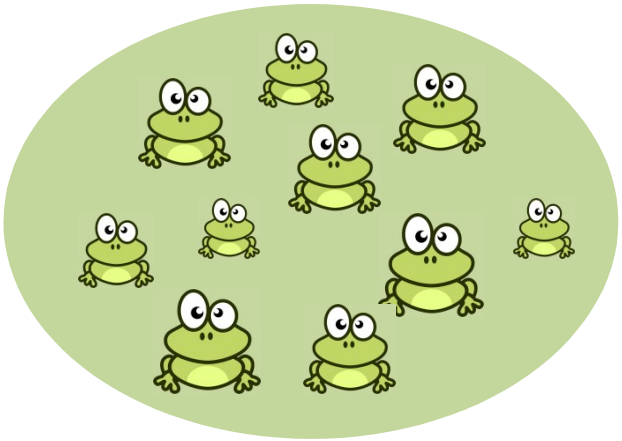


Максимальное правдоподобие; Цепи Маркова...



Статистическая модель: данные

Исходные данные: вес лягушек

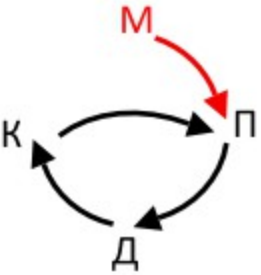


```
          [,1]
 [1,] 3.359673
 [2,] 2.796946
 [3,] 2.696572
 [4,] 3.314669
 [5,] 2.502077
 [6,] 2.804079
 [7,] 3.813665
 [8,] 2.323734
 [9,] 1.826915
[10,] 2.730826
```

Вопрос:

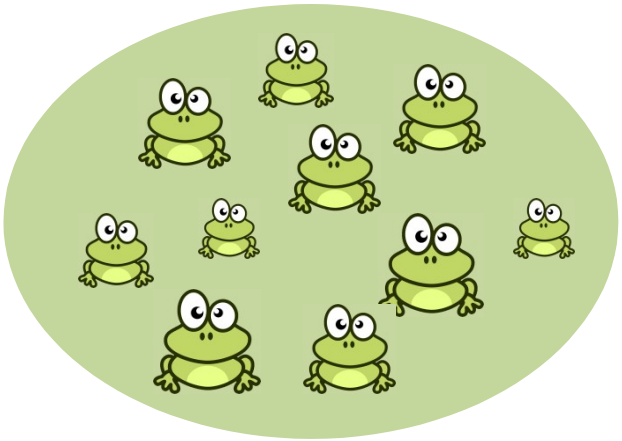
Каков средний вес лягушек?

Какова дисперсия веса лягушек?



Статистическая модель: данные

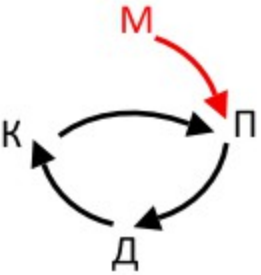
Исходные данные



	[,1]
[1,]	3.359673
[2,]	2.796946
[3,]	2.696572
[4,]	3.314669
[5,]	2.502077
[6,]	2.804079
[7,]	3.813665
[8,]	2.323734
[9,]	1.826915
[10,]	2.730826

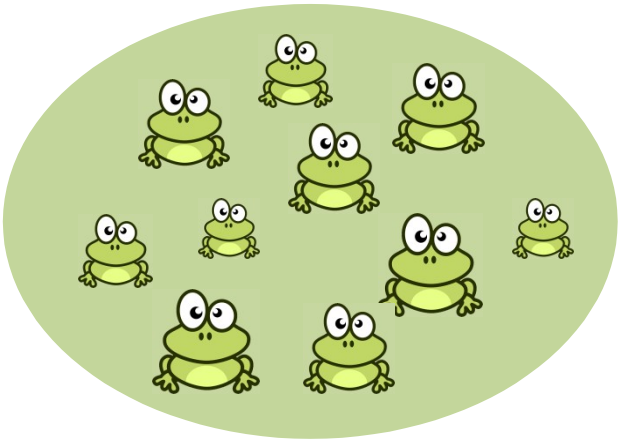
Решения:

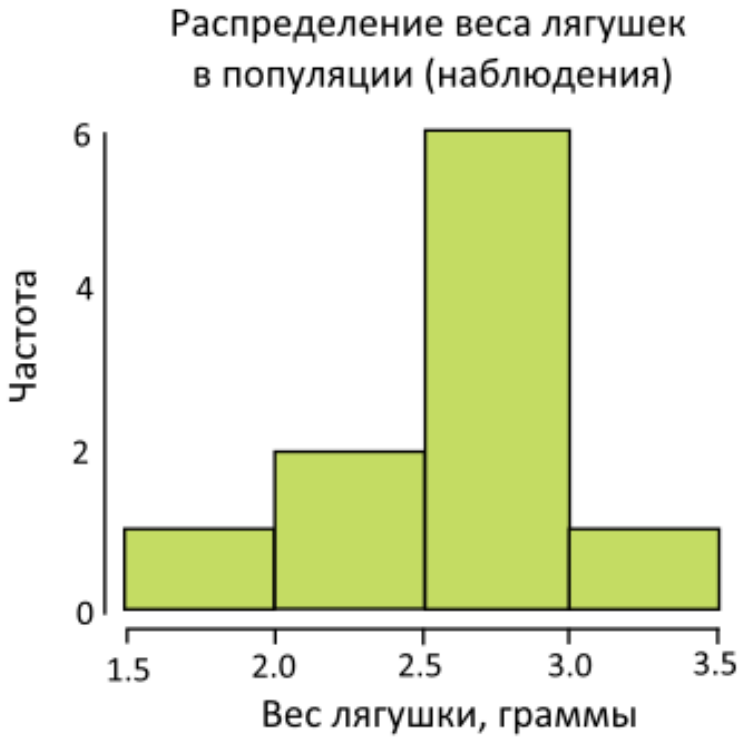
1. Посчитать вручную (среднее арифметическое, формула дисперсии)
2. Использовать модель и выяснить значение искомого параметра (среднего и дисперсии веса)

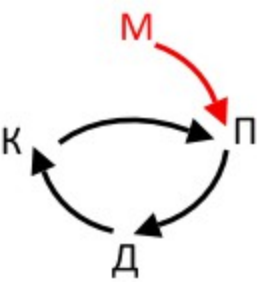


Статистическая модель: данные

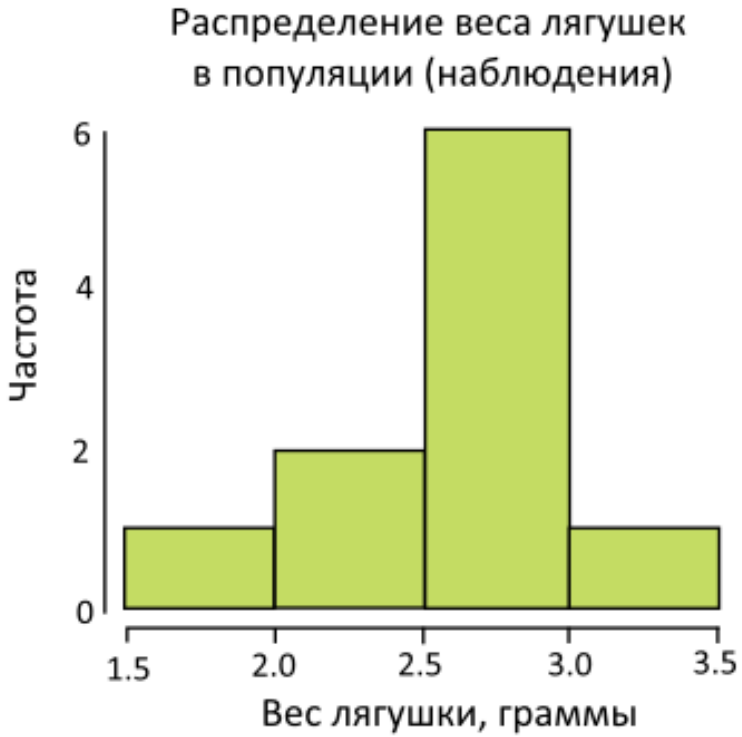
Исходные данные





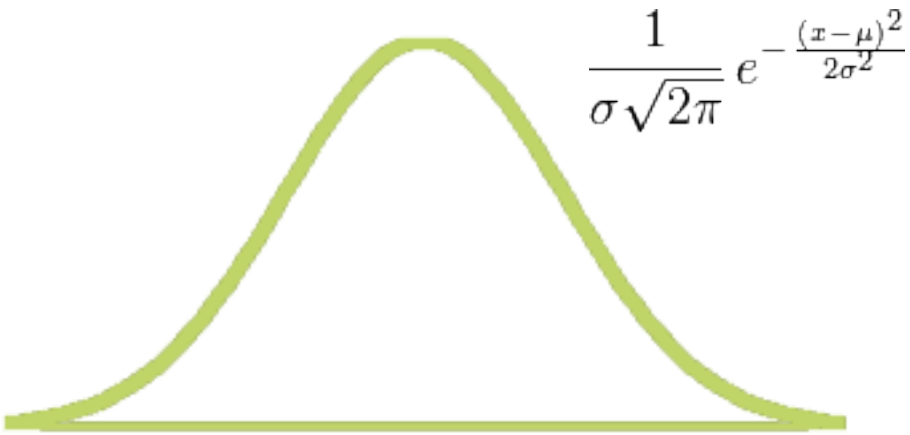


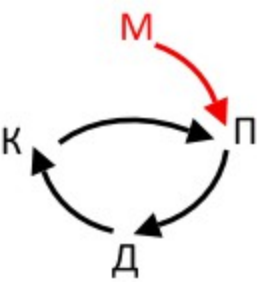
Статистическая модель: гипотеза



Наша гипотеза:

Модель, оптимально описывающая данное распредение – это нормальное распределение (мы могли бы взять любое другое!)





Статистическая модель: модель

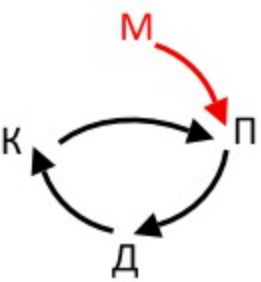
$$\frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

→

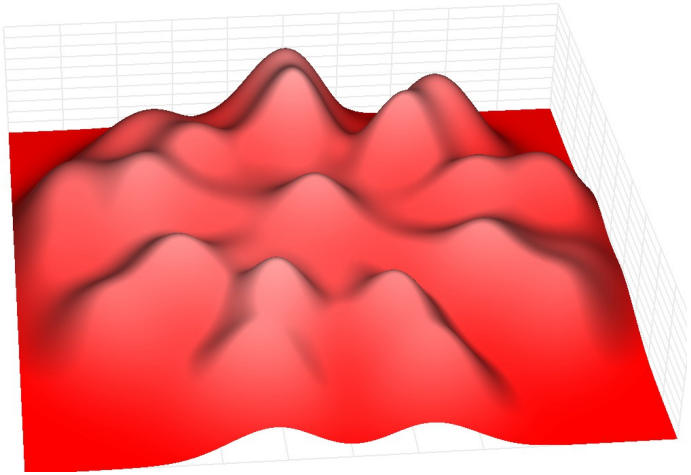
Модель:

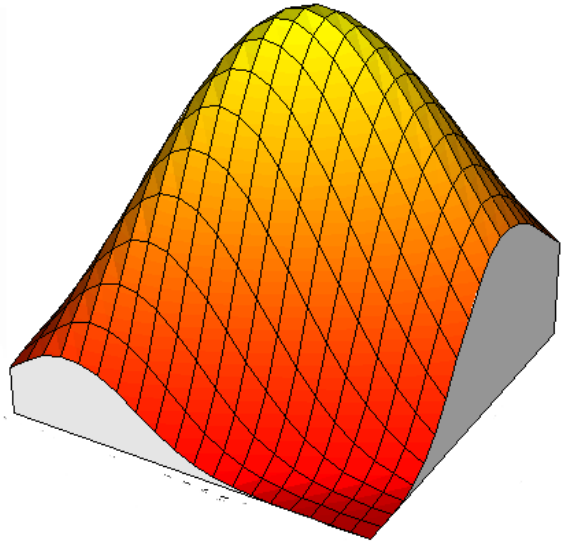
Нормальное распределение с неизвестными параметрами **σ** и **μ**

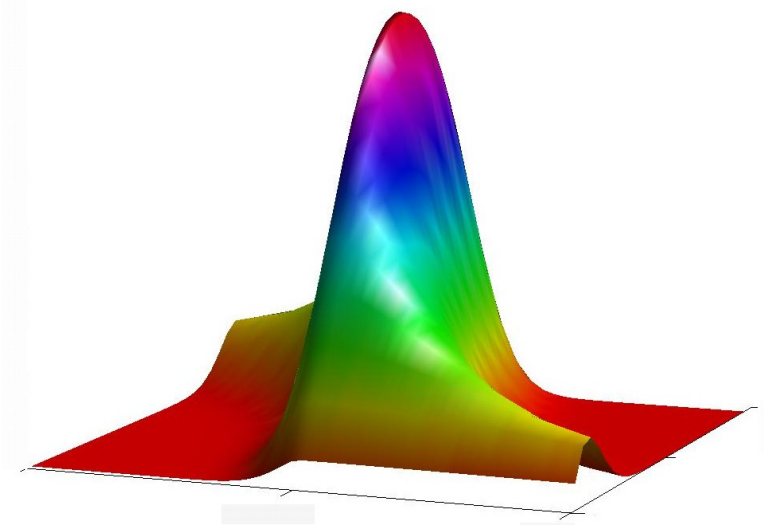
Как мы можем найти (подобрать) значения **σ** и **μ**?

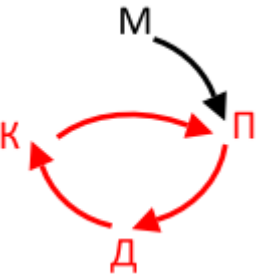


Примеры сложных распределений

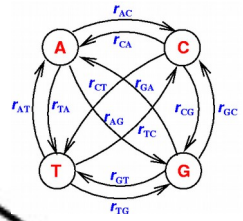








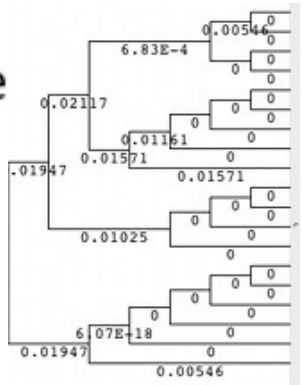
Данные, гипотеза и модель в контексте филогенетического дерева



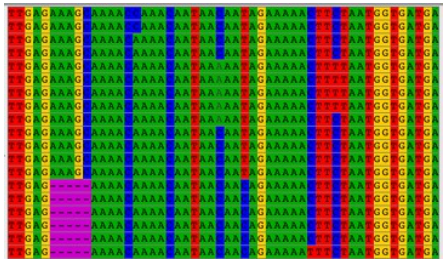
Гипотеза → Модель

Предсказание

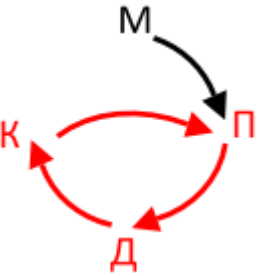
Коррекция модели (например, параметров)



Сравнение с эмпирическими данными



Максимальное правдоподобие;
Цепи Маркова...



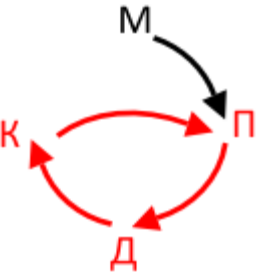
Как мы можем подобрать параметры модели?

Два наиболее часто используемых метода в филогенетике (и не только):

Максимальное правдоподобие

Монте-Карло с цепями Маркова

Максимальное правдоподобие



Максимальное правдоподобие

Fisher RA, 1922

L(Модель) = P(Данные|Модель + Параметры)

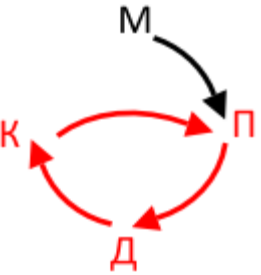
$$L = P(D|M)$$

Правдоподобие =
вероятность данных при заданной модели и параметрах

Максимальное правдоподобие =
такое значение параметров модели, при котором правдоподобие максимально

На практике чаще всего используется **логарифм правдоподобия**

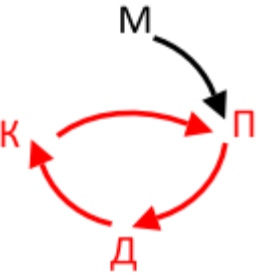
logL(Модель) = log(P(Данные|Модель + Параметры))



Как работает максимальное правдоподобие: простой пример

Данные: Серия бросков монетки: 7 орлов и 3 решки ($n = 10$, $x = 7$).

Вопрос: какова вероятность p выпадения орла при однократном подбрасывании монетки?



Как работает максимальное правдоподобие: простой пример

Данные: Серия бросков монетки: 7 орлов и 3 решки ($n = 10$, $x = 7$).

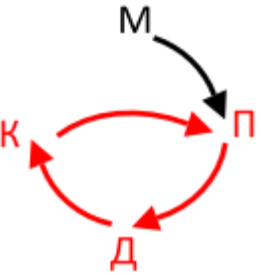
Вопрос: какова вероятность p выпадения орла при однократном подбрасывании монетки?

Модель: Биномиальное распределение

$$L(p) = \binom{n}{x} p^x (1-p)^{n-x}$$

Неизвестный параметр (вероятность выпадения орла):

$$0 < p < 1, p \in \mathbb{R}$$

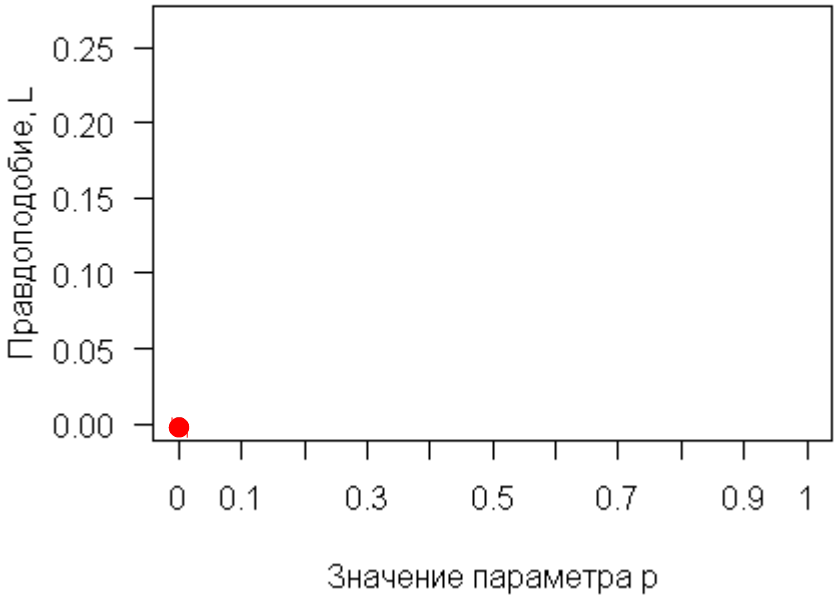


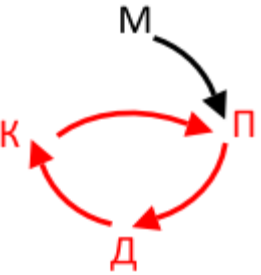
Максимальное правдоподобие: расчет вручную

$$L(p) = \binom{n}{x} p^x (1-p)^{n-x}$$

$p = 0.0$, $n = 10$, $x = 7$

$10!/(7! * 3!) * 0.1^7 + (1 - 0.1)^{(10-7)} = 0.000000000$



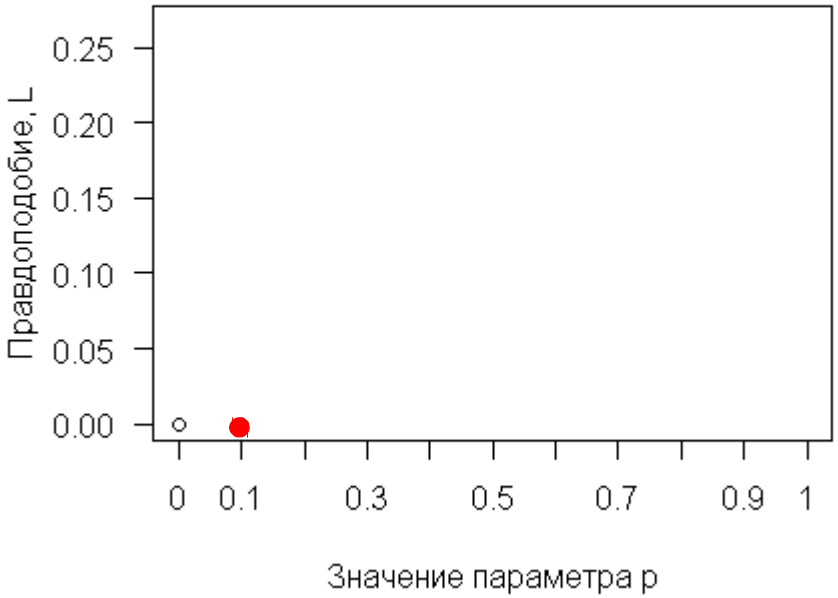


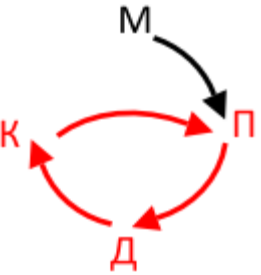
Максимальное правдоподобие: расчет вручную

$$L(p) = \binom{n}{x} p^x (1-p)^{n-x}$$

$p = 0.1$, $n = 10$, $x = 7$

$10!/(7! * 3!) * 0.1^7 + (1 - 0.1)^{(10-7)} = 0.000008748$



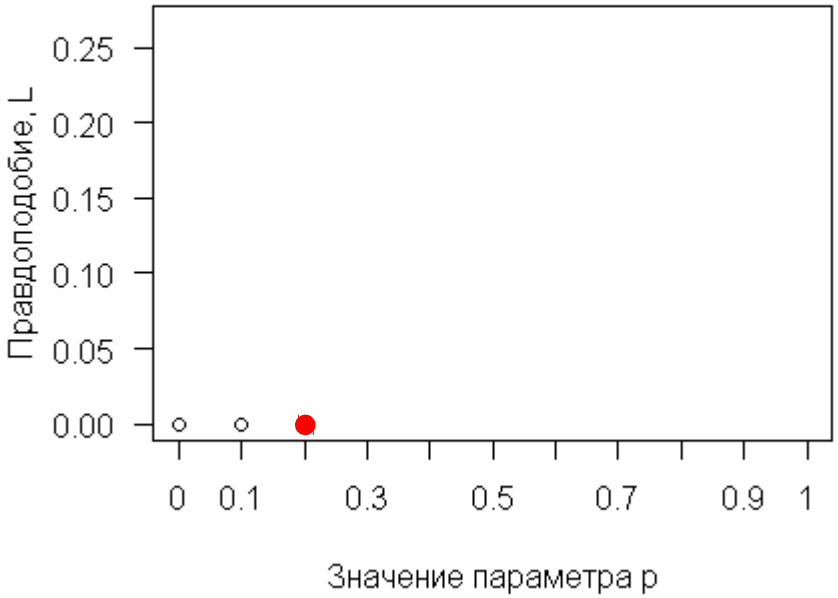


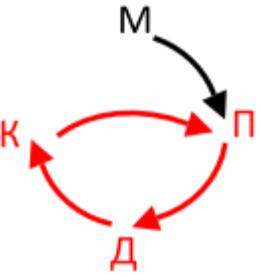
Максимальное правдоподобие: расчет вручную

$$L(p) = \binom{n}{x} p^x (1-p)^{n-x}$$

$p = 0.2$, $n = 10$, $x = 7$

$10!/(7! * 3!) * 0.1^7 + (1 - 0.1)^{(10-7)} = 0.000786432$



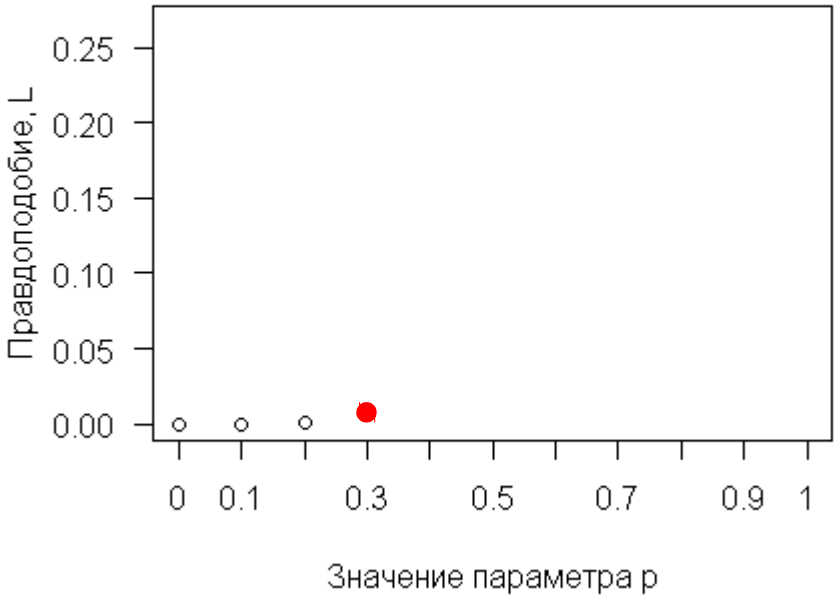


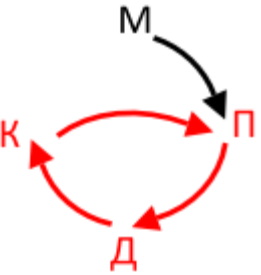
Максимальное правдоподобие: расчет вручную

$$L(p) = \binom{n}{x} p^x (1-p)^{n-x}$$

$p = 0.3$, $n = 10$, $x = 7$

$10!/(7! * 3!) * 0.1^7 + (1 - 0.1)^{(10-7)} = 0.009001692$



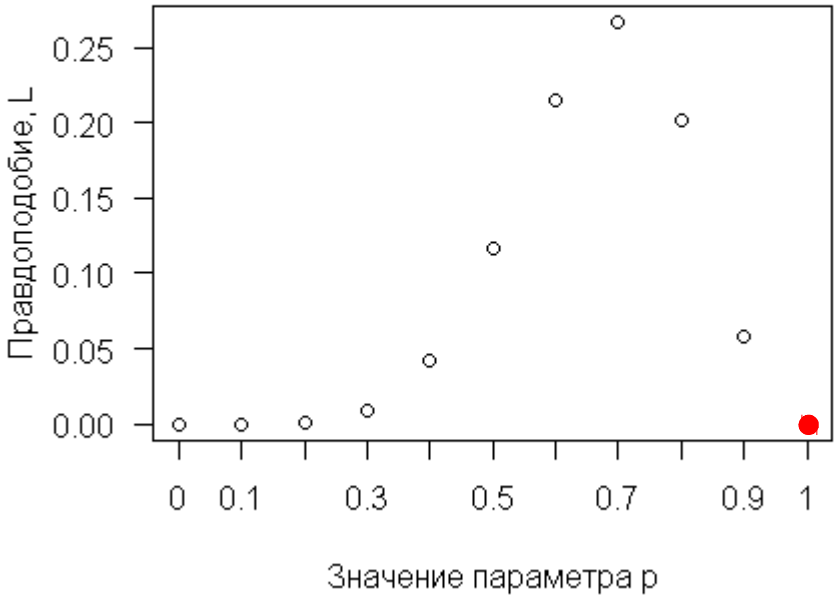


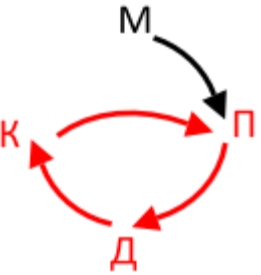
Максимальное правдоподобие: расчет вручную

$$L(p) = \binom{n}{x} p^x (1-p)^{n-x}$$

$p = 1$, $n = 10$, $x = 7$

$10!/(7! * 3!) * 0.1^7 + (1 - 0.1)^{(10-7)} = 0.000000000$



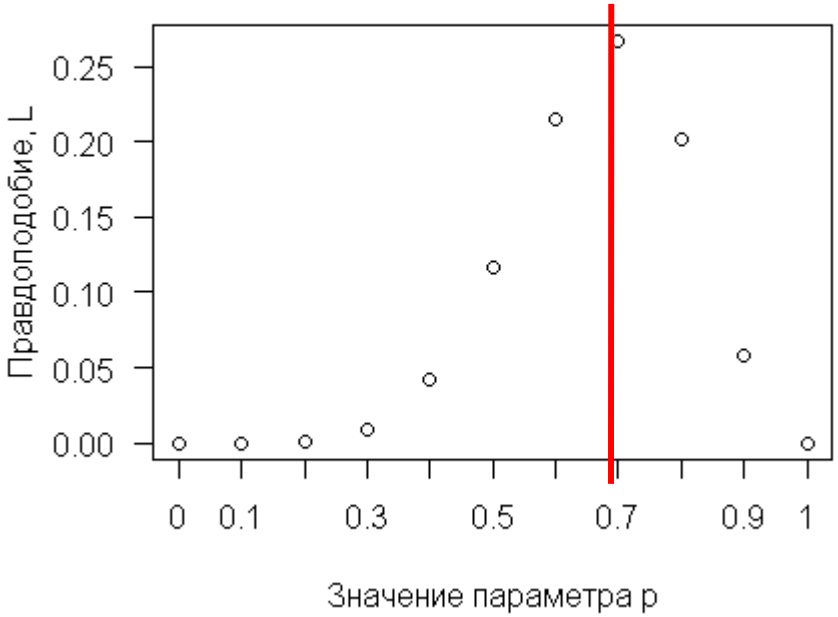


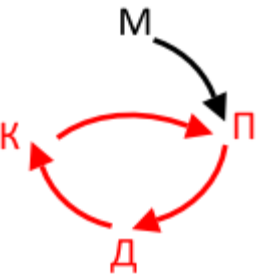
Максимальное правдоподобие: расчет вручную

$$L(p) = \binom{n}{x} p^x (1-p)^{n-x}$$

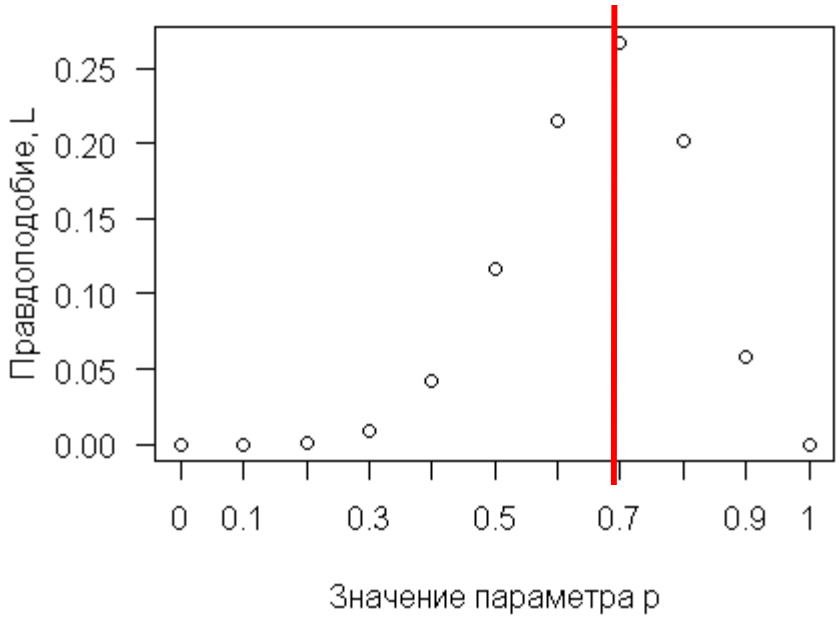
$p = 0.7$, $n = 10$, $x = 7$

$10!/(7! * 3!) * 0.1^7 + (1 - 0.1)^{(10-7)} = 0.266827932$



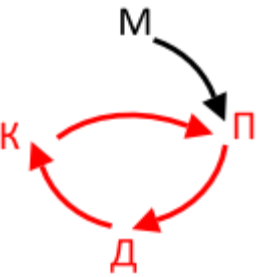


Максимальное правдоподобие: ответ и заключение

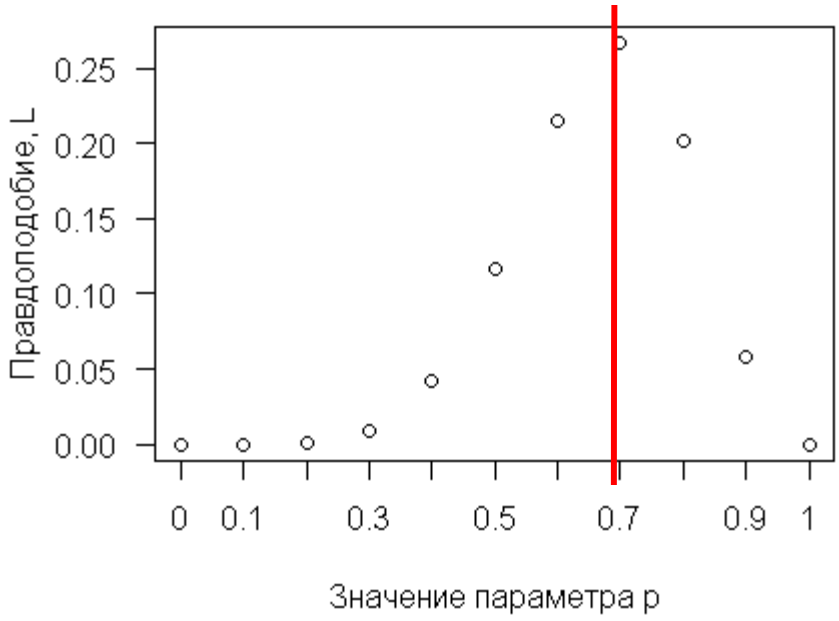


Вопрос:
какова вероятность выпадения орла при однократном подбрасывании монетки?

Ответ:



Максимальное правдоподобие: ответ и заключение



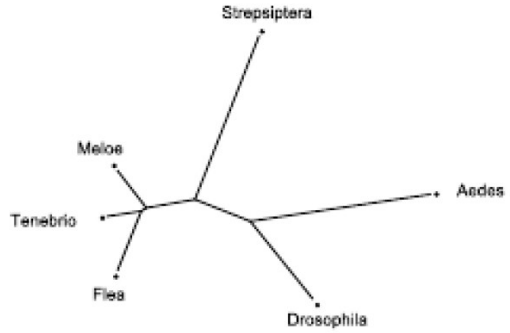
Вопрос:
какова вероятность выпадения орла при однократном подбрасывании монетки?

Ответ:

0.7

Максимальное правдоподобие для филогенетических деревьев

Максимальное правдоподобие для деревьев




```
Strepsiptera    AAGCTCATTAAATCGCTTTGGTTCCTTAGATAGTTGGAT...
Aedes           AGGCTCAGTATAACACTATAATTTACAAGATCATTGGAT...
Drosophila      AGGCTCATTATATCATTATGGTTCCTTAGATCGTTGGAT...
Flea            TGGCTCATTATATCATTATGGTTCATTAGATCGTTGGAT...
Meloe           AGGCTCATTAAATCATTATGGTTCCTTAGATCGTTGGAT...
Tenebrio        AGGCTCATTAAATCATTATGGTTCCTTAGATCGTTGGAT...
```

Дерево – это модель, состоящая из набора параметров: длин ветвей дерева, модели эволюции нуклеотидов и топологии

Мы хотим найти дерево, которое дает нам наибольшее правдоподобие при известных нуклеотидных последовательностях

L(Модель) = P(Выровненные последовательности | Модель эволюции нуклеотидов, Длины ветвей, Топология)

$$L(Q,B,T) = Prob(D|Q,B,T)$$

$L(Q, B, T) = Prob(D|Q, B, T)$

Максимальное правдоподобие для деревьев

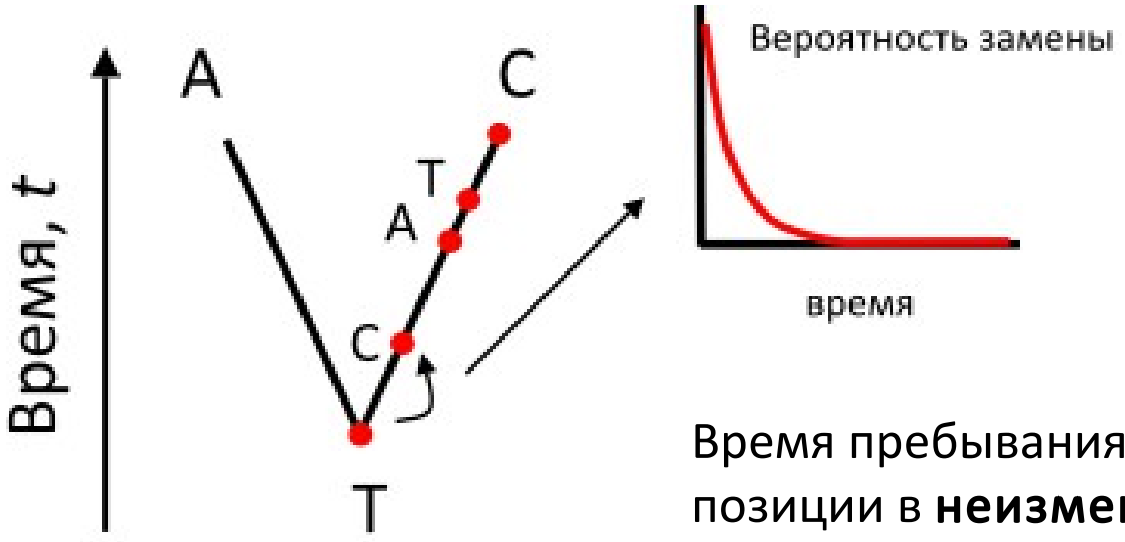
L(Модель) = P(Выровненные последовательности | Модель эволюции нуклеотидов, Длины ветвей, Топология)

```
Strepsiptera    AAGCTCATTAAATCGCTTTGGTTCCTTAGATAGTTGGAT...
Aedes           AGGCTCAGTATAACACTATAATTTACAAGATCATTGGAT...
Drosophila      AGGCTCATTATATCATTATGGTTCCTTAGATCGTTGGAT...
Flea            TGGCTCATTATATCATTATGGTTCATTAGATCGTTGGAT...
Meloe           AGGCTCATTAAATCATTATGGTTCCTTAGATCGTTGGAT...
Tenebrio        AGGCTCATTAAATCATTATGGTTCCTTAGATCGTTGGAT...
```

$L(Q,B,T) = Prob(D|Q,B,T)$

Максимальное правдоподобие для деревьев

L(Модель) = P(Выровненные последовательности | Модель эволюции нуклеотидов, Длины ветвей, Топология)

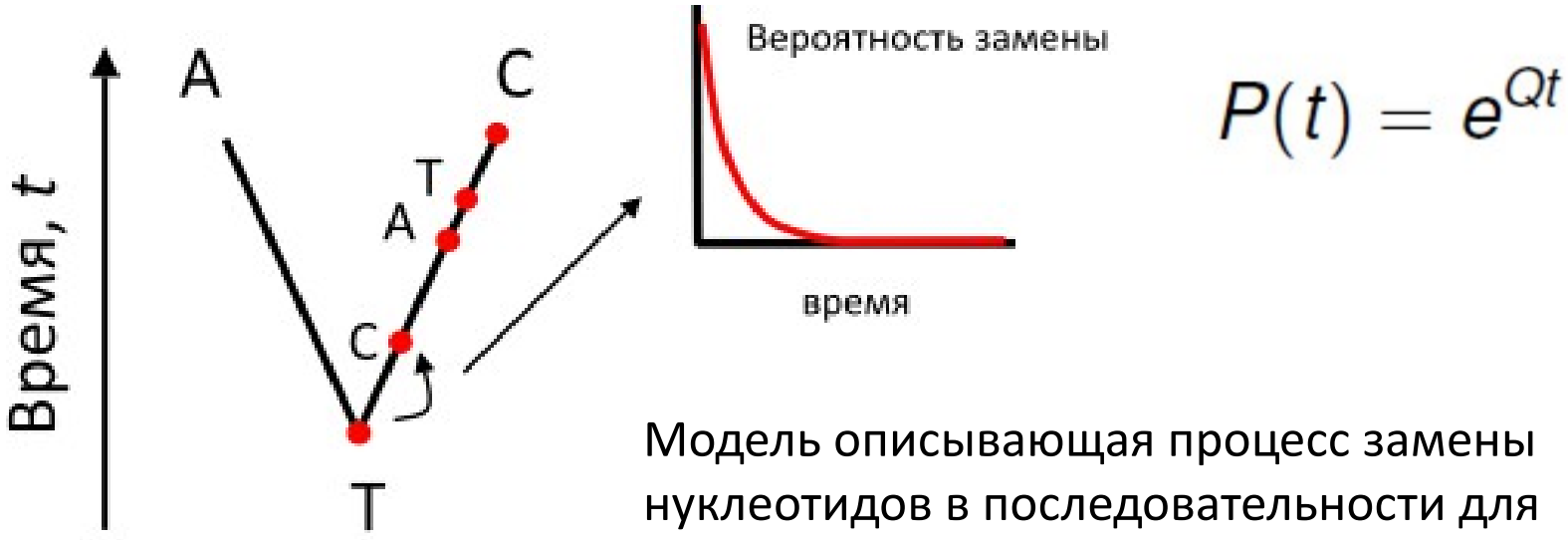


Время пребывания нуклеотидной позиции в **неизменном** состоянии распределено **экспоненциально**

$L(Q,B,T) = Prob(D|Q,B,T)$

Максимальное правдоподобие для деревьев

L(Модель) = P(Выровненные последовательности| Модель эволюции нуклеотидов, Длины ветвей, Топология)



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

Модель описывающая процесс замены нуклеотидов в последовательности для четырех состояний - процесс Маркова в непрерывном времени

$L(Q,B,T) = Prob(D|Q,B,T)$

Максимальное правдоподобие для деревьев

L(Модель) = P(Выровненные последовательности | Модель эволюции нуклеотидов, Длины ветвей, Топология)

Матрица мгновенных переходов Q

$$P(t) = e^{Qt} \qquad Q = \begin{bmatrix} - & r_{AC} & r_{AG} & r_{AT} \\ r_{CA} & - & r_{CG} & r_{CT} \\ r_{GA} & r_{GC} & - & r_{GT} \\ r_{TA} & r_{TC} & r_{TG} & - \end{bmatrix}$$

$$Q = \begin{bmatrix} - & 1 & \kappa & 1 \\ 1 & - & 1 & \kappa \\ \kappa & 1 & - & 1 \\ 1 & \kappa & 1 & - \end{bmatrix}$$

Kimura 2-parameters, 1981

$$Q = \begin{bmatrix} - & \pi_C & \pi_G & \pi_T \\ \pi_A & - & \pi_G & \pi_T \\ \pi_A & \pi_C & - & \pi_T \\ \pi_A & \pi_C & \pi_G & - \end{bmatrix}$$

Felsenstein, 1981

$$Q = \begin{bmatrix} - & \pi_C & \kappa\pi_G & \pi_T \\ \pi_A & - & \pi_G & \kappa\pi_T \\ \kappa\pi_A & \pi_C & - & \pi_T \\ \pi_A & \kappa\pi_C & \pi_G & - \end{bmatrix}$$

Hasegawa-Kishino-Yano, 1985

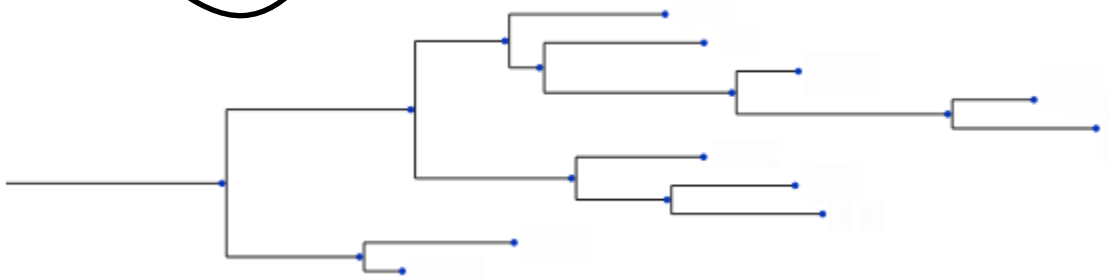
$L(Q,B,T) = Prob(D|Q,B,T)$

Максимальное правдоподобие для деревьев

L(Модель) = P(Выровненные последовательности|
Модель эволюции нуклеотидов, Длины ветвей, Топология)

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

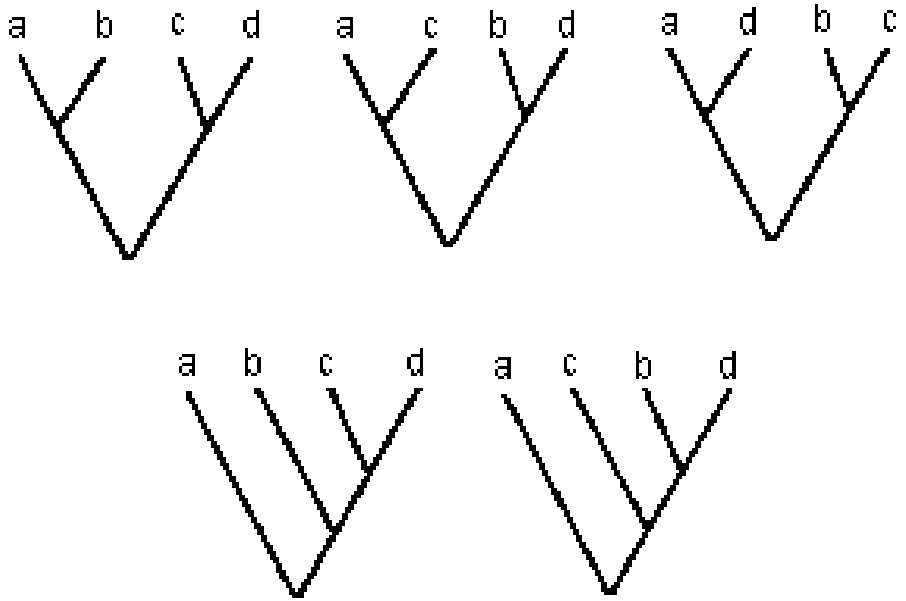
Длины ветвей деревьев



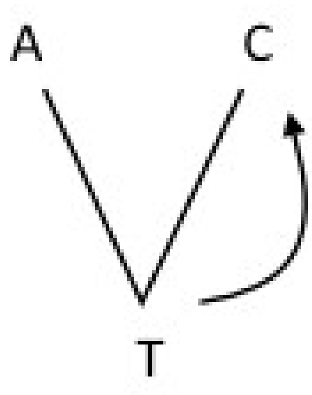
$L(Q, B, T) = Prob(D|Q, B, T)$

Максимальное правдоподобие для деревьев

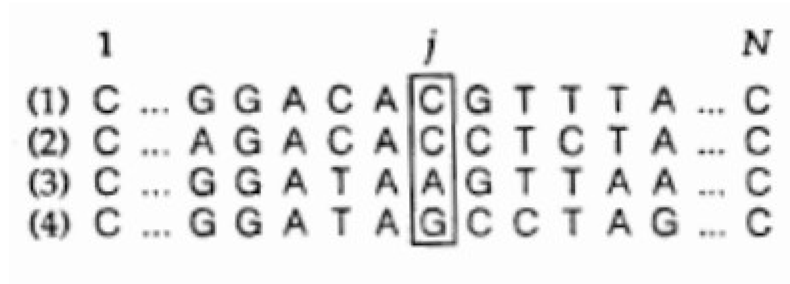
L(Модель) = P(Выровненные последовательности| Модель эволюции нуклеотидов, Длины ветвей, Топология)



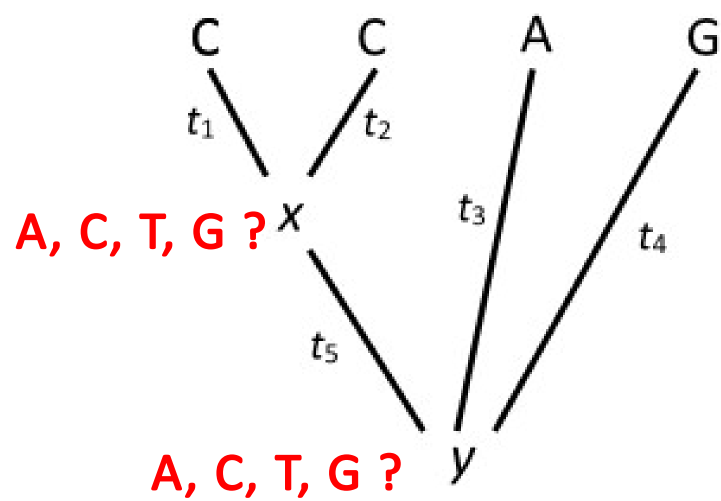
Максимальное правдоподобие для деревьев



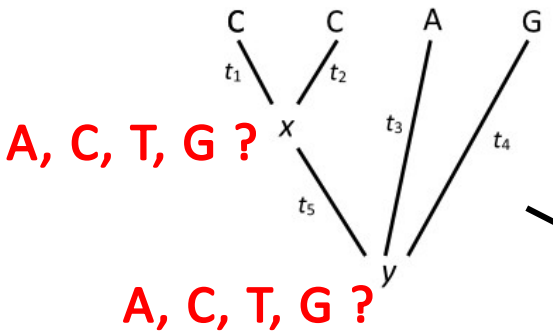
$$P(t) = e^{Qt}$$



Состояния в узлах дерева неизвестны, поэтому мы перебираем все возможные состояния в узлах



Максимальное правдоподобие для деревьев



$$L_{(j)} = \text{Prob}\left(\begin{array}{cccc} C & C & A & G \\ & A & & \\ & & A & \end{array}\right) + \text{Prob}\left(\begin{array}{cccc} C & C & A & G \\ & C & & \\ & & A & \end{array}\right)$$

$$+ \quad \ldots \quad + \text{Prob}\left(\begin{array}{cccc} C & C & A & G \\ & G & & \\ & & C & \end{array}\right)$$

$$+ \quad \ldots \quad + \text{Prob}\left(\begin{array}{cccc} C & C & A & G \\ & T & & \\ & & T & \end{array}\right)$$

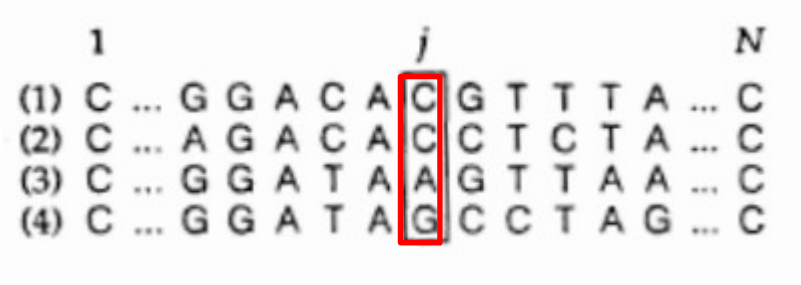
Максимальное правдоподобие для деревьев

$$P(D_j|Q,B,T) = \sum_{x}^{ACTG} \sum_{y}^{ACTG} P(C,C,A,G,x,y|Q,B,T)$$

$$P(D_j|Q,B,T) = \sum_{x}^{ACTG} \sum_{y}^{ACTG} P(y)P_{yx}(t_5)P_{xC}(t_1)P_{xC}(t_2)P_{yA}(t_3)P_{yG}(t_4)$$

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

Максимальное правдоподобие для деревьев



$$P(D|Q,B,T) = \sum_{j=1}^{N\ sites} P(D_j|Q,B,T)$$

Максимальное правдоподобие для деревьев

Заранее мы не знаем параметры матрицы переходов Q, длины ветвей деревьев B, и топологию дерева T. Поэтому:

1. Процесс начинается со случайных значений

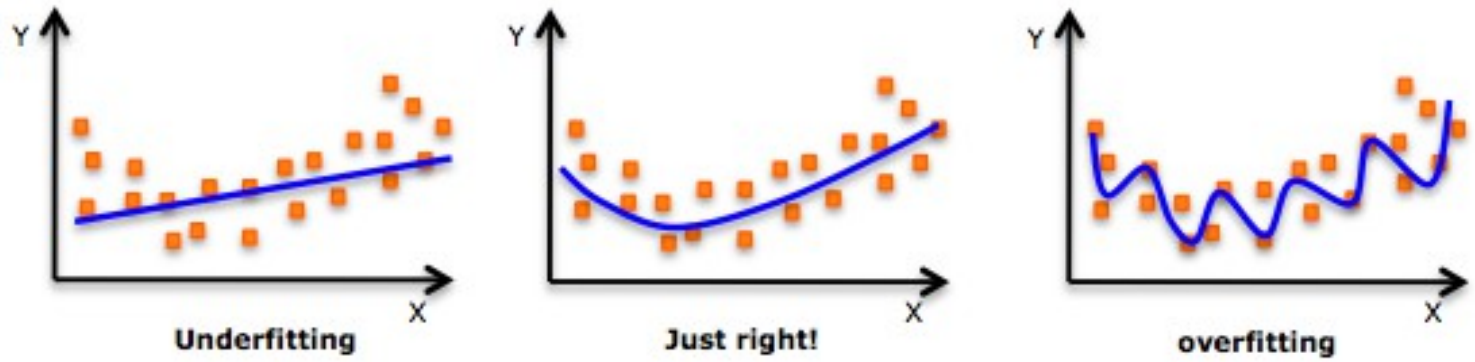
2. На каждой итерации параметры слегка изменяются (либо Q, либо B, либо T) и сравнивается правдоподобие текущего и предыдущего шага

3. Процесс повторяется до тех пор, как значение правдоподобия не достигнет максимального

$$P(D|Q,B,T) = \sum_{j=1}^{N\ sites} P(D_j|Q,B,T)$$

Максимальное правдоподобие и сравнение вложенных и невложенных моделей

Модель и данные



Вложенные (nested) и невложенные (non-nested) модели

$$y = \beta_0 + \beta_1 x_1$$
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x2$$

$$y = \beta_0 + \beta_1 x_1$$
$$y = \beta_1 x_1^2 + \beta_2 x_1 x_2$$

Сравнение вложенных моделей: Тест отношения правдоподобия (LRT)

Сравнение вложенных моделей: Тест отношения правдоподобия (LRT)

$$Q = \begin{bmatrix} - & \pi_C & \pi_G & \pi_T \\ \pi_A & - & \pi_G & \pi_T \\ \pi_A & \pi_C & - & \pi_T \\ \pi_A & \pi_C & \pi_G & - \end{bmatrix}$$

Felsenstein, 1981

?

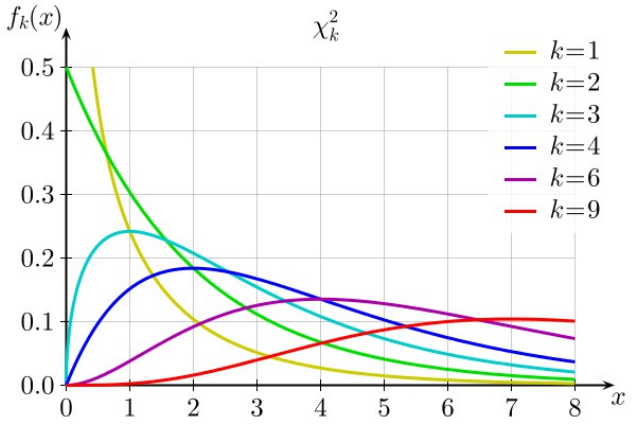
$$Q = \begin{bmatrix} - & \pi_C & \kappa\pi_G & \pi_T \\ \pi_A & - & \pi_G & \kappa\pi_T \\ \kappa\pi_A & \pi_C & - & \pi_T \\ \pi_A & \kappa\pi_C & \pi_G & - \end{bmatrix}$$

Hasegawa-Kishino-Yano, 1985

Сравнение вложенных моделей: Тест отношения правдоподобия (LRT)

$$D = -2\ln(\text{максимальное правдоподобие для модели } H0) + 2\ln(\text{максимальное правдоподобие для модели } H1)$$

$$2\left[lnL(\hat{\theta}) - lnL(\theta_0)\right] \approx \frac{(\theta_0 - \hat{\theta})^2}{\sigma} = \chi_1^2$$



Сравнение вложенных моделей: Тест отношения правдоподобия (LRT)

$$D = -2\ln(\text{максимальное правдоподобие для модели } H0) + 2\ln(\text{максимальное правдоподобие для модели } H1)$$

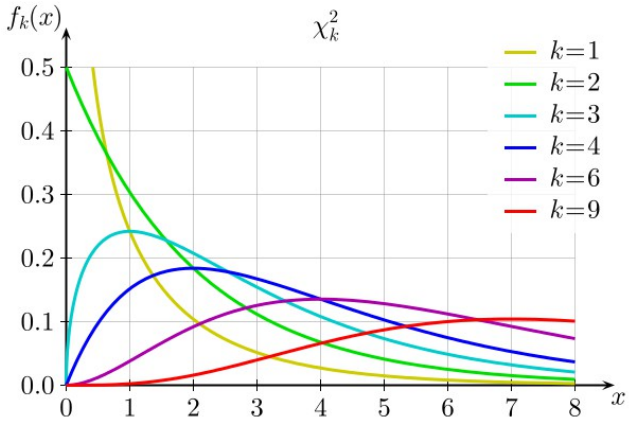
$$2\left[lnL(\hat{\theta}) - lnL(\theta_0)\right] \approx \frac{(\theta_0 - \hat{\theta})^2}{\sigma} = \chi_1^2$$

$$\begin{bmatrix} - & \pi_C & \pi_G & \pi_T \\ \pi_A & - & \pi_G & \pi_T \\ \pi_A & \pi_C & - & \pi_T \\ \pi_A & \pi_C & \pi_G & - \end{bmatrix}$$

F81

$$\begin{bmatrix} - & \pi_C & \kappa\pi_G & \pi_T \\ \pi_A & - & \pi_G & \kappa\pi_T \\ \kappa\pi_A & \pi_C & - & \pi_T \\ \pi_A & \kappa\pi_C & \pi_G & - \end{bmatrix}$$

HKY



Сравнение вложенных моделей: LRT Пример

$$2\left[lnL(\hat{\theta}) - lnL(\theta_0)\right] \approx \frac{(\theta_0 - \hat{\theta})^2}{\sigma} = \chi_1^2$$

F81, n = 3

$$\begin{bmatrix} - & \pi_C & \pi_G & \pi_T \\ \pi_A & - & \pi_G & \pi_T \\ \pi_A & \pi_C & - & \pi_T \\ \pi_A & \pi_C & \pi_G & - \end{bmatrix}$$

logL = -1523.25

HKY, n = 4

$$\begin{bmatrix} - & \pi_C & \kappa\pi_G & \pi_T \\ \pi_A & - & \pi_G & \kappa\pi_T \\ \kappa\pi_A & \pi_C & - & \pi_T \\ \pi_A & \kappa\pi_C & \pi_G & - \end{bmatrix}$$

logL = -1521.14

Degrees of Freedom	0.10	0.05	0.025	0.01
1	2.706	3.841	5.024	6.635
2	4.605	5.991	7.378	9.210
3	6.251	7.815	9.348	11.345
4	7.779	9.488	11.143	13.277
5	9.236	11.071	12.833	15.086
6	10.645	12.592	14.449	16.812
7	12.017	14.067	16.013	18.475
8	13.362	15.507	17.535	20.090
9	14.684	16.919	19.023	21.666
10	15.987	18.307	20.483	23.209

D = 2 * (-1521.14 – (-1523.25) = 4.22

d.f. = 4 - 3 = 1

p = 0.05, значимый порог = 3.841

Сравнение невложенных моделей: Информационный критерий Акаике (AIC)

Информационный критерий Акаике (AIC)

$$AIC = 2k - 2\ln(L)$$

K – число параметров в модели

F81, n = 3

$$\begin{bmatrix} - & \pi_C & \pi_G & \pi_T \\ \pi_A & - & \pi_G & \pi_T \\ \pi_A & \pi_C & - & \pi_T \\ \pi_A & \pi_C & \pi_G & - \end{bmatrix}$$

logL = -1523.25

AIC = 2*3 – 2 * (-1523.25)

AIC = 3052.5

HKY, n = 4

$$\begin{bmatrix} - & \pi_C & \kappa\pi_G & \pi_T \\ \pi_A & - & \pi_G & \kappa\pi_T \\ \kappa\pi_A & \pi_C & - & \pi_T \\ \pi_A & \kappa\pi_C & \pi_G & - \end{bmatrix}$$

logL = -1521.14

AIC = 2*4 – 2 * (-1521.14)

AIC = 3050.28

Разница AIC = 3052.5 - 3050.28 = 2.22

Правило:

< 2 модели неотличимы.

2 – 4 альтернативная модель имеет умеренную поддержку

4 – 6 альтернативная модель имеет сильную поддержку

> 6 – альтернативная модель предпочтительна

Akaike, 1978, 1979

Критерий Акаике скорректированный по размеру выборки и веса Акаике (AIC)

$$AIC_c = -2\,ln(L(\theta \mid y)) + 2K\left(\frac{n}{n-K-1}\right)$$

$$w_i = \frac{exp(-0.5\Delta_i)}{\sum_{r=1}^{N} exp(-0.5\Delta_r)} \qquad \text{where} \quad \Delta_i = AIC_i - AIC_{min}$$

Akaike, 1978, 1979