Угадай модель эволюции

Выполнил: Артемьев Святослав

Научный руководитель: Дмитрий Биба

Введение

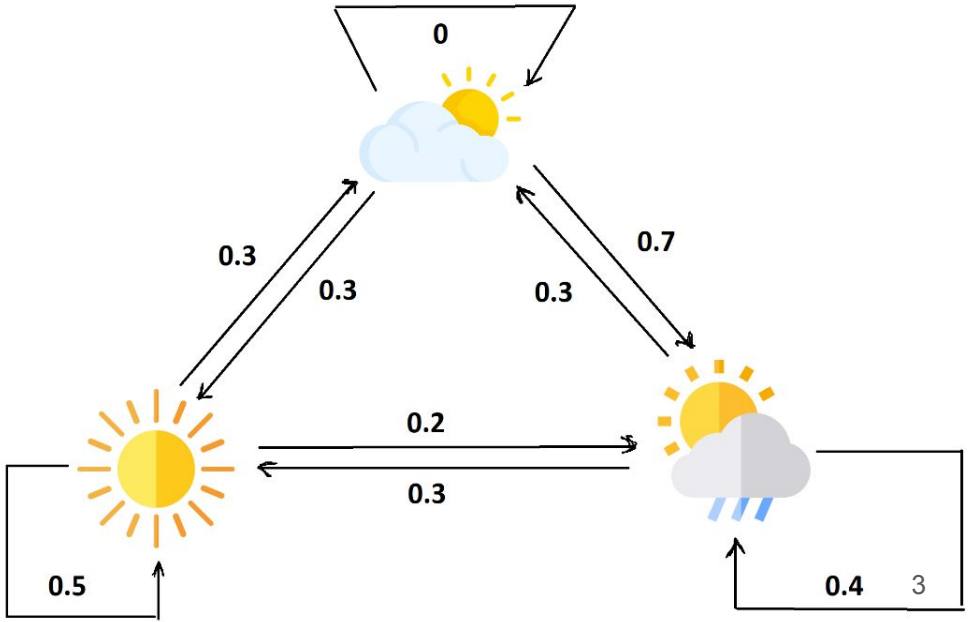
Модель эволюции представляет собой таблицу размером 4x4, где каждая ячейка содержит число, пропорциональное интенсивности замен каждого нуклеотида на другой. Например, если все нуклеотиды заменяются с одинаковой частотой, то все числа в таблице будут одинаковыми. Если же гуанин чаще всего заменяется на аденин, то число в соответствующей ячейке будет выше, чем в остальных. Знание модели эволюции может быть полезным для определения риска возникновения мутаций у пациента в случае рака определенного типа.

Задачи:

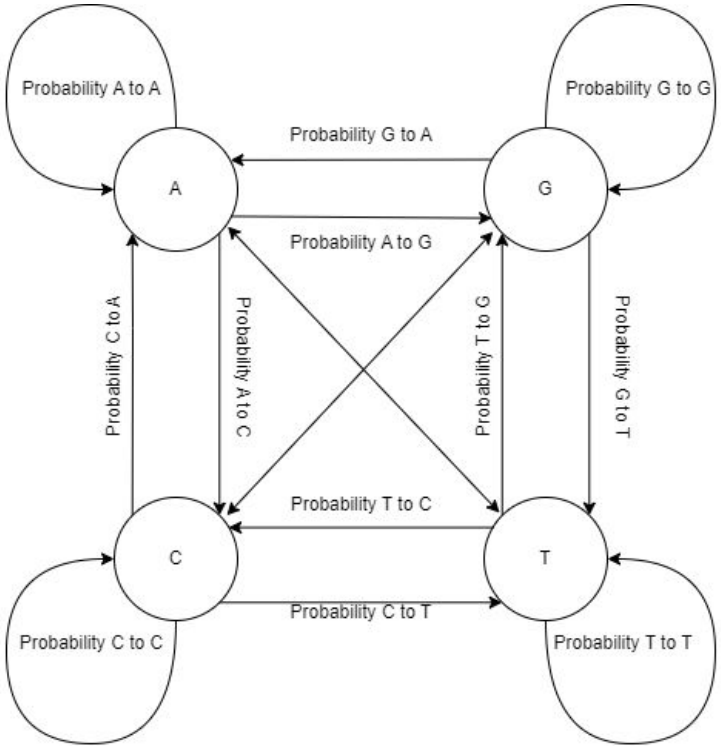
- Написать программу, которая будет изменять последовательность ДНК согласно заданной модели эволюции.
- Написать программу, которая будет угадывать модель эволюции на основании следующей информации: последовательность ДНК предка, последовательность ДНК потомка, и время, на протяжении которого эта последовательность эволюционировала.
- Оценить работу программы-предиктора.

Цепи Маркова. База проекта

$$P = \begin{bmatrix} 0.5 & 0.3 & 0.2 \\ 0.3 & 0 & 0.7 \\ 0.3 & 0.3 & 0.4 \end{bmatrix}$$



Цепи Маркова. В нашем проекте



В нашей работе мы будем использовать модификацию цепи Маркова, основанную на непрерывном времени. Для такое цепи характерно следующее:

1) Наличией Q-матрицы интенсивностей. Пример:

$$Q = \begin{array}{c} \\ A \\ G \\ C \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & G & C & T \end{array} \\ \begin{bmatrix} -3 & 1 & 1 & 1 \\ 1 & -3 & 1 & 1 \\ 1 & 1 & -3 & 1 \\ 1 & 1 & 1 & -3 \end{bmatrix} \end{array}$$

2) P-матрица высчитывается как: $P(t) = e^{Qt}$

Модель Джукса-Кантора (JC69)

Простейшая из всех моделей. Допущения модели состоят в том, что:

1. Интенсивности замен всех нуклеотидов равны
2. Равновесные частоты всех нуклеотидов равны

$$Q = \begin{bmatrix} -\frac{3}{4} * \mu & \mu * \frac{1}{4} & \mu * \frac{1}{4} & \mu * \frac{1}{4} \\ \mu * \frac{1}{4} & -\frac{3}{4} * \mu & \mu * \frac{1}{4} & \mu * \frac{1}{4} \\ \mu * \frac{1}{4} & \mu * \frac{1}{4} & -\frac{3}{4} * \mu & \mu * \frac{1}{4} \\ \mu * \frac{1}{4} & \mu * \frac{1}{4} & \mu * \frac{1}{4} & -\frac{3}{4} * \mu \end{bmatrix}$$

Модель Кимуры (K81)

Трехпараметрическая модель. Допущения модели:

1. Разные показатели интенсивностей переходов для транзиций и 2-х типов трансверсий.
    a. Переменная альфа отвечает за показатель интенсивностей для транзиций, то есть переходов A<—>G, C<—>T.
    b. Переменная бета отвечает за показатель интенсивностей для 1-го типа трансверсий, то есть переходов A<—>C, G<—>T
    c. Переменная лямбда отвечает за показатель интенсивностей для 2-го типа трансверсий, то есть переходов A<—>T, G<—>C
2. Равновесные частоты всех нуклеотидов равны

$$Q = \begin{bmatrix} -(\alpha+\beta+\gamma)*\frac{1}{4} & \alpha*\frac{1}{4} & \beta*\frac{1}{4} & \gamma*\frac{1}{4} \\ \alpha*\frac{1}{4} & -(\alpha+\beta+\gamma)*\frac{1}{4} & \gamma*\frac{1}{4} & \beta*\frac{1}{4} \\ \beta*\frac{1}{4} & \gamma*\frac{1}{4} & -(\alpha+\beta+\gamma)*\frac{1}{4} & \alpha*\frac{1}{4} \\ \gamma*\frac{1}{4} & \beta*\frac{1}{4} & \alpha*\frac{1}{4} & -(\alpha+\beta+\gamma)*\frac{1}{4} \end{bmatrix}$$

Модель Фельзенштейна (F81)

По сути является расширением модели Джукса-Кантора (JC69). Допущения модели:

1. Интенсивности замен всех нуклеотидов равны
2. Равновесные частоты всех нуклеотидов НЕ равны

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

SYM Модель (SYM)

Сложнейшая модель, рассматриваемая в этом проекте. Допущения модели состоят в том, что:

1. Интенсивности замен для каждой отдельно взятой пары нуклеотидов разные, однако вероятности перехода нуклеотидов внутри пары равны.
   a. Переменная альфа отвечает за вероятность переходов A<—>G
   b. Переменная бета отвечает за вероятность переходов A<—>C
   c. Переменная лямбда отвечает за вероятность переходов A<—>T
   d. Переменная дельта отвечает за вероятность переходов C<—>G
   e. Переменная эпсилон отвечает за вероятность переходов C<—>T
   f. Переменная эта отвечает за вероятность переходов G<—>T
2. Равновесные частоты всех нуклеотидов равны

$$Q = \begin{bmatrix} -(\alpha+\beta+\gamma)*\frac{1}{4} & \alpha*\frac{1}{4} & \beta*\frac{1}{4} & \gamma*\frac{1}{4} \\ \alpha*\frac{1}{4} & -(\alpha+\delta+\epsilon)*\frac{1}{4} & \delta*\frac{1}{4} & \epsilon*\frac{1}{4} \\ \beta*\frac{1}{4} & \delta*\frac{1}{4} & -(\beta+\delta+\eta)*\frac{1}{4} & \eta*\frac{1}{4} \\ \gamma*\frac{1}{4} & \epsilon*\frac{1}{4} & \eta*\frac{1}{4} & -(\gamma+\epsilon+\eta)*\frac{1}{4} \end{bmatrix}$$

Схема работы. Программа-генератор

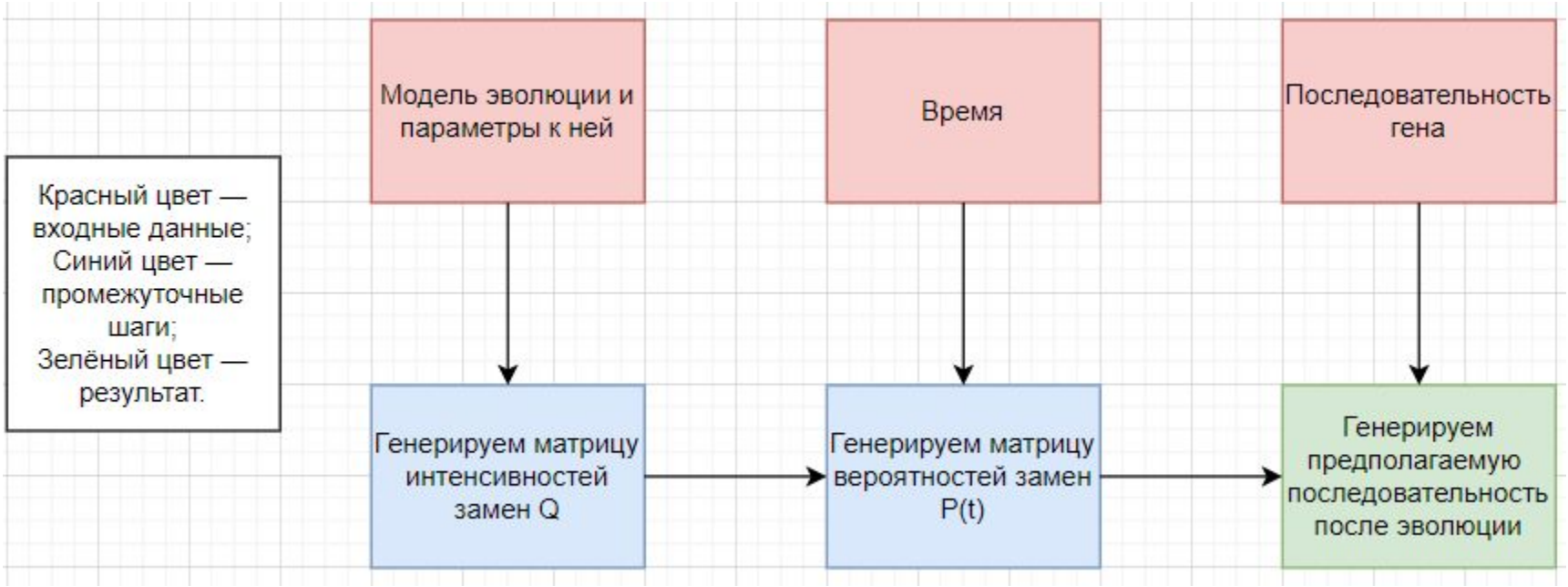
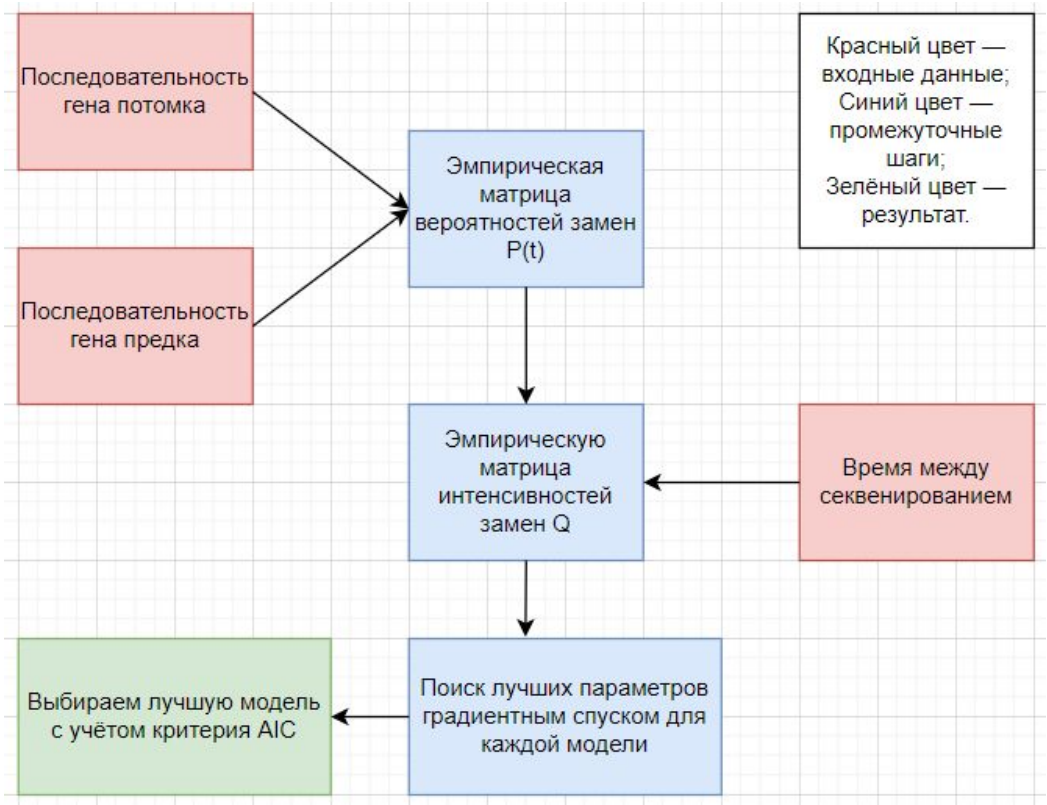


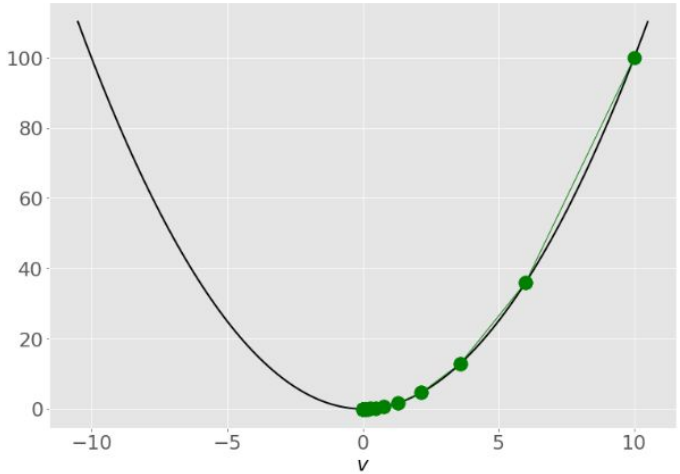
Схема работы. Программа-предиктор



Градиентный спуск. Общая идея

Градиентный спуск — численный метод нахождения локального минимума или максимума функции с помощью движения вдоль градиента, один из основных численных методов современной оптимизации. Пример реализации алгоритма для нахождения точки минимума функции y = x^2.

1. Выберем стартовую точку случайным образом. Пусть position = 10.
2. Найдем производную в окрестностях этой точки. (Мы взяли dx = 0.0001).
3. Пересчитаем position, с учетом полученной производной: $Pos = Pos - Z * f'(Pos)$



$$f'(x) = \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x}$$

Повторяя много раз, position будет стремиться к точке минимума.

Градиентный спуск. В нашем проекте

Для того, чтобы подбирать параметры за разумное время нам и понадобится градиентный спуск, ведь пользуясь только перебором — количество необходимых вычислений было бы очень большим. Чтобы использовать градиентный спуск в нашем проекте, мы искали нули следующего выражения:

$$\left(\begin{bmatrix} \alpha_0 & \alpha_1 & \alpha_2 & \alpha_3 \\ \alpha_4 & \alpha_5 & \alpha_6 & \alpha_7 \\ \alpha_8 & \alpha_9 & \alpha_{10} & \alpha_{11} \\ \alpha_{12} & \alpha_{13} & \alpha_{14} & \alpha_{15} \end{bmatrix} - \begin{bmatrix} \alpha_{16} & \alpha_{17} & \alpha_{18} & \alpha_{19} \\ \alpha_{20} & \alpha_{21} & \alpha_{22} & \alpha_{23} \\ \alpha_{24} & \alpha_{25} & \alpha_{26} & \alpha_{27} \\ \alpha_{28} & \alpha_{29} & \alpha_{30} & \alpha_{31} \end{bmatrix}\right)^2$$

Зелёная матрица — эмпирическая матрица

AIC

Мы нашли наилучшие параметры для каждой из возможных моделей. Перед нами встала задача: как выбрать правильную модель? Для этого мы использовали критерий Акаике (=AIC). Этот критерий задаётся формулой, в которой идёт противоборство между качеством работы модели (по сути результатом функции правдоподобия) и количества её параметров.

$$AIC = 2k - 2ln(P)$$

Демонстрация работы

Рабочий GitHub проекта:

https://github.com/furushigava/Guess-the-model-of-evolution

Демонстрация в Google Collab :

https://colab.research.google.com/drive/1lCqERRKLXMuvHgBktO-VEjjBLfNPJNlS

Заключение

В ходе нашего проекта все поставленные задачи были успешно выполнены, все программы написаны. Исследование качества программы-предиктора показало, что:

- При time в промежутке 0.05 до 3.5. Тогда процент сбоев составляет в среднем 0,047%, а средний процент угадываний — 91,125%
- Наблюдается закономерность, чем меньше время — тем чаще программа-предиктор будет выбирать простые модели.

Угадай модель эволюции

Выполнил: Артемьев Святослав

Научный руководитель: Дмитрий Биба

Цепи Маркова с дискретным временем

Последовательность X_n мы будем называть марковской цепью, с дискретным времнем если:

$$P(X_{n+1} = i_{n+1} | X_n = i_n, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) = P(X_{n+1} = i_{n+1} | X_n = i_n)$$

Таким образом, в простейшем случае условное распределение последующего состояния цепи Маркова зависит только от текущего состояния и не зависит от всех предыдущих состояний.

Матрицей вероятностей переходов мы будем считать матрицу $P_{ij}(n)$, если:

$$P_{ij} = P(X_{n+1} = j | X_n = i),\ \text{n} \text{ — количество шагов.}$$

Начальным (стартовым) распределением цепи Маркова будем считать вектор π, такой, что:

$$\pi = (p_1, p_2, p_3, \ldots, p_n),\ \text{где } p_i = P(X_0 = i)$$

В таком случае, получается, что вероятность перехода из некого состояния i в j за время n, при начальном распределении $\pi = (p_1, p_2, \ldots, p_n) : P_{ij}(n) = P^n$

Цепи Маркова с непрерывным временем

Пусть у нас есть матица вероятностей замен $P(t)$. Допустим, что t определено непрерывно. Найдём производную P'(t):

$$P'(t) = \frac{dP(t)}{dt} = \lim_{\Delta t \to 0} \frac{P(t+\Delta t) - P(t)}{\Delta t}$$

Так как $P(t+\Delta t) = P(t)P(\Delta t)$:

$$P'(t) = \lim_{\Delta t \to 0} \frac{P(t)P(\Delta t) - P(t)}{\Delta t} = \lim_{\Delta t \to 0} \frac{P(t)(P(\Delta t) - I)}{\Delta t}$$

$$P'(t) = P(t) * \boxed{\lim_{\Delta t \to 0} \frac{P(\Delta t) - I}{\Delta t}}$$ — назовём это Q матрицей

Q — матрица интенсивностей переходов. Она показывает, как изменяется P матрица, на очень маленьком промежутке времени $(t \to 0)$. В таком случае справедливо следующее:

$$\frac{dP(t)}{dt} = Q * P(t) \quad \Longrightarrow \quad \frac{1}{P(t)} * dP(t) = Q * dt$$

$$\int \frac{1}{P(t)} * dP(t) = \int Q * dt \quad \Longrightarrow \quad ln(P(t)) = Q * t$$

Получается, что: $\boxed{P(t) = e^{Qt}}$

I — единичная матрица, аналог 1 в скалярном представлении

I может быть, например, такими матрацами: $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}; \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$

То есть размер I матрицы зависит от размера P матрицы

Общий вид Q матрицы: $Q = \begin{bmatrix} -\alpha & \alpha \\ \beta & -\beta \end{bmatrix}$

Важно, что $\sum_{i=0}^{n} Q_{ji} = 0 \; \forall \; j$

Матрица P(t) в таком случае будет выглядеть так:

$$P(t) = e^{Qt} = \frac{1}{\alpha + \beta} \begin{bmatrix} \beta + \alpha e^{-(\alpha+\beta)t} & \alpha - \alpha e^{-(\alpha+\beta)t} \\ \beta - \beta e^{-(\alpha+\beta)t} & \alpha + \beta e^{-(\alpha+\beta)t} \end{bmatrix}$$