MCMC methods in Bioinformatics

Maria Chernigovskaya

26 июля 2018 г.

“Простое” сэмплирование

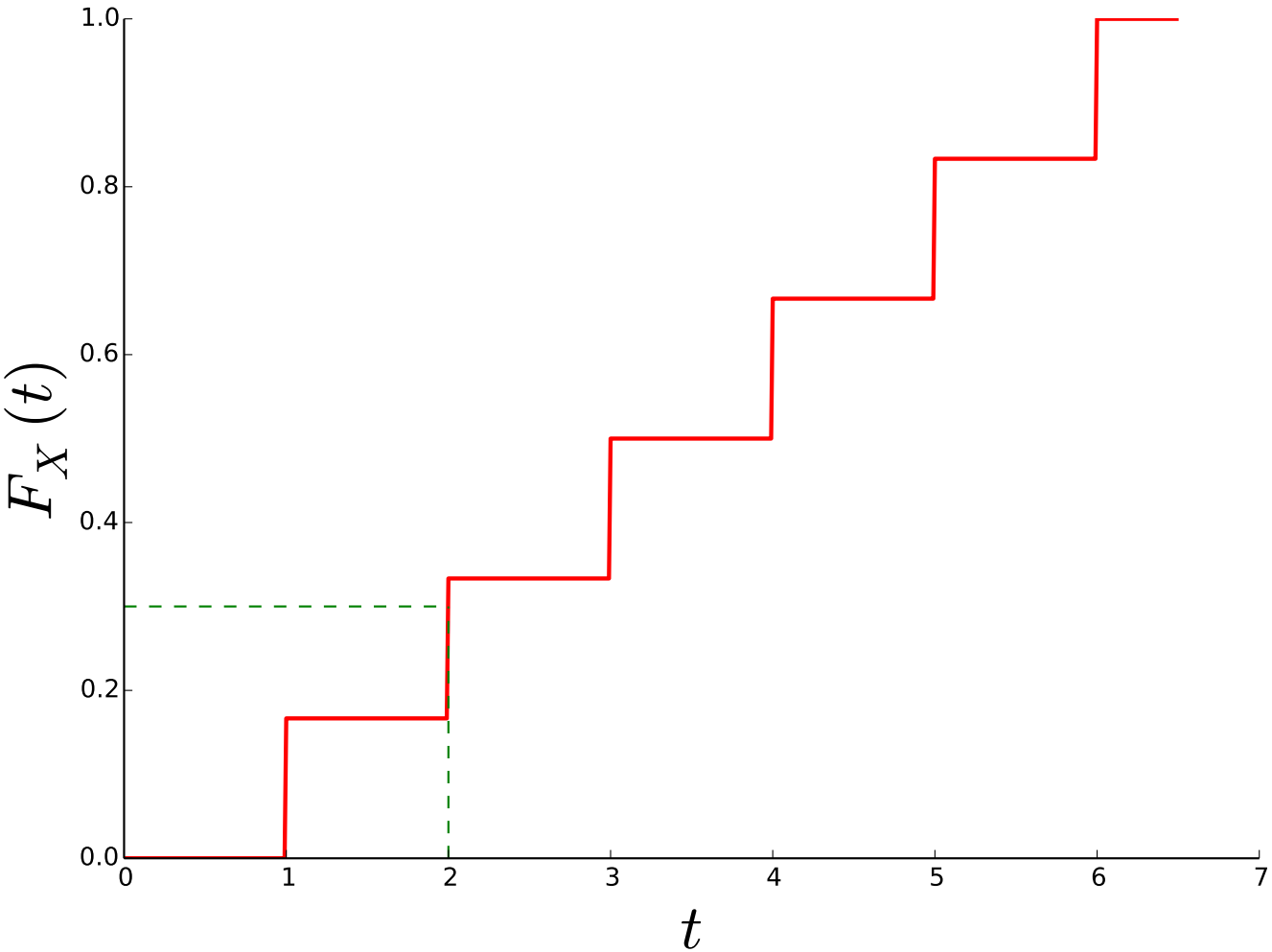
- Пусть есть U – равномерно случайное число из $[0, 1]$;
- Как с помощью этого симулировать подкидывание монетки?
- Как с помощью этого симулировать бросания кубика?
- Как с помощью этого симулировать непрерывную случайную величину (например, экспоненциальную)?
- Как с помощью этого симулировать дискретную случайную величину (например, Пуассона)?
- КАК ЖЕ получить случайное филогенетическое дерево?

Функция распределения

X - случайная величина, тогда

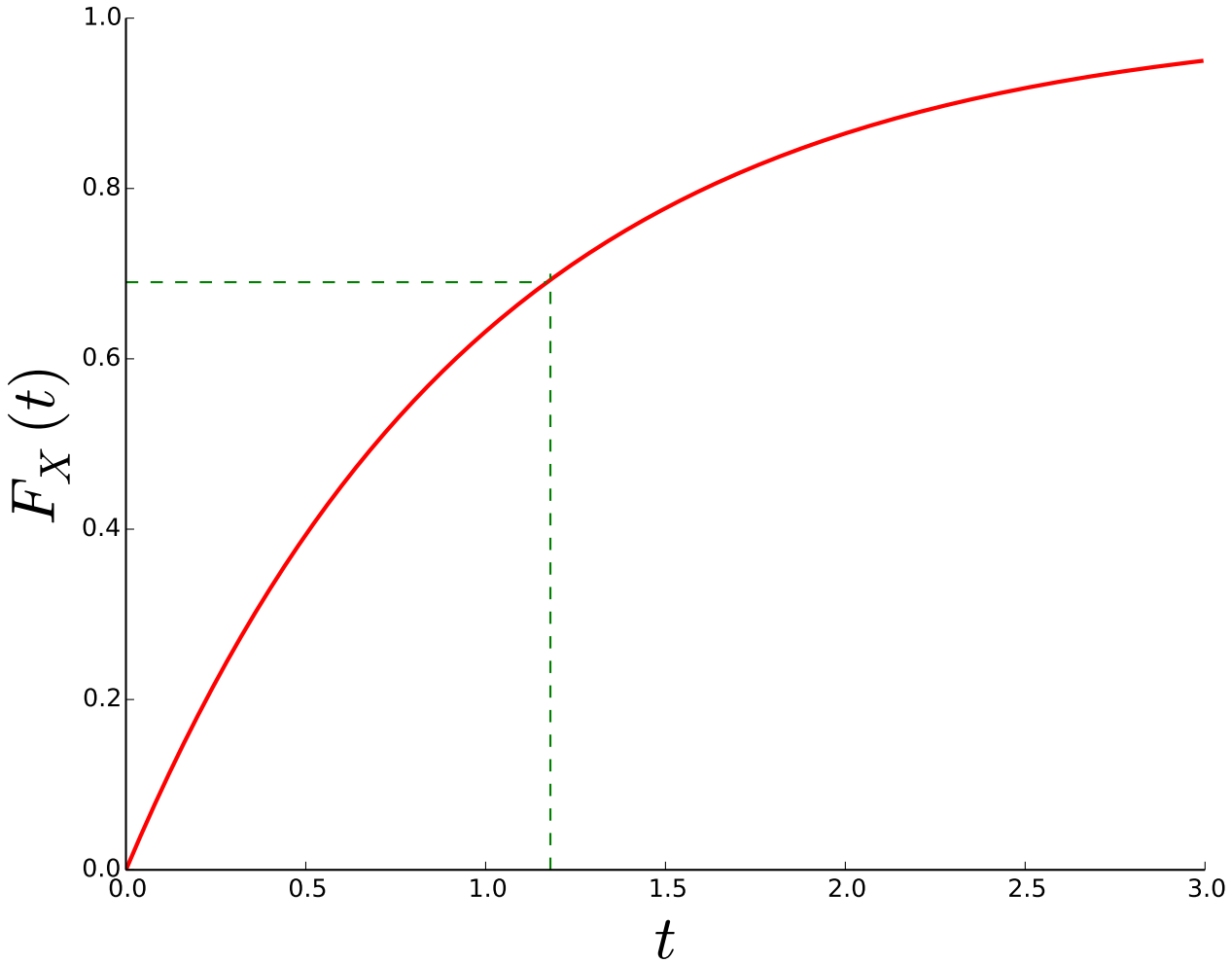
$$F_X(t) = P(X \leq t).$$

Дискретный случай



Распределение числа точек на кубике

Непрерывный случай



Экспоненциальное распределение $F_X(t) = 1 - e^{-x}$.

Дискретный случай с бесконечным числом состояний

Что делать, например, с распределением Пуассона ?

$$P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}$$

СЛОЖНО!

Что делать, если нужно симулировать сложный объект (например, граф, филогенетическое дерево, кристаллическую решетку)?

MCMC!

Марковские цепи

Состояния $i = 1, 2, \ldots$, вероятности перехода из состояния i в состояние j p_{ij}.
Пример: погода в стране Оз (дождь, хорошая погода, снег):

$$\begin{bmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \end{bmatrix}$$

Упр.: Найдите вероятность того, что погода изменится с хорошей на дождливую за 2 дня.

Metropolis–Hastings algorithm

Находимся в состоянии x_t, рассмотрим переход в следующее состояние.

1. предлагаем кандидата x' с вероятностью $Q(x_t, x')$;
2. Acceptance ratio (Метрополис, симметричное Q):

$$\alpha = \frac{P(x')}{P(x_t)}$$

   Или (Метрополис-Гастингс, несимметричное Q):

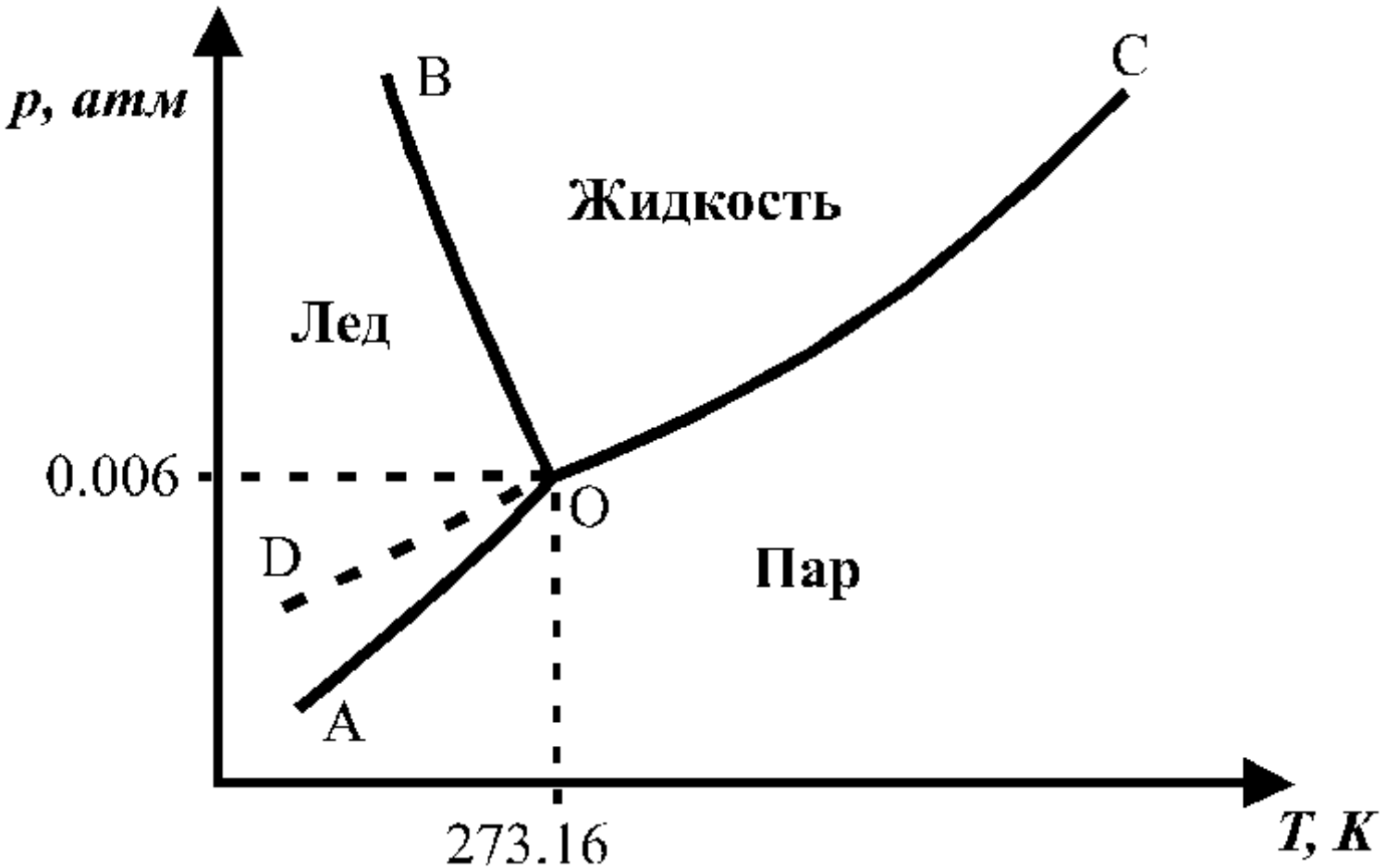
$$\alpha = \frac{P(x')}{P(x_t)} \frac{Q(x', x_t)}{Q(x_t, x')};$$

3. Если *alpha* > 1, то $x_{t+1} := x'$; если $\alpha < 1$, то

$$x_{t+1} := \begin{cases} x', & \text{с вероятностью } \alpha \\ x_t, & \text{с вероятностью } 1 - \alpha \end{cases}$$

Примеры на сегодня (??)

1. распределение Пуассона;
2. (внезапно) агрегатное состояние вещества



Шары и фазовый переход

Состояние – конфигурация шаров. Желаемое распределение: Все конфигурации равновероятны, т.е.

$$\frac{P(x')}{P(x)} = 1\,.$$

Функция $Q(x, x') = 1$, если положение всех шаров, кроме одного (B), совпадает, а положение B изменилось незначительно

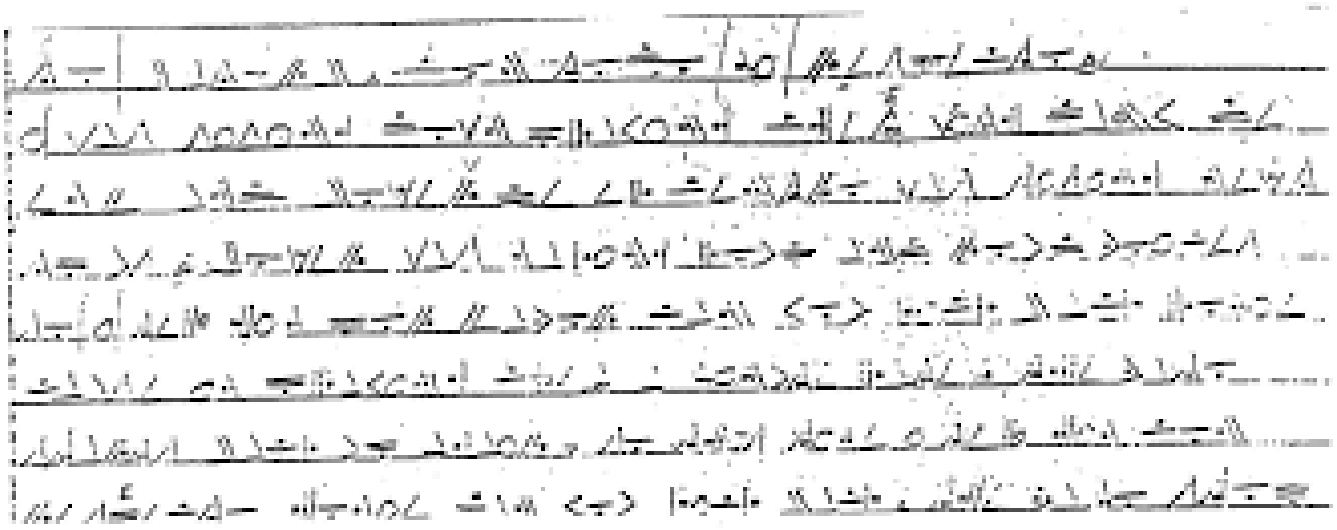
$$dist(B(x), B(x')) < \varepsilon$$

(и $Q(x, x') = 0$ в противном случае).
Заметим, что в данном случае Q – плотность вероятность, а не вероятность сама по себе, но для $Q(x, x')/Q(x', x)$ это не имеет принципиального значения.

Распределение Пуассона

Множество состояний $= \{0, 1, \dots\}$.
Правило перехода (Q) – переходим в предыдущее (если разрешено) или следующее число с равными вероятностями.

Криптография



Студенты Marc Coram and Phil Beineke at Stanford получили набор шифрованных сообщений из тюрьмы штата.

Состояние – код, т.е. отображение
$f : \{\text{символы из сообщения}\} \to \{\text{символы алфавита}\}$.
Сколько всего есть состояний?
Функция “хорошести” кода:

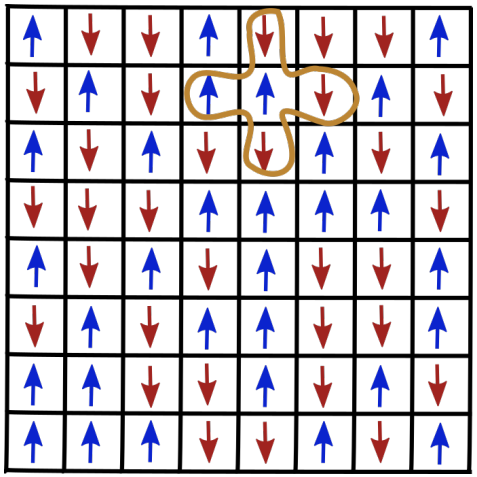
$$G(f) := \prod_i M\left(f(s_i), f(s_{i+1})\right),$$

где s_i – это i-й символ сообщения; M – это матрица частот паросочетаний букв в английском языке (обученная на **War and Peace**).

$$\frac{P(f')}{P(f)} = \frac{G(f')}{G(f)}$$

Переходы?

Модель Изинга: Намагничевание



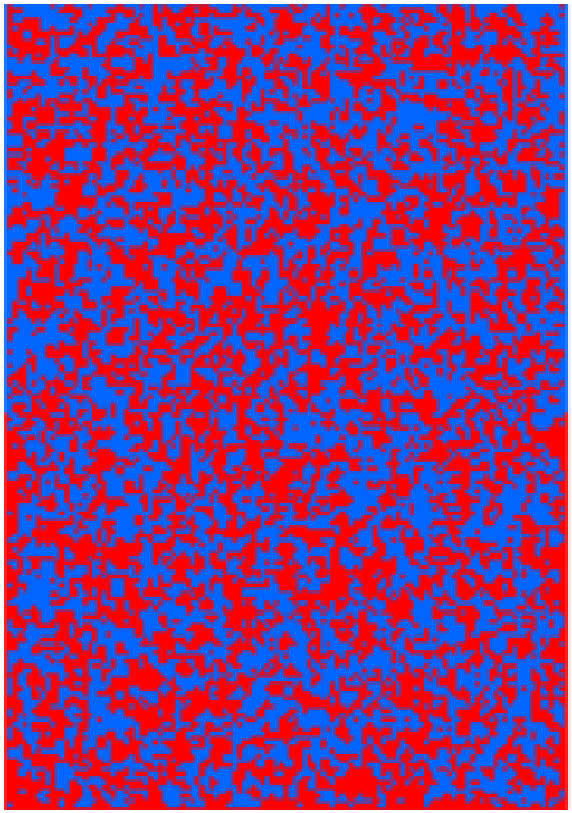
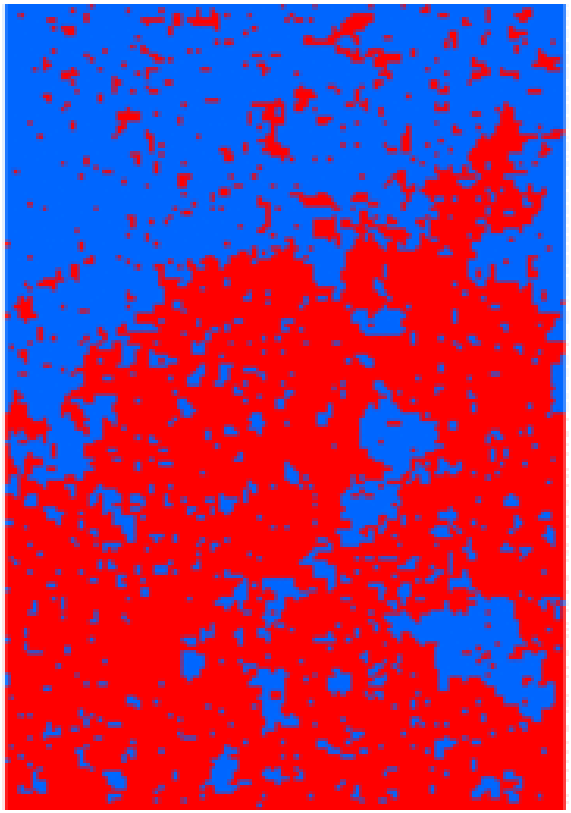
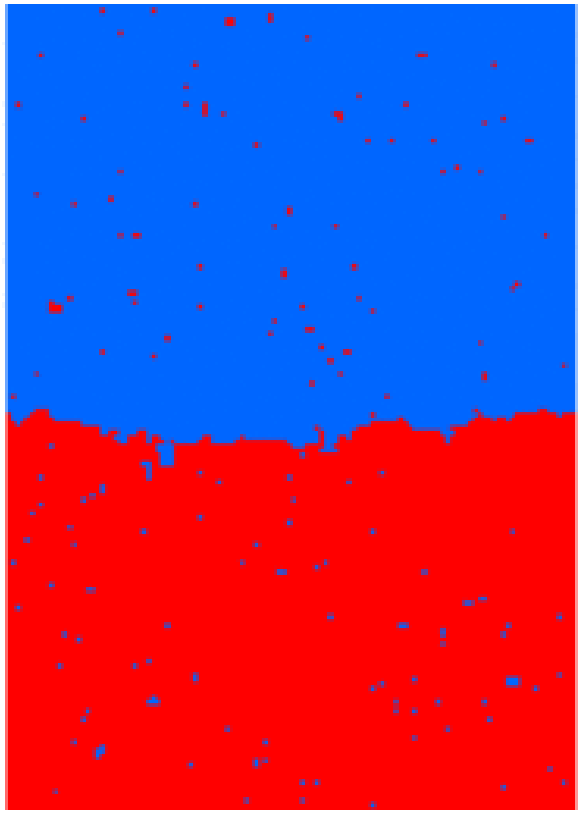
Спиновая решетка

$$\sigma(v) \in \{+, -\};$$

$$E(\sigma) = \#\{(v, w), v \sim w : \sigma(v) \neq \sigma(w)\};$$

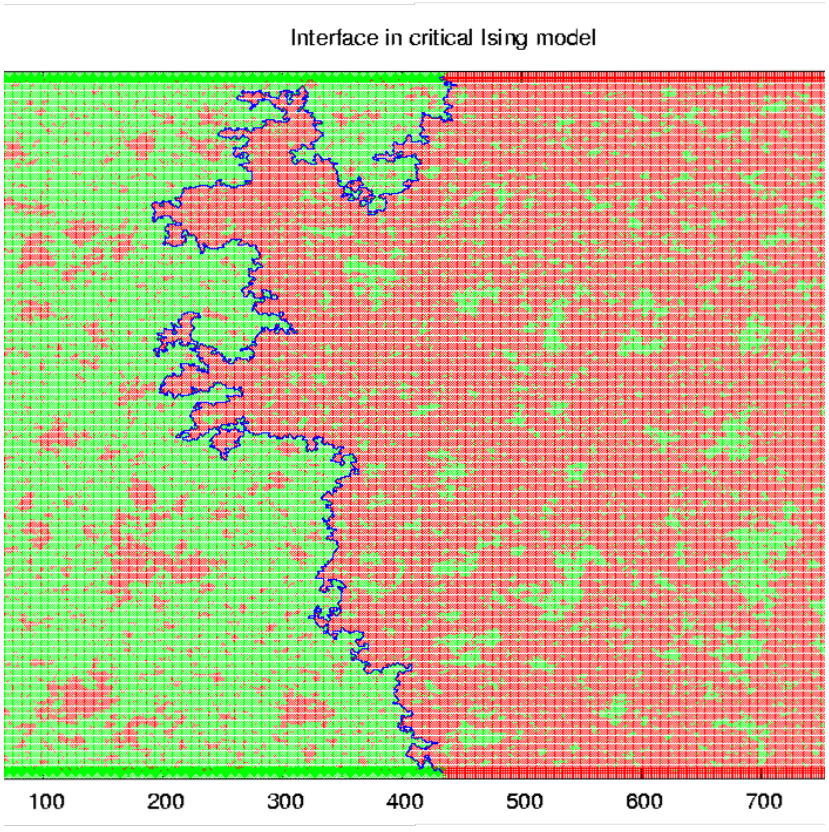
$$P(\sigma) = \frac{1}{Z_T} \exp\left(-\frac{1}{k_B T} E(\sigma)\right).$$

Модель Изинга, MCMC

Посткритическая ($T >> T_{cr}$), критическая $T = T_{cr}$, и докритическая ($T << T_{cr}$) фазы.

Модель Изинга



Модель Изинга

Jupyter notebook

mr.Bayes

Томас Байес (1702 — 7 апреля 1761) — английский математик, пресвитерианский священник.

mr.Bayes

Цель: построить филогенетическое дерево по некоторому набору геномов X.

- G – набор геномов (строки равной длины над алфавитом $A, G, C, T, -$),
- τ – филогенетическое дерево (топология),
- v – "длины" ребер,
- θ – параметры мутирования (например, матрица замен 5×5).

Формула Байеса:

$$p(\tau, v, \theta \mid G) = \frac{p(G \mid \tau, v, \theta) p(\tau, v, \theta)}{p(X)}.$$

Формула Байеса:

$$p(\tau, v, \theta \mid G) = \frac{p(G \mid \tau, v, \theta)p(\tau, v, \theta)}{p(G)}.$$

Как сделать переход?

$$(\tau, v, \theta) \mapsto (\tau', v', \theta') :$$

$$\alpha = \frac{p(\tau', v', \theta')}{p(\tau, v, \theta)} = \frac{p(G \mid \tau', v', \theta')}{p(G \mid \tau, v, \theta)},$$

так как предполагаем

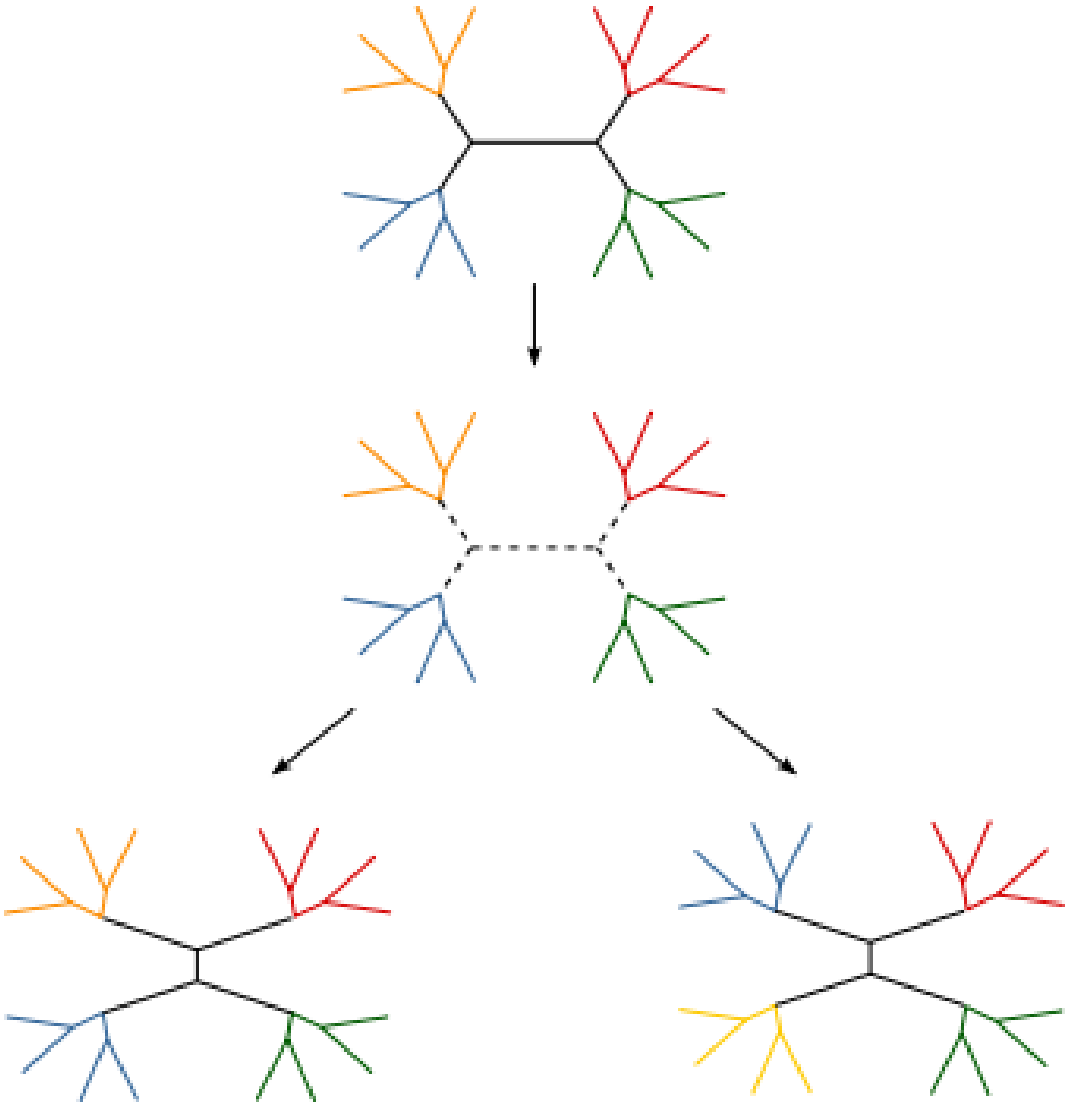
$$p(\tau, v, \theta) = p(\tau', v', \theta').$$

mr.Bayes: как делать переход?

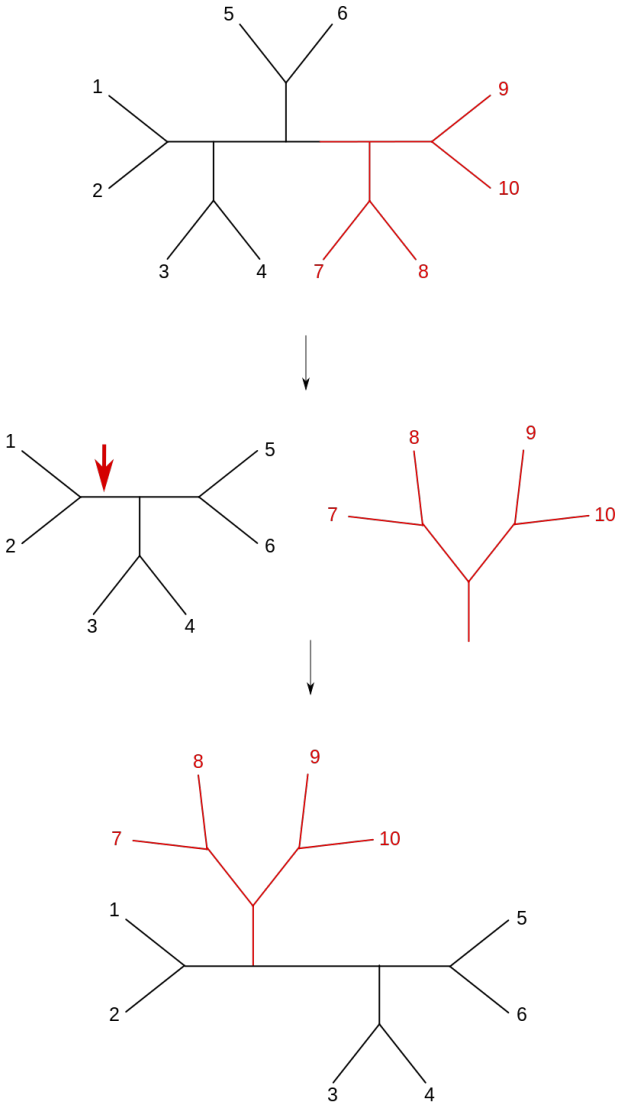
- v – вектор из e вещественных чисел, $v \in \mathbb{R}^e$; $v' \sim \mathcal{N}(v, \varepsilon E)$,
- θ – матрица мутаций, $\theta \in \mathbb{R}^{25}$; $\theta' \sim \mathcal{N}(\theta, \varepsilon E)$,
- τ – дерево, его можно менять ДВИЖЕНИЯМИ ДЕРЕВА.

mr.Bayes: движения дерева

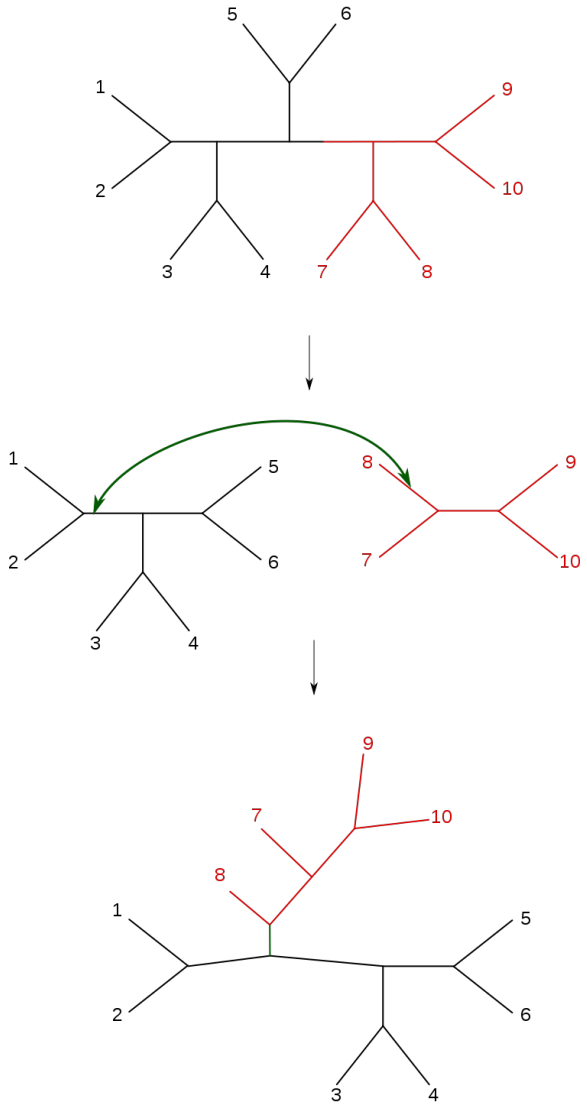
mr.Bayes: Nearest neighbor interchange (NNI)



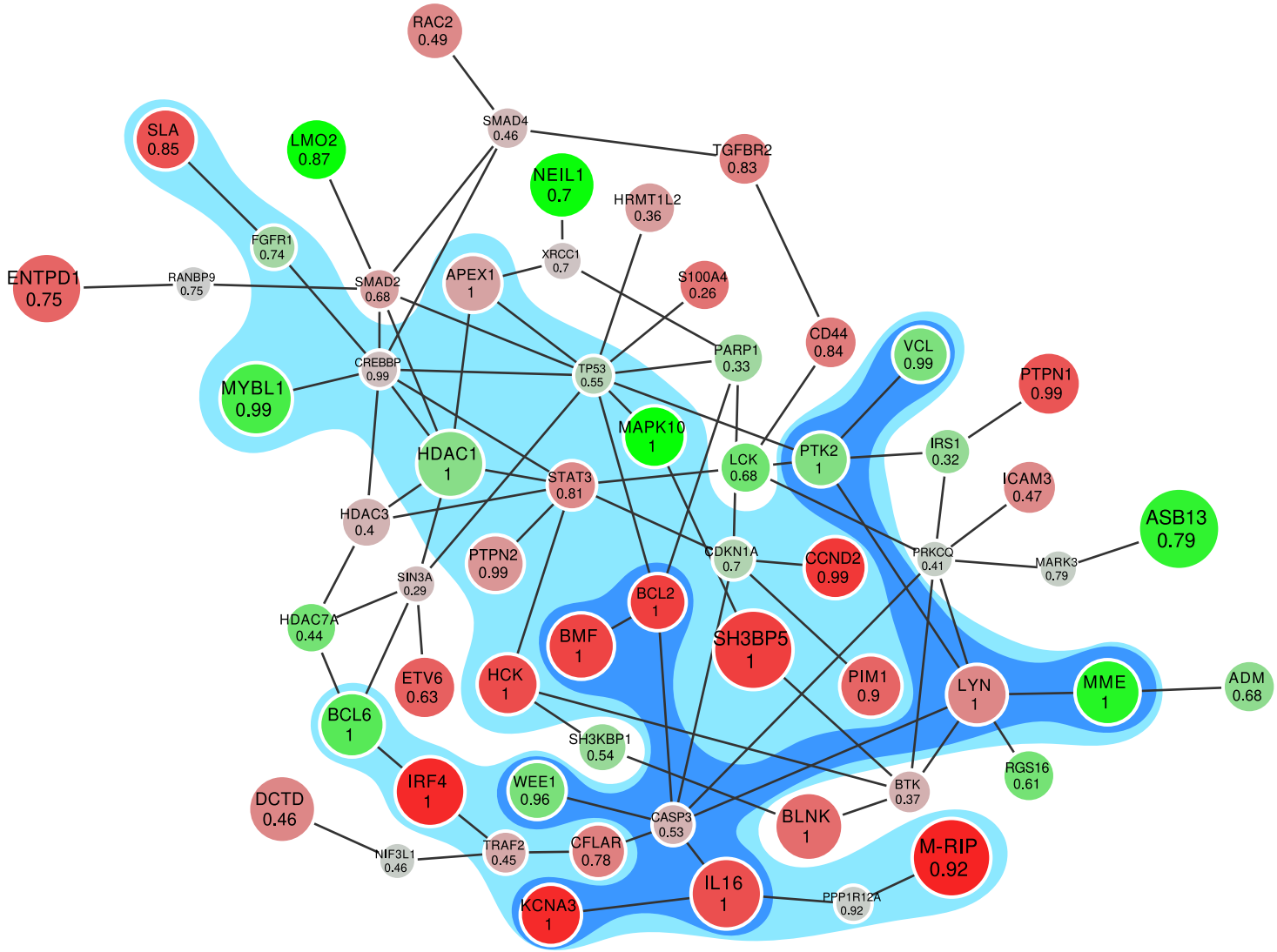
mr.Bayes: Subtree pruning and regrafting (SPR)



mr.Bayes: Tree bisection and reconnection (TBR)



Find active module in PPI network



Find active module in PPI network

Дано:

- Сеть белок-белковых взаимодействий;
- p-values для всех генов, кодирующих белки.

1. initialize S_0 as a random connected subgraph on $k = |V(M)|$ vertices
2. FOR $i = 0, 1, 2, \ldots$
   - Choose v_- from $V(S_i)$ and v_+ from $nei(S_i)$ uniformly;
   - Propose S' as an induced subgraph on $V(S_i) \setminus \{v_-\} \cup \{v_+\}$;
   - IF S' is connected: Acceptance Probability:

$$\rho(S_i, S') = \min\left\{1, \frac{p_{v_+}^{\alpha-1}}{p_{v_-}^{\alpha-1}} \frac{|\,\mathrm{nei}(S')|}{|\,\mathrm{nei}(S_i)|}\right\}$$

$$S_{i+1} := \begin{cases} S' \text{ with probability } \rho(S_i, S') \\ S_i \text{ with probability } 1 - \rho(S_i, S') \end{cases}$$

$$S_{i+1} := S_i$$