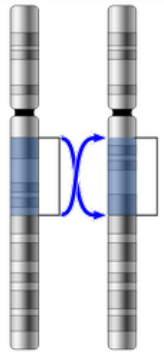
УНИВЕРСИТЕТ ИТМО

Анализ геномных перестроек с помощью случайных графов

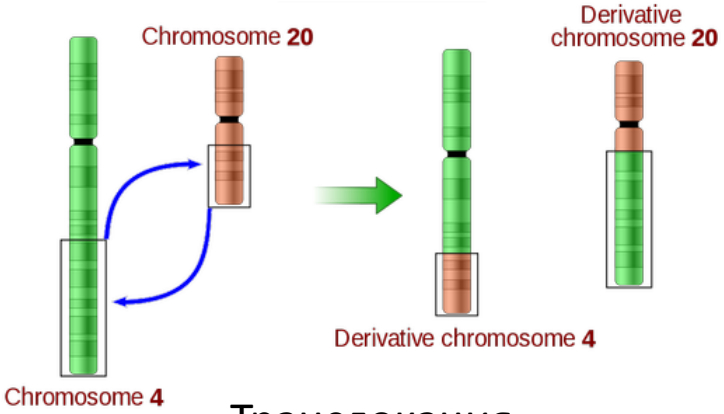
Забелкин А.А.
Научный руководитель: Алексеев Н.В.

Геномные перестройки

- Инверсия
- Транслокация
- Слияние
- Расщепление
- Транспозиция

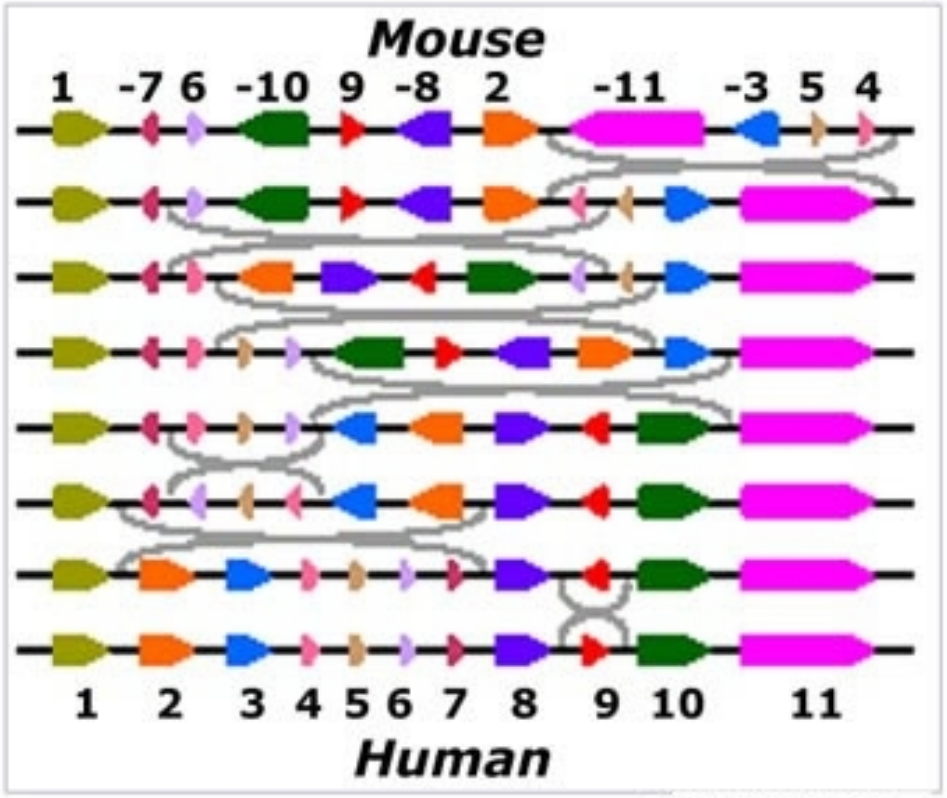


Инверсия



Транслокация

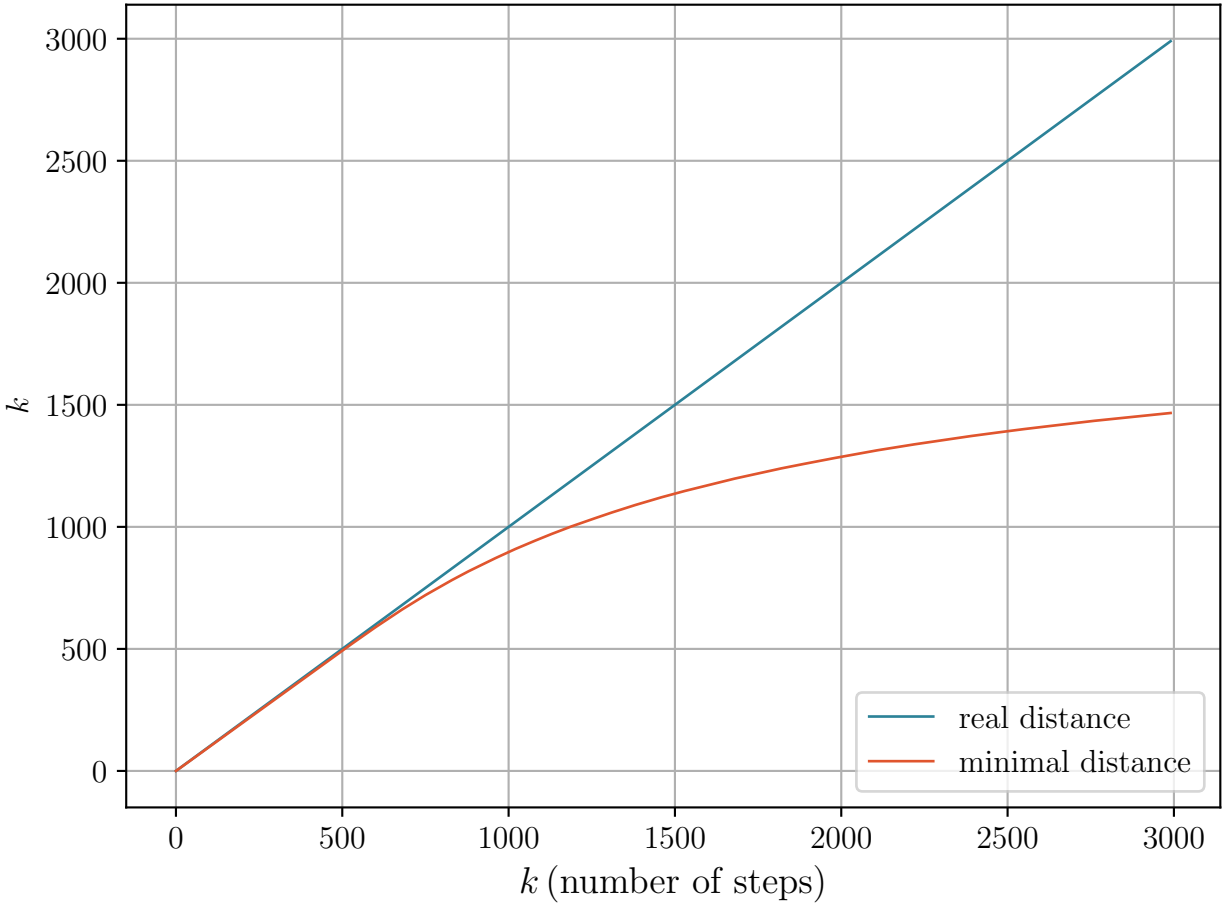
Эволюция генома



Оценка расстояния

- Предположение парсимонии
  - **Минимальное** расстояние, необходимое для преобразования одного генома в другой
- Истинное эволюционное расстояние
  - Оценка **реального** количество перестроек, произошедших между геномами в ходе эволюции

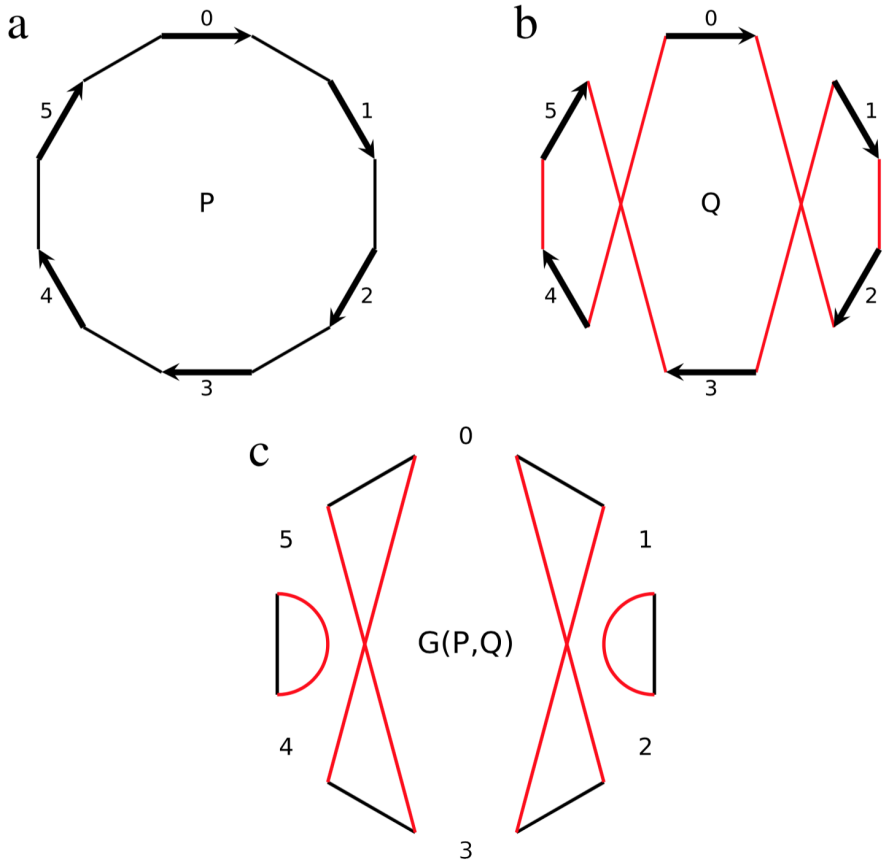
Могут сильно отличаться



Актуальность задачи

- Хотим знать реальное расстояние между видами
- Важно для многих филогенетических исследований
  - *Genome rearrangements and sorting by reversals, 1996* (600+ ссылок)
- В последние время появляется всё больше полностью собранных геномов

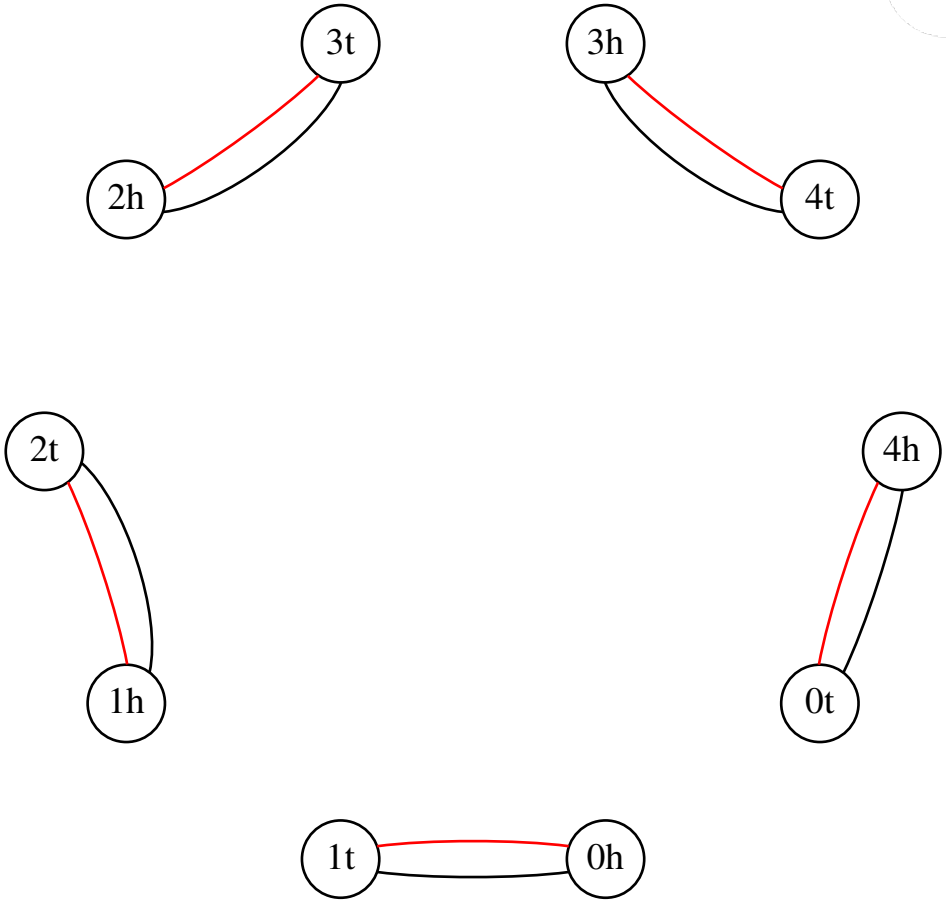
Можем представить геном как граф (breakpoint graph)



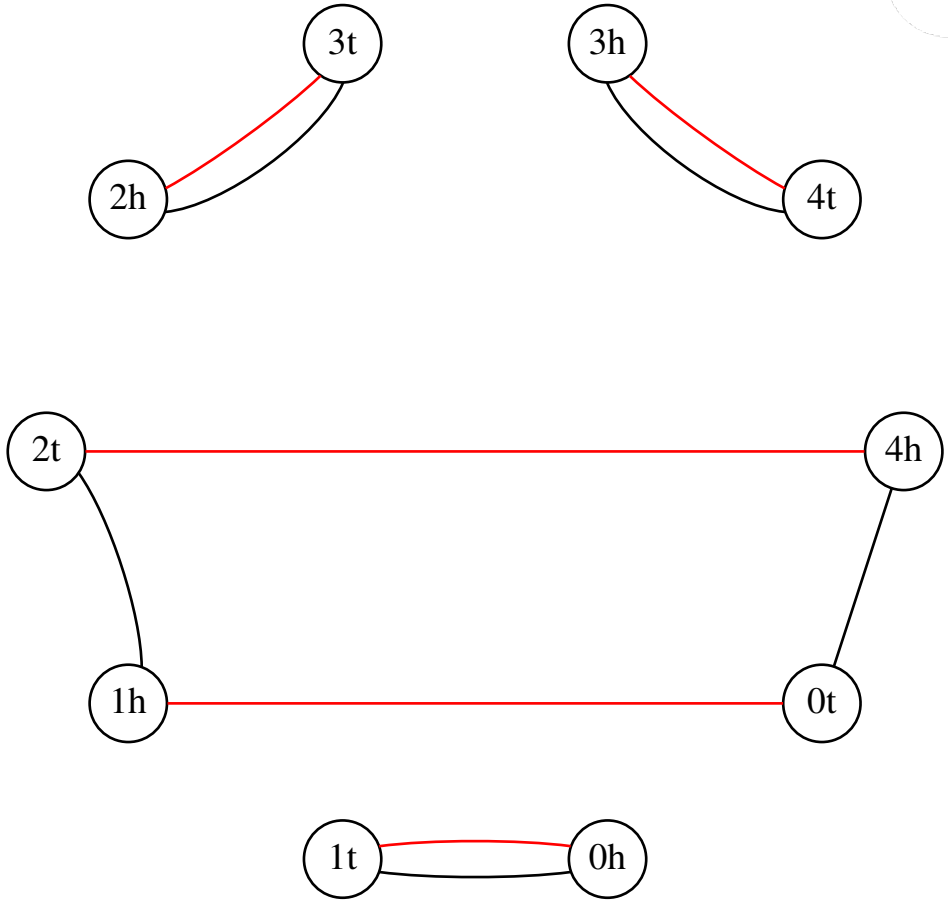
Дискретный марковский процесс

- Изначально оба генома одинаковы
- Моделируем на графе дискретный марковский процесс
  - Чёрные рёбра фиксированы
  - Перестройки совершаются на красных ребрах
- k - число шагов

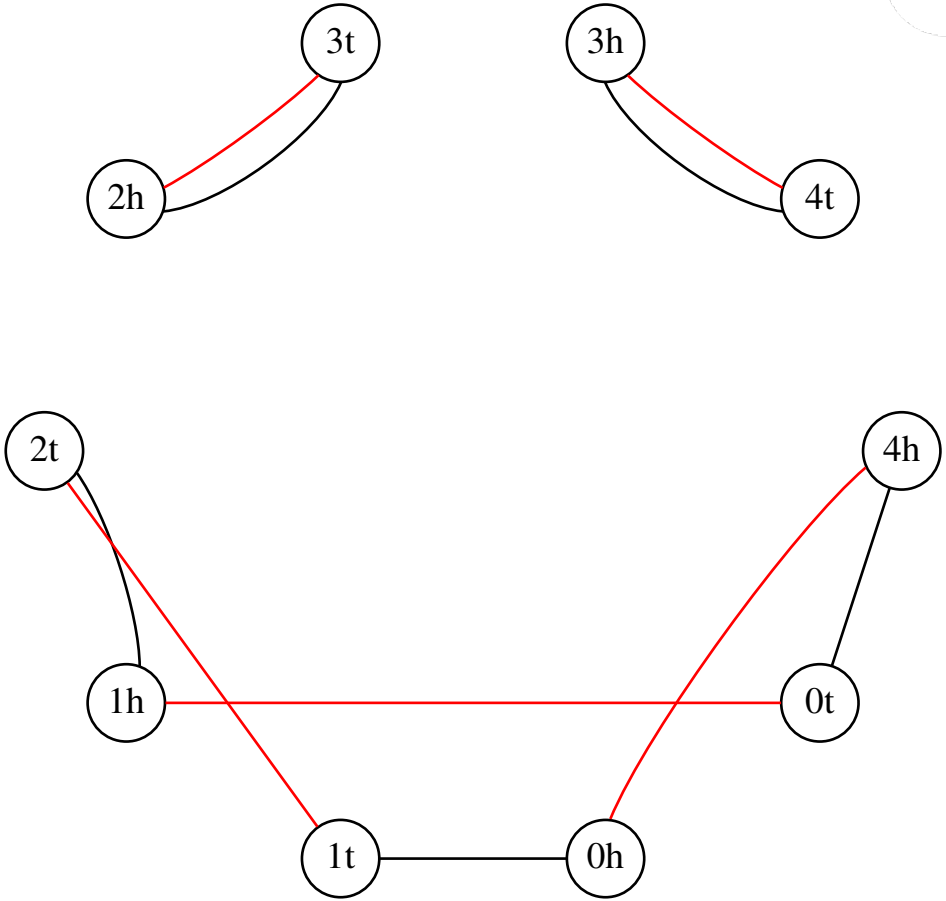
Марковский процесс k = 0



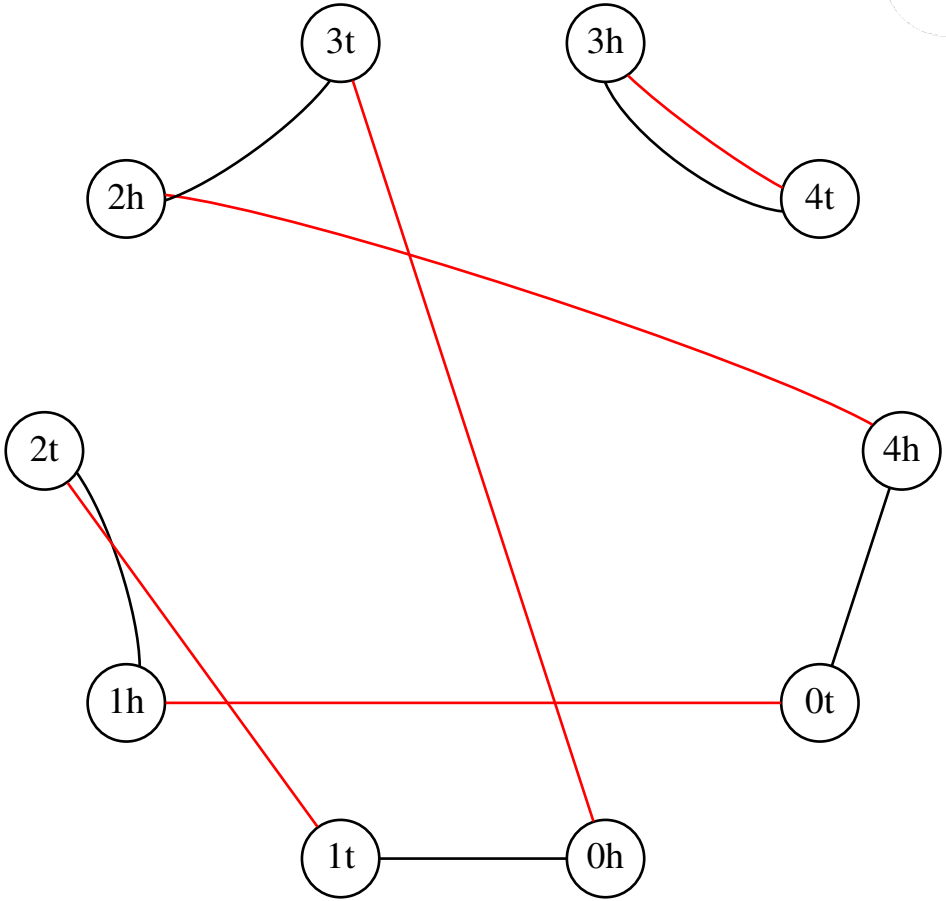
Марковский процесс k = 1



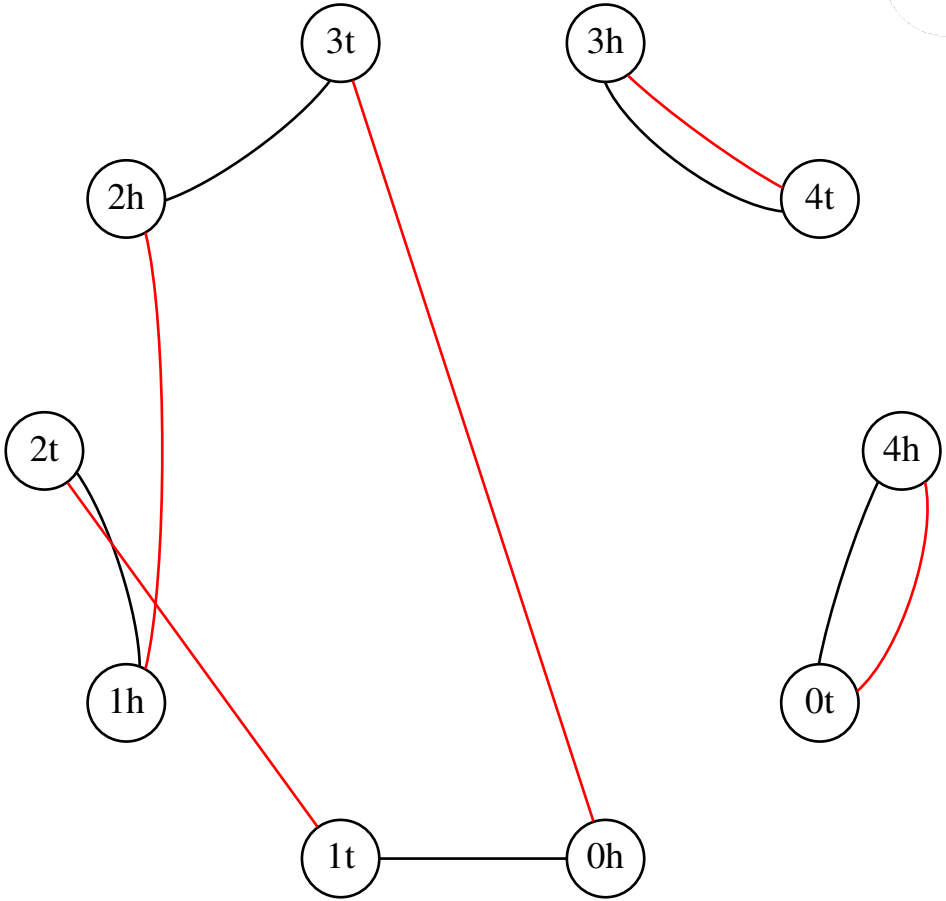
Марковский процесс k = 2



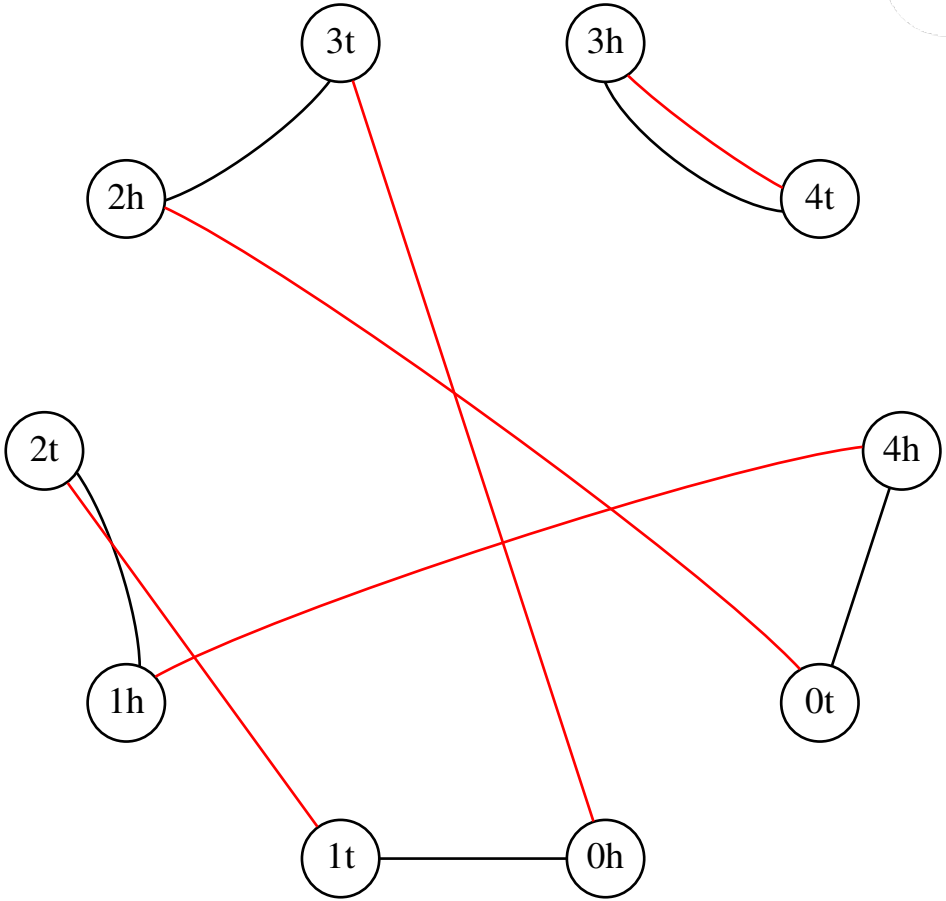
Марковский процесс k = 3



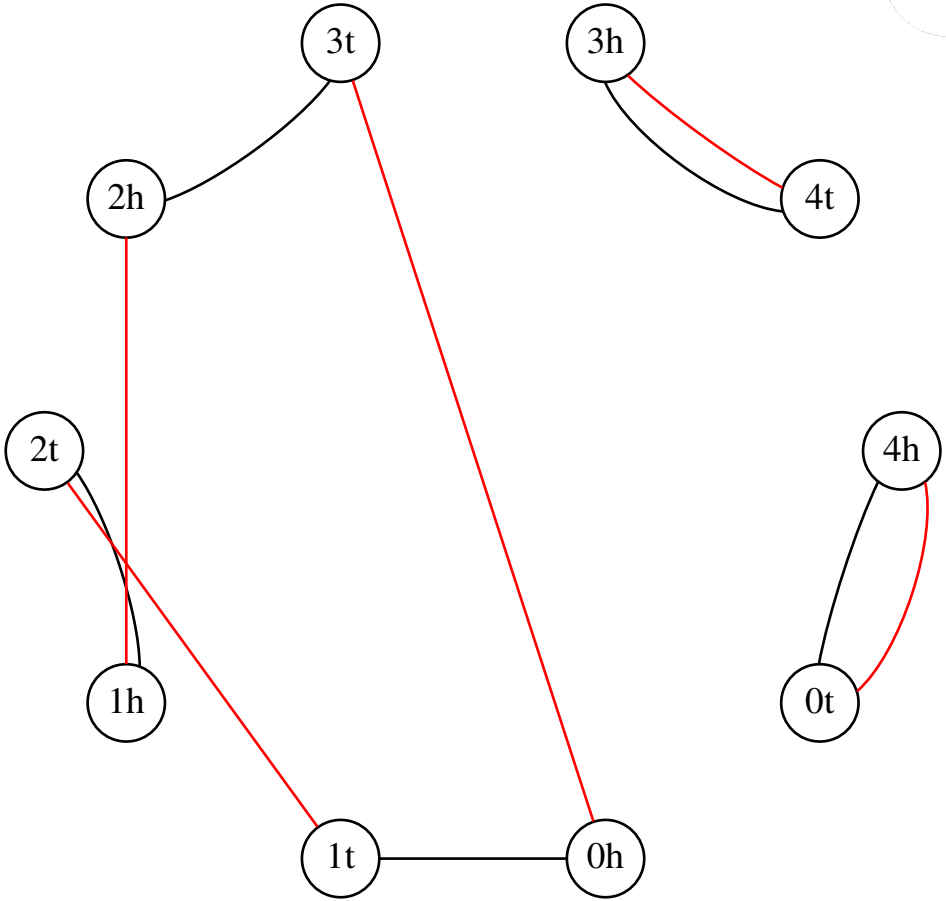
Марковский процесс k = 4



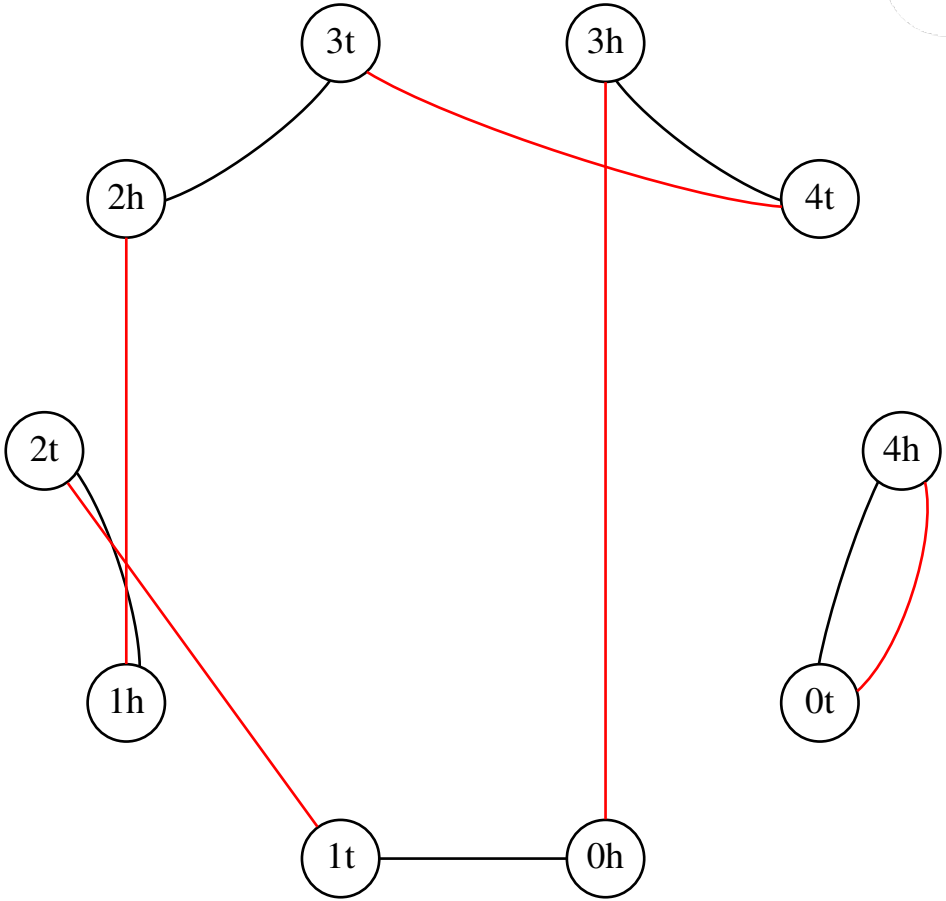
Марковский процесс k = 5



Марковский процесс k = 6



Марковский процесс k = 7



Задача оценки параметров модели

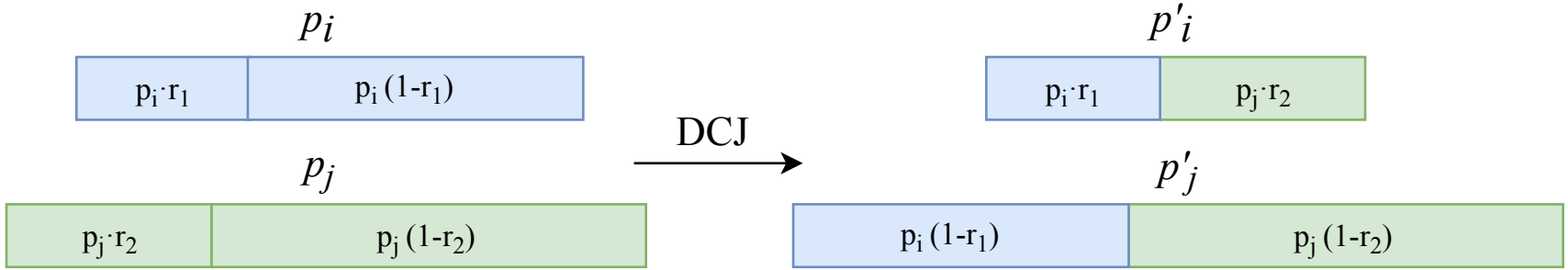
- Известно количество компонент каждой длины
- Предсказываем k

Сравнение подходов

- Как выбираются рёбра для перестройки:
  - Все ребра равновероятны
  - Все ребра имеют веса
- *Eric Tannier — Breaking Good: Accounting for Fragility of Genomic Regions in Rearrangement Distance Estimation, Genome Biology and Evolution. — 2016*
- В реальности разные хрупкие регионы имеют разную вероятность быть вовлеченными в перестройку

Модель, предложенная Танье

- Каждому ребру сопоставляется вероятность p_i быть вовлеченным в перестройку
- Веса пропорциональны длинам хрупких областей
- Перераспределение весов:



- Подобное перераспределение весов в пределе даёт равномерное распределение векторов на n-мерном симплексе

Минусы метода оценки, предложенного в статье

- Формулы для оценки сложно применить практически

$$c_2 = kn^2 \sum_{l=0}^{\infty} \sum_{m=0}^{\infty} \frac{(-2(k-1))^{l+m}(l+1)(m+1)}{\prod_{u=0}^{l+m+1}(n+u)}$$

- Оцениваемые параметры имеют высокую дисперсию
- Оценены только 2 компоненты, только для $k < n/2$

Делаем по-другому

- Производим асимптотический анализ **всех** компонент
- Опираемся на кумулятивные статистики => меньше дисперсия
- Нет ограничения на k

Результаты

- **Теорема**: среднее нормированное количество циклов длины m равно:

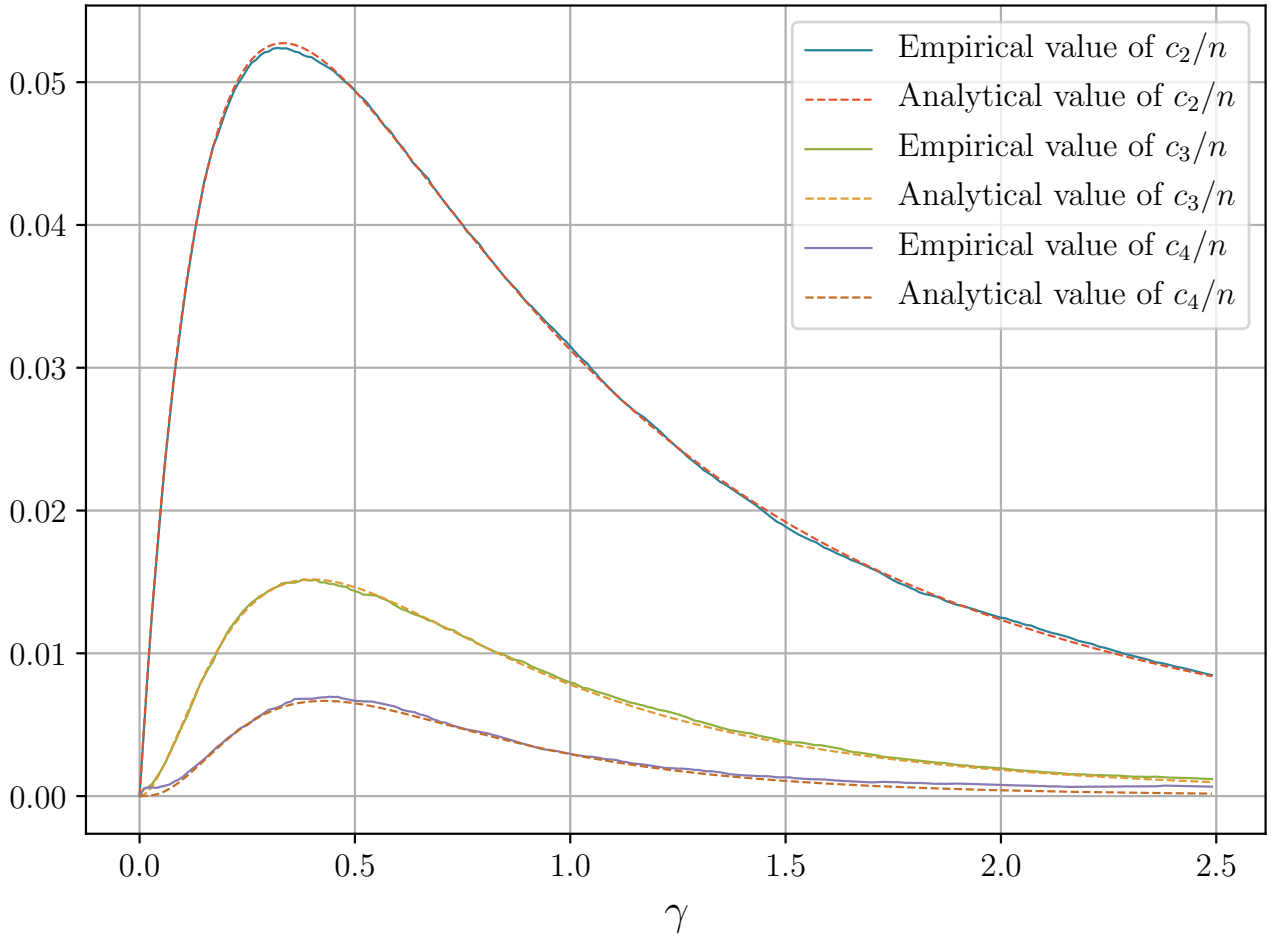
$$E\left(\frac{c_m}{n}\right) \xrightarrow[n\to\infty]{} \frac{(3m-3)!\gamma^{m-1}}{m!(2m-1)!(\gamma+1)^{3m-2}}$$

Схема доказательства

- Биекция задачи объединения в цикл с помеченными остовными деревьями, применение формулы Кэли
- Вычисление интеграла по соответствующей плотности вероятности

$$\int \cdots \int_{\mathbb{R}_+^m} \alpha_1 \cdot \ldots \cdot \alpha_m (\alpha_1 + \ldots + \alpha_m)^{m-2} e^{-\sum_{i=1}^m ((\gamma+1)\alpha_i)} \mathrm{d}\alpha_1 \ldots \mathrm{d}\alpha_m = \frac{(3m-3)!}{(2m-1)!(\gamma+1)^{3m-2}}$$

Соответствие эмпирических и теоретических результатов



	Метод Танье	Наш метод
Среднее время работы	3.02 сек.	0.00017 сек.
Средний модуль ошибки для $k < n/2$	1.99 %	0.68 %
Работа при $k \geq n/2$	Нет	Да
Реализация	Система нелин. урав., мод. градиентный спуск	Вещественный двоичный поиск
Оценка на c_1	$\sum_{l=0}^{\infty} \frac{(-2k)^l}{\prod_{u=0}^{l-1}(n+u)}$	$\frac{n^2}{2k+n}$
Оценка на c_2	$kn^2 \sum_{l=0}^{\infty} \sum_{m=0}^{\infty} \frac{(-2(k-1))^{l+m}(l+1)(m+1)}{\prod_{u=0}^{l+m+1}(n+u)}$	$\frac{kn^4}{(2k+n)^4}$
Оценка на c_m	Нет	Да

Применение к реальным данным

Пара геномов	Минимальное расстояние	Наш метод	Танье	Равновероятный метод
Prunus — Fragaria	273	297	284	283
Prunus — Malus	261	263	258	261
Fragaria — Malus	414	461	426	435

- *Whole genome comparisons of Fragaria, Prunus and Malus reveal different modes of evolution between Rosaceous subfamilies, 2012*
- Истинное расстояние отличается от минимального на 11%
- Уточнена граница парсимонии

Выводы

- Проведен теоретический анализ модели
  - Теорема 1 $E\left(\frac{c_m}{n}\right) \sim \frac{\gamma^{m-1}}{m!}\alpha_1 \dots \alpha_m (\alpha_1 + \dots + \alpha_m)^{m-2} e^{-\sum_{i=1}^{m} \alpha_i}$
  - Теорема 2 $E\left(\frac{c_m}{n}\right) \xrightarrow[n\to\infty]{} \frac{(3m-3)!\gamma^{m-1}}{m!(2m-1)!(\gamma+1)^{3m-2}}$
  - Уточнена граница применимости метода парсимонии
- Проведен эмпирический анализ модели
- Показана высокая согласованность эмпир. и теор. результатов
- Предложен более совершенный метод оценки
- Разработанный метод применён к реальным данным
- Оценки эволюционного расстояния уточнены на $\sim 10\%$
- Статья готовиться к подаче на конференцию RECOMB Comparative Genomics

Спасибо за внимание!

Вопросы?

ITMO more than a UNIVERSITY