УНИВЕРСИТЕТ ИТМО

Анализ геномных перестроек с помощью случайных графов

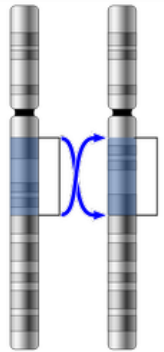
Забелкин А.А.
Науный руководитель: Алексеев Н.В.

Актуальность задачи

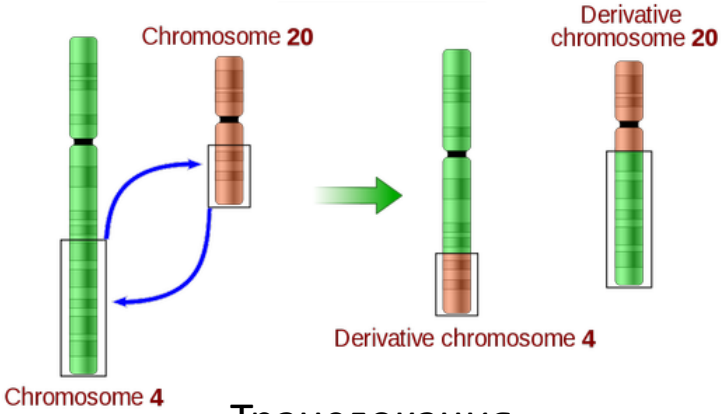
- Хотим знать реальное расстояние между видами
- Важно для многих филогенетических исследований
  - *M. Alekseyev, P. Pevzner* Multi-break rearrangements and chromosomal evolution // Theoretical Computer Science
- В последние время появляется всё больше «полностью» собранных геномов

Геномные перестройки

- Инверсия
- Транслокация
- Слияние
- Расщепление
- Транспозиция

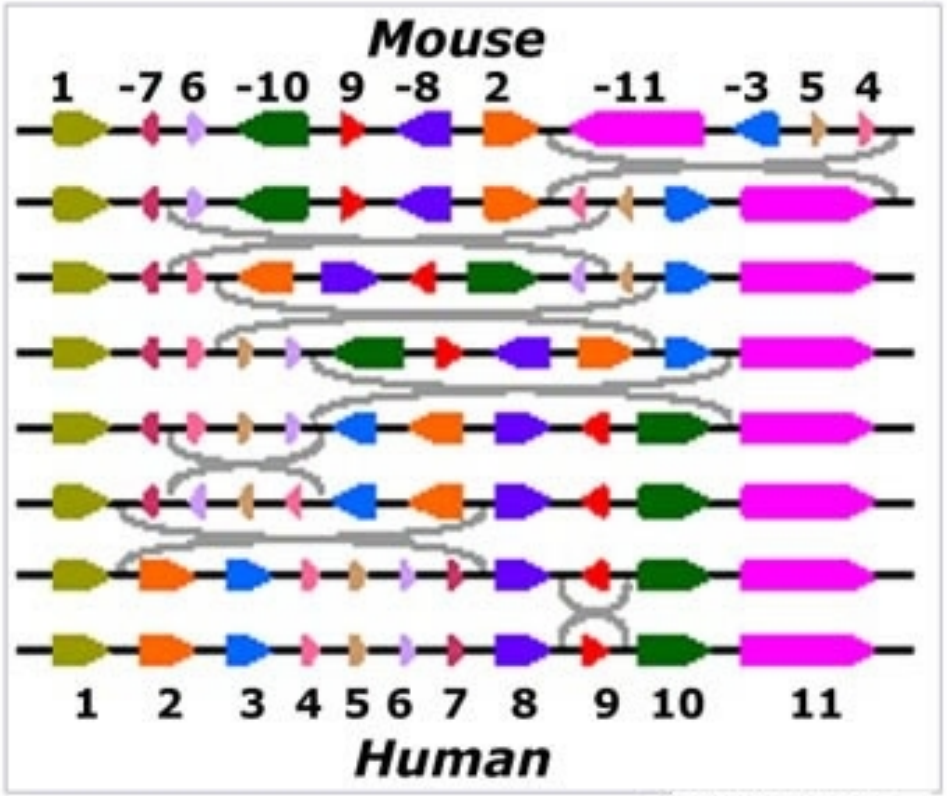


Инверсия



Транслокация

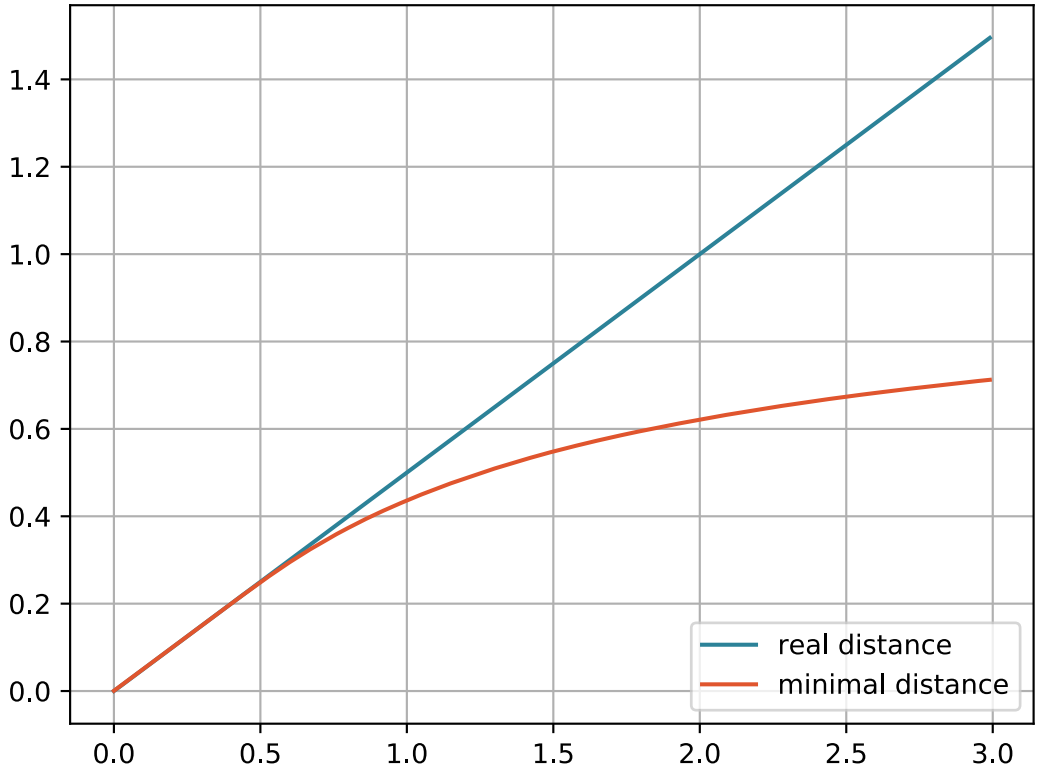
Эволюция генома



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

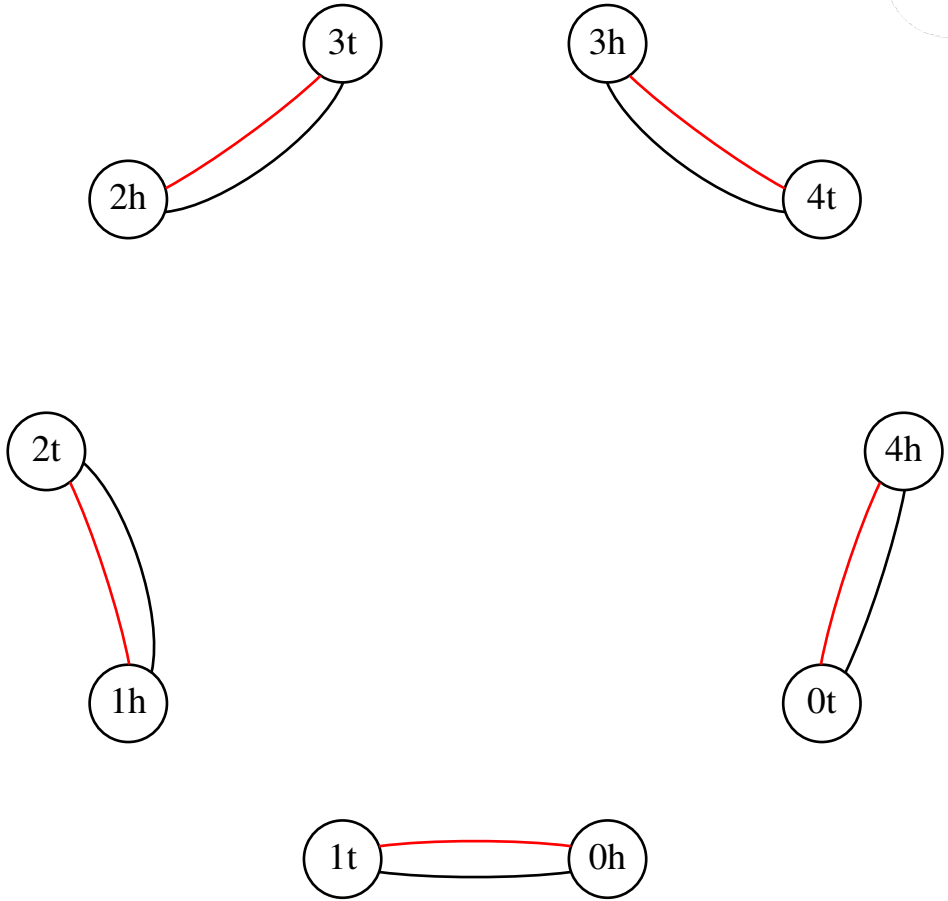
Могут сильно отличаться (нормировано от-но n)



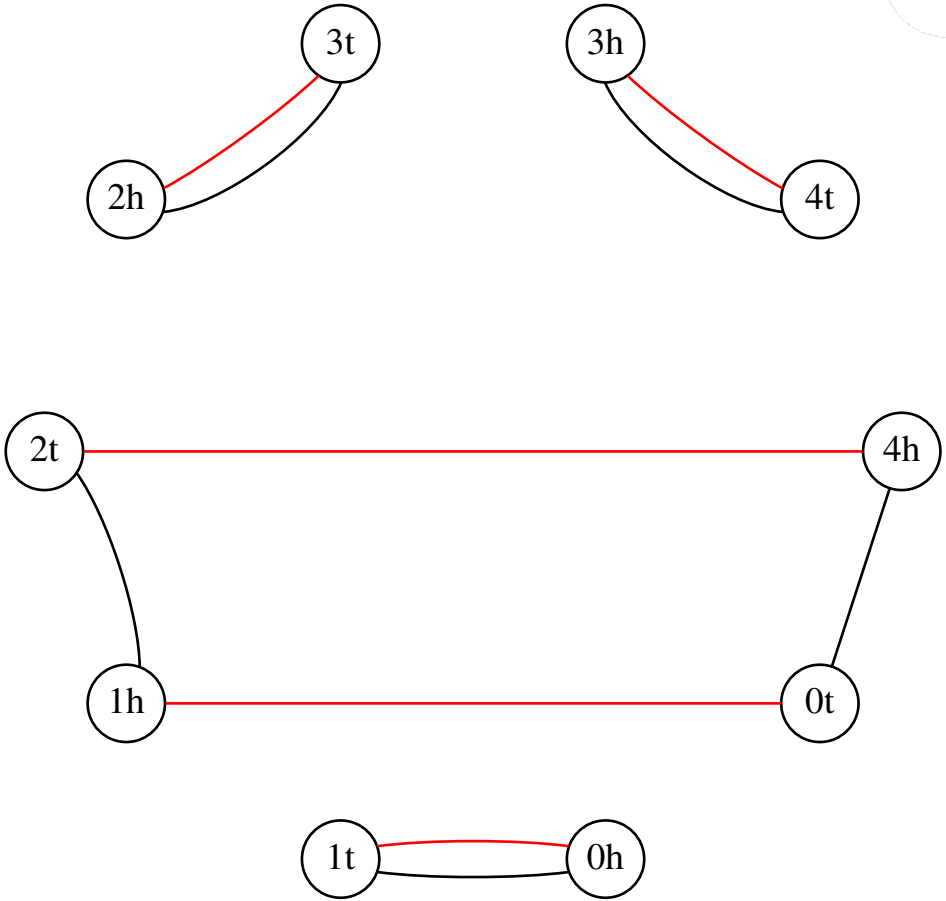
Граф точек разрыва

- Можем представить геном как граф
- Изначально оба генома одинаковы
- Моделируем на графе дискретный Марковский процесс
  - Чёрные рёбра фиксированы
  - Перестройки совершаются на красных ребрах
- k - число шагов

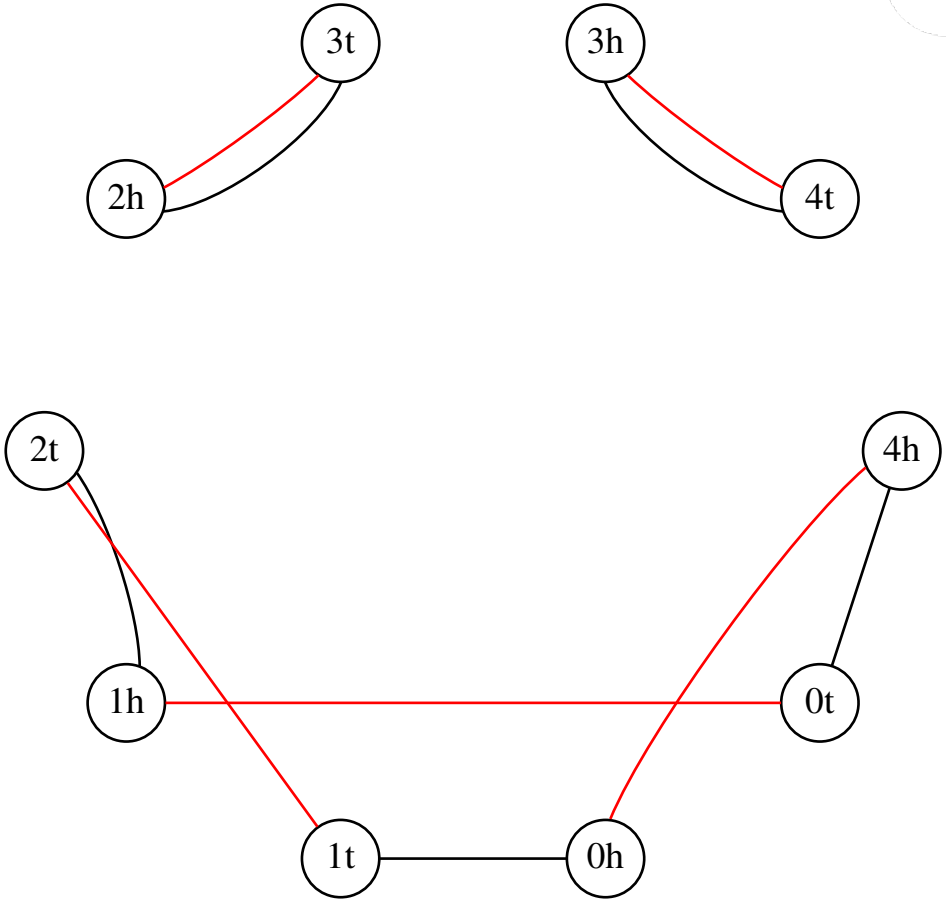
Марковский процесс k = 0



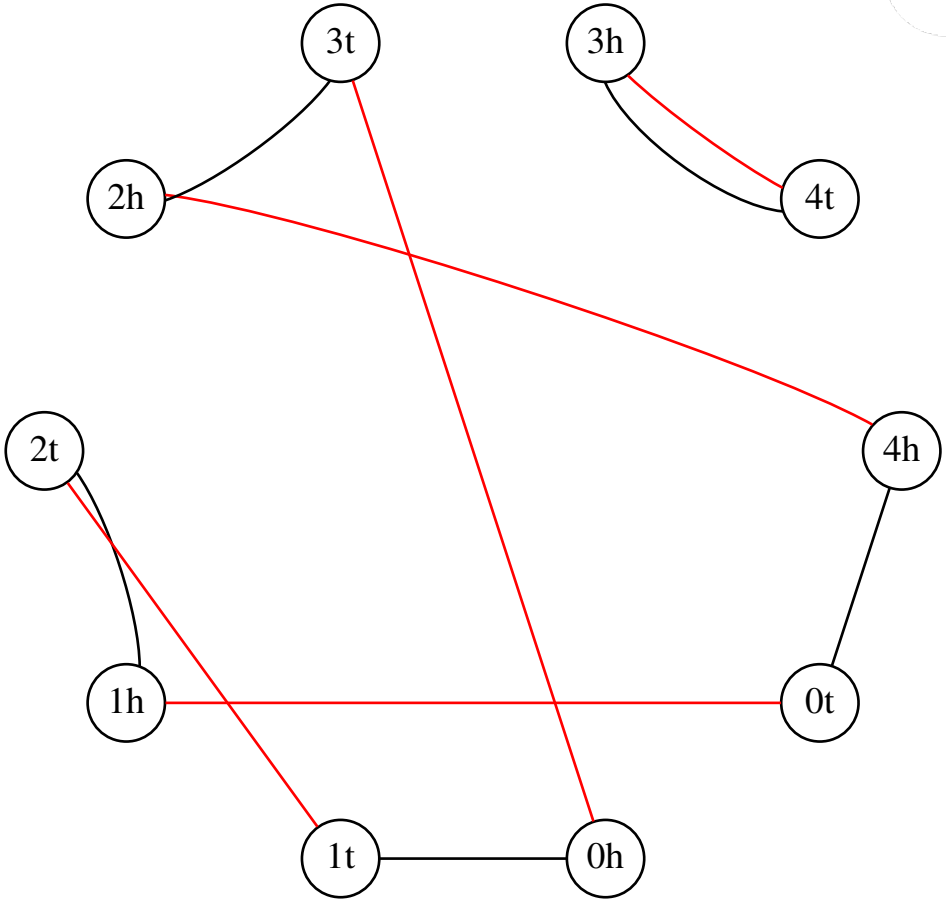
Марковский процесс k = 1



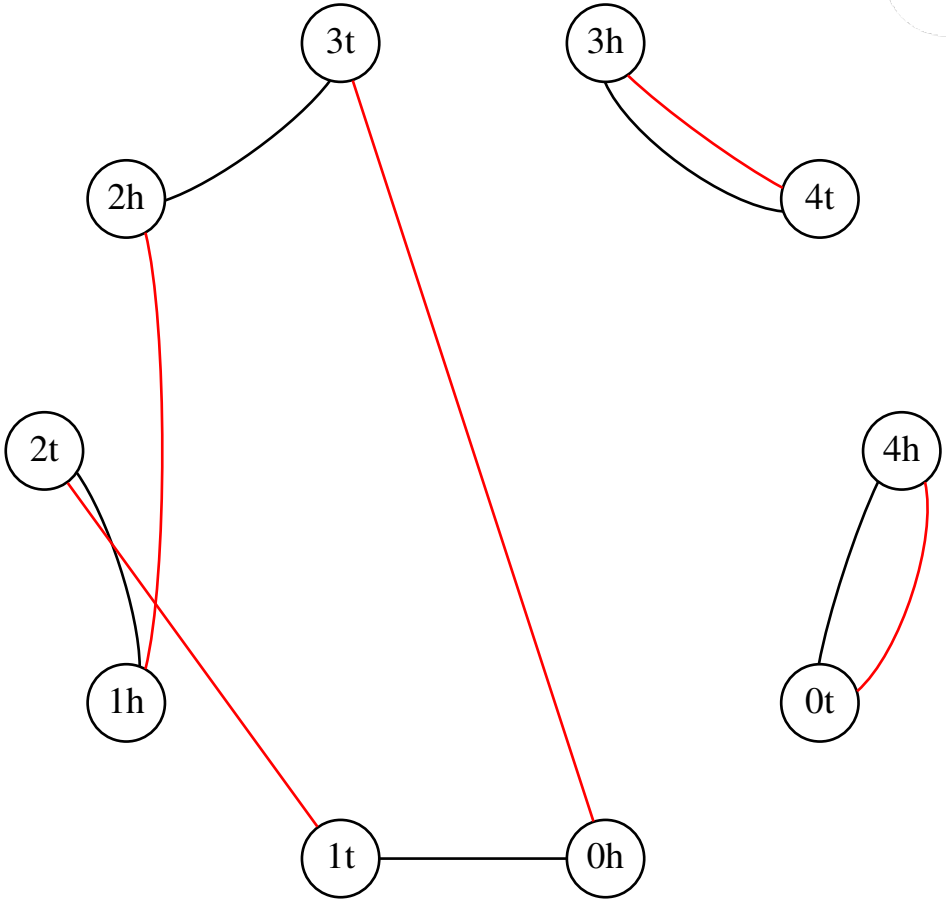
Марковский процесс k = 2



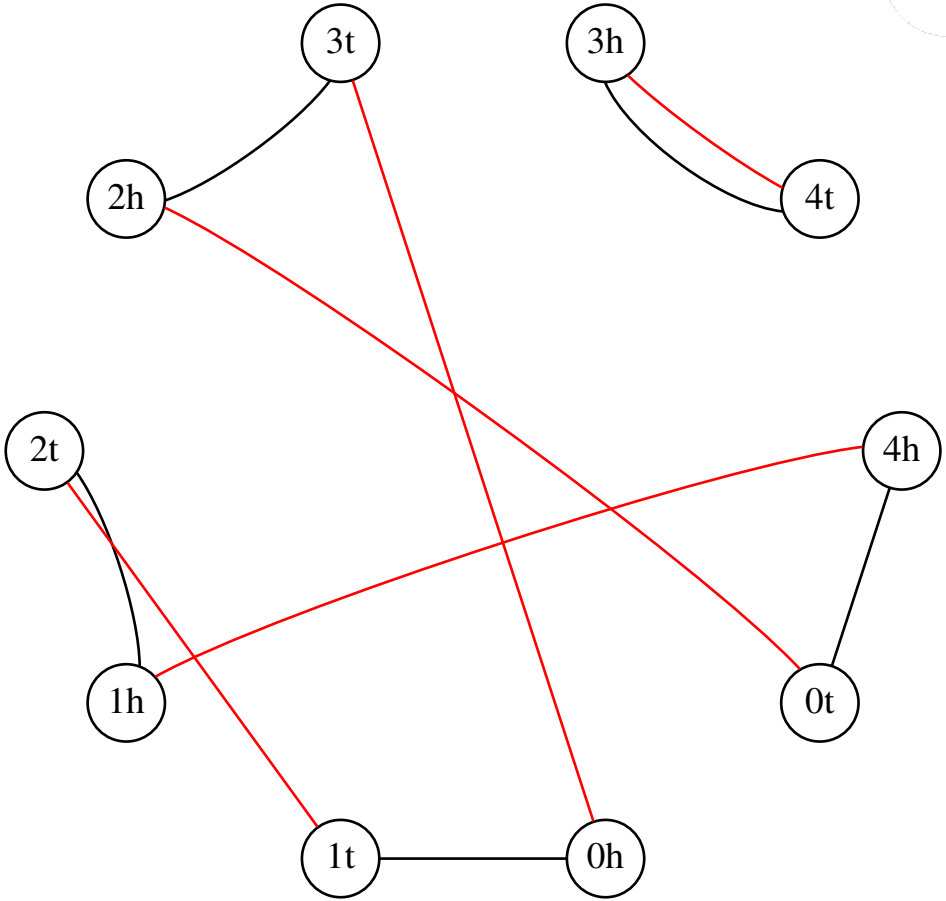
Марковский процесс k = 3



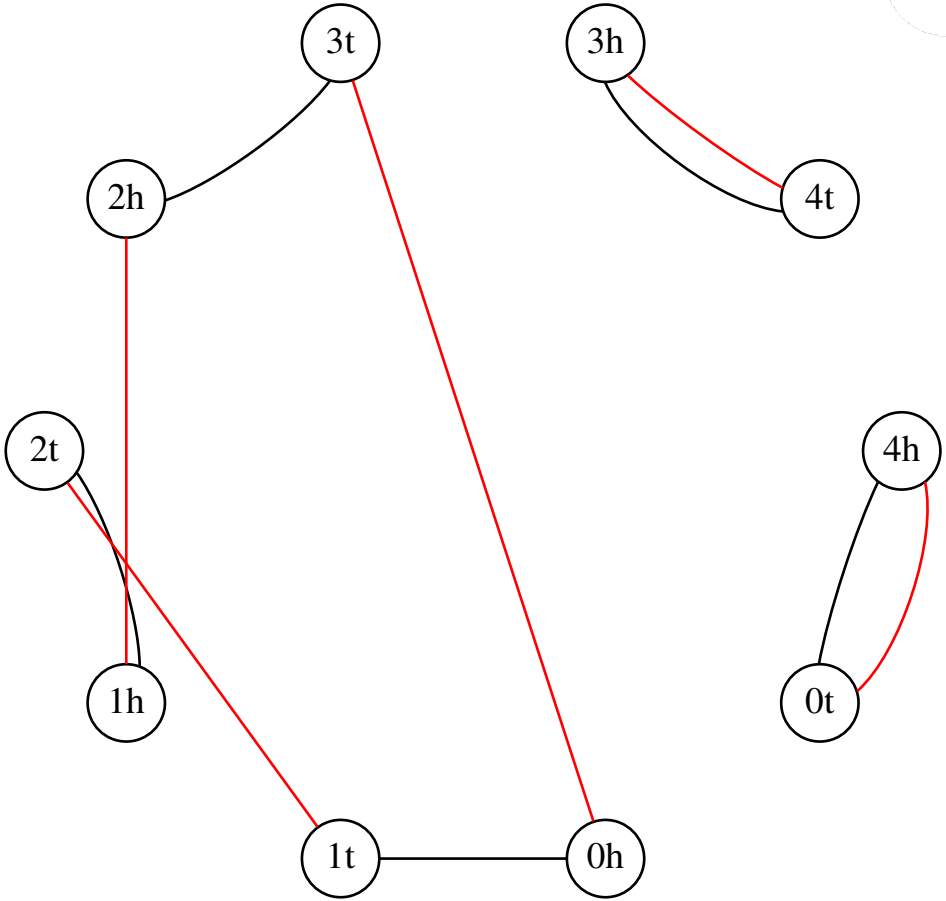
Марковский процесс k = 4



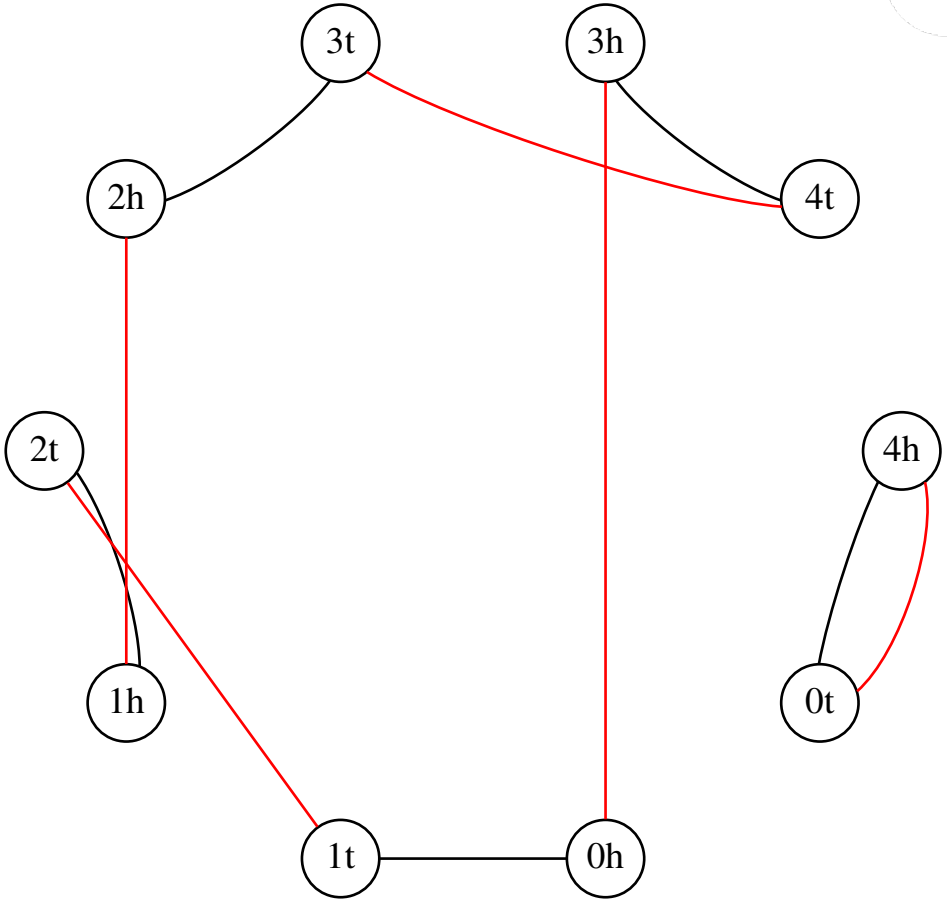
Марковский процесс k = 5



Марковский процесс k = 6



Марковский процесс k = 7



- Оцениваем необходимые компоненты
- $\frac{l \cdot c_l}{n} \approx e^{-\gamma l} \frac{(\gamma l)^{l-1}}{l!}$
- $\frac{b}{n} \approx 1 - e^{-\gamma}, \frac{d}{n} \approx 1 - \sum_{l=1}^{\infty} e^{-\gamma l} \frac{(\gamma l)^{l-1}}{l \cdot l!}$
- Предсказываем истинное эволюционное расстояние

Минусы подхода

- Breaking Good: Accounting for Fragility of Genomic Regions in Rearrangement Distance Estimation / P. Biller [и др.]
- Выбор рёбер происходит равновероятно
- В реальной разные регионы имею разную вероятность быть вовлеченными в перестройку

Модель Дирихле

- Каждому ребру сопоставляется вероятность p_i быть вовлеченным в перестройку
- Изначально все веса распределены по Дирихле
- В процессе перестройки необходимо правильно перераспределять веса
- Правильное перераспределение весов сохраняет распределение

Почему Дирихле?

- Это равномерное распределение на $(n-1)$ - мерном симплексе, т.е. на векторах $(x_1, x_2, \dots, x_n)$: $\sum x_i = 1$
  - Дробление отрезка длинной 1 на n частей
- В пределе Пуассон-Дирихле
  - Логарифмы простых сомножителей числа
  - Циклы в перестановках

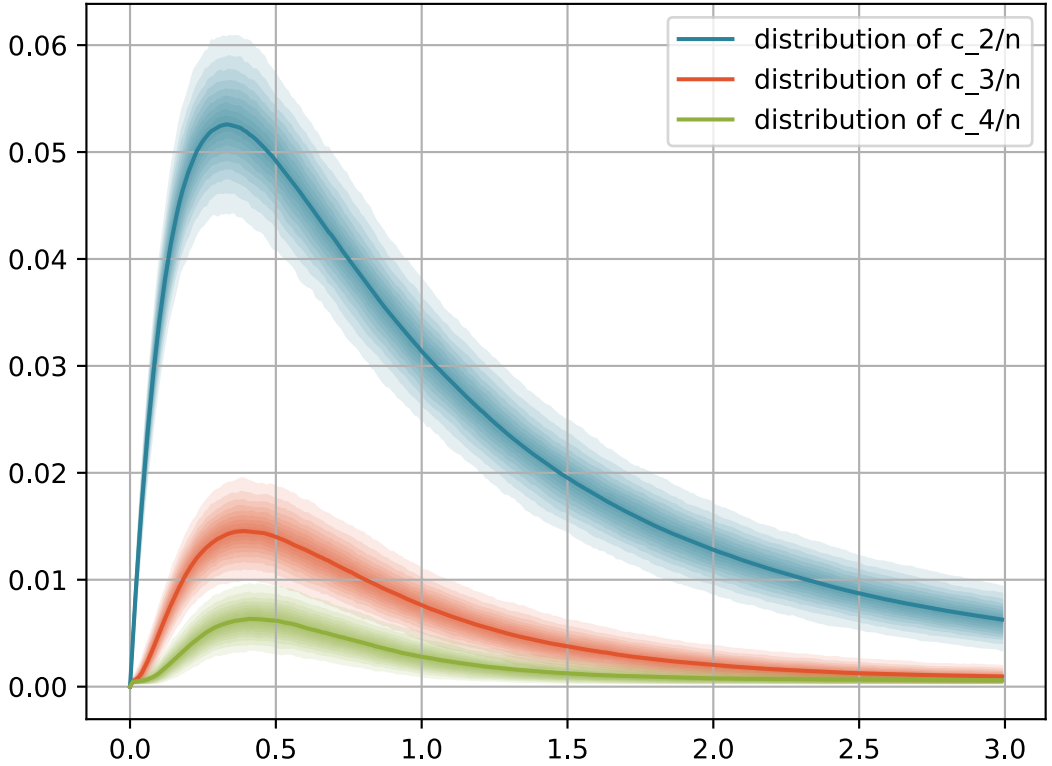
Минусы подхода, предложенного в статье

- Формулы для оценки сложно применить практически
- Оцениваемые параметры имеют высокую дисперсию

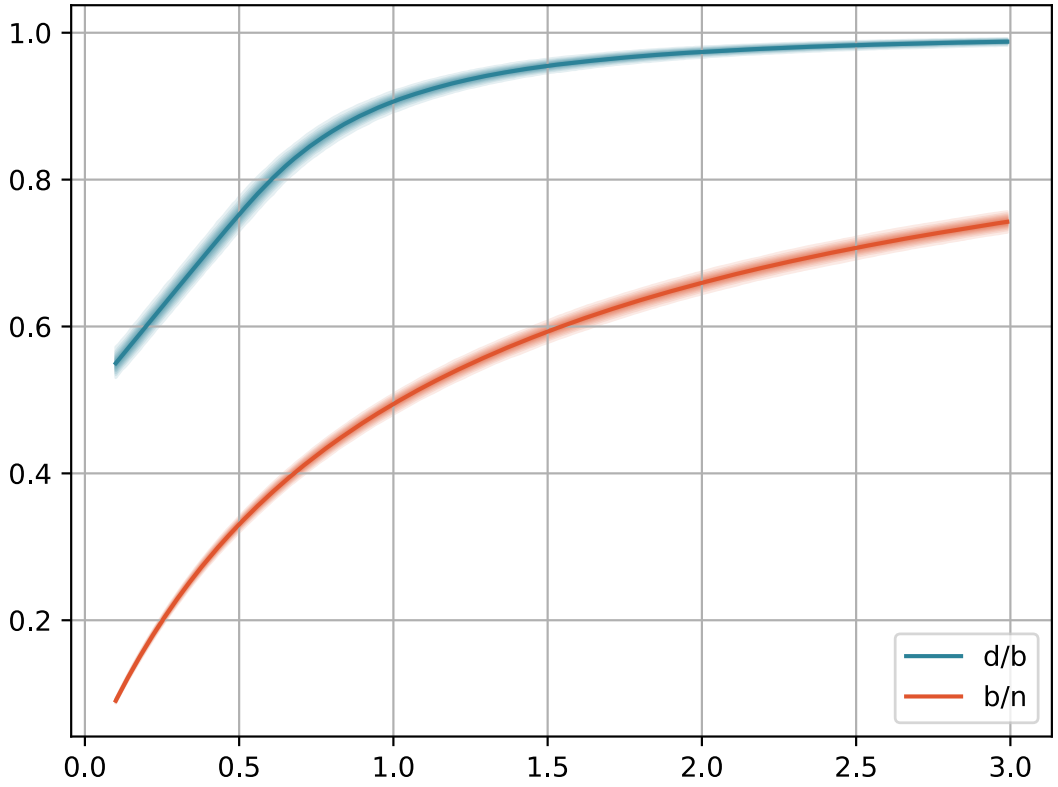
Делаем по-другому

- Опираемся на кумулятивные статистики
- Меньшая дисперсия, лучший результат
- Можно применить практически

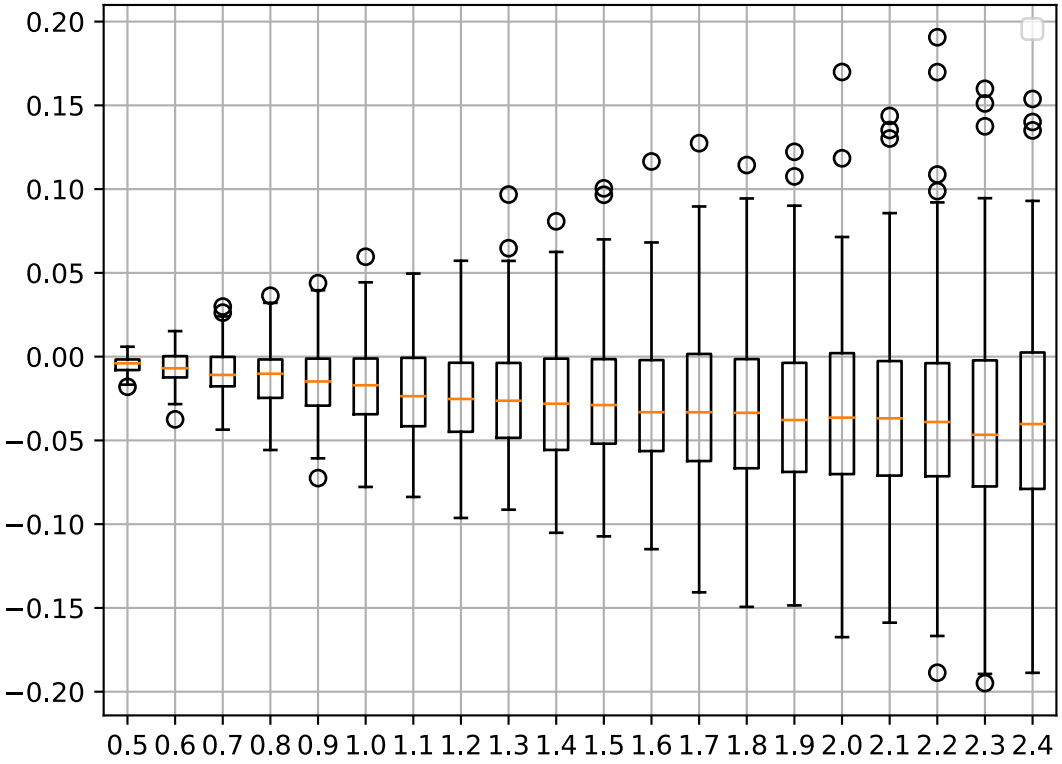
Эмпирический анализ необходимых комнонент



Кумулятивные статистики

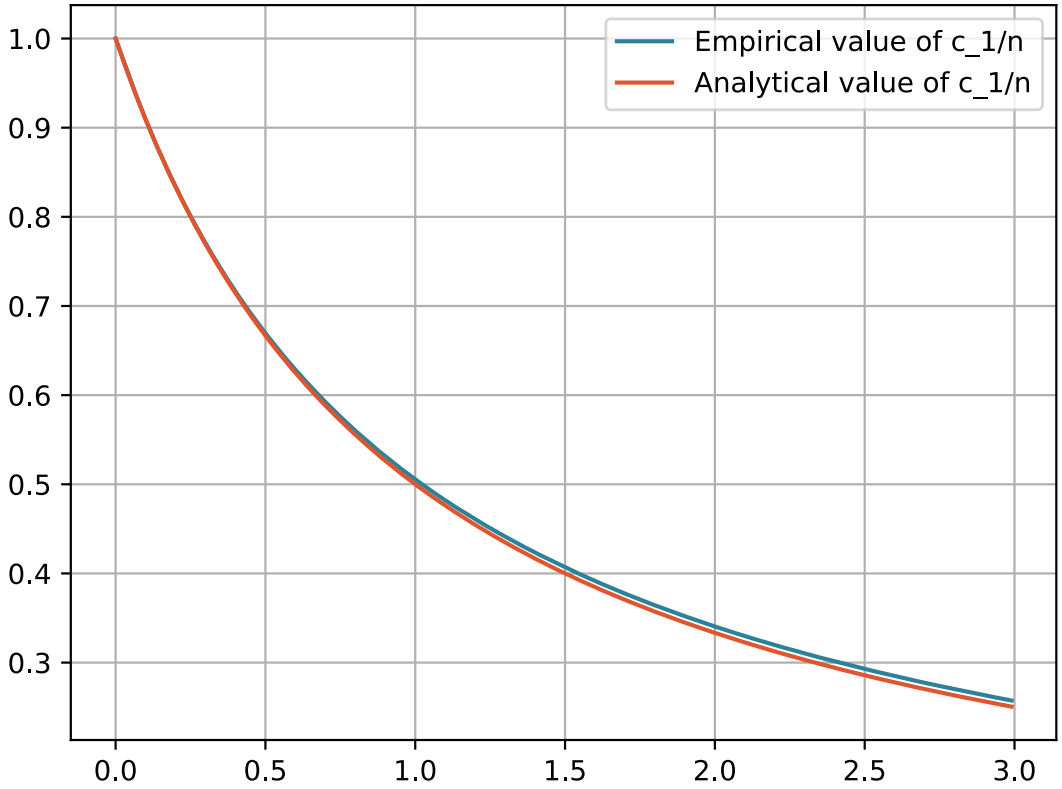


Алгоритм оценки истинного расстояния



Результаты

- На основании эмпирических данных реализован алгоритм оценивания истинного эволюционного расстояния
  - Для $k < n$, в более 90% случаях ошибка составляет не более 10%
  - Для больших k может достигать 15%
- Аналитически получена формула для количества циклов длины 1


Empirical value of c_1/n
Analytical value of c_1/n
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.0
0.5
1.0
1.5
2.0
2.5
3.0

Планы

- Теоретический анализ модели и оценка всех необходимых компонент

Спасибо за внимание!

ITMOre than a UNIVERSITY