**Министерство образования и науки Российской Федерации**
**ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ**
**«САНКТ-ПЕТЕРБУРГСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ»**

**ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА**

# «Анализ геномных перестроек с помощью случайных графов»

Автор: Забелкин Алексей Андреевич _____
Направление подготовки: 01.03.02 Прикладная математика и информатика
Квалификация: Бакалавр
Руководитель: Алексеев Н.В., канд. физ.-мат. наук _____

**К защите допустить**
Зав. кафедрой КТ Васильев В.Н., докт. техн. наук, проф._____

«___» _____ 20___ г.

Санкт-Петербург, 2018 г.

**Студент** Забелкин А.А. **Группа** М3436 **Кафедра** КТ **Факультет** ИТиП

**Направленность (профиль), специализация**
Математические модели и алгоритмы в разработке программного обеспечения

ВКР принята «___» _____ 20___ г.

Оригинальность ВКР _____%

ВКР выполнена с оценкой _____

Дата защиты «___» _____ 20___ г.

Секретарь ГЭК *Павлова О.Н.* _____

Листов хранения _____

Демонстрационных материалов/Чертежей хранения *отсутствуют*

## ОГЛАВЛЕНИЕ

# ВВЕДЕНИЕ

Для многих филогенетических исследований, изучающих эволюционные связи, является важной возможность оценивать эволюционное расстояние между различными видами. Самое первое решение подобной задачи заключалось в оценке минимального расстояния, которое требуется для преобразования одного генома в другой. Предположение о том, что для преобразования одного генома в другой было сделано минимальное число перестроек, называется предположением парсимонии.

Однако в реальном процессе эволюции данное предположение может не быть выполнено, поэтому необходимо иметь оценку, которая не будет опираться на предположение парсимонии. Подобную оценку, которая оценивает реальное расстояние между двумя геномами, а не минимальное, принято называть истинным эволюционным расстоянием. Данный термин впервые был предложен в [1].

На данный момент уже существует несколько методов оценки эволюционнного расстояния. В [2] показано, что истинное количество геномных перестроек между некоторыми видами дрожжей отличается от минимально возможного на 20%, а также разработан метод для оценки этого истинного расстояния. При этом в работе существенно использовали модель случайных графов Эрдеша-Реньи [3]. В статье [4] была высказана критика данной модели. В ней все геномные перестройки происходят равновероятно. Но в реальной жизни некоторые регионы имеют больший шанс быть вовлеченными в перестройку, а некоторые меньшую.

В данной работе рассмотрена новая модель генома, предложенная в [4]. Эта модель учитывает факт того, что разные регионы генома подвержены перестройкам в разной степени. Также эта модель является «хрупкой» — это означает, что только определенные «хрупкие» геномные области подвержены перестройкам. Проведён эмпирический и теоретический анализ данной модели. Построен алгоритм оценивания истинного эволюционного расстояния и проведено его сравнение с другими подходами.

# ГЛАВА 1. ОБЗОР СОВРЕМЕННЫХ РЕЗУЛЬТАТОВ В ОБЛАСТИ СРАВНИТЕЛЬНОЙ ГЕНОМИКИ И ГЕНОМНЫХ ПЕРЕСТРОЕК

## 1.1. Виды геномных перестроек

Геномные перестройки (англ. *genome rearrangements*) — это эволюционные события, которые меняют порядок генов на хромосомах. Некоторые хромосомные области более подвержены перестройкам, чем другие. Эта неустойчивость, как правило, обусловленая склонностью этих областей к смещению во время восстановления ДНК, усугубляется дефектами появления реплицирующих белков, которые повсеместно влияют на целостность генома.

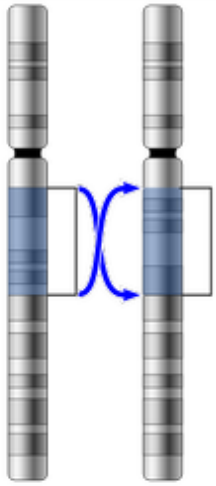


Рисунок 1 – Реверсия

Самые распространенные геномными перестройками являются:

а) Реверсия (инверсия) (англ. *reversal*) — разворот сегмента (рис. 1);
б) Транслокация — (анлг. *translocation*) попарная замена сегментов двух хромосом;
в) Слияние (англ. *fusion*) — объединение двух хромосом в одну;
г) Расщепление (англ. *fission*) — разделение одной хромосомы на две.

Все эти четыре типа перестановок могут быть смоделированы операцией Двойное-Разрезание-и-Склеивание (ДРС) (англ. *Double-Cut-and-Join*) [5], которая «разрезает» хромосому в двух локациях и склеивает полученные регионы в другом порядке.

Также существуют более сложные геномные перестройки, которые могут быть смоделированы как разрез хромосом в $k$ локациях, где $k > 2$, и склеивание полученных регионов в другом порядке. Например, транспозиция (англ.

*transposition*). В рамках данной работы эти перестройки рассматриваться не будут ввиду того, что происходят они достаточно редко, но при этом сильно усложняют получившуюся модель. Рассмотрены будут все виды геномных перестроек, моделируемые с помощью ДРС, то есть $k = 2$.

## 1.2. Существующие методы оценки эволюционного расстояния

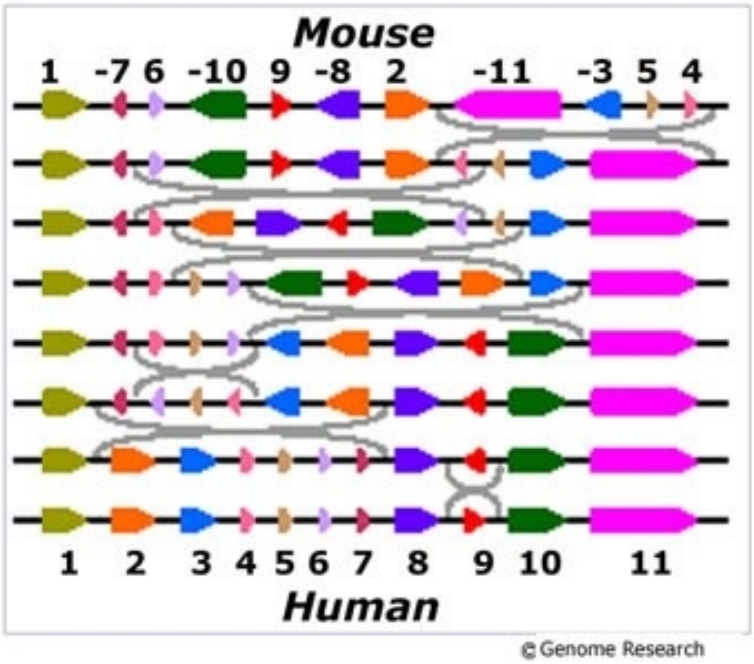



Рисунок 2 – Эволюционный сценарий

Для любой оценки расстояния между двумя геномами мы предполагаем, что они имеют в себе одинаковый набор блоков, организованных в разном порядке. Задача оценки расстояния сводится к задаче об оценке необходимого количества операций ДРС на соответствующих данным геномам графах.

### 1.2.1. Оценка через минимальное расстояние

Первым был предложен подход оценки расстояния как минимального необходимого, то есть минимального количества необходимых операций ДРС для перестройки одного генома в другой. Появление данного подхода весьма закономерно и на данный момент используется в качестве нижней границы для оценки результатов. Для геномов, которые слабо удалены друг от друга, данный метод даёт хорошие результаты.

Минимальное растоние определяется через «граф точек разрыва» (англ. *breakpoint graph*. Для восстановления эволюционного сценария решается задача сортировки. Подобная задача сортировки с ипользованием только инверсий является NP-полной [6].

Но если рассматривать геномы, которые достаточно далеко удалены друг от друга, то ошибка даного алгоритма становится заметно выше. Когда в реальности между генами могло произойти весьма большее количество перестроек, а природа их происхождения является случайной и не ведёт самым коротким путём, алгоритм оценки через минимальное расстояние будет же искать минимальный путь, который может отличаться от реального.

При большом количестве шагов ошибка данного метода в рамках модели, предложенной в данной работе, может достигать $50\,\%$. Но необходимо отметить, что с увеличением числа шагов происходит «насыщение» (англ. *saturation*) модели, и любой алгоритм оценки расстояния начинает давать более плохие результаты. В рамках построенной модели мы будем сравнивать получившийся алгоритм оценки с методом оценки через минимальное расстояние.

### 1.2.2. Модель поломки случайных регионов

В статье [1] был предложен метод для оценки истинного эволюционного расстояния, а также преложен сам термин «истинного эволюционного расстояния» (англ. *true evolutionary distance*). Однако, данный метод полагает, что геномы могут быть поломаны в любой позиции с равной вероятностью, то есть весь геном является «хрупким». Данное предположение, известное как модель поломок случайных регионов (ПСР-модель) (англ. *random breakage model*) эволюции хромосом, было опровергнуто в пользу более строгой модели поломок хрупких регионов (ПХР-модели) (англ. *fragile breakage model*) [7], в которой утверждается, что только определенные «хрупкие» (англ. *fragile*) геномные области подвержены к перестановкам. ПХР-модель поддерживается многими недавними исследованиями различных геномов (например, [8]). ПСР-модель можно рассматривать как экстремальный случай ПХР-модели, где каждая геномная область является хрупкой.

Таким образом, хотя в данной статье и был предложен алгоритм оценки, он не учитывает того факта, что некоторые регионы генома могут быть «прочными» (англ. *solid*) и не имеют возможности сломаться. С точки зрения биологии это означает, что подобные изменения критичны и, скорее всего, ведут к гибели организма.

### 1.2.3. Модель поломки хрупких регионов

В статье [2] предложен новый метод оценки истинного эволюционного расстояния между двумя геномами в рамках ПХР-модели (модель поломки

хрупких регионов). Произведено оценивание предложенного метода для имитируемых геномов, которые показывают его высокую точность.

Для оценки истинного эволюционного расстояния в данной модели используется анализ так называемого «графа точек разрыва». Аналитически выведены формулы для распределения необходимых для оценки компонент и на основании этих формул преложен аналитический метод оценки истинного эволюционного расстояния.

Так как изучаемая нами модель является развитием и уточнением ПХР-модели, далее ПХР-модель будет рассмотрена подробнее.

## 1.3. Анализ модели поломки хрупких регионов

### 1.3.1. Граф точек разрыва и двойной-разрез-и-склеивание

Анализ начинается с круговых геномов и позже обращается к линейным геномам. Геном с $n$ блоками представляется в виде графа генома, состоящего из $n$ направленных блоковых рёбер (англ. *block edges*), кодирующих блоки и их границы, и $n$ неориентированных рёбер смежности (англ. *adjacency edges*), кодирующих смежности между блоками.

Пусть $P$ и $Q$ — геномы, содержащие один и тот же набор блоков. Предположим, что в их графах геномов рёбра смежности $P$ окрашены в черный цвет (рис. 3a), а рёбра смежности $Q$ окрашены в красный цвет (рис. 3b). Граф точек разрыва $G(P, Q)$ является суперпозицией графов генома $P$ и $Q$ с удаленными блоковыми рёбрами (рис. 3c). Черные и красные рёбра смежности в $G(P, Q)$ образуют совокупность чередующихся черно-красных циклов.

Будем говорить, что черно-красный цикл является $l$-циклом, если он содержит $l$ черных ребер (и $l$ красных), а $c_l(P, Q)$ - число $l$-циклов в $G(P, Q)$. Важно отметить, что мы считаем рёбра только одного цвета, то есть на самом деле мы называем длиной половину числа ребер. Циклы длины $1$ называют тривиальными, остальные нетривиальными. Вершины нетривиальных циклов называются точками разрыва (англ. *breakpoints*).

Операция ДРС в геноме $Q$ заменяет любую пару красных рёбер смежности $\{x, y\}, \{u, v\}$ другой парой ребер, состоящих из тех же вершин, то есть: $\{x, u\}, \{y, v\}$ либо $\{u, y\}, \{v, x\}$. Говорится, что такая операция ДРС совершается на ребрах $\{x, y\}, \{u, v\}$ и их конечных точках (англ. *endpoints*) $x, y, u, v$. Операция ДРС в геноме $Q$, превращающая его в геном $Q'$, соответствует преобразованию графа точек разрава $G(P, Q)$ в граф точек разрыва $G(P, Q')$ (рис. 4).

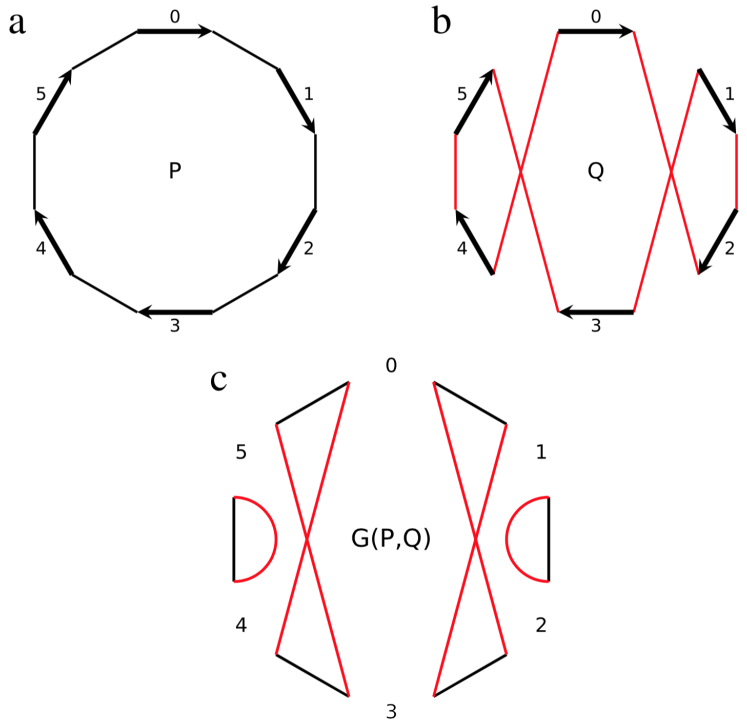



Рисунок 3 – **a** Геномный граф однохромосомного генома $P = (0, 1, 2, 3, 4, 5)$ с рёбрами смежности, окрашенными в черный цвет; **b** Геномный граф однохромосомного генома $Q = (0, -2, -1, 3, -5, -4)$ с рёбрами смежности, окрашенными в красный цвет; **c** Граф точек разрыва $G(P, Q)$ геномов $P$ и $Q$ представляет собой набор черно-красных циклов

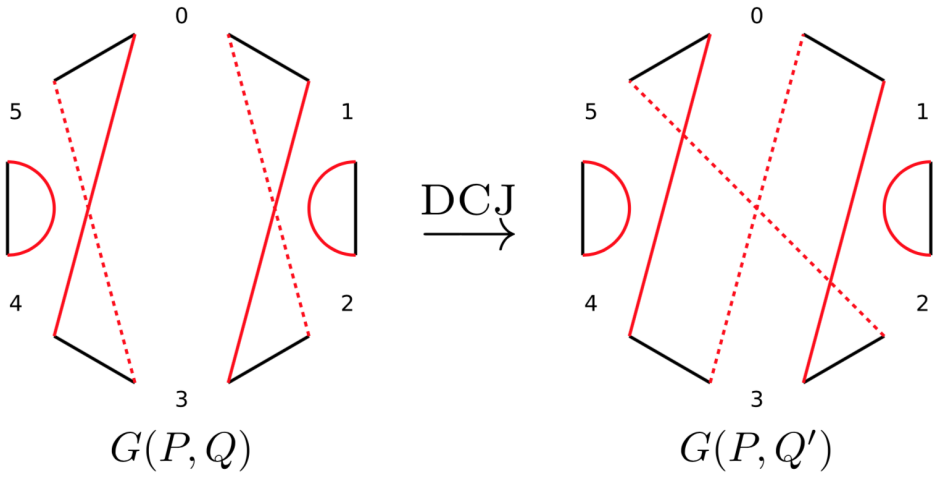



Рисунок 4 – операция ДРС в геноме $Q$ заменяет пару красных ребер в точке разреза $G(P, Q)$ другой парой красных ребер на тех же 4 вершинах

Если ребра $\{x, y\}, \{u, v\}$ принадлежат разным циклам, то операция ДРС, осуществляемая на графе точек разрыва, объединяет два цикла в один, а если ребра $\{x, y\}, \{u, v\}$ принадлежат к одному и тому же циклу, то операция ДРС разбивает один цикл на два или сохраняет текущее количество циклов. Минимальное

количество операций ДРС, которое потребуется для преобразования $Q$ в $P$ назовём расстоянием ДРС (обозначается $d(P,Q)$) между геномами $P$ и $Q$. Также введем ещё две компоненты: $b(P,Q) = \sum_{l \geqslant 2} l \cdot c_l(P,Q)$ — половина общей длины всех нетривиальных циклов, $c(P,Q) = \sum_{l \geqslant 2} c_l(P,Q)$ — количество циклов длины больше 1. Тогда для оценки на $d$ будет достаточно взять их разность: $d(P,Q) = b(P,Q) - c(P,Q)$.

### 1.3.2. Эволюционная модель

Задача эволюционной модели — оценка истинного эволюционного расстояния между геномами $P$ и $Q$. Будем считать, что геномы $P$ и $Q$ имеют один и тот же набор блоков, тогда в качестве процесса эволюции мы можем рассматривать дискретный марковский процесс. Каждая операция ДРС в таком процессе происходит независимо, то есть у модели нет «памяти», и с их помощью осуществляется последовательное превращение генома $P$ в геном $Q$. Начальной точкой данного процесса является геном $Z = P$, конечной точкой является геном $Z = Q$. Подобный процесс будет соответствовать преобразованию графа $G(P,P)$ в граф $G(P,Q)$. Истинное эволюционное расстояние между $P$ и $Q$ будем оценивать как количество операций ДРС в данном преобразовании (число $k$).

Также важным замечанием является то, что даже при наличии тривиальных циклов в $G(P,Q)$, их количество является неизвестным параметром. Это связано с тем, что мы не может точно сказать, является ли данный регион «прочным» или это «хрупкий» регион, который просто не был вовлечен в перестройку. Для того, чтобы учитывать этот факт будем считать, что геномы $P$ и $Q$ составлены из большого числа $n$ прочных регионов ($n$ неизвестно), сменяющихся хрупкими регионами, причём некоторые регионы могли сохраниться случайно. Подобные представления геномов $P$ и $Q$ обозначим $P_n$ и $Q_n$. Далее будем рассматривать преобразование генома $P_n$ в геном $Q_n$ последовательностью операцией ДРС, которые происходят только на хрупких регионах.

Также важно, что, хотя число $n$ прочных областей неизвестно, графы точек разрыва $G(P,Q)$ и $G(P_n,Q_n)$ имеют одну и ту же структуру циклов, за исключением тривиальных циклов. То есть, мы имеем $c_l(P_n,Q_n) = c_l(P,Q)$ для всех $l \geqslant 2$, что означает, в частности, $b(P_n,Q_n) = b(P,Q)$ и $c(P_n,Q_n) = c(P,Q)$, а следовательно $d(P_n,Q_n) = d(P,Q)$. Графы точек разрыва $G(P,Q)$ и $G(P_n,Q_n)$

могут различаться только числом тривиальных циклов, а в нашей модели эта компонента считается неизвестной.

В рамках рассматриваемой эволюционной модели следующие параметры считаются известными, то есть наблюдаемыми:

а) $c_l = c_l(P, Q)$ для $l \geqslant 2$ — число циклов длины $l$ в $G(P, Q)$;

б) $b = b(P, Q) = \sum_{l \geqslant 2} l \cdot c_l$ — половина от общей длины всех циклов где $l \geqslant 2$ в $G(P, Q)$;

в) $d = d(P, Q) = b - \sum_{l \geqslant 2} c_l(P, Q)$, — минимальное необходимое количество операций ДРС (для преобразования $P$ в $Q$).

Неизвестными, то есть ненаблюдаемыми, являются следующие параметры:

а) $c_1 = c_1(P_n, Q_n)$ — число тривиальных циклов в $G(P_n, Q_n)$;

б) $n$ — половина общей длины всех циклов в $G(P_n, Q_n)$ или число хрупких регионов в геномах $P$ и $Q$;

в) $k = k(P, Q)$ — чило ДРС операций в Марковском процессе или истинное эволюционное расстояние между $P$ и $Q$.

В качестве замечания необходимо отметить, что данная модель легко применяется к линейным геномам простым добавлением ребра между последним и первым блоками.

### 1.3.3. Теоретический анализ и оценка расстояния

В рамках данной модели основной задачей является оценка неизвестных параметров через известные. Для всех необходимых компонент аналитически выведем необходимые формулы.

Количество циклов заданной длины можно посчитать по формуле:

$$c'_{n,k,m} = \frac{\binom{k}{m-1}\binom{n-m}{2}^{k-m+1} m^{m-2} m!}{\binom{n}{2}^k},$$

где $n$ — число вершин, $k$ — число шагов, $m$ — длина цикла

Переходя к пределу, получим:

$$\frac{c_l}{n} = \frac{c'_{n,k,m}}{n} \xrightarrow[n\to\infty]{[k=\frac{n\gamma}{2}]} \frac{e^{-\gamma l}\gamma^{l-1} l^{l-2}}{l!},$$

где $\gamma = \frac{2k}{n}$ — число произошедших перестроек, нормированное относительно общего числа областей

Далее оцениваются нормированные величины $d$ и $b$:

$$\frac{b}{n} = 1 - e^{-\gamma} + o(1) \qquad \frac{d}{n} = 1 - \sum_{l=1}^{\infty} \frac{p_l}{l} + o(1),$$

$$p_l = e^{-\gamma l} \frac{(\gamma l)^{l-1}}{l!}$$

После этого мы можем оценить величину $\frac{d}{b}$ отдельно от $n$:

$$\frac{d}{b} \approx \frac{1 - \sum_{l=1}^{\infty} e^{-\gamma l} \frac{(\gamma l)^{l-1}}{l \cdot l!}}{1 - e^{-\gamma}}.$$

Но величины $d$ и $b$ известны в рамках данной модели, и мы знаем формулу зависимости $\gamma$ от этих величин. А также известно, как величина $\frac{b}{n}$ зависит от $\gamma$. Зная эти зависимости, получаем формулу для неизвестных нам $n$ и $k$:

$$n_e = \frac{b}{1 - e^{-\gamma_e}} \qquad k_e = \frac{\gamma_e \cdot n_e}{2},$$

где $n_e$ — оценка на количество хрупких регионов,

$k_e$ — оценка на истинное эволюционное расстояние

Для оценки качества работы данного метода проведём 200 независимых экспериментов с моделируемым Марковским процессом построим график вида «ящик с усами» (рис. 5).

На данном графике и во всех подобных графиках далее «ящикам» соответствует $50\,\%$ результатов, а усам $90\,\%$. Как мы видим, данный метод в $90\,\%$ случаев ошибается не более, чем на $7\,\%$, что является отличным показателем. Но, к сожалению, данная модель имеет и недостатки, критика была высказана в [4].

Дело в том, что выбор рёбер в Марковском процессе происходит равновероятно, в то время как в реальной жизни некоторые регионы имеют больший шанс быть вовлеченными в перестройку, а некоторые меньшую. Данное замечание послужило мотивацией для создания модели, представленной в данной работе.

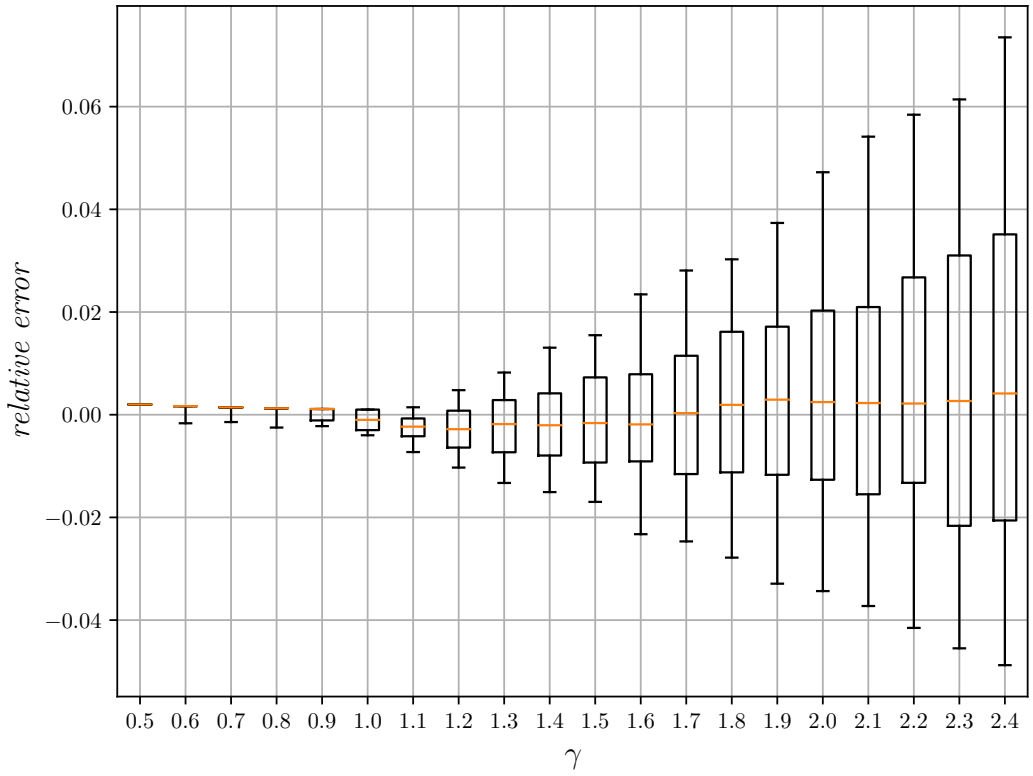



Рисунок 5 – Зависимость распределения относительной ошибки $\frac{k_e - k}{k}$ от $\gamma$

## 1.4. Описание модели Дирихле

### 1.4.1. Снабжения графа точек разрыва весами

Аналогично ПХР-модели модель Дирихле использует граф точек разрыва для представления геномов. В данной модели мы будем считать, что у каждого хрупкого региона есть некоторая вероятность быть вовлеченным в перестройку.

Пусть $n$ — это размер геномов $P$ и $Q$. Каждое красное ребро в графе снабдим соответствующим числом $p_i$ — его вероятностью быть вовлеченным в перестройку. $\sum_{i=0}^{n-1} p_i = 1$ по определению.

### 1.4.2. Модификация операции двойной-разрез-и-склеивание

Операция ДРС также производится на красных ребрах. Но теперь помимо рёбер смежности она принимает и вероятности, подписанные на этих ребрах.

Пусть операция ДРС производится на рёбрах с номерами $i$ и $j$ и соответствующие рёбра смежности — это $\{x, y\}, \{u, v\}$, а вероятности — $p_i$ и $p_j$. Эта операция аналогично заменяет данные рёбра другой парой рёбер, образующих паросочетания на тех же вершинах, что и исходные, то есть $\{x, u\}, \{y, v\}$ либо $\{u, y\}, \{v, x\}$. Также перераспределяются и веса. Новые веса $p'_i$ и $p'_j$ будут равны $r_1 \cdot p_i + r_2 \cdot p_j$ и $(1 - r_1) \cdot p_i + (1 - r_2) \cdot p_j$ соответственно, где $r_1$ и $r_2$ — случайные числа из отрезка $(0, 1)$.

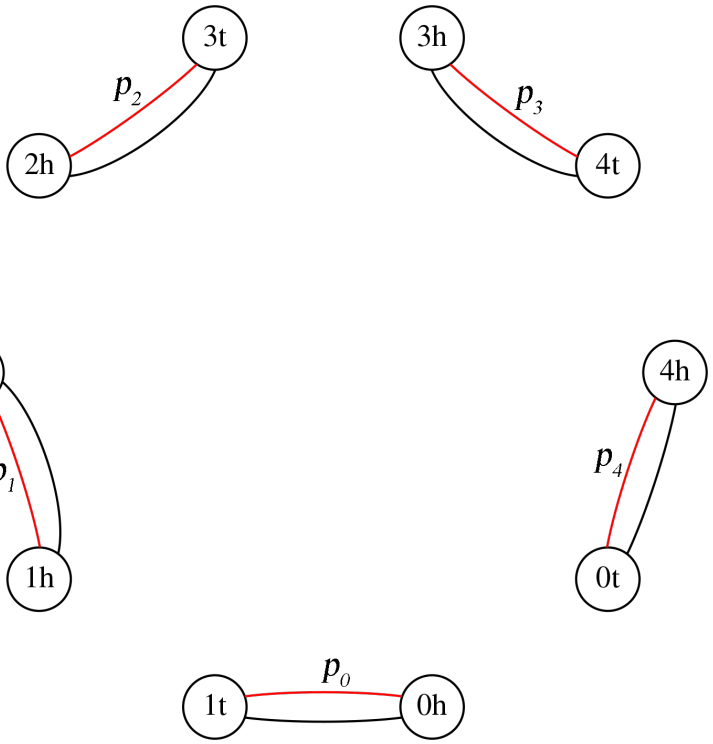



Рисунок 6 – Граф точек разрыва с весами

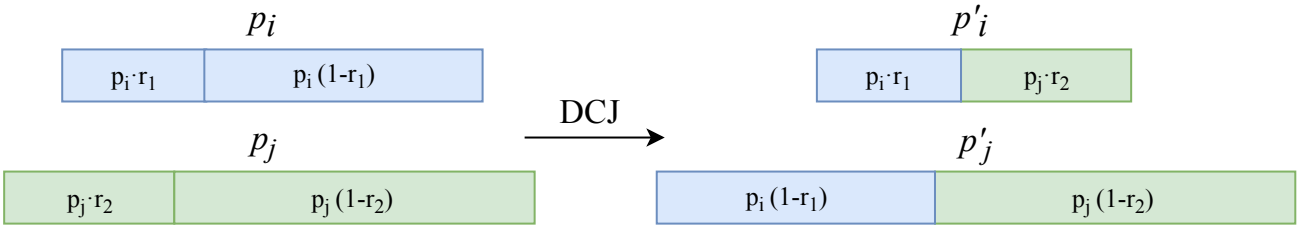



Рисунок 7 – Пример перераспределения весов

Указанное перераспределение весов предложено в статье [4], и соответствующие веса и правила перераспределения весов обусловлены биологически. В реальной жизни вероятность поломки какого-либо региона зависит от множества факторов. Геном имеет достаточно сложную трёхмерную структуру [9] и вероятность поломки региона тесно связана с физическим устройством генома. В данной работе мы придерживаемся мнения о том, что хрупкие регионы являются участками открытого хроматина. В предположении, что хрупкие регионы являются участками открытого хроматина, хрупкий регион может сломаться в любом своём месте. Считается, что вероятность быть вовлеченным в перестройку пропорциональна длине этого региона.

Каждый регион, вовлеченный в перестройку, равновероятно ломается в случайном месте своей длины, что соответствует числам $r_1$ и $r_2$. И далее веса

перераспределяются в соответствии с местами этих поломок, сохраняя общую сумму вероятностей (для примера см. Рис. 7).

### 1.4.3. Эволюционная модель и равновесное распределение

Как и в ПХР-модели, рассмотренной ранее, для оценки истинного эволюционного расстояния между геномами $P$ и $Q$, которые имеют один и тот же набор блоков, будем рассматривать процесс эволюции как дискретный Марковский процесс, который начинается с генома $P$ и заканчивается геномом $Q$. Подобное преобразование, как и ранее, осуществляется с помощью последовательности операций ДРС.

Отличие заключается в том, что на каждом шаге Марковского процесса выбор ребра происходит не равновероятно, а соответствует вероятностям, написанным на соответствующих рёбрах. То есть вероятность того, что 2 ребра с номерами $i$ и $j$ будут вовлечены в перестройку, будет равна $p_i \cdot p_j$ (выбор происходит независимо).

Полученный марковский процесс является:

а) Реверсивным, то есть время возвращения в некоторое состояние имеет конечное математическое ожидание;

б) Непериодичным, так как существует ненулевая вероятность остаться в текущем состоянии;

в) Неразложимым, то есть любое состояние процесса может быть достигнуто из любого другого состояния за конечное число шагов. Это свойство может быть проверено простым упорядочиванием состояний и математической индукцией.

Следовательно, этот марковский процесс сходится [10]. Как показано в [4], стационарное распределение этого процесса – это равномерное распределение на всех векторах $p = \{p_i\}$, сумма которых равна 1. Это распределение является плоским распределением Дирихле.

Для того, чтобы получить данное распределение, достаточно выбрать отдельные вероятности как распределенные по экспоненциальному закону и нормализовать [11]:

$$\text{Пусть } i \in \{1, 2, \ldots, n\}, \qquad \alpha_i = Exp(1), \qquad M = \sum_{i=0}^{n} \alpha_i,$$

тогда $(p_1, p_2, \ldots, p_n) = \left(\frac{\alpha_1}{M}, \frac{\alpha_2}{M}, \ldots, \frac{\alpha_n}{M}\right) = Dir(1, 1, \ldots, 1)$.

## Выводы по главе 1

В главе 1 были рассмотрены известные на данный момент методы для оценки эволюционного расстояния: оценка через минимальное расстояние, модель поломки случайных регионов, модель поломки хрупких регионов. Несмотря на то, что в рамках рассмотренных моделей можно получать достаточно точные результаты относительно истинного эволюционного расстояния, все они не учитывают некоторых биологических особенностей генома, например, факт того, что разные хрупкие регионы могут иметь разную вероятность быть вовлеченными в перестройку.

Была описана модель устройства генома, более точно учитывающая структуру ДНК (модель Дирихле). В дальнейшем будет проведён анализ именно этой модели.

# ГЛАВА 2. АНАЛИЗ МОДЕЛИ ДИРИХЛЕ

## 2.1. Математическое ожидание числа циклов заданной длины

Одним из самых важных параметров модели является число циклов заданной длины. Для простоты изложения сначала будут рассмотрены циклы с конкретными и малыми длинами (1 и 2). Далее эти рассуждения будут обобщены на циклы произвольной длины $m$.

### 2.1.1. Подсчет циклов длины один

Для начала оценим число циклов единичной длины. То есть это либо те циклы, которые не были затронуты в марковском процессе, либо те, которые были получены из рапада цикла большой длины.

Здесь и далее будем считать, что циклы заданной длины образуются только в результате слияния циклов меньшей длины, так как второй сценарий намного менее вероятен. Этот факт подтверждается сравнением аналитических и эмпирических результатов, которое будет приведено впоследствии. И это сравнение позволяет оценить погрешность, которая получается в условиях такого предположения. Это погрешность оказывается пренебрежимо малой.

Также считаем, что число хрупких регионов $n$ достаточно большое, и впоследствии переходим к пределу по $n$; в то же время, длины циклов $m$ полагаем фиксированными. Полагаем, что $k$ и $m$ имеют один порядок, а именно:

$$\exists\, \gamma \sim \frac{2k}{n} : \gamma \neq 0 \text{ и } \gamma \neq \infty\,.$$

Итак, зафиксируем ребро номером $i$, его вероятность быть вовлеченным в перестройку равна $p_i$. Так как шагов всего $k$, и на каждом шаге выбирается 2 ребра, вероятность, что $i$-ое ребро никогда не участвовало в перестройке равно $(1-p_i)^{2k}$.

Вычисляя математическое ожидание (подразумеваем математическое ожидание по Марковскому процессу) считаем, что $p_i$ фиксированы. Для учёта вероятностей $p_i$ далее берется интеграл по плотности вероятности. В данном случае мы можем считать $p_i$ фиксированными для всего Марковского процесса, так как оперируем стационарным распределением.

Общее число циклов длины один $c_1$ равно $\sum_i \mathbb{1}_{\{i\text{-е ребро не участвовало в перестройках}\}}$. Среднее нормированное число циклов длины 1 равно:

$$E\left(\frac{c_1}{n}\right) = \frac{1}{n}\sum_i (1-p_i)^{2k}\,.$$

Далее перейдём к пределу по $n$, а $p_i$ распишем как нормированные экспоненциальные величины $\alpha_i$, где $M = \sum_i \alpha_i$:

Заметим, что если $a$ конечно, то по центральной предельной теореме выполняется:

$$\left(\frac{M}{n}\right)^a = \left(\frac{n+\xi\sqrt{n}}{n}\right)^a = \left(1+\frac{\xi}{\sqrt{n}}\right)^a \xrightarrow[n\to\infty]{} 1\,.$$

По доказанному утверждению $\frac{2k}{M} \xrightarrow[n\to\infty]{} \gamma$, тогда

$$E\left(\frac{c_1}{n}\right) = \frac{1}{n}\sum_i (1-\frac{\alpha_i}{M})^{2k} \xrightarrow[n\to\infty]{[k=\frac{n\gamma}{2}]} \frac{1}{n}\sum_i e^{-\gamma\alpha_i}(1+o(1))\,.$$

Проинтегрируем по плотности вероятности:

$$\frac{1}{n}\sum_i \int_0^\infty e^{-\gamma\alpha_i}e^{-\alpha_i}\mathrm{d}\alpha_i \sim \frac{1}{n}\sum_i \int_0^\infty e^{-\alpha_i(\gamma+1)}\mathrm{d}\alpha_i \sim \frac{1}{n}\sum_i \frac{1}{1+\gamma} \sim \frac{1}{1+\gamma}\,.$$

### 2.1.2. Подсчет циклов длины два

Для того, чтобы посчитать число циклов длины 2, зафиксируем два ребра $i$ и $j$, образовавшие этот цикл. Всего шагов произошло $k$, и на каком-то из этих шагов эти два ребра были вовлечены в перестройку, следовательно, необходимо домножить на $k$. На протяжении всех остальных шагов эти рёбра затронуты не были, значит необходимо так же домножить на $(1-p_i-p_j)^{2(k-1)}$. Получаем формулы для вероятности $k\cdot p_i\cdot p_j(1-p_i-p_j)^{2(k-1)}$. Далее суммирием эти вероятности по всем возможным $i$ и $j$. Среднее нормированное число циклов длины 2 равно:

$$E\left(\frac{c_2}{n}\right) = \frac{1}{n}\sum_i\sum_j kp_ip_j(1-p_i-p_j)^{2(k-1)}\,.$$

Аналогично распишем $p_i$ через $\alpha_i$:

$$E\left(\frac{c_2}{n}\right) = \frac{1}{n}\sum_i\sum_j k\frac{\alpha_i\alpha_j}{M^2}(1-\frac{\alpha_i+\alpha_j}{M})^{2(k-1)} =$$

$$= \frac{k}{nM^2} \sum_i \sum_j \alpha_i \alpha_j (1 - \frac{\alpha_i + \alpha_j}{M})^{2(k-1)} \sim \frac{\gamma}{2M^2} \sum_i \sum_j \alpha_i \alpha_j e^{-\gamma\alpha_i} e^{-\gamma\alpha_j} (1 + o(1)) \,.$$

Проинтегрируем по плотности вероятности:

$$\frac{\gamma}{2M^2} \sum_i \sum_j \int_0^\infty \int_0^\infty \alpha_i \alpha_j e^{-\gamma\alpha_i} e^{-\gamma\alpha_j} e^{-\alpha_i} e^{-\alpha_j} \mathrm{d}\alpha_i \mathrm{d}\alpha_j \sim$$

$$\sim \frac{\gamma}{2M^2} \sum_i \sum_j \int_0^\infty \int_0^\infty \alpha_i \alpha_j e^{-\alpha_i(\gamma+1)} e^{-\alpha_j(\gamma+1)} \mathrm{d}\alpha_i \mathrm{d}\alpha_j \sim \frac{\gamma}{2M^2} \sum_i \sum_j \frac{1}{(1+\gamma)^4} \sim$$

$$\sim \frac{n^2\gamma}{2M^2(1+\gamma)^4} \sim \frac{\gamma}{2(1+\gamma)^4} \,.$$

### 2.1.3. Основная теорема

*Теорема* 1. Пусть геном $P_n$ — геном с $n$ хрупкими регионами и геном $Q_n$ получен из $P_n$ посредством $k = \frac{\gamma n}{2}$ операций ДРС для $\gamma > 0$.

Тогда для любого фиксированного $m$ среднее нормированное число циклов длины $m$ в $G(P_n, Q_n)$ равно:

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

*Доказательство*.

Для простоты прочные фрагменты будем называть блоками. Для начала необходимо выбрать блоки, из которых будет получен цикл, всего блоков $n$, нам необходимо $m$, зафиксируем необходимые блоки домножив на $\binom{n}{m}$. Для образования цикла длины $m$ нужно произвести $m-1$ шаг, всего шагов $k$, зафиксируем необходимые шаги домножив на $\binom{k}{m-1}$.

Далее, когда зафиксированы блоки и шаги, на которых они будут сливаться, нужно получить сумму вероятностей по всем возможным сценариям их слияния в цикл длины $m$. По лемме 2 эта вероятность равна $2^{m-1}(m-1)!p_1 \dots p_m(p_1 + \dots + p_m)^{m-2}$. И на всех остальных $k - m + 1$ шагах необходимо не затрагивать выбранные $m$ рёбер, это записывается как $(1 - \sum_{i=1}^m p_i)^{2(k-m+1)}$.

$$E\left(\frac{c_m}{n}\right) = \frac{1}{n}\binom{n}{m}\binom{k}{m-1}2^{m-1}(m-1)!p_1\dots p_m(p_1+\dots+p_m)^{m-2}\times$$

$$\times(1-\sum_{i=1}^{m}p_i)^{2(k-m+1)}\,.$$

Распишем $p_i$ через $\alpha_i$ и раскроем биномиальные коэффициенты:

$$E\left(\frac{c_m}{n}\right) \sim \frac{1}{n}\cdot\frac{n^m}{m!}\cdot\frac{k^{m-1}}{(m-1)!}2^{m-1}(m-1)!\frac{\alpha_1}{M}\dots\frac{\alpha_m}{M}\left(\frac{\alpha_1}{M}+\dots+\frac{\alpha_m}{M}\right)^{m-2}\times$$

$$\times(1-\sum_{i=1}^{m}\frac{\alpha_i}{M})^{2(k-m+1)} \sim$$

$$\sim \frac{n^{m-1}k^{m-1}}{m!}\cdot\frac{2^{m-1}}{M^{2m-2}}\alpha_1\dots\alpha_m(\alpha_1+\dots+\alpha_m)^{m-2}e^{-\sum_{i=1}^{m}\alpha_i} \sim$$

$$\sim \frac{1}{m!}\left(\frac{2k}{M}\right)^{m-1}\left(\frac{n}{M}\right)^{m-1}\alpha_1\dots\alpha_m(\alpha_1+\dots+\alpha_m)^{m-2}e^{-\sum_{i=1}^{m}\alpha_i} \sim$$

$$\sim \frac{\gamma^{m-1}}{m!}\alpha_1\dots\alpha_m(\alpha_1+\dots+\alpha_m)^{m-2}e^{-\sum_{i=1}^{m}\alpha_i}.$$

Проинтегрируем по плотности вероятности:

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

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

$$= [\text{по лемме 3}] = \frac{(3m-3)!\gamma^{m-1}}{m!(2m-1)!(\gamma+1)^{3m-2}}\,.$$

## 2.2. Вспомогательные леммы

### 2.2.1. Сведения задачи о слиянии в цикл к кодам Прюфера

*Лемма* 2. Сумма вероятностей по всем возможным сценариям слияния фиксированных блоков в цикл длины $m$ равна $2^{m-1}(m-1)!p_1\dots p_m(p_1+\dots+p_m)^{m-2}$.

*Доказательство*.

Сценарий объединения в цикл можно описать последовательностью упорядоченных пар $(i, j)$, которые будут записываться как $a_{ij}$, где $i$ и $j$ сообщают о том, какие именно блоки объединились на данном шаге, $i < j$. На рис. 8 приведён пример объединения последовательностью — $a_{13}, a_{23}, a_{34}$.

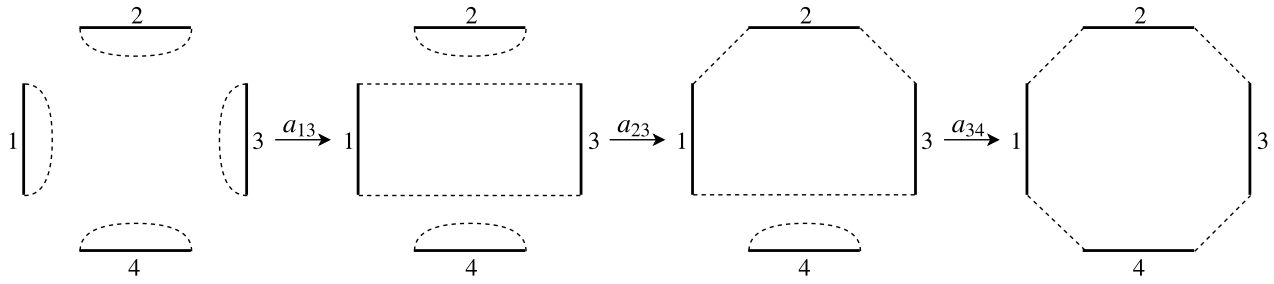



Рисунок 8 – Пример объеденения в цикл длины $4$

Отметим, что для наглядности, подписывая номера на блоках, мы имеем ввиду номера на соответствующих рёбрах. Под соответствующим рёбром понимается ребро, наиболее близкое при движении по часовой стрелке.

Величина $a_{ij}$ на конкретном шаге могла получиться двумя способами: сначала выбран блок $i$, потом блок $j$, или же наоборот, $j$ потом $i$. Поэтому итоговую формулу будет необходимо домножить на $2^{m-1}$, так как всего шагов $m - 1$.

Далее заметим, что для объединения в цикл не важен порядок операций, в котором они стоят. То есть сценарий $a_{13}, a_{23}, a_{34}$ равен сценарию $a_{23}, a_{13}a_{34}$ с точностью до перестановки. Следовательно, мы можем не учитывать конкретный порядок операций, а просто запоминать их множество, при этом домножив формулу на число перестановок, равное $(m - 1)!$.

Полученный объект можно интерпретировать как множество рёбер в соответствующем остовном дереве. Пример подобного соответствия представлен на рис. 9.

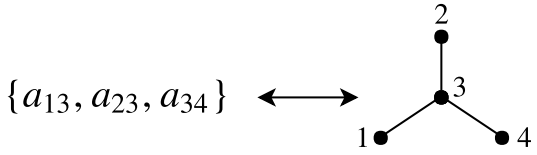



Рисунок 9 – Пример биекции с остовными деревьями

Как известно, остовные деревья в графе размера $m$, кодируются кодами Прюфера [12], состоящими из $m - 2$ чисел из множества $1,2,\ldots,m$. Нам необходимо получить сумму вероятностей по всем возможным сценариям слияния

фиксированных блоков в цикл длины $m$, мы свели эту задачу к сумме по всем остовным деревьям размера $m$. Степень вершины интерпретируется как число раз, когда соответствующее ребро было вовлечено в перестройку. Теперь можно воспользоваться обобщенной формулой Кэли [13]:

$$\sum_T \prod_{i=0}^{m} p_i^{d_i(T)} = p_1 \dots p_m (p_1 + \dots + p_m)^{m-2} .$$

С учётом возможных перестановок получим итоговую формулу:

$$2^{m-1}(m-1)! p_1 \dots p_m (p_1 + \dots + p_m)^{m-2} .$$

Важно отметить, что ввиду устройства операции ДРС, вероятности на рёбрах могут меняться. Но для объединения в цикл нас интересует только сумма вероятностей в компоненте, а не отдельные $p_i$. В процессе перераспределения сумма весов остаётся неизменной, а значит, наши рассуждения остаются верными.

### 2.2.2. Вычисление многомерного интеграла

*Лемма* 3.

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

$$= \frac{(3m-3)!}{(2m-1)!(\gamma+1)^{3m-2}}.$$

*Доказательство*. Введём замену $t_i = \alpha_i(\gamma+1)$:

$$\int \cdots \int_{\mathbb{R}_+^m} \frac{t_1}{\gamma+1} \cdot \ldots \cdot \frac{t_m}{\gamma+1} \left( \frac{t_1 + \ldots + t_m}{\gamma+1} \right)^{m-2} e^{-\sum_{i=1}^m t_i} \frac{\mathrm{d}t_1}{\gamma+1} \dots \frac{\mathrm{d}t_m}{\gamma+1} =$$

$$= \frac{1}{(\gamma+1)^{3m-2}} \int \cdots \int_{\mathbb{R}_+^m} t_1 \cdot \ldots \cdot t_m (t_1 + \ldots + t_m)^{m-2} e^{-\sum_{i=1}^m t_i} \mathrm{d}t_1 \dots \mathrm{d}t_m$$

Введём замену $u = t_1 + \ldots + t_m$:

$$\frac{1}{(\gamma+1)^{3m-2}} \int_0^\infty \int \cdots \int_{\sum_{i=1}^{m-1} t_i \leqslant u} t_1 \dots t_{n-1} \left( u - \sum_{i=1}^{m-1} t_i \right) u^{m-2} e^{-u} \mathrm{d}t_1 \dots \mathrm{d}t_{m-1} \mathrm{d}u =$$

$$= \frac{1}{(\gamma+1)^{3m-2}} \left( \int\limits_0^\infty \int \cdots \int_{\sum_{i=1}^{m-1} t_i \leqslant u} t_1 \cdot \ldots \cdot t_{m-1} u^{m-1} e^{-u} \mathrm{d}t_1 \ldots \mathrm{d}t_{m-1} \mathrm{d}u - \right.$$

$$\left. -(m-1) \int\limits_0^\infty \int \cdots \int_{\sum_{i=1}^{m-1} t_i \leqslant u} t_1^2 \cdot t_2 \cdot \ldots \cdot t_{m-1} u^{m-2} e^{-u} \mathrm{d}t_1 \ldots \mathrm{d}t_{m-1} \mathrm{d}u \right) =$$

$$= \frac{1}{(\gamma+1)^{3m-2}} \left( \int_0^\infty \left( \int \cdots \int_{\sum_{i=1}^{m-1} t_i \leqslant u} t_1 \cdot \ldots \cdot t_{m-1} \mathrm{d}t_1 \ldots \mathrm{d}t_{m-1} \right) u^{m-1} e^{-u} \mathrm{d}u \right.$$

$$\left. -(m-1) \int_0^\infty \left( \int \cdots \int_{\sum_{i=1}^{m-1} t_i \leqslant u} t_1^2 \cdot t_2 \cdot \ldots \cdot t_{m-1} \mathrm{d}t_1 \ldots \mathrm{d}t_{m-1} \right) u^{m-2} e^{-u} \mathrm{d}u \right) =$$

$$= [\text{по лемме 4 и лемме 5}] =$$

$$= \frac{1}{(\gamma+1)^{3m-2}} \left( \int_0^\infty \frac{u^{3m-3} e^{-u}}{(2m-2)!} \mathrm{d}u - (m-1) \int_0^\infty \frac{2u^{3m-3} e^{-u}}{(2m-1)!} \mathrm{d}u \right) =$$

$$= \frac{1}{(\gamma+1)^{3m-2}} \left( \frac{1}{(2m-2)!} - \frac{2(m-1)}{(2m-1)!} \right) \int_0^\infty u^{3m-3} e^{-u} \mathrm{d}u =$$

$$= \frac{1}{(\gamma+1)^{3m-2}} \left( \frac{2m-1-2m+2}{(2m-1)!} \right) \Gamma(3m-2) = \frac{(3m-3)!}{(2m-1)!(\gamma+1)^{3m-2}}$$

*Лемма* 4.

$$\int \cdots \int_{\sum_{i=1}^{n} t_i \leqslant u} t_1 \cdot \ldots \cdot t_n \mathrm{d}t_1 \ldots \mathrm{d}t_n = \frac{u^{2n}}{(2n)!}.$$

*Доказательство*. Доказательство проведём по индукции. База индукции, $n = 1$:

$$\int\limits_0^u t \mathrm{d}t = \frac{u^2}{2} - \frac{0^2}{2} = \frac{u^2}{2}$$

Шаг индукции, пусть выполняется:

$$\int \cdots \int_{\sum_{i=1}^{n} t_i \leqslant u} t_1 \cdot \ldots \cdot t_n \mathrm{d}t_1 \ldots \mathrm{d}t_n = \frac{u^{2n}}{(2n)!}.$$

Вычислим:

$$\int \cdots \int_{\sum_{i=1}^{n+1} t_i \leqslant u} t_1 \cdot \ldots \cdot t_{n+1} \mathrm{d}t_1 \ldots \mathrm{d}t_{n+1} = \int\limits_0^u \frac{(u - t_{n+1})^{2n}}{(2n)!} t_{n+1} \mathrm{d}t_{n+1} =$$

$$= -\frac{1}{(2n)!}\int\limits_0^u t_{n+1}\mathrm{d}\left(\frac{(u-t_{n+1})^{2n+1}}{2n+1}\right) =$$

$$= -\frac{1}{(2n)!}\left(\frac{(u-t_{n+1})^{2n+1}t_{n+1}}{2n+1}\bigg|_0^u - \int\limits_0^u \frac{(u-t_{n+1})^{2n+1}}{2n+1}\mathrm{d}t_{n+1}\right) =$$

$$= \frac{1}{(2n+1)!}\int\limits_0^u (u-t_{n+1})^{2n+1}\mathrm{d}t_{n+1} = \frac{1}{(2n+1)!}\left(-\frac{(u-t_{n+1})^{2n+2}}{2n+2}\bigg|_0^u\right) =$$

$$= \frac{u^{2n+2}}{(2n+2)!}$$

*Лемма* 5.

$$\int\cdots\int_{\sum_{i=1}^n t_i\leqslant u} t_1^2\cdot t_2\cdot\ldots\cdot t_n\mathrm{d}t_1\ldots\mathrm{d}t_n = \frac{2u^{2n+1}}{(2n+1)!}.$$

*Доказательство*. Доказательство проведём по индукции. База индукции, $n=1$:

$$\int\limits_0^u t^2\mathrm{d}t = \frac{u^3}{3} - \frac{0^3}{3} = \frac{u^3}{3}$$

Шаг индукции, пусть выполняется:

$$\int\cdots\int_{\sum_{i=1}^n t_i\leqslant u} t_1^2\cdot t_2\cdot\ldots\cdot t_n\mathrm{d}t_1\ldots\mathrm{d}t_n = \frac{2u^{2n+1}}{(2n+1)!}$$

Вычислим:

$$\int\cdots\int_{\sum_{i=1}^{n+1} t_i\leqslant u} t_1\cdot t_2\cdot\ldots\cdot t_{n+1}\mathrm{d}t_1\ldots\mathrm{d}t_{n+1} = \int\limits_0^u \frac{2(u-t_{n+1})^{2n+1}}{(2n+1)!}t_{n+1}\mathrm{d}t_{n+1} =$$

$$= -\frac{2}{(2n+1)!}\int\limits_0^u t_{n+1}\mathrm{d}\left(\frac{(u-t_{n+1})^{2n+2}}{2n+2}\right) =$$

$$= -\frac{2}{(2n+1)!}\left(\frac{(u-t_{n+1})^{2n+2}t_{n+1}}{2n+2}\bigg|_0^u - \int\limits_0^u \frac{(u-t_{n+1})^{2n+2}}{2n+2}\mathrm{d}t_{n+1}\right) =$$

$$= \frac{2}{(2n+2)!}\int_0^u (u - t_{n+1})^{2n+2} \mathrm{d}t_{n+1} = \frac{2}{(2n+2)!}\left(-\frac{(u-t_{n+1})^{2n+3}}{2n+3}\Big|_0^u\right) =$$

$$= \frac{2u^{2n+3}}{(2n+3)!}.$$

### 2.3. Построение метода оценки истинного эволюционного расстояния

Для оценки истинного эволюционного расстояния будем использовать кумулятивные статистики. Первая статистика $\frac{b}{n}$ — это нормированное число нетривиальных циклов:

$$\frac{b}{n} = 1 - \frac{c_1}{n} = 1 - \frac{1}{1+\gamma} = \frac{\gamma}{1+\gamma}.$$

Вторая статистика $\frac{d}{n}$ — это нормированнное минимальное эволюционное расстояние.

$$\frac{d}{n} = \sum_{m=2}^{\infty} \frac{c_m}{n}(m-1) = 1 - \frac{(1+\gamma)^2({}_2F_1\left(-\frac{2}{3}, -\frac{1}{3}, \frac{1}{2}, \frac{27\gamma}{4(1+\gamma)^3}\right) - 1)}{3\gamma}.$$

Чтобы оценить истинное эволюционное расстояние, узнаем реальные $d$ и $b$ на текущем геноме. Так как функция $\frac{d}{b}$ непрерывна и монотонна, мы можем найти её корень простым двоичным поиском, тем самым, мы узнаём $\gamma$. Далее, зная $\gamma$, находим значение $\frac{b}{n}$. Для того, чтобы предсказать $n$, достаточно разделить $b$ на $\frac{b}{n}$. Всё что осталось — вспомнить, что $k = \frac{\gamma n}{2}$.

В листинге 1 приведен код данного метода на языке *Python 3*:

Листинг 1 – Алгоритм оценки истинного эволючионного расстояния

```
def predict_k(d, b):
    d_over_n = lambda x: 1 - (1 + x) ** 2 * (hyp2f1(-2 / 3, -1 / 3,
        1 / 2, 27 * x / (4 * (1 + x) ** 3)) - 1) / (3 * x)
    b_over_n = lambda x: x / (1 + x)
    d_over_b = lambda r: lambda x: d_over_n(x) / b_over_n(x) - r

    gamma = optimize.bisect(d_over_b(d / b), 1e-6, 3, xtol=1e-4)
    b_n = b_over_n(gamma)
    n = b / b_n
    return n * gamma / 2
```

Далее, проведём симуляции геномных перестроек на языке Python и оценим работу алгоритма для $\gamma \in [0{,}5, 2.0)$. Граница парсимонии находится на $\gamma = 0{,}5$, поэтому меньшие значения нас не интересуют. $\gamma \geqslant 2$ не рассматриваются, так как настолько удаленные геномы очень редки.

Для оценки работы алгоритма построен график распределения относительной ошибки $\frac{k_e - k}{k}$ от $\gamma$ вида «ящик с усами». «Ящикам» соответсвуют $50\,\%$ результатов, «усам» соответствуют $90\,\%$. График приведен на рис. 10. Как мы видим, $50\,\%$ результатов оценки ошибаются не более, чем на $6\,\%$, а $90\,\%$ результатов ошибаются не более, чем на $10\,\%$. Что является хорошим показателем для оценки в рамках модели.

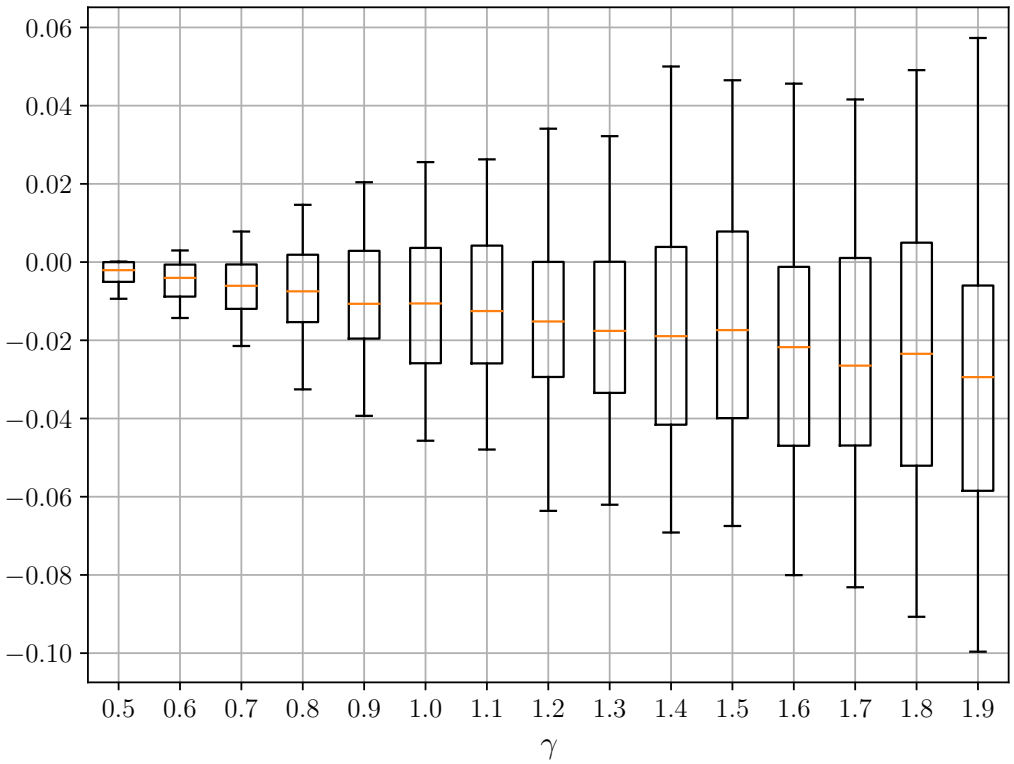



Рисунок 10 – Зависимость распределения относительной ошибки $\frac{k_e - k}{k}$ от $\gamma$ в новом методе оценки

Также, в таблице 1 приведено соответствие среднего модуля ошибки $\frac{k_e - k}{k}$ в процентах от $\gamma \in [0{,}5, 2.0)$.

Как видно из рис. 10, в данном методе оценки присутствует систематическая ошибка (предсказанное $k$ в среднем оказывается меньше, чем реальное). Это связано с тем, что асимптотические оценки учитывают только компоненты первого порядка. Но цикл заданной длины иногда может получаться ввиду распада цикла большей длины.

Таблица 1 – Средний модуль ошибки в процентах в зависимости от $\gamma$

| $\gamma$ | Средний модуль ошибки | $\gamma$ | Средний модуль ошибки |
|---|---|---|---|
| 0.5 | 0.3 % | 1.3 | 2.78 % |
| 0.6 | 0.58 % | 1.4 | 3.12 % |
| 0.7 | 0.86 % | 1.5 | 3.14 % |
| 0.8 | 1.24 % | 1.6 | 3.57 % |
| 0.9 | 1.59 % | 1.7 | 3.76 % |
| 1.0 | 1.88 % | 1.8 | 4.02 % |
| 1.1 | 2.1 % | 1.9 | 4.49 % |
| 1.2 | 2.43 % | 2.0 | 4.87 % |

Для того, чтобы учесть этот факт, эмпирически оценим данную погрешность и учтём её. Для $\gamma < 0.5$ погрешность для $\frac{d}{n}$ равняется 0, а для $\gamma \geqslant 0.5$ она составляет $\frac{0.1}{\sqrt{n}}$ Результаты работы метода оценки, учитывающего данную погрешность приведены на рис. 11.

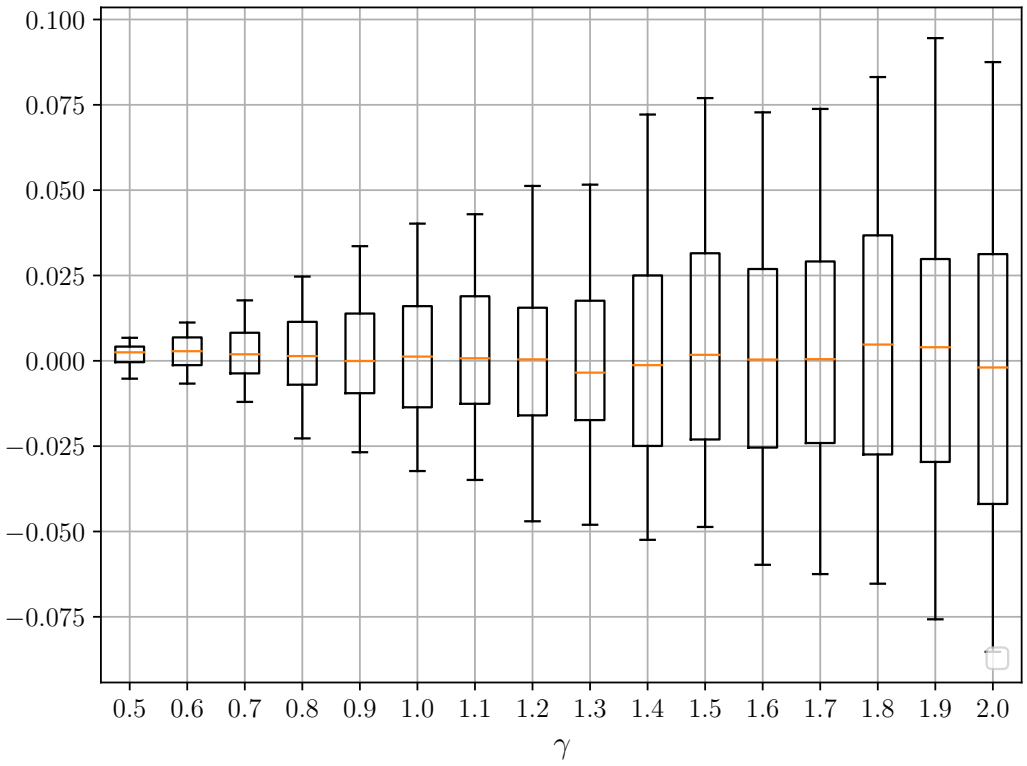



Рисунок 11 – Зависимость распределения относительной ошибки $\frac{k_e - k}{k}$ от $\gamma$ в новом методе оценки с учетом эмпирической оценки на погрешность

## Выводы по главе 2

В главе 2 проведён теоретический анализ модели Дирихле и произведена асимптотическая оценка всех необходимых компонент, построены комбинаторные формулы для среднего числа компонент в общем случае.

Предложен новый метод оценки истинного эволюционного расстояния в рамках этой модели, а также описана его реализация. Проведена оценка точности работы метода для $\gamma \in [0{,}5, 2.0)$.

# ГЛАВА 3. СРАВНЕНИЕ

## 3.1. Сравнение эмпирических и аналитических результатов

Для того, чтобы убедиться в правильности построения и анализа модели, проведём сравнение результатов, полученных эмпирическим и теоретическим способами. Для этого произведем симуляцию на языке Python 3 и сравним полученные результаты с результатами теоретическими. Помимо самого языка, для симуляции поведения модели использовалась библиотека для работы с графами networkx. В данном случае использовано 200 симулированных процессов.

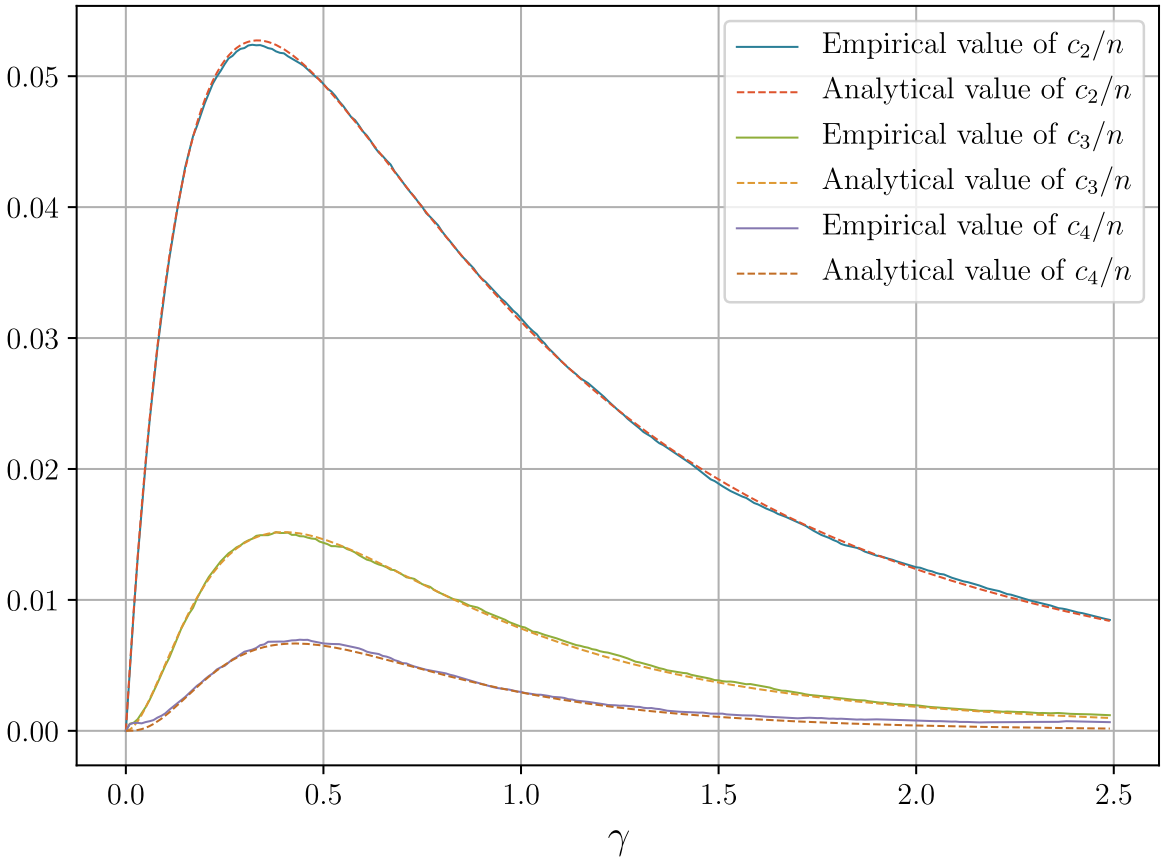



Рисунок 12 – Сравнение эмпирических и теоретических результатов для нормированного количества циклов заданной длины

На рис. 12 приведено сравнение результатов полученных эмпирическим и теоретическим способом. Как мы видим, результаты достаточно хорошо совпадают и могут только незначительно отличаться при $\gamma \geqslant 1.5$, что связано с тем, что циклы меньшей длины могут получаться в результате распадения циклов большей длины.

Так как метод оценки основывается на статистиках $\frac{b}{n}$ и $\frac{d}{n}$ и их частном $\frac{b}{b}$, построим подобный график и для них (рис. 13). Как мы видим, в данном слу-

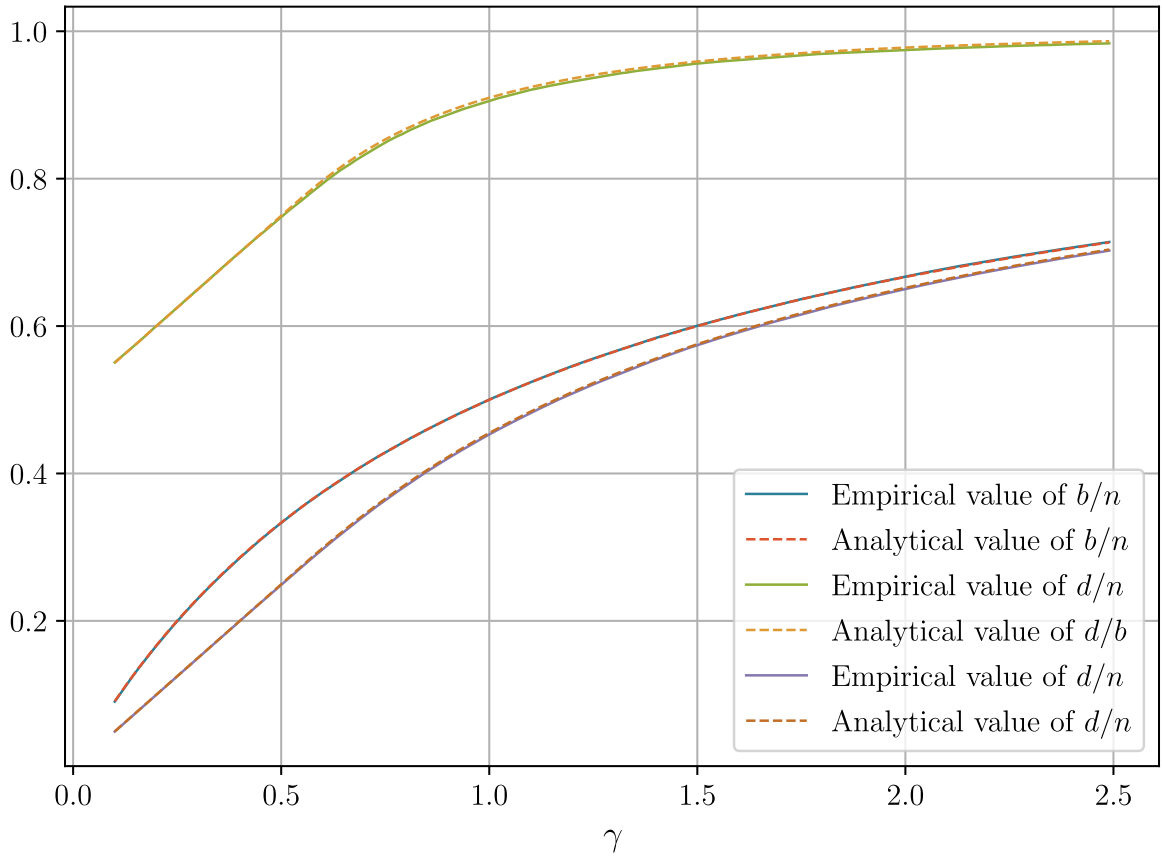


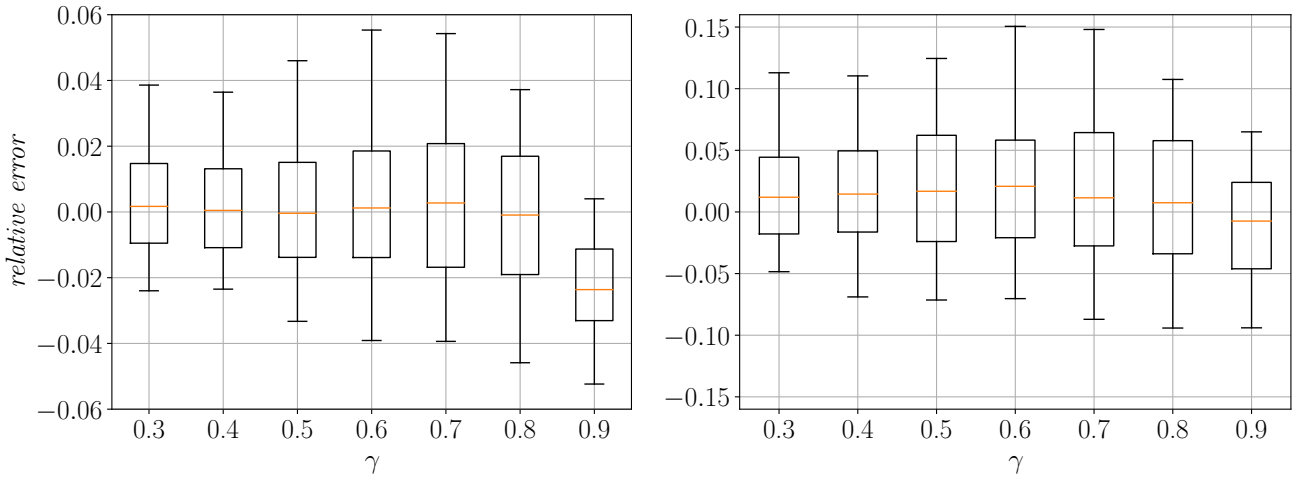

Рисунок 13 – Сравнение эмпирических и теоретических результатов для нормированного количества циклов заданной длины

чае результаты совпадают ещё больше, и в некоторых моментах графики даже неразличимы.

### 3.2. Сравнение с методом оценки Танье

Для сравнения методов также будем использовать симулированные данные. Для каждого $\gamma$ из отрезка $[0.3, 1)$ кратного $0.1$, попробуем предсказать, какое число шагов было сделано, и запишем соответствующую ошибку как $\frac{(k_e - k)}{k}$, где $k_e$ — предсказанное число шагов, а $k$ — реальное число шагов. Для каждого соответствующего $\gamma$ было проведено 200 симуляций и построены соответствующие графики вида «ящик с усами». Как и ранее, «ящик» соответствует $50\,\%$ результатов, а «усы», в свою очередь, $90\,\%$.

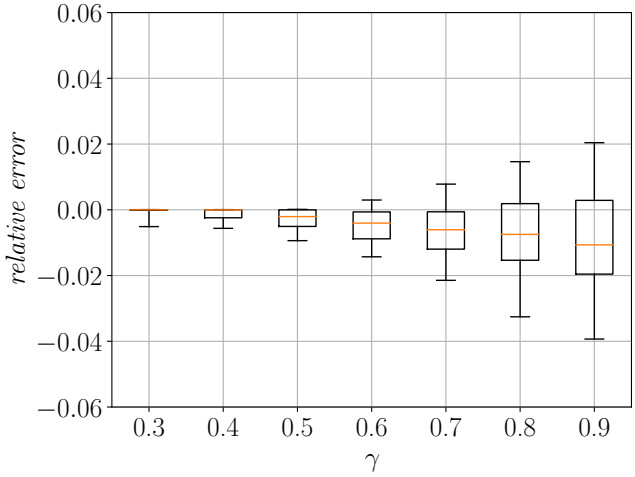
Важно отметить, что метод оценки не располагает информацией ни о реальном числе шагов, ни о количестве тривиальных циклов. Вся информация, известная методу — это количество циклов длин $l \geqslant 2$.



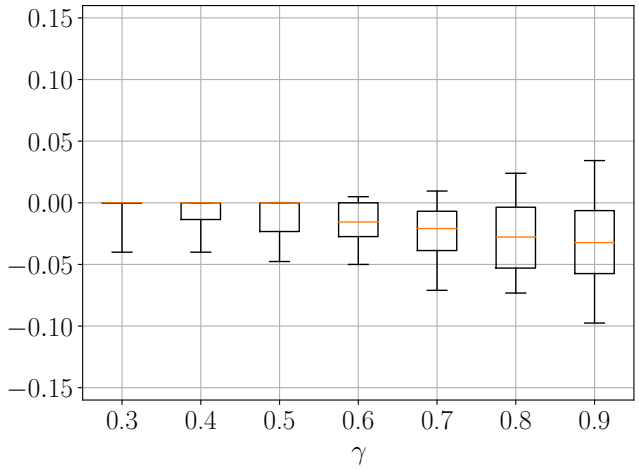

(а) Метод **Танье** на **больших** геномах




(б) Метод **Танье** на **малых** геномах




(в) **Наш** метод на **больших** геномах




(г) **Наш** метод на **малых** геномах

Рисунок 14 – Сравнение зависимости распределения относительной ошибки $\frac{k_e - k}{k}$ от $\gamma$ при оценке разными методами

На рис. 14 показаны результаты работы разных методов для геномов разных размеров. Для больших геномов $n$ находится в промежутке $[1000, 3000]$, для малых — в промежутке $[200, 400]$.

Как мы видим, ошибка метода Танье для больших геномов находится в пределах $4\,\%$ в $50\,\%$ случаев, погрешность нашего же метода находится в пределах $2\%$. А в $90\,\%$ случаев ошибка метода Танье находится в пределах $6\,\%$, наш же метод ошибается не более, чем на $4\,\%$.

Несмотря на то, что в нашей работе мы переходим к пределу по $n$, новый метод показывает лучшие результаты в сравнении с методом Танье.

Для малых геномов ошибка метода Танье находится в пределах $7\,\%$ в $50\,\%$ случаев, погрешность же нашего метода находится в пределах $6\%$. А в $90\,\%$ слу-

чаев ошибка метода Танье находится в пределах $15\,\%$, наш метод ошибается не более, чем на $10\,\%$.

При больших $\gamma$ сравнить методы оценки не представляется возможным, так как метод Танье работает только при $k < \frac{n}{2}$.

В главе 2 этой работы найдены теоретические оценки всех компонент, в то время как в работе [4] найдены оценки только для двух компонент, на которых построен метод оценивания. При этом одна из этих компонент $c_2$ имеет высокую дисперсию, что может негативно сказываться на точности оценки в случаях выбросов.

Оценки компонент, полученные в [4], являются нелинейными, и для поиска оптимального решения используется решение системы двух нелинейных уравнений модифицированными градиентными методами. Компонента $c_2$ немонотонна, что, однако, не влияет на результат.

В этой работе используются более простые и монотонные оценки, а для поиска корней - двоичный поиск.

Сравнение полученных оценок приведено в таблице 2:

Таблица 2 – Сравнение формул оценки необходимых компонент

| – | Оценка Танье | Наша оценка |
|---|---|---|
| Оценка на $c_1$ | $\sum\limits_{l=0}^{\infty} \frac{(-2k)^l}{\prod_{u=0}^{l-1}(n+u)}$ | $\frac{n^2}{2k+n}$ |
| Оценка на $c_2$ | $kn^2 \sum\limits_{l=0}^{\infty}\sum\limits_{m=0}^{\infty} \frac{(-2(k-1))^{l+m}(l+1)(m+1)}{\prod_{u=0}^{l+m+1}(n+u)}$ | $\frac{kn^4}{(2k+n)^4}$ |

Проведём сравнение реализаций данных методов. Оба метода были реализованы на языке программирования *Python 3*. Так как метод Танье основывается на модифицированном градиентом спуске и имеет два нелинейных уравнения, была посчитана соответствующая матрица Якоби. Как было предложено в статье [4], поиск оптимальных значений происходит с помощью метода *optimize.root* из библиотеки *scipy*. Наш же метод использует метод двоичного поиска *optimize.bisect* из той же библиотеки.

Измерения времени работы проведено на компьютере с процессором *Core i5-5257U* и установленными *Python* версии *3.6.1* и библиотекой *scipy* версии *1.0.1*.

Важно отметить, что время работы ни одного из рассматриваемых методов оценки асимптотически никак не зависит от размера графа. От размера графа зависит время, затрачиваемое на подсчёт необходимых компонент, и оно во всех случаях составляет $O(n)$, где $n$ — число вершин в графе. Подобная оценка достигается простым обходом в глубину.

Все измерения проводились на симулированных данных. Результаты приведены в таблице 3.

Таблица 3 – Сравнение работы методов

| – | метод Танье | наш метод |
|---|---|---|
| Среднее время работы на малых геномах | 2.98 сек. | 0.00016 сек. |
| Средний модуль ошибки на малых геномах | 4.8 % | 2.1 % |
| Среднее время работы на больших геномах | 3.02 сек. | 0.00017 сек. |
| Средний модуль ошибки на больших геномах | 1.99 % | 0.68 % |
| Работает при $k \geqslant \frac{n}{2}$ | Нет | Да |
| Реализация | Система уравнений и модифицированный градиентный спуск | Вещественный двоичный поиск |

## 3.3. Применение метода к реальным данным

### 3.3.1. Семейство *Rasacae*

Для рассмотрения возьмем следующие геномы видов из семейства *Rosaceae*: *Prunus* (слива), *Fragaria* (клубника), *Malus* (яблоко). Геномные данные взяты из статьи [14]. Рассмотрим всевозможные пары этих геномов и запишем результаты в таблицу 4.

Как можно увидеть из таблицы, истинное эволюционное растояние в рамках нашей модели может отличаться от минимального на 11 % (это достигается в паре «*Prunus* — *Fragaria*»). В то время как классический подход (равновероятная модель) к оценке истинного расстояния показывает разницы всего в 5 %. Заметим, что разница такого порядка является статистически значимой; вероятность того, что она является результатом погрешности измерений не превосходит 1 %.

Таблица 4 – Сравнения методов на реальных данных из семейства *Rasacae*

| Пара геномов | Минимальное расстояние | Наш метод | Танье | Равновероятная модель |
|---|---|---|---|---|
| *Prunus — Fragaria* | 273 | 297 | 284 | 283 |
| *Prunus — Malus* | 261 | 263 | 258 | 261 |
| *Fragaria — Malus* | 414 | 461 | 426 | 435 |

Аналогичная ситуация происходит на паре «*Prunus — Fragaria*», в рамках нашей модели разница с минимальным расстоянием достигает $8.7\,\%$, а в рамках равновероятной модели эта цифра составляет $4\,\%$.

Пара «*Prunus — Malus*» в нашей модели находится немного за границей парсимонии, а в рамках модели равновероятной перед ней. Поэтому различие в минимальном и истинном расстоянии мы видим только в рамках нашей модели.

Метод Танье на предложенных данных показывает себя чуть хуже. Результаты его работы примерно равны результатам оценщика классического. За исключением того факта, что в паре «*Prunus — Malus*» он показал результат меньший, чем минимальное расстояние. Вероятно это связно с тем, что произошёл выброс по компоненте $c_2$.

### 3.3.2. Класс *Mammalian*

Для следующего рассмотрения возьмем геномы видов из класса *Mammalian* (млекопитающие): *Rat* (крыса), *Chimpanzee* (шимпанзе), *Dog* (собака), *Mouse* (мышь), *Macaque* (макака), *Human* (человек). Рассмотрим всевозможные пары этих геномов и запишем результаты в таблицу 5.

Как можно увидеть из таблицы, в рамках модели с равновероятными поломками рёбер, все пары геномов находятся на стадии парсимонии. То есть оценка на истинное эволюционное расстояние совпадает с минимальным расстоянием.

В рамках нашей модели три пары геномов находятся за границей парсимонии и отличаются примерно на $1.5\,\%$. Этими парами являются: *Chimpanzee — Mouse*, *Dog — Mouse* и *Human — Mouse*. Это говорит о том, что *Mouse* более удалена от остальных видов из класса *Mammalian* (млекопитающие) взятого набора.

Метод Танье на рассматриваемых видах показывает себя неудовлетворительно. В $11$ парах из $15$ (*Chimpanzee — Dog*, *Chimpanzee — Mouse*, *Chimpanzee*

Таблица 5 – Сравнения методов на реальных данных из класса *Mammalian*

| Пара геномов | Минимальное расстояние | Наш метод | Танье | Равновер. модель |
|---|---|---|---|---|
| *Chimpanzee — Dog* | 312 | 312 | 287 | 312 |
| *Chimpanzee — Human* | 22 | 22 | 22 | 22 |
| *Chimpanzee — Mouse* | 420 | 426 | 379 | 420 |
| *Chimpanzee — Macaque* | 115 | 115 | 118 | 115 |
| *Chimpanzee — Rat* | 724 | 724 | 703 | 724 |
| *Dog — Human* | 304 | 304 | 278 | 304 |
| *Dog — Mouse* | 450 | 456 | 409 | 450 |
| *Dog — Macaque* | 301 | 301 | 275 | 301 |
| *Dog — Rat* | 756 | 756 | 733 | 756 |
| *Human — Mouse* | 408 | 414 | 370 | 408 |
| *Human — Macaque* | 106 | 106 | 109 | 106 |
| *Human — Rat* | 714 | 714 | 694 | 714 |
| *Mouse — Macaque* | 407 | 407 | 367 | 407 |
| *Mouse — Rat* | 454 | 454 | 464 | 454 |
| *Macaque — Rat* | 706 | 706 | 690 | 706 |

*— Rat*, *Dog — Human*, *Dog — Mouse*, *Dog — Macaque*, *Dog — Rat*, *Human — Mouse*, *Human — Rat*, *Mouse — Macaque*, *Macaque — Rat*) оценка на истинное расстояние меньше, чем минимальное расстояние. Причём, подобная ошибка достигает $9.8\,\%$ в случае пары *Mouse — Macaque* и некоторых других, что показывает неприменимость метода Танье к данным видам. Данная ошибка, вероятнее всего, связана с выбросом по компоненте $c_2$.

### 3.3.3. Род *Shigella*

Для последнего рассмотрения возьмем геномы видов из класса *Shigella*: *Shigella sonnei Ss046*, *Shigella boydii Sb227*, *Shigella boydii CDC 3083 94*, *Shigella flexneri 2a*, *Shigella flexneri 5 8401*, *Shigella flexneri 2a 2457T*. Рассмотрим всевозможные пары этих геномов и запишем результаты в таблицу 6.

Как можно увидеть из таблицы для видов из рода *Shigella*, в рамках модели с равновероятными поломками рёбер все пары геномов опять же находятся на

Таблица 6 – Сравнения методов на реальных данных из рода *Shigella*

| Пара геномов | Мин. расст. | Наш метод | Танье | Равновер. модель |
|---|---|---|---|---|
| *S. s. Ss046 — S. b. Sb227* | 32 | 32 | 31 | 32 |
| *S. s. Ss046 — S. b. CDC 3083 94* | 44 | 46 | 47 | 44 |
| *S. s. Ss046 — S. f. 2a* | 43 | 45 | 46 | 43 |
| *S. s. Ss046 — S. f. 5 8401* | 40 | 40 | 40 | 40 |
| *S. s. Ss046 — S.f. 2a 2457T* | 35 | 35 | 40 | 35 |
| *S. b. Sb227 — S. b. CDC 3083 94* | 42 | 42 | 45 | 42 |
| *S. b. Sb227 — S. f. 2a* | 45 | 45 | 46 | 45 |
| *S. b. Sb227 — S. f. 5 8401* | 40 | 40 | 40 | 40 |
| *S. b. Sb227 — S.f. 2a 2457T* | 35 | 35 | 34 | 35 |
| *S. b. CDC 3083 94 — S. f. 2a* | 9 | 9 | 9 | 9 |
| *S. b. CDC 3083 94 — S. f. 5 8401* | 14 | 14 | 15 | 14 |
| *S. b. CDC 3083 94 — S.f. 2a 2457T* | 28 | 28 | 29 | 28 |
| *S. f. 2a — S. f. 5 8401* | 11 | 11 | 12 | 11 |
| *S. f. 2a — S.f. 2a 2457T* | 29 | 29 | 30 | 29 |
| *S. f. 5 8401 — S.f. 2a 2457T* | 24 | 24 | 25 | 24 |

стадии парсимонии. То есть минимальное расстояние совпадает с оценкой на истинное эволюционное расстояние.

При этом в рамках нашего метода 2 пары (*S. s. Ss046 — S. b. CDC 3083 94*, *S. s. Ss046 — S. f. 2a*) выходят за границу парсимонии. Оценка на расстояние отличается от минимального отличается всего на 2 шага, но эти 2 шага дают отличие на 4.5 %, что является заметным различием.

Метод Танье на видах рода *Shigella* показывает себя лучше, чем на предыдущих. Только в случае одной пары оценка на истинное расстояние оказывается меньше, чем минимальное расстояние(*S. b. Sb227 — S.f. 2a 2457T*). На тех же двух парах, которые выходили за границу парсимониив в нашем методе, также достигается выход за границу. Но выход за подобную границу достигается и на многих других парах: *S. s. Ss046 — S.f. 2a 2457T*, *S. b. Sb227 — S. b. CDC 3083 94*, *S. b. Sb227 — S. f. 2a*, *S. b. CDC 3083 94 — S. f. 5 8401*, *S. b. CDC 3083 94 —*

*S.f. 2a 2457T*, *S. f. 2a — S. f. 5 8401*, *S. f. 2a — S.f. 2a 2457T*, *S. f. 5 8401 — S.f. 2a 2457T*.

### 3.3.4. Оценка производительности на реальных данных

Несмотря на то, что скорость работы метода оценки не зависит асимптотически от размера генома, для полноты картины проведём оценку на скорость работы методов на реальных данных. Для этого замерим среднее время работы каждого метода на всевозможных парах реальных данных рассмотренных выше, включая время на преобразование генома в граф точек и вычисление необходимых компонент. Результаты приведены в таблице 7.

Таблица 7 – Среднее время работы методов

| Набор геномов | Наш метод | Танье | Равновероятная модель |
|---|---|---|---|
| Семейство *Rasacae* | 0.012 сек. | 3.13 сек. | 0.015 сек. |
| Класс *Mammalian* | 0.027 сек. | 3.55 сек. | 0.037 сек. |
| Род *Shigella* | 0.004 сек. | 2.93 сек. | 0.008 сек. |

## Выводы по главе 3

В главе 3 был проведен эмпирический анализ модели и сравнение с теоретическими результатами. Также было проведено сравнение нового метода оценки с методом, предложенным в [4]. Это сравнение показало более высокие точность, эффективность и применимость указанного метода.

Разные методы оценки расстояния были применены к реальным данным. Основываясь на этих данных показано, что истинное эволюционное расстояние может отличаться от минимального до 11 %. И наконец, ввиду измененной границы парсимонии, показано, что в рамках рассматриваемой модели парсимония может не достигаться там, где она достигалась ранее.

## ЗАКЛЮЧЕНИЕ

В рамках выпускной квалификационной работы был предложен новый метод оценки истинного эволюционного расстояния между геномами в рамках модели, предложенной в [4]. Показана высокая точность данного метода.

Новый метод базируется на асимптотическом анализе комбинаторных формул для среднего числа компонент в общем случае. Все формулы выведены в общем виде для любых весов $p_i$ на рёбрах, при этом существенно использовалась формула Кэли [13]. Для получения конечных результатов, вычисляется многомерный интеграл по плотности вероятности для стационарного распределения рассматриваемой модели. При необходимости данные формулы могут быть применены и к другим распределениям весов на рёбрах.

В работе произведено сравнение эмпирических и теоретических результатов данной модели. Показана их высокая согласованность.

Также выполнено сравнение нового метода оценки с методом, предложенным в [4]. Показано, что новый метод дает более высокие точность, применимость и эффективность.

Разные методы оценки эволюционного расстояния были применены к реальным данным геномов из семейства *Rosaceae*, класса *Mammalian* и рода *Shigella*. Показано, что истинное эволюционное может существенно отличаться от минимального, уточнена граница парсимонии.

## СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ


1 *Lin Y.*, *Moret B.* Estimating true evolutionary distances under the DCJ model // Bioinformatics. — 2008. — Т. 24, № 13. — С. 114–122.

2 *Alexeev N.*, *Alekseyev M. A.* Estimation of the True Evolutionary Distance under the Fragile Breakage Model // BMC Genomics 18(Suppl 4). — 2017. — С. 19–27.

3 *Erdös P.*, *Rényi A.* On the Evolution of Random Graphs // Publ Math Inst Hung Acad Sci. — 1960. — Т. 5. — С. 17–61.

4 Breaking Good: Accounting for Fragility of Genomic Regions in Rearrangement Distance Estimation / P. Biller [и др.] // Genome Biology and Evolution. — 2016. — Май. — Т. 8, № 5. — С. 1427–1439.

5 *Yancopoulos S.*, *Attie O.*, *Friedberg R.* Efficient sorting of genomic permutations by translocation, inversion and block interchange // Bioinformatics. — 2005. — Авг. — Т. 21, № 16. — С. 3340–3346.

6 *Caprara A.* Sorting by reversals is difficult // Proceedings of the first annual international conference on Computational molecular biology. — 1997. — Янв. — Т. 20, № 23. — С. 75–83.

7 *Pevzner P.*, *Tesler G.* Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution // Proc Natl Acad Sci. — 2003. — Июнь. — Т. 100, № 13. — С. 7672–7677.

8 *Alekseyev M.*, *Pevzner P.* Comparative genomics reveals birth and death of fragile regions in mammalian evolution // Genome Biology. — 2010. — Нояб. — Т. 11, № 11. — С. 7672–7677.

9 The 3D organization of chromatin explains evolutionary fragile genomic regions / C. Berthelot [и др.] // Cell Rep. — 2015. — Март. — Т. 10, № 11. — С. 1913–1924.

10 *Бороков А.* Теория Вероятностей. — Москва «Наука», 1986. — С. 285.

11 *Devroye L.* Non-uniform random variate generation. — School of Computer Science, McGill University, 1986. — С. 593–599.

12 *Prüfer H.* Neuer Beweis eines Satzes über Permutationen // Arch. Math. Phys. — 1918. — Т. 27. — С. 742–744.

13 *Cayley A.* A theorem on trees // Quart. J. Math. — 1889. — Т. 23. — С. 376–378.

14 Whole genome comparisons of Fragaria, Prunus and Malus reveal different modes of evolution between Rosaceous subfamilies / S. Jung [и др.] // BMC Genomics. — 2012. — Апр. — Т. 13, № 129. — С. 1–12.