Национальный исследовательский университет ИТМО
(Университет ИТМО)

На правах рукописи

Забелкин Алексей Андреевич

Методы моделирования дискретных случайных процессов на основе комбинаторного анализа перестановок и представлении объектов реального мира в виде графов

Специальность 1.2.2 —
«Математическое моделирование, численные методы и комплексы программ (технические науки)»

Диссертация на соискание учёной степени
кандидата технических наук

Научный руководитель:
канд. техн. наук
Сергушичев Алексей Александрович

Санкт-Петербург — 2025

ITMO University

ITMO

As a manuscript

Zabelkin Alexey Andreevich

Modeling methods for discrete stochastic processes based on combinatorial analysis of permutations and representation of real-world objects as graphs

Specialty 1.2.2 —
Mathematical modeling, numerical methods and software packages (Engineering)

A thesis submitted in fulfillment of the requirements for the degree of
PhD in Engineering

Scientific advisor:
Doctor of Philosophy
Sergushichev Alexey Alexandrovich

Saint Petersburg — 2025

Содержание

Реферат

Общая характеристика работы

Актуальность темы исследования. В задачах прикладной математики часто требуется формальное описание и количественная оценка различий между сложными дискретными структурами — перестановками, графами и деревьями, а также анализ того, как эти структуры эволюционируют во времени или при многократных операциях над ними [6, 7]. Подобные задачи возникают в разных областях: от изучения текстовых данных до моделирования социальных сетей и сопоставления топологий [8–11]. Классическим примером является вычисление "расстояния" между двумя перестановками (минимального числа операций, переводящих одну конфигурацию в другую), что лежит в основе алгоритмов сортировки перестановок, анализа редактирования графов (graph edit distance), а также ряда других задач структурного сравнения [12].

Однако, когда речь заходит о динамике изменений, ситуацию усложняет тот факт, что операции над структурой (добавление рёбер в графы, перестановки элементов, модификации вершин дерева) могут происходить случайным образом с неоднородными вероятностями. В одних случаях вероятность операции считается одинаковой для всех элементов, как в классической модели Эрдёша–Реньи (равновероятное появление рёбер) [13], в других же требуется учесть разные "аффинности" отдельных элементов. Такое неравномерное распределение вероятностей оказывается востребованным при моделировании социальных сетей, соавторства текстов, взаимодействия порядка генов и т. п.

Кроме того, ещё одной важной проблемой является обнаружение параллельных (независимых) изменений на древовидном пространстве состояний. Если в вершинах дерева находятся разные версии исходного объекта (программного кода, текстовой традиции, биологической структуры и т. д.), то нередко интересуют изменения, которые возникли неоднократно и независимо друг от друга на разных ветвях дерева. Подобные конвергентные события важно выявлять в лингвистике (одинаковые языковые новации в независимых группах), в программной инженерии (одинаковые "патчи", реализованные параллельно), а также в биологии (повторные мутации в разных популяциях) [14]. Ранняя (парсимонийная) техника анализа обычно фиксирует минимальное число изменений, не давая количественной меры, отражающей степень параллельности. Проблема усложняется, если число различных ветвей велико, и требуется формализованная методика с ранжированием по "важности".

Биологические приложения занимают особое место в перечисленных задачах. Во-первых, при сравнении и эволюции геномов блоки (гены) можно рассматривать как перестановки, и расстояние между ними (количество инверсий или транспозиций) даёт оценку эволюционной близости [15, 16]. Во-вторых, при моделировании взаимодействий генов или клеточных состояний удобно использовать случайные графы, причём требуются модели, учитывающие различную

“интенсивность” связей [17]. Такие обобщённые модели (с аффинностями) могут предсказывать появление “гигантской компоненты” при иных порогах, чем классическая модель Эрдёша–Реньи. Это существенно влияет на интерпретацию биологических данных, когда слишком упрощённая модель недооценивает или переоценивает вероятность “слияния” крупных фрагментов в эволюционном процессе [18].

Таким образом, актуальными и востребованными **задачами**, объединяющими приложения из различных дисциплин, являются:

а) разработка и анализ математических моделей случайных операций над дискретными структурами с неоднородными вероятностями;
б) оценка расстояния между конфигурациями (включая перестановки, графы, деревья) с возможностью достоверно учитывать крупные масштабы изменений;
в) автоматизации анализа параллельных изменений на деревьях и введении количественных метрик степени их независимого возникновения.

Степень разработки проблемы. Разнообразные аспекты сравнения и эволюции дискретных структур были изучены в ряде фундаментальных и прикладных исследований.

Случайные графы и их обобщения. Классическая модель Эрдёша–Реньи, в которой каждое ребро возникает с одинаковой вероятностью, нашла широкое применение, описаное, в частности, в работах А.М. Райгородского [19, 20]. Позднее было показано, что во многих реальных сетях (социальных, биологических) важно учитывать неоднородность “аффинностей” вершин [18]. Данные обобщения позволяют точнее описывать системы с дифференцированным вкладом узлов. Однако итоговые формулы (например, для порога появления гигантской компоненты) сложны в вычеслении и применении и требуют новых комбинаторных и аналитических результатов [18].

Сравнение перестановок и вычисление расстояний. Для описания изменений последовательностей (в том числе геномных) широко применяются метрики на перестановках. Уже в 1990-х были сформулированы методы вычисления расстояния перестановок (например, через минимальное число операций инверсии/транспозиции) [15, 16], а также предложены статистические модели случайных перестроек (DCJ-модель, модель “хрупких” регионов) [12, 18]. Тем не менее, существующие подходы нередко опираются на бесконечные рядовые разложения и трудоёмкие итерационные алгоритмы, которые становятся неустойчивыми при большом количестве изменений [18]. Это затрудняет оценку истинной дистанции и требует поиска новых аналитических решений.

Обнаружение параллельных изменений на деревьях. В филогенетическом анализе, а также при изучении версий ПО, культурных традиций и других “древовидных” сценариях, давно известно, что один и тот же признак (исправление фрагмента кода, мутация, вставка текста и т. п.) может возникать неоднократно и независимо. Методы парсимонии (например, алгоритм Фитча) выявляют минимальное число таких изменений, но не дают количественной меры параллельно-

сти [21]. Ранние решения были фрагментарными и использовались, в основном, вручную, когда исследователь сам отмечает “зоны повторного возникновения”. Строгое формальное описание и автоматизация подобного анализа остаются открытой проблемой, особенно при больших масштабах данных.

Таким образом, к настоящему моменту накоплен значимый теоретический и прикладной инструментарий для исследования случайных дискретных структур, оценки расстояний и анализа эволюционных деревьев. Однако существенные ограничения всё ещё сохраняются:

- Неоднородность вероятностей далеко не всегда учитывается в традиционных моделях (например, классической модели Эрдёша–Реньи). При этом реальные системы (биологические, социальные) часто требуют более гибких параметров;
- Вычислительная сложность и расходимость рядов в существующих вероятностных моделях для перестановок и графов затрудняют получение точных оценок расстояния при больших масштабах изменений;
- Отсутствие формализованных алгоритмов выявления и количественной оценки параллельных изменений на деревьях: минимальное объяснение парсимонии не отражает “степень” и распределённость независимых появлений признака.

Всё это указывает на необходимость разработки новых математических методов, позволяющих (1) строить обобщённые случайные модели с учётом неоднородных вероятностей, (2) выводить аналитические формулы для расчёта расстояний, преодолевающие проблемы бесконечных рядов, и (3) автоматизировать обнаружение параллельных изменений с количественной оценкой их “независимости”. Результаты таких исследований востребованы как в теоретической математике (расширение классических моделей и методов комбинаторного анализа), так и в прикладных исследованиях, особенно в задачах эволюционной биологии, но и за её пределами — в лингвистике, анализе версий ПО, культурно-исторических исследованиях и других сферах.

Научная новизна состоит в том, что: (1) впервые получены аналитические выражения для оценки числа компонент в рамках случайных графов с индивидуальными вероятностями (обобщение модели Эрдёша–Реньи), устраняющие необходимость численного суммирования расходящихся рядов. (2) найден порог появления гигантской компоненты в модели случайных графов с неравномерными аффинностями, что вдвое меньше порога в классической модели Эрдёша–Реньи. (3) разработан метод оценки истинного расстояния между двумя конфигурациями с учётом неоднородностей, позволяющий устойчиво вычислять метрику при высоком уровне перестроек. (4) предложен алгоритм автоматического выявления параллельных изменений на деревьях. Введена новая комбинаторная метрика параллельности, позволяющая ранжировать независимые события по степени их распределённости на разных ветвях.

Теоретическая значимость определяется расширением классических вероятностных постановок путём введения неоднородных вероятностей, а также

количественной формализацией параллельных изменений на деревьях. В частности: (1) получены новые комбинаторные и асимптотические результаты, описывающие ожидаемое количество компонент заданного размера и появление гигантской компоненты для графов с индивидуальными аффинностями вершин; (2) предложены метод оценки расстояний между перестановками при больших масштабах изменений; (3) систематизирован подход к ранжированию случаев независимых изменений в древовидной топологии.

Практическая значимость работы определяется:

а) Повышение точности оценки расстояний при больших больших масштабах изменений, что важно для сравнительного анализа геномов, крупных текстовых данных.
б) Автоматизированная идентификация и ранжирование параллельных (независимых) изменений, востребованная в биоинформатике (выявление конвергентных мутаций), лингвистике (одинаковые инновации в родственных языках) и др.
в) Программная реализация (пакет *TruEst* для вычисления расстояний и *PaReBrick* для обнаружения параллельных событий), открытая для интеграции в другие исследовательские инструменты.

На защиту выносятся положения, обладающие научной новизной:

1. Комбинаторный метод описания структуры случайных графов с неравномерными аффинностями (обобщающий классическую модель Эрдёша–Реньи), отличающийся тем, что, с целью корректного учёта неоднородных вероятностей рёбер, предложены аналитические формулы для оценки числа компонент связности заданного размера и доказан новый порог появления гигантской компоненты, что расширяет применимость модели.
2. Метод оценки расстояния между объектами, представленными перестановками на основе случайных графов с неоднородными вероятностями состояний, отличающийся тем, что, с целью повышения точности вычислений на больших расстояниях, вместо численного суммирования потенциально расходящихся рядов используются аналитические выражения для ключевых характеристик циклограммы перестановки, что позволило реализовать устойчивое вычисление расстояния даже при высоком уровне эволюционных изменений.
3. Метод выявления и ранжирования независимых изменений в наборах перестановок на древовидных структурах, отличающийся тем, что, с целью автоматического и объективного выявления повторяющихся (конвергентных) событий, вводится новая комбинаторная метрика — «показатель параллельности», количественно отражающая как частоту и количество независимых изменений, так и их распределённость по вершинам дерева, что повышает достоверность и наглядность анализа параллельных эволюционных изменений.

Методы исследования. В работе использованы методы теории вероятностей и математической статистики, комбинаторные методы и алгоритмы на деревьях, методы численной оптимизации и анализа сходимости, экспериментальные тесты на синтетических и реальных данных (в первую очередь, геномных), оценивающие точность и скорость разработанных алгоритмов.

Достоверность научных результатов обеспечена: строгими математическими доказательствами корректности полученных формул, валидацией на симулированных данных, где истинные параметры известны заранее, сравнением с опубликованными результатами и моделями (включая классические алгоритмы оценки расстояния по перестановкам), открытым доступом к программному коду (`GitHub`-репозитории *TruEst* и *PaReBrick*), позволяющим независимо воспроизвести эксперименты.

Соответствие паспорту специальности. Полученные научные результаты соответствуют следующим пунктам паспорта специальности 1.2.2 — "Математическое моделирование, численные методы и комплексы программ (технические науки)":

Пункт 2 — "Разработка, обоснование и тестирование эффективных вычислительных методов с применением современных компьютерных технологий". В работе созданы и протестированы алгоритмы расчёта расстояний между сложными дискретными структурами (с неравномерной вероятностью перестроек) и методика обнаружения параллельных событий на деревьях, реализованные в виде программных модулей.

Пункт 4 — "Разработка новых математических методов и алгоритмов интерпретации натурного эксперимента на основе его математической модели". Предложенные методы позволяют интерпретировать результаты сравнительного анализа реальных данных (геномных, филологических и др.), используя математические модели случайных перестановок и графов с чётким формализмом выявления независимых изменений.

Апробация результатов работы

Основные результаты работы были представлены на следующих конференциях:

- RECOMB Comparative Genomics, 2022, онлайн;
- XI Конгресс молодых ученых, 2022, Университет ИТМО, Санкт-Петербург, Россия;
- Moscow Conference on Computational Molecular Biology, 2021, Москва, Россия;
- Пятидесятая научная и учебно-методическая конференция, 2021, Университет ИТМО, Санкт-Петербург, Россия;
- BiATA 2020 (Bioinformatics: From Algorithms to Applications), 2020, онлайн;
- RECOMB Comparative Genomics (постерный доклад), 2019, Монтпелье, Франция;

– Вероятностные методы в дискретной математике, 2019, Петрозаводск, Россия;
– Moscow Conference on Computational Molecular Biology, 2019, Москва, Россия;
– RECOMB Comparative Genomics, 2018, Шербрук, Канада;

Публикации автора по теме диссертации

Публикации в зарубежных изданиях, индексируемых в базах цитирования Web of Science или Scopus

1. ***Zabelkin A.***, *Avdeyev P.*, *Alexeev N.* TruEst: a better estimator of evolutionary distance under the INFER model // Journal of Mathematical Biology. — 2023. — Июль. — Т. 87, № 2. — ISSN 1432-1416. — DOI: 10.1007/s00285-023-01955-z. — URL: http://dx.doi.org/10.1007/s00285-023-01955-z.
2. ***Zabelkin A.***, *Yakovleva Y.*, *Bochkareva O.*, *Alexeev N.* PaReBrick: PArallel REarrangements and BReaks identification toolkit // Bioinformatics / под ред. R. Schwartz. — 2021. — Окт. — Т. 38, № 2. — С. 357–363. — ISSN 1367-4811. — DOI: 10.1093/bioinformatics/btab691. — URL: http://dx.doi.org/10.1093/bioinformatics/btab691.
3. ***Zabelkin A.***, *Alexeev N.* Estimation of the True Evolutionary Distance Under the INFER Model // Comparative Genomics. — Springer International Publishing, 2018. — С. 72–87. — ISBN 9783030008345. — DOI: 10.1007/978-3-030-00834-5_4. — URL: http://dx.doi.org/10.1007/978-3-030-00834-5_4.
4. *Seferbekova Z.*, ***Zabelkin A.***, *Yakovleva Y.*, *Afasizhev R.*, *Dranenko N. O.*, *Alexeev N.*, *Gelfand M. S.*, *Bochkareva O. O.* High Rates of Genome Rearrangements and Pathogenicity of Shigella spp. // Frontiers in Microbiology. — 2021. — Апр. — Т. 12. — ISSN 1664-302X. — DOI: 10.3389/fmicb.2021.628622. — URL: http://dx.doi.org/10.3389/fmicb.2021.628622.
5. *Petukhova N.*, ***Zabelkin A.***, *Dravgelis V.*, *Aganezov S.*, *Alexeev N.* Chromothripsis Rearrangements Are Informed by 3D-Genome Organization // Comparative Genomics. — Springer International Publishing, 2022. — С. 221–231. — ISBN 9783031062209. — DOI: 10.1007/978-3-031-06220-9_13. — URL: http://dx.doi.org/10.1007/978-3-031-06220-9_13.

Введение

Актуальность темы исследования. В задачах прикладной математики часто требуется формальное описание и количественная оценка различий между сложными дискретными структурами — перестановками, графами и деревьями, а также анализ того, как эти структуры эволюционируют во времени или при многократных операциях над ними [6, 7]. Подобные задачи возникают в разных областях: от изучения текстовых данных до моделирования социальных сетей и сопоставления топологий [8–11]. Классическим примером является вычисление "расстояния" между двумя перестановками (минимального числа операций, переводящих одну конфигурацию в другую), что лежит в основе алгоритмов сортировки перестановок, анализа редактирования графов (graph edit distance), а также ряда других задач структурного сравнения [12].

Однако, когда речь заходит о динамике изменений, ситуацию усложняет тот факт, что операции над структурой (добавление рёбер в графы, перестановки элементов, модификации вершин дерева) могут происходить случайным образом с неоднородными вероятностями. В одних случаях вероятность операции считается одинаковой для всех элементов, как в классической модели Эрдёша–Реньи (равновероятное появление рёбер) [13], в других же требуется учесть разные "аффинности" отдельных элементов. Такое неравномерное распределение вероятностей оказывается востребованным при моделировании социальных сетей, соавторства текстов, взаимодействия порядка генов и т. п.

Кроме того, ещё одной важной проблемой является обнаружение параллельных (независимых) изменений на древовидном пространстве состояний. Если в вершинах дерева находятся разные версии исходного объекта (программного кода, текстовой традиции, биологической структуры и т. д.), то нередко интересуют изменения, которые возникли неоднократно и независимо друг от друга на разных ветвях дерева. Подобные конвергентные события важно выявлять в лингвистике (одинаковые языковые новации в независимых группах), в программной инженерии (одинаковые "патчи", реализованные параллельно), а также в биологии (повторные мутации в разных популяциях) [14]. Ранняя (парсимонийная) техника анализа обычно фиксирует минимальное число изменений, не давая количественной меры, отражающей степень параллельности. Проблема усложняется, если число различных ветвей велико, и требуется формализованная методика с ранжированием по "важности".

Биологические приложения занимают особое место в перечисленных задачах. Во-первых, при сравнении и эволюции геномов блоки (гены) можно рассматривать как перестановки, и расстояние между ними (количество инверсий или транспозиций) даёт оценку эволюционной близости [15, 16]. Во-вторых, при моделировании взаимодействий генов или клеточных состояний удобно использовать случайные графы, причём требуются модели, учитывающие различную "интенсивность" связей [17]. Такие обобщённые модели (с аффинностями) могут предсказывать появление "гигантской компоненты" при иных порогах, чем

классическая модель Эрдёша–Реньи. Это существенно влияет на интерпретацию биологических данных, когда слишком упрощённая модель недооценивает или переоценивает вероятность “слияния” крупных фрагментов в эволюционном процессе [18].

Таким образом, актуальными и востребованными **задачами**, объединяющими приложения из различных дисциплин, являются:

а) разработка и анализ математических моделей случайных операций над дискретными структурами с неоднородными вероятностями;
б) оценка расстояния между конфигурациями (включая перестановки, графы, деревья) с возможностью достоверно учитывать крупные масштабы изменений;
в) автоматизации анализа параллельных изменений на деревьях и введении количественных метрик степени их независимого возникновения.

Степень разработки проблемы. Разнообразные аспекты сравнения и эволюции дискретных структур были изучены в ряде фундаментальных и прикладных исследований.

Случайные графы и их обобщения. Классическая модель Эрдёша–Реньи, в которой каждое ребро возникает с одинаковой вероятностью, нашла широкое применение, описаное, в частности, в работах А.М. Райгородского [19, 20]. Позднее было показано, что во многих реальных сетях (социальных, биологических) важно учитывать неоднородность “аффинностей” вершин [18]. Данные обобщения позволяют точнее описывать системы с дифференцированным вкладом узлов. Однако итоговые формулы (например, для порога появления гигантской компоненты) сложны в вычеслении и применении и требуют новых комбинаторных и аналитических результатов [18].

Сравнение перестановок и вычисление расстояний. Для описания изменений последовательностей (в том числе геномных) широко применяются метрики на перестановках. Уже в 1990-х были сформулированы методы вычисления расстояния перестановок (например, через минимальное число операций инверсии/транспозиции) [15, 16], а также предложены статистические модели случайных перестроек (DCJ-модель, модель “хрупких” регионов) [12, 18]. Тем не менее, существующие подходы нередко опираются на бесконечные рядовые разложения и трудоёмкие итерационные алгоритмы, которые становятся неустойчивыми при большом количестве изменений [18]. Это затрудняет оценку истинной дистанции и требует поиска новых аналитических решений.

Обнаружение параллельных изменений на деревьях. В филогенетическом анализе, а также при изучении версий ПО, культурных традиций и других “древовидных” сценариях, давно известно, что один и тот же признак (исправление фрагмента кода, мутация, вставка текста и т. п.) может возникать неоднократно и независимо. Методы парсимонии (например, алгоритм Фитча) выявляют минимальное число таких изменений, но не дают количественной меры параллельности [21]. Ранние решения были фрагментарными и использовались, в основном, вручную, когда исследователь сам отмечает “зоны повторного возникновения”.

Строгое формальное описание и автоматизация подобного анализа остаются открытой проблемой, особенно при больших масштабах данных.

Таким образом, к настоящему моменту накоплен значимый теоретический и прикладной инструментарий для исследования случайных дискретных структур, оценки расстояний и анализа эволюционных деревьев. Однако существенные ограничения всё ещё сохраняются:

- Неоднородность вероятностей далеко не всегда учитывается в традиционных моделях (например, классической модели Эрдёша–Реньи). При этом реальные системы (биологические, социальные) часто требуют более гибких параметров;
- Вычислительная сложность и расходимость рядов в существующих вероятностных моделях для перестановок и графов затрудняют получение точных оценок расстояния при больших масштабах изменений;
- Отсутствие формализованных алгоритмов выявления и количественной оценки параллельных изменений на деревьях: минимальное объяснение парсимонии не отражает “степень” и распределённость независимых появлений признака.

Всё это указывает на необходимость разработки новых математических методов, позволяющих (1) строить обобщённые случайные модели с учётом неоднородных вероятностей, (2) выводить аналитические формулы для расчёта расстояний, преодолевающие проблемы бесконечных рядов, и (3) автоматизировать обнаружение параллельных изменений с количественной оценкой их “независимости”. Результаты таких исследований востребованы как в теоретической математике (расширение классических моделей и методов комбинаторного анализа), так и в прикладных исследованиях, особенно в задачах эволюционной биологии, но и за её пределами — в лингвистике, анализе версий ПО, культурно-исторических исследованиях и других сферах.

Научная новизна состоит в том, что: (1) впервые получены аналитические выражения для оценки числа компонент в рамках случайных графов с индивидуальными вероятностями (обобщение модели Эрдёша–Реньи), устраняющие необходимость численного суммирования расходящихся рядов. (2) найден порог появления гигантской компоненты в модели случайных графов с неравномерными аффинностями, что вдвое меньше порога в классической модели Эрдёша–Реньи. (3) разработан метод оценки истинного расстояния между двумя конфигурациями с учётом неоднородностей, позволяющий устойчиво вычислять метрику при высоком уровне перестроек. (4) предложен алгоритм автоматического выявления параллельных изменений на деревьях. Введена новая комбинаторная метрика параллельности, позволяющая ранжировать независимые события по степени их распределённости на разных ветвях.

Теоретическая значимость определяется расширением классических вероятностных постановок путём введения неоднородных вероятностей, а также количественной формализацией параллельных изменений на деревьях. В частности: (1) получены новые комбинаторные и асимптотические результаты, опи-

сывающие ожидаемое количество компонент заданного размера и появление гигантской компоненты для графов с индивидуальными аффинностями вершин; (2) предложены метод оценки расстояний между перестановками при больших масштабах изменений; (3) систематизирован подход к ранжированию случаев независимых изменений в древовидной топологии.

Практическая значимость работы определяется:

а) Повышение точности оценки расстояний при больших больших масштабах изменений, что важно для сравнительного анализа геномов, крупных текстовых данных.

б) Автоматизированная идентификация и ранжирование параллельных (независимых) изменений, востребованная в биоинформатике (выявление конвергентных мутаций), лингвистике (одинаковые инновации в родственных языках) и др.

в) Программная реализация (пакет *TruEst* для вычисления расстояний и *PaReBrick* для обнаружения параллельных событий), открытая для интеграции в другие исследовательские инструменты.

На защиту выносятся положения, обладающие научной новизной:

1. Комбинаторный метод описания структуры случайных графов с неравномерными аффинностями (обобщающий классическую модель Эрдёша–Реньи), отличающийся тем, что, с целью корректного учёта неоднородных вероятностей рёбер, предложены аналитические формулы для оценки числа компонент связности заданного размера и доказан новый порог появления гигантской компоненты, что расширяет применимость модели.
2. Метод оценки расстояния между объектами, представленными перестановками на основе случайных графов с неоднородными вероятностями состояний, отличающийся тем, что, с целью повышения точности вычислений на больших расстояниях, вместо численного суммирования потенциально расходящихся рядов используются аналитические выражения для ключевых характеристик циклограммы перестановки, что позволило реализовать устойчивое вычисление расстояния даже при высоком уровне эволюционных изменений.
3. Метод выявления и ранжирования независимых изменений в наборах перестановок на древовидных структурах, отличающийся тем, что, с целью автоматического и объективного выявления повторяющихся (конвергентных) событий, вводится новая комбинаторная метрика — «показатель параллельности», количественно отражающая как частоту и количество независимых изменений, так и их распределённость по вершинам дерева, что повышает достоверность и наглядность анализа параллельных эволюционных изменений.

Методы исследования. В работе использованы методы теории вероятностей и математической статистики, комбинаторные методы и алгоритмы на деревьях, методы численной оптимизации и анализа сходимости, эксперименталь-

ные тесты на синтетических и реальных данных (в первую очередь, геномных), оценивающие точность и скорость разработанных алгоритмов.

Достоверность научных результатов обеспечена: строгими математическими доказательствами корректности полученных формул, валидацией на симулированных данных, где истинные параметры известны заранее, сравнением с опубликованными результатами и моделями (включая классические алгоритмы оценки расстояния по перестановкам), открытым доступом к программному коду (`GitHub`-репозитории *TruEst* и *PaReBrick*), позволяющим независимо воспроизвести эксперименты.

Соответствие паспорту специальности. Полученные научные результаты соответствуют следующим пунктам паспорта специальности 1.2.2 — "Математическое моделирование, численные методы и комплексы программ (технические науки)":

Пункт 2 — "Разработка, обоснование и тестирование эффективных вычислительных методов с применением современных компьютерных технологий". В работе созданы и протестированы алгоритмы расчёта расстояний между сложными дискретными структурами (с неравномерной вероятностью перестроек) и методика обнаружения параллельных событий на деревьях, реализованные в виде программных модулей.

Пункт 4 — "Разработка новых математических методов и алгоритмов интерпретации натурного эксперимента на основе его математической модели". Предложенные методы позволяют интерпретировать результаты сравнительного анализа реальных данных (геномных, филологических и др.), используя математические модели случайных перестановок и графов с чётким формализмом выявления независимых изменений.

Апробация результатов работы

Основные результаты работы были представлены на следующих конференциях:

- RECOMB Comparative Genomics, 2022, онлайн;
- XI Конгресс молодых ученых, 2022, Университет ИТМО, Санкт-Петербург, Россия;
- Moscow Conference on Computational Molecular Biology, 2021, Москва, Россия;
- Пятидесятая научная и учебно-методическая конференция, 2021, Университет ИТМО, Санкт-Петербург, Россия;
- BiATA 2020 (Bioinformatics: From Algorithms to Applications), 2020, онлайн;
- RECOMB Comparative Genomics (постерный доклад), 2019, Монтпелье, Франция;
- Вероятностные методы в дискретной математике, 2019, Петрозаводск, Россия;
- Moscow Conference on Computational Molecular Biology, 2019, Москва, Россия;

– RECOMB Comparative Genomics, 2018, Шербрук, Канада;

Финансирование

Автор признателен компании JetBrains Research за финансовую поддержку работы в 2017–2022 годах. Работа выполнена также благодаря финансированию от проекта 5-100.

Глава 1. Обзор предметной области

1.1. Модели случайных графов

1.1.1. Классическая модель Эрдеша–Реньи

Классическая модель случайного графа Эрдеша–Реньи представляет собой вероятностную модель неориентированного графа на n вершинах [13]. Существуют две эквивалентные версии модели. В модели $G(n, M)$ выбирается равновероятно один из всех графов с фиксированным числом вершин n и ровно M рёбрами. В более распространённой *биномиальной* модели $G(n, p)$ предполагается, что каждый из $\binom{n}{2}$ возможных ребёр присутствует независимо с вероятностью p.

Модели Эрдеша–Реньи заложили основу теории случайных графов и вероятностного метода в комбинаторике. При больших n такие графы демонстрируют резкие *пороговые явления*: многие свойства графа возникают почти наверняка, как только параметр p превышает некоторый критический порог (зависящий от n). Например, для достаточно малых p граф почти наверняка несвязен и разбит на множество мелких компонент, но при увеличении p происходит фазовый переход к связному графу. Ниже рассмотрен классический пример такого перехода – появление гигантской компоненты.

1.1.2. Порог появления гигантской компоненты

Одним из самых известных результатов Эрдеша–Реньи является порог появления гигантской связной компоненты в случайном графе. Под гигантской компонентой понимают связную компоненту размера порядка n (то есть содержащую положительную долю всех вершин графа). Для модели $G(n, p)$ при $n \to \infty$ существует критическое значение $p_c \sim \frac{1}{n}$, при превышении которого в графе почти наверняка присутствует единственная гигантская компонента. Точнее, если $p = \frac{c}{n}$, то наблюдается фазовый переход при $c = 1$. Когда $c < 1$ (то есть $p < \frac{1}{n}$), случайный граф со стремящейся к 1 вероятностью состоит лишь из множества малых компонент (размер каждой не более $O(\ln n)$). При $c > 1$ ($p > \frac{1}{n}$) почти наверняка возникает единственная большая компонента, содержащая $\Theta(n)$ вершин, тогда как все остальные компоненты остаются малыми. В точке $p \approx \frac{1}{n}$ происходит переходный режим: максимальная компонента имеет размер порядка $n^{2/3}$. Это явление аналогично перколяционному переходу в статистической физике; порог $p_c = 1/n$ часто называют критической точкой перколяции на полном графе.

Данный результат был впервые доказан Эрдешем и Реньи [13]. Он иллюстрирует свойственный случайным графам эффект: небольшое стохастическое изменение параметров (от $p = \frac{1-\varepsilon}{n}$ к $p = \frac{1+\varepsilon}{n}$) приводит к резкому структурному изменению графа (появлению «гиганта»). В дальнейшем мы увидим аналогичные идеи при рассмотрении разбиения генома на фрагменты под действием

случайных перестроек: там также наблюдается переход от сохранения большой цельной части структуры к её фрагментации на множество небольших блоков.

1.1.3. Аффинные модификации модели случайных графов

Классическая модель Эрдеша–Реньи предполагает однородность: все вершины и потенциальные рёбра статистически эквивалентны, вероятность появления любого ребра одинакова (p) и независима от других. Однако во многих реальных сетях (и тем более в структурных моделях геномов) такая простая случайность не соблюдается — некоторые связи возникают чаще других, степень вершин может подчиняться неоднородному распределению, наблюдается кластеризация и пр. Поэтому были предложены многочисленные обобщения модели случайного графа, вводящие *неоднородности* в вероятность ребер. Условно такие обобщения можно назвать «аффинными» модификациями модели, поскольку они сохраняют линейный характер зависимости вероятностей от некоторых параметров (например, от свойств вершин или уже существующих степеней), но отходят от строгой равновероятности всех связей.

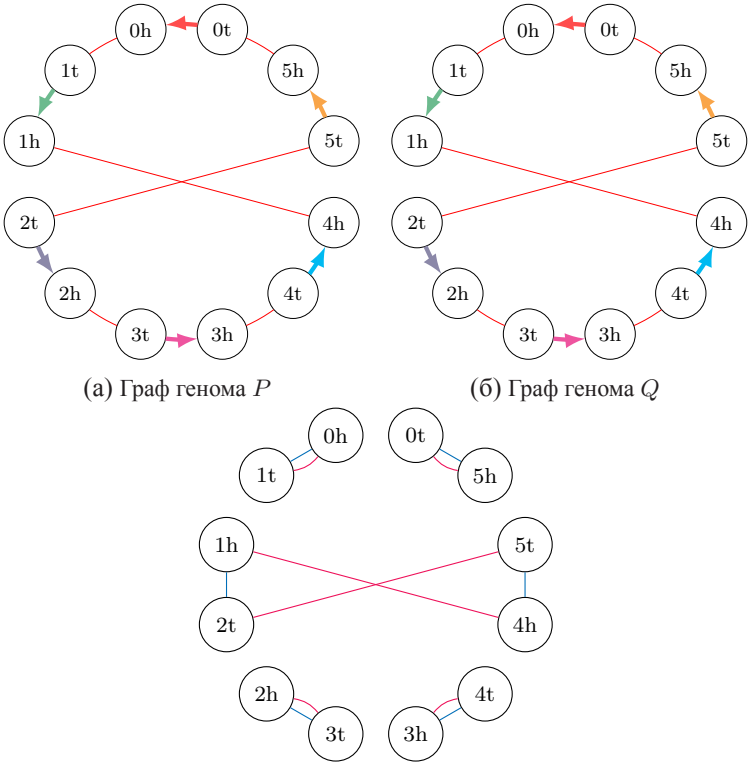
Одним из направлений обобщения являются модели с заданной степенной последовательностью. В работе Ньюмана, Строгатза и Уоттса [22] предложена генеративная модель случайного графа с произвольным заданным распределением степеней вершнин. В ней каждой вершине заранее приписывается случайная степень (например, согласно некоторому распределению), а затем вершины случайно спариваются по полурёбрам (англ. *half-edges*) до достижения требуемых степеней. Эта модель эквивалентна так называемой конфигурационной модели и позволяет получать случайные графы с заданными свойствами (например, с тяжёлыми хвостами распределения степеней), что является аффинной модификацией по отношению к модели Эрдеша–Реньи (где распределение степеней, напротив, близко к пуассоновскому и быстро убывает).

Другая известная модификация — модель предпочтительного присоединения (модель Барабаши — Альберт) [23]. В ней граф строится динамически: вершины добавляются последовательно, и каждая новая вершина соединяется с некоторым числом ранее добавленных вершин с вероятностями, пропорциональными степеням этих существующих вершин. Таким образом, вероятность образования нового ребра линейно (“аффинно”) зависит от текущей степени вершины: $\Pr(\text{новое ребро соединится с вершиной } i) \propto k_i + c$, где k_i — степень вершины i, а c — некоторая константа предпочтения. Данная модель генерирует “безмасштабные” сети с степенным распределением степеней вершин, что значительно отличает её от модели Эрдеша–Реньи.

Существуют и геометрические (пространственные) случайные графы, в которых вершины имеют случайные координаты, и рёбра возникают с вероятностью, зависящей от расстояния между вершинами (например, модель единичного диска). Это также вводит “аффинность” через функцию расстояния: близкие вершины имеют повышенный шанс связаться.

Обобщения модели случайного графа важны тем, что позволяют более адекватно моделировать сложные системы. В частности, при моделировании эволюции геномов нам потребуется учитывать неоднородности в вероятностях «сопряжённости» элементов генома (некоторые элементы чаще участвуют в эволюционных событиях, чем другие). Это является прямым аналогом отхода от простейшей равномерной случайности, подобно переходу от $G(n, p)$ к моделям с “горячими точками” (hot spots) или с индивидуальными вероятностями для различных потенциальных связей. Далее мы увидим, как такое введение “весов” и неоднородностей применяется к специальному графу, моделирующему структуры генома.

1.1.4. Граф точек разрыва и его модификации



(а) Граф генома P

(б) Граф генома Q

(в) Граф точек разрыва $G(P, Q)$

Рисунок 1 – Построение графа точек разрыва для пары циклических геномов и инверсией друг относительно друга

Формально геном можно представить как упорядоченную последовательность идентифицируемых блоков. Вершинами данного графа являются концы и начала геномных блоков. Данный граф включает направленные рёбра блоков (англ. *block edges*), задающие сами блоки и их ориентацию, и неориентированные рёбра соседств (англ. *adjacency edges*), которые отражают соседство между блоками. На рис. 1а приведён пример геномного графа для циклического генома P, содержащего блоки $(0, 1, 2, 3, 4, 5)$, а на рис. 1б — для генома $Q = (0, 1, -4, -3, -2, 5)$, в котором блоки 2, 3, 4 инвертированы относительно генома P.

Граф точек разрыва (брейкпоинт граф, англ. *breakpoint graph*) $G(P, Q)$ строится как объединение графов соседств для двух сравниваемых геномов после удаления направленных рёбер блоков (рис. 1в). При этом в итоговом графе присутствуют два типа рёбер: от генома P (синий цвет) и от генома Q (красный цвет). Если геномы состоят из одинаковых однокопийных блоков, то результатом объединения будет двуцветный граф с вершинами степени 2, компоненты связности которого — это циклы с чередующимися красными и синими рёбрами. Количество таких циклов определяет минимальное число шагов эволюции, необходимое для преобразования одного генома в другой.

Например, для приведённых геномов (рис. 1в), наличие циклов чётко показывает наличие инверсий: каждый цикл длины более двух отражает области, требующие дополнительных операций для согласования порядка блоков между геномами.

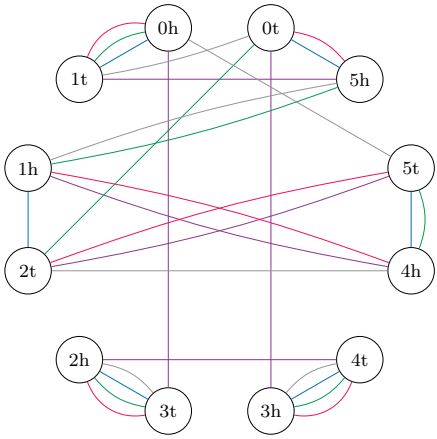


Рисунок 2 – Пример множественного графа точек разрыва для пяти штаммов; рёбра разных штаммов показаны различными цветами

Для анализа нескольких близкородственных геномов вводится *множественный граф точек разрыва* (англ. *multiple breakpoint graph*). Такой граф является расширением описанного выше подхода на случай множества организмов

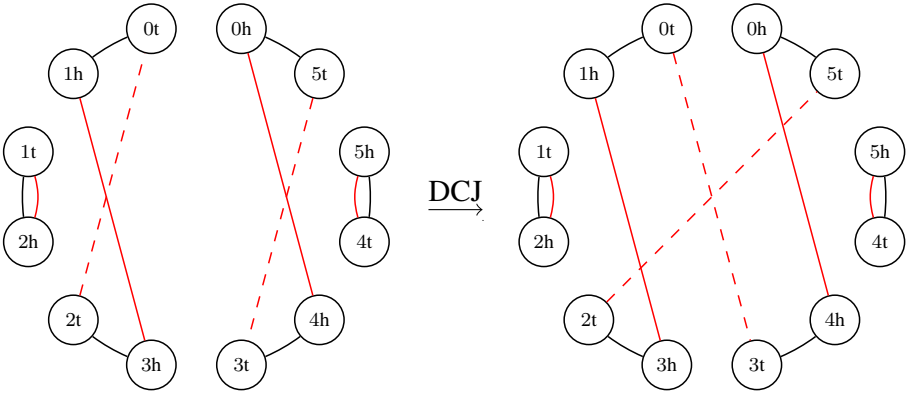


Рисунок 3 – Операция DCJ в геноме Q заменяет пару красных рёбер в графе точек разрыва $G(P, Q)$ на другую пару красных рёбер, образующую паросочетание на том же множестве из четырёх вершин.

и содержит рёбра от каждого рассматриваемого генома. В этом случае каждое ребро дополнительно снабжается меткой, идентифицирующей исходный организм (или штамм), а между двумя вершинами может возникнуть несколько параллельных рёбер от разных организмов. Такие «консенсусные» мульти-рёбра обозначают идентичные соседства блоков в нескольких геномах (рис. 2). Анализ этого мультиграфа позволяет выявлять эволюционные события, такие как потери соседств и независимые повторные перестройки (например, идентифицируя циклы длины 4 как признак инверсий). В следующем разделе будет рассмотрено, как использовать множественные графы точек разрыва для формализации признаков и идентификации параллельных изменений.

1.2. Постановка задачи оценки расстояний между структурами

1.2.1. Опеределение шага процесса

Операция двойной разрез и слияние (англ. *DCJ, Double Cut and Join*) является ключевым понятием в моделировании эволюционных перестроек геномов. Данная операция формализует различные типы реальных перестроек, таких как инверсии, транслокации, слияния и разрывы хромосом, посредством единой математической операции [15]. В терминах графа точек разрыва, каждая операция двойной разрез и слияние соответствует изменению структуры рёбер графа, которые отражают взаимное расположение геномных фрагментов. На рис. 3 показан пример операции двойной разрез и слияние, которая выполняется на графе точек разрыва $G(P, Q)$.

Формально, операция двойной разрез и слияние выполняется следующим образом:

а) Выбираются два ребра (или теломерные вершины) графа.
б) Оба выбранных ребра разрезаются, образуя четыре свободных конца.
в) Затем эти концы соединяются заново одним из возможных способов, который отличается от исходного состояния.

В зависимости от выбора исходных рёбер, операция двойной разрез и слияние может иметь несколько исходов:

– Если выбираются два ребра из одного и того же цикла или пути, то двойной разрез и слияние может разделить его на два меньших цикла или пути.
– Если выбираются рёбра из разных циклов или путей, то двойной разрез и слияние может привести к их слиянию.
– Если выбирается теломерная вершина, операция может привести к преобразованию линейной хромосомы в циклическую или наоборот.

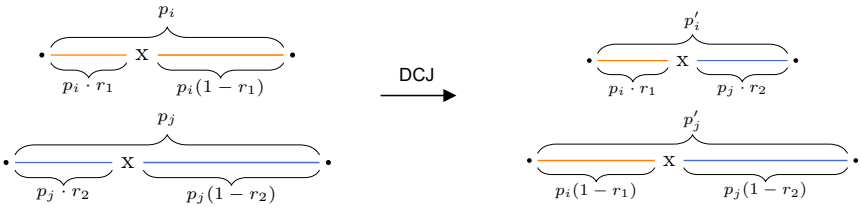


Рисунок 4 – Иллюстрация механизма перераспределения весов в графе точек разрыва

В случае модифицированного графа точек разрыва каждое ребро снабжается дополнительным весом, отражающим вероятность его участия в перестройке. Операция двойного разрыва и слияния (DCJ) в таком взвешенном графе осуществляется аналогично невзвешенному случаю, однако пара рёбер для перестройки выбирается с вероятностью, пропорциональной произведению их текущих весов.

Формально, пусть в результате операции двойной разрез и слияние разрываются два ребра смежности $\{x, y\}$ и $\{u, v\}$ с соответствующими весами p_i и p_j. Данные рёбра заменяются одной из двух возможных новых пар рёбер, образующих паросочетание на исходных четырёх вершинах: либо $\{x, u\}$ и $\{y, v\}$, либо $\{x, v\}$ и $\{y, u\}$. При этом, согласно модели, веса вновь образованных рёбер пересчитываются следующим образом. Случайно и независимо выбираются числа $r_1, r_2 \in (0, 1)$, соответствующие точкам разрыва внутри исходных рёбер. Новые веса p'_i и p'_j затем вычисляются согласно формулам:

$$p'_i = r_1 p_i + r_2 p_j, \quad p'_j = (1 - r_1)p_i + (1 - r_2)p_j.$$

Таким образом, общая сумма весов сохраняется, но сами веса перераспределяются в зависимости от случайно выбранных точек разрыва 4.

Использование взвешенного подхода при операции DCJ обосновано биологически и позволяет более реалистично отражать процесс геномных перестроек. В реальных геномах вероятность разрыва в конкретной области определяется множеством факторов, в частности, структурными особенностями хроматина. Например, вероятность перестройки участка пропорциональна длине соответствующего хрупкого региона, поскольку перестройка может произойти равновероятно в любой его точке. Данный подход существенно улучшает точность моделирования и позволяет учитывать неоднородность структуры генома, обеспечивая более точную количественную оценку геномных расстояний.

1.2.2. Эволюционная модель и её равновесное распределение

Для корректной оценки реального эволюционного расстояния между геномами P и Q, имеющими одинаковый набор геномных блоков, рассмотрим эволюцию как дискретный Марковский процесс. Данный процесс начинается с исходного состояния, заданного геномом P, и заканчивается целевым состоянием, задаваемым геномом Q. Переход между состояниями осуществляется посредством последовательности операций двойной разрез и слияние.

Основным отличием рассматриваемой здесь модели от классической модели случайных разрывов является способ выбора ребёр для выполнения операции перестройки. Если в классической модели ребра выбираются равновероятно, то здесь каждое ребро снабжено вероятностью (весом), и пара рёбер i и j выбирается независимо друг от друга с вероятностью, равной произведению их текущих весов: $p_i \cdot p_j$.

Полученный таким образом Марковский процесс обладает следующими ключевыми свойствами:

а) **Реверсивность.** Процесс обладает конечным математическим ожиданием времени возврата в исходное состояние.

б) **Непериодичность.** Вероятность остаться в текущем состоянии после очередного шага не равна нулю.

в) **Неразложимость.** Любое состояние процесса достижимо из любого другого состояния за конечное число шагов, что легко доказывается индуктивным упорядочением состояний.

Перечисленные свойства обеспечивают сходимость рассматриваемого марковского процесса к некоторому стационарному распределению [24]. Как показано ранее в [18], такое стационарное распределение является равномерным по всем векторам вероятностей $p = (p_1, p_2, \ldots, p_n)$ с условием нормировки $\sum_{i=1}^{n} p_i = 1$. Данное распределение соответствует плоскому распределению Дирихле ($Dir(1, 1, \ldots, 1)$).

Практическая генерация такого равновесного распределения выполняется посредством нормирования независимых случайных величин, распределённых экспоненциально с параметром 1 [25]. А именно, если независимо сгенерировать

набор $\alpha_i \sim Exp(1)$ для каждого $i \in \{1, 2, \ldots, n\}$ и вычислить их сумму $M = \sum_{i=1}^{n} \alpha_i$, то вектор вероятностей

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

будет иметь желаемое равномерное (плоское) распределение Дирихле.

1.2.3. Метрика минимального числа операций

Для количественной оценки эволюционных различий между двумя геномными структурами обычно вводится понятие расстояния, основанного на числе эволюционных событий. Парсимониальное расстояние определяется как минимальное число определённых операций перестройки, необходимое для превращения одной геномной последовательности в другую. Таким образом, расстояние в метрике двойной разрез и склеивание даёт наименьшее число разрезов/слияний, требуемых для преобразования одного генома в другой.

Расстояние, определённое как минимум операций (иногда его называют парсимониальное расстояние), широко используется благодаря вычислительной простоте во многих случаях. Для бездупликатных геномов парсимониальное расстояние вычисляется напрямую из графа разрывов, как отмечалось выше. В более простых случаях монохромосомных геномов без инверсий расстояние сводится к подсчёту количества циклов в графе. Во всех этих случаях такое расстояние действительно является метрикой на множестве геномных последовательностей.

Однако метрика минимальных операций адекватна эволюционному расстоянию лишь при условии небольшого числа различий. Для близкородственных геномов, которые эволюционировали от общего предка посредством относительно малого числа перестроек, минимальное число операций практически совпадает с истинным числом событий (поскольку маловероятны сложные случаи, когда несколько перестроек “компенсируют” друг друга). Но в случае отдалённых геномов (существенно расходившихся длительное время) парсимонианское расстояние становится ненадёжным: оно, как правило, занижает реальное число произошедших перестроек. Это связано с тем, что при значительном эволюционном расстоянии многие перестройки могут накладываться, повторно ломать уже однажды разорванные места или независимо затрагивать одни и те же области. В результате две сильно перестроенные геномные последовательности могут казаться ближе (в терминах минимальных операций), чем это было бы по факту эволюции. Например, если какой-то участок хромосомы переворачивался несколько, финальное сравнение покажет либо одно изменение, либо вообще отсутствие отличий (если вернулся исходный порядок), тогда как реально событий было больше одного.

1.2.4. Вероятностная модель поломки случайных регионов

Одним из основных вероятностных подходов к моделированию эволюции генома является модель случайных разрывов (англ *Random Breakage Model, RBM*) [26]. В рамках этой модели предполагается, что у генома нет предпочтительных мест для перестроек: каждое возможное место разрыва равновероятно может участвовать в перестройке, и события разрыва происходят независимо друг от друга. Иными словами, не существует *горячих точек* хромосомных перестроек, и распределение точек разрыва по геному однородно. Данная гипотеза восходит к работе Nadeau and Taylor (1984), в которой анализировалась длина сохранившихся сегментов между видами. Выводы указали, что распределение размеров синтенных блоков соответствует случайному (показательному) распределению разрывов, что подтвердило модель случайных поломок на том этапе исследований.

Модель случайных разрывов можно переформулировать на языке случайных графов: если представить каждый потенциальный разрыв (между двумя соседними основаниями генома или между блоками) как “ребро” между сегментами, то каждая перестройка соответствует случайному выбору такого ребра. За длительное эволюционное время множество перестроек приведёт к тому, что геном дробится на сегменты, и процесс можно уподобить случайному разбиению отрезка на части. Предсказания данной модели включают, например, экспоненциальное распределение размеров оставшихся цельных сегментов генома и линейную зависимость количества разрывов от эволюционного времени.

Однако последующие исследования поставили под сомнение универсальность модели случайных разрывов. В начале 2000-х с накоплением сравнительных данных по полным геномам было обнаружено, что разрывы далеко не всегда распределены равномерно: напротив, некоторые области генома разных видов совпадают по расположению точек разрыва гораздо чаще, чем ожидалось случайно. Так, Певзнер и Теслер [12] при сравнении геномов человека и мыши выявили кластеры повторно используемых точек разрыва, противоречащие модели случайных разрывов. Они предположили существование хрупких (англ. *fragile*) в геноме, более склонных к перестройкам, и сформулировали альтернативную гипотезу эволюции хромосом, получившую название модель хрупких разрывов (англ. *fragile breakage model*).

Тем не менее, модель случайных разрывов остаётся важной нулевой моделью: она проста и позволяет выводить явные формулы. Например, если геномы эволюционируют по этой модели, можно попытаться оценивать число перестроек на основе наблюдаемого числа разрывов и некоторых статистических гипотез. Однако, как было отмечено выше, такая оценка будет систематически заниженной при наличии повторных разрывов в одних и тех же местах. Для учета этого феномена нужны более сложные модели.

1.2.5. Вероятностная модель поломки хрупких регионов

Модель хрупких регионов (англ. *Fragile Breakage Model, FBM*) была предложена для объяснения отклонений от случайного распределения разрывов. В рамках FBM предполагается, что геном состоит из участков с различной хрупкостью: одни регионы могут многократно участвовать в перестройках (“хрупкие”), тогда как другие относительно устойчивы и разрываются редко (“устойчивые”). Таким образом, перестройки происходят не в случайных местах, а преимущественно в определённых горячих точках (англ. *hotspots*). Модель хрупких регионов более сложна, чем модель случайных регионов, но она позволяет получать оценки числа перестроек с учётом наблюдаемого повторного использования разрывов. Например, если в сравнении геномов выявлено меньше разрывов, чем ожидалось для данного числа операций, модель хрупких регионов объясняет это тем, что некоторые операции приходились на одни и те же места (“накладывались”). Для количественной оценки эволюционной дистанции в таких условиях разрабатываются статистические методы, основанные на вероятностных характеристиках графа разрывов. В частности, учитывается распределение циклов в графе разрывов: наличие необычно большого количества циклов определённых длин может свидетельствовать о неоднократных перестройках в одних и тех же местах.

Поставленная гипотеза получила развитие в количественных моделях: в частности, Танье и др. [18] предложили формализованную стохастическую модель эволюции, называемую INFER (англ. *Inversion History with Fragile Regions*). Хотя изначально INFER формулировалась для инверсий, она обобщается на любые операции эмулируемые с помощью двойного разреза и склеивания.

В модели INFER каждому потенциальному месту разрыва (каждому хрупкому региону и теломерам) приписывается вероятность p_i быть задействованным в перестройке (причём $\sum_i p_i = 1$). Эволюция генома моделируется как марковский процесс: на каждом шаге выбираются два места разрыва (скажем, i и j) с вероятностями p_i и p_j соответственно, после чего выполняется операция DCJ, затрагивающая эти два места. В результате образуются новые точки разрыва (например, если разрыв произошёл внутри региона, он разделяется на два новых региона), и вероятности ломкости обновляются для новых регионов по некоторому правилу. В версии [18] используется правило равномерного “разделения” вероятностей: грубо говоря, вероятность p_i разорванного региона распределяется между вновь образованными частями пропорционально случайным коэффициентам, чтобы их сумма равнялась p_i. Это означает, что регион, однажды разорвавшийся, может частично сохранить высокую ломкость в одной из своих частей. Несмотря на то, что в данной статье были предложена вероятностная модель более точно описывающая эволюцию генома, предложенные оценки являются лимитированными и выражены в виде бесконечных рядов.

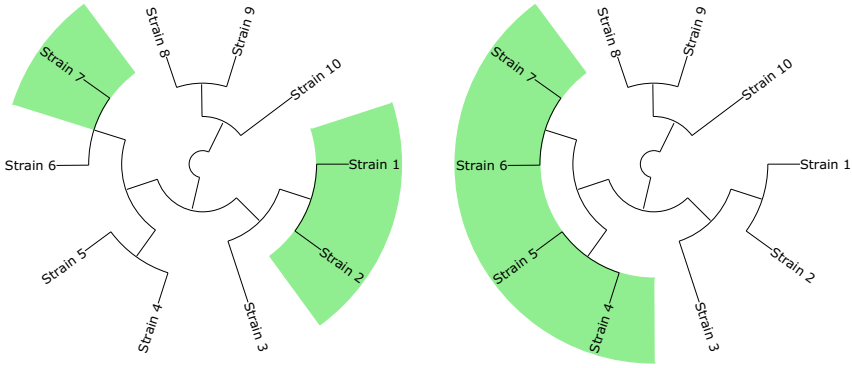
Таким образом, использование вероятностных моделей и соответствующих статистических оценивателей позволяет более надёжно измерять расстоя-

ния между геномными структурами, что важно для построения корректных филогенетических гипотез и понимания механизмов эволюции.

1.3. Методы анализа параллельных изменений в древовидных структурах

1.3.1. Параллельные изменения как выпуклые признаки на деревьях

В филогенетике признак (англ. *character*) называют выпуклым (англ. *convex*) на данном эволюционном дереве, если множество видов (листьев), обладающих этим признаком, образует связное поддерево. Эквивалентно, признак выпуклый (свободный от гомоплазии), если его можно объяснить одним появлением (или одним исчезновением) на некоторой ветви дерева 5б. В противном случае признак является невыпуслым (с гомоплазией), то есть его возникновение или утрата требуются в нескольких независимых точках дерева для согласования с наблюдаемым распределением 5а. В контексте геномных перестроек признаком может служить, например, наличие определённого геномного соседства или, напротив, факт разрыва между двумя геномными элементами. Если такой признак выпуклый, это означает, что соответствующая перестройка произошла единожды у общего предка группы организмов. Если же признак невыпуклый, значит похожие перестройки имели место неоднократно в разных филогенетических линиях (то есть налицо параллелизм).



(а) Признак с гомоплазией (невыпуклый)

(б) Признак, свободный от гомоплазии (выпуклый)

Рисунок 5 – Примеры состояния признака на дереве

Более формально: пусть дано множество таксонов X и филогенетическое X-дерево (T, ϕ), где $T = (V, E)$ — граф, а $\phi : X \to V$ — отображение, связывающее множество видов с листьями графа. Признак на множестве таксонов X

определяется как функция χ, отображающая некоторое непустое подмножество $X' \subseteq X$ в конечное множество состояний признака C:

$$\chi : X' \to C.$$

Признак χ называется выпуклым на дереве (T, ϕ), если существует такая функция расширения признака $\bar{\chi} : V \to C$, удовлетворяющая следующим условиям:

а) $\bar{\chi}(\phi(x)) = \chi(x)$ для всех $x \in X'$, то есть расширение согласовано с исходным распределением признака по листьям;

б) для каждого состояния признака $\alpha \in C$ индуцированный подграф дерева T, образованный вершинами множества $\{v \in V \mid \bar{\chi}(v) = \alpha\}$, является связным.

Алгоритмически выпуклость признака может быть проверена алгоритмом Фитча. Этот алгоритм для каждого признака на данном дереве вычисляет минимальное число изменений состояния ($0 \leftrightarrow 1$, где 1 — признак присутствует) вдоль ветвей, необходимое для воспроизведения наблюдаемого распределения 0/1 на листьях. Если минимальное число изменений больше 1, признак не может быть объяснён одним появлением — следовательно, он параллельный.

Важно отметить, что невыпуклость признака может быть следствием как реальных параллельных процессов, так и артефактов (например, неточного построения дерева). Признак, требующий два изменения, иногда можно сделать выпуклым, чуть изменив топологию дерева. По этой причине, анализируя параллельные перестройки, следует убедиться в надёжности филогенетической основы и при возможности использовать дополнительные данные (например, информацию о функциях разрываемых регионов), чтобы исключить ложные совпадения.

1.3.2. Литературные примеры параллельных изменений

Параллельные перестройки геномов наиболее ярко задокументированы у микроорганизмов, где сравнительно небольшие геномы и обилие штаммовых данных позволяют точно отследить независимые события. Например, в популяциях бактерий *Pseudomonas aeruginosa* наблюдалась инверсия крупного фрагмента хромосомы, фланкированного рибосомными оперонами, которая возникала независимо в разных изолятах [28]. Показано, что эта перестройка (переворот сегмента между двумя копиями rRNA-гена) приводит к изменениям фенотипа – влиянию на устойчивость к окислительному стрессу, метаболизм и вирулентность; то есть, вероятно, она подвергалась отбору в сходных условиях, возникнув параллельно у разных потомков без недавнего общего предка.

В работе, посвящённой изучению эволюции стрептококков, выявлены независимые инверсии, связанные с паралогичными генами PhtD и PhtB [27].

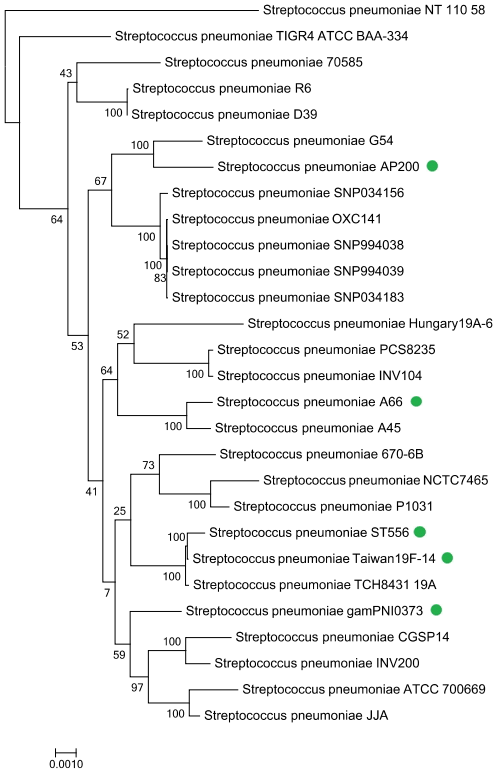


Рисунок 6 – Генетическое дерево демонстрирует распределение инверсий по генам PhtD и PhtB, штаммы с такими инверсиями выделены зелёным [27]

Эти перестройки обнаружены у различных штаммов, филогенетически разделённых и не образующих общую кладу по данным генам. Несмотря на то, что механизм таких инверсий, вероятно, связан с гомологичной рекомбинацией, филогенетический анализ показал, что деревья, построенные на последовательностях генов, задействованных в инверсии, согласуются с общими филогениями этих штаммов, подтверждая независимый характер возникновения событий. Аналогичные результаты были получены и для других видов рода *Streptococcus*, где параллельные инверсии по указанным паралогам также были выявлены.

Другой пример касается патогенного стрептококка *Streptococcus pyogenes*. У этого вида обнаружены параллельные изменения числа копий определённых блоков, связанные с фаговыми инсерциями [27]. В частности, блок, содержащий рРНК-оперон, в большинстве геномов представлен несколькими (шестью) копиями, но в некоторых эволюционно удалённых линиях независимо наблюдались либо утраты одной копии, либо, наоборот, приобретение дополнительных копий

(до четырёх сверх нормы). Анализ геномного окружения показал, что эти изменения копийности связаны с независимыми интеграциями и потерями фаговых последовательностей в разных кладах *S. pyogenes*. Таким образом, хотя у общего предка всех штаммов было шесть копий оперона, последующие перестройки в различных ветвях привели к разным вариантам — классический случай параллельной эволюции геномной структуры.

Высокие скорости и параллельность перестроек отмечены и в ряде других патогенов.

Хотя большинство исследований параллельных перестроек сосредоточены на прокариотах, аналогичные явления наблюдаются и в эукариотических геномах. Особенно интересны случаи, в которых крупномасштабные хромосомные инверсии и слияния возникают независимо в различных филогенетических линиях.

Так, в работе [29] были описаны участки генома человека, демонстрирующие повторяющееся переключение ориентации (англ. *inversion toggling*). Такие инверсии, возникающие независимо у разных особей, охватывают сотни килобаз и часто обогащены генами. Интересно, что значительная часть этих регионов совпадает с участками, ранее инвертировавшимися в ходе эволюции приматов, что говорит о древнем и устойчивом характере таких перестроек.

Особый интерес вызывает обнаруженная автором предвзятость по отношению к половым хромосомам: 45% всех повторяющихся инверсий были локализованы на хромосомах X или Y, что может объясняться особенностями репарации ДНК в непарных участках этих хромосом. Эти наблюдения подтверждают, что половые хромосомы представляют собой “горячие точки” структурной нестабильности, где независимо могут происходить сходные перестройки, включая разрывы и инверсии [29].

Стоит отдельно упомянуть гипотезу сети аберрантных филогений, основанная на отборе (англ. *SNAP, Selection-driven Network of Aberrant Phylogenies*), предложенную для объяснения параллельных перестроек [30]. Согласно этой гипотезе, перестройки могут возникать параллельно под действием сходных селекционных давлений при адаптации к новой нише, особенно если в геноме имеются места, ломкость которых обеспечивает быстрый адаптивный ответ. Иными словами, сходная среда может “направлять” эволюцию разных популяций по похожим структурным путям, вызывая независимые перестройки в аналогичных локусах. Примеры с инверсиями около rRNA-генов и фаговыми инсерциями в патогенах согласуются с этой идеей, так как соответствующие перестройки дают преимущество в определённых условиях (иммунное уклонение, регуляция вирулентности) и потому появляются в разных линиях независимо.

1.3.3. Задача оценки степени параллельности

После выявления параллельных геномных перестроек естественным следующим шагом является количественная оценка степени такого параллелизма.

Понимание, насколько широко распространены независимые повторения структурных изменений, важно для интерпретации их биологического значения и выявления возможных адаптивных механизмов, формирующих эволюцию вида.

В простейшем случае такая оценка может сводиться к подсчёту числа признаков, возникших независимо на разных ветвях филогенетического дерева. Однако подобный подход не учитывает того, что отдельные события могут повторяться неодинаковое количество раз, и поэтому не всегда отражает истинный масштаб параллельности.

В связи с этим возникает потребность в более информативных агрегированных показателях, способных учитывать как частоту, так и «силу» повторения перестроек. Например, таким агрегированным показателем может служить доля параллельных событий от общего числа всех зафиксированных геномных перестроек в данной группе организмов. Другим примером является индекс гомоплазии, отражающий степень повторности появления одинаковых структурных изменений на эволюционном дереве. Подобные метрики позволяют дать более глубокое представление о характере и частоте параллельных эволюционных событий и, как следствие, помочь понять их потенциальную адаптивную роль или выявить механизмы структурной нестабильности, присущие определённым геномным регионам.

Выводы по главе 1

В данной главе был проведён обзор фундаментальных концепций и математических моделей, используемых для изучения эволюционных процессов на примере структурных перестроек геномов.

В частности, были рассмотрены классические модели случайных графов, такие как модель Эрдеша–Реньи и её модификации, включая конфигурационную модель и модель предпочтительного присоединения. Было показано, что данные модели способны описывать важные пороговые явления, такие как появление гигантской компоненты. Также рассмотрены аффинные модификации, позволяющие более адекватно моделировать реальные системы с неоднородными вероятностями появления связей.

Далее был представлен граф точек разрыва, используемый для формализации геномных перестроек. Граф точек разрыва и его модификации (в том числе взвешенный вариант) обеспечивают удобный формализм для оценки расстояния между геномами. На основе данного графа была подробно описана операция двойного разреза и слияния, являющаяся основной единицей эволюционных изменений в рамках рассматриваемых моделей.

Также были подробно рассмотрены вероятностные модели эволюции, такие как модель случайных разрывов и модель хрупких регионов. Последняя модель учитывает неоднородность геномов в плане вероятностей разрывов и является более реалистичной с биологической точки зрения. В рамках вероятностного подхода была описана эволюционная модель с весами, представленная как

марковский процесс, и показано, что такая модель обладает стационарным равномерным распределением (распределением Дирихле).

Наконец, были рассмотрены методы анализа параллельных изменений на филогенетических деревьях. В частности, было формализовано понятие выпуклости признаков на деревьях, являющееся важным инструментом для выявления параллельных эволюционных событий. Приведены литературные примеры параллельных перестроек в геномах бактерий и эукариот, продемонстрировавшие, что параллельные изменения являются широко распространённым явлением. Также была поставлена задача количественной оценки степени параллельности таких изменений и предложены возможные подходы к её решению.

Таким образом, представленный обзор позволяет сформулировать основные задачи дальнейших исследований, связанные с разработкой и применением новых вероятностных и комбинаторных методов для более точного моделирования и анализа параллельных структурных перестроек геномов.

Глава 2. Анализ случайных графов и оценка расстояний между структурами

2.1. Математическое ожидание числа компонент заданного размера

Одним из важнейших параметров рассматриваемой модели является количество компонент (связных частей графа) заданного размера. Для ясности сначала рассмотрим случаи небольшого размера (например, компоненты из одного элемента), а затем получим общую формулу для произвольного размера m. Наконец, будет обсуждена асимптотика появления "гигантской" компоненты при увеличении размера.

2.1.1. Компоненты размера 1

Рассмотрим вероятность того, что определённая связь (ребро) остаётся неизменной после случайных преобразований. Пусть всего имеется n элементов (блоков) и соответственно n возможных связей между ними. Каждая связь i (адресованная соответствующим параллельным фрагментам в двух структурах) характеризуется показателем "хрупкости" $p_i \in [0,1]$, отражающим вероятность участия этой связи в случайной перестройке на каждом шаге. Предполагается, что значения p_i нормированы и суммарно образуют распределение вероятностей на множестве связей: $\sum_{i=1}^{n} p_i = 1$. На каждом элементарном шаге происходит случайное преобразование структуры, в ходе которого выбираются две связи (не обязательно различные, если операция разрыва-вставки затрагивает одну связь дважды) для перестройки. Таким образом, на каждом шаге вероятность того, что именно i-я связь будет вовлечена в операцию, равна $2p_i$.

Пусть за всю историю произошло k случайных операций. Тогда вероятность того, что фиксированное i-е ребро *никогда* не участвовало ни в одной из k перестроек, равна $(1-p_i)^{2k}$ (так как на каждом из k шагов данная связь должна остаться вне двух задействованных связей). Введём индикаторную случайную величину

$$\mathbf{1}_{\{i\text{-я связь не затронута}\}},$$

которая равна 1, если связь i осталась нетронутой, и 0 иначе. Общее число компонент размера 1 (то есть отдельных связей, сохранившихся неизменными) обозначим $c_1 = \sum_{i=1}^{n} \mathbf{1}_{\{i\text{-я связь не затронута}\}}$. Тогда по определению математического ожидания

$$E(c_1) = \sum_{i=1}^{n} P(i\text{-я связь не участвовала в перестройках}) = \sum_{i=1}^{n} (1-p_i)^{2k}.$$

Нас будет интересовать нормированное значение, то есть средняя доля таких связей:

$$E\Big(\frac{c_1}{n}\Big) = \frac{1}{n}\sum_{i=1}^{n} (1-p_i)^{2k}.$$

Для дальнейшего анализа будем считать число элементов n большим и перейти к предельному случаю $n \to \infty$. Одновременно будем полагать, что число операций k линейно зависит от n. А именно, предположим, что существует постоянная $\gamma > 0$ такая, что

$$\frac{2k}{n} \;\to\; \gamma \quad \text{при } n \to \infty\,,$$

то есть $k \sim \frac{\gamma n}{2}$. Таким образом, γ характеризует *интенсивность* перестроек (ожидаемое число операций на элемент структуры).

Заметим, что при указанных предположениях величины p_i сами по себе случайны и, вообще говоря, различны для разных i. В модели предполагается, что значения p_i независимы и одинаково распределены: интенсивности "хрупкости" фрагментов подчиняются некоторому распределению. Мы будем рассматривать *равномерный случай*, когда показатели p_i имеют так называемое "плоское" распределение Дирихле, соответствующее отсутствию априорных предпочтений (все фрагменты структуры в среднем одинаково хрупки). В этом случае можно считать, что случайные величины $\alpha_i = np_i$ являются независимыми экспоненциальными величинами с единичным математическим ожиданием. Иными словами, при $n \to \infty$ суммарная мера $M = \sum_{i=1}^{n} \alpha_i$ по закону больших чисел будет близка к n (более строго, $\frac{M}{n} \to 1$ при $n \to \infty$). Это даёт приближение $p_i = \frac{\alpha_i}{M} \approx \frac{\alpha_i}{n}$ для каждого i. Используя это и предположение о линейном росте k, можем оценить:

$$(1-p_i)^{2k} \;=\; \Big(1 - \frac{\alpha_i}{M}\Big)^{2k} \;\approx\; \Big(1 - \frac{\alpha_i}{n}\Big)^{\gamma n} \;\xrightarrow[n\to\infty]{}\; e^{-\gamma\alpha_i}\,,$$

поскольку $\frac{M}{n} \to 1$. Таким образом,

$$E\Big(\frac{c_1}{n}\Big) \;\approx\; \frac{1}{n}\sum_{i=1}^{n} e^{-\gamma\alpha_i}\,.$$

Так как α_i независимы и распределены по экспоненциальному закону с плотностью $f(\alpha) = e^{-\alpha}$ на $\mathbb{R}_+$, то можно проинтегрировать полученное выражение по распределению α_i. Практически это означает, что при больших n сумма $\frac{1}{n}\sum_i e^{-\gamma\alpha_i}$ близка к математическому ожиданию $E(e^{-\gamma\alpha})$ для одной экспоненциальной величины α. Последнее легко находится напрямую:

$$E(e^{-\gamma\alpha}) \;=\; \int_0^\infty e^{-\gamma\alpha} e^{-\alpha} d\alpha \;=\; \int_0^\infty e^{-(\gamma+1)\alpha}\, d\alpha \;=\; \frac{1}{\gamma+1}\,.$$

Отсюда получаем предельную долю неизменных связей:

$$\lim_{n\to\infty} E\Big(\frac{c_1}{n}\Big) \;=\; \frac{1}{1+\gamma}\,.$$

Другими словами, при большом количестве элементов n и фиксированном отношении $\gamma = 2k/n$, в среднем доля компонент размером 1 равна $\frac{1}{1+\gamma}$. Например,

если число операций k примерно равно числу элементов ($\gamma \approx 2$), то лишь порядка $1/(1+2) \approx 33\%$ связей остаются нетронутыми; если же операций существенно меньше ($\gamma \ll 1$), то значительная часть ($\approx 100\%$ при $\gamma \to 0$) исходных связей сохранится.

Аналогичным образом можно получить выражение для числа компонент, состоящих из двух элементов. В контексте задачи такие компоненты размера 2 соответствуют появлению цикла длины 2 (пары связанных элементов) в результирующей структуре. Пусть фиксированы два конкретных ребра (связи) i и j. Для того чтобы из них образовалась отдельная компонента (цикл длины 2), необходимо, чтобы на одном из шагов эти связи были одновременно выбраны для перестройки и при этом склеены вместе, а на всех остальных шагах ни i, ни j не участвовали в операциях. Вероятность того, что на некотором заданном шаге выбраны именно i и j, приблизительно равна $p_i p_j$ (для простоты принимая независимость выборов, поскольку $p_i, p_j \ll 1$ при большом n). Поскольку подходящий шаг может быть любым из k, суммарная вероятность того, что *существует* шаг, на котором произошла перестройка с участием именно этих двух связей, оценится как $k\, p_i p_j$. Зафиксировав такой шаг, требуется, чтобы на всех остальных $(k-1)$ шагах связи i и j не затрагивались; вероятность этого равна $(1-p_i-p_j)^{2(k-1)}$ (на каждом из остальных шагов ни одна из двух связей не должна попасть в выбранную пару). Перемножая указанные факторы, получаем вероятность конкретного сценария образования компоненты $\{i,j\}$:

$$P(\text{связи } i,j \text{ образуют цикл длины } 2) \;\approx\; k\, p_i\, p_j \left(1-p_i-p_j\right)^{2(k-1)}.$$

Просуммировав эту вероятность по всем упорядоченным парам (i,j) (где $1 \leqslant i < j \leqslant n$), получим математическое ожидание общего числа таких компонентов. Нормируя на n, запишем:

$$E\Big(\frac{c_2}{n}\Big) \;=\; \frac{1}{n} \sum_{1 \leqslant i < j \leqslant n} k\, p_i\, p_j \left(1-p_i-p_j\right)^{2(k-1)}.$$

Далее, аналогично случаю $m=1$, можно провести предельный переход $n \to \infty$. При больших n и $\gamma = 2k/n = \text{const}$ величины p_i малые, и вышеуказанная сумма может быть проинтегрирована, заменив $p_i = \alpha_i/n$, $p_j = \alpha_j/n$ и используя экспоненциальное распределение α_i. В результате получается некоторое явное выражение в замкнутом виде через γ. Тем не менее, вместо того чтобы отдельно вычислять частные случаи $m = 1, 2$, целесообразно сразу приступить к выводу общей формулы для компоненты произвольного размера m. При этом перечисленные частные случаи (для $m=1$ и $m=2$) будут получены как следствия общей формулы.

2.1.2. Компоненты произвольного размера

Теперь рассмотрим произвольное фиксированное число $m \geqslant 1$ и вычислим математическое ожидание $E(c_m)$, где c_m – случайное число компонент (в результ-

тирующей структуре) размера m. В терминах графовой модели это означает, что нас интересует число циклов длины m, образовавшихся после k случайных перестроек. Для простоты дальнейшего изложения будем называть базовые прочные фрагменты структуры *блоками*. Тогда цикл длины m соответствует объединению m различных блоков в одну компоненту.

Рассмотрим некоторое фиксированное подмножество из m определённых блоков (скажем, блоки с номерами $1, 2, \ldots, m$) и исследуем вероятность того, что именно эти m блоков в итоге образуют единый цикл (компоненту размера m). Очевидно, что для образования цикла из m блоков необходимо совершить ровно $m-1$ операций, каждая из которых соединяет между собой две разрозненные группы блоков. Поскольку всего совершается k операций, то нам нужно сначала выбрать, на каких шагах происходили эти $m-1$ объединений. Число способов выбрать $m-1$ шагов из k равно $\binom{k}{m-1}$. Далее, когда выбраны шаги, необходимо определить, каким образом на этих шагах происходило слияние указанных m блоков в единый цикл. Этот процесс можно описать последовательностью упорядоченных пар (i, j), где i и j – номера блоков, которые объединяются на данном шаге (причём договариваемся всегда указывать $i < j$). Такая последовательность пар кодирует *сценарий* объединения блоков. Например, на рис. 7 показан сценарий объединения 4 блоков с номерами $1, 2, 3, 4$ в цикл длины 4, задаваемый последовательностью a_{13}, a_{23}, a_{34} (здесь a_{ij} обозначает операцию слияния блоков i и j на соответствующем шаге).

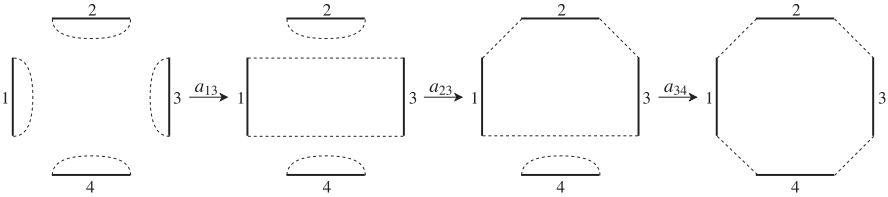


Рисунок 7 – Пример объединения в цикл длины 4 (сценарий: a_{13}, a_{23}, a_{34}).

При такой визуализации блоки пронумерованы для наглядности, но следует иметь в виду, что эти номера соответствуют определённым исходным связям (потенциальным разрывам) между блоками. Иными словами, операции происходят над связями, соединяющими блоки в исходной конфигурации: когда говорится, что выполняется операция a_{ij}, подразумевается, что на данном шаге выбраны связи, инцидентные блокам i и j (каждый блок соединён в исходной структуре двумя связями с соседями, если рассматривать структуру как граф). Таким образом, слияние блоков i и j в рамках одной операции означает, что разрываются связи, ранее соединявшие эти блоки с их соседями, и затем блок i и j становятся соседними в новой структуре (образуя часть цикла).

Каждый сценарий объединения m фиксированных блоков в цикл можно однозначно отобразить на помеченное дерево с m вершинами (так называемое

остовное дерево, соединяющее эти m элементов). Блоки $1, 2, \ldots, m$ станут вершинами такого дерева. Проведём ребро между вершинами i и j в дереве, если в сценарии имеется шаг a_{ij} (то есть на одном из шагов были объединены блоки i и j). На рис. 8 представлен пример остовного дерева, соответствующего приведённому выше сценарию (для блоков $1, 2, 3, 4$). Легко видеть, что построенное соответствие является биекцией: каждому возможному сценарию объединения m блоков соответствует единственное дерево на этих m вершинах, и наоборот, любое остовное дерево на m узлах задаёт ровно один сценарий слияния блоков. Однако эта биекция не является однозначной, а имеет определённый коэффициент кратности. Действительно, дерево не отражает порядка, в котором производились операции (то есть все $m-1$ шагов могут идти в разном хронологическом порядке, приводя к одному и тому же конечному результату). Перестановок $m-1$ шагов существует $(m-1)!$ штук, что даёт соответствующий множитель. Кроме того, каждая отдельная операция слияния могла происходить в одном из двух "направлений", то есть при заданных двух выбранных связях есть два варианта, как именно соединить их концами при перестройке (обе возможности приводят к объединению циклов, но различаются переставляемыми концами разорванных связей). Для цикла произвольной длины каждый из $m-1$ шагов обладает двумя такими вариантами, что даёт множитель 2^{m-1}. Таким образом, одному остовному дереву соответствует $2^{m-1}(m-1)!$ различных сценариев слияния.

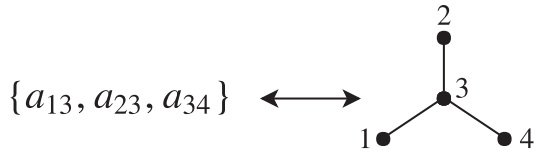


Рисунок 8 – Пример биекции сценария с остовным деревом для $m = 4$.

Теперь оценим суммарную вероятность всех возможных сценариев, приводящих к объединению заданных блоков $1, \ldots, m$ в один цикл. Как отмечалось выше, на определённом шаге операция затрагивает связи, инцидентные некоторым блокам. Предположим, что показатель хрупкости каждого из выбранных m блоков оставался постоянным на всём протяжении процесса и равен p_i для i-го блока. Тогда вероятность конкретного сценария S можно разложить по шагам: на каждом шаге, когда объединяются блоки i и j, в вероятность входит множитель $p_i p_j$ (пропорционально вероятности выбора связей этих блоков). Таким образом, вероятность сценария S равна произведению $p_i^{d_i}$ по всем i из $\{1, \ldots, m\}$, где d_i – степень вершины i в соответствующем дереве T_S. Иными словами, d_i равно числу операций, в которых участвовал i-й блок при данном сценарии. Поэтому вероятность сценария S запишется как

$$P(S) \;=\; \prod_{i=1}^{m} p_i^{\,d_i}\,.$$

Чтобы найти искомую суммарную вероятность, нужно просуммировать $P(S)$ по всем сценариям S, приводящим к объединению блоков $1,\dots,m$. Благодаря описанной выше биекции, вместо суммирования по сценариям можно суммировать по всем возможным помеченным деревьям T на m вершинах, учитывая множитель кратности $2^{\,m-1}(m-1)!$ сценариев на одно дерево. То есть

$$\sum_{S} P(S) \;=\; 2^{\,m-1}(m-1)!\,\sum_{T}\prod_{i=1}^{m} p_i^{\,d_i(T)}\,,$$

где сумма справа берётся по всем остовным деревьям T, соединяющим вершины $1,\dots,m$, а $d_i(T)$ – степень вершины i в дереве T. Известный результат комбинаторики (следствие кодировки Прюфера для labeled-деревьев) гласит, что

$$\sum_{T}\prod_{i=1}^{m} p_i^{\,d_i(T)} \;=\; (p_1+p_2+\cdots+p_m)^{\,m-2}\,p_1p_2\cdots p_m\,.$$

Действительно, число различных помеченных деревьев на m заданных вершинах равно $m^{\,m-2}$ (формула Кэли), а указанное взвешенное суммирование приводит именно к указанной комбинации (её можно вывести напрямую из представления дерева кодом Прюфера длины $m-2$). Таким образом, мы получили значение суммарной вероятности:

$$\sum_{S} P(S) \;=\; 2^{\,m-1}(m-1)!\,p_1p_2\cdots p_m\,(p_1+p_2+\cdots+p_m)^{\,m-2}\,.$$

Именно эта величина и представляет собой вероятность того, что выбранные блоки $1,\dots,m$ объединятся в единый цикл длины m (через все возможные сценарии). Формально данный результат можно сформулировать в виде леммы.

Лемма 1. Для фиксированного набора из m блоков с показателями хрупкости $p_1,p_2,\dots,p_m$ суммарная вероятность всех сценариев, приводящих к слиянию этих блоков в одну компоненту (цикл) размера m, равна

$$2^{\,m-1}(m-1)!\;p_1p_2\cdots p_m\;(p_1+p_2+\cdots+p_m)^{\,m-2}\,.$$

Итак, для фиксированных m блоков мы вывели вероятность того, что они образуют цикл длины m в результате k перестроек. Теперь получим выражение для математического ожидания общего числа таких компонент c_m. Заметим, что различных наборов из m блоков (которые могут образовать компоненту) всего $\binom{n}{m}$. Вероятность для каждого из них определяется формулой из леммы 1 выше. При суммировании по всем комбинациям нужно также учесть, что указанные события (образования конкретных циклов) не являются независимыми; однако

при вычислении *математического ожидания* мы можем воспользоваться принципом линейности ожидания и просуммировать вероятности без учёта пересечений. Поэтому:

$$E(c_m) \;=\; \binom{n}{m}\binom{k}{m-1}\,2^{\,m-1}(m-1)!\,\cdot$$

$$\sum_{1\leqslant i_1<\cdots<i_m\leqslant n} p_{i_1}p_{i_2}\cdots p_{i_m}\left(p_{i_1}+\cdots+p_{i_m}\right)^{m-2}\Bigl(1-\sum_{j=1}^{m}p_{i_j}\Bigr)^{2(k-m+1)}.$$

Здесь $\binom{n}{m}$ отвечает выбору m блоков, $\binom{k}{m-1}$ – выбору шагов для их слияния, а оставшийся множитель из леммы даёт сумму вероятностей всех сценариев слияния этих блоков. Множитель $(1-\sum_{j=1}^{m}p_{i_j})^{2(k-m+1)}$ представляет вероятность того, что на остальных $k-m+1$ шагах ни одна из выбранных m связей не была задействована (т.е. они не прерывались и не участвовали в посторонних операциях, кроме тех $m-1$ шагов, на которых происходило их целенаправленное объединение в цикл). Наконец, нормируем на общее число элементов n:

$$E\Bigl(\frac{c_m}{n}\Bigr) \;=\; \frac{1}{n}\binom{n}{m}\binom{k}{m-1}\,2^{\,m-1}(m-1)!\,\cdot$$

$$\sum_{1\leqslant i_1<\cdots<i_m\leqslant n} p_{i_1}\cdots p_{i_m}\left(p_{i_1}+\cdots+p_{i_m}\right)^{m-2}\Bigl(1-\sum_{j=1}^{m}p_{i_j}\Bigr)^{2(k-m+1)}. \qquad (1)$$

Выражение (2.1.2) ещё весьма громоздкое. Однако оно значительно упростится, если снова рассмотреть предельный случай $n\to\infty$ при фиксированном m и доле операций $\gamma = 2k/n$. Перейдём к пределу, аналогично тому, как мы это делали для $m=1$. В пределе можно заменить сумму по всем наборам $\{i_1,\dots,i_m\}$ интегралом по совместному распределению величин $\alpha_{i_1},\dots,\alpha_{i_m}$ (напомним, что $p_i=\alpha_i/M$ с $M\approx n$). Каждая из α_i имеет экспоненциальное распределение. После раскрытия биномиальных коэффициентов и замены суммирования интегралом, подробные выкладки приводят к следующему результату:

$$\lim_{n\to\infty} E\Bigl(\frac{c_m}{n}\Bigr) \;=\; \frac{(3m-3)!\,\gamma^{\,m-1}}{m!\,(2m-1)!\,(\gamma+1)^{\,3m-2}}\,.$$

Этот результат можно оформить как основной теоретический вывод для рассматриваемой модели.

Утверждение 1. Пусть $\gamma=\frac{2k}{n}$ есть асимптотическая интенсивность перестроек. Тогда для любого фиксированного $m\geqslant 1$ верно следующее:

$$\lim_{n\to\infty} E\Bigl(\frac{c_m}{n}\Bigr) \;=\; \frac{(3m-3)!\,\gamma^{\,m-1}}{m!\,(2m-1)!\,(\gamma+1)^{\,3m-2}}\,,$$

где c_m обозначает число компонент размера m, образовавшихся после k случайных перестроек.

Доказательство. Основные шаги доказательства уже изложены выше. Благодаря лемме 1, нам удалось получить комбинаторное выражение (2.1.2) для математического ожидания $E(c_m)$. Далее используется асимптотический анализ при $n \to \infty$. В этом предельном режиме величины $\alpha_i = np_i$ аппроксимируют экспоненциально распределённые независимые случайные величины, и сумма по всем наборам индексов $\{i_1, \dots, i_m\}$ заменяется интегрированием по $\alpha_1, \dots, \alpha_m \in \mathbb{R}_+^m$. Критическим шагом является вычисление следующего многомерного интеграла:

$$I_{m,\lambda} \;=\; \int\limits_{\mathbb{R}_+^m} \cdots \int \alpha_1 \alpha_2 \cdots \alpha_m \left(\alpha_1 + \alpha_2 + \cdots + \alpha_m\right)^{m-2} \cdot$$

$$e^{-\lambda(\alpha_1+\alpha_2+\cdots+\alpha_m)}\, d\alpha_1\, d\alpha_2 \cdots d\alpha_m\,,$$

где $\lambda = \gamma+1$. Вычисление интеграла $I_{m,\lambda}$ можно выполнить методом поэтапного интегрирования. Удобно перейти к новым переменным: $S = \alpha_1 + \cdots + \alpha_m$ (суммарная длина), и относительно пропорций $\beta_i = \alpha_i / S$ (так что $\beta_1 + \cdots + \beta_m = 1$). Якобиан такого перехода равен $S^{\,m-1}$. Тогда интеграл факторизуется:

$$I_{m,\lambda} \;=\; \int_0^\infty S^{\,m-1}\, S^{\,m-2}\, S^m\, e^{-\lambda S} dS \;\times \int\limits_{\substack{\beta_i \geqslant 0 \\ \beta_1+\cdots+\beta_m=1}} \cdots \int \left(\beta_1 \beta_2 \cdots \beta_m\right)^{1} d\beta_1 \cdots d\beta_{m-1}\,.$$

Здесь первый множитель возникает из $d\alpha_1 \cdots d\alpha_m \;=\; S^{\,m-1}\, dS\, d\beta_1 \cdots d\beta_{m-1}$, второй $S^{\,m-2}$ из $(\sum \alpha_i)^{m-2}$, а третий S^m из произведения $\alpha_1 \cdots \alpha_m$. Таким образом,

$$I_{m,\lambda} \;=\; \int_0^\infty S^{\,3m-3} e^{-\lambda S} dS \;\times \int_{\Delta_m} \beta_1^1 \beta_2^1 \cdots \beta_m^1\, d\beta_1 \cdots d\beta_{m-1}\,,$$

где Δ_m обозначает стандартный симплекс $\{\beta_i \geqslant 0, \sum \beta_i = 1\}$. Первый интеграл легко вычисляется как интеграл Гамма:

$$\int_0^\infty S^{\,3m-3} e^{-\lambda S} dS \;=\; \frac{(3m-3)!}{\lambda^{\,3m-2}}\,.$$

Второй интеграл представляет собой бета-функцию от параметров $(2, 2, \dots, 2)$ (по одному для каждого β_i). Известно, что

$$\int_{\Delta_m} \beta_1^{a_1-1} \beta_2^{a_2-1} \cdots \beta_m^{a_m-1} d\boldsymbol{\beta} \;=\; \frac{\Gamma(a_1)\Gamma(a_2)\cdots\Gamma(a_m)}{\Gamma(a_1+a_2+\cdots+a_m)}\,,$$

где $\boldsymbol{\beta} = (\beta_1, \dots, \beta_m)$ и Γ обозначает гамма-функцию. В нашем случае $a_i = 2$ для всех i, поэтому

$$\int_{\Delta_m} \beta_1^1 \beta_2^1 \cdots \beta_m^1 d\boldsymbol{\beta} \;=\; \frac{\Gamma(2)^m}{\Gamma(2m)} \;=\; \frac{1!^m}{(2m-1)!} \;=\; \frac{1}{(2m-1)!}\,.$$

Перемножив два найденных фактора, получаем

$$I_{m,\lambda} \;=\; \frac{(3m-3)!}{(2m-1)!\,\lambda^{\,3m-2}}\,.$$

Подставляя $\lambda = \gamma + 1$, находим главный член в выражении (2.1.2):

$$\frac{1}{n}\binom{n}{m}\binom{k}{m-1}2^{m-1}(m-1)!\;\frac{(3m-3)!}{(2m-1)!(\gamma+1)^{\,3m-2}} \;\approx$$

$$\approx\; \frac{n^m}{m!}\,\frac{(n\gamma/2)^{m-1}}{(m-1)!}\,2^{m-1}(m-1)!\;\frac{(3m-3)!}{(2m-1)!(\gamma+1)^{\,3m-2}}\,,$$

что при упрощении и делении на n действительно приводит к заявленной формуле. Более строгое доказательство требует оценки пренебрегаемых членов и обоснования замены суммирования интегрированием, однако качественный результат не меняется. □

Теорема 1 даёт полное асимптотическое описание распределения числа компонент фиксированного размера m в рассматриваемой модели. Отметим, что при $m = 1$ и $m = 2$ она согласуется с ранее рассмотренными частными случаями. Так, при $m = 1$ получаем

$$\lim_{n\to\infty} E(c_1/n) = \frac{(3\cdot 1-3)!\gamma^0}{1!\cdot(2\cdot 1-1)!(\gamma+1)^1} = \frac{1}{\gamma+1}\,,$$

как и было получено выше. При $m = 2$ имеем

$$\lim_{n\to\infty} E(c_2/n) = \frac{(3\cdot 2-3)!\gamma^{\,1}}{2!\cdot(2\cdot 2-1)!(\gamma+1)^4} = \frac{3!\,\gamma}{2\cdot 3!(\gamma+1)^4} = \frac{\gamma}{2(\gamma+1)^4}\,.$$

2.1.3. Асимптотика гигантской компоненты

Выше мы рассмотрели компоненты фиксированного размера m, который не растёт с увеличением n. Однако когда число операций k достаточно велико (в частности, $\gamma = 2k/n$ превышает некоторый порог), в системе могут появляться компоненты весьма крупных размеров, сравнимых с n. В теории случайных графов известно явление *гигантской компоненты*: при достижении критического числа связей появляется единственная связная компонента, охватывающая положительную долю от всех вершин графа, тогда как остальные компоненты остаются малыми (размер порядка $\log n$ или даже ограничен сверху константой). В классической модели Эрдиша–Реньи $G(n,p)$ порог возникновения гигантской компоненты соответствует средней степени вершин равной 1. В модели, изучаемой в данной главе, роль “вершин графа” играют базовые блоки структуры, а роль “случайных ребёр” – произошедшие перестройки (объединения и разбиения циклов). Несмотря на то, что данная модель не сводится напрямую к стандартному случайному графу, можно ожидать аналогичного фазового перехода

при некотором критическом значении $\gamma = \gamma_c$. Теоретический анализ [0] и полученные нами формулы позволяют утверждать следующее:

Утверждение 1. Существует критическое значение $\gamma_c = 0.5$, при превышении которого в предельном случае $n \to \infty$ в системе появляется гигантская компонента. А именно, если $\gamma < 0.5$, то все компоненты имеют ограниченный размер (с ростом n ни одна компонента не содержит существенной доли от общего числа блоков). Если же $\gamma > 0.5$, то с вероятностью, стремящейся к 1 при $n \to \infty$, существует ровно одна компонента размера $\Theta(n)$ (то есть пропорционального n) и она единственна, а остальные компоненты значительно меньше.

Доказательство. Утверждение о $\gamma < 0.5$ непосредственно следует из предыдущих результатов: когда γ мал, вероятность слияния большого числа блоков экспоненциально мала, и практически все перестройки происходят независимо, затрагивая разные области структуры. Формально, можно показать, что при $\gamma < 1/2$ справедливо асимптотическое равенство

$$\sum_{m=1}^{\infty} \frac{m\, E(c_m)}{n} \;=\; 1\,,$$

что означает, что сумма долей всех компонент (с учётом их размера m) исчерпывает 100% элементов. Отсюда видно, что подавляющее большинство блоков распределено по малым компонентам и ни одна компонента не содержит существенной части блоков. Наоборот, при $\gamma > 0.5$ указанная сумма (доля охваченных малыми компонентами) становится меньше 1, так что остаётся некоторый ненулевой остаток доли элементов, не входящих в компоненты ограниченного размера. Эти оставшиеся блоки и формируют гигантскую компоненту. Критическое значение $\gamma_c = 0.5$ можно считать точкой фазового перехода, при которой минимальное расстояние между структурами перестаёт совпадать с реальным (см. ниже). □

Отметим, что в рассматриваемой модели гигантская компонента имеет смысл «цикла большой длины» в результирующем графе несоответствия структур. Появление такой компоненты означает, что существенная часть исходных связей (блоков) оказалась вовлечена в одну большую перестроечную цепочку. Интересно, что критическое значение $\gamma_c = 0.5$ (то есть $k = n/4$ операций) совпадает с границей применимости парсимонийных методов оценки расстояния, о которой будет сказано ниже.

2.2. Вспомогательные леммы

2.2.1. Вычисление суммарной вероятности через коды Прюфера

В данном подразделе был доказан вспомогательный результат (лемма 1), использованный при выводе общей формулы для $E(c_m)$ в предыдущем разделе.

Лемма была сформулирована и доказана в рамках рассмотрения биекции между последовательностями слияния и остовными деревьями, при помощи кодирования Прюфера.

2.2.2. Аналитическое вычисление многомерных интегралов

При предельном переходе $n \to \infty$ для получения явных формул мы сталкиваемся с интегралами высокой размерности, например с интегралом $I_{m,\lambda}$, приведённым в доказательстве теоремы 1. Ниже формулируется обобщённый результат, лежащий в основе вычисления подобных интегралов.

Лемма 2. Для любых целых $m \geqslant 2$ и $\lambda > 0$ выполнено тождество:

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

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

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

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

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

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

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

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

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

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

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

$$= [\text{по лемме } 3 \text{ и лемме } 4] =$$

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

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

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

□

Лемма 2 позволяет напрямую получить формулу теоремы 1 из выражения (2.1.2), что и было сделано выше.

Лемма 3.

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

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

$$\int_0^u t \mathrm{d}t = \frac{u^2}{2} - \frac{0^2}{2} = \frac{u^2}{2}$$

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:

$$\int \cdots \int_{\sum_{i=1}^{n+1} t_i \leqslant u} t_1 \cdot \ldots \cdot t_{n+1} \mathrm{d}t_1 \ldots \mathrm{d}t_{n+1} = \int_0^u \frac{(u-t_{n+1})^{2n}}{(2n)!} t_{n+1} \mathrm{d}t_{n+1} =$$

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

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

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

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

□

Лемма 4.

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

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

$$\int_0^u t^2 \mathrm{d}t = \frac{u^3}{3} - \frac{0^3}{3} = \frac{u^3}{3}$$

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:

$$\int \cdots \int_{\sum_{i=1}^{n+1} t_i \leqslant u} t_1 \cdot t_2 \cdot \ldots \cdot t_{n+1} \mathrm{d}t_1 \ldots \mathrm{d}t_{n+1} = \int_0^u \frac{2(u - t_{n+1})^{2n+1}}{(2n+1)!} t_{n+1} \mathrm{d}t_{n+1} =$$

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

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

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

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

□

2.3. Метод оценки истинного расстояния между структурами

Теперь, когда мы провели подробный анализ распределения компонент различного размера, можно перейти к основной задаче – оценке реального расстояния (числа перестроек k) между двумя структурами на основе наблюдаемых различий в их организации. Предположим, что мы имеем две структуры (например, две геномных последовательности), одна из которых могла получиться из

другой в результате некоторой последовательности случайных операций (перестроек) по описанной модели. Мы можем сравнить эти две структуры и построить их *граф несоответствия* (breakpoint graph), вершинами которого являются исходные блоки, а рёбрами – пары блоков, соседствующие друг с другом в одной из структур (чёрные рёбра соответствуют соседству в первой структуре, красные – во второй). Связные компоненты в таком графе имеют вид черно-красных циклов (если структуры замкнуты по кругу) или черно-красных цепочек (если имеются свободные концы). Минимальное число перестроек (операций DCJ), необходимое для превращения одной структуры в другую, называют *парсимонийным расстоянием* d. Оно может быть вычислено непосредственно по графу несоответствия. В частности, в случае двух кольцевых структур без свободных концов выполняется простая формула:

$$d \;=\; n - c\,,$$

где n – число блоков (вершин графа), а c – количество циклов в графе (включая тривиальные циклы длины 1). Количество "разрывов" (breakpoints) между двумя структурами, обозначаемое b, определяется как число пар блоков, которые соседничают в одной структуре, но не соседничают в другой. Этот показатель тоже легко вычисляется по графу: каждый нетривиальный цикл длины m в графе несоответствия вносит m разрывов, а каждый тривиальный цикл (совпадающее соседство) не даёт разрывов. Таким образом, общее число breakpoints равно

$$b \;=\; n - c_1\,,$$

где c_1 – число тривиальных циклов (компонент размера 1) в графе.

Таким образом, мы можем считать, что для данных двух структур (после выравнивания их блоков) наблюдаемыми величинами являются b и d. Очевидно, всегда $d \leqslant b$ (парсимонийное расстояние не превосходит числа разрывов). Причина неравенства в том, что одна перестройка может ликвидировать сразу два разрыва, если она соединяет две ранее несоседних пары блоков. В частности, при небольшом количестве перестроек (когда γ мало) большинство операций устраняют лишь один разрыв (или даже ни одного, если операция разрывает уже правильно соседние блоки), и тогда $d \approx b$. С другой стороны, при большом числе перестроек возможны ситуации, когда d значительно меньше b за счёт того, что многие разрывы могли "случайно" исправиться позднейшими операциями. В таких случаях парсимонийный подход (d) будет существенно недооценивать реальное число операций k, и требуется статистическая коррекция.

Опираясь на результаты предыдущих разделов, разработаем метод оценки истинного расстояния k по наблюдаемым b и d. Основная идея состоит в следующем: если мы предполагаем, что рассматриваемые структуры эволюционировали по описанной случайной модели с параметром $\gamma = 2k/n$, то можно выписать теоретические выражения (через γ) для математического ожидания доли разрывов b/n и доли парсимонийного расстояния d/n. При больших n соответствующие отношения должны выполняться приблизительно и для конкретных

наблюдаемых значений b, d. Таким образом, зная b и d, мы можем восстановить (оценить) параметр γ, а затем вычислить k.

Согласно полученным ранее результатам, при больших n и известном γ математическое ожидание (и типичное значение) нормированного числа разрывов b/n равно

$$\frac{b}{n} \;\approx\; \frac{\gamma}{1+\gamma}\,,$$

поскольку $c_1/n \approx \frac{1}{1+\gamma}$ и $b/n = 1 - c_1/n$. Что касается доли парсимонийного расстояния, по определению $d = n - c$ (с учётом всех циклов). Выразим c через сумму по циклам различной длины:

$$\frac{c}{n} \;=\; \sum_{m=1}^{\infty} \frac{c_m}{n}\,,$$

где c_m – число циклов длины m (компонент размера m) в графе несоответствия. Теорема 1 даёт асимптотическое значение каждого слагаемого при известных γ. Поэтому можно записать:

$$\frac{d}{n} \;=\; 1 - \frac{c}{n} \;\approx\; 1 - \sum_{m=1}^{\infty} \frac{(3m-3)!\,\gamma^{\,m-1}}{m!\,(2m-1)!\,(\gamma+1)^{\,3m-2}}\,.$$

Правая часть представляет собой функцию от γ, которую мы обозначим $f(\gamma)$. Это выражение можно свести к более компактной форме, используя специальные функции. Вычитая из единицы полученный бесконечный ряд и преобразуя, авторы приходят к следующему представлению:

$$\frac{d}{n} \;=\; 1 - f(\gamma) \;=\; \frac{(1+\gamma)^2\Big({}_2F_1\big(-\tfrac{2}{3}, -\tfrac{1}{3}; \tfrac{1}{2}; \tfrac{27\gamma}{4(1+\gamma)^3}\big) - 1\Big)}{3\gamma}\,, \qquad (2)$$

где ${}_2F_1(a, b; c; z)$ – гипергеометрическая функция Гаусса. В данной формуле мы опустили знак $\approx$ ради краткости, подразумевая, что выражение справедливо в пределе $n \to \infty$. На практике же для больших, но конечных n формула служит хорошей аппроксимацией для математических ожиданий $E(d/n)$ и $E(b/n)$.

Функции $b/n = \frac{\gamma}{1+\gamma}$ и $d/n = f(\gamma)$ являются монотонно возрастающими функциями своего аргумента γ (что неудивительно: чем больше истинных перестроек, тем больше наблюдаемых разрывов и тем больше парсимонийное расстояние). Однако отношение

$$r(\gamma) \;=\; \frac{d/n}{b/n} \;=\; \frac{f(\gamma)}{\gamma/(1+\gamma)}$$

оказывается тоже монотонно растущим по γ и, кроме того, *не зависит от* n. Действительно, подставляя выражения, получим

$$r(\gamma) \;=\; \frac{f(\gamma)(1+\gamma)}{\gamma}\,,$$

где $f(\gamma)$ задано формулой (2) выше. При $\gamma \to 0$ нетрудно проверить, что $f(\gamma) \sim \gamma/2$ (для малых γ цикл длины 1 преобладает, и каждое событие либо не меняет d, либо уменьшает b на 1, так что $d \approx k = \gamma n/2$). Тогда $r(\gamma) \to 1$ при $\gamma \to 0$. С другой стороны, при γ большого порядка $r(\gamma)$ стремится к некоей константе < 1 (парсимонийное расстояние составляет лишь долю от всех разрывов). На рис. **??** в приложении показан вид зависимости $r(\gamma)$ для различных условий, подтверждающий её монотонность.

Таким образом, оценивание истинного расстояния сводится к следующей процедуре. Вычисляем для наблюдаемых данных величину

$$r_{\text{obs}} \;=\; \frac{d}{b}\,,$$

то есть отношение парсимонийного расстояния к числу разрывов. Поскольку $r_{\text{obs}} = r(\gamma)$ для некоторого (неизвестного) γ, а функция $r(\gamma)$ монотонна и гладка, уравнение $r(\gamma) = r_{\text{obs}}$ имеет единственное решение по γ. Обозначим это решение γ_e (оценка для истинного γ). Найти его можно численно, используя, например, метод половинного деления на интервале $\gamma \in [0, +\infty)$ (практически поиск можно ограничить разумным максимумом, например $\gamma = 10$, так как большие γ редко встречаются и дают значения r близкие к асимптотическому пределу). Получив оценку γ_e, далее рассчитываем оценку для n:

$$\hat{n} \;=\; \frac{b(1+\gamma_e)}{\gamma_e}\,.$$

Это равенство получается из соотношения $\frac{b}{n} = \frac{\gamma}{1+\gamma}$, решённого относительно n. Наконец, оцениваем k как

$$\hat{k} \;=\; \frac{\hat{n}\,\gamma_e}{2}\,.$$

Итоговая величина $\hat{k}$ и принимается за оценку истинного числа перестроек (эволюционного расстояния) между рассматриваемыми структурами.

Обсудим корректность и границы применимости описанного метода. Во-первых, заметим, что при $\gamma_e < 0.5$ решение уравнения $r(\gamma_e) = r_{\text{obs}}$ тривиально: в этом случае $r(\gamma) = 1$ для $0 \leqslant \gamma \leqslant 0.5$ (то есть $d = b$), и минимальное расстояние равно реальному. Это соответствует случаю умеренно эволюционировавших структур, когда парсимонийный метод не даёт ошибки. Наш метод автоматически даст $\gamma_e = r_{\text{obs}} = 1$, откуда $\hat{k} = d = b$, то есть мы вернёмся к парсимонии как частному случаю. При $\gamma_e > 0.5$ описанная процедура вводит статистическую поправку. Теоретически, при больших γ наш метод немного *недооценивает* k из-за допущения о пренебрежимо малом количестве “расщеплений” циклов. Действительно, при очень больших k (когда один и тот же блок мог участвовать более чем в одной перестройке) становится заметным эффект образования некоторых циклов заданной длины не только слиянием малых циклов, но и распадом больших циклов. Предложенная модель учитывает только процессы слияния, поэтому в

результатах при γ больших порядка присутствует небольшое систематическое отклонение: $\hat{k}$ получается чуть меньше, чем k в симуляциях. Однако практически этот эффект незначителен для разумных значений γ (меньше 5–10). В заключительном разделе мы рассмотрим, как можно эмпирически поправить данную погрешность.

Ещё одним ограничением метода является необходимость знания или оценки n (числа блоков) для применения формул. В ряде задач (например, геномных) n известно заранее. Если же n неизвестно, метод позволяет оценить n как промежуточную величину $\hat{n}$, что, впрочем, строго корректно только при выполнимости предположений модели.

2.4. Оценка точности метода на модельных данных

Разработанный метод был реализован алгоритмически и протестирован на симулированных данных. Генерировались пары структур (геномов) с различным числом перестроек $\gamma = 2k/n$ в диапазоне от 0.5 до 2.0. Для каждой фиксированной γ проводилось большое число независимых симуляций, и по каждой паре структур вычислялись значения b, d, а затем оценка $\hat{k}$ по описанному методу. После этого анализировалась относительная ошибка оценки, определяемая как

$$\frac{\hat{k} - k}{k},$$

то есть отклонение оценённого числа перестроек от фактического, нормированное на истинное значение. На рис. 9 представлен обобщённый график результатов (в виде “ящиков с усами”) для разных значений параметра γ. Каждый ящик охватывает центральные 50% результатов (межквартильный размах), а усы – 90% результатов. Видно, что медианная ошибка практически равна нулю (ящики сосредоточены около горизонтальной оси), а разброс ошибок относительно невелик. В частности, медианное отклонение не превышает 6% во всём диапазоне $\gamma \in [0.5,\ 2.0)$, а у 90% оценок отклонение не превышает 10%. Это указывает на высокую точность метода оценки в рамках рассматриваемой модели.

В таблице 1 численно приведены средние абсолютные ошибки (в процентах) для нескольких выбранных значений γ. Можно заметить, что ошибка слегка возрастает с увеличением γ, но остаётся в пределах 5% даже при γ вплоть до 2.0. Кроме того, при $\gamma < 0.5$ ошибка равна нулю (что соответствует режиму применимости парсимонии). Рост ошибки при больших γ объясняется упомянутым систематическим эффектом недоучёта расщепления циклов в нашей аналитической модели. В реальных симуляциях не все циклы образуются только слиянием мелких компонентов: некоторая доля появляется из-за дробления крупных компонентов, что методически приводит к занижению оценённого k.

Как было отмечено, наш аналитический метод имеет тенденцию незначительно занижать оценки k при больших γ. На рис. 9 это проявляется в том, что медианные значения ошибки (серии внутри ящиков) лежат чуть ниже нуля при

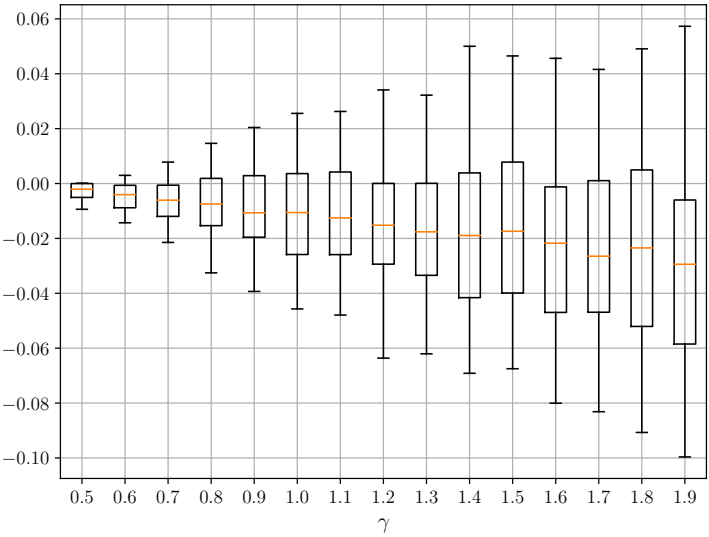


Рисунок 9 – Распределение относительной ошибки $\frac{\hat{k}-k}{k}$ в зависимости от параметра γ для разработанного метода оценки (без учёта поправки на расщепления). Ящики показывают границы 25% и 75% квантилей, усики – 5% и 95% квантилей.

Таблица 1 – Средний модуль ошибки оценки $\hat{k}$ в процентах для разных значений γ.

γ	Сред. ошибка	γ	Сред. ошибка
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	2.10%	1.9	4.49%
1.2	2.43%	2.0	4.87%

$\gamma > 1.0$, то есть $\hat{k}$ в среднем меньше настоящего k. Для компенсации этого эффекта можно ввести небольшую поправку. Эмпирически было найдено, что основная систематическая ошибка связана с недоучётом d/n на величину порядка $0.1/\sqrt{n}$ при $\gamma \geqslant 0.5$, тогда как при $\gamma < 0.5$ ошибка отсутствует (что согласуется с тем, что при малом γ расщепления действительно пренебрежимо редки). Поэтому можно скорректировать оценку, увеличивая d на $\frac{0.1}{\sqrt{n}}$ (для $\gamma \geqslant 0.5$) перед

вычислением отношения $r = d/b$. Эта поправка практически не влияет на случаи с большим n или умеренным γ, но для крайних режимов слегка повышает оценку k. Результаты работы метода с учётом данной эмпирической правки показаны на рис. 10. Видно, что систематический сдвиг устранён: медианы ошибок теперь проходят практически по нулевой линии даже при максимальных γ. Таким образом, скорректированный метод выдаёт асимптотически несмещённые оценки k во всём диапазоне параметров модели.

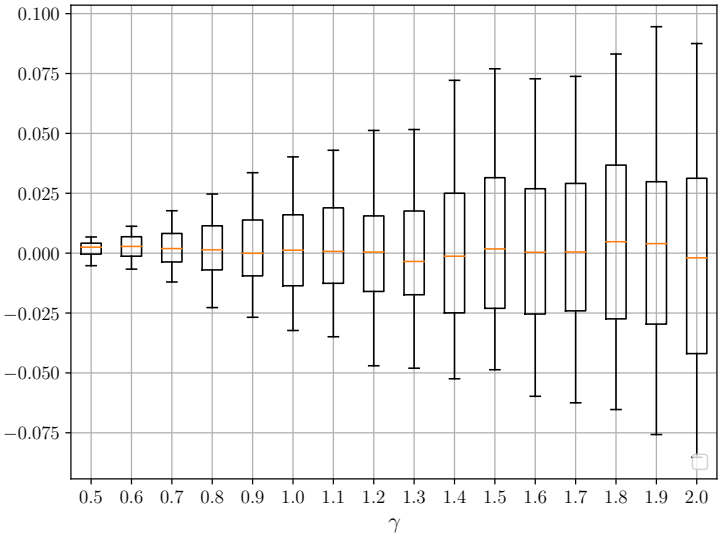


Рисунок 10 – Распределение относительной ошибки $\frac{\hat{k}-k}{k}$ после введения поправки на систематическое смещение (добавление $0.1/\sqrt{n}$ к d/n при $\gamma \geqslant 0.5$). Сравните со случаем без поправки на рис. 9.

Таким образом, на синтетических данных подтверждается высокая точность предлагаемого метода оценки истинного эволюционного расстояния. В частности, метод корректно работает в той области параметров ($\gamma > 0.5$), где классический парсимонийный подход систематически недооценивает расстояние. На модельных примерах получено, что даже для существенно эволюционировавших структур (например, $\gamma = 2.0$, что соответствует $k \approx n$) новая оценка $\hat{k}$ отличается от реального значения k менее чем на 5% в среднем, тогда как парсимонийная оценка d в таких случаях отличается на десятки процентов.

Следует подчеркнуть, что описанные эксперименты проводились на *модельных данных*, то есть на синтетических структурах, эволюционировавших строго в соответствии с предложенной случайной моделью (с заданными параметрами γ и распределением хрупкостей). Это необходимо для валидации метода и соответствия условий его вывода. В следующей главе будет рассмотрено

применение метода к реальным данным и оценено, насколько хорошо модель описывает реальные ситуации.

Выводы по главе 2

В данной главе были разработаны и математически обоснованы методы анализа случайных графов, моделирующих геномные перестройки, и методы оценки истинного эволюционного расстояния между структурами. Основные результаты, полученные в этой главе, можно суммировать следующим образом:

а) Разработан строгий комбинаторный подход к вычислению математического ожидания количества компонент заданного размера в графе несоответствия, основанный на биекции между сценариями объединения блоков и помеченными деревьями. Получены аналитические выражения для ожидаемого числа компонент произвольного размера.

б) Выполнен асимптотический анализ поведения компонент в графе несоответствия при увеличении числа элементов n и числа операций k. В частности, показано существование фазового перехода (при критическом значении $\gamma_c = 0.5$), после которого в системе возникает гигантская компонента, охватывающая существенную часть элементов структуры.

в) Полученные формулы позволили построить новый статистический метод оценки истинного эволюционного расстояния между двумя структурами на основе наблюдаемых характеристик графа несоответствия (числа разрывов и парсимонийного расстояния). Метод учитывает нелинейную связь между числом разрывов и реальным числом перестроек, характерную для больших эволюционных расстояний.

г) Проведена оценка точности предложенного метода на синтетических данных, моделирующих реалистичные сценарии эволюции геномных структур. Показано, что разработанный подход существенно превосходит парсимонийный метод в условиях большого числа перестроек (при $\gamma > 0.5$), давая ошибки оценки не более 5%, тогда как классические подходы существенно недооценивают реальное число перестроек.

д) Вспомогательные математические результаты (например, вычисления многомерных интегралов и использование кодов Прюфера для подсчёта вероятностей сценариев) имеют самостоятельное значение и могут быть использованы при решении других комбинаторных задач в области стохастического анализа графовых моделей.

Таким образом, результаты второй главы обеспечивают строгую математическую основу для количественной оценки эволюционных перестроек, подготавливая почву для применения и валидации этих методов на реальных биологических данных, чему посвящены последующие главы.

Глава 3. Алгоритмы детектирования и количественной оценки параллельных изменений

3.1. Алгоритмы предварительной обработки данных

3.1.1. Реконструкция древовидных структур состояний

3.1.2. Выделение линейных блоков консервативности

3.1.3. Построение признакового описания перестановок

3.2. Обнаружение параллельных событий

3.2.1. Оценка согласованности признаков с топологией древовидной структуры

3.2.2. Оценка степени параллельности и ранжирование событий

3.3. Кластеризация признаков по топологическим паттернам

3.4. Асимптотический анализ предлагаемых алгоритмов

Глава 4. Программная реализация разработанных методов и экспериментальная проверка

4.1. Описание программного пакета TruEst

4.1.1. Структура и модули

4.1.2. Интерфейс и применение на реальных данных

4.2. Описание программного пакета PaReBrick

4.2.1. Структура и модули

4.2.2. Интерфейс и визуализация результатов

4.3. Результаты экспериментальной проверки

4.3.1. Точность и скорость вычислений на модельных данных

4.3.2. Применение методов

Заключение

Список литературы

1. ***Zabelkin A.***, *Avdeyev P.*, *Alexeev N.* TruEst: a better estimator of evolutionary distance under the INFER model // Journal of Mathematical Biology. — 2023. — Июль. — Т. 87, № 2. — ISSN 1432-1416. — DOI: 10.1007/s00285-023-01955-z. — URL: http://dx.doi.org/10.1007/s00285-023-01955-z.
2. ***Zabelkin A.***, *Yakovleva Y.*, *Bochkareva O.*, *Alexeev N.* PaReBrick: PArallel REarrangements and BReaks identification toolkit // Bioinformatics / под ред. R. Schwartz. — 2021. — Окт. — Т. 38, № 2. — С. 357–363. — ISSN 1367-4811. — DOI: 10.1093/bioinformatics/btab691. — URL: http://dx.doi.org/10.1093/bioinformatics/btab691.
3. ***Zabelkin A.***, *Alexeev N.* Estimation of the True Evolutionary Distance Under the INFER Model // Comparative Genomics. — Springer International Publishing, 2018. — С. 72–87. — ISBN 9783030008345. — DOI: 10.1007/978-3-030-00834-5_4. — URL: http://dx.doi.org/10.1007/978-3-030-00834-5_4.
4. *Seferbekova Z.*, ***Zabelkin A.***, *Yakovleva Y.*, *Afasizhev R.*, *Dranenko N. O.*, *Alexeev N.*, *Gelfand M. S.*, *Bochkareva O. O.* High Rates of Genome Rearrangements and Pathogenicity of Shigella spp. // Frontiers in Microbiology. — 2021. — Апр. — Т. 12. — ISSN 1664-302X. — DOI: 10.3389/fmicb.2021.628622. — URL: http://dx.doi.org/10.3389/fmicb.2021.628622.
5. *Petukhova N.*, ***Zabelkin A.***, *Dravgelis V.*, *Aganezov S.*, *Alexeev N.* Chromothripsis Rearrangements Are Informed by 3D-Genome Organization // Comparative Genomics. — Springer International Publishing, 2022. — С. 221–231. — ISBN 9783031062209. — DOI: 10.1007/978-3-031-06220-9_13. — URL: http://dx.doi.org/10.1007/978-3-031-06220-9_13.
6. *Penny D.* Inferring Phylogenies.—Joseph Felsenstein. 2003. Sinauer Associates, Sunderland, Massachusetts. // Systematic Biology. — 2004. — Авг. — Т. 53, № 4. — С. 669–670. — ISSN 1063-5157. — DOI: 10 . 1080 / 10635150490468530. — URL: http://dx.doi.org/10.1080/10635150490468530.
7. *Bunke H.* On a relation between graph edit distance and maximum common subgraph // Pattern Recognition Letters. — 1997. — Авг. — Т. 18, № 8. — С. 689–694. — ISSN 0167-8655. — DOI: 10.1016/s0167-8655(97)00060-3. — URL: http://dx.doi.org/10.1016/S0167-8655(97)00060-3.
8. *Baret P.* Phylogenetic Analysis of Gregory of Nazianzus' Homily 27 // Le poids des mots: Actes des Journées d'étude. — 2004. — Exact conference details or editors not provided—please update if available.
9. *McCollum J.*, *Turnbull R.* Using Bayesian phylogenetics to infer manuscript transmission history // Digital Scholarship in the Humanities. — 2023. — Дек. — Т. 39, № 1. — С. 258–279. — ISSN 2055-768X. — DOI: 10 . 1093 / llc / fqad089. — URL: http://dx.doi.org/10.1093/llc/fqad089.

10. *Piñar G.*, *Tafer H.*, *Schreiner M.*, *Miklas H.*, *Sterflinger K.* Decoding the biological information contained in two ancient Slavonic parchment codices: an added historical value // Environmental Microbiology. — 2020. — Май. — Т. 22, № 8. — С. 3218–3233. — ISSN 1462-2920. — DOI: 10.1111/1462-2920.15064. — URL: http://dx.doi.org/10.1111/1462-2920.15064.

11. *Newman M. E. J.* The Structure and Function of Complex Networks // SIAM Review. — 2003. — Янв. — Т. 45, № 2. — С. 167–256. — ISSN 1095-7200. — DOI: 10.1137/s003614450342480. — URL: http://dx.doi.org/10.1137/S003614450342480.

12. *Pevzner P.*, *Tesler G.* Genome Rearrangements in Mammalian Evolution: Lessons from Human and Mouse Genomes // Genome Research. — 2003. — Т. 13, № 1. — С. 37–45.

13. *Erdős P.*, *Rényi A.* On Random Graphs // Publicationes Mathematicae. — 1959. — Т. 6. — С. 290–297.

14. *Rokas A.*, *Carroll S. B.* Frequent and Widespread Parallel Evolution of Protein Sequences // Molecular Biology and Evolution. — 2008. — Июнь. — Т. 25, № 9. — С. 1943–1953. — ISSN 1537-1719. — DOI: 10.1093/molbev/msn143. — URL: http://dx.doi.org/10.1093/molbev/msn143.

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

16. *Braga M. D.*, *Willing E.*, *Stoye J.* Genomic distance with DCJ and indels // International Workshop on Algorithms in Bioinformatics. — Springer. 2010. — С. 90–101.

17. *Barabási A.-L.*, *Oltvai Z. N.* Network biology: understanding the cell's functional organization // Nature Reviews Genetics. — 2004. — Февр. — Т. 5, № 2. — С. 101–113. — ISSN 1471-0064. — DOI: 10.1038/nrg1272. — URL: http://dx.doi.org/10.1038/nrg1272.

18. *Biller P.*, *Gueguen L.*, *Knibbe C.*, *Tannier E.* Breaking Good: Accounting for Fragility of Genomic Regions in Rearrangement Distance Estimation // Genome Biology and Evolution. — 2016. — Май. — Т. 8, № 5. — С. 1427–1439.

19. *Райгородский А.* Модели случайных графов и их применения // Труды Московского физико-технического института. — 2010. — Т. 2, № 4. — С. 130–140.

20. *Райгородский А.* Модели случайных графов. — Litres, 2022.

21. *Avdeyev P.*, *Jiang S.*, *Aganezov S.*, *Hu F.*, *Alekseyev M. A.* Reconstruction of Ancestral Genomes in Presence of Gene Gain and Loss // Journal of Computational Biology. — 2016. — Март. — Т. 23, № 3. — С. 150–164. — ISSN 1557-8666. — DOI: 10.1089/cmb.2015.0160. — URL: http://dx.doi.org/10.1089/cmb.2015.0160.

22. *Newman M. E. J.*, *Strogatz S. H.*, *Watts D. J.* Random graphs with arbitrary degree distributions and their applications // Physical Review E. — 2001. — Июль. — Т. 64, № 2. — ISSN 1095-3787. — DOI: 10.1103/physreve.64.026118. — URL: http://dx.doi.org/10.1103/PhysRevE.64.026118.

23. *Barabási A.-L.*, *Albert R.* Emergence of Scaling in Random Networks // Science. — 1999. — Окт. — Т. 286, № 5439. — С. 509–512. — ISSN 1095-9203. — DOI: 10.1126/science.286.5439.509. — URL: http://dx.doi.org/10.1126/science.286.5439.509.

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

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

26. *Lin Y.*, *Moret B. M.* Estimating true evolutionary distances under the DCJ model // Bioinformatics. — 2008. — Июль. — Т. 24, № 13. — С. i114–i122. — ISSN 1367-4803. — DOI: 10.1093/bioinformatics/btn148. — URL: http://dx.doi.org/10.1093/bioinformatics/btn148.

27. *Shelyakin P. V.*, *Bochkareva O. O.*, *Karan A. A.*, *Gelfand M. S.* Micro-evolution of three Streptococcus species: selection, antigenic variation, and horizontal gene inflow // BMC Evolutionary Biology. — 2019. — Март. — Т. 19, № 1. — ISSN 1471-2148. — DOI: 10.1186/s12862-019-1403-6. — URL: http://dx.doi.org/10.1186/s12862-019-1403-6.

28. *Irvine S.*, *Bunk B.*, *Bayes H. K.*, *Spröer C.*, *Connolly J. P. R.*, *Six A.*, *Evans T. J.*, *Roe A. J.*, *Overmann J.*, *Walker D.* Genomic and transcriptomic characterization of Pseudomonas aeruginosa small colony variants derived from a chronic infection model // Microbial Genomics. — 2019. — Апр. — Т. 5, № 4. — ISSN 2057-5858. — DOI: 10.1099/mgen.0.000262. — URL: http://dx.doi.org/10.1099/mgen.0.000262.

29. *Porubsky D.* [и др.]. Recurrent inversion polymorphisms in humans associate with genetic instability and genomic disorders // Cell. — 2022. — Май. — Т. 185, № 11. — 1986–2005.e26. — ISSN 0092-8674. — DOI: 10.1016/j.cell.2022.04.017. — URL: http://dx.doi.org/10.1016/j.cell.2022.04.017.

30. *Brandis G.*, *Hughes D.* The SNAP hypothesis: Chromosomal rearrangements could emerge from positive Selection during Niche Adaptation // PLOS Genetics / под ред. E. P. C. Rocha. — 2020. — Март. — Т. 16, № 3. — e1008615. — ISSN 1553-7404. — DOI: 10.1371/journal.pgen.1008615. — URL: http://dx.doi.org/10.1371/journal.pgen.1008615.

0. *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.

Список иллюстраций

Список таблиц