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:

а) 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.
б) 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.
в) 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;
- 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:

а) 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.

б) 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.

в) 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;
- 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

При сравнении двух геномов, подвергшихся перестройкам, удобным формализмом является граф точек разрыва (англ. *breakpoint graph*). Этот граф строится следующим образом. Сначала геномы представляют в виде последовательностей идентифицируемых фрагментов (блоков синтении) с учётом ориентации. Вершинами графа служат концы этих блоков (точки разрыва между блоками). Далее для каждого генома проводится соединение концов блоков, следующих друг за другом в данном геноме, при помощи рёбер (то есть каждое *соседство* двух блоков в геноме даёт ребро в графе). Таким образом, если сравниваются два генома P и Q, то в граф добавляются рёбра двух типов: рёбра типа P (отражающие соседства в первом геноме) и рёбра типа Q (соседства во втором геноме). В итоге получается 2-расцветочный мультиграф с $2n$ вершинами (если в обоих геномах n блоков) и, в общем случае, $2n$ рёбрами (по n рёбер каждого типа, если считать, что каждый геном однокрохромосомный без разрывов). Каждый узел этого графа имеет степень 2 (по одному ребру от каждого генома), поэтому компоненты графа образуют циклы различной длины.

Количество циклов в графе точек разрыва тесно связано с расстоянием между геномами в перестройках. В частности, для моделей без дупликаций справедливо, что чем больше циклов, тем ближе геномы. Например, в случае одной хромосомы без разрывов (то есть геномы представимы как перестановки), расстояние в операциях DCJ можно вычислить как $d_{\text{DCJ}} = n - c$, где n — число блоков, а c — количество циклов в объединённом графе. Интуитивно это означает: каждый цикл свидетельствует о том, что соответствующие элементы уже находятся на «правильных местах» друг относительно друга, и не требуют дополнительных операций для преобразования одного генома в другой; оставшиеся несовпадения требуют операций перестройки. Таким образом, граф точек разрыва служит ключевым инструментом для вычисления минимального числа необходимых операций, т.е. расстояния перестройки между двумя геномами.

Модель случайных перестроек генома часто предполагает, что разрывы в геноме происходят случайно. Если предположить, что каждый возможный разрыв (точка между геномными элементами) с равной вероятностью участвует в

каждой операции перестройки, то граф точек разрыва между геномами после k случайных операций будет обладать статистическими свойствами, близкими к свойствам случайного графа. Например, при достаточном числе перестроек можно ожидать, что большинство циклов будут короткими (2 или 4 вершины), и лишь немногие компоненты свяжут значительную часть вершин графа (аналогия с гигантской компонентой) — иначе говоря, исходный порядок блоков будет почти полностью "перемешан". Однако реальные эволюционные процессы могут отличаться от простого равномерного случайного разрыва: существуют области генома, более устойчивые к перестройкам, и наоборот, участки, ломкие и перестраивающиеся неоднократно. Это приводит к модификациям классического графа разрывов, в которые включаются *веса* или дополнительные параметры, отражающие неоднородность генома.

Однако стандартный граф точек разрыва не учитывает неоднородность генома: все разрывы считаются равновероятными. В реальных биологических данных некоторые области генома склонны к частым перестройкам (т. н. хрупкие регионы), тогда как другие остаются стабильными. Чтобы учесть эту особенность, вводится модифицированный граф точек разрыва с весами. В таком графе каждому ребру (точке разрыва) приписывается вес, отражающий вероятность или частоту взаимодействия именно в этом месте [18]. Весовая схема может быть основана на эмпирических данных или результатах предварительного статистического анализа геномных перестроек.

Другой модификацией является построение мульти-графа разрывов (англ. *breakpoint multigraph*) для множества геномов. Если имеется коллекция близкородственных геномов (например, виды или штаммы бактерий), можно обобщить граф разрывов на несколько геномов: вершины по-прежнему соответствуют концам блоков, а рёбра добавляются для *каждого* генома из коллекции, отражая соседства блоков в этом геноме. При этом каждое ребро помечается меткой, указывающей, какому геному (или листу филогенетического дерева) оно принадлежит. В результате между двумя вершинами может быть несколько рёбер (по одному от некоторых геномов) — так формируются консенсусные мульти-рёбра (англ. *consensus multi-edge*), отражающие тот факт, что несколько геномов имеют одинаковое соседство блоков. Такой мультиграф позволяет одновременно анализировать сходства и различия в упорядочении блоков у всех рассматриваемых видов Более того, он служит основой для выявления эволюционных событий: разрывы (отсутствие определённого соседства у части видов) и перестройки (например, наличие 4-цикла, свидетельствующего об инверсии в некоторых геномах). В следующем разделе будет показано, как на базе мультиграфа разрывов можно формализовать признаки (англ. *characters*) и идентифицировать параллельные изменения, сравнивая распределение этих признаков на филогенетическом дереве.

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.

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.

В случае модифицированного графа точек разрыва, в котором рёбра имеют веса, операция двойной разрез и слияние применяется аналогичным образом, однако вероятность выбора тех или иных рёбер теперь определяется их весами. Конкретно, пары рёбер выбираются с вероятностью, пропорциональной произведению их весов. После выполнения операции DCJ, веса вновь образованных рёбер пересчитываются согласно заданным правилам модели, таким образом отражая изменение состояния генома и перераспределение вероятностей будущих перестроек.

Использование весового подхода в операции двойной разрез и слияние позволяет более точно отражать реальные биологические процессы, в частности учитывая неоднородность геномов по склонности к перестройкам, что существенно повышает точность оценки эволюционного расстояния между структурами.

1.2.2. 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.

1.2.3. Вероятностная модель поломки случайных регионов

Одним из основных вероятностных подходов к моделированию эволюции генома является модель случайных разрывов (англ *Random Breakage Model, RBM*) [24]. 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.

1.2.4. 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.

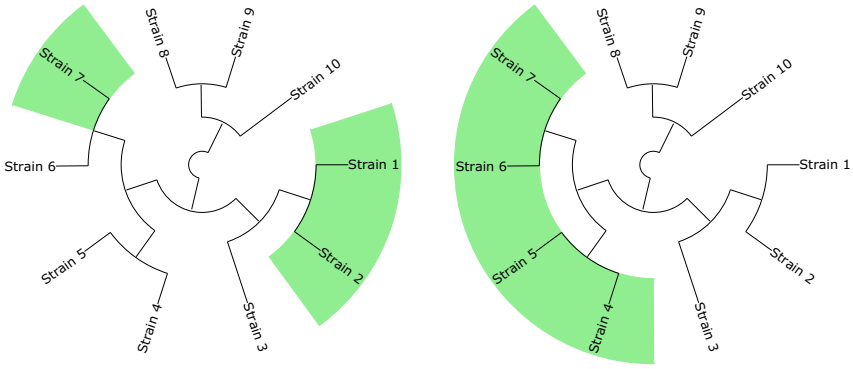
Таким образом, использование вероятностных моделей и соответствующих статистических оценивателей позволяет более надёжно измерять расстояния между геномными структурами, что важно для построения корректных филогенетических гипотез и понимания механизмов эволюции.

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 1б. 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 1a. 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 1 – 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 [26]. 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 [25]. 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 [25]. 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

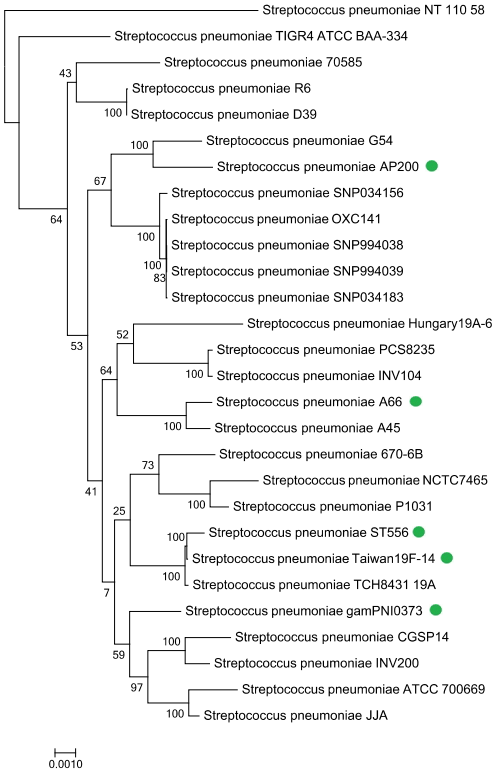


Рисунок 2 – Генетическое дерево демонстрирует распределение инверсий по генам PhtD и PhtB, штаммы с такими инверсиями выделены зелёным [25]

(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 [27] 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 [27].

Стоит отдельно упомянуть гипотезу сети аберрантных филогений, основанная на отборе (англ. *SNAP, Selection-driven Network of Aberrant Phylogenies*), предложенную для объяснения параллельных перестроек [28]. 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.

Глава 2. Комбинаторные и асимптотические методы анализа случайных графов и оценки расстояний

2.1. Математическое ожидание числа компонентов заданного размера

2.1.1. Подсчет числа компонентов начального размера

2.1.2. Подсчет числа компонентов произвольного размера

2.1.3. Асимптотический анализ и порог возникновения гигантской компоненты

2.2. Вспомогательные леммы

2.2.1. Редукция задачи к кодированию деревьев

2.2.2. Аналитический расчёт многомерных интегралов

2.3. Построение метода оценки истинного расстояния

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. *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.

25. *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.

26. *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.

27. *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.

28. *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.

Список иллюстраций

Список таблиц