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

1.1.2. Порог появления гигантской компоненты

1.1.3. Аффинные модификации модели случайных графов

1.1.4. Граф точек разрыва и его аффинные модификации

1.2. Постановка задачи оценки расстояний между структурами

1.2.1. Метрика минимального числа операций

1.2.2. Вероятностная модель поломки случайных регионов

1.2.3. Вероятностная модель поломки хрупких регионов

1.3. Методы анализа параллельных изменений в древовидных структурах

1.3.1. Снабжение графа точек разрыва весами

1.3.2. Выпуклые признаки на деревьях

1.3.3. Литературные примеры параллельных изменений

1.3.4. Задача оценки степени параллельности

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.

Список иллюстраций

Список таблиц