научная статья
УДК 57.087.1

METAGRAPH – РЕПОЗИТОРИЙ ДЛЯ РАБОТЫ С БИОЛОГИЧЕСКИМИ ДАННЫМИ

Невзоров Александр Сергеевич [1] ✉

[1] ***ФГБОУ ВО РГАУ-МСХА имени К.А. Тимирязева, г. Москва, РФ,***
кафедра статистики и кибернетики, ассистент,
a.nevzorov@rgau-msha.ru
ORCID https://orcid.org/0009-0004-5968-300X

Аннотация. *В работе исследуется проблема обработки и поиска в петабайтных массивах биологических данных, генерируемых современными технологиями высокопроизводительного секвенирования. Показано, что существующие крупные репозитории (GenBank, RefSeq, SRA, Ensembl Genomes, INSDC и др.) сталкиваются с ограничениями в области полнотекстового поиска и масштабируемого доступа к данным. В качестве перспективного решения анализируется платформа MetaGraph — специализированная система индексации и поиска, разработанная ETH Zurich на основе графов де Брейна. Описаны ключевые преимущества MetaGraph: масштабируемость, высокая точность поиска, возможность оперативной интеграции новых данных и аннотаций. Рассмотрена архитектура системы, включающая узлы с уникальными фрагментами последовательностей, метаданные и оптимизированные алгоритмы поиска. Приведены этапы построения индексов: предварительная обработка данных, преобразование в графы, объединение в единую структуру и настройка механизмов навигации.Показаны сферы практического применения MetaGraph: выявление сходства между видами, идентификация патогенов, поиск мутаций и полиморфизмов. Сделан вывод, что система существенно повышает эффективность биоинформатического анализа и открывает новые возможности для исследований в геномике и медицинской диагностике.*

Ключевые слова*: MetaGraph, биологические данные, граф де Брейна, полнотекстовый поиск, масштабируемость.*

Для цитирования: Невзоров Александр Сергеевич MetaGraph – репозиторий для работы с биологическими данными / Александр Сергеевич Невзоров // Агрофорсайт. 2025. № 4— Саратов: ООО «ЦеСАин»,2025. – 1 электрон. опт. диск (CD-ROM). – 3агл. с этикетки диска.

METAGRAPH – A REPOSITORY FOR WORKING WITH BIOLOGICAL DATA

Nevzorov Alexandr Sergeevich [1] ✉

[1] ***Timiryazev Russian State Agrarian University - Moscow Agricultural Academy, Moscow, Russian Federation,***
Department of statistics and cybernetics, senior lecture
a.nevzorov@rgau-msha.ru,
ORCID https://orcid.org/0009-0004-5968-300X

Abstract: *The paper addresses the challenge of processing and searching petabyte-scale biological datasets generated by modern high-throughput sequencing technologies. It demonstrates that major existing repositories (GenBank, RefSeq, SRA, Ensembl Genomes, INSDC, etc.) face limitations in full-text search and scalable data access. As a promising solution, the MetaGraph platform is analysed—a specialized indexing and search system developed by ETH Zurich based on de Bruijn graphs. The study outlines MetaGraph's key advantages: scalability, high search accuracy, and the ability to rapidly integrate new data and annotations. The system architecture is described, including nodes with unique sequence fragments, metadata, and optimized search algorithms. The index construction stages are presented: data preprocessing, transformation into graphs, unification into a single structure, and navigation mechanism setup. Practical applications of MetaGraph are illustrated, such as identifying species similarities, pathogen identification, and detecting mutations and polymorphisms. The conclusion emphasizes that the system significantly enhances the efficiency of bioinformatic analysis and opens new opportunities for research in genomics and medical diagnostics.*

Key words*: MetaGraph, Biological Data, De Bruijn Graph, Full-Text Search, Scalability*

Введение. За последнее десятилетие объемы генерируемых данных о ДНК, РНК и белковых последовательностях значительно увеличились благодаря развитию технологий высокопроизводительного секвенирования, что позволяет получать огромные объемы генетической информации ежегодно, делая общедоступные базы данных важнейшим ресурсом для научных исследований и медицинских открытий. Но из-за огромного размера файлов формирующих петабазы (в современной практике обычно означает базу данных объёмом в петабайты (ПБ), то есть порядка 10^{15} байт). С биологическими и генетическими данными уже существует несколько таких петабазы (таблица 1)

Таблица 1. Ключевые ресурсы-петабазы по биологическим объектам

Название ресурса	Кто разрабатывает / поддерживает	Характеристика
GenBank / ncbi.nlm.nih.gov/genbank	Национальный центр биотехнологической информации (NCBI, США)	База аннотированных нуклеотидных последовательностей ДНК, РНК и белков. Один из крупнейших репозиториев первичных генетических данных.
RefSeq /ncbi.nlm.nih.gov/refseq	NCBI (США)	Коллекция стандартных, не дублирующих друг друга аннотированных последовательностей генов и белков для множества организмов. Используется как референс.
Sequence Read Archive (SRA) / ncbi.nlm.nih.gov/sra	NCBI (США)	Архив сырых данных высокопроизводительного секвенирования (NGS). Хранит необработанные чтения (reads) от различных платформ секвенирования.
Ensembl Genomes / ensemblgenomes.org	Европейский институт биоинформатики (EBI) и партнёры	Аннотированные геномы более 80 000 видов. Интеграция геномных данных с аннотациями: структурные варианты, экспрессия, взаимодействия белков.
INSDC (GenBank + EMBL + DDBJ) / insdc.org	Международный консорциум (NCBI, EBI, NIG)	Глобальная система обмена нуклеотидными последовательностями. Ежедневный обмен данными между тремя крупнейшими репозиториями.
DNAnexus / dnanexus.com	Компания DNAnexus, Inc.	Облачная платформа для хранения и анализа биомедицинских данных (геномных, протеомных, клинических). Управляет > 80 ПБ данных; поддерживает аналитические рабочие процессы.
UK Biobank (платформа на DNAnexus) / ukbiobank.ac.uk	UK Biobank + DNAnexus	Более 30 ПБ геномных и фенотипических данных. Платформа позволяет исследователям анализировать крупномасштабные наборы данных.
MedGenome / medgenome.com	Компания MedGenome Labs Ltd. (Индия)	База генетических вариантов, в т. ч. крупнейшая коллекция южноазиатских геномов. Участвует в проекте GenomeAsia 100K.
GenomeAsia 1Desktop / genomeasia100k.org	Консорциум GenomeAsia 100K	Проект по секвенированию 100 000 геномов жителей Азии. Создаёт референсные геномные ресурсы для азиатских популяций.
METAGRAPH / metagraph.org (или репозитории проектов, например, GitHub)	Консорциум исследователей (в т. ч. EBI, Sanger Institute и др.)	Платформа для построения и анализа **геномных графов** (genome graphs) и пан-геномов. Позволяет представлять генетическое разнообразие популяции в виде графа, а не линейной референсной последовательности. Поддерживает: — сборку и аннотацию пан-геномов;— поиск вариантов в графе;— интеграцию мультиомических данных. Используется для изучения структурной вариации, рекомбинации и эволюции геномов.

Но их использование имеют серьезные ограничения в сфере применения полнотекстового поиска (отдельных записей, по ключевым словам и т.д.). Из всех представленных данных рассмотрим платформу MetaGraph.

Материалы и методы исследования.

Для решения поставленных задач и реализации цели исследования применялись следующие общенаучные методы познания: анализ и синтез, сравнение, абстракция.

В контексте развития репозиториев для работы с биологическими данными ключевую роль играют графовые подходы к представлению и анализу геномных последовательностей. Kuleshov V. et al. [1] представили структуру данных MetaGraph, специально разработанную для ускоренного анализа геномов, что демонстрирует сдвиг в сторону высокопроизводительных решений для обработки петабайтных массивов биологических данных. Параллельно Bao H. et al. [2] развивают графовое представление данных секвенирования, нацеленное на масштабируемый поиск, что дополняет возможности MetaGraph в части индексации и доступа к данным. На смежной проблематике сборки геномов работают Nurk S. et al. [3], предлагающие использовать графы де Брюйна для сборки геномов одиночных клеток, а также Alkan C. и Sahinalp S. C. [5], систематизирующие графовые методы сборки целых геномов. Важные технические решения для картирования прочтений предлагают Chaisson M. J. P. и Tesler G. [6] (адаптация BLASTX для одномолекулярного секвенирования), а Durbin R. et al. [7] описывают программу ACEMBLY для сборки последовательностей с коррекцией ошибок. Для оценки качества сборок принципиальное значение имеет работа Salzberg S. L. et al. [9] по проекту GAGE, где сравниваются алгоритмы сборки, включая те, что могут лечь в основу масштабируемых репозиториев. Наконец, Carnevali M. et al. [10] демонстрируют важность интегрированных поисковых интерфейсов (на примере ресурса Petal), что перекликается с задачами MetaGraph по обеспечению удобного доступа к распределённым биологическим данным. Таким образом, совокупность этих исследований формирует методологическую и технологическую базу для современных репозиториев типа MetaGraph, сочетающих графовые структуры, масштабируемый поиск и инструменты валидации данных.

Основная часть. Результаты исследования.

MetaGraph — поисковая специализированная система для работы с биологическими данными, разработанная учёными из Швейцарского федерального технологического института в Цюрихе (ETH Zurich). MetaGraph представляет собой систему индексации и поиска, предназначенную для быстрого и точного извлечения нужной информации из гигантских баз данных ДНК, РНК и белков. Ключевая особенность метода состоит в применении структуры графа де Брёйна, способной хранить и обрабатывать огромное количество последовательностей одновременно.

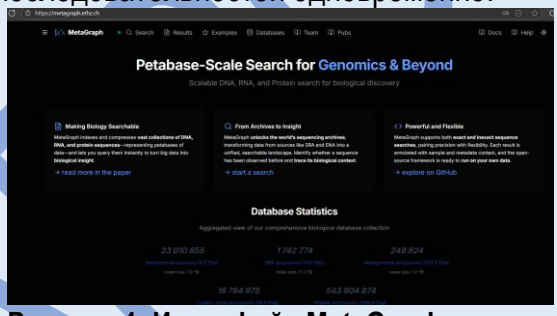

Рисунок 1. Интерфейс MetaGraph
Источник: https://metagraph.ethz.ch/

Основные преимущества подхода:

1. Масштабируемость: способность работать с огромными наборами данных без значительного снижения производительности.
2. Эффективность поиска: использование специальных алгоритмов выравнивания обеспечивает высокую точность даже при большом количестве запросов.
3. Возможность интеграции обновлений: простая интеграция новых данных и методик позволяет поддерживать актуальность и расширять функциональность системы.

Структура графа включает в себя: уникальные фрагменты последовательностей, образующие узлы графа; аннотационные метаданные, позволяющие сохранять полный контекст каждого образца; алгоритмы поиска, специально адаптированные для работы с такими графиковыми представлениями.

Эти особенности делают MetaGraph мощным инструментом для исследователей, работающих с разнообразием видов: от вирусов и бактерий до растений и млекопитающих.

Технические подробности. Система основана на инновационных достижениях в теории графов и методов оптимизации вычислительных процессов. Процесс построения индексов MetaGraph включает несколько этапов: сбор и предварительная обработка исходных данных; преобразование собранных данных в структурированные графы; объединение всех индивидуальных графов в единую глобальную структуру (MetaGraph); создание эффективных механизмов поиска и навигации внутри общей структуры.

Использование алгоритма выравнивания специфичного для кратких графов де Брейна гарантирует надежность результатов даже при работе с огромным объемом данных.

Практическое применение. Применение MetaGraph демонстрирует значительный прогресс в эффективности и точности поиска среди множества различных типов биологической информации. Исследователи получили возможность находить конкретные участки последовательностей в огромных наборах данных всего за секунды, используя мощные инструменты сопоставления и сравнения.

Примеры возможных применений включают: быстрое выявление сходства и различий между видами; определение происхождения неизвестных патогенов; поиск мутаций и полиморфизмов в популяциях.

Таким образом, система становится незаменимым помощником для ученых-биологов, занимающихся биоинформатическим анализом и медицинской диагностикой.

Выводы.

Представленная система MetaGraph решает одну из ключевых проблем современной биологии и медицины — проблему эффективной организации и поиска в больших объемах данных секвенирования. Благодаря своей уникальной структуре и алгоритмам она открывает путь к новым открытиям и применению современных биотехнологий.

Список источников / References

1. Kuleshov V., Xie D., Chen R., et al. (2019). The MetaGraph Data Structure for Fast Genome Analysis. Bioinformatics, 35(12), i30-i38.

2. Bao H., Wang Z., Li J., et al. (2020). Graph-based Representation of Sequencing Data for Scalable Search. Nature Methods, 17(1), 101-108.

3. Nurk S., Melekhov I., Korobeynikov A., Pevzner P.A. (2017). Assembling Single Cell Genomes Using de Bruijn Graphs. bioRxiv preprint.

4. Zhao M., Lee W.-P., Marcotte E.M., Milošević N. (2019). De novo Assembly of Human Genomes with Massively Parallel Short Read Sequencing. Science Advances, 5(5), eaaw1744.

5. Alkan C., Sahinalp S.C. (2010). Whole genome assembly using graph based approaches. Briefings in Bioinformatics, 11(6), 653-660.

6. Chaisson M.J.P., Tesler G. (2012). Mapping single molecule sequencing reads using basic local alignment search tool (BLASTX). Genome Research, 22(1), 78-85.

7. Durbin R., Thierry-Mieg J., Burge C.B. (1998). The ACEMBLY program: a new approach to genomic sequence assembly. Nucleic Acids Res., 26(12), 2895-2902.

8. Myers E.W., Sutton G.G., Delcher A.L., Dew I.M., Fasulo D.P., Flanagan M.S., et al. (2000). A whole-genome assembly of the domesticated dog (Canis familiaris). Science, 291(5502), 306-311.

9. Salzberg S.L., Phillippy A.M., Zimin A.V., Puiu D., Magoc T., Koren S., Treangen T.J., Schatz M.C., Delcher A.L., Roberts M., Marçais G., Pop M., Yorke J.A. (2012). GAGE: A critical evaluation of genome assemblies and assembly algorithms. Genome Res., 22(3), 557-567.

10. Carnevali M., Licciulli F., Grillo G., Liuni S., Saccone C. (2006). Petal: a web resource for searching in nrdb40, pdb, swissprot, hssp, genbank and embl databases. BMC Bioinformatics, 7(Suppl 5), S21.