Обзорная статьи
УДК 004.04

METAGRAPH – РЕПОЗИТОРИЙ ДЛЯ РАБОТЫ С БИОЛОГИЧЕСКИМИ ДАННЫМИ

Невзоров Александр Сергеевич [1]

[1] ***ФГБОУ ВО РГАУ-МСХА имени К.А. Тимирязева**, **127550, г. Москва, ул. Тимирязевская, д. 49**, **a.nevzorov@rgau-msha.ru, 0009-0004-5968-300X***

Аннотация. *методы молекулярного секвенирования позволили достичь существенного прогресса в исследовании живых организмов, однако сложность работы с большими объемами полученных данных создает значительные проблемы. Данная статья посвящена представлению системы MetaGraph, позволяющей эффективно организовать хранение и поиск в огромных массивах генетической информации.*

***Ключевые слова**: MetaGraph, Биологические данные, Граф де Брейна, Полнотекстовый поиск, Масштабируемость.*

Для цитирования: Для цитирования: Невзоров Александр Сергеевич MetaGraph – репозиторий для работы с биологическими данными/ Невзоров Александр Сергеевич // Агрофорсайт. 2025. № 5— Саратов: ООО «ЦеСАин»,2025. – 1 электрон. опт. диск (CD-ROM). – Загл. с этикетки диска.

METAGRAPH – A REPOSITORY FOR WORKING WITH BIOLOGICAL DATA

Nevzorov Alexandr Sergeevich [1]

[1] ***Russian State Agrarian University - Moscow Timiryazev Agricultural Academy, 127550, Moscow, Timiryazevskaya St., 49, a.nevzorov@rgau-msha.ru, 0009-0004-5968-300X***

Abstract: *molecular sequencing methods have enabled significant progress in the study of living organisms, but the complexity of working with large volumes of data creates significant challenges. This article presents the MetaGraph system, which enables efficient storage and search of vast amounts of genetic information***.**

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

Введение.

За последнее десятилетие объемы генерируемых данных о ДНК, РНК и белковых последовательностях значительно увеличились благодаря развитию технологий высокопроизводительного секвенирования. Эти технологии позволяют получать огромные объемы генетической информации ежегодно, делая общедоступные базы данных важнейшим ресурсом для научных исследований и медицинских открытий.

Однако существующие методы хранения и поиска данных имеют серьезные ограничения. Например, традиционный подход основан на поиске отдельных записей, по ключевым словам, или метаданным, после чего полученные записи загружаются

локально для дальнейшего анализа. Это создает дополнительные трудности, особенно учитывая огромный размер файлов (петабазы). Методы полнотекстового поиска остаются практически неприменимы для столь больших объемов данных.

Материалы и методы исследования.

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

Основная часть. Результаты исследования.

Огромные объемы данных требуют принципиально новых подходов к поиску и обработке, иначе большая часть информации останется неиспользованной.

Структура MetaGraph.

MetaGraph представляет собой специализированную 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 Брейна, способной хранить и обрабатывать огромное количество последовательностей одновременно.

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

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.

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.

Информация об авторе (авторах)

А.С. Невзоров – ***старший преподаватель кафедры статистики и кибернетики института экономики и управления АПК***

Information about the author

A.S. Nevzorov – ***Senior Lecturer, Department of Statistics and Cybernetics, Institute of Economics and Management of the Agro-Industrial Complex.***