УДК 004.4

МЕТОДЫ ПРОГРАММИРОВАНИЯ И ОБРАБОТКИ БОЛЬШИХ ДАННЫХ В БИОИНФОРМАТИКЕ

Домбровский Я.А.

старший преподаватель

Калужский государственный университет им. К.Э. Циолковского,

Калуга, Россия

Салдаева А.А.

магистрант,

Калужский государственный университет им. К.Э. Циолковского,

Калуга, Россия

Аннотация.

Биоинформатика представляет собой междисциплинарную область, объединяющую биологию, информатику и статистику для анализа и интерпретации биологических данных. В последние годы наблюдается стремительный рост объемов биологической информации благодаря развитию технологий секвенирования нового поколения (NGS), масс-спектрометрии и других методов молекулярной биологии. Обработка и анализ таких больших данных требуют разработки новых подходов и инструментов, основанных на современных методах программирования и вычислительных технологиях. В данной статье рассматриваются основные методы программирования и обработки больших данных, применяемые в биоинформатике, включая параллельные вычисления, машинное обучение и облачные технологии. Приводятся примеры успешных проектов и исследований, демонстрирующих эффективность этих методов в решении различных задач биоинформатики.

Ключевые слова: Биоинформатика, большие данные, методы программирования, параллельные вычисления, машинное обучение, нейронные

сети, облачные технологии, выравнивание последовательностей, анализ экспрессии генов, варианты последовательности.

METHODS OF PROGRAMMING AND PROCESSING BIG DATA IN BIOINFORMATICS

Dombrovsky Y.A.
Senior Lecturer
Kaluga State University named after K.E. Tsiolkovsky,
Kaluga, Russia

Saldaeva A.A.
Graduate student,
Kaluga State University named after K.E. Tsiolkovsky,
Kaluga, Russia

Annotation.
Bioinformatics is an interdisciplinary field that combines biology, computer science, and statistics to analyze and interpret biological data. In recent years, there has been a rapid increase in the volume of biological information due to the development of new generation sequencing technologies (NGS), mass spectrometry and other methods of molecular biology. Processing and analyzing such big data requires the development of new approaches and tools based on modern programming methods and computing technologies. This article discusses the main methods of programming and processing big data used in bioinformatics, including parallel computing, machine learning and cloud technologies. Examples of successful projects and studies demonstrating the effectiveness of these methods in solving various bioinformatics problems are given.
Keywords: Bioinformatics, big data, programming methods, parallel computing, machine learning

Биоинформатика является одной из наиболее быстро развивающихся областей науки, играющей ключевую роль в современной медицине, фармакологии и биотехнологиях. С развитием технологий высокопроизводительного секвенирования (High Throughput Sequencing, HTS) объемы генерируемых биологических данных стремительно увеличиваются. Эти данные включают последовательности ДНК/РНК, протеомы, метаболомы и другие типы информации, необходимые для понимания механизмов функционирования живых организмов. Однако, чтобы извлечь полезные знания из этих огромных массивов данных, требуются мощные вычислительные ресурсы и эффективные алгоритмы обработки информации [6].

Современные методы программирования играют важную роль в биоинформатике, обеспечивая возможность автоматизации сложных процессов анализа и визуализации данных. Развитие параллельных вычислений, использование кластеров и суперкомпьютеров позволяет значительно ускорить выполнение ресурсоемких задач, таких как выравнивание последовательностей, моделирование белков и анализ экспрессии генов [1]. Кроме того, применение методов машинного обучения открывает новые перспективы для предсказания структур белков, классификации паттернов экспрессии генов и выявления генетической предрасположенности к заболеваниям.

История развития биоинформатики тесно связана с эволюцией компьютерных наук и информационных технологий. Ниже представлены ключевые этапы развития информационных технологий в биоинформатике (таблица 1).

Таблица 1 – Ключевые этапы развития информационных технологий в биоинформатике

Период	Основные события и достижения
1950-е	Начало использования компьютеров для анализа биологических данных. Формализация алгоритмов сравнения последовательностей.
1960-е	Появление первых программ для выравнивания последовательностей. Создание баз данных нуклеотидных и аминокислотных последовательностей.
1970-е	Разработка метода быстрого преобразования Фурье (FFT). Ускоренное сравнение последовательностей.
1980-е	Создание первого суперкомпьютера Cray. Разработка алгоритмов динамического программирования.
1990-е	Запуск проекта «Геном человека». Появление высокопроизводительных серверов и кластеров.
2000-е	Развитие технологий секвенирования нового поколения (NGS). Широкое внедрение параллельных вычислений.
2010-е	Расцвет облачных технологий и машинного обучения. Использование глубоких нейронных сетей для анализа биологических данных.
2020-е	Интеграция искусственного интеллекта в биоинформатические исследования. Повышение роли квантовых вычислений.

Целью настоящей работы является обзор основных методов программирования и обработки больших данных, применяемых в биоинформатике, а также демонстрация их эффективности на примере конкретных научных исследований. Мы рассмотрим подходы к организации параллельных вычислений, особенности использования облачных платформ, а также применение методов машинного обучения для решения актуальных задач биоинформатики.

Рассмотрим основные методы программирования и обработки больших данных в биоинформатике.

Параллельные вычисления стали неотъемлемой частью современных вычислительных систем, особенно в контексте обработки больших объемов данных. Применение параллельных алгоритмов позволяет существенно сократить время выполнения трудоемких операций, таких как выравнивание последовательностей нуклеиновых кислот, аминокислот или белков, построение филогенетических деревьев и анализ больших баз данных (рисунок 1).

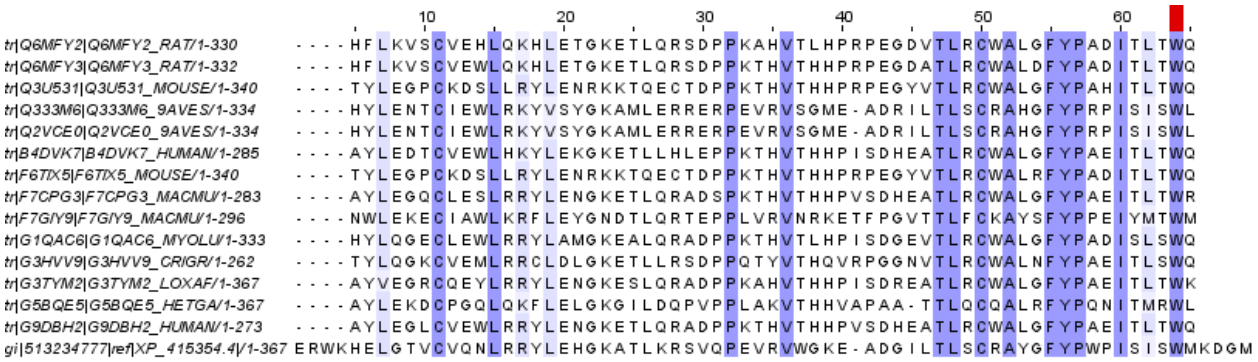

Рис. 1 – Пример представления множественного выравнивания белковых последовательностей (составлено авторами)

Message Passing Interface (MPI) и Open Multi-Processing (OpenMP) являются двумя основными стандартами для организации параллельных вычислений. MPI используется для распределенных вычислений между несколькими процессорами или узлами сети, тогда как OpenMP предназначен для многопоточной обработки внутри одного узла. Оба подхода находят широкое применение в биоинформатике. Например, программы для выравнивания последовательностей, такие как BLAST, были адаптированы для работы в параллельном режиме с использованием MPI, что позволило значительно увеличить скорость поиска гомологичных последовательностей среди миллионов записей в базах данных (рисунок 2).

```
hemoglobin subunit beta [Papio anubis]
Length=147

 GENE ID: 100137310 HBB | hemoglobin, beta [Papio anubis]
(10 or fewer PubMed links)

 Score =  114 bits (284),  Expect = 4e-24, Method: Compositional matrix adjust.
 Identities = 63/145 (44%), Positives = 86/145 (60%), Gaps = 8/145 (5%)

Query  3    LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-DLSH-----GSAQV  56
            L+P +K  V A WGKV  +  E G EAL R+ + +P T+ +F  F DLS       G+ +V
Sbjct  4    LTPEEKNAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFDSFGDLSSPAAVMGNPKV  61

Query  57   KGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPA  116
            K HGKKV  A ++ + H+D++    + LS+LH  KL VDP NFKLL + L+  LA H
Sbjct  62   KAHGKKVLGAFSDGLNHLDNLKGTFAQLSELHCDKLHVDPENFKLLGNVLVCVLAHHFGK  121

Query  117  EFTPAVHASLDKFLASVSTVLTSKY  141
            EFTP V A+  K +A V+  L  KY
Sbjct  122  EFTPQVQAAYQKVVAGVANALAHKY  146
```

Рис. 2 – Пример парного выравнивания последовательностей аминокислот, выполненного программой BLAST (составлено авторами)

MapReduce – это парадигма программирования, разработанная Google для обработки больших наборов данных на кластерах компьютеров. Эта модель была реализована в виде платформы Apache Hadoop, которая стала де-факто стандартом для управления большими данными в распределенной среде. В биоинформатике Hadoop применяется для задач, связанных с обработкой и анализом NGS-данных, таких как сборка генома, определение вариаций последовательности и аннотирование генов.

Машинное обучение становится одним из ключевых инструментов в биоинформатике благодаря своей способности выявлять скрытые закономерности в данных и делать прогнозы на основе ранее накопленных знаний. Современные модели глубокого обучения, такие как сверточные нейронные сети (CNN) и рекуррентные нейронные сети (RNN), успешно применяются для анализа изображений, предсказания структуры белков и идентификации регуляторных элементов в геноме.

Глубокие нейронные сети демонстрируют высокую эффективность в анализе изображений, распознавании речи и обработке естественного языка. В биоинформатике CNN используются для анализа микроскопических изображений клеток и тканей, а также для определения локализации белков в клетке. RNN находят применение в задачах, связанных с анализом временных рядов, например, при изучении динамики экспрессии генов [4].

Методы классификации и регрессии широко используются в биоинформатике для предсказания функциональных свойств белков, классификации типов раковых опухолей и выявления мутаций, ассоциированных с заболеваниями. Примером успешного применения таких методов является проект The Cancer Genome Atlas (TCGA), где использовались алгоритмы машинного обучения для анализа данных экспрессии генов и выявления специфических маркеров рака [3].

Облачные платформы предоставляют доступ к мощным вычислительным ресурсам и инструментам для хранения и обработки больших данных. Это особенно важно для биоинформатиков, работающих с огромными объемами данных, такими как геномы целых популяций или транскриптомы отдельных клеток.

Amazon Web Services предлагает широкий спектр сервисов для работы с большими данными, включая хранилища данных (S3), базы данных (DynamoDB) и инструменты для аналитики (EMR). В биоинформатике AWS активно используется для запуска вычислительно интенсивных приложений, таких как сборка геномов и анализ экспрессии генов.

Google Cloud Platform предоставляет аналогичные сервисы, включая BigQuery для анализа больших данных и Compute Engine для виртуальных машин. Одним из примеров успешного применения GCP в биоинформатике является проект DeepVariant от Google AI, который использует глубокое обучение для точного вызова вариантов последовательности из данных NGS [2].

Примеры успешных проектов в области биоинформатики наглядно демонстрируют, как современные методы программирования и обработки больших данных способствуют решению важных научных проблем. Рассмотрим несколько таких проектов подробнее.

Проект «Геном человека» (Human Genome Project) завершённый в 2003 году, имел целью расшифровку полной последовательности человеческого генома. Для его реализации потребовались значительные вычислительные мощности и специализированные алгоритмы. Одной из главных задач стало выравнивание миллиардов коротких фрагментов ДНК, полученных в результате секвенирования. В процессе проекта были разработаны новые методы параллельного программирования, позволяющие эффективно использовать суперкомпьютерные кластеры. Программа BLAST, одна из важнейших в биоинформатике, была оптимизирована для работы в параллельном режиме, что

значительно ускорило поиск гомологичных последовательностей среди миллионов записей в базах данных [5].

The Cancer Genome Atlas (TCGA) – это международный проект, направленный на всесторонний анализ генетического материала различных видов рака. Основная цель проекта — определить мутации, гены и пути, участвующие в развитии онкологических заболеваний. В рамках TCGA были собраны огромные объёмы данных, включающих полногеномные профили секвенирования, экспрессию генов, метилирование ДНК и другие характеристики опухолевых клеток. Для анализа этих данных использовались методы машинного обучения, включая глубокие нейронные сети, что позволило идентифицировать новые биомаркеры и потенциальные цели для таргетированной терапии.

DeepVariant от Google AI – это инструмент для высокоточного вызова вариантов последовательности (SNP и InDel) из данных секвенирования следующего поколения (NGS). Проект основан на применении глубоких нейронных сетей, которые обрабатывают изображения, полученные из необработанных данных секвенирования. Использование графических процессоров (GPU) для ускорения вычислений позволило добиться высокой точности и скорости работы. DeepVariant стал важным инструментом для исследователей, занимающихся геномикой и персонализированной медициной.

Metagenomics Analysis with Galaxy – это веб-платформа для биоинформатического анализа, предоставляющая исследователям удобный интерфейс для выполнения сложных вычислительных задач. Один из ярких примеров её применения – анализ метагеномных данных. Метагеномика изучает совокупность всех геномов микроорганизмов, присутствующих в определённой экосистеме. Для обработки метагеномных данных требуются большие вычислительные ресурсы и специализированные алгоритмы. Galaxy предоставляет удобные инструменты для сборки геномов, аннотации генов и

таксономического анализа, позволяя учёным проводить комплексный анализ метагеномов.

Европейская инфраструктура для биологических данных (ELIXIR) – это европейский консорциум, созданный для координации усилий по управлению и использованию биологических данных. Проект направлен на создание единой инфраструктуры для обмена, хранения и анализа биологических данных. ELIXIR объединяет усилия исследовательских институтов, университетов и компаний, предлагая доступ к мощнейшим вычислительным ресурсам и специализированному ПО. Важнейшими направлениями деятельности ELIXIR являются разработка стандартов для представления данных, обеспечение совместимости разных баз данных и поддержка исследовательских проектов в области биоинформатики.

Современная биоинформатика требует применения инновационных методов программирования и обработки больших данных для эффективного анализа и интерпретации биологических данных. Параллельные вычисления, машинное обучение и облачные технологии открывают новые возможности для решения сложнейших задач, стоящих перед учеными-биоинформатиками. Успех таких проектов, как Human Genome Project, TCGA и ICGC, демонстрирует важность интеграции этих подходов в научные исследования. Дальнейшее развитие методов программирования и вычислительной инфраструктуры позволит ускорить прогресс в области медицины, фармакологии и биотехнологий.

Библиографический список:

1. Аксютина, Е. М. Анализ оптимизированного алгоритма выравнивания биологических последовательностей / Е. М. Аксютина, Ю. С. Белов // Электронный журнал: наука, техника и образование. – 2017. – № 2. – С. 100–106.

2. Исаев, Е. А. Проблема обработки и хранения больших объемов научных данных и подходы к ее решению / Е. А. Исаев, В. В. Корнилов // Математическая биология и биоинформатика. – 2013. – Т. 8. – № 1. – С. 49–65.

3. Никонорова, М. Л. Методы машинного обучения в биоинформатике / М. Л. Никонорова // Региональная информатика и информационная безопасность. – 2022. – С. 411–414.

4. Свешникова, А. Н. Экспрессия генов и микрочипы: проблемы количественного анализа / А. Н. Свешникова, П. С. Иванов // Российский химический журнал. – 2007. – Т. 51. – № 1. – С. 127–135.

5. Спринджук, М. В. Обработка и визуализация данных, полученных с ДНК-матриц / М. В. Спринджук и др. // Инновационные технологии в медицине. – 2015. – № 2–3. – С. 98–110.

6. Min, S. Deep learning in bioinformatics / S. Min, B. Lee, S. Yoon // Briefings in bioinformatics. – 2017. – Vol. 18. – No. 5. – Pp. 851–869.

Оригинальность 80%