Материалы международной научно-практической конференции,посвященной 130-летию С. Сейфуллина = С. Сейфуллиннің 130 жылдығына арналған халықаралық ғылыми - практикалық конференциясының материалдары. - 2024. – Ч.II.- С. 37-41.

УДК 004.4

СОВРЕМЕННЫЕ МЕТОДЫ И ИНСТРУМЕНТАЛЬНЫЕ СРЕДСТВА В ОБРАБОТКЕ ГЕНОМНЫХ ДАННЫХ

Байжанов Е.И., 1 курс магистрант
Казахский агротехнический исследовательский университет им. С.Сейфуллина, г. Астана

Комиссаров А.С. в своих научных трудах изучил ДНК мышей и классифицировал тандемные повторы в геноме, подробно рассмотрел и проанализировал их распределение и структурные особенности. Проделанная работа позволила ему выявить сравнительно новые тандемные повторы, которые позволили автору определить новый метод поиска тандемных повторов, до этого не известных. Несомненно, вклад Комисарова А.С. в развитие биоинженерии неоспорим [1].

Среди казахстанских авторов заслуживает внимания исследования А.А. Исмаиловой, которая в своих трудах проанализировала и разработала биоинформационное приложение для идентификации тандемных повторов на базе полногеномного секвенирования. Разработанное Исмаиловой А.А. приложение позволяет вводить разные варианты цепочки ДНК животных, растений, а также человека, обрабатывать большой объем генетических данных, обнаруживать тандемные повторы по заданным параметрам. Данное приложение подходит для научных исследований [2].

Проблемы идентификации и анализа тандемных повторов при генотипировании в бактериальной клетке рассмотрел Шевцов В.А. Его труды положили основу для дальнейших исследований в этой области [3].

Ученых волновал вопрос о том, какова распространенность тандемных повторов в геномах эукариотов и прокариотов, какова их особенность. К.Усдин провел сравнительный анализ прокариотов и эукариотов. Были установлены функциональные различия, эти различия были систематизированы и обобщены определенным образом, дана характеристика тандемным повторам и их наличие в кодирующих и некодирующих областях изучаемых генов. На ряду с этим, К.Усдин исследовал и выявил тандемные повторы, которые способны вызывать заболевания и генетические нарушения у человека, рассмотрел особенности их биологических процессов [4].

Поиск тандемных повторов предполагает апеллирование огромным объёмом информации, в связи с этим, исследователю приходиться вручную перерабатывать большое количество данных. В этом случае возможны неточности в связи с человеческим фактором. Эту проблему частично решил Сперлинг А.К., который создал программу для поиска в длинной цепочке

ДНК нужных повторов, которые можно задавать системе. Такая программа может обнаружить два или более повторов любой последовательности [5].

Современные методы обработки геномных данных

1) Секвенирование нового поколения (NGS).

Разработанные технологии NGS позволяют своевременно и качественно проводить работы по секвенированию клонально амплифицированных ДНК-матриц. Процесс состоит из четырех этапов, один из которых – создание библиотек и клональная амплификация. Это позволяет проводить научные исследования в объемах, которые ещё недавно были невозможны из-за своей трудоемкости и энергозатратности. Основными методами рассматриваемой технологии NGS являются секвенирование всего генома, а также секвенирование экзома и транскриптомное секвенирование (RNA-seq).

2) Инструменты для анализа NGS-данных.

Эти инструменты содержат программы: 1) BWA (Burrows-Wheeler Aligner) – это инструмент для выравнивания прочтений на референсный геном; 2) GATK (Genome Analysis Toolkit) – нацелена на анализ геномных вариаций, также может применяться для определения однонуклеотидных полиморфизмов (SNP) и инделов. Обе программы прошли апробацию и показали себя как эффективный инструмент для работы в области медицины, биологии и генной инженерии. К основным недостатком данных программ относится их высокая стоимость.

3) Методы машинного обучения.

На данный момент применяются нейронные сети и алгоритмы кластеризации для анализа биомедицинских данных, особенно в области онкологических исследований, они нацелены на поиск закономерностей в геномных данных, а также прогнозирования функциональных элементов генома и диагностики заболеваний.

4) Технология CRISPR/Cas9.

CRISPR/Cas9 - уникальный инструмент, который позволяет не только изучать последовательность в генетическом коде, но и дает возможность целенаправленно изменять эту последовательность ДНК по заданным условиям. Эта технология дает возможность лечения и предупреждения заболеваний у людей. Казахстанские исследователи активно работают в области применения технологии CRISPR для терапии наследственных заболеваний населения.

Инструментальные средства

1) Bioinformatics Workbench

Платформа, которая используется для анализа биоинформатических данных. Она предоставляет инструменты для анализа последовательностей, поиска генов, визуализации и управления большими геномными данными.

2) Galaxy

Galaxy — это платформа для проведения биоинформатических анализов, которая позволяет исследователям выполнять анализы геномных

данных без необходимости разбираться в сложностях программирования. Этот инструмент активно используется в Казахстане для проведения различных видов анализов NGS.

3) Программный инструмент Tandem Repeat Finder.

Данный программный инструмент удобен для работы с ДНК растений и животных. Программа также имеет удобный интерфейс и хороший функциональные возможности. Инструмент Tandem Repeat Finder быстро и точно находит заданные последовательности тандемных повторов в цепочке ДНК. В Республике Казахстан разработан ряд улучшений к методам поиска тандемных повторов, особенно в связи с изучением редких и эндемичных видов флоры и фауны.

4) DNASTAR

Это программное обеспечение применяется для работы с биологическими последовательностями, такими как ДНК, РНК и белки. Оно предоставляет инструменты для выравнивания, анализа вариабельности и филогенетического анализа генетических данных.

В таблице 1 можно увидеть сравнение инструментальных средств для обработки геномных данных

Таблица 1 - Сравнение инструментов для обработки геномных данных

Инструмент	Основные возможности	Преимущества	Недостатки
BWA	Выравнивание коротких прочтений на референсный геном	Высокая точность выравнивания	Требует значительных вычислительных ресурсов

Продолжение таблицы 1

GATK	Анализ генетических вариаций, SNP и инделов	Широкая поддержка для различных типов данных	Сложный для новичков, требуются навыки в командной строке
Galaxy	Визуальный интерфейс для анализа геномных данных	Удобство использования без программирования	Ограничена масштабируемость при обработке больших данных
CRISPR/ Cas9	Редактирование генов	Возможность изменения генома с высокой точностью	Необходимы высокие знания для безопасного применения
Tandem	Поиск тандемных	Специфичен для	Ограниченная

Repeat Finder	повторов в геномных данных	поиска повторов, полезен в генетике растений и животных	область применения
DNASTAR	Анализ последовательностей ДНК, РНК и белков	Мощный инструмент для фило-генетического анализа и выравнивания последовательностей	Платное программное обеспечение

Проблемы биоинформатики

1. Неоднородность и неполнота данных

Геномные данные часто содержат пропуски, ошибки или неполные участки. Это может быть связано с ограничениями технологий секвенирования или сложностью работы с определёнными участками ДНК (например, с повторами или сложно структурированными участками). Ошибки в данных могут привести к неправильной интерпретации генетических вариаций или биологических процессов. Не до конца решён вопрос о создании единой базы данных, которая бы могла хранить и корректно интерпретировать все геномные данные.

2. Ограничения машинного обучения и искусственного интеллекта

Машинное обучение используется для предсказания функциональных элементов генома и поиска паттернов в данных. Однако модели машинного обучения часто зависят от обучающих данных, и они могут быть слишком узко ориентированными, что ограничивает их обобщающую способность. Это может привести к получению результатов, которые не всегда применимы к новым или непредсказуемым биологическим системам. Алгоритмы машинного обучения не в состоянии точно прогнозировать и определить все взаимодействия между генами и белками, тем более в сложных биологических системах, таких как человеческий организм.

3. Масштабируемость обработки данных

Современные методы, такие как секвенирование нового поколения (NGS), генерируют огромные объёмы данных, которые трудно обрабатывать и хранить. Инфраструктура для хранения геномных данных часто оказывается дорогой и сложной в реализации. Это приводит к трудностям в анализе данных, особенно в небольших лабораториях и научных группах, которые не имеют доступа к большим вычислительным мощностям. Разработка эффективных алгоритмов для анализа больших данных, с учётом будущего роста объёмов геномных данных и генетических исследований, на данном этапе развития, является важной задачей.

4. Этические и правовые вопросы

Геномные исследования связаны с важными этическими аспектами, такими как конфиденциальность данных и возможное использование геномной информации для дискриминации. Это ограничивает доступ к некоторым геномным данным и вызывает опасения относительно возможных злоупотреблений, связанных с персонализированной генетической информацией. Не до конца проработаны международные стандарты и законы, касающиеся защиты геномных данных, что делает эту область уязвимой для манипуляций и неправомерного использования.

5. Низкая воспроизводимость исследований

Результаты биоинформатических исследований нередко оказываются трудно воспроизводимыми. Это может быть связано с различиями в данных, программных версиях и вычислительных средах. Это снижает доверие к выводам исследований и замедляет прогресс в области биоинформатики. Требуется больше усилий для стандартизации методов и обеспечения воспроизводимости биоинформатических анализов.

В результате проведённого анализа можно сделать вывод, что наука о тандемных повторах продолжает развиваться, особенно в свете современных биоинформационных технологий. Вклад таких учёных, как А.С. Комиссаров, А.А. Исмаилова, В.А. Шевцов, и К.Усдин, имеет важное значение для понимания особенностей тандемных повторов, их роли в генетических процессах, а также для разработки методов их идентификации и анализа. Эти исследования создают фундамент для будущих исследований и технологий, связанных с геномными данными, предоставляя более эффективные способы работы с большими объёмами информации.

Одним из ключевых достижений последних лет является появление высокоэффективных методов секвенирования, таких как NGS, которые позволяют значительно ускорить процесс анализа геномов. Однако, несмотря на технологический прогресс, остаются актуальными проблемы точности данных, их неоднородности, а также сложности в обработке большого объёма генетической информации. Программы, такие как BWA, GATK, и Tandem Repeat Finder, предоставляют исследователям мощные инструменты для анализа, но их высокая стоимость и требовательность к вычислительным ресурсам ограничивают доступность для небольших лабораторий и исследовательских групп.

Особое внимание также следует уделить машинному обучению, которое всё активнее применяется в биоинформатике. Нейронные сети и алгоритмы кластеризации помогают выявлять закономерности и прогнозировать биологические процессы, однако их использование сопряжено с определёнными трудностями, такими как зависимость от объёмов и качества данных.

Среди современных технологий редактирования генома CRISPR/Cas9 выделяется как один из самых перспективных инструментов, который уже находит своё применение в лечении генетических заболеваний. В Казахстане

активно развиваются исследования в этой области, что открывает возможности для создания новых терапевтических методов.

Тем не менее, проблемы, такие как недостаточная масштабируемость методов обработки данных и этические вопросы, остаются ключевыми вызовами. Необходима разработка более эффективных и доступных решений, которые позволили бы всем исследователям вне зависимости от их ресурсов участвовать в геномных исследованиях.

Таким образом, дальнейшее развитие методов биоинформатики и технологий анализа геномных данных должно быть направлено на улучшение воспроизводимости исследований, оптимизацию существующих программных инструментов, решение проблем конфиденциальности и безопасности генетической информации. Эти меры позволят сделать геномные исследования более доступными, эффективными и безопасными.

Список литературы

1.Комиссаров, АС, Кузнецова, ИС, Подгорная, ОИ. (2010). Центромерные тандемные повторы мыши in silico и in situ. *Генетика*. 46: 9, 1217–1221.

2.Yekaterina, G., Ismailova, AA. Shaushenova, AG, Mutalova, ZhS, Dossalyanov, D, Ainagulova, A, Naizagarayeva, A, (2023). Interpretation of laboratory results through comprehensive automation of medical laboratory using OpenAI Eastern. *European Journal of Enterprise Technologies,* 4: 2-124, 26-34.

3.Шевцов, ВА. (2022). Биоинформационные алгоритмы для идентификации тандемных повторов в бактериальных геномах при генотипировании. Сейфуллинские чтения – 18: "Молодежь и наука – взгляд в будущее". I: IV, 66-68.

4.Sperling, AK, Li, RW, (2013) Repetitive Sequences. Brenner's Encyclopedia of Genetics: Second Edition. 150-154.

5.Kumari, D, Lokanga, RA, McCann, C, Ried T, Usdin, K. (2024) The fragile X locus is prone to spontaneous DNA damage that is preferentially repaired by nonhomologous end-joining to preserve genome integrity (external link) iScience. 27.