Лекция 4. «Сборка» полной последовательности генома. Постановка задачи восстановления геномной последовательности.

1 «Сборка» полной последовательности генома.

2 Постановка задачи восстановления геномной последовательности.

1. «Сборка» полной последовательности генома. Геномная последовательность является фундаментальной информацией для понимания биологии организма. С развитием технологий секвенирования и биоинформатики сборка полного генома стала реальной задачей, открывающей широкие перспективы в области медицины, сельского хозяйства и экологии. В последние годы произошли значительные изменения в подходах к сборке генома благодаря появлению новых технологий секвенирования и алгоритмов сборки.

Современные технологии секвенирования, такие как секвенирование нового поколения (Next-Generation Sequencing, NGS), позволили значительно увеличить объемы и скорость получения генетической информации. Однако короткие риды, генерируемые такими платформами, как Illumina, создают сложности при сборке генома из-за наличия повторяющихся последовательностей и структурных вариаций. Для преодоления этих проблем были разработаны новые подходы и инструменты.

Одним из ключевых достижений стало появление длинночитающих технологий секвенирования, таких как PacBio и Oxford Nanopore. Эти платформы позволяют получать риды длиной от нескольких тысяч до сотен тысяч пар оснований. Длинные риды существенно облегчают сборку генома, позволяя перекрывать повторяющиеся области и точно определять структурные вариации. Несмотря на более высокую частоту ошибок по сравнению с короткими ридами, новые алгоритмы коррекции ошибок и гибридные методы сборки позволяют эффективно использовать эти данные.

Гибридная сборка, которая комбинирует данные коротких и длинных ридов, стала стандартным подходом для получения высококачественных геномных сборок. Короткие риды обеспечивают высокую точность последовательности, в то время как длинные риды помогают разрешать сложные регионы генома. Интеграция данных разных технологий позволяет компенсировать недостатки каждой из них и улучшать качество сборки.

Важным направлением развития стали алгоритмы сборки, основанные на графах де Брейна и методах с использованием глубокого обучения. Новые инструменты, такие как Flye и Shasta, оптимизированы для работы с длинными ридами и способны быстро и точно собирать геномы большого размера. Применение машинного обучения позволяет улучшать процессы коррекции ошибок и прогнозирования структурных особенностей генома.

Кроме того, трехмерная организация генома и эпигенетические модификации становятся все более значимыми для понимания функциональной структуры генома. Методы, такие как Hi-C и ATAC-seq, позволяют изучать пространственную организацию хроматина и интегрировать эту информацию в процессы сборки. Это особенно важно для эукариотических геномов с высокой сложностью и повторяющимися элементами.

Автоматизация и стандартизация процессов сборки генома также являются важными факторами. Разработка пайплайнов, таких как nf-core и Galaxy, позволяет унифицировать подходы и облегчает воспроизводимость результатов. Это способствует более широкому распространению геномных технологий и повышает качество научных исследований.

В области медицинской геномики сборка генома человека и патогенных микроорганизмов открывает новые возможности для диагностики и терапии заболеваний. Индивидуальные геномные сборки позволяют выявлять редкие генетические вариации, связанные с наследственными болезнями, и разрабатывать персонализированные подходы к лечению. В эпидемиологии сборка геномов вирусов и бактерий помогает отслеживать распространение инфекций и разрабатывать эффективные меры противодействия.

В сельском хозяйстве и экологии сборка геномов растений и животных способствует селекции новых сортов и пород с улучшенными характеристиками, а также пониманию адаптационных механизмов в изменяющихся условиях окружающей среды. Метагеномные исследования, основанные на сборке геномов из смешанных образцов, позволяют изучать биоразнообразие и взаимодействия в экосистемах.

Несмотря на значительный прогресс, остаются и вызовы. Геномы многих организмов содержат большое количество повторов и сложных структур, что затрудняет их полную и точную сборку. Технологии секвенирования продолжают совершенствоваться, уменьшая стоимость и увеличивая качество данных. Разработка новых алгоритмов и методов анализа, учитывающих специфические особенности различных геномов, является актуальной задачей.

Основная стратегия создания сборки включает (1) изоляцию геномной ДНК из биологического образца и (2) фрагментацию ДНК на небольшие части, которые затем секвенируются по отдельности. После получения последовательностей небольших частей — называемых **прочтениями — исследователи собирают их, как крошечные части гигантской головоломки, в постепенно увеличивающиеся непрерывные части последовательности (называемые контигами ).** Этот подход называется  секвенированием **методом дробовика всего генома ( WGS ).**

Контиги являются первым уровнем в иерархии геномной сборки. Следующий шаг — построение каркасов (суперконтигов). Чтобы построить каркас, исследователи размещают несколько контигов в правильном порядке и ориентации. Чтобы сделать каркас единой единицей последовательности (единой записью последовательности), они представляют собой пробелы в последовательности между контигами в каркасе с помощью серий NNN (вместо последовательности [ДНК](https://www.ncbi.nlm.nih.gov/books/NBK21106/def-item/app37/) A, T, G и C). Сборки на уровне каркаса , как правило, будут иметь несколько записей каркаса плюс несколько записей контигов. Следующий шаг — правильно упорядочить, сориентировать и собрать в последовательность хромосом каркасы, принадлежащие одной хромосоме. Опять же, исследователи представляют любые пробелы в последовательности в собранной хромосоме с помощью NNN. Сборка на уровне хромосомы , как правило, будет иметь одну запись для каждой хромосомы. Нелокализованные \* и неразмещенные  записи контигов и каркасов могут сопровождать записи хромосом и вместе составлять первичную сборку . В дополнение к первичной сборке, сборка генома может содержать другие записи последовательностей, такие как заплатки (для исправления областей первичной сборки) и/или альтернативные локусы (для предложения альтернативных моделей для высоковариабельных областей хромосом). Наконец, неядерный геном (такой как митохондрия) может завершить сборку для сборки. Полные сборки не будут иметь пробелов в последовательности в собранных хромосомах. Хотя возможно генерировать полные сборки для  [прокариот](https://www.ncbi.nlm.nih.gov/taxonomy/?term=prokaryotes) — где геномы представляют собой отдельные кольцевые хромосомы — и низших [эукариот](https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=2759) (например, дрожжей), этот уровень в настоящее время недостижим для сложных геномов высших эукариот (включая человека). Обратите внимание, что исследователи могут прекратить свои усилия по секвенированию/сборке, как только соберут необходимую информацию. Поэтому многие сборки остаются неопределенно долго как коллекции контигов и/или каркасов независимо от сложности генома организма.

Сборка генома относится к реконструкции всей последовательности генома путем выравнивания и слияния прочтений последовательностей, полученных с помощью современных технологий [секвенирования генома](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/genome-sequencing) . Это необходимо (a) потому что современные[технологии NGS](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/next-generation-sequencing) в основном генерируют короткие длины прочтений ДНК (25–400 п.н. в зависимости от платформы NGS), (b) для обработки терабайт данных секвенирования, (c) для исправления ошибок, полученных во время секвенирования, и (d) для разрешения повторяющихся последовательностей, что, возможно, является самой большой проблемой, с которой сталкиваются методы сборки генома. Было написано большое количество компьютерных программ сборки генома, которые сшивают целые хромосомы из коротких фрагментированных прочтений ДНК. Программы сборки генома используют данные из одиночных и парных прочтений для сборки генома. Одиночные прочтения представляют собой непрерывные секвенированные фрагменты, которые можно объединить с помощью перекрывающихся областей в «контиги». Парные считывания — это два конца одной и той же молекулы ДНК, которые получаются путем секвенирования одного конца ДНК, а затем секвенирования ее с другого конца. Данные парных считываний могут указывать на размер повторяющихся областей.

Программы сборки генома используют два класса алгоритмов: перекрытие–разметка–консенсус (OLC) и де-брейн-граф.

В заключение, сборка полной последовательности генома является динамично развивающейся областью, объединяющей достижения молекулярной биологии, биоинформатики и компьютерных наук. Современные технологии и методы позволяют получать все более точные и полные геномные последовательности, что расширяет наши знания о жизни и предоставляет инструменты для решения важных практических задач.

**2. Постановка задачи восстановления геномной последовательности.** Решение задачи восстановления геномной последовательности, или сборки генома, является одной из важнейших проблем современной геномики и биоинформатики. Это задача состоит в восстановлении полной последовательности генома из набора коротких фрагментов ДНК (ридов), полученных с помощью методов секвенирования. Сборка генома имеет огромное значение для понимания биологии и эволюции организмов, для медицины, сельского хозяйства, экологии и метагеномики. Сегодня речь пойдет о ключевых аспектах задачи восстановления геномной последовательности: её сложности, методах и подходах, а также текущих достижениях и нерешенных проблемах.

Анализ последовательностей ДНК стал критически важным аспектом фундаментальных биологических исследований в различных прикладных областях, таких как медицинская диагностика, биотехнология, судебная биология, вирусология и биологическая систематика. Идентификация заболеваний, таких как различные виды рака, возможна посредством сравнения стабильных, мутировавших последовательностей ДНК использоваться в качестве руководства для лечения пациентов. Персонализированная медицинская помощь может быть предоставлена ​​посредством быстрого подхода к секвенированию ДНК и путем распознавания и перечисления большего количества организмов.

Сборка *de novo* относится к секвенированию нового генома, где нет референсной последовательности для выравнивания. Чтения последовательностей собираются как контиги, и качество покрытия данных сборки *de novo* зависит от размера и непрерывности контигов. Точная реконструкция генома является обязательной, поскольку согласованность и базовая точность сборки будут влиять на результаты всех последующих анализов. Проблема сборки становится более сложной и требующей больших вычислительных затрат с увеличением усилий по секвенированию и сборке геномов большего количества видов, особенно с короткими, неточными прочтениями последовательностей и геномными повторами. По сравнению с традиционными подходами, такими как секвенирование по Сэнгеру, секвенирование следующего поколения (NGS) позволяет быстрее и точнее характеризовать любой вид.

Быстрое секвенирование, достигнутое с помощью современной технологии секвенирования ДНК, сыграло важную роль в секвенировании полной последовательности ДНК или геномов различных типов и видов, включая людей, организмы, растения и микробы. Получение геномных последовательностей теперь намного проще и дешевле, чем во время проекта «Геном человека», благодаря современным методам, которые были разработаны за последние два десятилетия.

В исследовании генома сборка генома *de novo* является фундаментальным начинанием, которое привело к созданию высококачественных референсных геномов многих гаплоидных или высокоинбридинговых видов и способствовало открытию генов, сравнительной геномике и другим исследованиям. Все более мощные алгоритмы анализа необходимы, чтобы идти в ногу с растущей доступностью данных о последовательностях. Это имеет особое значение, поскольку собираются большие геномы, где наборы данных могут достигать сотен гигабайт в размере.

Поскольку сборка *de novo* обычно позволяет выполнять запросы по всему набору считываний последовательностей, огромные наборы данных представляют практическую проблему для разработчиков и пользователей программного обеспечения для сборки. В настоящее время для ассемблера требуется один компьютер с большим объемом памяти, обычно в сотни гигабайт или большой распределенный кластер соединенных компьютеров.

Одним из преимуществ метода сборки *de novo* является то, что он позволяет создавать точные референтные последовательности даже для сложных или полиплоидных поколений, предоставлять ценную информацию для картирования геномов новых организмов или завершения геномов известных организмов, разрешать чрезвычайно похожие или повторяющиеся области для точной сборки *de novo* и распознавать структурные варианты и сложные перестройки.

Секвенирование всего генома остается сложной задачей. Одной из самых критических и сложных проблем в биоинформатике является проблема сборки последовательностей. Целью сборки генома является воссоздание полного генома из нескольких относительно коротких последовательностей. Перекрытия могут быть объединены для формирования контигов в прочтениях из одной и той же области генома, но геномные повторы, более длинные, чем перекрытия, вызывают неясную реконструкцию и фрагментацию сборки. Большинство геномов, особенно эукариотические геномы, сильно повторяются и усложняют сборку, скрывая взаимосвязи прочтений множеством ложных вариантов.

Существуют две стратегии для решения этого фундаментального ограничения: увеличение эффективной длины прочтения и разделение неточных повторов на основе копий-специфичных вариантов.

В повторяющихся областях генома сложно точно собрать короткие риды, поэтому могут быть получены неточные или нерешенные сборки. Повторение областей генома было улучшено с использованием технологий длиннопрочитанных одиночных молекулярных последовательностей (SMS), таких как Pacific Biosciences и Oxford Nanopore. Однако несколько длинных участков повторяющейся ДНК не поддаются этим подходам.

С недавним развитием технологии длинных чтений стало возможным иметь почти готовые сборки. Однако извлечение информации в длинных чтениях по-прежнему подвержено ошибкам, поскольку повторы должны быть последовательно преодолены. Попытки преодолеть повторы, которые по сути неразрешимы из имеющихся чтений, приведут к неправильным сборкам и в конечном итоге повлияют на последующий научный анализ: процедура, которая может быть мотивирована обещанием более высокого балла N50. Однако, учитывая данные, консервативный подход, который прерывает сборку в точках очевидной неопределенности, может не дать самых длинных контигов, которые могут быть построены. В этом смысле ассемблер, способный распознавать и разрешать все и только те повторяющиеся шаблоны, которые разрешаемы при наличии данных чтения, должен быть оптимальным ассемблером.

Сборка генома привлекла повышенный интерес с появлением технологий NGS. Хотя представлено несколько сборщиков генома, сборка генома *de novo* с использованием считываний следующего поколения по-прежнему сталкивается с четырьмя ключевыми проблемами. Первая проблема — ошибки секвенирования, которые способствуют включению артефактов в результаты сборки. Ошибки секвенирования обычно приводят к сложному графу де Брейна. Окончательные результаты сложного графа де Брейна, как правило, неудовлетворительны. Вторая проблема — смещение секвенирования. Например, смещение состава оснований (в пользу GC-сбалансированных регионов) платформы секвенирования Illumina обычно приводит к неравной глубине секвенирования по всему геному. Третья проблема — топологическая сложность повторяющихся регионов в геноме. Большинство геномов, особенно геномы млекопитающих, имеют некоторые повторы, которые составляют около 25–50% всего генома. Повторы создают не только неправильные представления или расхождения в результатах сборки, но и несовместимую глубину данных о последовательностях. Четвертая проблема — использование больших объемов вычислительных ресурсов. Несмотря на то, что сборка небольших геномов, таких как бактериальные геномы, занимает всего несколько минут, для сборки больших геномов, таких как геномы млекопитающих, обычно требуется несколько дней или даже недель.

При восстановлении геномной последовательности возникает ряд фундаментальных сложностей, связанных с особенностями геномного материала и ограничениями технологий секвенирования. Современные методы, такие как Next-Generation Sequencing (NGS) и длинночитающие технологии (например, PacBio и Oxford Nanopore), позволяют получать огромные объемы данных, но при этом генерируют риды длиной от нескольких сотен до нескольких тысяч пар оснований. Такие риды значительно короче полного генома, что делает необходимой задачу их объединения в последовательную геномную структуру.

**Основные сложности** заключаются в следующих аспектах:

1. **Повторы**. Геномные последовательности содержат множество повторяющихся элементов, что делает процесс восстановления последовательности неоднозначным. Примером могут служить тандемные повторы и мобильные генетические элементы, длина которых часто превосходит длину ридов.
2. **Ошибки секвенирования**. Технологии секвенирования, особенно те, что работают с длинными ридами, имеют повышенную частоту ошибок, что приводит к ложным перекрытиям и осложняет сборку.
3. **Неоднородное покрытие**. Случайное распределение ридов по геному может приводить к недостаточному покрытию определенных участков, что в свою очередь ухудшает полноту сборки и качество конечной последовательности.

Эти проблемы делают задачу восстановления геномной последовательности не просто задачей простого объединения данных, но сложной вычислительной задачей, требующей методов анализа данных и статистической обработки.

Основные методы сборки генома разделяются на два основных подхода: **Overlap-Layout-Consensus (OLC)** и **графы де Брейна**. Эти подходы базируются на различных алгоритмических концепциях и по-разному подходят к решению задачи восстановления последовательности.

Метод **OLC** наиболее эффективен для длинных ридов с низким уровнем ошибок. В рамках этого метода сначала находятся все перекрытия между ридами, затем формируется макет последовательности (layout), и на основе согласования (consensus) создается единая последовательность. Этот метод, используемый в таких инструментах, как Canu и Flye, хорошо решает проблему сборки для больших геномов, но требует высоких вычислительных ресурсов.

Метод **графов де Брейна** более эффективен для коротких ридов и больших объемов данных, как при NGS-секвенировании. В этом подходе риды разбиваются на короткие последовательности (к-меры), а затем строится ориентированный граф, узлы которого представляют собой к-меры, а ребра соединяют перекрывающиеся к-меры. Несмотря на высокую производительность, этот метод чувствителен к ошибкам в ридах, что приводит к ложным ветвлениям в графе и усложняет сборку. Инструменты, такие как SPAdes, эффективно используют данный подход.

В последние годы все чаще используются **гибридные методы**, объединяющие данные коротких и длинных ридов, что позволяет компенсировать недостатки каждой из технологий. Например, в гибридной сборке длинные риды помогают справиться с повторами и разрывами, тогда как короткие риды обеспечивают высокую точность конечной последовательности. Эта стратегия доказала свою эффективность для сборки сложных геномов и метагеномных образцов.

С ростом сложности геномов и объемов данных, появление алгоритмов глубокого обучения и нейронных сетей открыло новые горизонты в области сборки генома. Системы, использующие искусственный интеллект, способны анализировать огромные объемы данных, улучшая предсказание и выравнивание последовательностей. Глубокое обучение помогает корректировать ошибки и анализировать сложные повторяющиеся области генома, обеспечивая более точные результаты. Одним из примеров является метод hifiasm, который использует фазированные графы для высокой точности восстановления геномов с гаплотипами.

После сборки генома крайне важно оценить её качество. Одной из ключевых метрик является **N50**, которая показывает минимальную длину последовательности, при которой контиги, равные или более длинные этого значения, составляют половину сборки. Другие важные показатели включают **полноту генома**, определяемую с помощью набора основных генов, таких как BUSCO, и **соответствие референсной последовательности**, если таковая доступна. Инструменты для оценки качества позволяют исследователям принимать обоснованные решения о достоверности геномной сборки.

Точные и полные геномные сборки открывают огромные возможности в биологии и смежных науках. В медицине сборка геномов помогает выявлять редкие генетические вариации, которые могут быть связаны с наследственными заболеваниями и реакцией на лекарства. В эпидемиологии и микробиологии сборка геномов патогенов позволяет отслеживать происхождение и распространение инфекций, что особенно важно для пандемий и контроля заболеваний.

В сельском хозяйстве и экологии сборка геномов помогает в селекции растений и животных, а также в изучении биоразнообразия экосистем. Метагеномные исследования позволяют получать информацию о структуре и функциях микробных сообществ, влияющих на здоровье человека, животных и растений.

Задача восстановления геномной последовательности — это не просто технический процесс, но сложная вычислительная проблема, требующая инновационных подходов и методов. Несмотря на значительный прогресс, остаются нерешенные задачи, связанные с точностью и полнотой сборки геномов, особенно для организмов с большими и сложными геномами. Сборка генома продолжает развиваться благодаря новым алгоритмам, методам глубокого обучения и гибридным подходам, что открывает новые возможности для науки и практических приложений.

Вопросы для самоконтроля:

1. Какие основные сложности возникают при сборке полной геномной последовательности из коротких ридов, и как они связаны с особенностями геномного материала?
2. В чем заключаются ключевые различия между алгоритмами Overlap-Layout-Consensus (OLC) и графов де Брейна? Какие задачи решают эти подходы, и в каких случаях они применимы?
3. Как использование длинночитающих технологий, таких как PacBio и Oxford Nanopore, помогает решать проблемы, связанные с повторяющимися последовательностями и структурными вариациями?
4. Что такое гибридная сборка генома, и какие преимущества она дает по сравнению с использованием данных одного типа (только коротких или только длинных ридов)?
5. Какие алгоритмы и методы, основанные на глубоких нейронных сетях, используются для коррекции ошибок и анализа структурных вариаций при сборке геномов? Как искусственный интеллект влияет на точность сборки?
6. Как оценка качества геномной сборки с использованием метрик, таких как N50 и полнота генома, помогает исследователям принимать обоснованные решения о достоверности полученной последовательности?
7. Каковы перспективы использования полных геномных сборок в областях медицины, сельского хозяйства и экологии? Какие практические задачи могут быть решены с их помощью?