Лекция 14. Биоинформатика в исследовании генома.

1 Биоинформатика в генетике.

2 Биоинформатика в исследовании генома.

1. Биоинформатика в генетике. Биоинформатика, связанная с генетикой и геномикой, является научной поддисциплиной, которая включает использование компьютерных технологий для сбора, хранения, анализа и распространения биологических данных и информации, таких как последовательности ДНК и аминокислот или аннотации к этим последовательностям. Ученые и врачи используют базы данных, которые организуют и индексируют такую ​​биологическую информацию, чтобы улучшить наше понимание здоровья и болезней и, в некоторых случаях, как часть медицинской помощи.

Одной из областей, где биоинформатика особенно полезна, является геномика, которая может генерировать огромные объемы информации. Биоинформатика помогает придать смысл данным, которые могут быть использованы для постановки диагноза у пациента с редким заболеванием, для отслеживания и мониторинга инфекционных организмов по мере их перемещения в популяции или для определения наилучшего лечения для пациента с раком.

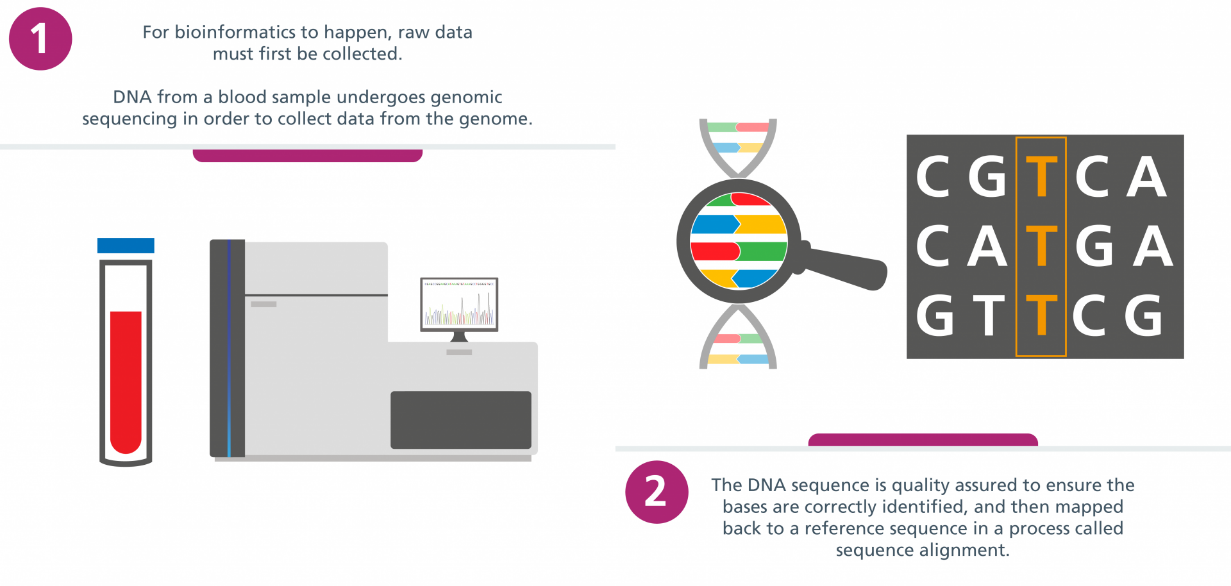
Биоинформатика используется в различных геномных тестах, но она полезна при секвенировании генома, и особенно при WGS, из-за объема данных, которые генерируются из генома человека.

После того, как образец был взят у пациента и его ДНК была извлечена, он будет секвенирован машиной для создания набора файлов данных. Вот тут-то и вступает в дело биоинформатика.

Затем эти файлы фильтруются и анализируются биоинформатиками в конвейерах, каждый из которых включает в себя различную специализированную серию шагов в зависимости от клинического вопроса и типа образца, который был секвенирован.

Результатом всей работы, проделанной биоинформатиком, является список аннотированных вариантов, который затем интерпретируется клиническим ученым для составления клинического отчета.

Этот отчет составляется индивидуально для каждого случая с целью помочь в организации медицинской помощи пациенту и его семье. Однако важно помнить, что результаты не всегда являются окончательными.



*Рисунок 1. Биоинформатический путь ДНК-тестирования пациента на наличие определенного редкого заболевания.*

Хотя мощные технологии секвенирования и опытные биоинформатики могут предоставить нам детальное понимание генома человека, знать все невозможно. Геномика — это быстро развивающаяся область, и биоинформатики ограничены точностью имеющихся у них ссылок. Другими словами, даже самые лучшие данные по-прежнему точны лишь настолько, насколько точно мы их понимаем.

2. Биоинформатика в исследовании генома.

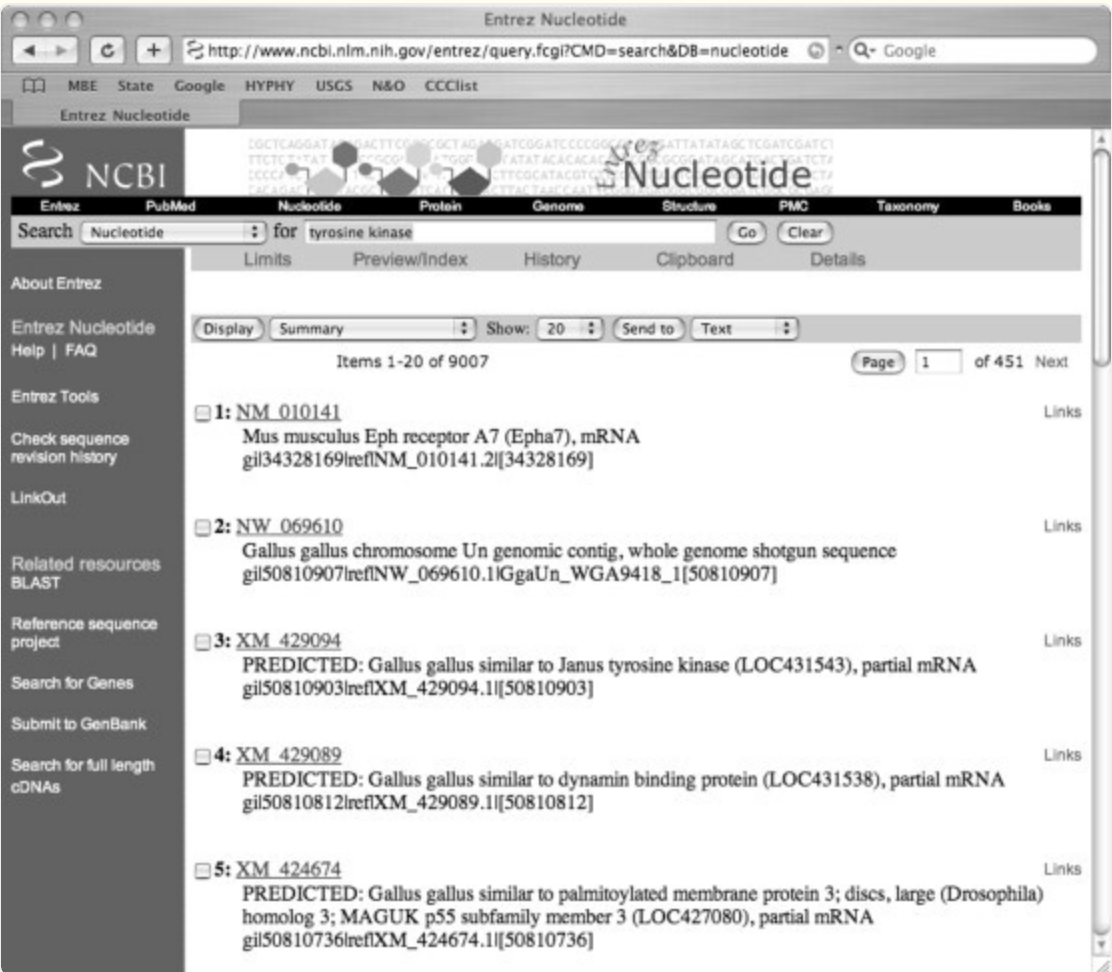
По мере совершенствования технологий сбора геномных данных росла и потребность в новых методах управления и анализа огромных объемов накопленных данных. Термин биоинформатика эволюционировал и стал включать математический, статистический и вычислительный анализ геномных данных. Работа в области биоинформатики охватывает проектирование баз данных, системную инженерию, искусственный интеллект, прикладную математику и статистику, и все это с основным фокусом на геномной науке.

Различные темы биоинформатики могут быть проиллюстрированы с использованием основных технологий, описанных в предыдущем разделе. Необходимо выполнить выравнивание последовательностей, чтобы собрать фрагменты последовательностей. Все эти последовательности, вместе с важной информацией об их источниках, функциях и т. д., должны храниться в базах данных, которые должны быть легко доступны пользователям в различных местах.

После получения последовательности необходимо аннотировать ее функцию. Одной из самых фундаментальных задач аннотации является задача вычислительного поиска генов, в которой последовательность генома или хромосомы вводится в алгоритм, который впоследствии выводит предсказанное местоположение генов. Последовательность гена, предсказанная или экспериментально определенная, должна иметь предсказанную функцию, и для этой задачи доступно множество инструментов биоинформатики. После того, как данные микрочипов станут доступны, необходимо идентифицировать подмножества корегулируемых генов и идентифицировать гены, которые по-разному экспрессируются между двумя или более методами лечения или типами тканей. Данные полиморфизма из SNP используются для поиска корреляций, например, с наличием или отсутствием заболевания в семейных родословных. Все эти вопросы имеют фундаментальное значение и затрагивают множество различных областей. По необходимости биоинформатика является в высшей степени междисциплинарной областью.

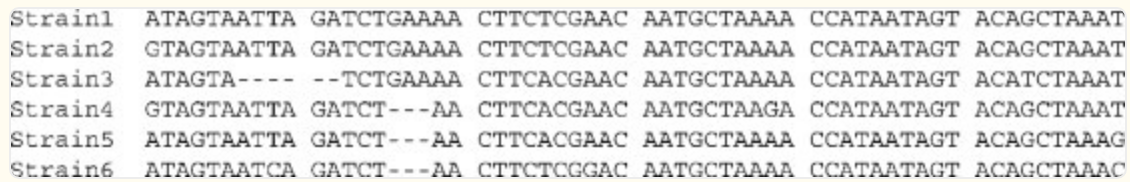
### Геномные базы данных. Геномные проекты предполагают далеко идущее сотрудничество между многими исследователями во многих областях по всему миру, и крайне важно, чтобы полученные данные были легко доступны как для участников проекта, так и для всего научного сообщества. В свете этого требования появился ряд ключевых центральных хранилищ данных. Помимо предоставления хранения и поиска последовательностей генов, некоторые из этих баз данных также предлагают передовые методы анализа последовательностей и мощные инструменты визуализации.

В Соединенных Штатах основным публичным геномным ресурсом является Национальный центр биотехнологической информации (NCBI). Веб-сайт NCBI ( [http://www.ncbi.nlm.nih.gov](http://www.ncbi.nlm.nih.gov/) ) предоставляет, казалось бы, бесконечную коллекцию данных и инструментов анализа данных. Возможно, самым важным элементом коллекции NCBI является база данных последовательностей ДНК и РНК GenBank. NCBI предоставляет различные инструменты для поиска в GenBank, а элементы в GenBank связаны с другими базами данных как внутри, так и за пределами NCBI. На рисунке 1 показаны некоторые результаты простого запроса к базе данных нуклеотидов GenBank.



***Рисунок 2.*** *Результат простого запроса базы данных GenBank в NCBI.*

### *Выравнивание последовательности.* Самым фундаментальным вычислительным алгоритмом в биоинформатике является алгоритм парного выравнивания последовательностей. Он не только имеет непосредственную практическую ценность, но и лежащий в его основе алгоритм динамического программирования также служит концептуальной основой для многих других важных методов биоинформатики. Цель выравнивания последовательностей — принять в качестве входных данных две или более последовательностей ДНК, РНК или аминокислот; определить области последовательностей, которые похожи друг на друга в соответствии с некоторой мерой; и вывести последовательности со схожими позициями, выровненными в столбцах. Выравнивание шести последовательностей из штаммов ВИЧ показано на рисунке 2.



***Рисунок 2****. Множественное выравнивание последовательностей шести последовательностей ВИЧ. Вставки или удаления нуклеотидов в существующие последовательности или из них обозначены символом «-».*

Выравнивания последовательностей имеют многочисленные применения. Выравнивания пар последовательностей помогают нам определить, имеют ли они одинаковые или схожие функции. Области выравниваний с небольшими вариациями последовательностей, вероятно, соответствуют важным структурным или функциональным областям генов, кодирующих белки. Изучая закономерности сходства в выравнивании генов нескольких видов, можно сделать вывод об эволюционной истории вида и даже реконструировать последовательности ДНК или аминокислот, которые присутствовали в предковых организмах. Многие методы аннотации, включая назначение функции белка и идентификацию участков связывания факторов транскрипции, полагаются на множественные выравнивания последовательностей в качестве входных данных.

### *Генное предсказание.* Задача предсказания генов концептуально проста для описания: имея очень длинную последовательность ДНК, определить местоположение генов. К сожалению, решение этой проблемы не так просто. В качестве первого прохода можно просто найти все пары стартовых (ATG) и стоповых (TAG, TGA, TAA) кодонов. Блоки последовательности длиннее, скажем, 300 нуклеотидов, которые фланкированы стартовыми и стоповыми кодонами и имеют длину, кратную трем, скорее всего, будут генами, кодирующими белок. Хотя этот простой метод, скорее всего, найдет много генов, он, вероятно, будет иметь высокий уровень ложноположительных результатов (неправильно предсказать, что последовательность является геном), и он, безусловно, будет иметь высокий уровень ложноотрицательных результатов (не предсказать реальные гены). Например, метод не учитывает возможность интронов и не может предсказать короткие гены. Алгоритмы поиска генов опираются на разнообразную дополнительную информацию для составления прогнозов, включая известную структуру генов, распределение длин интронов и экзонов в известных генах и консенсусные последовательности известных регуляторных последовательностей.

Новый и мощный подход к прогнозированию расположения генов использует подход сравнительной геномики. Теперь доступна вся последовательность генома человека, и известно расположение десятков тысяч генов. Предположим, что лаборатория сейчас секвенирует геном гепарда. Поскольку люди и гепарды являются млекопитающими, у них должны быть достаточно похожие геномы. В частности, большинство последовательностей генов должны быть довольно похожими. Прогнозирование генов может осуществляться путем выполнения попарного выравнивания последовательностей двух геномов, а затем предсказания того, что позиции в геноме гепарда, соответствующие расположениям известных человеческих генов, также являются генами гепарда. Этот подход является чрезвычайно эффективным, хотя он, очевидно, пропустит гены, которые являются уникальными для одного вида или другого.

Функциональная аннотация. После того, как геном секвенирован и его гены найдены или предсказаны, следующим шагом в биоинформатическом конвейере является определение биологической функции генов. В идеале молекулярно-биологическая работа должна проводиться в лаборатории для изучения функции каждого гена, но очевидно, что такой подход неосуществим. Будут описаны два основных вычислительных подхода, один с использованием сравнительной геномики, а другой с использованием HMM.

***Сравнительные геномные подходы*** к назначению функций генам основаны на простом логическом предположении: если ген вида A очень похож на ген вида B, то два гена, скорее всего, имеют одинаковые или связанные функции. Эта логика давно применяется на более высоких биологических уровнях (например, почки разных видов имеют одну и ту же базовую биологическую функцию, хотя точные детали могут различаться у двух видов). На уровне генов вывод менее точен, особенно если виды, участвующие в сравнении, не являются близкородственными, но подход тем не менее полезен и обычно эффективен.

Простые поиски в базах данных являются наиболее простым сравнительным геномным подходом к функциональной аннотации. Недавно обнаруженная последовательность гена, которая возвращает совпадения с генами цитохромоксидазы при вводе в BLAST, скорее всего, сама будет геном цитохромоксидазы. Сложности возникают, когда совпадения относятся к отдаленно родственным видам, когда соответствующие области очень короткие или когда последовательность соответствует членам многогенного семейства. В первом случае функции генов могли измениться за десятки или сотни миллионов лет с тех пор, как два организма разделяли общего предка. Однако, если два или более таких отдаленно родственных организма имеют последовательности генов, которые почти идентичны, можно привести веский аргумент, что ген имеет решающее значение в обоих организмах и что одна и та же функция сохранялась на протяжении всей эволюционной истории.

Короткие совпадения могут возникать просто в результате элементарной структуры белка. Например, две последовательности могут иметь области, которые совпадают просто потому, что они обе кодируют альфа-спиральные области. Такие совпадения предоставляют полезную структурную информацию, но более сильный вывод об общей функции не оправдан. Мультигенные семейства являются результатом дупликации генов, за которой следует функциональная дивергенция. Примерами служат семейства генов глобина и амилазы. В какой-то момент в прошлом один ген в одном организме был полностью дублирован в геноме. В этот момент дублированная копия могла свободно развить новую, но часто связанную функцию. Последующие дупликации позволяют таким семействам расти и диверсифицироваться. Из-за их общего происхождения все члены семейства генов, как правило, имеют схожие последовательности ДНК. Этот факт затрудняет назначение функции с высокой точностью, когда совпадения появляются в поиске по базе данных, но часто он обеспечивает общий класс функций для последовательности запроса.

Были предприняты попытки классифицировать все известные белки по функциональным группам с использованием сравнительной геномики.

**Выявление дифференциально экспрессируемых генов.** Распространенным экспериментом является использование технологии микрочипов или олигонуклеотидных матриц для измерения уровня экспрессии нескольких тысяч генов при двух различных «обработках». Часто бывает так, что одна обработка является контролем, а другая — стимулом окружающей среды, таким как лекарство, химикат или изменение физической переменной, такой как температура или pH. Другие возможности включают сравнение двух типов тканей (например, мозг против сердца), между больными и здоровыми тканями (например, опухоль против нормы) или между образцами на двух фазах развития (например, эмбрион против взрослого). Одной из основных причин проведения такого эксперимента является выявление генов, которые по-разному экспрессируются между двумя обработками.

Биоинформатика стала незаменимым мостом между необработанными генетическими данными и значимыми знаниями о структуре, функциях и значении генома. Он включал в себя разнообразный набор вычислительных инструментов, алгоритмов и баз данных, специально предназначенных для извлечения знаний из генетического кода. По своей сути биоинформатика стремилась преобразовать A, T, C и G последовательностей ДНК в понятную информацию, которая могла бы способствовать научным открытиям.

Одной из основных задач биоинформатики был анализ последовательностей. Учитывая огромное количество доступных генетических данных, исследователям потребовались инструменты для сравнения, сопоставления и выявления закономерностей в последовательностях ДНК. Биоинформатики разработали алгоритмы, способные идентифицировать гены, регуляторные элементы и генетические вариации внутри геномов. Эти инструменты позволили ученым точно определить конкретные гены, ответственные за черты характера, болезни и эволюционные адаптации.

Функциональная геномика стала еще одним центром биоинформатики. Целью проекта было понять, как гены функционируют и взаимодействуют в сложных биологических системах. Чтобы добиться этого, биоинформатики объединили данные из различных источников, включая транскриптомику (изучение экспрессии генов), протеомику (изучение белков) и метаболомику (изучение метаболитов). Этот целостный подход позволил исследователям разгадать сложные сети генетической регуляции и белковых взаимодействий, которые лежат в основе клеточных процессов.

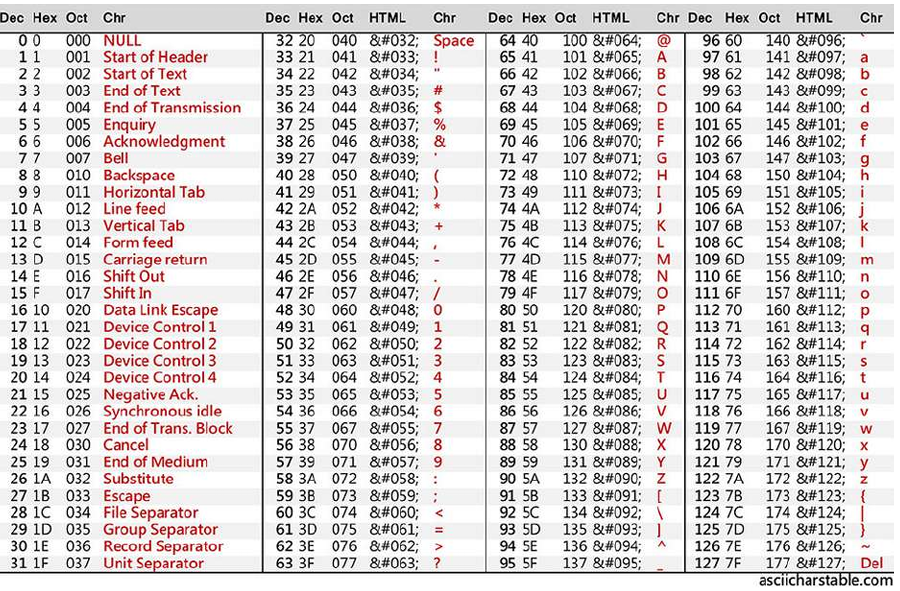
В сфере персонализированной медицины биоинформатика сыграла ключевую роль в фармакогеномике. Эта дисциплина исследовала, как генетический состав человека влияет на его реакцию на наркотики. Анализируя генетические вариации, инструменты биоинформатики помогли выявить генетические маркеры, связанные с эффективностью лекарств и побочными реакциями. Эти знания проложили путь к индивидуальному лечению, учитывающему генетический профиль пациента, что в конечном итоге повысило эффективность и безопасность медицинских вмешательств.

Кроме того, биоинформатика была незаменима при изучении геномики заболеваний. Ученые полагались на биоинформатический анализ, чтобы расшифровать генетическую основу заболеваний, начиная от редких генетических нарушений и заканчивая сложными состояниями, такими как рак и диабет. Инструменты биоинформатики облегчили идентификацию генов, связанных с заболеваниями, что позволило глубже понять механизмы заболевания и разработать таргетные методы лечения.

Эволюционная биология также извлекла пользу из биоинформатики. Сравнивая генетические последовательности разных видов, биоинформатики внесли свой вклад в наше понимание эволюционных взаимоотношений и дивергенции. Этот сравнительный анализ помог реконструировать древо жизни, проливая свет на историю эволюции организмов и генетические адаптации, которые позволили видам процветать в соответствующей среде обитания.

В заключение, биоинформатика послужила катализатором, который превратил необработанные генетические данные в бесценные знания о геноме. Его инструменты и алгоритмы позволили исследователям расшифровать генетический код, раскрыть функциональные аспекты генов, продвинуть персонализированную медицину, разгадать генетическую основу болезней и изучить тонкости эволюционной биологии. Поскольку геномная революция продолжает развиваться, биоинформатика остается на переднем крае, стимулируя научные открытия и раскрывая глубокие тайны, закодированные в нашей ДНК.

Контроль качества ридов и предварительная обработка Как уже было сказано, высокопроизводительное секвенирование на разных платформах не расшифровывает всю ДНК организма как один длинный отрезок, а вместо этого создает большой набор коротких фрагментов, каждый из которых содержит лишь малую часть генетической информации. Эти фрагменты называются «последовательным чтением» и хранятся в так называемых файлах «FASTQ». Риды, полученные в результате секвенирования, могут быть собраны de novo в полный геном или сопоставлены с уже собранным эталонным геномом родственного организма. Однако ни одна технология секвенирования не является идеальной, и сырые риды неизбежно содержат ошибки секвенирования. Вероятность ошибки для каждого нуклеотида при каждом чтении всегда записывается в файл FASTQ. Поэтому самым первым шагом фрагментарного анализа является контроль качества и фильтрация. Этот шаг направлен на удаление некачественных ридов. Файлы FASTQ всегда содержат 4 строки на последовательность (Рис. 3). В первой строке отображается идентификатор последовательности и описание. Вторая строка содержит последовательность нуклеотидов. Третья строка обычно содержит только символ «+» и иногда тот же идентификатор и описание последовательности, что и первая строка. В четвертой строке отображается оценка качества каждого нуклеотида, показанного во второй строке.



***Рисунок 3.*** *Структура файла FASTQ.*

# **Применение биоинформатики в геномных исследованиях.** Слияние биологии и информационных технологий, известное как биоинформатика, оказало глубокое влияние на геномные исследования. Эта междисциплинарная область открыла новую эру понимания тонкостей генома, позволив ученым исследовать разнообразные приложения, от персонализированной медицины до эволюционной биологии.

**1. Анализ последовательностей.** Одним из фундаментальных применений биоинформатики является анализ последовательностей. Исследователи используют вычислительные инструменты для выравнивания, сравнения и интерпретации последовательностей ДНК и РНК. Этот процесс позволяет идентифицировать гены, регуляторные элементы и генетические вариации внутри геномов. Это основа для понимания генетических кодов и выявления того, как они влияют на черты характера и болезни.

**2. Функциональная геномика.** Биоинформатика способствует функциональной геномике, целью которой является расшифровка того, как гены работают в биологических системах. Объединяя данные из различных источников, таких как транскриптомика (экспрессия генов), протеомика (взаимодействие белков) и метаболомика (анализ метаболитов), исследователи получают представление о сложных сетях, которые управляют клеточными процессами. Эти знания имеют решающее значение для понимания механизмов заболеваний и биологических функций.

**3. Фармакогеномика.** Персонализированная медицина является отличительной чертой влияния биоинформатики на здравоохранение. С помощью фармакогеномики инструменты биоинформатики анализируют генетический состав человека, чтобы предсказать, как он будет реагировать на лекарства. Такой индивидуальный подход повышает эффективность лечения, сводя к минимуму побочные реакции, совершая революцию в уходе за пациентами.

**4. Геномика болезней.** Биоинформатика играет жизненно важную роль в раскрытии генетической основы болезней. Исследователи используют биоинформатический анализ для выявления генов, связанных с заболеваниями, и генетических вариаций. Эти знания углубляют наше понимание механизмов заболеваний и помогают разрабатывать таргетные методы лечения, особенно для сложных состояний, таких как рак и неврологические расстройства.

**5. Эволюционная биология.** Сравнительная геномика, раздел биоинформатики, способствует нашему пониманию эволюционных взаимоотношений и адаптаций. Сравнивая генетические последовательности разных видов, исследователи реконструируют древо жизни, выявляя генетические изменения, которые привели к дивергенции видов и адаптации к различным средам.

**6. Структурная геномика.** Инструменты биоинформатики используются в структурной геномике для прогнозирования трехмерных структур белков и нуклеиновых кислот. Эти прогнозы необходимы для понимания того, как молекулы функционируют и взаимодействуют, помогая в разработке лекарств и разработке белков.

**7. Метагеномика.** В исследованиях окружающей среды и микробов метагеномика опирается на биоинформатику для анализа сложных смесей генетического материала различных организмов. Этот подход позволяет изучать целые микробные сообщества, проливая свет на разнообразные экосистемы, которые играют решающую роль в биогеохимических циклах и здоровье человека.

**8. Функциональная аннотация.** Биоинформатика помогает в функциональной аннотации, приписывая генам и белкам биологические функции. Эта аннотация жизненно важна для аннотирования недавно секвенированных геномов, понимания функций генов и определения потенциальных мишеней для лекарств.

**9. Эпигеномика.** Биоинформатика играет важную роль в эпигеномике, изучении химических модификаций ДНК и гистонов, которые регулируют экспрессию генов. Это помогает идентифицировать эпигенетические маркеры, связанные с заболеваниями, и проливает свет на то, как факторы окружающей среды могут влиять на регуляцию генов.

Применение биоинформатики в геномных исследованиях далеко идущее и преобразующее. От раскрытия генетической основы болезней до развития персонализированной медицины и изучения сложностей эволюционной биологии, биоинформатика произвела революцию в нашей способности использовать генетические данные для улучшения здоровья человека и нашего понимания жизни на Земле. Поскольку технологии продолжают развиваться, влияние биоинформатики на исследования в области геномики будет расти еще больше, обещая новые идеи и прорывы в ближайшие годы.

# **Вызовы и будущие направления.** Хотя биоинформатика произвела революцию в геномных исследованиях, она не лишена проблем и новых возможностей. Заглядывая в будущее, понимание этих проблем и предвидение будущих направлений имеют важное значение для использования всего потенциала биоинформатики в геномике.

**1. Объем и сложность данных.** Главной проблемой биоинформатики является огромный объем и сложность генетических данных. Наборы геномных данных продолжают расти в геометрической прогрессии благодаря достижениям в технологиях секвенирования. Управление, хранение и обработка этих огромных наборов данных требуют постоянных инноваций в инфраструктуре данных и вычислительных ресурсах.

**2. Качество данных.** Обеспечение качества и точности геномных данных имеет решающее значение. Ошибки в секвенировании, сборке или аннотации могут привести к неверным выводам. Разработка надежных методов контроля качества и процессов проверки данных необходима для получения надежных результатов исследований.

**3. Разработка алгоритмов.** По мере развития геномики растет и потребность в сложных алгоритмах и вычислительных методах. Разработка эффективных алгоритмов для таких задач, как вызов вариантов, анализ транскриптома и структурное предсказание, остается важнейшей областью исследований. Кроме того, интеграция мульти-омных данных и подходов машинного обучения создает захватывающие возможности и проблемы.

**4. Этические соображения.** Этические последствия геномных исследований и биоинформатики сложны. Проблемы конфиденциальности, обмен данными и информированное согласие являются постоянными этическими соображениями. Нахождение баланса между доступностью данных для исследований и защитой генетической информации людей является постоянной проблемой.

**5. Междисциплинарное сотрудничество.** Биоинформатика процветает благодаря сотрудничеству биологов, компьютерщиков и статистиков. Поощрение междисциплинарного общения и сотрудничества имеет важное значение для развития этой области и использования разнообразного опыта.

**6. Воспроизводимость:** Обеспечение воспроизводимости биоинформатических анализов является приоритетом. Исследователи должны всесторонне документировать свои методы, инструменты и параметры, чтобы позволить другим повторить их результаты. Стандарты и лучшие практики в области биоинформатики играют жизненно важную роль в достижении воспроизводимости.

**7. Клинический перевод.** Преодоление разрыва между геномными исследованиями и клиническими применениями является постоянной проблемой. Включение геномных данных в рутинную медицинскую помощь требует устранения нормативных, этических и логистических препятствий, обеспечивая при этом достоверность и клиническую полезность генетических тестов.

**8. Обучение и образование.** Экспертиза в области биоинформатики пользуется большим спросом, но существует нехватка квалифицированных специалистов. Инвестиции в программы обучения и образования в области биоинформатики имеют решающее значение для создания квалифицированной рабочей силы, способной удовлетворить растущие потребности исследований в области геномики.

**9. Точная медицина.** Будущее медицины – за точной медициной, где лечение подбирается с учетом генетического профиля человека. Реализация этого видения требует интеграции геномных данных в клиническую практику и решения проблем, связанных с интерпретацией данных и принятием клинических решений.

**10. Аналитика больших данных.** Область биоинформатики должна постоянно адаптироваться к достижениям в области анализа больших данных. Для анализа и интерпретации все более сложных наборов данных необходимы масштабируемые и эффективные вычислительные решения, позволяющие исследователям извлекать значимую информацию.

В ближайшие годы биоинформатика продолжит развиваться как динамичная область на стыке биологии и науки о данных. Решение этих проблем и использование возможностей биоинформатики позволят исследователям раздвинуть границы геномных исследований, что приведет к прорывам в понимании генетики, болезней и персонализированной медицины. Более того, биоинформатика будет играть ключевую роль в расширении наших знаний о роли генома в эволюции, экологии и более широком контексте жизни на Земле. Поскольку технологии и вычислительная мощь продолжают развиваться, биоинформатика остается на переднем крае научных открытий, обещая будущее, в котором генетический код станет мощным инструментом для улучшения здоровья человека и улучшения нашего понимания мира природы.

Вопросы для самоконтроля:

1. Как методы биоинформатики помогают в анализе больших объемов геномных данных, полученных при секвенировании?
2. В чем заключается значение алгоритмов выравнивания последовательностей для идентификации гомологичных генов между различными видами?
3. Какие методы биоинформатики используются для аннотации геномов, и какие данные они предоставляют для функциональной интерпретации генов?
4. Как биоинформатические подходы помогают в выявлении генетических маркеров, связанных с наследственными заболеваниями?
5. Как анализ сетей взаимодействий белков и генов способствует пониманию молекулярных механизмов в клетках?