
ЛЕКЦИЯ 4

ДНК. МЕТОДЫ СЕКВЕНИРОВАНИЯ

1. Секвенирование геномов

Секвенирование происходит в 3 этапа: выделение геномной ДНК, прочтение и создание библиотеки фрагментов; сборка; аннотация и анализ.

Любой метод изучения веществ никогда не работает с одной молекулой. Для секвенирования, масс-спектрометрии и тех методов, откуда получается большое количество данных, необходимы количества порядка 10^5 - 10^8 .

Наиболее часто используемая масса для сиквенса – несколько нанограмм. Сколько в нанограмме ДНК? Пусть, длина ДНК примерно 1,5 млн. нуклеотидов ($1,5 * (10^6)$), а масса 1 нуклеотида примерно 500 Da, тогда масса ДНК примерно равна $1,5 * (10^6) * 500 = 10^9$ г/моль. $(10^{-9}) / (10^9) = 10^{-18}$ моль, $10^{-18} / 10^{23} = 10^5$ – цепей ДНК.

Цепи фрагментируют ультразвуком, рестриктазами или любым другим способом. Главное знать размер полученных фрагментов, например, по методу дробовика (shotgun sequencing). Полученные прочитанные фрагменты надо собрать в исходную последовательность. Исходная цепь восстанавливается благодаря тому, что у фрагментов, полученных от разных цепей, есть общие перекрывающиеся части. Для начала, из этих фрагментов нужно создать библиотеку — набор из последовательностей длиной в 2–3 тыс. нуклеотидов.

Для хорошего сигнала во время секвенирования, надо много одинаковой ДНК в пробирке. Копии фрагментов получают путем их клонирования.

2. Клонирование

Плаزمида — это кольцевая ДНК длиной 2–15 тыс. нуклеотидов, не являющаяся обязательной частью генома большой клетки, в которой она находится. Плазмиды живут внутри клетки и размножаются за счет ее строительного материала и ферментов. При делении клетки, плазмиды оказываются во всех дочерних клетках. Кроме того, они способны переходить из клетки в клетку путем горизонтального переноса.



С помощью методов генной инженерии в плазмиду встраивают случайные фрагменты ДНК, полученные методом дробовика. Этими генетическими конструкциями обрабатывают колонию *E. coli*. Концентрация плазмид подобрана таким образом, что статистически 1 плазида может попасть в 1 бактерию. Кроме того, в последовательности плазмиды содержится ген TetM, который обуславливает устойчивость клетки к тетрациклину.

В чашку Петри, обработанную тетрациклином, помещают культуру. Те бактерии, у которых нет плазмиды и, соответственно, нет гена TetM, погибают, а бактерии с плазмидами остаются, растут и размножаются. На чашке, в месте, где осталась бактерия с плазмидой образуется колония из прямых потомков этой клетки. Во всех потомках будет плазида. Далее колонии рассеиваются по чашкам Петри, вместе с кишечной палочкой размножаются и интересующие фрагменты ДНК.

Этап клонирования необходим, так как изначальное количество выделенной ДНК мало, чтобы увидеть четкий сигнал с приборов.

Свойство ДНК размножаться в необходимых количествах используют, например, при исследовании останков древних людей. В ископаемых ДНК находится в плохом состоянии, но благодаря тому, что ее можно реплицировать, получается достаточное количество материала для дальнейшей работы.

Далее из клеток кишечной палочки выделяется вся ДНК, а потом и сами участки плазмид. Это делается за счет предварительно внесенных по краям фрагмента сайтов рестрикции для известных рестриктаз. Рестриктаза вырезает интересующий фрагмент. В итоге в 1 пробирке от одной колонии получается много одинаковых фрагментов ДНК. Таким способом можно увеличить количество ДНК на 6 порядков.

3. Секвенирование по Сэнгеру

Секвенировать неизвестную последовательность можно двумя способами:

- можно отщеплять по одному нуклеотиду и определять каждый, но нужно следить, чтобы не отщеплялось по нескольку букв.
- можно достраивать цепь ДНК.

Первый способ не очень эффективен. Сейчас более распространен второй метод. Для него в реакционную смесь добавляют нуклеозид-трифосфаты (dNTP) и фермент, который будет по правилу комплементарности достраивать вторую цепь на матричной ДНК. Особенность метода в том, что в реакционную смесь добавляют **дидезоксинуклеотиды** (ddNTP: ddA, ddT, ddG, ddC). Если такой нуклеотид встроится в ДНК, то синтез цепи терминируется. И так как дидезоксинуклеотиды встраиваются полимеразой случайно, в результате получится набор цепей ДНК разных длин.

Пользуясь этим свойством, реакцию разносят по 4 пробиркам, и в каждую добавляют только по одному ddNTP и все 4 dNTP (например, в первую пробирку добавлены А, Т, G, С, ddA, во вторую — А, Т, G, С, ddT, и т. д.). Тогда в первой пробирке обрывы цепи будут происходить только по аденину, во второй — по тимину и т. д..

Далее образцы из каждой пробирки закапывают в отдельную лунку в геле для электрофореза и разгоняют цепи электрическим полем. Чем длиннее фрагмент, тем медленнее



! Конспект не проходил проф. редактуру, создан студентами и, возможно, содержит смысловые ошибки. Следите за обновлениями на lectoriy.mipt.ru.

он будет ползти и останется недалеко от лунки. После электрофореза в геле видны полосы ДНК, соответствующие каждой своей длине (см. рис.4.1). Сравнивая положение этих полос, достраивают исходную последовательность ДНК.

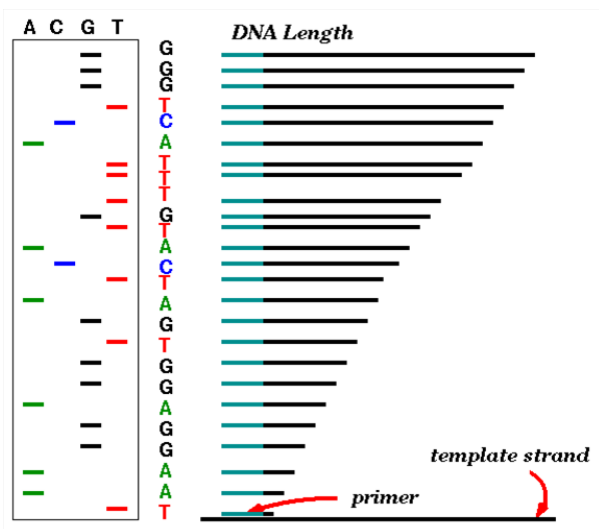


Рис. 4.1: Электрофорезный гель

В случае автоматического секвенирования, можно подобрать такие нуклеотиды, которые при присоединении к цепи будут давать квант света. Тогда добавляя каждый раз один нуклеотид, фиксировать вспышку света над реакционной камерой и по этому оценивать присоединился ли нуклеотид или нет. Это принцип пиросеквенирования (при пиросеквенировании вспышку света дает не присоединившийся нуклеотид, а отщепившийся и расщепленный ферментом пирофосфат, прим. автора).

Результат работы секвенатора — хроматограмма (см. рис.4.2), в которой совмещены вспышки с 4 пробирок сразу.

Таким образом, прочитываются все начальные фрагменты (2–3 тыс. н.), полученные методом дробовика.

Техника позволяет секвенировать только по 500 нуклеотидов с обоих концов фрагмента. Если изначально фрагмент был длиной 2 тыс. нуклеотидов, то посередине остаются неизвестными 1000 нуклеотидов. Помечают, что эти два чтения по краям связаны между собой пробелом в 1 тыс. нуклеотидов. Пусть эти чтения «R1-left» и «R1-right».

Какие-то участки из разных фрагментов одной и той же ДНК могут наложиться друг на друга. Таким образом, восстанавливается небольшой кусочек ДНК и закрываются небольшие пробелы (сбор контигов).

Часто бывает, что многие пробелы не закрываются таким способом. Чтобы попался нужный рид, закрывающий какой-то пробел, нельзя секвенировать бесконечно наугад. Для этого используется другой метод, который описан ниже.

На этом этапе получены контиги. Но неизвестно, как они располагаются друг относительно друга в исходной строке. Из библиотеки ридов выбирают те пары ридов, которые располагаются по краям контигов (и по краям пробелов). И если пробел между этими ридом составляет не больше 1 тыс. нуклеотидов, используется технология ПЦР (полимеразная цепная реакция).

! Для подготовки к экзаменам пользуйтесь учебной литературой. Об обнаруженных неточностях и замечаниях просьба писать на pulsar@phystech.edu



Конспект не проходил проф. редактуру, создан студентами и, возможно, содержит смысловые ошибки. Следите за обновлениями на lectoriy.mipt.ru.

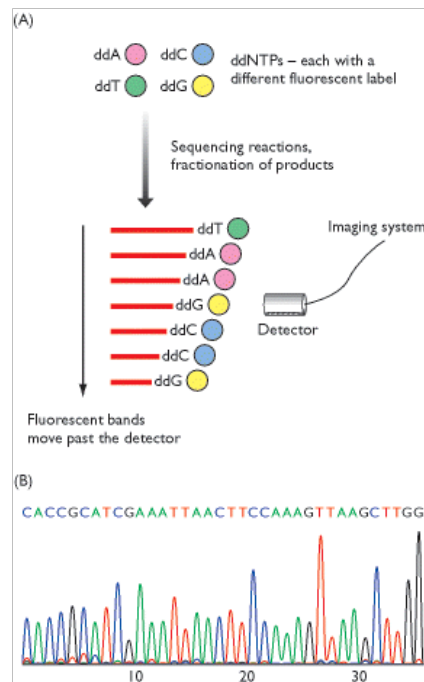


Рис. 4.2: Хроматограмма

Это ключевая технология в исследовании молекул ДНК и РНК, она основывается на естественных процессах, происходящих в живой клетке. Перед делением в клетке есть фермент полимераза, который на расходящихся цепях материнской ДНК достраивает дочерние цепи. В итоге получается 2 копии ДНК.

Для работы полимеразы, как в клетке так и в амплификаторе (в аппарате ПЦР), нужна матричная ДНК и затравки (праймеры) для ее работы. Праймер связывается с ДНК комплементарно (праймер — это РНК-фрагмент, прим. автора), на этот комплекс садится полимераза и достраивает недостающую цепь ДНК. Можно синтезировать другой праймер, который будет садиться на ДНК с противоположного конца и достраивать тот же фрагмент, но в обратном направлении. При этом важно соблюдать температурный режим:

- при повышенной температуре ($70 - 80^\circ$) происходит плавление ДНК, ее цепи расходятся
- на цепи садятся праймеры (этот этап называется «отжиг праймеров» и он происходит при более низкой температуре, которая зависит от самого праймера, прим. автора)
- далее полимеразы достраивает цепи (при этом снова нужна смена температуры. Она зависит от того, откуда взят фермент, например, чаще всего используется taq-полимераза с температурным оптимумом в районе $70 - 80^\circ$, прим. автора) и цикл повторяется.

За каждый цикл количество ДНК удваивается.



Для подготовки к экзаменам пользуйтесь учебной литературой. Об обнаруженных неточностях и замечаниях просьба писать на pulsar@phystech.edu

! Конспект не проходил проф. редактуру, создан студентами и, возможно, содержит смысловые ошибки. Следите за обновлениями на lectoriy.mipt.ru.

Чтобы искусственно синтезировать праймеры к незаполненному участку, необходимо знать кусок этой последовательности длиной примерно в 20 нуклеотидов. Эта информация содержится в библиотеке ридов.

На данном этапе имеются исходная ДНК, последовательность нуклеотидов которой необходимо восстановить, контиги с известным расположением друг относительно друга и неизвестные между ними. Далее синтезируют праймеры, которые будут комплементарны пограничным участкам контигов. К тотальной ДНК добавляют праймеры и полимеразу, которая в ходе ПЦР достраивает неизвестный участок много раз. Полученный участок секвенируют. Если неизвестный участок был длиной примерно 1 тыс. нуклеотидов, то за один такой цикл он заполнится. Если же пробел был длиной 10 тыс. н., то количество циклов «создание праймеров–ПЦР–секвенирование» увеличится.

Может быть такая неприятная ситуация, когда осталось всего 5 контигов, но каким образом они друг относительно друга ориентированы все еще не известно. То есть не нашлось ни одного фрагмента из библиотеки, который соответствовал одним концом одному контигу, а другим концом – другому. Тогда снова используют технологию ПЦР. Для этого на оба конца каждого контига подбирают по праймеру, всего получается 10 праймеров. Затем ставят ПЦР с тотальной ДНК и всеми возможными комбинациями праймеров. По наличию продукта после ПЦР можно определить расположение контигов относительно друг друга.

Собрать много больших контигов можно довольно быстро и автоматизировано, а заполнение пробелов происходит вручную. Это дорого, долго и может быть ненужно, потому что 99% последовательности уже известно. На сегодняшний день в генбанке все последовательности находятся именно в таком виде.

4. Геном человека

Программа по секвенированию генома человека стартовала в 1990 году, это обошлось в 3,5 млрд USD. В феврале 2011 получен черновой сиквенс. Работало огромное количество лабораторий, но российские лаборатории участия не принимали (в отличие от российских ученых). Тогда еще не было разработан метод дробовика.

Методика заключалась в том, что сначала поделили хромосомы, потом сами хромосомы порезали на фрагменты по 10 тыс. п. н., их клонировали в одних бактериях, очистили один фрагмент от другого и раздробили на еще более мелкие куски, которые затем снова клонировали в *E.coli*, выделили, прочитали и склеили друг с другом.

Крейг Вентер — автор метода дробовика. Он считал, что не нужно кучу раз пере-севать и клонировать куски ДНК, а можно просто разбить сразу на куски и потом их собрать. Крейг Вентер с помощью придуманного им метода занимался секвенированием параллельно с проектом Геном человека с 1999 года. К 2011 году вышли 2 журнала, Nature и Science. В одном из них опубликовался консорциум о том, что они собрали геном человека и с разницей в несколько месяцев опубликовался Крейг Вентер. Он сделал это один и во много раз дешевле и быстрее чем, остальные (он за 2 года, а консорциум 11 лет).

В 2003 году был получен полный сиквенс. В 1995 году был сделан первый геном организма длиной 1,8 млн. пар нуклеотидов. Затем были получены геномы *Mycoplasma*, *Helicobacter*, в 1996 — дрожжи, 2000 — дрозофила и червь. В период с 1995 по 2003 люди

! Для подготовки к экзаменам пользуйтесь учебной литературой. Об обнаруженных неточностях и замечаниях просьба писать на pulsar@phystech.edu



Конспект не проходил проф. редактуру, создан студентами и, возможно, содержит смысловые ошибки. Следите за обновлениями на lectoriy.mipt.ru.

научились секвенировать от 1 млн. до 1 млрд. последовательностей (возможно, лектор имеет ввиду секвенирование геномов длиной 1 млн.-1 млрд. нуклеотидов, прим. автора), а до этого с момента открытия структуры ДНК ничего подобного не происходило. Теперь же можно легко получать информацию о геномах, с малыми затратами времени и денег. Отличия генома человека, полученного в 2001 и в 2003 годах:

- было закончено 99% ген-содержащих областей генома (а черновой 90%);
- точность не более 1 ошибки на 10 тыс. позиций (в 2001 1 ошибка на 1000);
- пробелов осталось 400, а было 150 тыс;
- размер самой длинной непрерывной строки — 27 млн позиций (ранее около 80 тыс);
- оставшиеся 1,5% длины генома и 400 брешей будут закрываться вне рамок общемирового скоординированного проекта (институт UCSC, который занимается поддержанием информации о геноме человека, на его сайте есть все последние версии генома человека (последняя «Human Genome 19»)).

Общие характеристики генома человека:

- длина ДНК в вытянутом состоянии почти 2 метра;
- число нуклеотидов примерно $3 * (10^9)$;
- число хромосом — 23 (первая самая большая, последняя самая маленькая).

Гранты в области секвенирования генома человека:

- 2004 г. — NIH (самый большой институт США (национальный институт здоровья), в структуру которого входят все остальные государственные институты) предложил грант в 70 млн. USD на снижение стоимости генома человека с 3 млрд. USD до 1 тыс. USD;
- 2006 г. — X Prize Foundation объявили о гранте в 10 млн. USD за разработку метода, позволяющего секвенировать 100 геномов человека в течение 10 дней, и чтобы стоимость каждого не превышала 10 тыс. USD.

5. Другие методы секвенирования

Новый тренд исследований в этой области — сделать секвенирование быстрее, дешевле и надежнее. Появились 4 системы **секвенирования второго поколения (NGS)**: Genome sequencer 20/FLX, Solexa 1G, SOLiD system, Polonator G.007. Некоторые остались на рынке до сих пор, но под другими названиями.

Ключевые принципы работы секвенаторов второго поколения заключаются в следующем. Исходные ДНК нарезаются на куски по методу дробовика. Далее, исключая этап пересевания через клетки, работают с получившимся набором. На шарик-носитель диаметром 10 нм специальным образом прикрепляется фрагмент ДНК. На шарике может уместиться до 1000 таких последовательностей. Шарик содержит металл, который



Для подготовки к экзаменам пользуйтесь учебной литературой. Об обнаруженных неточностях и замечаниях просьба писать на pulsar@phystech.edu

притягивается магнитом. Это удобно, если надо носители с ДНК промыть или переместить. Далее ставят ПЦР: шарики помещаются в раствор с нуклеотидами, полимеразой и праймерами. При этом получившаяся копия ДНК тоже прикрепляется к соответствующему шартику. После нескольких циклов ПЦР получается шарик с прикрепленными к нему одинаковыми фрагментами ДНК.

Далее на подложку с порами наносятся шарики, в каждую пору попадает по одному шартику. Затем добавляют необходимые ферменты для достраивания ДНК, и по очереди добавляют нуклеотиды. Когда происходит включение нуклеотида в цепь, выделяется квант света. Над подложкой расположена камера, которая фиксирует эти вспышки.

В итоге получается большая серия фотографий, которые идут последовательно друг за другом. Важно при этом разрешение камеры. Поначалу миниатюризация процесса ограничивалась разрешением камеры, которое было на тот момент в распоряжении.

По фотографиям можно для каждого шартика определить нуклеотидную последовательность. С одной подложки можно прочесть $20 * (10^9)$ шариков, в каждом шарике по 50 букв, итого $100 * (10^9)$, то есть на современных приборах можно получить стократное покрытие человеческого генома.

Вся процедура секвенирования на аппаратах нового поколения длится неделю, так как нужно производить много манипуляций связанных с наливанием реагентов, перемешиванием, сливанием и т. д.. Таких циклов может быть около 1000 чтобы прочесть 50 нуклеотидов. 1 запуск прибора обходится за 10 тыс. долларов США, в основном за счет стоимости реагентов.

Есть еще одна интересная технология фирмы **Nanopore sequencing**, которая пока находится на стадии разработки. Она заключается в протаскивании ДНК через мембрану. У разных нуклеотидов в стороны торчат разные функциональные группы. Если на мембране установить что-нибудь подвижное, то оно начинает по-разному колебаться, в зависимости от того, какой нуклеотид проходит через просвет мембраны.

Разработчики этого метода стремились уйти от принципа достраивания комплементарной цепи, чтобы не тратить деньги на реагенты и работать с отдельными молекулами. Этот прибор можно откалибровать по известной последовательности и протаскивать далее все что угодно. Но это пока идея будущего.

Есть еще технология чтения с помощью **сканирующего электронного микроскопа**. Для этого надо ДНК расположить на подложке, вытянуть и пройти по последовательности. При этом кантилевер по-разному подпрыгивает за счет взаимодействия с разными функциональными группами.

В секвенировании по принципу достраивания ДНК тоже происходит удешевление. Например, есть тот же шарик с прикрепленными фрагментами ДНК, и регистрируется не свет (так как светящиеся нуклеотиды дороги), а локальное изменение pH, потому что при полимеразной реакции выделяется протон. То есть, ко дну колодца с шариком нужно подвести микросхемы, которые будут измерять pH. Реакция идет с обычными нуклеотидами, выделяется протон, измеряется pH, и, если он изменился, произошло присоединение нуклеотида к цепи.

На таком приборе за один запуск можно покрыть 10-кратный геном человека. Каждый геном, секвенированный с помощью этой технологии, стоит примерно 500 долларов (эта технология применяется в секвенаторах **Ion Torrent**). В случае работы с бактериями экономия налицо, так как их геном в 1000 раз меньше.