ЛЕКЦИЯ 5

ЧТО ЗАКОДИРОВАНО В ДНК

Лектор показывает чип от Ion Torrent 360 (см. лекцию №4) на примерно миллион колодцев, каждый с pH-метром для регистрации включения нуклеотида в цепь. Длина одного чтения примерно 200 оснований. Если на чипе примерно миллион колодцев, то производительность получается равной 200 млн. нуклеотидов за 3–4 часа работы. Также лектор принес с собой номер журнал Природа (Nature) с интересными статьями.

1. Определение белок-кодирующей последовательности

Отсеквенировав геном, получаем последовательность нуклеотидов, но что же там закодировано? Помогает в понимании этого то обстоятельство, что генетический код по большей части универсален и можно легко транслировать последовательность нуклеотидов в последовательность аминокислот.

Полученную последовательность разбиваем на 6 рамок (считывания, прим. автора): с первого нуклеотида триплеты переводим в аминокислоты, затем, сдвигая рамку считывания, переводим в аминокислоты триплеты со второго нуклеотида и далее с третьего. Проделываем то же самое с другого конца последовательности (см. рис.5.1). Получается 6 вариантов длинных аминокислотных последовательностей, которые периодически прерываются, так как на пути встречаются **стоп-кодоны**.

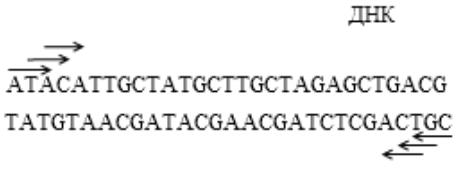

Рис. 5.1: Рамки считывания на ДНК

Очевидно, что прерываться последовательности будут в разных местах, поэтому и куски последовательностей по размерам получатся разные. Важно, чтобы в кодирующих

последовательностях не было стоп-кодонов, поэтому кодирующей рамкой чаще является самая длинная рамка.

Может возникнуть ситуация, когда в одну и в другую сторону есть длинная рамка (в направлении и 3'-5' и 5'-3' концов ДНК, прим. автора). Для упрощения ситуации рассматривают только последовательности с размером превышающим 100 нуклеотидов (100 нуклеотидов — 33 аминокислот, 1 аминокислота примерно 100 Da, тогда масса белка 3,3 kDa, что даже меньше чем, масса самых маленьких белков (рибосомальные — самые маленькие белки — могут весить 5 kDa)).

Среди оставшихся последовательностей необходимо узнать, кодируют ли они белки, и если да, то какие. В случае бактерий все просто, так как сейчас есть 1500 аннотированных полных геномов и примерно столько же частично секвенированных. Частично секвенированные геномы тоже подходят для такого анализа, поскольку, наверняка пробелы отсеквенированы у родственных бактерий, или же пропущенные участки содержат в себе повторы и некодирующие белок последовательности. Сравнивая 2 получившиеся перекрывающиеся рамки с генетическим банком данных с помощью алгоритма Blast, можно понять на какие гены и насколько сильно они похожи.

Консенсусные ситуации (те, при которых последовательность действительно белок-кодирующая):

- Не перекрывается с другими рамками;
- Высокая гомология по Blast;
- **Консервативные домены** соответствуют (если взять родственные белки с одинаковой функцией, то окажется, что у них есть очень сильно похожие участки — консервативные домены). В похожих участках находятся последовательности, которые участвуют в формирование трехмерной структуры белка и его активного центра, и, следовательно, отвечают за его функционирование.
- Наличие/отсутствие **трансмембранных частей**. Хороший инструмент — программа THMM, которая строит профиль последовательности белка с учетом вероятности наличия трансмембранных частей, частей расположенных кнаружи и ковнутри от мембраны. Мембрана — липидный бислой, в котором расположен белок. Внутренняя и внешняя слои мембраны отличаются друг от друга по химическому составу. Поэтому есть некоторые участки белка, которые «любят» находиться:
  - во внутренней мембране,
  - в пространстве внутри мембраны (гидрофобные участки),
  - во внешней мембране в интерфейсе с наружной средой клетки,
  - участки, которые находятся внутри цитоплазмы или снаружи клетки.

  Можно оценить свою последовательность аминокислот с помощью большой базы данных известных трансмембранных участков. При этом важно учитывать не только схожесть по аминокислотному составу, но и наличие доменов, за счет которых белок удерживается в мембране).

2. Описание белка

Следующий этап работы с белком — его описание. Если белок обладает ферментативной активностью, то необходимо присвоить ему номер из четырех чисел: 1.1.2.3 **(enzyme comission (EC))**, где каждая цифра обозначает класс, подкласс, подподкласс и явного представителя, соответственно. Этот номер указывает на реакцию, в которой участвует фермент и на его субстраты.

Идентификатор **COG** (так называемые, **кластеры ортологичных групп**) — это тоже число. Белки со схожей функцией и последовательностью аминокислот относятся к одному COG, например, COG002. Если получен белок, который не похож на уже изученные и идентификатор ему присвоить не удается, то для создания нового COG необходимо как минимум 3 белка из разных бактерий со сходными функциями и последовательностями.

Работа с эукариотическими белками усложняется тем, что у эукариот происходит сплайсинг. Поэтому нельзя просто транслировать геном по рамкам. Облегчает работу то, что в ДНК по краям интронов и экзонов есть известные специфические сигналы для фермента, который вырезает интроны и соединяет экзоны. К плюсам в работе с эукариотами можно отнести тот факт, что белки эукариот не сильно видоизменяются (не понятно, речь об эволюционных изменениях или посттрансляционных модификациях, прим. автора), в отличие от белков прокариот.

3. Проблема увеличения количества данных в ГенБанке

Основной проблемой, сопряженной с ростом количества данных, считают ошибки аннотирования. Например, при публикации гена в ГенБанке, дали название и присвоили EC, но с ошибкой. Новые исследователи, опираясь на неправильные данные, свои последовательности называют также неправильно. Таким образом, из-за случайных ошибок при общем росте количества информации, происходит увеличение количества неверно аннотированных белков.

Такие ошибки выявляются после анализа строения и функций этого белка. Например, АТФ-аза — это сложный белковый комплекс, состоящий из десятков субъединиц внутри и снаружи мембраны. Для удобства, гены, кодирующие эти субъединицы расположены в ДНК друг за другом. Ферменты, участвующие в одном метаболическом пути, тоже расположены друг за другом. Поэтому, окружение гена может помочь в сомнительных ситуациях и в обнаружении ошибок в базе данных.

4. Предсказание функций белка с помощью сигнальных последовательностей

Для компартментализации белка в клетке, в составе его последовательности есть небольшой пептид, который отвечает за взаимодействие с каким-то ферментом и перенос в определенное место. Например, у всех мембранных белков такой сигнальный пептид будет одинаковым. Потом он удаляется. Наличие такого сигналинга в клетке является

большим подспорьем в предсказании функций белка.

5. Идентификация с помощью масс-спектрометрии

Для начала необходимо выделить все белки из клетки и с помощью метода **shot-gun** разрезать их на короткие пептиды. Один из ферментов, который режет белок — трипсин — обладает высокой специфичностью и разрывает пептидную цепь после положительно заряженных аминокислот (лизина и аргинина). Получаются фрагменты длиной 5–20 аминокислот (500–2000 Da) и это удобно для проведения масс-спектрометрии. Вообще с помощью этого метода можно измерять белки с массой до 5 kDa.

Современные масс-спектрометры измеряют массу иона в вакууме, поэтому пептид для измерения нужно «испарить и зарядить».
Метод ионизации **ESI** (2001) (electrospray ionization электро спрей ионизэйшн).

Из отрицательно зараженной иглы впрыскивается через положительно заряженные «ворота» капли раствора с пептидными фрагментами, окруженными водной оболочкой. Далее вода испаряется, площадь поверхности капли с сконцентрированными на ней зарядами уменьшается. Заряд на единицу площади становится больше, вступают в силу отталкивание одноименных зарядов и капелька делится на 2, потом еще на 2 и т. д. В итоге получается 1 ион-пептид, с зарядом, полученным от свойств самих аминокислот или от присоединенных протонов из воды.

Поведение иона в электрическом поле будет определятся отношением его массы к его заряду. Далее можно пропускать ионы в постоянном электрическом поле и смотреть, какой из них придет быстрее к фотоэлектроумножителю. Если известны количества ионов и их массы, то можно получить график по времени, где каждый пик соответствует своему типу иона. Откалибровав этот график по массе, можно запускать интересующие ионы и по положению пика судить об отношении их массы к заряду.

Для того чтобы найти из отношения массу помогает период полураспада. В природе в небольших количествах есть изотоп C^{13}, во всем живом он содержится в известной пропорции (1%). Рассмотрим полипептиды, содержащие по 100 атомов углерода. Есть вероятность того, что в пептиде не содержится ни одного C^{13}. Также есть вероятность того, что есть пептиды, в которых по одному атому C^{13}, по два C^{13} и т. д. Чем больше углеродов в пептиде, тем ниже вероятность, что в нем нет ни одного C^{13}. Картинка, которая получается с фотоумножителя (см. рис. 5.2), будет содержать пики, и первым будет пик, который соответствует ситуации, когда в пептиде нет ни одного C^{13}. Его относительная высота будет падать с увеличением количества углеродов в пептиде. После этого пика следуют другие, соответствующие одному C^{13} в пептиде, двум C^{13}, трем, и т. д.

Зная, что разница между этими пиками по массам в 1 Da можно определить заряд интересующего пептида. Это метод моноизотопного расщепления.

Квадруполи (лат. «4 палочки») способны улавливать и удерживать большое количество ионов. Изменяя настройки масс-спектрометра можно заставить ионы двигаться по какой-то траектории, а также удерживать или выбрасывать наружу только ионы с определенным отношением m/z. То есть можно выделять ионы с определенной массой, что понадобится для второго этапа shot-gun. Возьмем 1000 пептидных ионов одной массы

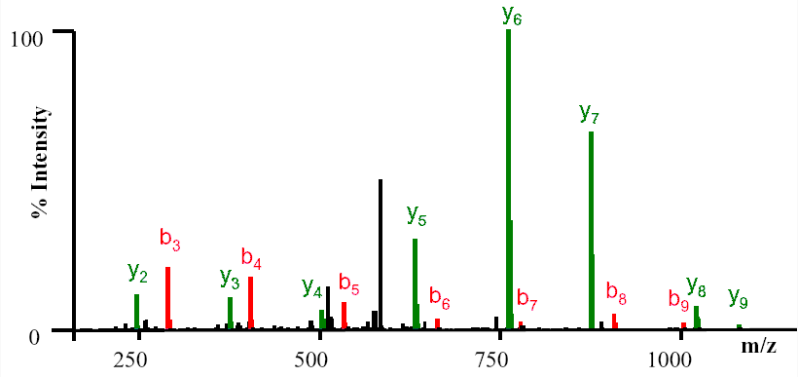

Рис. 5.2: График интенсивности от m/z

($m_{родительская}$) и будем бомбардировать их инертным газом (collision-induced dissociation, диссоциация за счет соударения), так, чтобы самые слабые связи разорвались.

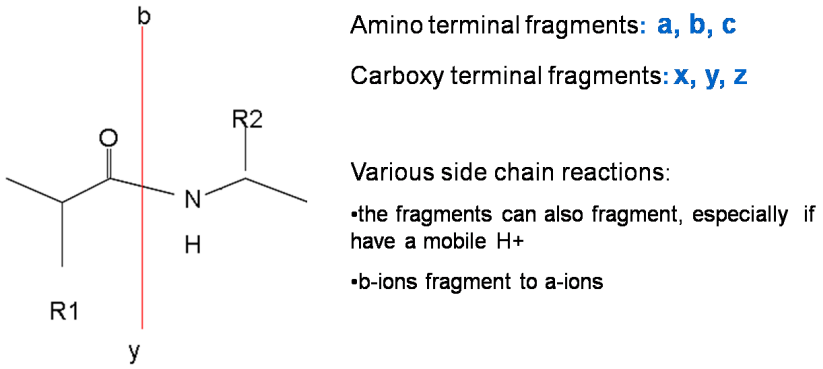

Рис. 5.3: Номенклатура пептидных ионов

Чаще пептид рвется между карбокси- и аминогруппами, но есть вероятность разрыва любой связи. Есть специальная номенклатура получаемых ионов (см. рис. 5.3). Из 100 одинаковых пептидов может образоваться 200 разных ионов. Получится некий спектр зависимости m/z от концентрации определенного иона. Чем выше пик, тем больше вероятность разрыва пептида по определенной связи. Часть пиков соответствует y-ионам, часть — b-ионам, часть — ионам, образовавшимся в ходе разрыва a–x или c–z связи. Но преобладать будет набор из y– и b–ионов. По такому спектру можно определять состав пептида.

Могут возникать сложности с большим количеством вариантов таких составов или с наличием a-, x-, c– и z–ионов.

Таким образом, можно секвенировать один пептид, проводить несколько анализов, но для массового применения этот метод пока не подходит.

6. Идентификация белка по базе данных

Другой способ определения белка — поиск соответствий по базе данных. Например, есть секвенированный геном с аннотированными белковыми последовательностями и

получен спектр фрагментации белка того же организма. Очевидно, что все выделенные белки из этого организма были синтезированы на основе его генома.

Для фрагментации белка используют специфический фермент (трипсин), который режет после лизина и аргинина. При ферментативном гидролизе трипсин может пропустить сайт разрезания (Missed cleavage site)или разрезать в неспецифичном месте (No missed cleavage site), когда, например, один трипсин может разрезать другой и тем самым нарушить его правильную работу. Процент неспецифичности колеблется в переделах от 1 до 5%, остальное зависит от подобранных условий протекания реакции гидролиза.

Из белковой базы данных составляют список всех триптических пептидов, то есть всех пептидов, ограниченных лизинами и аргининами. Из 1000 белков получается примерно 25000 пептидных последовательностей. Можно составить по списку триптических пептидов наборы, которые по массе соответствуют $m_{родительская}$. Чаще всего вариант будет только один, все зависит от того насколько точно измерена родительская масса.

Если все же вариантов несколько, нужно обратиться к спектру фрагментации белка. Для всех подходящих вариантов наборов строим теоретические спектры фрагментации белка и сравниваем с реальным спектром (см. рис. 5.4). Наиболее подходящий спектр

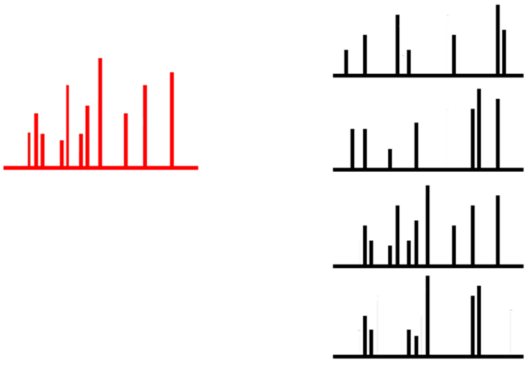

Рис. 5.4: Сравнение спектров фрагментации

соответствует более правильной последовательности белка.

Биоинформатическая задача состоит в том, чтобы придумать такой алгоритм, который мог бы хорошо оценивать, насколько схожи спектры фрагментации между собой. Можно оценивать в процентах или в баллах (scoring).

Проверка алгоритма на реальных данных.
Для известной смеси пептидов составляют экспериментальные и теоретические спектры, запускают алгоритм и по его результатам подправляют его.

Существует несколько наиболее популярных алгоритмов: Sequest, Mascot (Matrix Science) (платная), OMSSA (NCBI), X!Hunter (Global Protein Machine Organization).

Не обязательно восстанавливать последовательность аминокислот в белке полностью. Все полученные варианты белковой последовательности могут иметь биологическое значение при поиске в базе данных, потому что для базы данных используются только те белки, которые есть в геноме.

Это высокопроизводительный метод (удобнее, чем выстраивание комбинаций для каждого спектра). Его минус в том, что базы данных не полные и содержат ошибки.

Кроме того, при поиске можно учитывать посттрансляционные модификации белка (навешивание фосфора, сахаров, окисление метионина, наличие двойной связи между

углеродами цистеинов (возможно, имеются в виду дисульфидные мостики между цистеинами, прим.автора)).

Пример: экспериментально получено 10000 спектров, для каждого из базы данных нашли по 10 возможных последовательностей, и каждой присвоили какое-то количество баллов (scoring). Распределение баллов при достаточном количестве спектров стремится к бимодальному. Это распределение хорошо объясняется тем, что есть истинные пептиды, которые реально были в смеси и получили самые высокие баллы и есть ненастоящие пептиды, которые тоже набрали какие-то очки. Эти 2 пика распределения полностью друг от друга не расходятся. Нужно выбрать такой порог, чтобы правее него находилось не очень много неправильных идентификаций, например не больше 5%. Это есть пороговый скоринг.

Для того чтобы считать, что белок был идентифицирован нужно найти как минимум 2 его пептида с 95%-ной вероятностью. Тогда вероятность того, что он идентифицирован верно, равна $1 - (0.05^2)$ ($>$99%).

Но у двух гомологичных белков может быть один и тот же пептидный состав, поэтому сложно определить, какая из этих изоформ реальна. Тогда необходимо найти тот пептид в смеси, который есть только у одной изоформы, и, если он найдется, то это как раз та, реальная изоформа белка. Может быть и так, что они обе есть, просто уникальные пептиды для одной изоформы найдены, а для другой — нет.

7. Кодирующие участки человеческого генома

Считается что лишь 5–10% генома человека кодирует белки. Из них экзонами является менее 1%. Какие функции выполняют остальные 90–95%? Или это генетический «мусор»?

Проект Encode: было взято около 100 клеточных линий и они были исследованы с помощью большого количества экспериментальных методик: секвенирование РНК, определение участков эухроматина, связывание транскрипционных факторов, сшивки взаимодействующих участков ДНК; то есть, исследовались все возможные взаимодействия ДНК с самой собой и другими белками. Экспериментально подтверждено, что 80% человеческого генома участвует в регуляции. Остальные 20% могут быть выключены в этой ткани, так как эксперименты проводились на дифференцированных клетках.

С эмбриональными же клетками работать пока не умеют. А в них могут быть активны совсем другие регуляторные гены. Еще надо понимать, что культуры человеческих клеток, с которыми проводятся исследования, часто имеют раковое происхождение, поскольку их легче всего культивировать.