ЛЕКЦИЯ 7

ВИЗУАЛИЗАЦИЯ ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ

Все содержание этой и предыдущей лекции описано почти целиком в журнале Nature Methods за 2010 год.

1. Инструменты визуализации

(Продолжение лекции о визуализации экспериментальных данных.) Целью этой лекции является представление основных методов визуализации данных, ознакомление с принципами работы того или иного инструмента для применения и лучшего понимания какой-либо области. На прошлой лекции говорили:

- об основных каналах восприятия информации (цвета, формы, размеры);
- про визуализацию информации, связанной с геномами (от целого генома к отдельным буквам и показателям самого прибора);
- о принципах организации UCSC Genome Browser, который визуализирует данные о человеческом геноме;
- о том, что для сравнения геномов можно пользоваться точечной диаграммой выравнивания;
- про многомерную визуализацию и тепловые карты;
- договорились подумать про GC-skew и почему в кольцевом геноме бактерий в первой половине все гены закодированы на одной цепи, а во второй — на обратной;
- про Circos, про то, как на круговой диаграмме можно отобразить большое количество информации;

- про визуализацию в масс-спектрометрии по трем параметрам (время, интенсивность и масса);
- про карты плотности и их наглядность, про метод главных компонент;
- о том, что многопараметрические данные можно отображать разным образом, например, с помощью метода главных компонент, сокластеризации паттернов, тепловой карты;
- о том, что удобно при множественных выравниваниях добавлять цвет;
- про построение филогенетического дерева, о том, что расстояние на дереве показывает, например, разницу в нуклеотидном полиморфизме.

2. Метаболические пути

Метаболическая реакция — это превращение в организме веществ с высоким порогом активации. Порог активации понижается с помощью катализаторов (в неживой природе) или ферментов (в живой природе). В итоге организм получает выигрыш в энергии или строительные материалы. **Метаболический путь** состоит из метаболических реакций и отражает путь превращений простейшей молекулы в клетке. **Карта метаболизма** представляет собой разветвленную сеть метаболический путей организма, из произвольной точки которой можно перейти в любую другую точку.

3. Cytoscape. Описание работы, примеры

Под элементы сети (кружочки и стрелочки) можно закодировать любые элементы и Cytoscape визуализирует ее. Плюс этой программы в том, что она связана с биологией и ее постоянным развитием занимаются несколько институтов биологического направления. У Cytoscape есть множество плагинов, которые помогают наносить данные на уже существующие сети, также она может взаимодействовать с различными базами данных.

В метаболических картах Cytoscape кружочками изображены не метаболиты, а ферменты, а стрелками обозначается их связь в каком-то процессе. Цветом может быть помечен уровень экспрессии этого фермента в тканях. (Посмотреть на такие сети можно на сайте `http://cytoscapeweb.cytoscape.org/demo`). Можно изменять способ подачи информации через другие графики, диаграммы и пр.

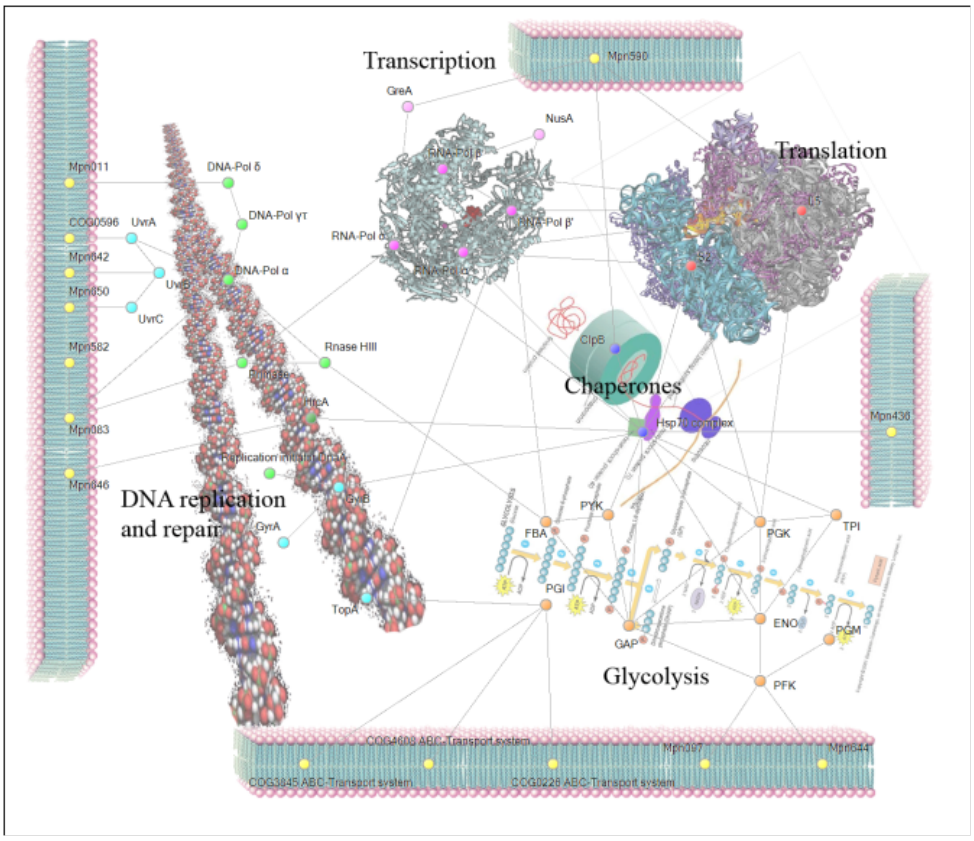

Рис. 7.1: Наложение данных о взаимодействиях белков на некоторые клеточные структуры и метаболические пути. Точками обозначены белки, их цвет маркирует разные функции, например, голубые обозначают участие белка в гликолизе, желтые — шапероны, зеленые — связанные с транскрипцией белки, темно-голубые связаны с трансляцией, синие — «боковые» белки в клетке, красные — репарация ДНК, фиолетовые — воспроизводство ДНК; стрелками нанесены данные о взаимодействии белков друг с другом.

4. Белковые взаимодействия

Экспериментально экспрессируется в большом количестве один и тот же белок, и он связываются с любыми молекулами в клетке, с которыми вообще может связаться. Далее клетку убивают в мягких условиях, чтобы не разрушить слабые взаимодействия, возникшие между молекулами. Смесь прогоняется через колонку с антителами против интересующего белка. Белок соединяется с антителом и при этом остается связан слабыми взаимодействиями с другими белками. Метод не очень достоверен, но с помощью него многие белковые взаимодействия можно так определить.

На слайде 7.1 видно, что большое количество точек связаны между собой. То есть, белки образуют комплексы, образуя одну большую машину. Для того, чтобы посмотреть в каких процессах участвуют некоторые известные белки в клетке, можно в Cytoscape ввести их названия, дальше программа свяжется с базами данных белок-белковых взаимодействий и скачает все известные взаимодействия для интересующих белков.

Перспективным направлением визуализации является объединение всех уровней изображения в одном инструменте. Например, легкое на уровне тканей, гистология, клеточ-

ный уровень, трехмерная структура молекул, уровень экспрессии и т. д. объединенные в одном месте.

5. Статистика в биологии

Всегда, когда проводится эксперимент, *исследователь имеет дело не со всей совокупностью потенциальных испытуемых, а только с выборкой.* То есть, надо понимать, что любые экспериментально полученные данные — это данные о какой-то выборке, и на основе нее ученые экстраполируют какое-то явление на всю совокупность. Например, есть общее количество людей в России, и исследуется N пациентов из общего числа M, или для эксперимента взяты N бактерий из числа M, или N молекул из M; короче говоря, всегда выбирается малое количество из совокупности. Естественно, что по-другому, охватывая всю совокупность, эксперимент поставить нельзя. Нельзя, например, проверить абсолютно все клетки *E.coli* на устойчивость в гипертоническом растворе. Вопрос в том, сколько нужно взять представителей, чтобы сделать правильный вывод об общем числе, то есть, нужно определить размер выборки.

Основные понятие статистики:

- значение **генеральной совокупности** (общее число представителей);
- **выборка** из генеральной совокупности (всегда больше 30);
- **доверительная вероятность** (интервал, %), которая с заданной вероятностью покрывает заданное значение;
- **ошибка выборки.**

Например, взяли 10000 клеток *E.coli*, применили к ним 20 мМ NaCl и выяснили, что 7000 умерли. Посчитали ошибку выборки (допустим ± 200), доверительный интервал равен 95%. И далее, обобщая все эти данные, говорим, что при воздействии 20 мМ соли с 95%-ной вероятностью 70% (± 2 %) клеток *E.coli* умрут. Очевидно, остается какая-то вероятность наблюдения отклонений в генеральной совокупности. Вопрос: почему, наблюдая за выборкой, можно сказать что-то достоверное о совокупности? В основе нашего мира лежит закон о правильных, нормальных и равномерных распределениях, и какую бы физическую величину мы не взяли, обычно ее распределение нормально (см. рис. 7.2)(**Центральная предельная теорема**, прим.автора).

Например, нормально распределение роста человека. Но если взять 10 человек и построить на их основе распределение, то насколько точно оно будет соответствовать реальному контуру распределения?

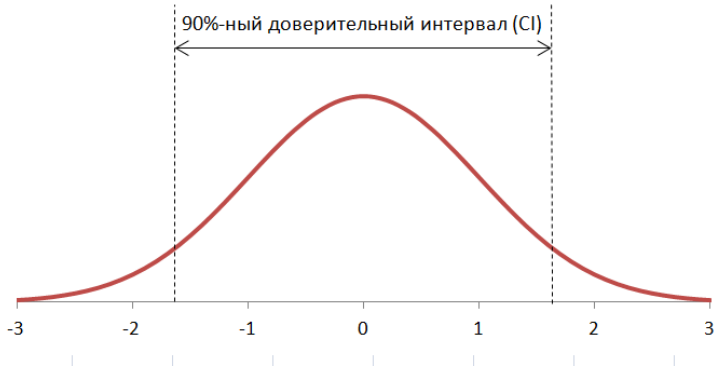

Рис. 7.2: График нормального распределения величины

6. Неправильная выборка

Описан случай, когда в Америке испытуемых обзванивали по телефону и на основе этого делали какие-то выводы обо всех американцах. Очевидно, что это неправильная статистика, поскольку на звонок ответили только те, кто в этот момент находился дома. Также нельзя судить о возрастном составе города, просто стоя на одном месте и отмечая возраст прохожих. В экспериментах с культурой клеток нужно брать материал не так, как удобно (например, с поверхности, потому что в этом случае мы захватим больше клеток, каким-то образом больше стремящихся к поверхности), а максимально случайно. Для того чтобы делать какие-то выводы о совокупности нужно случайным образом организовать выборку.

В биологии часто используются **процентили**. 25-ая процентиль — это линия, которая отсекает 25% всех значений, 50-ая — 50% (она же медиана) и т. д.

7. Критерий Стьюдента

Критерий Стьюдента используют для оценки достоверности различия. Но нужно иметь ввиду, что критерий Стьюдента работает только тогда, когда сравниваются между собой 2 группы. И если есть 3 группы, то нельзя между собой попарно измерить критерий Стьюдента и делать на основе этого выводы. Для этого есть многомерные критерии, которые учитывают распределение трех величин одновременно.

8. Коэффициент Пирсона

Этот коэффициент говорит о том, как 2 переменные зависят друг от друга. Он вычисляется через средние величины.

9. Большие массивы данных

Метагеном кишечника человека содержит 10 трлн. бактерий, у каждой бактерии геном состоит из примерно 5 млн. п. н.. На каждого человека читают только 5 млрд. оснований. Вопрос в том, какую часть от общего числа составляет бактерия, которую

зафиксировали в эксперименте.

10. Ложноположительные определения

Касаясь протеомики, речь заходила о **ложноположительных определениях**. Например, из смеси, состоящей из *E.coli* и *B.subtilis*, нужно определить клетки кишечной палочки. Для начала задаем вопрос: «Это клетка *E.coli*?». Далее можно иметь 4 варианта исходов:

1. Это *E.coli*, определенная как *E.coli* — true positive (TP, правильный положительный ответ на вопрос);

2. Это *B.subtilis*, определенный как *E.coli* — false positive (FP, неправильный отрицательный ответ на вопрос);

3. Это *E.coli*, определенная как *B.subtilis* — false negative (FN, неправильный положительный ответ);

4. Это клетка *B.subtilis*, определенная как *B.subtilis* — true negative (TN, приавильный отрицательный ответ).

Идеальная ситуация, когда TP+TN=100Если при определении пептида есть какой-то процент FP (см. рис. 7.3), то этот процент увеличивается при переходе к определению белка. Чтобы добиться низкого FP, часто жертвуют FN.

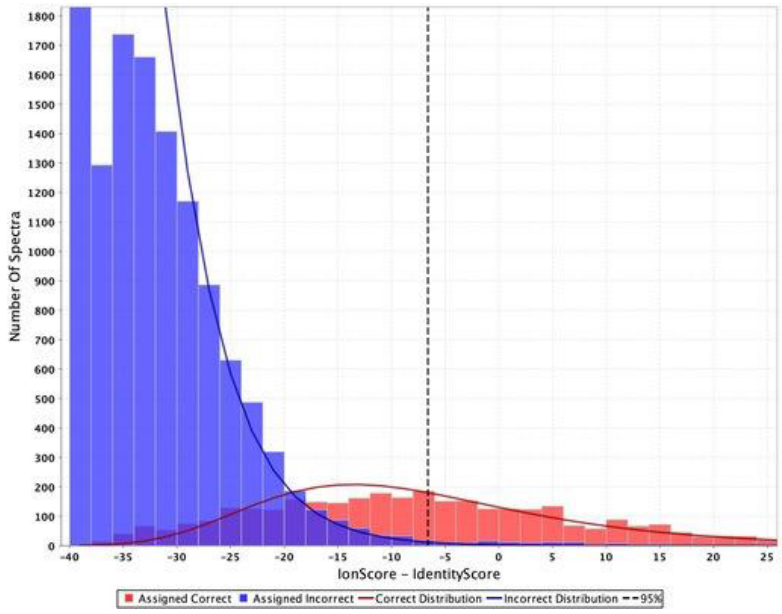

Рис. 7.3: Распределение, отражающее пересечение TP и FP

11. Задача про ЕГЭ

Студентам было предложено построить распределение школьников по баллам, набранным ими на ЕГЭ части А. Теоретически распределение должно получиться нормальным, но так как есть люди, которые знают материал плохо, распределение получается бимодальным. При этом одна вершина в районе 60 баллов соответствует ученикам, которые ставили ответы наугад, а вторая вершина в районе 95 баллов — ученикам, подготовившимся к экзамену. И есть некое пороговое значение между этими вершинами, которое соответствует школьникам либо ошибившимся, но относящимся к подготовленной группе, либо особо удачно ставящим ответы наугад. Также могут быть резкие пики в районе максимального балла или минимального проходного балла, которые соответствуют списывающим ученикам.

Наблюдаются такие же распределения голосов на выборах. Это связано с тем, что человек не может выбрать совершенно случайное число и выбирает что-то преимущественно одно.

12. P-value

В биологии часто используется **P-value**, или **E-value** (взаимозаменяемые понятия). Это значение отражает вероятность того, что наблюдаемое явление случайно.

Например, при измерении экспрессии генов в раковых клетках установлено, что 3 гена, которые находятся рядом на хромосоме, имеют уровень экспрессии выше порогового. Это может быть как случайное, так и значимое явление. Как оценить E-value этого события? Необходимо взять нормальное распределение экспрессии всех генов в клетке и присвоить интересующим нас генам случайные значения экспрессии из этого распределения. Проделав эту операцию 100000 раз, посмотреть сколько раз получается такая ситуация, когда у трех рядом лежащих генов экспрессия выше порогового значения. Многое зависит от порогового значения, но если оно подобрано верно, то такой ситуации ни разу не встретится. Например, P-value= 10^{-10}, то есть 1 раз на 10 млрд. встретится такое событие и, следовательно, оно не случайно. Такое значение P-value очень хорошо для значимости эксперимента. Этот показатель — статистическое обоснование того, что событие значимо.

Если взять низкий порог, выше которого будет 90% генов, то P-value не будет значимым и всегда найдется 3 гена, расположенных рядом с экспрессией выше порога.

Некоторые выводы из лекции:

- Если распределение какой-то величины получается не нормальным, необходимо понять почему;
- Выборку для хорошей значимости результата необходимо делать хотя бы равной 30;
- Любой результат эксперимента — это всего лишь вероятность в 95% случаях попасть в какой-то интервал от заданного числа;
- Если сравниваются больше двух выборок, то нельзя использовать критерий Стьюдента. Для этого разработаны другие многомерные критерии.