Лекция: **Статистический анализ микрочиповых данных**

Слайд 1.

Слайд 2. Обработка данных, полученных методом ДНК-биочипов

Обработка включает в себя получение исходных данных – сканированные изображения чипов после гибридизации, матрицы измерений – оцифрованные по специальным алгоритмам распознавания значения сигналов, и матрицы данных экспрессии генов, по которым производится дальнейший анализ.

Слайд 3. Фрагмент матрицы данных

Матрица состоит из строк и столбцов. Строки – профили экспрессии генов. Общее число профилей – **23232**. В данном фрагменте первые три столбца – номера проб на чипе, следующие два – текстовые идентификаторы профилей и генов, следующие четыре – числовые значения профилей (два – значения сигналов в красном свете (Cy5) – образец и два – в зеленом (Cy3) – контроль).

Слайд 4. Визуализация исходны данных

Простейший способ визуализации – построение графика зависимости значений одного столбца от другого по всем профилям. На графике видно, что бОльшей экспрессии генов по одному столбцу, как правило, соответствует и бОльшая экспрессия по другому, как и должно быть. Видно также, что значения распределены крайне неравномерно. Основная масса значений экспрессии сосредоточена в области малых значений, а некоторые отклоняются очень сильно. Поэтому практически всегда значения экспрессии логарифмируются по основанию 2, т. е. данные преобразуются по формуле: $y=log_2(x)$**.** Этот этап называется **предобработкой**. Визуализация применима на всех этапах анализа микрочиповых данных.

Слайд 5. Этапы анализа микрочиповых данных

1. Предобработка
2. Нормализация
3. Нахождение дифференциально экспрессирующихся генов
4. Многомерный анализ

Слайд 6. Предобработка: логарифмирование

Слайд 7. Визуализация логарифмированных данных

Тот же график после логарифмирования. Видно, что распределение точек более равномерно. Видно также, что зависимость между значениями нелинейна. Если повернуть график с помощью преобразования

Слайд 8. Поворот

$$\mathbf{M = \log_2 R - \log_2 G}$$

$$\mathbf{A = \log_2 R + \log_2 G}$$

Слайд 9. Локально взвешенная (lowess) регрессия

… и рассчитать эту зависимость, например, с помощью локально взвешенной (lowess) регрессии, то, вычитая ее из каждого значения,

Слайд 10. Локально взвешенная (lowess) регрессия - результат

… получим следующий график, на котором уже нет смещения.

Примечание: во многих работах используется преобразование

$$\mathbf{M = log2R - log2G}$$

$$\mathbf{A = (log2R + log2G)/2}$$

Разница несущественна.

Слайд 11. Нахождение дифференциально экспрессирующихся генов

Если отсечь 95%-ми или 99%-ми границами отклоняющиеся значения, то мы таким способом обнаружим дифференциально экспрессирующихся гены. Значения выше верхнего пунктира соответствуют высокоэкспрессирующимся генам, ниже нижнего – низкоэкспрессирующимся. Сильные отклонения могут быть просто выбросами, каждый случай требует отдельного анализа.

Слайд 12. Обратный поворот

Обратное преобразование

$$\boldsymbol{\log_2 R = (M+A)/2}$$

$$\boldsymbol{\log_2 G = (A-M)/2}$$

Слайд 13. Нормализация (логарифмический масштаб)

…приводит к прежним, но уже скорректированным на lowess-регрессию, переменным. Все эти преобразования в идеале должны привести к переменным, у которых нет систематического смещения и которые более или менее соответствуют нормальному (гауссову) распределению. Поэтому они называются нормализацией.

Часто в нормализацию включают еще и нормировку, но мы ее рассмотрим отдельно. Иногда полученные переменные еще и потенцируют ($y=e^x$), чтобы вернуться совсем к первоначальным, но скорректированным значениям экспрессии, и в таком виде они хранятся в базах данных, но при обработке их все равно приходится логарифмировать.

Слайд 14. Многомерный анализ: фрагмент матрицы данных

Многомерный анализ применяется, когда рассматривается одновременно несколько образцов, анализируемых на одном и том же типе чипов. Обычно их несколько десятков. Образцы могут быть взяты в различные моменты времени или в различных состояниях ткани, органа, организма или у разных организмов. Выявление групп коэкспрессирующихся генов используется для нахождения генов, имеющих общие функции. Различия в экспрессии генов между контрольными и экспериментальными (или болезненными) состояниями могут указывать на гены, функционально связанные с отклонениями от нормы. После установления такой связи микрочиповые данные могут использоваться для диагностики образцов неизвестного типа, что открывает широкие перспективы для применения этой технологии в клинических целях. Таким образом, есть две задачи: а) нахождение похожих друг на друга профилей; б) поиск профилей с резко различной экспрессией на разных группах образцов.

В приведенном примере матрица состоит из строк и столбцов. Строки – профили экспрессии генов, столбцы – образцы (пробы) (Borovecki et al., 2005). Файлы с данными этого эксперимента были взяты из базы данных GEO (Barrett T. et al., 2005; http://www.ncbi.nlm.nih.gov/geo/). Общее число образцов – **31**, профилей – **22283**. В данном фрагменте первые два столбца – текстовые идентификаторы профилей и генов, следующие восемь – числовые значения профилей. Четыре из них относятся к образцам, взятым у здоровых индивидуумов, следующие четыре – у больных болезнью Хантингтона.

Слайд 15. Стандартизация: центрирование и нормирование

Если мы умножим значения любого количественного признака на любую ненулевую константу и прибавим к ним любую константу, то это никак не изменит относительных расстояний между объектами по этому признаку. Поэтому мы можем использовать преобразования сдвига и масштаба для приведения разных признаков в соответствие друг с другом. Преобразование:

$$x_i' = (x_i - x^*)$$

где $x^* = \Sigma x_i'/N$ – среднее значение, называется *центрированием*. После центрирования новое среднее признака равно 0 ($x'^* = \Sigma x_i'/N = 0$). Преобразование:

$$xi' = x_i / s$$

где $s^2 = \Sigma(x_i - x^*)^2/N$ – дисперсия признака (вместо N часто применяется $N-1$), называется *нормированием*. После такого преобразования все признаки становятся безразмерными, а новая дисперсия равна 1 ($s'^2 = \Sigma(x_i' - x'^*)^2/N = 1$).

Нормировать можно не только на сигму, но и на корень из суммы квадратов. Разница только в множителе $\sqrt{N}$**.** Стандартизировать можно как столбцы, так и строки матрицы экспрессии генов.

Центрируем каждую строку матрицы ее средним арифметическим и нормируем на корень из суммы квадратов, соответственно. После такого преобразования среднее по каждой строке равно нулю, сумма квадратов – единице, а скалярное произведение строк равно обычному линейному коэффициенту корреляции Пирсона ***r***, который логично принять за меру сходства между профилями.

Слайд 16. Мера различия - евклидово расстояние

Если мы хотим работать с дистанциями, то за меру различия между профилями можно взять евклидово расстояние между строками, т.е. корень из суммы квадратов покоординатных разностей. После несложных преобразований получим

$$D = \sqrt{\sum (x_i - y_i)^2} = \sqrt{\sum x_i^{\,2} + \sum y_i^{\,2} - 2\sum x_i y_i} = \sqrt{2 - 2r}\,,$$

т.е., евклидово расстояние между строками является в этом случае монотонной функцией коэффициента корреляции. Чем выше коэффициент корреляции, тем меньше расстояние. Если *r=1*, то *D=0*, как и следовало ожидать. Максимальное расстояние *D=2* достигается при *r= -1.* Следовательно, результаты кластерного анализа для методов типа ближайшего или дальнего соседа, где используются только сравнения меры сходства-различия, и при выборе коэффициента корреляции и при выборе евклидова расстояния будут совпадать. Для методов, где вычисляется средняя мера сходства-различия, результаты могут несколько различаться. Однако, если в качестве меры различия выбрать квадрат евклидова расстояния $D^2=2(1-r)$ (но заметим, что он уже не является евклидовым расстоянием), то результаты кластеризации будут совпадать и для усредняющих меру сходства-различия методов

Слайд 17. Термокарты (heatmaps) и иерархическая кластеризация

Простейший способ получить визуальное представление о всей таблице сразу – это ее раскрасить. Раскраска осуществляется следующим образом. Каждому значению таблицы сопоставляется отдельная клетка. Клетка раскрашивается в зеленый (или синий) цвет, если значение меньше среднего, и в красный (или желтый), если значение больше. Причем, чем больше значение по абсолютной величине, тем цвет ярче. В черный (или серый, или белый) красятся клетки, значении в которых близки к среднему.

Но раскрашенная таблица выглядит очень пестро, если ее не упорядочить. Один из самых популярных способов упорядочения – это иерархическая классификация. Иерархическая означает, что каждый класс вложен в некоторый

другой. Самый известный и часто используемый алгоритм иерархической классификации – алгоритм ближайшего соседа. Вначале каждый объект (в данном случае профиль экспрессии) считается отдельным классом. На следующем шаге ищется пара самых близких объектов, которая объединяется в новый класс. Расстояния (или меры сходства) для нового класса со старыми пересчитываются по следующему правилу: расстоянием между классами считается расстояние между ближайшими объектами в этих классах (отсюда и название). Далее все повторяется до тех пор, пока не останется ровно один класс, содержащий все объекты.

Общепринятым способом отобразить иерархическую классификацию является дендрограмма. Объекты играют роль листьев и расположены каждый на своей ветке. Если объекты объединяются в один класс, то и их ветви объединяются в одну, причем длина равна расстоянию (или сходству) между классами. Чтобы дендрограмму можно было нарисовать, объекты надо переставить местами. Если одновременно переставить строки таблицы и термокарты, то результат будет более нагляден. Классифицировать можно и образцы, в этом случае надо переставлять столбцы таблицы и термокарты.

Слайд 18. Термокарты (heatmaps) и иерархическая кластеризация (2)

Данные со слайда 14 – две различных платформы.

Слайд 19. Алгоритм К-средних

На начальном этапе случайным образом выбирается К объектов (К задается исследователем). Они объявляются центрами классов. Остальные объекты разносятся по классам по следующему правилу: каждый объект попадает в тот класс, к центру которого он находится ближе всего. После этого в каждом классе определяется новый центр. Снова все объекты разносятся по классам и так до тех пор, пока процесс не сойдется.

В отличие от иерархической классификации, все классы равноправны и находятся на одном уровне. Есть критерии, позволяющие оценить удачность разбиения на классы. Если разбиение оказалось не очень удачным, К меняется и весь процесс повторяется с другим К. Рекомендуется для начала брать К равным квадратному корню из числа генов, однако это сильно зависит от исследуемого множества.

Слайд 20. Изображение одного кластера

Профили экспрессии, попавшие в один кластер, можно изобразить в виде графика. Поскольку профилей в таблице очень много, то и в каждом кластере их тоже много. Поэтому при их изображении они сливаются, как ручейки в реке. На слайде видно, что в кластер действительно попали профили со сходной экспрессией.

Слайд 21. Изображение нескольких кластеров

На рисунке изображено сразу несколько кластеров. Ясно, что в качестве меры сходства использовался коэффициент корреляции, так как в один кластер попадают профили, сходные по поведению, но различные по масштабу. Ясно также, что кластеры 1 и 3, а также 2 и 5 противоположны по своему поведению, т.е. гены, входящие в эти кластеры, экспрессируются на одних образцах и не проявляют активности на других, а гены из противоположных кластеров – наоборот. Такая информация может оказаться полезной при конструировании генных сетей.

Слайд 22. Самоорганизующиеся карты Кохонена (Self-Organizing Maps – SOM)

Одним из широко используемых методов анализа микрочиповых данных являются самоорганизующиеся карты Кохонена (SOM-анализ) (Kohonen, 1997; Tamayo et al., 1999). Суть метода заключается в нелинейной трансформации множества точек, представляющих, например, профили экспрессии генов, в визуализируемое пространство малой размерности с одновременной кластеризацией этих точек. Под центры будущих кластеров заранее выделяются узлы одномерной или двумерной прямоугольной или гексагональной решетки. Число узлов также выбирается исследователем. На первом шаге каждый узел заполняется следующим образом: с помощью датчика случайных чисел либо генерируются координаты центра, либо выбирается один из имеющихся профилей. Далее запускается итерационный процесс. Для каждого профиля координаты ближайшего к нему центра (*в исходном пространстве*) пересчитываются таким образом, чтобы приблизить его к этому профилю. В этом же пространстве заодно приближаются, хотя и несколько меньше, соседи этого центра *по решетке*. Процесс повторяется заданное число раз. После этого каждый профиль относится к ближайшему центру кластера. Идея состоит в том, чтобы отобразить множество объектов в двумерное или даже одномерное пространство с максимальным сохранением отношения соседства между точками.

Слайд 23. Метод главных компонент 1

В SOM-анализе впервые появляется понятие взаимного расположения точек и кластеров, правда, не в исходном пространстве, а в некотором построенном искусственно. Вообще говоря, ниоткуда не следует, что оно действительно отображает какие-то существенные взаимоотношения между точками и кластерами, хотя на практике этот метод работает довольно успешно. Что касается методов кластеризации, то в них понятие взаимного расположения точек не используется вообще, и это является их большим недостатком.

Если рассматривать каждый столбец матрицы экспрессии как точку в многомерном пространстве, то можно попытаться увидеть их взаимное расположение относительно друг друга. Хотя координат у такой точки – несколько десятков тысяч, но, поскольку число образцов мало, реальная размерность не превышает числа образцов. Существуют способы, позволяющие спроектировать облако точек в визуализируемое пространство меньшей размерности, например, на плоскость, с минимальными потерями. К таким методам относится метод главных компонент. На рисунке показано расположение образцов на плоскости первых двух компонент, на которые приходится ~20% общей дисперсии. Это не очень много, но

вполне достаточно, чтобы увидеть, что все образцы распались на три группы: больных болезнью Хантингтона, предрасположенных к этой болезни и здоровых, причем в группу предрасположенных попало четверо здоровых (назовем их субнормальными). Надо подчеркнуть, что информация о принадлежности пациентов к тем или иным группам в методе главных компонент никак не использовалась, они распались на группы по экспрессии генов.

Слайд 24. Метод главных компонент 2

В методе главных компонент (Principal component analysis – PCA) осуществляется поворот в исходном пространстве точек таким образом, чтобы в проекции на новые оси – главные компоненты – дисперсия всего множества точек была максимальна. Вследствие этого, главных компонент ровно столько же, сколько исходных осей. Однако дисперсия сосредоточена, в основном, в первых компонентах. Поэтому можно рассматривать только их, отбрасывая остальные. Этот метод как разновидность проекционных методов, иногда называют также *singular value decomposition (SVD).* На слайде приведено то же множество образцов в проекции на плоскость, образованную первой и третьей главными компонентами (~17.5% общей дисперсии). Видно, что здесь предрасположенные к болезни находятся между больными и здоровыми, а субнормальные попали в группу здоровых. Никакого противоречия с предыдущим слайдом на самом деле нет, это взгляд на одно и то же множество объектов с разных точек зрения.

Слайд 25. Метод главных компонент 3

Преобразование исходной таблицы в методе главных компонент приводит к тому, что вместе со столбцами преобразуются и строки – профили экспрессии генов. Каждый профиль получает новые координаты на новых осях, которые в этом пространстве называются собственными векторами. Каждому собственному вектору однозначно соответствует главная компонента с тем же номером, а каждому направлению в пространстве собственных векторов – направление в пространстве главных компонент. Поэтому в пространстве профилей экспрессии можно выделить то же самое направление, по которому на предыдущем слайде различаются больные и здоровые. Профили, которые наиболее далеко отстоят от центра (отделены линией), и соответствуют генам-кандидатам. Подтверждением этому служит тот факт, что все 12 генов–маркеров болезни Хантингтона, найденных в работе (Borovecki et al., 2005), оказались именно в этой зоне

Слайд 26. Многомерное шкалирование (Multidimensional scaling - MDS)

Существует еще один метод уменьшения размерности исходного множества точек – многомерное шкалирование. В этом методе каждой точке исходного множества ставится в соответствие точка в пространстве меньшей размерности, чаще всего, на плоскости. Далее точки на плоскости начинают передвигаться таким образом, чтобы матрица расстояний между ними как можно лучше (в смысле некоторого критерия, называемого “стрессом”) соответствовала матрице сходства (различий, расстояний) между точками исходного множества. Из-за вычислительных ограничений этот метод применим к небольшим множествам объектов (около сотни).

На слайде показан результат применения этого метода к той же группе образцов. Видно, что, по существу, есть два направления различий. Одно – между здоровыми и больными вместе (первая ось) и субнормальными и предрасположенными вместе. Второе – между здоровыми и больными (вторая ось) и оно же – между субнормальными и предрасположенными. Поскольку преобразование нелинейно, аналогичное преобразование в пространстве признаков не ищется.

Слайд 27. «Обучение с учителем» (supervised data analysis)

Все рассмотренные до сих пор методы относились к категории “обучения без учителя” *(unsupervised data analysis)*. Информация о принадлежности объектов к группам в расчетах никак не использовалась. Если же эти различия обнаруживались на графиках, то это являлось свойством самих групп, следствием их явной обособленности друг от друга (рисунок слева). Однако возможна ситуация (рисунок справа), когда явной обособленности на самом деле нет, но мы хотим все же научиться различать объекты разных групп, например, в целях диагностики. Для этого мы сначала обучаемся на известных примерах, а потом применяем полученные знания к новым объектам с неизвестной принадлежностью *(supervised data analysis)*. Эта задача не безнадежна, можно выделить области пространства, где встречаемость объектов некоторой группы выше, чем других. Если неизвестный объект попадает в эту область, то мы его и относим к этой группе. Это не гарантирует от ошибки, но минимизирует ее вероятность.

Слайд 28. t-критерий Уэлша

Рассмотрим задачу выделения профилей экспрессии генов, максимально различающихся на двух группах образцов (например, больных и здоровых). График таких профилей может выглядеть примерно так, как в нижней части слайда. Для этой цели можно использовать *t*-критерий Уэлша. Для каждой группы образцов по каждому профилю вычисляются средние и дисперсии. Чем больше разница между образцами, тем больше должен быть этот критерий. Профили с максимальными значениями t-критерия наиболее перспективны для дальнейшего содержательного анализа.

Примечание. В отечественной биометрической литературе произошла некоторая путаница и обычно этот критерий приводится под именем *t*-критерия Стьюдента. Настоящий критерий Стьюдента имеет несколько другой вид. Применять можно оба, но критерий Уэлша имеет некоторые преимущества по сравнению с критерием Стьюдента, например, в нем не требуется равенства неизвестных генеральных дисперсий.

Слайд 29. Дискриминантный анализ 1

Можно искать направления в исходном пространстве точек, в проекции на которые максимально различаются объекты разных групп. Критерием служит отношение межгрупповой и объединенной внутригрупповой дисперсий (критерий

Фишера). Для случая двух групп этот критерий эквивалентен критерию Стьюдента. Число направлений равно числу групп без единицы.

Слайд 30. Дискриминантный анализ 2

Так же, как и в случае главных компонент, этим направлениям однозначно соответствуют направления в пространстве признаков, в нашем случае, – пространстве профилей экспрессии генов. Профили с максимальной проекцией на дискриминантные оси наиболее перспективны в качестве генов-кандидатов. Нахождение такого направления реализуется с помощью дискриминантного анализа (Discriminant analysis - DA).

Слайд 31. Спасибо за внимание

Литература:

Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика: Исследование зависимостей. –М.: Финансы и статистика, 1985. -487с.

И.Гайдышев. Анализ и обработка данных: специальный справочник. СПб: Питер, 2001. 752с.

Дж.Джефферс. Введение в системный анализ: применение в экологии. М.: Мир, 1981. 252с.

Л.А.Животовский. Популяционная биометрия. М.: Наука, 1991. 271с.

Кендалл М., Стьюарт А. Статистические выводы и связи. -М.: Наука, 1973. 899с.

Дж.О.Ким, Ч.У.Мьюллер, У.Р.Клекка и др. Факторный, дискриминантный и кластерный анализ. М.: Финансы и статистика, 1989. 215с.

А.В.Коросов. Экологические приложения компонентного анализа. Петрозаводск: Изд-во Петрозаводского ун-та, 1996. 152с.

М.Уильямсон. Анализ биологических популяций. М.: Мир, 1975. 271с.

Allison DB, Cui X, Page GP, Sabripour M. Microarray data analysis: from disarray to consolidation and consensus // Nature Rewiews|Genetics, 2006, V. 7. P. 55-65.

Alter O., Brown P.O., Botstein D. Singular value decomposition for genome-wide expression data processing and modeling. Proc. Natl. Acad. Sci. USA. V.97(18), pp.10101-10106 (2000).

Barrett T. et al. (2005) NCBI GEO: mining millions of expression profiles--database and tools. Nucleic Acids Res. 33(Database issue):D562-566.

Borovecki F. et al., (2005) Genome-wide expression profiling of human blood reveals biomarkers for Huntington's disease. Proc Natl Acad Sci U S A. 102(31):11023-11028.

de Hoon M.J. et al. (2004) Open source clustering software. Bioinformatics. 20(9):1453-1454.

Hand D.J., Heard N.A. (2005) Finding groups in gene expression data. J Biomed Biotechnol. 2005(2):215-225.

Tamayo P., Slonim D., Mesirov L., Zhu Q., Kitareewan S., Dmitrovski E., Lander E.S., Golub T.R. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. // Proc. Natl. Acad. Sci. USA. V.96, pp.2907-2912 (1999).

Kohonen T. Self-Organizing Maps, Berlin-Heidelberg, 1997, 420pp.

Liu Z, Chen D, Bensmail H, Xu Y. Custering gene expression data with kernel principal components // Journal of Bioinformatics and Computational Biology. Vol. 3. No. 2. 2005. P. 303-316.

Simon R. Using DNA microarrays for diagnostic and prognostic prediction // Expert. Rev. Mol. Diagn., 2003. V. 3(5). P. 587-595.

P. Törönen, M. Kolehmainen, G. Wong, E. Castrén, Analysis of gene expression data using self-organizing maps, FEBS Letters, 451, 142-146, 1999

Wall M.E. et al. (2001) SVDMAN-singular value decomposition analysis of microarray data. Bioinformatics. 17(6):566-568.

Yeung, K. Y., and Ruzzo, W. L. (2001). Principal Component Analysis for clustering gene expression data // Bioinformatics, 17, 763-774.