ПРИМЕНЕНИЕ НЕМЕТРИЧЕСКОГО МНОГОМЕРНОГО ШКАЛИРОВАНИЯ ДЛЯ МУЛЬТИПЛАТФОРМЕННОЙ ОБРАБОТКИ МИКРОЧИПОВЫХ ЭКСПРЕССИОННЫХ ДАННЫХ

В.М. Ефимов, А.В. Катохин

Учреждение Российской академии наук Институт цитологии и генетики Сибирского отделения РАН, Новосибирск, Россия, e-mail: efimov@bionet.nsc.ru

Разработан многомерный метод согласованного мультиплатформенного поиска дифференциально экспрессирующихся генов-маркеров в массивах микрочиповых экспрессионных данных для одного и того же набора образцов. Метод применен для совместного анализа двух массивов микрочиповых экспрессионных данных, полученных с применением платформ CodeLink и Affymetrix (Borovecki *et al*., 2005). Построены списки дифференциально экспрессирующихся генов-маркеров, соответствующих каждой из четырех групп образцов периферической крови (пациенты с болезнью Хантингтона, с предсимптоматикой этой болезни, субнорма, норма).

Ключевые слова: профили экспрессии генов, неметрическое шкалирование, болезнь Хантингтона.

Введение

К числу многообещающих направлений современной биологии относится анализ экспрессии генов с помощью микрочипов. Микрочипы – упорядоченные комплекты фрагментов ДНК или РНК (зондов), иммобилизованные на специальных носителях (пластинках из стекла, пластика или кремния, каплях геля), называемых платформами (вместе с сопутствующими технологиями, как правило, разных производителей). Технология микрочипов позволяет одновременно анализировать экспрессию десятков тысяч генов в нескольких десятках проб (образцов), обычно представляемую в виде матрицы уровней экспрессии, в которой столбцам отвечают образцы, а строкам – гены. Столбцы и строки этой матрицы называются профилями экспрессии соответственно образцов или генов. (Иногда возникает терминологическая путаница: термин «профиль экспрессии генов» в некоторых работах может означать «профиль экспрессии образца по всем генам».).

Одной из существенных проблем, возникающих при анализе микрочиповых экспрессионных данных, является несогласованность результатов, получаемых при анализе одних и тех же образцов на разных платформах. Основная трудность заключается в том, что вследствие различия технологий не удается напрямую сравнивать профили экспрессии генов. Во-первых, наборы генов разных платформ совпадают лишь частично. Во-вторых, даже одни и те же гены разных платформ можно лишь условно отождествить друг с другом из-за того, что на разных платформах иммобилизуются, вообще говоря, несколько разные зонды, хотя и относящиеся к одному и тому же гену (изоформы, клоны), что естественно приводит к различиям в профилях экспрессии. Имеется и ряд других межплатформенных различий, например, в интенсивности гибридизации, уровне «шума», характере случайных ошибок и т. д. (Cheadle *et al*., 2007; Bemmo *et al*., 2008). Тем не менее необходимость объединения результатов, получаемых на разных платформах, остается крайне актуальной.

Существует несколько подходов к решению этой проблемы. Простейший и наиболее часто применяемый способ – это ее обойти. По каждому массиву микроэкспрессионных данных отдельно решается одна и та же задача, например, поиск генов-кандидатов, которые можно использовать в качестве маркеров онкологи-

ческих или иных заболеваний. Далее списки генов-кандидатов сравниваются и выбираются те, которые присутствуют в нескольких списках. Большим недостатком такого подхода является то, что при любых уровнях значимости гены, представленные только на одной платформе, не имеют никакого шанса попасть в общий список. Например, в работе P. Pedotti *et al*. (2008) уровень экспрессии генов в 10 образцах (5 – трансгенные мыши и 5 – дикий тип) измерен с помощью 5 разных платформ, и на 10 %-м уровне значимости по критерию FDR (Benjamini, Hochberg, 1995) отобраны гены с дифференциальной экспрессией на разных группах мышей. Барьер преодолели 4, 130, 3071, 54 и 13 генов от каждой платформы соответственно. Только два гена оказались общими для всех 5 списков и 4 встретились по 4 раза.

Другой способ – ограничиться только одними и теми же зондами одних и тех же генов и состыковать их профили по всем платформам. Недостатки: приходится уделять специальное внимание отождествлению зондов по их описаниям, которые не всегда имеются или достаточны для этого (Barnes *et al*., 2005). Однако эту подзадачу можно решать эмпирически. В работе H.K. Lee *et al*. (2004), например, проанализировано 60 массивов и выбраны пары зондов, корреляция между которыми значима в трех и более списках (максимум 31). Такие зонды считаются совпадающими. Различия в технологиях тоже остаются, что влечет необходимость различного рода предварительных трансформаций. Кроме того, так же, как и в предыдущем случае, серьезным недостатком является то, что при этом теряется значительная часть уникальной информации.

Мы предлагаем новый способ. Для каждой платформы вычисляется матрица различий между профилями экспрессии образцов по всему набору генов. Далее конструируется объединенная матрица различий, по ней методом неметрического шкалирования строится общее для всех платформ евклидово пространство профилей экспрессии образцов малой размерности, выявляются интересующие нас направления в этом пространстве (например между больными и контролем) и из всех платформ выбираются профили экспрессии генов, максимально коррелирующие с этими направлениями.

Материалы и методы

Анализировался массив микрочиповых экспрессионных данных, полученных с применением платформы CodeLink. Массив после процедуры фильтрования (строки с отсутствующими значениями удалялись) содержал 17525 полных профилей экспрессии генов (строки) по 31 образцу периферической крови. Эти же образцы анализировались с применением платформы Affymetrix – 22282 полных профиля. Файлы с данными этих экспериментов извлечены из базы данных GEO (Barrett *et al*., 2005; http://www.ncbi.nlm.nih.gov/geo/). Образцы периферической крови взяты у 12 пациентов с болезнью Хантингтона, 5 – с предсимптоматикой этой болезни (предрасположенных), 14 – здоровых. Болезнь вызывается мутацией по гену, кодирующему белок хантингтин, и проявляется с возрастом. Больные характеризуются наличием мутации и клиническими проявлениями болезни Хантингтона, предрасположенные – наличием мутации, но отсутствием клинических проявлений, здоровые – отсутствием и мутации, и клинических проявлений. Уровень клинических проявлений определялся опытным неврологом по унифицированной шкале болезни Хантингтона, наличие мутации – генетическим тестированием (Borovecki *et al*., 2005). Массивы неоднократно анализировались другими авторами и фактически стали тестовым множеством для проверки различных методов анализа микрочиповых данных (http://www.genesifter.net/web/webinars.php). Оба массива логарифмировались, затем центрировались и нормировались сначала по столбцам для устранения неоднородности по образцам, затем по строкам для устранения эффектов масштаба.

После этого для каждой платформы вычислялась матрица попарных евклидовых расстояний между профилями экспрессии образцов. (Следует подчеркнуть, что метод работает для любых мер сходства–различия, а не только для расстояний, удовлетворяющих аксиомам метрики. Поэтому далее в любом случае будем именовать их различиями, чтобы отличать от евклидовых расстояний, генерируемых методом неметрического многомерного шкалирования.). Межплатформенный коэффициент корреляции между матрицами различий равен 0,61, а

зависимость различий друг от друга оказалась явно нелинейная (рис. 1). Полученные матрицы ранжировались и усреднялись. Средний ранг различий достаточно хорошо воспроизводит исходные различия между образцами по каждой платформе (корреляция с Affymetrix – 0,89, с CodeLink – 0,88).

Матрица средних рангов различий между профилями экспрессии образцов обрабатывалась методом неметрического многомерного шкалирования (Taguchi, Oono, 2005). В этом методе каждому объекту исходного множества ставится в соответствие точка в евклидовом пространстве малой размерности, чаще всего на плоскости. Далее точки в этом пространстве передвигаются таким образом, чтобы матрица расстояний между ними как можно лучше соответствовала матрице различий между объектами исходного множества. За критерий соответствия принят коэффициент ранговой корреляции между расстояниями и различиями, что эквивалентно критерию, приведенному в работе Taguchi, Oono (2005).

Результаты

Все профили экспрессии образцов распались на группы, соответствующие пациентам, страдающим болезнью Хантингтона, предрасположенным к этой болезни и здоровым (рис. 2). Надо подчеркнуть, что имеющаяся информация об изначальной принадлежности пациентов к тем или иным группам при обработке методом неметрического шкалирования никак не использовалась. Тем не менее четверо пациентов из числа здоровых образовали группу, которая достаточно далеко отстоит от других здоровых и примыкает к группе предрасположенных. Это послужило основанием для выделения их в отдельную группу, которая была названа нами «субнормальной». Все четыре группы занимают отдельные области на плоскости.

Для каждой группы вычислены центроиды и направления на них от общего центра тяжести. Для всех профилей экспрессии генов обеих платформ вычислены и отложены на плоскости коэффициенты корреляции с осями неметрического

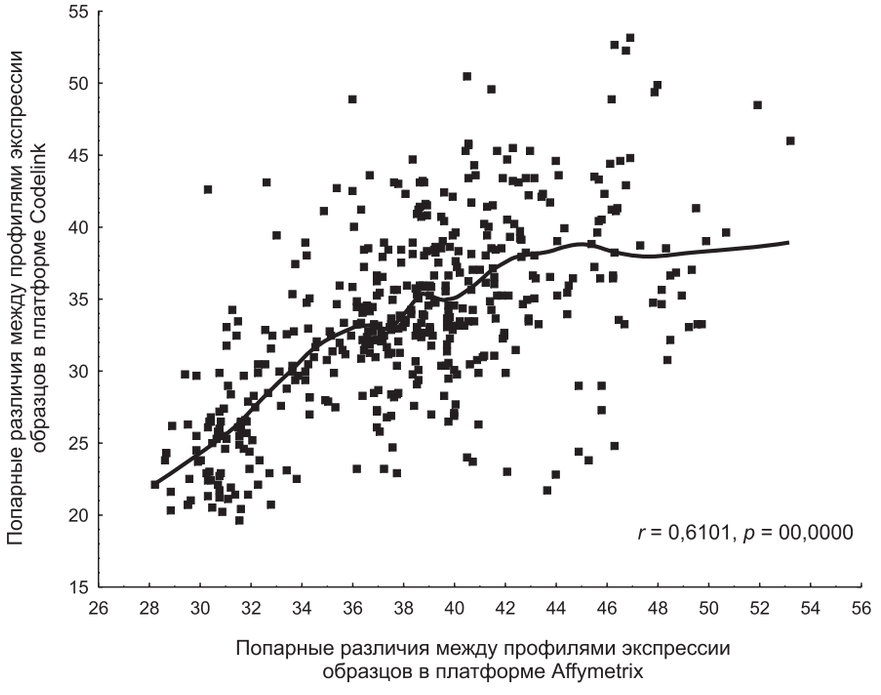

Рис. 1. Соответствие попарных различий между профилями экспрессии образцов для платформ Affymetrix и CodeLink (линия – lowess-регрессия).

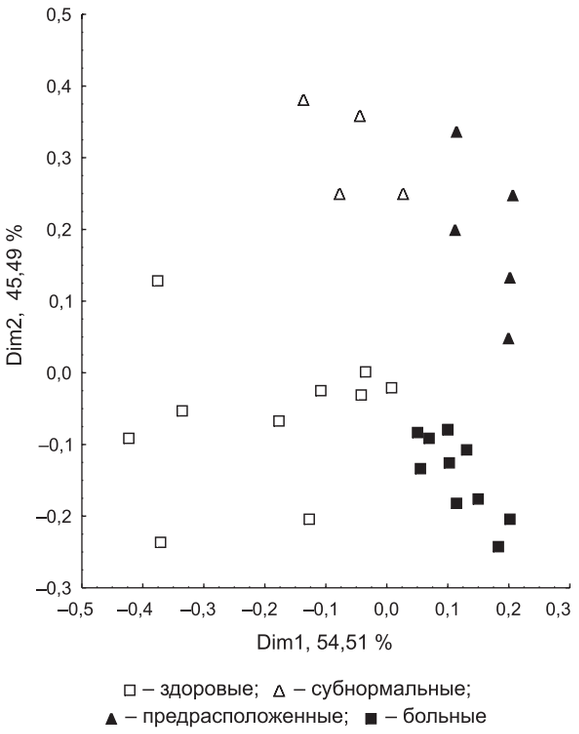

Рис. 2. Расположение профилей экспрессии образцов на плоскости неметрического двумерного шкалирования.

двумерного шкалирования (рис. 3). Каждому направлению на плоскости экспрессии образцов (рис. 2), в частности на центроид любой группы, взаимно однозначно соответствует направление на плоскости экспрессии генов (рис. 3). Профили экспрессии генов, наиболее отстоящие от центра вдоль этого направления, характеризуются самыми высокими значениями на объектах этой группы.

Профили, расположенные на противоположной стороне рисунка (не показано), наоборот характеризуются самыми низкими значениями. Поэтому для каждой группы образцов можно выделить соответствующую группу профилей экспрессии генов. Границу можно проводить из разных соображений. На рис. 3 граница проведена таким образом, чтобы вероятность ее превышения для максимального по модулю случайного коэффициента корреляции из 39807 (22282+17525) была не более 0,05. Следует заметить, что это достаточно жесткий барьер, практически гарантирующий отсутствие случайных профилей экспрессии генов в выделенной области. Тем не менее его преодолели 3670 генов (750 – Affymetrix, 2920 – CodeLink), что является довольно неожиданным. Общими для обеих платформ являются 8600 генов, из них совместно преодолели барьер 241. Крестиками и ромбиками обозначены профили экспрессии генов-кандидатов, выделенных в работе (Borovecki *et al*., 2005), которые, как и следовало ожидать, почти все попадают в выделенную область. Часть генов платформы Affymetrix представлены несколькими изоформами, поэтому некоторые из них оказались разбросаны по всей плоскости.

Расположение групп позволяет высказать гипотезу о возможности последовательного перехода по круговой траектории: здоровые → субнормальные → предрасположенные → больные, причем каждой группе соответствует свой набор профилей экспрессии генов. Метод неметрического многомерного шкалирования позволяет выделить эти наборы для дальнейшего анализа средствами генной онтологии.

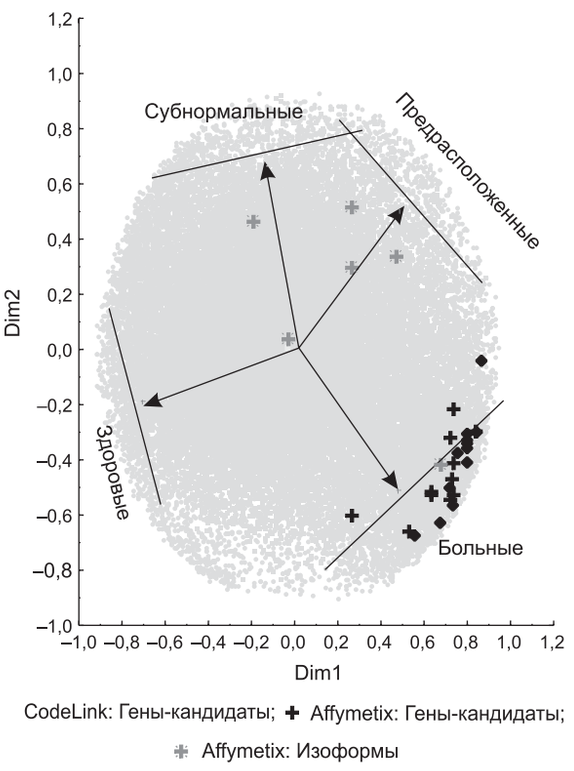

Рис. 3. Расположение профилей экспрессии генов на плоскости коэффициентов корреляции с осями неметрического двумерного шкалирования (Affymetrix + CodeLink).

Обсуждение

В чем состоит идея предлагаемого способа обработки? Схема рассуждений довольно проста. Есть одни и те же образцы, и для них на разных платформах получены профили экспрессии по всем генам. Если эти образцы отличаются друг от друга по экспрессии генов, то эти отличия должны проявляться на каждой платформе. И наборы отличий должны быть хоть в чем-то подобны друг другу, иначе пропадает сама идея применения микрочипов. С другой стороны, поскольку каждая платформа, безусловно, имеет свою специфику, они не обязаны совпадать. Первая задача состоит в том, чтобы выявить то общее, что есть в наборах отличий между профилями образцов для разных платформ.

Метод неметрического многомерного шкалирования оказался исключительно подходящим для решения этой задачи. В этом методе входной информацией служит матрица ранжированных различий между профилями образцов. Мы любым удобным для себя образом определяем различия между профилями образцов по одной из платформ, вычисляем матрицу этих различий, упорядочиваем и заменяем каждое различие его рангом. То есть нас интересует не конкретное значение различия между двумя профилями образцов, а только его ранг среди всех других различий. Естественно предположить, что этот ранг должен быть близок к рангу различия между профилями этих же образцов, вычисленного по другой платформе, т. е. должна наблюдаться корреляция между рангами различий в разных платформах. И действительно, в рассматриваемом случае корреляция и между различиями, и между их рангами составила 0,61. Корреляция не слишком высока, что указывает на значительную разницу между обеими изучаемыми платформами. Но оба коэффициента корреляции одинаковы, и это означает, что переход к рангам не привел к заметной потере информации.

Развивая эту идею дальше, мы можем просто вычислить средний ранг различий для каждой пары образцов по обеим платформам и снова его упорядочить, взяв его ранг. Тем самым мы усилим то общее, что есть в обеих матрицах ранжированных различий, и ослабим специфику каждой отдельной платформы.

Очевидно, что число платформ совершенно непринципиально, так же, как и способы измерения различия. Ранжирование, и в этом его несомненный плюс, все приводит к единой шкале с равномерным распределением, избавляя от проблем, связанных с нелинейностью, разномасштабностью и неоднородностью распределений.

Далее к матрице новых рангов применим собственно сам метод неметрического многомерного шкалирования в варианте Шепарда-Тагучи-Ооно (Taguchi, Oono, 2005). В этом методе с каждым образцом сначала случайным образом сопоставляется точка в евклидовом пространстве малой размерности, например на плоскости. Между всеми точками вычисляется попарная матрица расстояний, которые ранжируются. Если ранг различия между какой-то парой образцов меньше, чем ранг расстояния между соответствующими точками, то эти точки несколько приближаются друг к другу, если больше – то удаляются друг от друга. Процедура повторяется до тех пор, пока не сойдется. Если коэффициент корреляции между рангами различий и расстояний не слишком высок, то размерность евклидова пространства увеличивается на единицу и весь процесс повторяется сначала. Получившиеся на предыдущем шаге координаты используются в качестве начальной конфигурации для нового шага. В конечном итоге точки располагаются так, что матрица расстояний между ними максимально соответствует матрице различий между объектами исходного множества. Мы получаем представление совокупности образцов множеством точек в многомерном евклидовом пространстве и можем анализировать их взаимное расположение относительно друг друга как визуально, так и с помощью стандартной техники многомерного анализа.

То, что в результате применения этого метода все образцы распались на группы, соответствующие различным стадиям заболевания, говорит о том, что предлагаемая процедура действительно позволила уловить существенные различия между образцами (рис. 2). Но анализ взаимного расположения соответствующих им точек на графике говорит гораздо больше. Точки, соответствующие предрасположенным и части нормальных, которых мы назвали суб-

нормальными, заметно сдвинулись вверх от точек, соответствующих здоровым и больным. Поскольку каждому направлению на плоскости профилей образцов (рис. 2) взаимно однозначно отвечает направление на плоскости профилей генов (рис. 3), это означает, что у субнормальных и предрасположенных экспрессируются совсем не те гены, что у здоровых и больных. Между здоровыми и больными больше общего, чем между здоровыми и предрасположенными или субнормальными и больными. Из рис. 2 очевидно, что задачу поиска генов-кандидатов, перспективных для выявления болезни Хантингтона, можно ставить по-разному. Можно искать гены, экспрессия которых различается у больных и здоровых, что обычно и делается. Но очевидно, что при этом в группу больных попадет и какая-то часть предрасположенных. А можно искать гены, экспрессия которых специфична именно для больных и выделяет их среди всех остальных, включая и субнормальных, и предрасположенных, что, собственно, и сделано в нашей работе. Наши результаты подтвердили выводы работы F. Borovecki *et al*. (2005), несмотря на разницу в методах обработки. Однако в работе H. Runne *et al*. (2007) на другой группе людей статистической разницы между больными и контролем по этим 12 генам найдено не было. Возможные причины расхождения с результатами H. Runne *et al*. (2007), на наш взгляд, могут заключаться или в различиях критериев отнесения пациентов к больным и контролю, или в несколько иной генетической природе анализируемых групп. Этот вопрос требует дальнейших исследований.

Природа вновь выявленной группы, названной нами субнормальной, достаточно загадочна. Во-первых, требует тщательного содержательного анализа то обстоятельство, что точки, соответствующие группе предрасположенных вместе с группой субнормальных, лежат не между точками, соответствующим группам здоровых и больных, как логично было бы ожидать, а заметно сдвинуты по отношению к ним (рис. 2). Во-вторых, по критериям отбора в группу здоровых четыре субнормальных индивидуума не должны быть носителями мутации по гену, кодирующему белок хантингтин. Именно увеличенное число повторов CAG в этом гене, переданное по наследству и продолжающее увеличиваться в нейронах у потомков, проявляется обычно после 50 лет в виде болезни Хантингтона – сильного нейродегенеративного расстройства, прогрессирующего с возрастом (Landles, Bates, 2004). Однако недавно стало известно, что количество повторов в этом гене может увеличиваться с возрастом и в соматических тканях, в частности в нейронах головного мозга у формально здоровых людей, и превышать порог в 35 повторов, после которого оно считается мутацией (Shelbourne *et al*., 2007). Таким образом, анализ различий в экспрессии генов между нормальными особями и индивидами из выделенной нами субнормальной группы мог бы пролить свет на природу факторов (по-видимому, общих для соматических и генеративных тканей), приводящих к мутированию в результате нарушений в процессе репликации или репарации ДНК участков генов, содержащих тринуклеотидные повторы. Поэтому проблема с подбором людей для контрольной группы становится очень существенной для адекватных сравнений профилей экспрессии генов и надлежащего выбора биомаркеров такого сложного метаболического заболевания, как болезнь Хантингтона.

Работа поддержана грантами НШ-2447. 2008.4, РФФИ 07-04-00441-а и программой РАН «Молекулярная и клеточная биология» (проект № 10.7).

Литература

Barnes M., Freudenberg J., Thompson S. *et al*. Experimental comparison and cross-validation of the Affymetrix and Illumina gene expression analysis platforms // Nucl. Acids Res. 2005. V. 33. № 18. P. 5914–5923.

Barrett T., Suzek T.O., Troup D.B. *et al*. NCBI GEO: mining millions of expression profiles – database and tools // Nucl. Acids Res. 2005. V. 33. P. 562–566.

Bemmo A., Benovoy D., Kwan T. *et al*. Gene expression and isoform variation analysis using affymetrix exon arrays // BMC Genomics. 2008. V. 9. P. 529–543.

Benjamini Y., Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing // J. Roy. Stat. Soc. B. 1995. V. 57. P. 289–300.

Borovecki F., Lovrecic L., Zhou J. *et al*. Genome-wide expression profiling of human blood reveals biomarkers for Huntington's disease // Proc. Natl

Acad. Sci. USA. 2005. V. 102. P. 11023–11028.
Cheadle C., Becker K.G., Cho-Chung Y.S. *et al*. A rapid method for microarray cross platform comparisons using gene expression signatures // Mol. and Cellular Probes. 2007. V. 21. P. 35–46.
Landles Ch., Bates G.P. Huntingtin and the molecular pathogenesis of Huntington's disease // EMBO Rep. 2004. V. 5. № 10. P. 958–963.
Lee H.K., Hsu A.K., Sajdak J. *et al*. Coexpression analysis of human genes across many microarray data sets // Genome Res. 2004. V. 14. P. 1085–1094.
Pedotti P., Hoen P.A., Vreugdenhil E. *et al*. Can subtle changes in gene expression be consistently detected with different microarray platforms? // BMC Genomics. 2008. V. 9. P. 124–136.
Runne H., Kuhn A., Wild E.J. *et al*. Analysis of potential transcriptomic biomarkers for Huntington's disease in peripheral blood // PNAS. 2007. V. 104. № 36. P. 14424–14429.
Shelbourne P.F., Keller-McGandy C., Bi W.L. *et al*. Triplet repeat mutation length gains correlate with cell-type specific vulnerability in Huntington disease brain // Hum. Mol. Genet. 2007. V. 16. № 10. P. 1133–1142.
Taguchi Y.H., Oono Y. Relational patterns of gene expression via non-metric multidimensional scaling analysis // Bioinformatics. 2005. V. 21. № 6. P. 730–740.

APPLICATION OF NONMETRIC MULTIDIMENSIONAL SCALING FOR ANALYSIS OF CROSS-PLATFORM GENE EXPRESSION MICROARRAY DATA

V.M. Efimov, A.V. Katokhin

Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia, e-mail: efimov@bionet.nsc.ru

Summary

The multidimensional method of the conformed multiplatform search of genes-markers with differential expression in the microarray data for the same set of samples is developed. The method is applied to the joint analysis of two different microarray platforms, CodeLink and Affymetrix (Borovecki *et al*., 2005). Lists of genes-markers with differential expression corresponding to each of four groups of samples of peripheral blood (symptomatic Huntington disease patients, presymptomatic, subnorm, norm) are built.