**УДК** 519.25

# СНИЖЕНИЕ РАЗМЕРНОСТИ В ЗАДАЧЕ АНАЛИЗА ДАННЫХ ЭКСПРЕССИИ ГЕНОВ НА МИКРОЧИПАХ

**Воронцов А. С.**, студент гр. КС-71-10
**Михальский А. И.**, д.б.н., профессор, в.н.с., Email: ipuran@yandex.ru
МГТУ МИРЭА, Кафедра биомедицинской электроники №540, Москва
Институт проблем управления им. В.А. Трапезникова РАН, Москва

**Аннотация.** Рассматривается проблема выбора информативных генов при анализе их экспрессии на микрочипах. Для решения поставленной задачи вводится характеристика, аналогичная симметризованному расстоянию Кульбака-Лейблера, для расстояния между распределениями экспрессии в образцах контрольной и подверженной онкологическому заболеванию ткани. Описываются методика оценки расстояния между распределениями и результаты расчётов по реальным данным для рака печени. Показано, что в рассмотренных образцах всего 10% генов отвечает за 90% расстояния между распределениями, что говорит о высокой информативной избыточности использования всех генов.
**Ключевые слова:** экспрессия генов на микрочипах; расстояние между распределениями; снижение размерности; информационная избыточность.

# DIMENSION REDUCTION OF MICROCHIP GENE EXPRESSION DATA

***Vorontsov A.S.,*** student КС-71-10
***Michalski A.I***., D.ofSci., proff**.,** Email: ipuran@yandex.ru
MGTU MIREA, Faculty biomedical electronics №540, Moscow
Institute of Control Sciences V.A. Trapeznikov RAS, Moscow

**Abstract.** A problem of informative genes selection is considered. A divergence between gene expression distribution in control and cancer tissues is constructed like the symmetric Kulback-Leibler distance. An estimate of the divergence between gene expression distributions is proposed and the results for data on hepatic cancer are presented. It is shown that no more than 10% of presented genes demonstrate 90% of divergence between gene expression distributions. This fact indicates an informative redundancy of the total set of genes in the presented data.
**Keywords:** microchip gene expression data; distributions divergence; dimension reduction; informative redundancy.

## Введение

Совершенствование технологической базы биологических исследований привело к революционным изменениям в области генетических исследований, исследований метаболизма и открыло новые направления развития медицины [1]. Наряду с возможностью оценки риска развития наследственных заболеваний и патологий, доступность большого числа генетических данных и различных биомаркеров позволила приблизиться к пониманию механизмов возникновения и развития онкологических заболеваний, заболеваний сердечно-сосудистой системы,

многих других хронических патологий и открыла перспективы создания новых лекарств и методов лечения [4-7].

В настоящее время промышленно производится большое число разнообразных генных микрочипов, используемых для выяснения предрасположенности человека к различным заболеваниям, оценки реакции организма на различные лекарственные препараты, и многое другое [2]. Генные микрочипы позволяют с высокой точностью диагностировать заболевание не только на ранних стадиях, но и до проявления этого заболевания.

Развитие технологии генных микрочипов породило и новые проблемы в области анализа данных, полученных с их помощью. Генные микрочипы позволяют за одно исследование в автоматическом режиме получить информацию об экспрессии тысяч генов. Из-за погрешностей приборов, человеческого фактора, несовершенства методов регистрации сигналов, наличия молчащих генов и многого другого, в выходных данных появляются погрешности и ошибки. Кроме того, для получения статистически надёжного результата при регистрации большого числа генов необходимо иметь достаточное количество повторных независимых наблюдений, то есть проводить генетический анализ большого числа людей, страдающих одинаковой болезнью, что проблематично не только из-за высокой стоимости исследований, но и по причине малой распространённости конкретных патологий в популяции.

Таким образом, возникает задача повышения качества анализа результатов и уменьшения времени обработки данных генного микрочипа путём обработки не всех генов, а наиболее информативных для изучения конкретной. В статье рассматривается метод выделения информативных генов путём сравнения распределений экспрессии генов в образцах контрольной и подверженной заболеванию тканей.

**Оценка информативности генов**

Для выделения информативных генов рассмотрим величину, характеризующую расстояние между распределениями экспрессии гена в образцах контрольной и подверженной заболеванию тканей. В качестве такой характеристики примем дивергенцию Ку́льбака — Ле́йблера, называемую в теории информации информационной дивергенцией, либо относительной энтропией [2]. Эта характеристика является несимметричной мерой удаленности друг от друга двух вероятностных распределений. Обычно, одно из сравниваемых распределений — это «истинное» распределение, второе — предполагаемое (проверяемое), являющееся приближением первого.

Для дискретных распределений расстояния Кульбака-Лейблера вычисляется по формуле:

$$D_{KL} = \sum_x p(x) \ln \frac{p(x)}{q(x)}$$

где $p(x)$, $q(x)$ – функции вероятности для двух распределений дискретной случайной величины $X$. При решении задачи анализа экспрессии генов функции вероятности заменим гистограммами, построенными на наборе дискретных значений гена в двух классах. Расстояние Кульбака-Лейблера для характеристики различия распределений запишем как

$$D_{KL} = -\frac{1}{2} \sum_x \left( p_1(x) \ln p_2(x) + p_2(x) \ln p_1(x) \right)$$

здесь суммирование проводится по множеству дискретных значений экспрессии гена, использованных для построения гистограмм в двух классах, $p_1(x)$ - значения гистограммы, построенной в первом классе, $p_2(x)$ - значения гистограммы, построенной во втором классе.

Вычислим характеристику $D_{KL}$ для каждого гена и упорядочим гены по мере убывания этой величины. Полученный ряд даёт представление об информативности генов при сравнении двух классов и может использоваться в дальнейшем анализе, например, для отбора признаков при построении правила классификации.

**Результаты анализа реальных данных**

Описанная методика оценки информативности генов применялась к реальным предварительно обработанным данным, состоявшим из двух наборов. В первом наборе были представлены данные по экспрессии генов в опухоли при раке печени, отнесенные к экспрессии генов в метастазе. Во втором наборе содержались данные по экспрессии генов в метастазе, отнесенные к экспрессии генов в нормальной ткани. Оба набора включали данные для 63 человек по 7581 генам.

На рис. 1 представлена плотность распределения расстояния Кульбака-Лейблера, вычисленного по всем 7581 гену. Эта плотность строилась с помощью процедуры density из статистического пакета R.

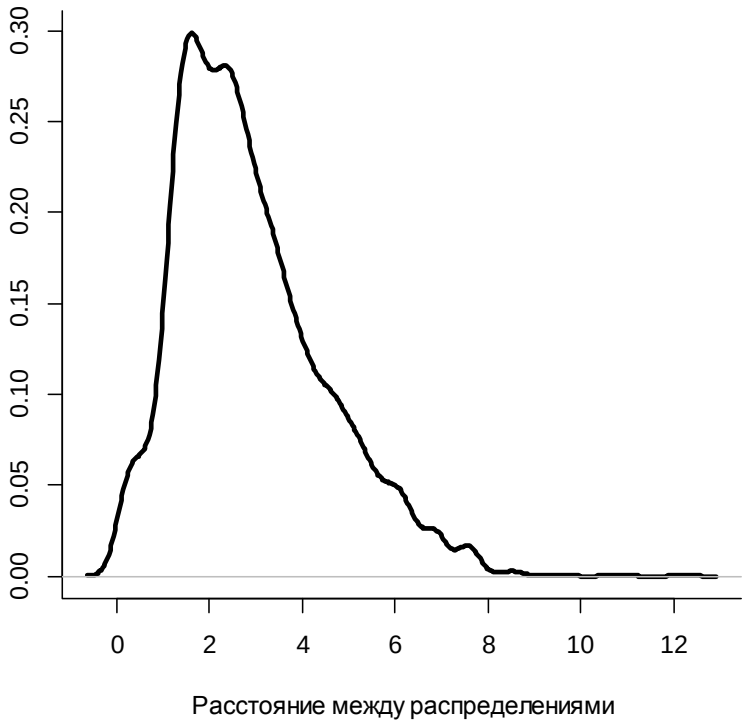


Рис. 1. Плотность распределения значений характеристики $D_{KL}$ для 7581гена.

На рис. 2 для иллюстрации представлены примеры распределений в двух классах экспрессий генов с маленькой и с большой величиной характеристики $D_{KL}$.

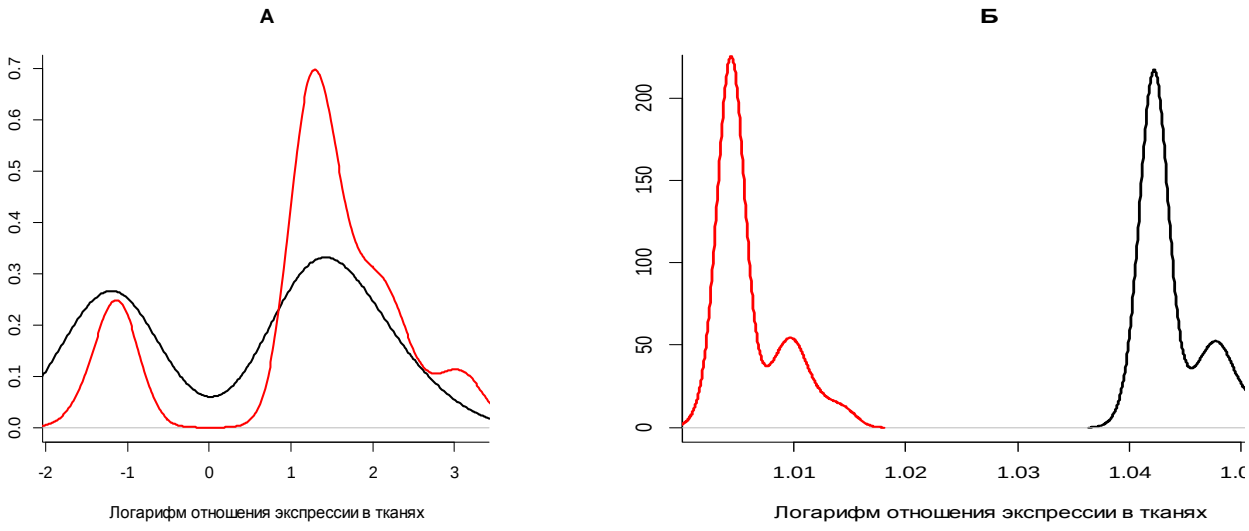


Рис. 2. Плотность распределения экспрессии генов в двух классах.
А - для гена с $D_{KL} = 2.6$, Б - для гена с $D_{KL} = 12.2$.

Из рисунка видно, что маленькая величина характеристики $D_{KL}$ соответствует плохо разделимым распределениям, а большая величина - хорошо разделимым распределениям.

Чтобы выделить гены, распределение экспрессии которых в двух классах различаются в наибольшей степени, строилась кумулятивная функция распределения. График кумулятивной функции, построенной с помощью процедуры ecdf из статистического пакета R, представлен на рис. 3. На рисунке вертикальная пунктирная линия указывает для значения характеристики $D_{KL}$ порог, выше которого лишь 10% генов имеют большие значения.

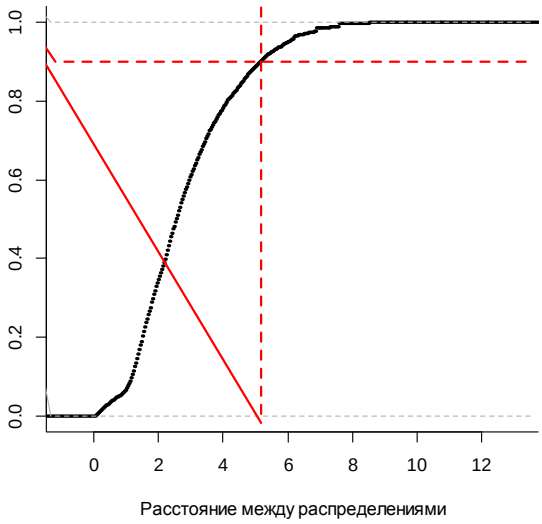


Рис. 3. Кумулятивная функция распределения характеристики $D_{KL}$ для 7581гена.

**Заключение**

Проведенный анализ распределений экспрессии генов показал, что с помощью расстояния Кульбака-Лейблера можно произвести отбор информативных генов, для которых характерны существенные различия в распределении их экспрессии в двух классах. Такая процедура проводится автоматически и позволяет существенно сократить размерность решаемой задачи, а именно число генов, которые необходимо исследовать на следующих стадиях изучения данных. Отобранные гены могут использоваться при построении решающих правил для классификации типов тканей и для выявления структурных взаимосвязей генов (построения генных сетей). Методика применения расстояния Кульбака-Лейблера для оценки информативности признаков универсальна и может применяться для анализа данных любой природы.

**Список литературы**

1. Баранов В.С. Программа "Геном человека" и научная основа профилактической медицины // Вестник РАМН − 2000. − №10. − С. 27-36.

2. Кульбак С. Теория информации и статистика. − М.: Наука, 1967. −408 с.

3. Свешникова А.Н., Иванов П.С. Экспрессия генов и микрочипы: проблемы количественного анализа // Рос. хим. ж. − 2007. − т.LI, № 1. − С. 127-135.

4. Au W.W., Ruchirawat M. Biomarkers in population studies: environmental mutagenesis and risk for cancer // Rev. Environ. Health. − 2009. − Vol.24, №2. − P. 117-127.

5. Carrara S., Ghoreishizadeh S., Olivo J. Fully integrated biochip blatforms for advanced healthcare // Sensors. − 2012. − Vol.12, №8. − P. 11013–11060.

6. Goldman M.A. Digital drug discovery // Genome Biology. − 2005. − Vol.6 − P. 348-350.

7. Napoli C., Lerman L.O., Sica V. Microarray analysis: a novel research tool for cardiovascular scientists and physicians // Heart. − 2003. − Vol.89 − P. 597–604.