

РАЗДЕЛЕНИЕ ТРАНСКРИПЦИОННЫХ СИГНАЛОВ КЛЕТОЧНЫХ ПОПУЛЯЦИЙ ОПУХОЛЕВЫХ ТКАНЕЙ РАКА ЛЕГКОГО С ИСПОЛЬЗОВАНИЕМ МЕТОДА НЕЗАВИСИМЫХ КОМПОНЕНТ

А. Л. Одинец

*Белорусский государственный университет, г. Минск;
odinets96@mail.ru;*

*науч. рук. – Н. Н. Яцков, канд. физ.-мат. наук, доц.
консультант – П. В. Назаров, канд. физ.-мат. наук, доц.*

В данной работе рассматривается разделение транскрипционных сигналов клеточных популяций с использованием метода независимых компонент. Вызванный стремительным развитием технологий секвенирования, быстрый рост экспериментальных массивов данных в области анализа генетической информации создаёт необходимость в создании и применении новых алгоритмов по поиску полезной их составляющей. Описанный в статье метод независимых компонент позволяет упростить анализ данных генной экспрессии. В качестве экспериментальных данных используется набор экспрессий генов пациентов с раком лёгкого. Приводятся результаты разделения транскрипционных сигналов для набора из 20531 гена 553 пациентов. Установлена возможность использования матрицы коэффициентов независимых компонент для классификации пациентов с высокой точностью.

Ключевые слова: транскрипционные сигналы; метод независимых компонент; карцинома; ДНК; генная экспрессия.

ВВЕДЕНИЕ

С появлением генных микрочипов высокого разрешения стало возможно изучение и одновременный мониторинг экспрессии всего клеточного генома [1]. Однако размеры экспериментальных данных затрудняют использование классических алгоритмов анализа данных [2], например, для разделения транскрипционных сигналов клеточных популяций. Одним из перспективных направлений, позволяющим существенно улучшить разделение транскрипционных сигналов, является метод независимых компонент [3].

Целью работы является исследование возможностей применения метода независимых компонент для разделения транскрипционных сигналов на примере рака лёгкого – карциномы. В ходе выполнения работы решаются следующие задачи:

1. Программная реализация метода независимых компонент.
2. Анализ экспериментальных данных, представленных набором из библиотеки атласа генома рака (TCGA) [4].

3. Оценка устойчивости и точности разработанных алгоритмов.

МЕТОД НЕЗАВИСИМЫХ КОМПОНЕНТ

В методе независимых компонент предполагается, что исходные сигналы s_i смешаны коэффициентами a_{ij} , где j – номер сигнала, i – номер переменной [5]. В эксперименте регистрируются смешанные сигналы:

$$X = S A \quad (1)$$

где X – матрица размера $(N \times K)$ регистрируемых данных, где N – общее число генов в эксперименте, K – количество наборов измерений; A – матрица размера $(F \times K)$ коэффициентов a_{ij} ; S – матрица размера $(N \times F)$, где F – количество независимых компонент, формирующая независимые компоненты или источники сигналов s_i . Если известны коэффициенты a_{ij} , то задача решается линейной системой, построенной на основе матриц формулы (1). Однако s_i и a_{ij} не известны.

Возможный подход к решению подобной ситуации – использовать для оценки s_i и a_{ij} некоторую информацию о статистических свойствах исходных сигналов s_i . Основопологающим предположением является то, что сигналы s_i негауссовы [5]. Производится поиск поворота экспериментальных данных X , чтобы результирующие сигналы s_i были наименее гауссовыми [5]. Для этого, после первого случайного поворота, каждый последующий уточняется посредством вычисления разницы энтропии случайной гауссовой величины и искомого вектора – независимой компоненты. Из-за предварительного шкалирования данных корреляция и ковариация обоих векторов должна совпадать. Разница энтропий – негэнтропия – стремится к максимуму по мере приближения к независимой компоненте и является опорным критерием в ряде реализаций метода независимых компонент [5].

Приняв во внимание указанные предположения, модель метода независимых компонент определена, возможно нахождение смешивающей матрицы и независимых компонент с точностью, определяющейся относительным вкладом независимых компонент в общую вариацию регистрируемых данных.

АНАЛИЗ ДАННЫХ ГЕННОЙ ЭКСПРЕССИИ

В работе рассмотрены данные генной экспрессии для лёгочного рака – карциномы, 20531 генов 553 пациентов. Априорно известны пациенты с опухолями – 502 человека. Требуется разделить транскрипционные сигналы больных и здоровых людей. В качестве среды обработки выбран язык R [6].

Применение метода независимых компонент требует выбора количества независимых компонент F . В качестве критерия оптимального выбора используется квадрат относительной ошибки главных компонент, представляющий собой отношение остаточной дисперсии к выборочной. Вычислены сигналы S и матрицы коэффициентов A для наборов независимых компонент от 1 до 50. Анализ кривой относительной ошибки главных компонент, представленной на левой половине рисунка, позволил принять решение о достаточности $F = 20$ взятых независимых компонент. Разбиение данных на два класса (здоровые и больные пациенты) выполнено методом k -средних [7]. Наилучшие результаты получены для $F \geq 10$ независимых компонент, что подтверждает правильность выбора.

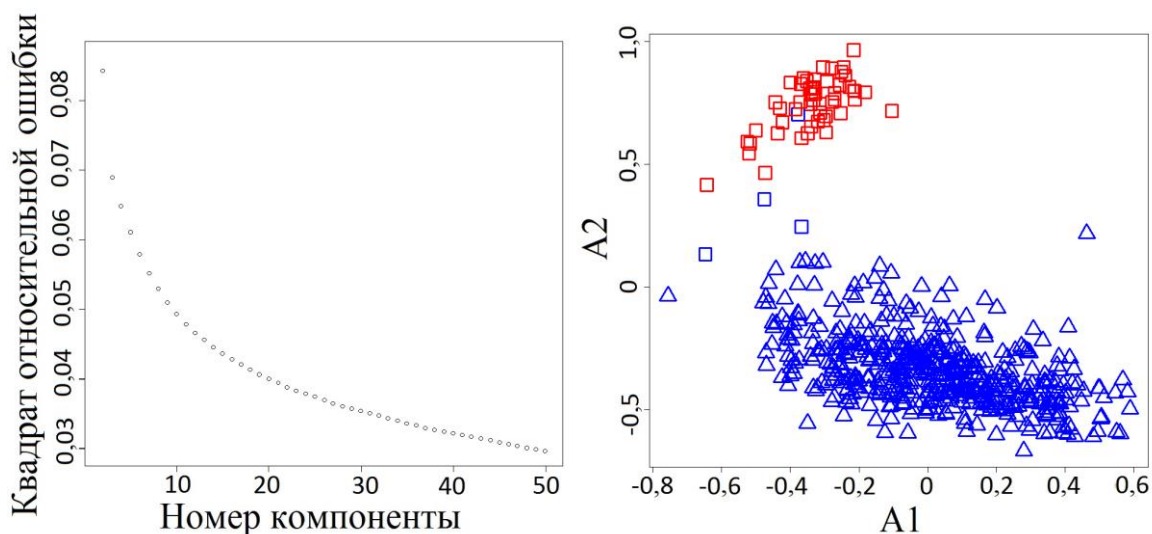


Рис. 1. Квадрат относительной ошибки главных компонент и значения первых двух векторов коэффициентов матрицы A

Успешное разделение пациентов по признаку наличия опухоли, основанное на матрице коэффициентов A , представлено на правой части рисунка. Выделенная красным область отражает группу априори здоровых пациентов, квадратами обозначены пациенты соответствующего кластера, полученного методом k -средних.

Первая независимая компонента сопоставлена сигналу клеточной популяции на формирование опухоли. Анализ показал, что гены с уровнем сигнала, близким к нулю, участие в заболевании не принимают. Таким образом, метод независимых компонент позволяет не только выделить транскрипционные сигналы клеточных популяций, но и определить набор генов-маркеров.

Анализ полученных методом независимых компонент матриц S и A указал на разделение информации между ними. Матрица независимых компонент S позволяет выделить профили транскрипционных сигналов, в частности отвечающий за образование опухоли, следовательно и отве-

чающие за этот процесс гены. Матрица же коэффициентов A содержит информацию непосредственно о самих пациентах, позволяя выделять пациентов в группы по разным признакам. Это объясняется размерностью начальных и конечных данных: если исходный массив состоял из генов-строк и пациентов-столбцов, т. е. «гены» $\times$ «пациенты», то полученные S – это «гены» $\times$ «компоненты», в то время как A – это «компоненты» $\times$ «пациенты».

ЗАКЛЮЧЕНИЕ

Разработан и применен к данным генной экспрессии метод независимых компонент. Проведен анализ генома рака лёгкого – карциномы. Методом независимых компонент разделены транскрипционные сигналы здоровых и больных пациентов. На основе критерия ошибки метода главных компонент выбрано оптимальное число независимых компонент – 20. На основе априорных данных проведена оценка разделения. Достигнута точность в 97,8 % верно классифицированных пациентов с перспективой улучшения посредством замены метода кластеризации.

Библиографические ссылки

1. Shefa U., Jung J. Comparative study of microarray and experimental data on Schwann cells in peripheral nerve degeneration and regeneration: big data analysis // *Neural Regen Res.* 2019. Vol. 14, P. 1099–1104. DOI: 10.4103/1673-5374.250632.
2. Gauch H.G. Jr, Qian S., Piepho H.P., et al. Consequences of PCA graphs, SNP codings, and PCA variants for elucidating population structure // *PLoS One.* 2019. Vol. 14. DOI: 10.1371/journal.pone.0218306
3. Nazarov P.V., Wienecke-Baldacchino A.K., Zinovyev A., et al. Deconvolution of transcriptomes and miRNomes by independent component analysis provides insights into biological processes and clinical outcomes of melanoma patients // *BMC Med Genomics.* 2019. Vol. 12. DOI: 10.1186/s12920-019-0578-4
4. Cooper L.A., Demicco E.G., Saltz J.H., et al. PanCancer insights from The Cancer Genome Atlas: the pathologist's perspective // *J Pathol.* 2018. Vol. 244, P. 512–524. DOI: 10.1002/path.5028
5. Sompairac N., Nazarov P.V., Czerwinska U., et al. Independent Component Analysis for Unraveling the Complexity of Cancer Omics Datasets // *Int J Mol Sci.* 2019. Vol. 20. DOI: 10.3390/ijms20184414
6. Kosinski M. Package ‘RTCGA’ // Electronic resource. URL: <http://bioconductor.org/packages/release/bioc/manuals/RTCGA/man/RTCGA.pdf> (date of access: 20.10.2019)
7. Яцков Н. Н., Шингарёв И. П. Интеллектуальный анализ данных: методические указания к лабораторным работам. Минск: БГУ, 2012.