УДК 577.2:616-006

У.Е. Каиров[1,2], А.Ю. Зиновьев[3], Т.А. Карпенюк[1*], Е.М. Раманкулов[2]
[1]Казахский национальный университет им. аль-Фараби, Казахстан, г. Алматы
[2]Национальный центр биотехнологии РК, Казахстан, г. Астана
[3]Институт Кюри, Франция, Париж
[*]E-mail: ukairov@gmail.com

Поиск воспроизводимых независимых компонент в транскриптомах пациентов с раком молочной железы

Аннотация. С использованием Метода независимых компонент, разработана методика поиска воспроизводимых независимых компонент в транскриптомных наборах рака молочной железы на основе микрочипов Affymetrix HG-U133A. Данная методика может быть использована для построения корреляционных графов, выявления генных взаимодействий и новых сигнальных путей, характерных для раковых заболеваний.
Ключевые слова: рак молочной железы, метод независимых компонент, корреляционные графы экспрессии генов, генные взаимодействия, сигнальные пути

В настоящее время, широкое применение в биомедицине и биотехнологии получили технологии высокоплотных микрочипов [1,2]. Высокоплотные микрочипы представляют собой высокопроизводительные средства для одновременного определения уровня экспрессии тысяч генов, различных типов тканей и клеток при различных физиологических состояниях, обработанных химическими или лекарственными препаратами [3-7]. Наиболее активно высокоплотные микрочипы используют для определения экспрессии генов в раковых тканях человека. Такой транскриптомный анализ является одним из важных направлений современной функциональной геномики и биоинформатики.

Накопление огромных массивов данных, генерируемых различными исследованиями на основе технологий высокоплотных микрочипов, требует применения новых статистических и математических подходов для обработки данных. К примеру, один микрочип для измерения экспрессии генов человека компании Affymetrix, несет информацию о более чем 25 тысячах генах [8,9], а количество образцов в современных исследованиях с использованием микрочипов, может превышать десятки сотен.

Классическая статистика хорошо оперирует данными в случаях, где мы имеем большое количество наблюдений для небольшого числа переменных. Однако существующие сегодня методы анализа, особенно в геномных исследованиях, генерируют чрезмерно большое количество разных переменных. В таких случаях применяются методы «обучения без учителя», использующие технику уменьшения размерности данных для сокращения многомерности транскриптомных данных и выделения значимых паттернов экспрессии. Одним из перспективных и применяемых математических методов для анализа данных большой размерности является Метод независимых компонент (МНК) [10].

Первая работа с применением МНК для анализа транскриптомных данных была опубликована в 2001 году [11]. Затем, в более детальной работе, был освещен потенциал МНК, примененный для анализа данных, приближенных к биологическим реалиям клетки [12]. Несмотря на наличие исследований с применением МНК, ранее не проводили тщательный анализ и поиск воспроизводимых независимых компонент, необходимый для дальнейшего правильного извлечения и построения генных сетей взаимодействия.

В задачи нашего исследования входил поиск и анализ воспроизводимых независимых компонент после применения Метода независимых компонент для транскриптомных наборов данных рака молочной железы.

© 2012 al-Farabi Kazakh National University

В качестве материала для исследования использовались необработанные серии наборов данных GSE1456, GSE2034, GSE2990, GSE3494 (*.CEL и *.DAT файлы) микрочипов Affymetrix (HG-U133A) из банка данных GEO (Gene Expression Omnibus, NCBI) [14]. Для анализа были выбраны микрочипы пациентов с раком молочной железы. Статистическая обработка значений сигналов была проведена с помощью R2.8.1 (Bioconductor) [15] с применением метода нормализации GCRMA [16]. Для центрирования, объединения пробсетов по медианному значению, вычисления коэффициентов корреляции Пирсона и анализа обработанных данных использовалась математическая среда разработки и программирования Matlab 2008b (MathWorks) [17]. Для расчета независимых компонент использовался пакет функций и процедур Icasso [18] для среды Matlab.

Для исследования была отобрана выборка из общедоступных наборов микрочипов, состоящая из четырех серий данных разных экспериментов с общим количеством образцов, равным 885 (Таблица 1).

Изначально необработанные *.CEL файлы для четырех раковых наборов подвергались статистической процедуре нормализации с помощью метода GCRMA (рисунок). В нашем случае, применение GCRMA нормализации позволило учитывать поправку на GC-содержание в пробсетах, расположенных на микрочипе. Затем, значения сигналов с микрочипов были отцентрированы в Matlab по средней величине отдельного пробсета для каждого отдельного набора данных. Дополнительно, с целью избежания избыточной зашумленности данных проводился отбор наиболее вариабельных генов по пороговому значению от 3.0 и более. Для микрочипов Affymetrix HG-U133A характерно соответствие нескольких пробсетов одному гену [8,9]. Для унификации и приведения пробсетов к одному гену, в среде Matlab была применена процедура объединения нескольких пробсетов, соответствующих одному гену, по медианному значению. После применения такого подхода спектр распределения сигналов кардинально не меняется и позволяет избежать потери информативности при статистической обработке сигналов (Рисунок).

Таблица 1 - Общая характеристика выборки данных

№ серии набора данных	Тип рака	Количество образцов в серии	Количество выбранных микрочипов	Количество пробсетов на микрочипе	Количество генов после статистической обработки
GSE1456	Рак молочной железы	318	159	22 285	8775
GSE2034	Рак молочной железы	286	286	22 285	13016
GSE2990	Рак молочной железы	189	189	22 285	8941
GSE3494	Рак молочной железы	502	251	22 285	9085

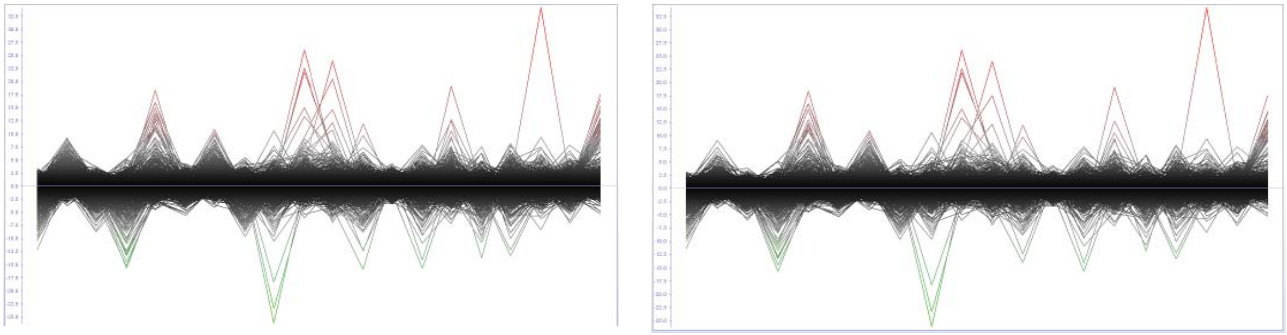

Рисунок - Общая картина распределения сигналов экспрессии на примере набора данных GSE3494 до процедуры медианного объединения пробсетов (сверху) и после применения процедуры (снизу). По оси абсцисс – кривые значений сигналов с пробсетов, а по оси ординат – количественные значения уровней сигналов

Таблица 2 - Корелляционная таблица значений независимых компонент для наборов данных GSE1456, GSE2034, GSE2990, GSE3494

Независимая переменная	Тип взаимодействия	Зависимая переменная	Абсолютное значение коэффициента корреляции
1IC2	корреляция	2IC1	0.708856
1IC4	корреляция	2IC2	0.768026
1IC5	корреляция	2IC6	0.653702
1IC10	корреляция	2IC13	0.568184
1IC12	корреляция	2IC4	0.689559
1IC15	корреляция	2IC10	0.546145
1IC19	корреляция	2IC12	0.596774
1IC20	корреляция	2IC16	0.57104
1IC1	корреляция	3IC6	0.521852
1IC2	корреляция	3IC3	0.719649
1IC4	корреляция	3IC1	0.809597
1IC5	корреляция	3IC2	0.780374
1IC8	корреляция	3IC10	0.517197
1IC9	корреляция	3IC11	0.516189
1IC10	корреляция	3IC7	0.632663
1IC12	корреляция	3IC8	0.733949
1IC13	корреляция	3IC15	0.519835
1IC16	корреляция	3IC14	0.654042
1IC1	корреляция	4IC3	0.555495
1IC2	корреляция	4IC20	0.767128
1IC3	корреляция	4IC13	0.548447
1IC4	корреляция	4IC1	0.836501
1IC5	корреляция	4IC4	0.741107
1IC9	корреляция	4IC10	0.548755
1IC10	корреляция	4IC6	0.632444
1IC12	корреляция	4IC5	0.766037
1IC13	корреляция	4IC9	0.585166
1IC16	корреляция	4IC15	0.574019
1IC17	корреляция	4IC7	0.55048
1IC20	корреляция	4IC8	0.579017
2IC1	корреляция	3IC3	0.686068
2IC2	корреляция	3IC1	0.770041
2IC4	корреляция	3IC8	0.715829
2IC6	корреляция	3IC2	0.647796
2IC8	корреляция	3IC10	0.709816
2IC9	корреляция	3IC11	0.514548
2IC10	корреляция	3IC14	0.510483
2IC11	корреляция	3IC15	0.572404
2IC13	корреляция	3IC7	0.511726
2IC16	корреляция	3IC13	0.548153
2IC17	корреляция	3IC16	0.573692
2IC1	корреляция	4IC20	0.684927
2IC2	корреляция	4IC1	0.80812
2IC4	корреляция	4IC5	0.74372
2IC6	корреляция	4IC4	0.646009
2IC8	корреляция	4IC7	0.713021
2IC9	корреляция	4IC10	0.575606
2IC13	корреляция	4IC6	0.544601
2IC17	корреляция	4IC14	0.550909
4IC1	корреляция	3IC1	0.868915
4IC2	корреляция	3IC4	0.597187
4IC3	корреляция	3IC6	0.702172
4IC4	корреляция	3IC2	0.795974
4IC5	корреляция	3IC8	0.861986
4IC6	корреляция	3IC7	0.759029
4IC7	корреляция	3IC10	0.778385
4IC9	корреляция	3IC15	0.597743
4IC10	корреляция	3IC11	0.658879
4IC11	корреляция	3IC12	0.662245
4IC12	корреляция	3IC13	0.636641
4IC14	корреляция	3IC16	0.704727
4IC15	корреляция	3IC14	0.641869
4IC16	корреляция	3IC14	0.583524
4IC18	корреляция	3IC17	0.594559

Цифра перед каждой независимой компонентой (IC) соответствует набору данных: 1 – GSE1456, 2 – GSE2034, 3 – GSE2990, 4 – GSE3494.

Таким образом, была разработана методика поиска воспроизводимых независимых компонент среди разных наборов транскриптомных данных на основе микрочипов Affymetrix HG-U133A. Данная методика применения Метода независимых компонент и дальнейший анализ могут использоваться для построения корреляционных графов, генных взаимодействий и выявления новых сигнальных путей в раковых заболеваниях.

Литература

1 Schena M., Shalon D., Davis R.W., Brown P.O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray // Science. – 1995. – V.270. - P. 467-70.

2 DeRisi J., Penland L., Brown P.O. et al. Use of a cDNA microarray to analyse gene expression patterns in human cancer // Nature Genetics. – 1996. – V.14. – P. 457-460.

3 Rolph M.S., Sisavanh M., Liu S.M., Mackay C.R. Clues to asthma pathogenesis from microarray expression studies // Pharmacol.Ther. – 2006. - V.109. - P. 284-294.

4 Evans S.J., Choudary P.V., Vawter M.P., et al. DNA microarray analysis of functionally discrete human brain regions reveals divergent transcriptional profiles // Neurobiol. Dis. – 2003. – V14. – P. 240-250.

5 Golub T.R., Slonim D.K., Tamayo P., et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring // Science. – 1999. – V.286. – P. 531-537.

6 Van Veer L.J., Dai H., Van de Vijver M.J., et al. Gene expression profiling predicts clinical outcome of breast cancer // Nature. – 2002. – V. 415. – P. 530-536.

7 Sorlie T., Tibshirani R., Parker J., et al. Repeated observation of breast tumor subtypes in independent gene expression data sets // Proc. Natl. Acad. Sci. USA. – 2003.-V.100 (14). – P. 8418-8423.

8 http://www.affymetrix.com

9 Каиров У.Е., Зиновьев А.Ю., Карпенюк Т.А., Раманкулов Е.М. ДНК-микрочипы: от основ технологии к анализу данных // Вестник КазНУ. Серия биологическая. -2012. - №4 (56). – С. 270-274.

10 Comon P., Independent Component Analysis: a new concept // Signal Processing. – 1994. – 36 (3). – P. 287–314.

11 Liebermeister W. Linear modes of gene expression determined by independent component analysis // Bioinformatics. – 2002. – V. 18 (1). – P. 51-60.

12 Lee S. and Batzoglou S. Application of independent component analysis to microarrays // Genome Biol. – 2003. -V. 4 (11). – P. 1-21.

13 Hori G., Inoue M., Nishimura S., and Nahakara H. Blind gene classification an application of a signal separation method // Genome Informatics. – 2001. – V. 12. – P. 255-256.

14 http://www.ncbi.nlm.nih.gov/geo/

15 http://www.bioconductor.org/

16 Wu Z., Irizarry R.A., Gentleman R. et al. A model based background adjustment for oligonucleotide expression arrays // Journal of the American Statistical Association. - 2004. – V. 99. – P. 909-917.

17 http://www.mathworks.com/

18 Himberg J., Hyvarinen A. and Esposito F.. Validating the independent components of neuroimaging time series via clustering and visualization // Neuroimage. – 2004. – V.22 (3). – P. 1214-1222.

19 Jackson D. Stopping rules in principal component analysis: a comparison of heuristical and statistical approaches // Ecology. – 1993. – V.74 (8). – P. 2204-2214.

У.Е. Каиров, А.Ю. Зиновьев, Т.А. Карпенюк, Е.М. Раманкулов

Сүт безі ісігімен ауыратын пациенттердің транскриптомаларында қалпына келтірілетін тәуелсіз компоненттерді іздеу

Тәуелсіз компоненттер әдісті қолданып, Affymetrix HG-U133A микрочиптер негізінде сүт безі ісігінің транскриптомдардан тәуелсіз компоненттерді іздеу әдісі шығарылды. Бұл әдіс рак ауруларына тән корреляциялық графтарды тұрғызу үшін, гендік байланыстарды және жаңа сигналдық жолдарды анықтау үшін қолданылуы мүмкін.

Түйін сөздер: сүт безі ісігі, тәуелсіз компоненттер әдісі, ген экспрессиясының корреляциялық графтары, гендік әрекеттесу, сигналдық жолдар.

U.Ye. Kairov, A.Yu. Zinovyev, T.A. Karpenyuk, Ye.M. Ramanculov

Search of reproducible independent components in transcriptomes of patients with breast cancer

The technique with using of Independent component analysis for searching of reproducible independent components from Affymetrix HG-U133A array transcriptomes of patients with breast cancer has been developed. This technique can be used to construct correlation graphs of gene expression, to reveal gene interactions and new signaling pathways specific to cancer.

Keywords: breast cancer, independent component analysis, correlation graphs of gene expression, gene interactions, signaling pathways.