АНАЛИЗ ВОЗДЕЙСТВИЯ INT-γ НА КЛЕТКУ С ИСПОЛЬЗОВАНИЕМ ПРОГРАММНОГО ПАКЕТА *GENEEXPRESSIONANALYSER*

А.В. Саечников[1], P.V. Nazarov[2], Н.Н. Яцков[1], В.В. Апанасович[1]

[1]*Белорусский государственный университет, Минск, Беларусь;*

[2]*Genomics Research Unit, Centre de Recherche Public de la Sante, Luxembourg, Luxembourg*

saetchnikov.anton@tut.by

В работе приводятся результаты исследования изменений в клетке меланомы под воздействием INT-γ с течением времени. Представлен разработанный пакет GeneExpressionAnalyser для анализа данных микрочипов ДНК и описана алгоритмическая составляющая.

Ключевые слова: Микрочип ДНК, INF-γ, Matlab, Кластеризация, Метод главных компонент, Significance Analysis of Microarrays (SAM).

ЭКСПЕРИМЕНТАЛЬНЫЕ ДАННЫЕ

Для анализа воздействия INT-γ на клетку были использованы данные, полученные в работе [1], которые находятся в свободном доступе и могут быть загружены из хранилища ArrayExpress (индекс E-MEXP-3720).

МЕТОДОЛОГИЯ ПРОГРАММНОГО ПАКЕТА *GENEEXPRESSIONANALYSER*

В качестве среды разработки и реализации пакета выбрана программная среда Matlab, библиотека Bioinformatics. Для построения графического интерфейса используется система GUIDE пакета MATLAB 7.11.0 (R2010b) для ОС Windows®. Существенным преимуществом пакета MATLAB является оптимизация ядра, реализованного на языке программирования C++, для математических вычислений с матрицами, что значительно увеличивает скорость анализа и моделирования больших объемов многомерных данных. Программный пакет *GeneExpressionAnalyser* включает следующие методы и алгоритмы для анализа биочипов ДНК:

- загрузка данных, получаемых после обработки микрочипа
- предварительная обработка и фильтрация данных
- нормировка [2]
- восстановление пропущенных значений
- поиск дифференциально-выраженных генов с использованием статистического метода SAM (Significance Analysis of Microarrays [3])
- иерархическая и неиерархическая кластеризация генов [4],
- анализ методом главных компонент [5],
- выделение статистически значимых биофункций в ходе анализа генных аннотаций GO (GeneOntology)-анализ методами точного теста Фишера и случайных перестановок [6].

Большинство ключевых параметров анализа вводятся пользователем с помощью стандартного оконного интерфейса ОС Windows®. Вывод промежуточных и итоговых результатов осуществляется в виде графиков и внешних объектов (баз данных для возможности анализа промежуточных и конечных результатов, итоговых таблиц, дендрограмм). Предусмотрена возможность сохранения результатов анализа данных в графические файлы. Промежуточные и конечные результаты анализа можно сохранить в специальном формате, с возможностью последующего использования. Отдельные составляющие данного программного пакета были ранее протестированы на примере опубликованных экспериментальных данных, а также на смоделированных данных [7,8].

РЕЗУЛЬТАТЫ

Загружалось 17 различных значений экспрессии (разные временные точки и различные технические репликанты) для 33252 гена. Данные проверялись на наличие пропусков, которых выявлено не было, поэтому в последующем этапы удаления генов с большим количеством пропусков и восстановления пропущенных значений не производилось. Далее была произведена RMA-нормализация для каждого технического репликанта для уменьшения факторов, связанных с каждой конкретной матрицей и уменьшения влияния шумовой компоненты на значения экспрессии. После чего построены матрицы уровней экспрессиии для каждого технического репликанта, над которыми в дальнейшем и производились вычислительные процедуры.

На следующем этапе (поиск значимых генов) использованы следующие варианты анализа методом SAM (FDR < 0.05): oneclass, two class paired, two class unpaired, one class timecourse, multiclass, результаты которого приведены в Таблице 1.

Численные результаты анализа методом SAM

Таблица 1

Время	03H		12H		24H		48H		72H		JII ctrl	
	Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg
One class	59	0	158	0	2634	5791	2103	6923	4740	8580	3	0
Two class paired	71	10	337	40	1549	3887	2510	6281	3361	5954	8	1
Two class unpaired	60	0	427	0	772	1279	1030	3607	2306	4942	131	85

One class timecourse	pos	neg	FDR	Multiclass	Significant	FDR
	2673	481	0,038		1242	0,04

Pos (Positive regulated) – выраженные гены, Neg (Negative regulated) – подавленные гены.

Конечной целью анализа методом SAM является выделение как можно большего количества генов с наименьшим количеством ошибочно выделенных. Поэтому на основании результатов, представленных в Таблице 1, для анализа воздействия INF-γ на клетку, использовались результаты, полученные two class paired и one class анализом.. Максимальный эффект воздействия на клетку INF-γ в обоих случаях наблюдался в промежуток времени с 12 до 24 часов после начала лечения, причем после 24 часов около 70% всех значимых генов являлись подавленными генами. Первая реакция клетки на INF-γ наблюдается по прошествии определенного промежутка времени.

Эффективность воздействия интерферона в случае SAM two class paired падает после 48 часов (изменение количества значимых генов с 48 до 72 часов мало), а в случае SAM one class эффективность воздействия меньше в промежуток времени с 24 до 48 часов, а после 48 часов улучшается. Несмотря на то, что количество значимых генов в случае one class больше, для дальнейшего анализа были оставлены результаты SAM two class paired, т.к. сами входные данные больше подходят под определение two class-данных: класс данных до лечения и класс данных после лечения. В случае two class paired характер изменения количества значимых генов похож на полученный ранее в рамках анализа пакетом limma в программе R [9].

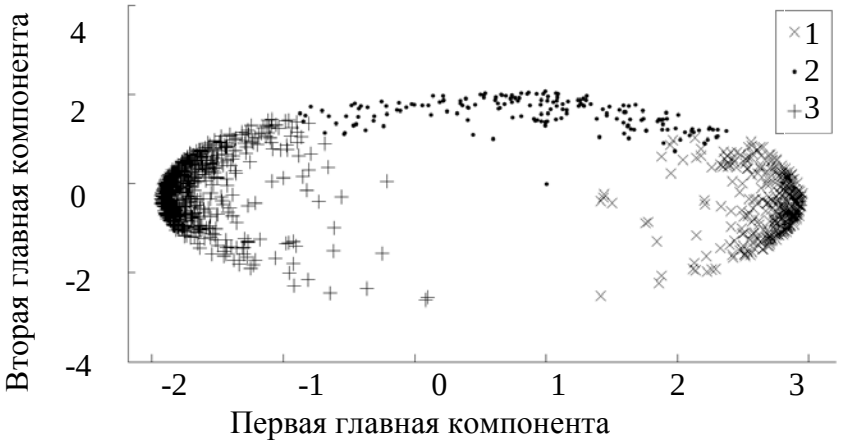

Рис. 1. – Результаты для SAM multiclass: первая компонента – 70,1% дисперсии всех признаков; вторая компонента – 15,4% дисперсии всех признаков

Рис. 2. – Результаты для SAM one class timecourse: первая компонента – 62,7% дисперсии всех признаков; вторая компонента – 15,6% дисперсии всех признаков

Для кластеризации были использованы результаты полученные методом SAM multiclass и SAM one class timecourse. Была проведена оценка качества кластеризации при помощи метода главных компонент, где было сделано предпочтение результату, полученному методом SAM multiclass (*Рис. 1, 2*) в силу пространственной разделенности кластеров генов и четкой выраженности областей кластеров. В случае SAM one class timecourse наблюдается отсутствие части кластера генов, который четко выражен на *Рис.1* и отсутствуют четкие границы областей кластеров. Кластеризация выполнена различными комбинациями методов связывания и метрик сравнения, в результате чего, определено, что максимальный кофенетический коэффициент 0.9228 достигается при использовании метода расчета расстояния между объектами по фор-

муле Эвклида и центроидной метрики сравнения. С помощью данной комбинации было определено три главных кластера генов.

Первый кластер (305 генов) характеризует первоначально подавленные гены, экспрессия которых оставалась почти неизменной в течение первых 12 часов, а потом экспрессии этих генов достигают определенного значения в промежуток времени от 12 до 48 часов после начала лечения, причем эти гены становятся выраженными и уже не изменяют свое состояние. Похожую реакцию на лечение можно наблюдать при анализе воздействия INF-γ на 2 кластер генов (637 генов), только гены со временем подавляются. Гены, относящиеся к третьему кластеру (149 генов), начинают реагировать с INF-γ сразу же после начала лечения и меняют свое состояние с подавленного на выраженное за 12 часов после начала лечения, после чего, с 12 до 24 часов практически не меняют свое состояние, а потом в промежуток времени с 24 до 72 часов становятся опять подавленными.

Необходимо отметить, что результаты для генов, обработанных ингибитором, блокирующим сигнал от интерферона, аналогичны полученным с самого начала обработки. Это подтверждает, что изменения экспрессии обусловлены воздействием INF-γ на клетку, а не изменениями в самой клетке со временем.

На последнем этапе анализа данных были определены значимые биофункции для дифференциальных генов, полученных методом SAM two class paired для качественного анализа того, какие биологические функции и как изменялись с течение времени. Кроме того, данный анализ был приведен над кластерами генов, полученных методом SAM multiclass для определения значимых биофункций, активность которых изменялась в соответствии с профилями экспрессии кластеров. Статистическая значимость биофункций определялась критерием $p < 0.01$. В соответствии с результатами SAM two class paired, было определено 3159 значимых биофункций, которые характерны для выраженных генов и определено 3680 значимых биофункций, которые характерны для подавленных генов. Полученные результаты можно разделить на 5 основных групп биофункций в соответствии с тем, что биофункции имеют доминирующую выраженность/подавленность в определенный момент времени. В 3 часа, после начала лечения, пик выраженности достигается 6,6 процентами всех статистически значимых биофункций, в 12 часов – 22,9%, в 24 часа – 44,9%, в 48 часов – 14,2%, в 72 часа – 9,2%, в 72 часа с ингибитором – 2,2% всех выраженных биофункций. Для подавленных биофункций пик подавленности в 12 часов после начала лечения достигается 3,5 процентами биофункций, в 24 часа – 19,2%, в 48 часов – 25,3%, в 72 часа – 52% всех выраженных биофункций.

Из полученных результатов следует, что абсолютное большинство биофункций достигают пика подавленности в конце эксперимента, тогда как, вплоть до 24 часов после начала эксперимента, фактически нет подавленных биофункций. При этом промежуток времени с 3 до 24 часов после начала эксперимента происходит наибольший рост количества биофункций, достигших пика выраженности, а через 72 часа после начала эксперимента пика выраженности достигают лишь 9,2% всех выраженных биофункций. Т.е., в целом, наблюдается антисимметричный процесс изменения количества значимых биофункций с течением времени.

Следует отметить, что в начале процесса лечения (3-12 часов) выражены биофункции, связанные с реакцией на INF-γ (response to interferon-gamma), с реакцией иммунной системы (immune response). Через 12 часов большинство выраженных биофункций составляют биофункции сигнального пути (signaling pathway), через 24 часа - преимущественно выражены биофункции связывания (binding), причем, в ос-

новном, это биофункции связывания протеинов (protein binding), при этом похожее доминирование наблюдается через 72 часа, но уже для подавленных биофункций. Через 48 часов можно отдельно выделить группу выраженных биофункций, связанных с положительной регуляцией на процессы и активность (positive regulation), через 72 часа наиболее значимые выраженные биофункции, связаны с процессами, происходящими в ядре клетки (nucleus) и транскрипциями ДНК, мРНК, рРНК (transcription). Промежутку времени с 48 до 72 часов характерны выраженные биофункции, связанные с регуляцей экспресссии генов (regulation of gene expression). Для подавленных генов через 48 часов характерны биофункции, связанные с метаболическими процессами в клетке (metabolic process). Через 24 часа - характерны также некоторые биофункции, связанные с метаболическими процессами, но значимее биофункции, связанные с процессами в мембране ретикулума (reticulum membrane).

Для кластеров получено, что гены, попавшие в первый кластер, имеют 654 значимых биофункций, относящиеся ко второму кластеру имеют 1323 значимые биофункции, а в третьем кластере имеют 1149 значимых биофункций. В первом кластере большинство наиболее значимых биофункций составляют процессы биосинтеза аминокислот (amino acid biosynthetic process). Это значит, что процессы биосинтеза аминокислот не активны в начальный промежуток времени (до 12 часов) и активны с 48 часов после начала лечения. Во втором кластере нет группы похожих биофункций, но можно отдельно выделить функцию клеточного компонента (cellular component) и межклеточной адгезии (cell-cell adhesion). Если проанализировать список биофункций, которые попали в третий кластер, то можно заметить, что это биофункции, которые характеризуют иммунную реакцию клетки («innate immune response», «immune response», «immune response to tumor cell» и т.д.) и пути передачи сигнала (interferon-gamma-mediated signaling pathway, tumor necrosis factor-mediated signaling pathway и т.д.). Получается, что иммунная реакция клетки постоянно нарастает с начала лечения и в промежуток времени с 12 до 24 часов INF-γ максимально активна, после 24 часов активность иммунной системы постепенно падает, а после 72 часов иммунная система неактивна.

ЛИТЕРАТУРА

[1] *Nazarov, P.V.* Interplay of microRNAs, transcription factors and target genes: linking dynamic expression changes to function / Nazarov, P.V. [и др.] // Nucleic Acids Research 2013. 1-15

[2] *Irizarry, R.A.* Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data / Irizarry, R.A. [и др.] // Biostatistics V. 4. 2003. 249–264.

[3] *Tusher, V. G.* Significance analysis of microarrays applied to the ionizing radiation response / Virginia Goss Tusher, Robert Tibshirani, Gilbert Chu. // PNAS.,2001. Т. 98, 9. pp. 5116-5121.

[4] *Speed, T.* Statistical Analysis of Gene Expression Microarray Data: Clustering Microarray Data/ Speed, T. // Chapman and Hall/CRC, 2005. – 240 p.

[5] Айвазян, С. Прикладная статистика: Классификация и снижение размерности: Справ.изд. / С. А. Айвазян, В. М. Бухштабер, Е. С. Енюков, Л. Д. Мешалкин; Под ред. С. А. Айвазяна. М.: Финансы и статистика, 1989. 607 с.

[6] *Zeeberg B.R.* High-Throughput GoMiner, an 'industrial-strength' integrative gene ontology tool for interpretation of multiple-microarray experiments, with application to studies of Common Variable Immune Deficiency (CVID) / Zeeberg B.R. [and etc.] // BMC Bioinformatics. 2005.Vol. 6;. 168. p. 1-18.

[7] *Саечников, А. В.* Разработка метода главных компонент для анализа микрочипов ДНК / Саечников, А. В. // Сборник работ 69-й научной конференции студентов и аспирантов БГУ, 2012. С. 268-272

[8] *Саечников А. В.* Программный пакет GeneExpressionAnalyser для анализа микрочипов ДНК / Саечников А. В., Яцков Н.Н., Апанасович В.В. // Сборник тезисов докладов «Медэлектроника 2012», Мн., 2012. С. 79-81