ИНФОРМАТИКА

2014 **апрель-июнь** **№ 2**

ПРИКЛАДНЫЕ ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

УДК 004.67, 575.112:004

А.В. Саечников [1], Н.Н. Яцков [1], П.В. Назаров [2], Л. Валлар [2], В.В. Апанасович [1]

АНАЛИЗ ЭКСПРЕССИИ ГЕНОВ В РЕЗУЛЬТАТЕ ВОЗДЕЙСТВИЯ ИНТЕРФЕРОНА IFN-γ НА КЛЕТКУ С ИСПОЛЬЗОВАНИЕМ ПРОГРАММНОГО ПАКЕТА GeneExpressionAnalyser

Предлагается программный пакет GeneExpressionAnalyser для анализа данных, полученных в ходе проведения экспериментов с использованием биочипов ДНК. Детально рассмотрены алгоритмы предварительного анализа данных, выделения дифференциально-выраженных генов и анализа биологических функций клетки. Работоспособность пакета исследуется на примере опубликованных экспериментальных данных, представляющих результаты эксперимента по исследованию изменений экспрессии генов в клетке меланомы под воздействием интерферона IFN-γ с течением времени.

Введение

Развитие биотехнологий напрямую связано с разработкой эффективных методов и алгоритмов обработки большого объема информации, получаемой от биологических микрочипов различного назначения. Олигонуклеотидные микрочипы ДНК позволяют изучать экспрессию генов [1]. Анализ данных, полученных в ходе проведения экспериментов с использованием микрочипов ДНК, дает возможность выявить функционально связанные и несвязанные гены, определить доминирующие биофункции клетки. Реализация данной задачи требует значительной рутинной обработки и количественной интерпретации экспериментальных данных, однако алгоритмические возможности для их реализации лишь ограниченно предложены в стандартных открытых статистических пакетах обработки данных [2, 3]. Практически каждый из пакетов имеет определенные недостатки, например не позволяет исследовать данные с наличием пропущенных значений, дубликатов, выбросов; содержит ограниченный круг алгоритмов кластерного анализа экспрессии генов и их аннотаций [4, 5]. В экспериментах с микрочипами ДНК даже результаты предварительной обработки изображений зависят от типа микрочипа. Это условие накладывает ограничения на типы принимаемых программным обеспечением данных и дальнейшую обработку результатов. Существует множество разработанных программных пакетов, как лицензионных, так и находящихся в свободном доступе. Во многих пакетах, таких как BETR, CAGED, GeneNetWeaver, SNOMAD, GeneMAPP, реализован только определенный алгоритм анализа данных [6–8]. Достойной альтернативной платформой является среда статистического программирования *R* [9]. Среда *R* имеет ряд преимуществ: используется в открытом и бесплатном проекте Bioconductor для программных разработок, включает широкий набор статистических функций, в том числе функций для анализа биочипов ДНК [10–12]. Однако проект предоставляет лишь ограниченный набор программных средств для создания графических интерфейсов (например, tcltk, RGtk2, rpanel, gWidgets, Shiny), облегчающих работу пользователя, что в конечном счете автоматически требует от пользователя знания языка программирования *R* или хотя бы частичного понимания программного кода.

Существуют и другие пакеты программного обеспечения для комплексного анализа данных, получаемых с помощью биочипов ДНК, например Coral [13], DMET-Analyzer [14], eXframe[15]. Некоторые из пакетов являются бесплатными при условии дополнительного использования сторонних коммерческих программ, другие позволяют работать только со специализированными типами микрочипов. Программные пакеты Partek [16] и IPA [17] выполняют определенные этапы полного анализа данных. Поэтому их приходится использовать совместно

с другими программными продуктами, что требует дополнительных затрат на проведение работ по совместимости результатов входных и выходных данных. Пакет GoMiner [18] – мощный инструмент для выделения активных функций клетки, однако он имеет ряд недостатков. Во-первых, для получения списка значимых функций требуется использование дополнительных программных пакетов обработки данных и предоставление списка значимых генов. Во-вторых, архитектура пакета GoMiner основана на использовании интернет-подключения к удаленным базам данных, что приводит к значительному замедлению вычислений.

Целью данной работы является разработка программного комплекса для анализа широкого набора биочипов ДНК, интегрирующего основные этапы анализа данных: загрузку данных, предварительную обработку, выделение значимых генов, определение доминирующих функций клетки. Для всестороннего исследования разработанного программного пакета используются опубликованные наборы данных, полученные в ходе проведения эксперимента по исследованию воздействия интерферона IFN-γ на живую клетку человека [19].

1. Экспериментальные данные

Для демонстрации работоспособности пакета GeneExpressionAnalyser используются опубликованные экспериментальные данные [19], размещенные в хранилище ArrayExpress (индекс E-MEXP-3720). В работе изучается влияние интерферона IFN-γ на клеточную линию A375 меланомы. Клетки помещались в пробирки с питательной средой и культивировались в течение 96 ч. В определенные моменты времени клетки однократно подвергались воздействию IFN-γ (в концентрации 50 нг/мл). В качестве контрольных использовались клетки, не подвергавшиеся воздействию интерферона, а также клетки, в которых сигнальная система, состоящая из янус-киназы сигнального белка-трансдуктора и активатора транскрипции (JAK – STAT), заблокирована с помощью ингибитора янус-киназы I (JAK inhibitor I). Таким образом, были задействованы три биологических репликанта: клетки в отсутствии воздействия IFN-γ (ctrl), клетки с заблокированным сигнальным путем JAK – STAT (JII ctrl), а также культивированные клетки – в течение 3, 12, 24, 48 и 72 ч (далее моменты времени обозначены как 03H, 12H, 24H, 48H, 72H) после добавления IFN-γ. В точке нулевого момента времени рассматривается контрольный образец ctrl, а в точке 72 ч – контрольный образец JII ctrl. Для временных точек 03H, 12H проведены по два эксперимента с биочипами, для временных точек 24H, 48H и 72H – по три эксперимента для каждого образца. Из клеточного материала выделена РНК, которая после соответствующей обработки была помещена на микрочипы Affymetrix Gene Chip Human Gene 1.0 ST Array. Детальное описание эксперимента и его результатов приведено в [19]. Выполнен анализ 17 биочипов, представляющих временные точки и различные технические репликанты. На каждом биочипе размещено по 33 252 гена.

2. Программный пакет GeneExpressionAnalyser

В качестве основы среды разработки и реализации пакета выбрана программная среда Matlab с библиотекой Bioinformatics. Для построения графического интерфейса использовалась система GUIDE пакета MATLAB 7.11.0 (R2010b) для ОС Windows®. Среда Matlab выбрана по следующим причинам: 1) из-за возможности разработки хорошо зарекомендованных standalone-приложений, дружественного графического интерфейса и работы в режиме офлайн (достаточно нескольких подключений для обновления базы данных биофункций), что, несомненно, приставляет интерес для работы пользователей из стран СНГ; 2) надежности и оптимизации ядра среды, реализованного на языке программирования C++, для реализаций операций математических вычислений с матрицами, что значительно увеличивает скорость анализа и моделирования больших объемов многомерных данных; 3) наличия широкого набора отлаженных функций и программных инструментов для анализа биочипов ДНК, часть из которых существенно программно оптимизирована (что увеличивает эффективность и точность анализа). Разработанный пакет является бесплатным при условии распространения скомпилированного исполнительного файла с набором необходимых для работы библиотек.

Также бесплатно распространяется файл среды Matlab при условии наличия предварительно установленного пакета Matlab на компьютере-клиенте.

GeneExpressionAnalyser осуществлет следующие функции (рис. 1):

– загрузку и предварительную фильтрацию некачественных данных;

– предварительную обработку данных (опционально);

– нормировку данных [20–23];

– восстановление пропущенных значений [24];

– поиск дифференциально-выраженных генов (с использованием метода анализа значимости биочипов, далее используется англоязычная аббревиатура SAM (от Significance Analysis of Microarrays) [25]);

– иерархическую и неиерархическую кластеризацию [26];

– визуализацию данных с использованием метода главных компонент [27];

– выделение статистически значимых функций в ходе анализа аннотаций генов (англ. GeneOntology-анализ, GO [28]) с применением методов Фишера [18] и случайных перестановок [29].

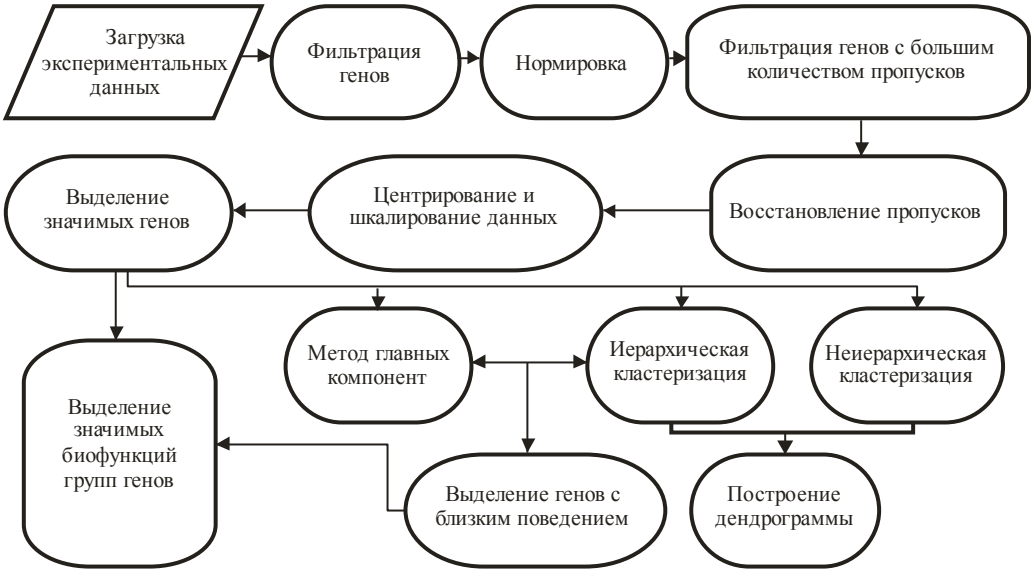


Рис. 1. Функциональная схема обработки данных с использованием GeneExpressionAnalyser

Рассмотрим подробно этапы анализа данных.

Загрузка данных. В программе предлагается выбрать один из трех вариантов формата загрузки данных:

1. Формат данных двухканальных микрочипов (*.gpr-файлы) [30]. После загрузки производится удаление из рассмотрения данных, для которых значение параметра *Flag* = 0 [30].

2. Формат данных чипов Affymetrix [31]. В этом случае требуется загрузка *.cel-файлов, которые содержат информацию об интенсивности люминесценции ячеек микрочипа, а также *.cdf-библиотеку, в которой интегрирована информация о ДНК-мишенях микрочипа. После загрузки предлагается выполнить фильтрацию генов с низким значением интенсивности экспрессии.

3. Загрузка данных из таблиц Excel, содержащих MA-значения [32] или непосредственно значения экспрессии генов [33].

Предварительная обработка данных (на примере двухканальных микрочипов). Процедура включает вычитание шума в каналах, определение уровня экспрессии и средней интенсивности в каналах:

$$M = log_2\left(\frac{R–Rb}{G–Gb}\right),\ A = \frac{1}{2}log_2((R–Rb)(G–Gb)),$$

где M – оценка уровня экспрессии гена; A – значение средней интенсивности флуоресценции сигнала гена в красном и зеленом каналах для некоторой ячейки биочипа; R, G – значения интенсивности ячейки биочипа в красном и зеленом каналах соответственно; Rb, Gb – фоновые значения интенсивности ячейки в красном и зеленом каналах соответственно.

Для чипов Affymetrix данная процедура не проводится, так как данные представляют собой логарифм интенсивности, пропорциональный логарифму концентрации РНК.

Нормировка. Для данных, снятых с помощью двухцветных матриц, применяется нормировка методом локального сглаживания графиков разброса (locally weighted scatter plot smoothing [20]). Этот шаг необходим для снижения влияния факторов, непосредственно связанных с конкретным биочипом. Для Affymetrix-данных производится RMA(Robust Multichip Average)-нормировка [21, 22] по каждому из чипов.

RMA-нормировка состоит из трех этапов. На первом этапе производится расчет фоновой компоненты для каждого биочипа в отдельности. (Предполагается, что фоновая компонента шума подчиняется нормальному распределению, а сигнальная – экспоненциальному распределению.) Фоновая компонента вычитается из набора данных.

На втором этапе производится квантильная нормировка по всему набору значений, в результате которой выравниваются эмпирические распределения интенсивностей проб для каждого биочипа. На последнем этапе производится суммирование интенсивностей проб в пробсеты с использованием метода медианной очистки (Median Polish [21, 22]).

Корректировка матрицы уровней экспрессии. Данная процедура связана с устранением пропущенных значений в таблице исходных данных. Пропущенные значения являются результатом повреждения биочипа или неполной гибридизации биологического материала. Если у гена пропущено больше некоторого определенного процента значений экспрессии (как правило, 33 % значений) в сводной таблице экспериментальных данных, то в дальнейшем данный ген исключается из последующего анализа. В разработанном пакете реализована процедура восстановления пропущенных значений. Для восстановления пропущенных значений используются метод k-ближайших соседей и его различные модификации [24]. Эффективность и качество методов восстановления пропущенных значений падают в случае анализа времязависимых данных [24]. Как правило, для чипов Affymetrix данная процедура не используется в силу высокого качества изготовления чипов, высокой плотности записи и точности.

В ходе выполнения данного этапа производится центрирование и шкалирование данных с целью устранения неоднородности в данных:

$$M_{ij} = \frac{x_{ij} - \overline{M}_j}{\delta(\overline{M}_j)},$$

где $\overline{M}_j$ и $\delta(\overline{M}_j)$ – математическое ожидание и среднеквадратическое отклонение по j-й колонке (биочип); M_{ij} – значение экспрессии i-го гена на j-м биочипе. Данная нормировка проводится по среднему всех значений экспрессии генов по каждому из биочипов в отдельности.

Для дальнейшего анализа производится отбор наиболее информативных данных. Простейший способ состоит в исключении генов с низкими уровнями экспрессии. Пороговое значение выбирается произвольно или экспертным решением. Производится группировка генов (усреднение значений экспрессии для экземпляров одного гена).

Выделение значимых генов методом SAM [25]. Метод SAM основан на анализе флуктуаций в экспериментальных данных. Для определения неслучайных флуктуаций вводится статистика

$$d_i = \frac{r_i}{s_i + s_0},$$

где r_i – нормированное значение экспрессии i-го гена; s_i – стандартное отклонение значения экспрессии гена; s_0 – малая постоянная величина, введенная для уменьшения зависимости d_i от уровня экспрессии гена.

Величина s_0 рассчитывается итерационно на основе значений r_i и s_i. Производится упорядочивание значений $d(i)$ в порядке убывания. Для каждого гена из набора случайно сгенерированных данных рассчитывается d_{pi}, затем гены выстраиваются в порядке убывания значений d_{pi}. Ожидаемое относительное различие в уровнях экспрессии d_e определяется как среднее по всем случайно сгенерированным данным. Следующий этап – построение диаграммы рассеяния наблюдаемого относительного различия от ожидаемого относительного различия в уровнях экспрессии. Гены, которые расположены на диаграмме рассеяния на расстоянии, большем чем Δ (данный параметр подбирается) от диагональной прямой, считаются значимыми. Оценочное значение ошибочно найденных значимых генов (false discovery rate – FDR) рассчитывается как среднее значение числа генов, названных значимыми для всех перестановок. Варианты метода SAM различаются в методике расчета величин r_i и s_i. Например, для однофакторного анализа значимости (SAM one class)

$$r_i = x_i = \sum_j \frac{x_{ij}}{n},\ s_i = \sqrt{\sum_j \frac{(x_{ij}-x_i)^2}{n(n-1)}},$$

где j – номер экспериментального значения для одного гена; x_i – среднее значение экспрессии гена; x_{ij} – значение экспрессии i-го гена в j-м биочипе; n – количество биочипов. Для многофакторного анализа значимости (SAM-multiclass)

$$x_{im} = \sum_j \frac{x_{ij}^m}{n};$$

$$r_i = \sqrt{(\{\frac{\sum n_m}{\prod n_m}\} \sum_{m=1}^{C} n_m (x_{im}-x_i)^2)};$$

$$s_i = \sqrt{(\frac{1}{\sum (n_m-1)}(\sum \frac{1}{n_m}) \sum_{m=1}^{C} \sum_{j \in C_m} (x_{ij}-x_{im})^2)},$$

где C – количество классов данных; C_m – индекс наблюдений в классе данных m; n_m – количество наблюдений в классе m. Подробная информация по методике расчета SAM представлена в работе [29].

В программном пакете реализованы различные варианты анализа методом SAM (однофакторный, двухфакторный парный (непарный) и временной анализы). Метод SAM реализован на базе опубликованной *R*-функции [34]. В результате анализа данных с использованием метода SAM формируются списки дифференциально выраженных генов.

Кластеризация. Профили экспрессии отражают изменение транскрипции в зависимости от внешних условий или концентрации веществ, поэтому в пакете GeneExpressionAnalyser реализована возможность классификации генов и биочипов. Реализованы методы неиерархического (метод *k*-средних [26]) и иерархического кластерного анализа [27, 35]. В результате кластерного анализа формируются основные профили генной экспрессии.

Анализ данных с использованием метода главных компонент. Метод главных компонент [36] реализован для оценки качества кластеризации данных, визуализации данных в пространстве низкой размерности, определения выбросов в пакете. Характерной особенностью метода являются выбор и оценка значимости не отдельных переменных, а информативных по совокупности групп переменных. Пользователю предоставляется возможность построения графиков разброса в пространстве различных главных компонент (рис. 2).

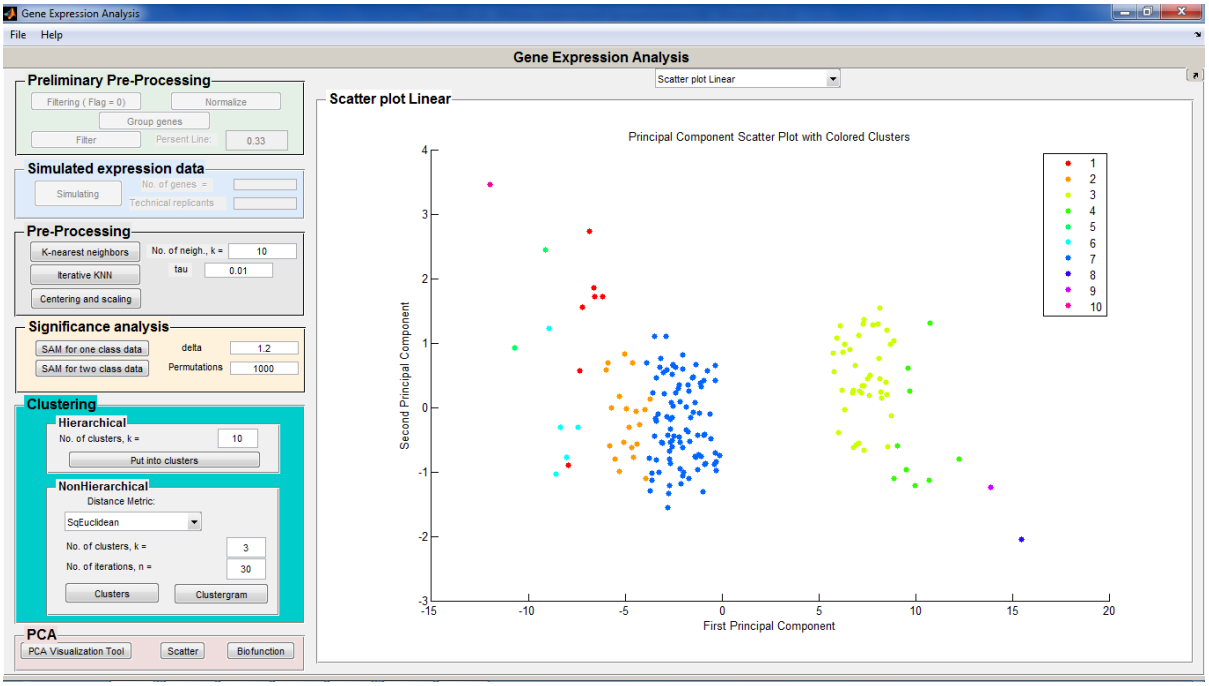


Рис. 2. Рабочее окно GeneExpressionAnalyser

Выделение значимых функций генов в ходе анализа генных аннотаций. Данный этап достаточно полно реализован в программном пакете GoMiner [18, 29], а также в программном пакете IPA® (Ingenuity Pathway Analysis). За алгоритмическую основу в разработанном пакете взяты (и усовершенствованы) подходы, используемые в названных выше пакетах. При этом оставлена возможность подключения к удаленным GO-базам данных. Достоинством GeneExpressionAnalyser является возможность определения значимых аннотаций для подгрупп дифференциально выраженных генов. В программе реализован метод одностороннего точного критерия Фишера [27]. Метод используется для определения статистической значимости категорий аннотаций онтологии генов, характеризуемой p-величиной [27]. Пользователь выбирает требуемый уровень p-значения (например, p = 0,05), после чего формируется список GO-аннотаций.

Большинство ключевых параметров анализа вводятся пользователем с помощью стандартного оконного интерфейса ОС Windows® (см. рис. 2). Вывод промежуточных и итоговых результатов осуществляется в виде графиков и внешних объектов (баз данных для возможности анализа 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 открытия и исследования. Работоспособность отдельных составляющих программного пакета подтверждена и исследована на примерах смоделированных и опубликованных экспериментальных данных [37, 38].

3. Имитационное моделирование

В GeneExpressionAnalyser интегрирована имитационная модель, генерирующая выражения экспрессии генов в эксперименте с биочипами ДНК. Имитационная модель предназначена с целью оценки устойчивости и проверки работоспособности разрабатываемых алгоритмов пакета, а также моделирования пользователем различных условий и возможностей проведения экспериментов с биочипами ДНК. Модель воспроизводит матрицы уровней экспрессии генов. Этапы обработки данных до этапа центрирования и шкалирования не учитываются. В модели задается вид профиля экспрессии группы генов, затем накладывается шум [39]. Добавление гауссова шума к профилю экспрессии производится следующим образом:

$$M_{sim}=M_{def}+\sigma*z,$$

где M_{def} – априорный профиль экспрессии в логарифмической шкале; σ – среднеквадратическое отклонение шума (выбирается исходя из оценки параметров распределения шума, наблюдаемого в конкретном эксперименте); z – реализация случайной величины со стандартным нормальным распределением $N(0,1)$.

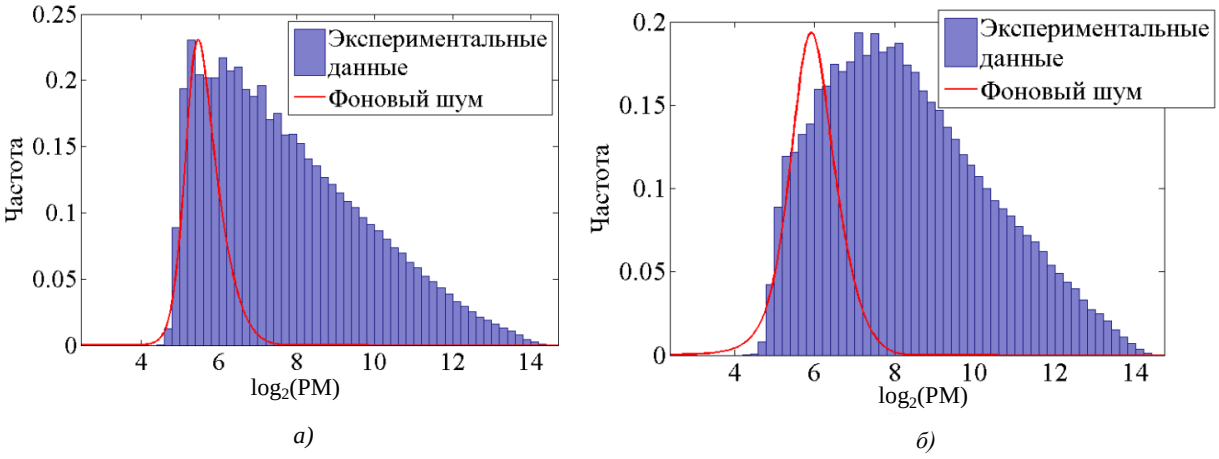
Модель способна генерировать данные времязависимых экспериментов, для чего задается изменяющийся во времени априорный профиль экспрессии, а также биочипы размером до 50 000 ячеек. Количество дифференциально выраженных генов не должно превышать 20 % от общего числа генов. Ограничение обусловлено трудоемкостью вычислений при проведении многофакторного и однофакторного временного SAM-анализов, так как для работы данного метода необходимо одновременно работать с данными большого количества биочипов. В работе производилось моделирование данных для проведения проверки работоспособности программного обеспечения в целом и отдельных методов, входящих в состав пакета. Модель можно применять для моделирования значений экспрессии PM(Perfect Match)-проб генов (25-мерных последовательностей, комплементарных эталонной на биочипе). В данном варианте моделирования значения экспрессии характеризуются экспоненциальной плотностью распределения.

4. Результаты исследования

В ходе проверки работоспособности пакета GeneExpressionAnalyser выполнен анализ 17 наборов экспериментальных данных. Рассмотрим основные результаты, полученные на каждом из этапов анализа. Ввиду высокого качества данных фильтрация с большим количеством пропусков не производилась. Выполнена RMA-нормализация набора данных. На первом этапе нормировки определены плотности интенсивности фоновой компоненты для каждого биочипа, относительно которых в дальнейшем были отнормированы значения экспрессии генов.

На рис. 3. показано распределение значений интенсивности $\log_2$(PM) [33], полученной после предварительной обработки изображений микрочипов. В результате нормировки преобразованы матрицы уровней экспрессии для каждого технического репликанта.

На этапе выделения дифференциально значимых генов выполнены следующие варианты метода SAM (FDR < 0,05): однофакторный, двухфакторный парный, двухфакторный непарный, однофакторный временной, многофакторный.



а) *б)*

Рис. 3. Гистограммы распределения интенсивности $\log_2$(PM) и рассчитанной плотности фоновой компоненты:
а) первый технический репликант в момент времени 12 ч после начала воздействия;
б) второй технический репликант в момент времени 48 ч после начала воздействия

Однофакторный, двухфакторный парный, двухфакторный непарный SAM-методы использовались для выделения значимых генов в рамках каждой временной точки эксперимента, однофакторный временной и многофакторный SAM-методы – для получения списка значимых генов по всему эксперименту. Для метода SAM one class значения экспрессии нормировались относительно среднего значения контроля, что давало возможность воспринимать данные в рамках одной временной точки как данные одного класса.

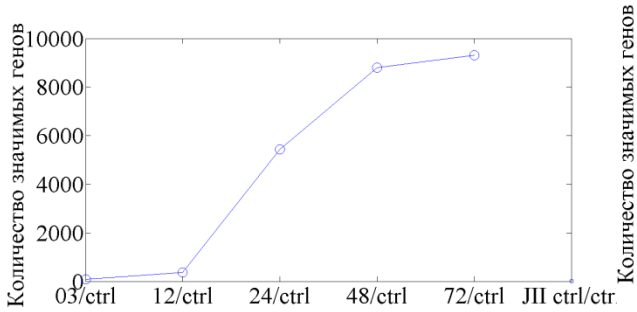
Для двухфакторного парного и двухфакторного непарного SAM-методов для каждой временной точки строились два класса данных: контрольные и подверженные влиянию IFN-γ. В случае парного метода были синхронизированы значения для различных микрочипов. Для микрочипа, представляющего третий контрольный образец (культивированные клетки после добавления IFN-γ), было сформировано контрольное значение как среднее между контрольными значениями первого и второго микрочипов). Конечной целью SAM-метода является выделение как можно большего числа значимых генов с сохранением пропорции ошибочных генов ниже заданного порога, поэтому исходя из полученных данных (рис. 4, *а*) было решено использовать результаты, полученные вариантами двухфакторного парного и однофакторного методов.

Результаты обработки методом SAM

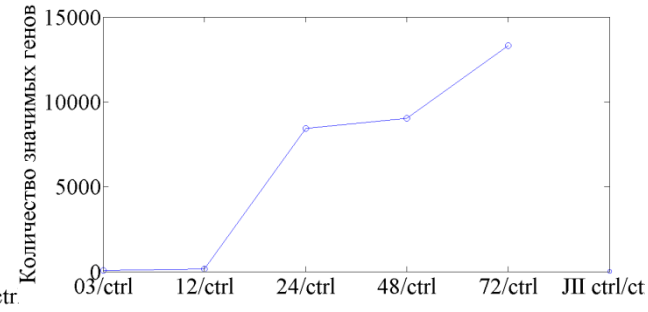
Тип метода SAM	03H		12H		24H		48H		72H		JII ctrl	
	Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg
Однофакторный	59	0	158	0	2634	5791	2103	6923	4740	8580	3	0
Двухфакторный парный	71	10	337	40	1549	3887	2510	6281	3361	5954	8	1
Двухфакторный непарный	60	0	427	0	772	1279	1030	3607	2306	4942	131	85

Однофакторный временной	Pos	Neg	FDR	Многофакторный	Significant	FDR
	2673	481	0,038		1242	0,04

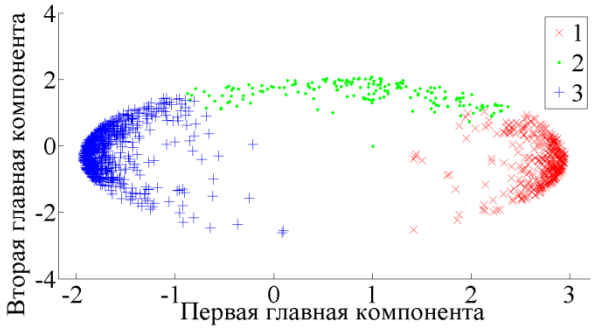
а)



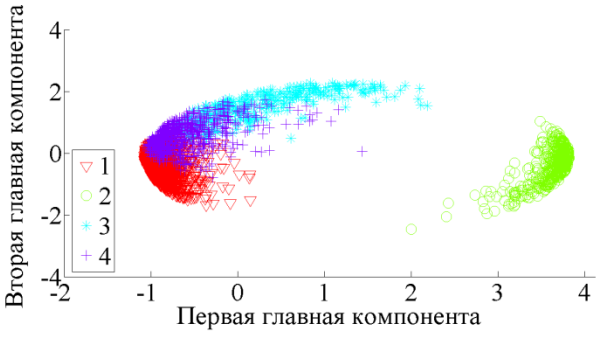
б)



в)



г)



д)

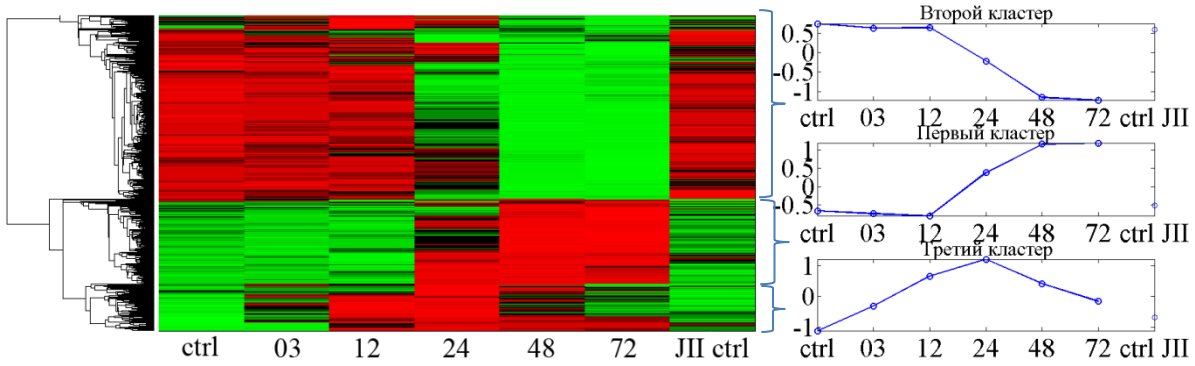
Рис. 4. Результаты дифференциального анализа с использованием SAM-метода: *а)* результат обработки SAM-методом (Pos – выраженные гены, Neg – подавленные гены); зависимости количества значимых генов от времени:*б)* SAM two class paired; *в)* SAM one class; графики разброса экспрессий генов в первых двух главных компонентах (маркерами обозначены различные кластеры генов): *г)* результаты для многофакторного SAM: первая компонента – 70,1 % дисперсии всех признаков; вторая – 15,4 % дисперсии всех признаков; *д)* результаты для однофакторного временного SAM: первая компонента – 62,7 % дисперсии всех признаков; вторая – 15,6 % дисперсии всех признаков

Необходимо отметить, что максимальный эффект воздействия на клетку IFN-γ в обоих случаях наблюдался в промежуток времени с 12 до 24 ч после начала воздействия. При этом в 24 ч около 70 % всех значимых генов являлись подавленными генами. Первая реакция клетки на IFN-*γ* наблюдается по прошествии определенного промежутка времени. Максимальный эффект воздействия интерферона падает после 48 ч (изменение количества значимых генов с 48 до 72 ч мало́). Результаты однофакторного SAM-метода демонстрируют ослабление воздействия интерферона в промежуток времени с 24 до 48 ч и последующее его усиление (рис. 4, *б, в*). Несмотря на то что количество значимых генов в случае однофакторного анализа больше, для дальнейшего анализа оставлены результаты двухфакторного парного SAM-метода, так как сами входные данные больше подходят под определение двухфакторных данных: класс данных до воздействия и класс данных после воздействия. Следует отметить, что в случае двухфакторных данных производились наименьшие изменения матриц уровней экспрессии, что является дополнительным фактором в пользу выбора данного результата. В случае двухфакторного парного SAM-метода характер изменения количества значимых генов аналогичен характеру изменений, полученному с использованием пакета limma [19].

Для последующей кластеризации используются дифференциально выраженные гены, полученные многофакторным и однофакторным временным SAM-методами. В ходе анализа результатов многофакторного SAM-метода каждая временная точка и контрольные данные воспринимались как отдельный класс данных, а для однофакторного временного SAM-метода все данные нормированы с учетом контрольных значений и обрабатывались в совокупности. Выполнена иерархическая кластеризация дифференциально выраженных генов. Оценка качества иерархической кластеризации произведена при помощи кофенетического корреляционного коэффициента и визуально с использованием дендрограмм и графиков разброса в пространстве первых главных компонент. В качестве критерия определения количества кластеров был выбран коэффициент несоответствия (inconsistency) [40]. Установлено предпочтение результату, полученному многофакторным SAM-методом (рис. 4, *г, д*), в силу пространственной разделенности кластеров генов и четкой выраженности областей кластеров. В результате иерархического кластерного анализа определены три главных кластера генов (рис. 5).

Первый кластер (305 генов) характеризует первоначально подавленные гены, экспрессия которых оставалась почти неизменной в течение первых 12 ч. Затем экспрессии этих генов достигают максимума в промежуток времени от 12 до 48 ч после начала воздействия, причем гены становятся выраженными и уже не изменяют свое состояние. Аналогичные результаты наблюдаются при анализе воздействия IFN-γ на второй кластер генов (637 генов), только гены со временем подавляются. Гены, относящиеся к третьему кластеру (149 генов), начинают реагировать на IFN-γ сразу же после начала воздействия и меняют свое состояние с подавленного на выраженное за 12 ч после начала воздействия, после чего с 12 до 24 ч практически не меняют свое состояние, в промежуток времени с 24 до 72 ч – подавлены.

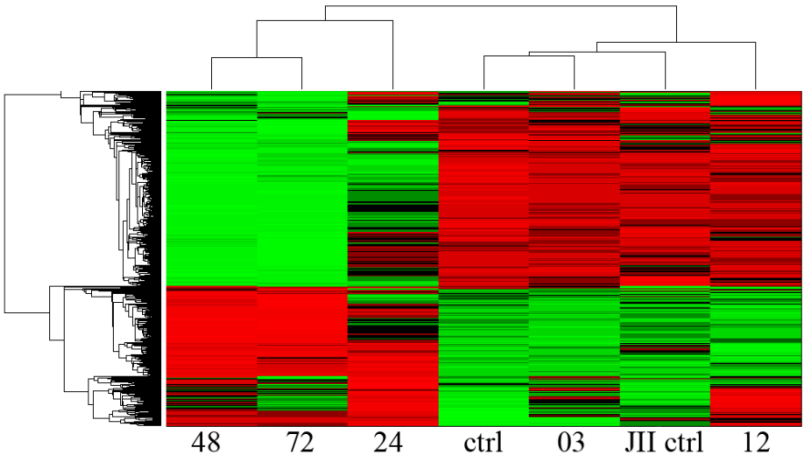
Значения экспрессии генов, которые получены для клеток, обработанных блокирующим сигнал от интерферона ингибитором, аналогичны результатам контрольных необработанных клеток. Это подтверждает вывод о том, что изменения экспрессии обусловлены воздействием IFN-γ на клетки, но не изменениями в самих клетках с течением времени. Данный вывод также подтверждается результатами анализа SAM-методом по каждой временной точке, где количество значимых генов для JII ctrl мало. Дендрограмма на рис. 5, *в* показывает кластеризацию генов не только по изменениям в течение воздействия, но и по временным точкам. Если остановиться на втором уровне разветвления дерева кластеризации, можно выделить четыре основных кластера: первый – 48 и 72 ч; второй – 24 ч; третий – ctrl, 03, JII ctrl; четвертый – 12 ч после начала воздействия. На рис. 5, *б* указан список наиболее известных генов, транскрипционные изменения которых типичны для воздействия клетки IFN-γ и были подтверждены в работах других авторов [19].



а)

Кластеры		
первый	второй	третий
AAK1	A2M	ACSL5
AARS	AASS	ADAP1
ATG12	COL5A2	CXCL10
ATL3	CPVL	CXCL11
DEPDC7	CRISPLD 1	IRF1
HDAC9	IFT122	SECTM1
HERPUD 1	IGFBP2	SERPING 1
HRH1	IGFBP5	SLC12A7
IER2	MMP1	SMOX
IL1A	SOX5	SOCS3
TMCC3	THYN1	SP110
TMEM15 4	TIAM1	SPRY4
TMEM88	VCAN	STAT1

б)



в)

Рис. 5. Результаты кластеризации: *а)* дендрограмма кластеризации, показывающая изменения экспрессии генов с течением времени (справа представлены профили экспрессии кластеров); *б)* список наиболее характерных генов для каждого кластера, названия генов представлены в виде HGNC(HUGO GeneNomenclatureCommittee)-кодов; *в)* дендрограмма кластеризации профилей экспрессии и временных точек (слева – дерево кластеризации профилей, сверху – дерево кластеризации временных точек)

На завершающем этапе анализа данных определены активные биофункции клетки, контролируемые выраженными генами, полученными двухфакторным парным SAM-методом. Дополнительный анализ выполнен над кластерами генов, полученных многофакторным SAM-методом для определения значимых функций, активность которых изменялась в соответствии с профилями экспрессии кластеров. Статистическая значимость биофункций определялась с помощью точного критерия Фишера ($p < 0{,}01$). В результате анализа группы генов, полученных двухфакторным парным SAM-методом, определены 2975 значимых функций (2419 биологических и 556 молекулярных), характерных для выраженных генов, и 2706 значимых функций (1487 биологических и 1219 молекулярных функций), характерных для подавленных генов. Часть полученных результатов, которая качественно описывает общую совокупность функций (указаны для каждой группы наиболее значимые функции), можно найти по адресу http://sablab.net/students/saetchnikov/ в табл. 1. Полученные результаты можно разделить на пять основных групп функций в соответствии с тем, что функции выраженных либо подавленных генов имеют доминирующую значимость в определенный момент времени. В 3 ч после начала воздействия пик выраженности достигают 6,6 % всех статистически значимых функций, в 12 ч – 22,9 %, в 24 ч – 44,9 %, в 48 ч – 14,2 %, в 72 ч – 9,2 %, в 72 ч с ингибитором – 2,2 % всех выраженных функций. Для подавленных функций пик подавленно-

сти в 12 ч после начала воздействия достигают 3,5 % функций, в 24 ч – 19,2 %, в 48 ч – 25,3 %, в 72 ч – 52 % всех выраженных функций.

Абсолютное большинство функций достигают пика подавленности в конце эксперимента, тогда как вплоть до 24 ч после начала эксперимента фактически нет подавленных функций. При этом в промежуток времени с 3 до 24 ч после начала эксперимента происходит наибольший рост количества функций, достигших пика выраженности, а в 72 ч после начала эксперимента пика выраженности достигают лишь 9,2 % всех выраженных функций. В целом наблюдается асимметричный процесс изменения количества значимых функций с течением времени.

В начале процесса воздействия (3–12 ч) выражены биологический процесс реакции на IFN-γ (response to interferon-gamma) и биологический процесс реакции иммунной системы (immune response). Через 12 ч после начала воздействия большинство выраженных функций составляют биологические процессы сигнального пути (signaling pathway). В 24 ч преимущественно выражены молекулярные функции связывания (binding), причем в основном это молекулярные функции связывания белков (protein binding), при этом похожее доминирование можно наблюдать в 72 ч после начала воздействия, но уже для подавленных функций. В 48 ч можно отдельно выделить группу выраженных функций, связанных с положительной регуляцией процессов активности (positive regulation), в 72 ч после начала воздействия наиболее значимы выраженные молекулярные функции связывания ионов (ion binding) и транскрипции (transcription). Для подавленных генов в 12 ч характерны функции, связанные с положительной либо отрицательной регуляцией процессов в клетке (regulation), в 48 ч – связанные с метаболическими процессами в клетке (metabolic process). Для 24 ч после начала воздействия характерны также некоторые функции, связанные с метаболическими процессами, но более значимые, чем функции биосинтетического процесса (biosynthetic process).

Для кластеров получены следующие результаты:

1) гены, которые попали в первый кластер, имеют 638 (554 биологических, 84 молекулярных) значимых функций;

2) гены, относящиеся ко второму кластеру, имеют 1200 (788 биологических, 412 молекулярных) значимых функций;

3) гены третьего кластера имеют 1096 (838 биологических, 258 молекулярных) значимых функций.

В первом кластере большинство активных функций отражают процесс биосинтеза аминокислот (amino acid biosynthetic process). Активность процессов биосинтеза аминокислот увеличивается с 12 ч и достигает максимума в 48 ч после начала воздействия. Во втором кластере можно отдельно выделить функции сборки (assembly) и метаболических процессов (metabolic process). В третьем кластере функции характеризуют иммунную реакцию клетки (innate immune response, immune response, immune response to tumor cell etc) и пути передачи сигнала (interferon-gamma-mediated signaling pathway, tumor necrosis factor-mediated signaling pathway etc). Иммунная реакция клетки постоянно нарастает с начала воздействия и в промежуток времени с 12 до 24 ч с начала воздействия IFN-γ максимально активна. После 24 ч активность иммунной системы постепенно падает, а в 72 ч после начала воздействия иммунная система неактивна.

5. Сравнение результатов

Приведем сравнение полученных и опубликованных в [19] результатов. В силу того что конечные данные (значимые функции) получены с помощью разных методик и в различных программных обеспечениях, напрямую сравнить их затруднительно. Для сравнения количества выраженных функций в заданные моменты времени (03H, 12H, 24H, 48H, 72H) по отношению к нулевому отсчету был использован веб-ресурс http://biocompendium.embl.de/. Использование данного ресурса позволяет исключить неопределенность, связанную с вопросами стандартизации протокола выделения биофункций в работе [19]. Анализ выполнен как для списков генов, полученных в работе [19], так и для результатов данной работы. Результаты представлены в таблице. Необходимо отметить, что в результате работы

GeneExpressionAnalyser почти по всем временным точкам было получено большее число значимых функций (в среднем на 44 %), чем в работе [19]. Процент общих значимых функций в среднем составляет 68 % от количества значимых функций, которые были получены в результате анализа, проведенного в работе [19].

Количество значимых функций по временным точкам

Источники сравниваемых данных	03Н	12Н	24Н	48Н	72Н
Работа [19]	4	16	24	47	34
Результаты данных исследований	3	17	47	66	48
Общие результаты для данных исследований и работы [19]	3	12	15	31	24

Вероятно, большее количество значимых функций, выделенных программным пакетом GeneExpressionAnaliser, нежели функций, выделенных с помощью ресурса http://biocompendium.embl.de/, является результатом обновления базы данных GO-аннотаций, а также учета молекулярных функций.

Заключение

Разработан программный комплекс GeneExpressionAnalyser для анализа широкого набора биочипов ДНК, интегрирующий основные этапы анализа данных, такие как: загрузка данных, предварительная обработка, выделение значимых генов, определение доминирующих функций клетки. Работоспособность алгоритмов программного пакета подтверждена на примерах анализа смоделированных и экспериментальных данных. Пакет GeneExpressionAnalyser имеет следующие преимущества над существующим программным обеспечением в данном сегменте экспериментальных исследований:

1) свободно распространяется в пределах Республики Беларусь, требует своевременного обновления базы данных аннотаций. Распространяется программный пакет как исполнительный файл с набором библиотек, что значительно упрощает работу пользователя;

2) позволяет исследовать биочипы Affymetrix, двухцветные биочипы, а также наборы данных, представленные в табличном виде;

3) содержит широкий набор методов многомерного анализа данных;

4) выполняет эффективный анализ аннотаций списка дифференциально выраженных генов.

В работе приведены результаты исследований изменений экспрессии выраженности генов в клетке меланомы A375 под воздействием IFN-γ с течением времени с помощью программного пакета GeneExpressionAnalyser. Полученные результаты воспроизводят опубликованные ранее данные, что подтверждает работоспособность разработанных методов анализа данных и программного обеспечения. Дополнительный материал по теме данных исследований можно найти по адресу http://sablab.net/students/saetchnikov/.

Список литературы

1. Свешникова, А.Н. Экспрессия генов и микрочипы: проблемы качественного анализа / А. Н. Свешникова, П.С. Иванов // Рос. хим. ж. – 2007. – Т. 51 (1). – С. 127–135.

2. Maciejewski, H. Gene set analysis methods: statistical models and methodological differences / H. Maciejewski // Brief Bioinform. – 2013. – Feb. 14. – P. 1–15.

3. Assessment of gene set analysis methods based on microarray data / H.A. Majd [et al.] // Gene. – 2013. – Vol. 534. – P. 383–389.

4. SplicerEX: a tool for the automated detection and classification of mRNA changes from conventional and splice-sensitive microarray expression data / T.J. Robinson [et al.] // RNA. – 2012. – Vol. 18(8). – P. 1435–1445.

5. BEAT: Bioinformatics Exon Array Tool to store, analyze and visualize Affymetrix GeneChip Human Exon Array data from disease experiments / A. Consiglio [et al.] // BMC Bioinformatics. – 2012 – 13(Suppl 4): S21. – P. 1–14.

6. Bar-Joseph, Z. Studying and modelling dynamic biological processes using time-series gene expression data / Z. Bar-Joseph, A. Gitter, A. Simon // Nature Reviews Genetics. – 2012. – Vol. 13. – P. 552–564.

7. Mehta, J.P. Software and tools for microarray data analysis / J.P. Mehta, S. Rani // Methods Mol Biol. – 2011. – Vol. 784. – P. 41–53.

8. Pathway analysis software: annotation errors and solutions / N.K. Henderson-Maclennan [et al.] // Mol Genet Metab. – 2010. – Vol. 101(2–3) – P. 134–140.

9. Brian, S. E. Handbook of Statistical Analyses Using R / S.E. Brian, T.A. Hothorn. – Chapman and Hall/CRC, 2009. – 376 p.

10. Bioconductor: open software development for computational biology and bioinformatics / R.G. Gentleman [et al.] // Genome Biology – 2004. – Feb. 14. – P. 80.1–80.16

11. Count-based differential expression analysis of RNA sequencing data using R and Bioconductor / S. Anders [et al.] // Nat Protoc. – 2013. – № 8(9). – P. 1765–1786.

12. BeadArray Expression Analysis Using Bioconductor / M. Dunning [et al.] // PLoSComput Biol. – 2011. – № 7(12). – P. 1–39.

13. Coral: an integrated suite of visualizations for comparing clusterings/ D. Filippova, A. Gadani, C. Kingsford // BMC Bioinformatics – 2012. – Vol. 13:276. – P. 1–13.

14. DMET-Analyzer: automatic analysis of Affymetrix DMET Data / P.N. Guzzi [et al.] // BMC Bioinformatics – 2012. – Vol. 13:258. – P. 1–10.

15. eXframe: reusable framework for storage, analysis and visualization of genomics experiments/ A.U. Sinha [et al.] // BMC Bioinformatics – 2011. – Vol. 12:452. – P. 1–13.

16. Next Generation Sequencing & Microarray Data Analysis Software | Partek Incorporated [Electronic resource]. – 2013. – Mode of access : http://www.partek.com. – Date of access : 15.11.2013.

17. Ingenuity IPA – Integrate and understand complex omnics data [Electronic resource]. – 2013. – Mode of access : http://www.ingenuity.com/products/ipa. – Date of access : 16.11.2013.

18. GoMiner: a resource for biological interpretation of genomic and proteomic data / B.R. Zeeberg [et al.] // Genome Biol. – 2003. – Vol. 4 (4). – Art. R28. – P. 1–8.

19. Interplay of microRNAs, transcription factors and target genes: linking dynamic expression changes to function / P.V. Nazarov [et al.] // Nucleic Acids Research. – 2013. – Vol. 41(5). – P. 2817–2831.

20. Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data / R.A. Irizarry [et al.] // Biostatistics. – 2003. – Vol. 4(2). – P. 249–264.

21. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation / S. Dudout [et al.] // Nucleic Acids Research. Oxford University Press. – 2002. – Vol. 30, 4 e15. – P. 1–10.

22. Comparison of Affymetrix data normalization methods using 6,926 experiments across five array generations / R. Autio [et al.] // BMC Bioinformatics. – 2009. – 10 (Suppl 1): S24. – P. 1–12.

23. Analysis of boutique arrays: A universal method for the selection of the optimal data normalization procedure / B. Uszczyńska [et al.] // Mol. Med. – 2013. – Sep. 32(3). – P. 668–684.

24. Bra´s, Lı´gia P. Improving cluster-based missing value estimation / Lígia P. Brás, José C. Menezes // Biomolecular Engineering. – 2007. – T. 24. – P. 273–282.

25. Tusher, V.G. Significance analysis of microarrays applied to the ionizing radiation response / V.G. Tusher, R. Tibshirani, G. Chu // PNAS. – 2001. – Vol. 98,9. – P. 5116–5121.

26. Speed, T. Statistical Analysis of Gene Expression Microarray Data: Clustering Microarray-Data / T. Speed // Chapman and Hall/CRC. – 2005. – 240 p.

27. Прикладная статистика: классификация и снижение размерности : справ. изд. / С.А. Айвазян [и др.]. – М. : Финансы и статистика, 1989. – 607 с.

28. The Gene Ontology [Electronic resource]. – 2013. – Mode of access : http://www.geneontology.org. – Date of access : 15.11.2013.

29. 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) / B.R. Zeeberg [et al.] // BMC Bioinformatics. – 2005. – Vol. 6; 168. – P. 1–18.

30. Molecular Devices Launches the GenePix(R) SL50 Slide Loader and GenePix(R) Pro 7.2 Software // PR Newswire Europe Including UK Disclose [Electronic resource]. – 2011. – Mode of access : http://www.prnewswire.co.uk/news-releases/molecular-devices-launches-the-genepixr-sl50-slide-loader-and-genepixr-pro-72-software-145284555.html. – Date of access : 16.11.2013.

31. Home | Affymetrix [Electronic resource]. – 2013. – Mode of access : http://www.affymetrix.com. – Date of access : 13.11.2013.

32. Advanced spot quality analysis in two-colour microarray experiments / M. Yatskou [et al.] // BMC Research Notes. – 2008. – Vol. 1:80. – P. 1–13.

33. Dynamic regulation of microRNA expression following Interferon-γ-induced gene transcription / S. Reinsbach [et al.] // RNA Biology. – 2012. – Vol. 9:7. – P. 978–989.

34. Samr: SAM: Significance Analysis of Microarrays [Electronic resource]. – 2011. – Mode of access : http://cran.r-project.org/web/packages/samr/samr.pdf. – Date of access : 13.11.2013.

35. Data Mining Practical Machine Learning Tools and Techniques / Ian H. Witten [et al.]. – The Morgan Kaufmann Series in Data Management Systems, 2011. – 664 p.

36. Hyvarinen, F. Independent Component Analysis / F. Hyvarinen. – Wiley series, 2001. – 505 p.

37. Разработка метода главных компонент для анализа микрочипов ДНК / А.В. Саечников // 69-я научная конф. студентов и аспирантов БГУ : тез. докл. – Минск, 2012. – С. 268–272.

38. Саечников, А.В. Программный пакет GeneExpressionAnalyser для анализа микрочипов ДНК / А.В. Саечников, Н.Н. Яцков, В.В. Апанасович // Медэлектроника 2012 : тез. докл. – Минск, 2012. – С. 79–81.

39. Novikov, E. An algorithm for automatic evaluation of the spot quality in two-color DNA microarray experiments / E. Novikov, E. Barillot // BMC Bioinformatics. – 2005. – Vol. 6: 293. – P. 1–18.

40. Uragun, B. The discrimination of interaural level difference sensitivity functions: development of a taxonomic data template for modelling / B. Uragun, R. Rajan // BMC Neuroscience. – 2013. – Vol. 14: 144. – P. 1–19.

Поступила 26.12.2013

[1]*Белорусский государственный университет,*
Минск, ул. Курчатова, д. 1
e-mail: saetchnikov.anton@tut.by

[2]*Центр геномных исследований,*
L-1526 Люксембург
e-mail: petr.nazarov@crp-sante.lu

A.V. Saetchnikov, M.M. Yatskou, P.V. Nazarov, L. Vallar, V.V. Apanasovich

ANALYSIS OF CELLULAR REACTION TO IFN-γ STIMULATION BY A SOFTWARE PACKAGE GeneExpressionAnalyser

The software package GeneExpressionAnalyser for analysis of the DNA microarray experimental data has been developed. The algorithms of data analysis, differentially expressed genes and biological functions of the cell are described. The efficiency of the developed package is tested on the published experimental data devoted to the time-course research of the changes in the human cell under the influence of IFN-γ on melanoma. The developed software has a number of advantages over the existing software: it is free, has a simple and intuitive graphical interface, allows to analyze different types of DNA microarrays, contains a set of methods for complete data analysis and performs effective gene annotation for a selected list of genes.