Программа GeneExpressionAnalyser предназначена для комплектного анализа данных, полученных с микрочипа.

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

- Формат данных двухканальных микрочипов (*.gpr-файлы). После загрузки производится удаление из рассмотрения данных, для которых значение параметра *Flag* = 0. Выбор соответствующих файлов производится в произвольной последовательности, производится до тех пор пока пользователь не нажмет кнопку отмена в диалоговом окне выбора.
- Формат данных чипов Affymetrix . 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. Сначала пользователю предлагается выбрать папку с *.cel-файлами, в следующем диалоговом окне необходимо выбрать *.cdf-библиотеку ДНК.
- Загрузка данных из таблиц Excel, содержащих MA-значения или непосредственно значения экспрессии генов.

После успешной загрузки данных область предварительной обработки данных становится доступной для пользователя. Фильтрация (Flag = 0) выполняется для *.gpr-данных. Нормализация производится для всех типов данных. Методы нормировки подробно описаны в сопровождающей программный пакет статье. В результате нормировки выводится графический результат нормировки в области вывода графиков программы. Численные результаты нормировки сохраняются в оперативную память.

Последующий этап выполнения анализа *.gpr данных является группировка генов, которая заключается в поиске по всему биочипу копий одного гена и узреднение значения экспрессии. Данный этап самый затратный, поэтому по выполнению данного этапа рекомендуется произвести сохранение рабочего пространства Matlab.

Корректировка матрицы уровней экспрессии. Данная процедура связана с устранением пропущенных значений в таблице исходных данных. Пропущенные значения могут появляться в результате повреждения биочипа (дефекты рабочей поверхности биочипа, обусловленные сбоями в работе устройств и т.д.). Процент допустимого процента пропуска экспрессии генов выбирается в соответствующем окне и после нажатия кнопки Filter производится удаление генов, у которых процент пропуска значений больше заданного.

После этого необходимо производить восстановление пропущенных значений для оставшихся генов. Подробное описание метода восстановления приведено в [1], [5].

Далее выполняется центрирование и шкалирование, После выполнения производится сохранение результатов, которые в дальнейшем используются для анализа методом SAM. Данные сохраняются в таблицу, где:

Первый лист содержит все данные, которые были получены после центрирования и шкалирования

Второй лист содержит нецентрированные и нешкалированные данные, которые были получены после нормировки

Третий лист содержит информацию о каждом столбце данных (т.е. биочипа).

Перед выполнением SAM анализа необходимо сохраненную таблицу, преобразовать таким образом, чтобы третий лист содержал необходимую информацию для правильного выполнения выбранного типа SAM-анализа. Более подробно как оформлять данную информацию описано в [2].

После выполнения SAM анализа пользователю предоставляется для анализа график построенный в ожидаемое различие/реальное различие. Кроме того формируется база данных *.mat с необходимой информацией о результатах анализа методом SAM и таблица с 3 листами, где на первом листе содержатся значении экспрессии значимых генов, на

втором – список значимых генов, а на третьем идентификаторы классов, которым соответствуют столбцы с первого листа.

После выполнения SAM-анализа может производиться кластеризация результатов. Для выполнения иерархической или неиерархической кластеризации надо выбрать таблицу с результатами SAM, где в третьем листе должны быть численные идентификаторы каждого класса данных, необходимых для построения дендрограммы. По результатам кластеризации в области вывода графической информации будут выведены графики профилей центров кластеров и графики содержащие профили всех генов каждого кластера. По результатам кластеризации производится сохранение результатов кластеризации в базу данных matlab.

Дендрограмму кластеризации можно построить отдельно по нажатии соответствующей кнопки, предварительно загрузив сохраненную на этапе кластеризации базу данных.

Для оценки качества кластеризации можно запустить метод главных компонент, предварительно загрузив данные кластеризации, таким же образом можно запустить внешний графический объект для анализа качества кластеризации в различных главных компонентах.

Заключительная функция – выделение значимых функций рассматриваемой биологической системы. На данном этапе есть возможность выбрать для анализа как данные, полученные в результате SAM анализа, так и на этапе последующей кластеризации. По умолчанию подгрузка базы данных генных аннотаций производится в режиме онлайн, по запросу можно запустить анализ в оффлайн режиме, предварительно сохранив базу данных аннотаций на ПК.

Задание параметров используемых при предварительном анализе данных микрочипа, а также для моделирования значений экспрессии осуществляется с помощью элементов, обозначенных на Рис. 18.

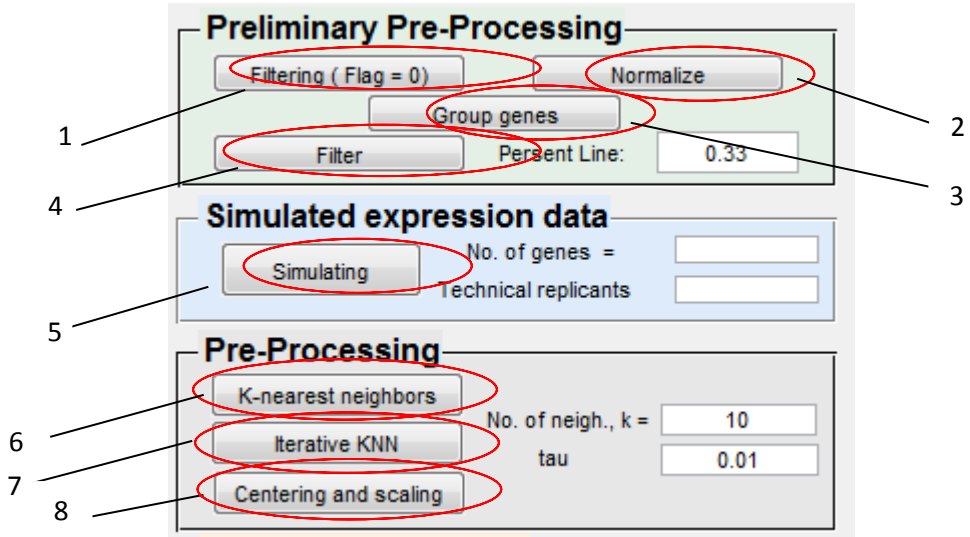

Рис. 1. Область задания параметров предварительной обработки и моделирования данных.

Краткое описание назначения этих элементов:

Блок простой предварительной обработки:

1) Кнопка фильтрации данных с низким качеством.
2) Кнопка нормировки загруженных данных. Осуществляется нормировка, по схеме, описанной в [5]. Необходимо отметить, что после нажатия данной кнопки открывается внешний объект, позволяющий контролировать результат, полученный двумя типами нормировок. Также данный объект можно открыть после выполнения всех нормировок, если нажать на кнопку вывода текущего графика во внешнее окно.

3) Кнопка группировки генов. В данной кнопке заложена функция корректировки значений по каждой микроматрице в отдельности, которая заключается в определении среднего значения экспрессии гена, в том случае, когда на микроматрице было предусмотрено несколько спотов для анализа одного гена.
4) Кнопка фильтрации значений, для которой задается параметр характеризующий процентную часть пропусков из всех значений. Функционирование данной операции подробно описано ранее.

Блок моделирования данных экспрессии генов:

5) Кнопка построения матрицы экспрессии по схеме, описанной в [5]. Для выполнения данной процедуры выведены области для ввода количества смоделированных генов, а также количества технических репликантов в смоделированном эксперименте.

Блок предварительной обработки :

6) Кнопка восстановления пропущенных значений методом *k*-ближайших соседей с модификацией, описанной с данной работе, для выполнения которого реализована возможность варьирования параметром количества ближайших соседей, с целью получения наилучшего результата.
7) Кнопка восстановления пропущенных значений итерационным методом *k*-ближайших соседей, для которого дополнительно выведена возможность регулирования параметра завершения итерационного процесса.
8) Кнопка центрирования и шкалирования данных, которая преобразует данные по схеме, описанной ранее.

Задание параметров используемых при статистическом анализе данных осуществляется с помощью элементов, обозначенных на Рис. 1.

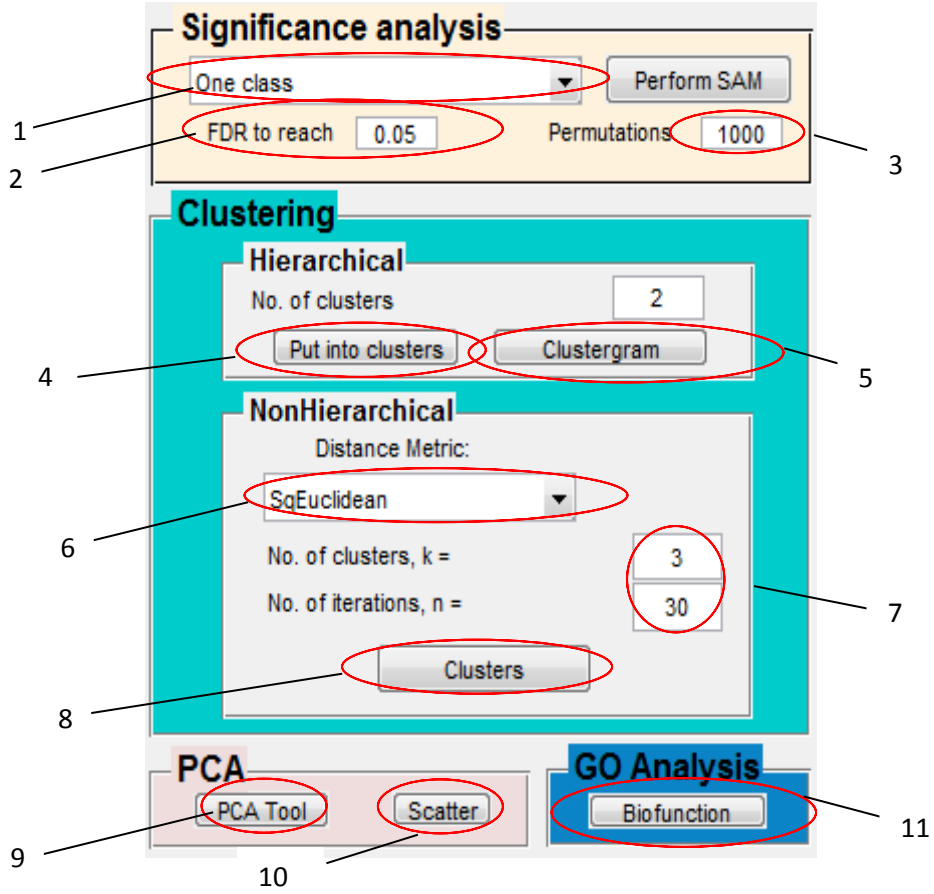

Рис. 2. Область задания параметров методов статистического анализа.

Краткое описание назначения этих элементов:

Блок метода SAM:

1) Кнопка выбора типа метода SAM для анализа данных.
2) Область выбора приемлемого FDR для SAM.

3) Количество выполняемых перестановок.

Блок иерархического кластерного анализа данных:

4) Кнопка иерархической кластеризации, результатом которой будет график профилей экспрессии генов каждых кластеров. Для работы необходимо задание количества кластеров.
5) Кнопка для построения дендрограммы по полученным результатам кластеризации.

Блок неиерархического кластерного анализа данных:

6) Пункт выбор метрики близости между генами.
7) Пункт задания количества кластеров, а также итераций. Данные параметры выведены в силу необходимости для использования метода k-средних.
8) Кнопка для запуска алгоритма неиерархической кластеризации

Блок выполнения метода главных компонент:

9) Кнопка запуска во внешнем окне объекта для просмотра и выделения генов с близким поведением, с возможностью представления в виде диаграммы рассеяния в любых интересующих главных компонентах.
10) Кнопка построения стационарной диаграммы рассеяния в первых двух главных компонентах с различными цветовыми метками для генов различных кластеров.
11) Кнопка для запуска анализа генных аннотаций.

Список литературы

1. Разработка метода главных компонент для анализа микрочипов ДНК / А.В. Саечников // Сб. работ 69-й научной конференции студентов и аспирантов БГУ. – 2012. – С. 268–272.

2. 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.

3. Саечников, А.В. Программный пакет GeneExpressionAnalyser для анализа микрочипов ДНК / А.В. Саечников, Н.Н. Яцков, В.В. Апанасович // Сб. тезисов докл. «Медэлектроника 2012». – 2012. – С. 79–81.

4. Анализ воздействия INT-γ на клетку с использованием программного пакета GENEEXPRESSIONANALYSER / А.В. Саечников [и др.] // International congress on computer science: information systems and technologies CSIST-2013, 4-7 November 2013 Minsk, – 2013. – С. 153–158

5. Анализ экспрессии генов в результате воздействия интерферона IFN-γ на клетку с использованием программного пакета GENEEXPRESSIONANALYSER / А.В. Саечников [и др.] // Информатика апрель-июнь 2014. [в печати] – 2014. – С. 5-18