ПРОГРАММНЫЙ ПАКЕТ *GENEEXPRESSIONANALYSER* ДЛЯ АНАЛИЗА МИКРОЧИПОВ ДНК

Саечников А.В., Яцков Н.Н., Апанасович В.В.

Белорусский государственный университет, кафедра системного анализа и компьютерного моделирования, Минск, Беларусь, E-mail: saetchnikov.anton@tut.by

Abstract. The software package ***GeneExpressionAnalyser*** for analysis of the DNA microarraydata has been developed. ***GeneExpressionAnalyser*** performs complete analysis of microarray data, real-time modeling of gene expression profiles in microarray, and calculation of the statistical significance for the biological functions of studied organisms. ***GeneExpressionAnalyser*** has been tested on the sets of simulated data and published experimental data. The designed package is a good alternative to commercial projects in the analysis and interpretation of multidimensional datasets of genetic information in biomedical research.

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 с 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]. Причем практически каждый из пакетов имеет определенные недостатки. Многие некоммерческие пакеты обработки данных не позволяют исследовать данные с пропусками. Архитектура пакета GoMiner основана на использовании интернет-подключения к удаленным базам данных, что приводит к значительному замедлению вычислений. Во многих пакетах реализован только определенный алгоритм анализа данных [2]. Достойной альтернативной платформой является среда статистического программирования R [3]. Пакет R имеет ряд преимуществ – является открытым и бесплатным проектом для программных разработок, включает широкий набор статистических функций и функций для анализа микроматриц ДНК. Однако проект не предоставляет программных средств для создания графических интерфейсов для пользователей. Что автоматически требует от пользователя знания языка программирования R или хотя бы частичного понимания программного кода. Существуют и другие менее эффективные аналоги программного обеспечения для комплексного анализа данных с биочипов ДНК, например: Plus 2.0 Array [4], NexusExpression™[5], GenomeStudioSoftware [6]. Данные пакеты являются коммерческими, некоторые из них позволяют работать только со специализированными типами микрочипов.

Целью работы является разработка программного обеспечения для комплексного анализа микрочипов ДНК, моделирования профилей выраженности генов в режиме реального времени, а также для определения значимости биологических функций исследуемых организмов.

Разработка 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 функции для анализа биочипов ДНК: загрузка данных, предварительная обработка и фильтрация данных, поиск

дифференциально-выраженных генов с использованием статистического метода SAM (SignificanceAnalysisofMicroarrays [7]), иерархическая и неиерархическая кластеризация генов с близким поведением [8], метод главных компонент [9], различные способы визуализации данных и профилей генов, выделение статистически значимых биофункций в ходе анализа генных аннотаций GO (GeneOntology)-анализ [10] методами точного теста Фишера [11] и случайных перестановок [12], экспорт и сохранение результатов анализа. Стоит отметить, что анализ генных аннотаций достаточно полно реализован в проекте GoMiner [12]. Однако GoMiner это зарубежный ресурс, требующий удаленного подключения и устойчивой линии связи. Что значительно снижает эффективность анализа данных в пределах нашей Республики. Функциональная схема работы программы *GeneExpressionAnalyser* представлена на **Рис 1**.

Рисунок 1 – Функциональная схема работы программного обеспечения.

Большинство ключевых параметров анализа вводятся пользователем с помощью стандартного оконного интерфейса ОС Windows® (*Рис. 1*). 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 открытия и исследования. Главное окно программы представлено на **Рис. 2**.

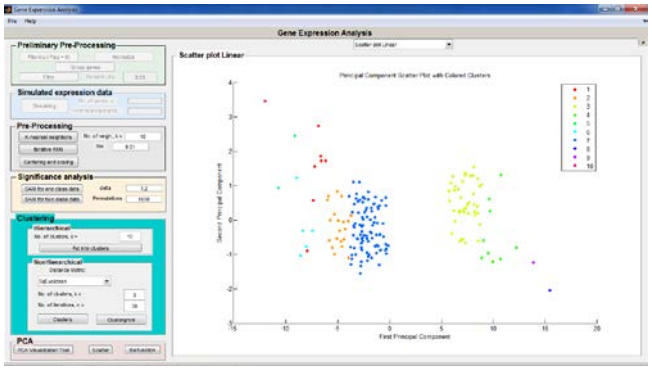

Рисунок 2 – Главное окно программы GeneExpressionAnalysis.

Результаты тестирования программного пакета *GeneExpressionAnalyser*

Программный пакет протестирован на примере опубликованных экспериментальных данных [13], а также на смоделированных данных. В результате анализа 11710 генов с 8 технических биочипов-репликантов [13] выявлено 182 значимых гена. Смоделированные данные для тестирования сгенерированы методом составления матрицы уровней экспрессии генов. Суть моделирования заключается в том, что изначально априори задается вид

профиля экспрессии, затем накладывается шум [14]. На следующем этапе, в результате анализа данных с помощью метода главных компонент выделено 6 подгрупп генов с близким поведением. Для каждой из выделенных подгрупп, а также для выраженных и подавленных генов, определены статистически значимые GO-аннотации. Для смоделированных данных оценка значимости GO-аннотаций не проводилось.

Стоит отметить, что выполнение процесса предварительной обработки (а именно, исключения генов с процентом пропуска значений больше заданного, а также поиска и усреднения множества копий одного и того же гена) требует значительное количество машинного времени. В ходе проведенных исследований, вышеназванная процедура потребовала 35 минут (тактовая частота процессора– 2,1 ГГц). Целесообразно, после выполнения данной процедуры воспользоваться предлагаемой возможностью сохранения текущего рабочего состояния. Ресурсоемкой операцией является поиск и выделение биофункций генов в пакете *GeneExpressionAnalyser*. На исследованных данных в среднем требовалось 25 мин для определения значимых GO-аннотаций интересующих групп генов. Однако ресурс GoMiner [12] в ходе анализа этих же наборов данных и получения подобных результатов потребовал 4.5 часа. Увеличение производительности достигнуто за счет введения операции сортировки GO-аннотаций в базе данных.

Заключение

В работе представлен программный пакет *GeneExpressionAnalyser* для комплексного анализа микрочипов ДНК. Разработанный пакет является достойной альтернативой коммерческим проектам в области анализа и интерпретации многомерных наборов данных генетической информации. В дальнейшем планируется усовершенствование и автоматизация программного пакета, всестороннее тестирование пакета как на опубликованных данных, так и на смоделированных данных с целью выявления и устранения технических ошибок, улучшение методов анализа данных, а также расширение перечня обрабатываемых микрочипов.

Литература

[1] **Свешникова А.Н., Иванов П.С**.Экспрессия генов и микрочипы: проблемы качественного анализа // Рос. Хим. Ж., 2007. Т. 51(1). с. 127-135.

[2] **Ziv Bar-Joseph, Anthony Gitter, Itamar Simon**. Studying and modelling dynamic biological processes using time-series gene expression data // Nature Reviews Genetics. 2012. Vol. 13, p. 552-564.

[3] **Интернет-адрес**: http://www.r-project.org.

[4] **Интернет-адрес**: http://www.biocompare.com/19143-Human-Whole-Genome-Microarrays/84204-Human-Genome-U133-Plus-20-Array/

[5] **Интернет-адрес**: http://www.biodiscovery.com/software/nexus-expression/

[6] **Интернет-адрес**: http://www.illumina.com/software/genomestudio_software.ilmn

[7] **Virginia Goss Tusher, Robert Tibshirani, Gilbert Chu**. Significanceanalysisofmicroarrays applied to the ionizing radiation response // PNAS.,2001. Т. 98, 9. pp. 5116-5121.

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

[9] **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.

[10] **Интернет-адрес**:http://www.geneontology.org/

[11] **Zeeberg BR, Feng W, Wang G, Wang MD, Fojo AT[and etc.]**GoMiner: a resource for biological interpretation of genomic and proteomic data. // Genome Biol. 2003. Vol. 4(4), Art. R28. p. 1-8.

[12] **Barry R Zeeberg, Haiying Qin, SudarshanNarasimhan[and etc.]** 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) // BMC Bioinformatics. 2005.Vol. 6;. 168. p. 1-18.

[13] **Yatskou M, Novikov E, Vetter G, Muller A, Barillot E, Vallar L, Friederich E.** “Advanced spot quality analysis in two-colour microarray experiments” // BMC Research Notes, 2008 Sep 17;1(1):80, p 1-13.

[14] **E. Novikov, E. Barillot**. An algorithm for automatic evaluation of the spot quality in two-color DNA microarray experiments // BMC Bioinformatics. 2005.Vol. 6: 293. p. 1-18.