используемых кумулянтов, что приводит к повышению точности оценок параметров. Нахождение корней полинома позволяет сократить время нахождения оценок параметров по сравнению со временем решения системы уравнений с помощью численных методов.

Литература

1. *Lakowicz J. R.* Principles of Fluorescence Spectroscopy. Singapore, 2006.
2. *Krichevsky O., Bonnet G*. Fluorescence correlation spectroscopy: the technique and its applications // Reports on Progress in Physics. 2002, V. 65, P. 251-297.
3. *Muller J. D.* Cumulant Analysis in Fluorescence Fluctuation Spectroscopy // Biophys. J. 2004, V. 86, P. 3981-3992.
4. *Chen Y., Müller J., So P., Gratton E*. The Photon Counting Histogram in Fluorescence Fluctuation Spectroscopy// Biophys. J. 1999, V. 77, P. 553-567.
5. *Shingaryov I., Skakun V., Apanasovich V*. Photon Counts Simulation in Fluorescence Fluctuation Spectroscopy // Pattern Recognition and Information Processing (PRIP), 2009, Minsk. P. 178-182.

АНАЛИЗ БИОЛОГИЧЕСКИХ ДНК-МИКРОЧИПОВ С ИСПОЛЬЗОВАНИЕМ СРЕДЫ R

А. С. Рындин

1. ВВЕДЕНИЕ

Биочипы, или микроматрицы ДНК, оказали огромное влияние на развитие медико-биологических дисциплин, связанных с исследованием генов, включая онкологию, токсикологию, фармакологию, биологию развития [1]. Эксперименты с участием биочипов позволяют изучать функции генов, их взаимосвязь, биологические процессы с их участием, а также проводить множество других биологических исследований [1, 2].

Эксперимент с биочипами выдаёт данные об экспрессии десятков тысяч генов. Проанализировать такой огромный объём данных можно только с помощью вычислительной мощи компьютеров. Подобный анализ огромных массивов биологических данных составляет предмет самостоятельной области науки – биоинформатики, научной дисциплины на стыке биологии, информатики и вычислительной математики.

Целью данной работы является изучение статистических методов обработки экспериментальных биологических данных, полученных с помощью биочипов, и их реализация в среде R на примере опубликованных данных [3].

2. МЕТОДОЛОГИЯ

Обработка данных включает в себя фильтрацию ячеек данных (спотов ДНК) с низким качеством, нормировку и фильтрацию данных с большим количеством пропусков, восстановление пропущенных значений, выделение статистически значимых генов и кластерный анализ выделенных генов.

После загрузки и фильтрации данных значения интенсивностей каждого гена преобразуются в MA-значения по формулам (1) и (2):

$$M_g = \log_2\left(\frac{R_g - Rb_g}{G_g - Gb_g}\right), \tag{1}$$

$$A_g = \frac{1}{2}\log_2\left(\left(R_g - Rb_g\right)\cdot\left(G_g - Gb_g\right)\right), \tag{2}$$

где индекс g – номер гена, R и G – интенсивности спота в красном и зеленом каналах соответственно, Rb и Gb – фоновые интенсивности спота в красном и зеленом каналах соответственно (данные значения содержатся во входном файле).

Величину M называют уровнем экспрессии гена, A – средней интенсивностью.

Нормировка – это преобразование уровней экспрессии генов (M значений) с целью устранения систематических вариаций небиологической природы [4]. Нормировка снижает зашумлённость данных, вызванную неоднородностью поверхности микроматрицы, неравномерным распределением флуоресцентных меток по молекулам образцов, неравномерной концентрацией самих молекул образцов в зондах микроматрицы и другими экспериментальными факторами. Согласно [4], для корректировки эффектов, вызванных пространственной неоднородностью микроматрицы, наиболее надёжным является учет полного набора генов.

В эксперименте обычно используется несколько биочипов, так называемые технические репликанты, поэтому для каждого гена получается набор из M значений:

$$M_{g1}, M_{g2}, M_{g3}, \ldots, M_{gr}, \tag{3}$$

где g – номер гена, r – количество репликантов.

После нормировки необходимо отфильтровать гены с большим количеством пропусков в наборе M значений (см. формулу (3)). Если у гена пропущено небольшое число значений экспрессии (обычно менее 33%), то пропущенные значения можно аппроксимировать по методу ближайших k соседей [5].

Выделение статистически значимых генов, т.е. представляющих интерес для дальнейшего исследования, можно выполнить с помощью метода SAM (significance analysis of microarrays) [6]. SAM метод контролирует *FDR* (false discovery rate, частота ошибок первого рода):

$$FDR = \left\langle \frac{m}{N} \right\rangle, \tag{4}$$

где m – число генов, ошибочно отнесённых к значимым, N – число всех генов, отнесённых к значимым.

После выделения значимых генов проводят их кластерный анализ.

Основным результатом обработки данных является набор значимых генов, разделённых по кластерам, и значение FDR для данного набора значимых генов.

3. РЕЗУЛЬТАТЫ

Среда программирования R – это свободная и открытая среда статистического анализа. В R хорошо развиты векторно-матричная обработка данных и статистический аппарат. Для R написано множество динамически загружаемых библиотек-расширений, называемых R-пакетами (R packages).

Для загрузки и фильтрации данных использованы функции *readTargets*, *read.maimages* и *wtflags* R-пакета limma (http://bioconductor.org). Загруженные данные, около 1350 генов, снятые с трёх биочипов представляют собой 4050 RG значений (*R, Rb, G* и *Gb*). Из них отфильтровано 1554 значения (38.37%) как некачественные, для которых параметр качества *Flags<75* (этот параметр содержится во входных файлах).

Для нормировки экспериментальных данных использована функция *normalizeWithinArrays* (пакет limma) с параметрами по умолчанию. Эта функция преобразовала RG значения в MA значения (см. формулы (1) и (2)) и пронормировала полученные MA значения по методу print-tip LOWESS. Далее выполнена глобальная нормировка по среднему по M значениям каждой микроматрицы отдельно.

Затем были отфильтрованы 636 (47.11%) генов, имеющие хотя бы одно пропущенное значение. 714 генов оставлены для последующего анализа. Восстановление пропущенных значений экспрессии не проводилось, т.к. в эксперименте участвовало всего три биочипа, чего недостаточно для надёжного восстановления пропущенных значений.

Значимыми, т.е. представляющими интерес, являются дифференциально выраженные гены. С помощью метода SAM [6], реализованного в

R пакете siggenes (http://bioconductor.org), из 714 выделены 46 значимых генов с *FDR=8.25%.*

Над выделенными 46 генами проведен кластерный анализ иерархическим методом с помощью функции *hclust* из стандартного R пакета stats. Для анализа качества кластеризации применены кофенетические коэффициенты корреляции. Кофенетическое расстояние рассчитано с помощью функции *cophenetic* (R пакет stats). Расчёт кофенетических коэффициентов корреляции для различных метрик и методов связывания функции *hclust* показал, что наилучшим сочетанием является метрика максимального значения и метод средней связи с кофенетическим коэффициентом корреляции равным 0.913.

4. ВЫВОДЫ

В данной работе изучены биочипы ДНК и методы обработки данных об экспрессии генов. Методы реализованы в свободной и открытой среде статистического анализа R и исследованы на примере опубликованных экспериментальных данных [3].

В результате анализа из 1350 генов отфильтровано 714 генов, из которых выделено 46 значимых с *FDR=8.25%* (из 46 генов ожидается 35 дифференциально выраженных и 11 ошибочно причисленных к дифференциально выраженным).

Полученные данные могут быть использованы в дальнейшем процессе исследования генов. Конечной целью такого исследования являются изучение биологических функций выделенных генов, их взаимосвязей, процессов с участием этих генов.

Среда R является удобным инструментом для решения задач статистической обработки данных об экспрессии генов, полученных с микроматриц ДНК.

Литература

1. *Свешникова А.Н., Иванов П.С.* Экспрессия генов и микрочипы: проблемы количественного анализа // Рос. хим. ж. 2007. №51. С. 127-135.
2. *Hoheisel J.D.* Microarray technology: beyond transcript profiling and genotype analysis // Nat. Rev. Genet. 2006. 7 марта. №7. С. 200-210.
3. *Yatskou M., Novikov E., Vetter G., Muller A., Barillot E., Vallar L., Friederich E.* Advanced spot quality analysis in two-colour microarray experiments // BMC Res. Notes. 2008. 17 сент. №1. С. 80.
4. *Yang Y.H., Dudoit S., Luu P., Lin D.M., Peng V., Ngai J., Speed T.P*. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation // Nucleic Acids Res. 2002. 15 февр. №30. С.15.