Информационные процессы, Том 14, № 2, 2013, стр. 118–123.

Выбор функции плотности вероятности распределения экспрессии генов при обработке данных в методе RNA-Seq[1]

Г. А. Андрианов*, О. С. Кременецкая**

*Национальный исследовательский университет "Высшая школа экономики", Москва, Россия
**Центр теоретических проблем физико-химической фармакологии РАН, Москва, Россия

Поступила в редколлегию 19.06.2013

Аннотация—В настоящий момент для выравнивания гистограммы распределения ридов, полученных в результате обработки транскриптомов различных особей, предлагают использовать отрицательное биномиальное распределение. В данной работе проанализировано "физическое" обоснование возникновения уширения пуассоновского распределения и сделан вывод, что истинный вид распределения есть действительно сложное распределение Пуассона (частным случаем которого является отрицательное биномиальное распределение), однако представляет собой другой частный случай данного распределения — случай n-кратной (n — случайная величина с распределением Пуассона) свертки случайных величин с экспоненциальным распределением, а не логарифмически распределенных случайных величин. Показано, что распределение интенсивности экспрессии генов у группы лиц, вычисленное по опубликованным данных, лучше описывается сверткой с экспоненциальным ядром.

Ключевые слова: отрицательное биномиальное распределение, RNA-Seq, сложное распределение Пуассона, экспрессия генов.

1. ВВЕДЕНИЕ

При обработке данных в методе RNA-Seq на этапе анализа необходимо различать, произошли ли изменения в экспресси тех или иных генов. Для этого строят распределение ридов и выравнивают полученную гистограмму теоретической кривой. На основе знаний о характере кривой можно сделать вывод о том, изменилась или нет экспрессия генов в одной выборке по сравнению с другой и рассчитать достоверность результата.

Распределение ридов характеризует неоднородности между образцами, а также неоднородности, возникающей при измерении. В методе RNA-Seq различают [1] три уровня неоднородности между образцами:

1. Неоднородность между различными субъектами: предметом исследования является смесь образцов, полученных для выборки индивидуумов (экземпляров, особей или лиц), обладающих индивидуальными особенностями;
2. Неоднородность образцов РНК: возникает при получении РНК из клеток;
3. Неоднородность фрагментов РНК: амплификация происходит с определенным образом фрагментированной РНК, при секвенировании считываются не все фрагменты, загруженные в ячейки прибора для секвенирования.

Характер данных [2, 3], проанализированных в настоящей работе, предполагает наличие неоднородностей между образцами на всех трех уровнях. Однако, в дальнейшем в данной

[1] Работа поддержана Министерством образования и науки РФ в рамках ФЦП «Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса Российской Федерации на 2007–2013 годы», ГК No. 14.512.11.0042.

работе будет показано, что при исследовании транскриптомов смесей образцов, взятых от различных индивидуумов, решающее значение имеет неоднородность первого типа, заключающаяся в различиях между рассматриваемыми индивидуумами.

Пусть некоторое свойство, являющееся общим для группы индивидуумов, связано с экспрессией генов. Тогда различия в обладании данным свойством будет выражаться в том, что для разных групп индивидуумов будут наблюдаться различные наборы данных об интенсивности экспрессии гена. Данные в пределах исследуемого набора будут иметь распределение, похожее по форме на распределение Пуассона. Анализ литературы показывает, что данное распределение, вообще говоря, имеет дисперсию, большую, чем математическое ожидание, и, таким образом, его функция плотности вероятности (ФПВ) "шире", чем ФПВ распределения Пуассона. Если между группами существуют различия, то мы наблюдаем различимые между собой ФПВ для каждой из групп. Очевидно, что задача выравнивания гистограммы распределения актуальна для установления такого различия.

Для описания кривой ФПВ распределения интенсивности экспрессии разные авторы предлагают различные аналитические выражения [4–6]. Исследователи пришли к выводу, что лучше всего подходит для решения задачи параметризованное отрицательное биномиальное распределение [7]. Функция плотности вероятности данного распределения имеет следующий вид:

$$f(y;\mu,\phi) = \frac{\Gamma(y+\phi^{-1})}{\Gamma(\phi^{-1})\Gamma(y+1)} \left(\frac{1}{1+\mu\phi}\right)^{\phi^{-1}} \left(\frac{\mu}{\phi^{-1}+\mu}\right)^{y}, \tag{1}$$

где μ представляет собой математическое ожидание, а $\mu+\phi\mu^2$ – дисперсию. При $\phi \to 0$ данное распределение сводится к своему частному случаю – распределению Пуассона. В работе [8] указано, что пуассоновская модель не учитывает биологическое разнообразие материала, и должна использоваться модель распределения, для которого дисперсия больше, чем математическое ожидание, например, модель отрицательного биномиального распределения (1).

2. ПРЕДЛАГАЕМАЯ МОДЕЛЬ

Соглашаясь с авторами [8], мы также предлагаем рассматривать распределение, имеющее дисперсию, вообще говоря, превышающую среднее значение. Вклад в данную дисперсию, действительно, может быть разделен на две части – техническую и биологическую. С технической частью все просто, она определяет пуассоновское распределение некого свойства (в случае анализа данных RNA-Seq – это число ридов того или иного гена, определяющее интенсивность экспрессии гена), обусловленное тем, что в методе приходится работать с измерениями величины – выборками, по которым составляется представление о значении данной величины. Но само свойство на наборе данных также оказывается распределенным. Распределение свойства между индивиддумами соответствует биологическому вкладу в результирующую дисперсию.

Распределения рассматриваимаемого типа хорошо известны. В [9] они назывются сложными распределениями Пуассона:

$$f(y) = e^{-\lambda} \sum \frac{\lambda^n}{n!} g(y)^{n*}, \tag{2}$$

где $g(y)^{n*}$ – n-кратная свертка ФПВ $g(y)$ – распределения биологического свойства по индивидуумам, которое как раз и обеспечивает биологический вклад в дисперсию. Для данного распределения преобразование Лапласа-Стилтьеса (ПЛС) имеет вид:

$$F(s) = e^{\lambda(G(s)-1)}, \tag{3}$$

где $G(s)$ – ПЛС функции распределения биологического свойства по индивидуумам.

В.Феллер показал, что столь широко используемое при обработке результатов в методе RNA-Seq отрицательное биномиальное распределение является сложным распределением Пуассона, причем в данном случае распределение биологического свойства $g(y)$ представляет собой логарифмическое распределение с ПЛС $G(s) = \frac{1}{\lambda} \ln \frac{1}{1-qs}$ [9].

Мы предположили, что распределение интенсивности экспрессии гена по различным индивидуумам в группе должно иметь более простой характер. Например, $g(y)$ может представлять собой хорошо известное экспоненциальное распределение с параметром μ. Здесь уместно напомнить, что расстояние между событиями в пуассоновском, абсолютно случайном, потоке событий также распределено экспоненциально.

ПЛС для экспоненциального распределения имеет вид:

$$G(s) = \frac{\mu}{\mu + s}. \tag{4}$$

Подставив (4) в (3) и произведя некоторые вычисления, а затем, применив к (3) обратное преобразование Лапласа-Стилтьеса, получим:

$$f(y) = \delta(y)e^{-\lambda} + e^{-(\lambda+\mu y)} \sqrt{\frac{\lambda\mu}{y}} I_1(2\sqrt{\lambda\mu y}), \tag{5}$$

где $\delta(y)$ – дельта-функция Дирака, $I_1(x)$ – модифицированная функция Бесселя первого рода первого порядка [10].

Зависимость (5) также, как и отрицательное биномиальное распределение (1) представляет собой сложное распределение Пуассона, только данное распределение оказывается сформированным путем n-кратной свертки (n имеет распределение Пуассона) не лагарифмически распределенных вкладов, а вкладов, имеющих более простое и физически обоснованное экспоненциальное распределение.

3. АНАЛИЗ ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ

В работе были построены гистограммы распределения интенсивности экспрессии для 40 различных генов. Анализировали данные о 129 различных индивидуумах: 60 – в одном наборе данных, люди с белым цветом кожи [2], и 69 – в другом наборе данных, люди с черным цветом кожи [3]. Иногда наблюдалось явное бимодальное распределение интенсивности экспрессии гена. В некоторых случаях, например, для гена SPPL2B и гена BZRAP1, после разделения общей выборки в 129 человек на две группы 60 и 69 человек с разным цветом кожи, бимодальное распределение разделялось на два колоколообразных пика с различными средними значениями. Это свидетельствует о том, что некоторые биологические свойства выражены в разной степени у людей разных рас.

В ряде случаев бимодальность сохранялась после разделения по цвету кожи, так как она была обусловлена иными причинами. В большинстве случаев мы имели и для неразделенных выборок одну колоколообразную гисторамму распределения. Данная гистограмма удовлетворительно выравнивалась сложным распределением Пуассона с ФПВ, выраженной формулой (5). Проведенный анализ продемонстрировал, что различия между рассматриваемыми индивидуумами являются основной причиной возникновения неоднородности выборки в ее биологическом, не техническом аспекте.

На рис. 1 изображена типичная гистограмма распределения интенсивности экспрессии гена (на примере гена HCCS), построенная по данным, опубликованным в [2,3]. При выравнивании данной гистограммы последние два столбца данных, содержащие лишь по одному отсчету, объединяли. Распределение, выраженное данной гистограммой, проверяли на соответствие

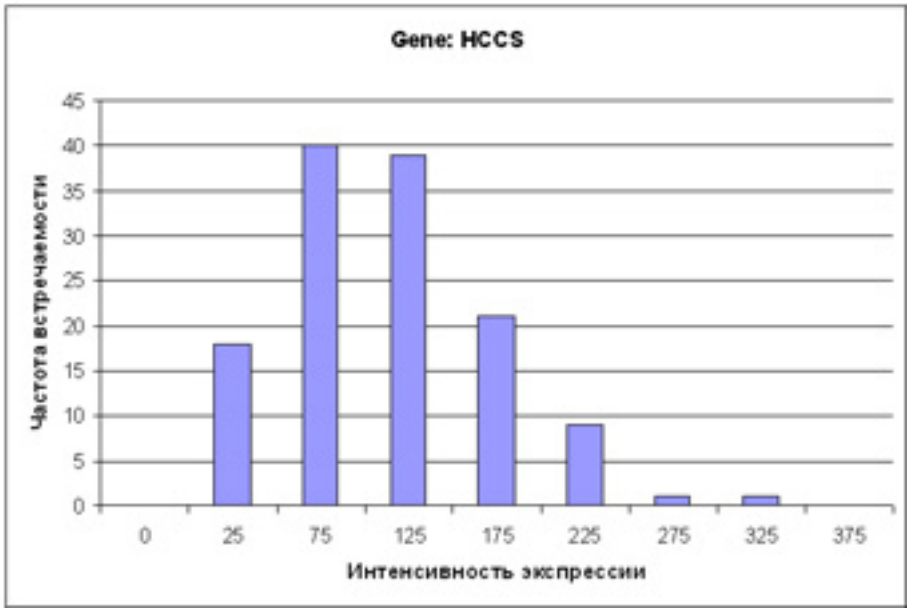

Рис. 1. Гистограмма распределения интенсивности экспрессии гена HCCS.

сложному распределению Пуассона с экспоненциальными вкладами (5) и параметризованному отрицательному биномиальному распределению (1).

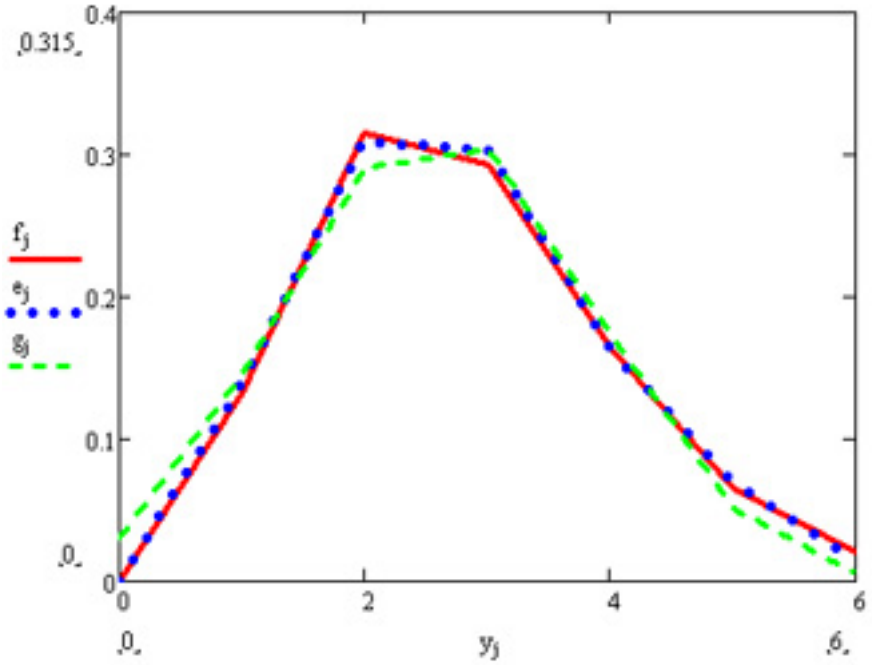

Рис. 2. Выравнивание гистограммы распределения интенсивности экспрессии гена HCCS.

Выяснилось, что погрешность, определяемая средним квадратом отклонения теоретической зависимости от соответствующего значения столбца гистограммы, в случае предлагаемого в данной работе распределения (5) составляет $3 \cdot 10^{-5}$, а в случае отрицательного биномиально-

го распределения $3 \cdot 10^{-4}$, то есть, на порядок больше. Рис.2 иллюстрирует, насколько лучше выравниваются экспериментальные данные (показаны точками, e_j) предлагаемой нами зависимостью (сплошная линия, f_j), по сравнению с параметризованным отрицательным биномиальным распределением (пунктирная линия, g_j).

Полученная теоретическая кривая сложного распределения Пуассона с экспоненциальными вкладами (СРПЭ) обладает несколько более тяжелым хвостом распределения по сравнению с параметризованным отрицательным биномиальным распределением (ПОБР). Математическое ожидание в данном примере для предлагаемого в настоящей работе распределения составляет 140.1, для отрицательного биномиального оно всего 130.9, прямой расчет по данным дает 139.4, то есть СРПЭ в сравнении с ПОБР лучше отражает математическое ожидание случайной величины.

В данном типичном случае распределения интенсивности экспрессии (ген HCCS) наблюдались следующие параметры двухпараметрической зависимости (5): $\lambda = 9.97$, $\mu = 0.0712$.

Приведем еще один пример. Для гена SNX11 параметр $\lambda = 10.0$, а параметр $\mu = 0.0129$. Среднее значение здесь совсем другое: 775 (вычисленное как $\frac{\lambda}{\mu}$) или 774.3 (напрямую по данным). Можно предположить, что в распределении Пуассона среднее значение равно 10 не случайно, и это как-то связано с техническими особенностями метода. Тем не менее, необходимо заметить, что $\lambda = 10$ в точности выполняется лишь для 15% исследованных генов. В большинстве случаев этот параметр лежит в интервале от 2.7 до 15.1.

4. ЗАКЛЮЧЕНИЕ

ФПВ распределения интенсивности экспрессии определяется зависимостью (5) и представляет собой СРПЭ – сложное распределение Пуассона с экспоненциальными вкладами. Параметризованное отрицательное биномиальное распределение (ПОБР), широко используемое в настоящее время в программном обеспечении для анализа результатов RNA-Seq, близко по форме к предложенному в данной работе СРПЭ. Таким образом, ошибка при проведении автоматизированных расчетов с использованием ПОБР вместо СРПЭ будет невелика. Однако необходимо заметить, что при использовании ПОБР, обладающего менее тяжелым хвостом, математическое ожидание случайной величины будет несколько (на 7%) занижено. Великолепное выравнивание экспериментальной гистограммы при помощи теоретической зависимости СРПЭ косвенно подтверждает гипотезу об экспоненциальном распределении биологического свойства, выражаемого интенсивностью экспрессии гена. Для разработки программного обеспечения, позволяющего делать вывод об изменениях в экспрессии генов, представляется целесообразным использовать СРПЭ – сложное распределение Пуассона с экспоненциальными вкладами (5). Данное распределение более точно, чем параметризованное отрицательное биномиальное распределение, отражает физический смысл наблюдаемой неоднородности.

5. БЛАГОДАРНОСТИ

Авторы благодарят д. ф.-м. н. Всеволода Юрьевича Макеева (Институт общей генетики им. Н. И. Вавилова РАН) за внимательное прочтение данной работы и ценные замечания.

СПИСОК ЛИТЕРАТУРЫ

1. Auer P. L. and Doerge R. W. Statistical design and analysis of RNA sequencing data. *Genetics*, 2010, 185: 405-416.

2. Montgomery S. B., Sammeth M., Gutierrez-Arcelus M., Lach R. P., Ingle C., Nisbett J., Guigo R., Dermitzakis E. T. Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature*, 2010, 464(7289): 773-7.

3. Pickrell J. K., Marioni J. C., Pai A. A., Degner J. F., Engelhardt B. E., Nkadori E., Veyrieras J. B., Stephens M., Gilad Y., Pritchard J. K. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, 2010, 464(7289): 768-72.

4. Marioni J. C., Mason C. E., Mane S. M., Stephens M., and Gilad Y., RNA-Seq: an assessment of technical reproducibility and comparison with gene expression arrays, *Genome Research*, 2008, vol. 18, no. 9, pp. 1509вАУ1517.

5. Costa V., Angelini C., Feis I. D., and Ciccodicolo A. Uncovering the Complexity of Transcriptomes with RNA-Seq. *Journal of Biomedicine and Biotechnology*, 2010, Article ID 853916, 19 pp.

6. Fang Zh., Martin J., and Wang Zh. Statistical methods for identifying differentially expressed genes in RNA-Seq experiments. *Cell & Bioscience*, 2012, 2:26.

7. Robinson M. D. and Smyth G. K. Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics*, 2008, 9, 2, pp. 321вАУ332.

8. McCarthy D. J., Chen Y., and Smyth G. K.Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research*, 2012, Vol. 40, No. 10, pp. 4288вАУ4297.

9. Феллер В. *Введение в теорию вероятностей и ее приложения. Том 1.* М.: Мир, 1964.

10. Ватсон Г. *Теория бесселевых функций. Т. 1, 2.* М.: ИЛ, 1949.

On the Probability Density Function Choice for the Gene Expression Distribution in the RNA-Seq Data Analysis

G. A. Andrianov* and O. S. Kremenetskaya**

***National Research University Higher School of Economics, Moscow, Russia**
****Center for Theoretical Problems of Physicochemical Pharmacology RAS, Moscow, Russia**

At the moment, for the equalization of reads histogram, which derived from the treatment of the transcriptome of different individuals, it is suggested to use a negative binomial distribution. In this paper we analyze the "physical" basis of a broadening of Poisson distribution, and conclude that the true form of the distribution is really compound Poisson distribution (a special case of which is the negative binomial distribution), but the true choice is another special case of this distribution, i.e. n-times convolution (n is a random variable with Poisson distribution) of random variables with the exponential (not logarithmical) distribution. It is shown that a distribution of gene expression intensity in a group of individuals calculated from the published data is described better by the convolution of exponential functions.

KEYWORDS: negative binomial distribution, RNA-Seq, compound Poisson distribution, gene expression.