Оценка произвольно малых P-значений в тесте представленности функциональных наборов генов

Сухов В.Д., аспирант, факультет информационных технологий и программирования, университет ИТМО, vdsukhov@ya.ru

Короткевич Г.В., аспирант, факультет информационных технологий и программирования, университет ИТМО
Сергушичев А.А., научный сотрудник, факультет информационных технологий и программирования, университет ИТМО

Аннотация

В настоящее время тест представленности функциональных наборов генов является широко распространенным подходом при анализе данных экспрессии генов. В работе проведено исследование ошибки приближенного алгоритма при анализе представленности функциональных наборов генов. Получена теоретическая оценка для ошибки получаемых P-значений. Проведено сравнение работы приближенного и точного алгоритмов на целочисленных рангах и определено оптимальное число шагов для эффективной работы приближенного алгоритма.

Введение

При проведении теста представленности функциональных наборов генов основной задачей является нахождение из априори определенной коллекции функциональных наборов генов тех, что имеют неслучайное поведение в рассматриваемом эксперименте. При решении данной задачи часто используют простое сэмплирование. Недостатком такого подхода является то, что размер сгенерированной выборки задаёт ограничения на вычисляемые P-значения [1]. В работе [2] были предложены два алгоритма, лишенные вышеуказанного недостатка. Так, первый – точный – работает на целочисленных рангах и основан на применении динамического программирования, второй – приближенный – основан на применении многоуровневого подхода Монте-Карло.

Важным моментом при работе с приближенным алгоритмом является понимание того, какая ошибка образуется при вычислении P-значений. В работе предложен метод для оценки ошибки, а также определено необходимое число шагов для эффективной работы алгоритма.

Постановка задачи

При проведении теста представленности функциональных наборов генов входными данными являются вектор рангов генов $S = (S_1, S_2, \ldots, S_N)$ размера N и коллекция функциональных наборов генов P. Для сравнения и поиска значимых наборов генов используется статистика представленности [3]. Она определяется следующим образом:

$$s(p) = \max_i |ES_i|,$$

где p – произвольный набор из P, а значение ES_i в свою очередь определено как

$$ES_i = \begin{cases} 0, & \text{если } i = 0, \\ ES_{i-1} + \dfrac{|S_i|}{\sum\limits_{i \in p} |S_i|}, & \text{если } 1 \leqslant i \leqslant N \text{ и } i \in p, \\ ES_{i-1} - \dfrac{1}{N-k}, & \text{если } 1 \leqslant i \leqslant N \text{ и } i \notin p. \end{cases}$$

Теперь для определения, имеет ли некоторый набор p неслучайное поведение, необходимо вычислить вероятность $\mathrm{P}\left(s(X) > s(p)\right)$, где X - случайный набор генов такого же размера, что и набор p.

Приближенный алгоритм

Приближенный алгоритм основан на многоуровневом подходе Монте-Карло [4]. Для вычисления вероятности $\mathrm{P}(s(X) > s(p))$ вводится расщепление $-1 = l_0 < l_1 < \ldots < l_t = s(p)$ по значениям статистики представленности. Затем вычисляются следующие вероятности:

$$\begin{gathered} \mathrm{P}\left(s\left(X\right) > l_1 \middle| \, s\left(X\right) > l_0\right) = \alpha_1 \\ \mathrm{P}\left(s\left(X\right) > l_2 \,\middle|\, s\left(X\right) > l_1\right) = \alpha_2 \\ \ldots \\ \mathrm{P}\left(s\left(X\right) > l_t \,\middle|\, s\left(X\right) > l_{t-1}\right) = \alpha_t. \end{gathered}$$

После этого вероятность $\mathrm{P}(s(X) > s(p))$ оценивается как $\prod\limits_{i=1}^{t} \alpha_i$. Для вычисления α_i применяется условное сэмплирование. Так, при помощи алгоритма Метрополиса, для каждого уровня i равномерно генерируются выборки нечётного размера Z из условного распределения

$\mathrm{P}\left(\cdot \mid s\left(X\right) > l_{i-1}\right)$. В качестве $l_1, l_2, \ldots, l_{t-1}$ берётся медиана выборки, поэтому $\alpha_1 = \alpha_2 = \ldots = \alpha_{t-1} = 1/2$. Значение α_t определяется как отношение числа наборов генов в выборке из распределения $\mathrm{P}\left(\cdot \mid s\left(X\right) > l_{t-1}\right)$ со значением статистики представленности больше l_t к размеру выборки.

Оценка ошибки

Известно, что медиана m выборки размера Z из равномерного распределения является случайной величиной из бета-распределения с параметрами $\alpha = (Z+1)/2,\ \beta = (Z+1)/2$. Это позволяет воспользоваться следующими свойствами бета-распределения:

$$\mathrm{E}[\log(m)] = \psi\left(\frac{Z+1}{2}\right) - \psi\left(Z+1\right), \quad \psi - \text{дигамма-функция},$$

$$\mathrm{D}[\log(m)] = \psi_1\left(\frac{Z+1}{2}\right) - \psi_1\left(Z+1\right), \quad \psi_1 - \text{тригамма-функция}.$$

Тем самым можно перейти к вычислению логарифма вероятности

$$\log \mathrm{P}(s(X) > s(p)) = (t-1)\left(\psi\left(\frac{Z+1}{2}\right) - \psi\left(Z+1\right)\right) + \log(\alpha_t),$$

и получить теоретическую оценку сверху для ошибки в следующем виде

$$\mathrm{error} = \sqrt{t \cdot \mathrm{D}[\log(m)]} = \sqrt{t \cdot \left(\psi_1\left(\frac{Z+1}{2}\right) - \psi_1\left(Z+1\right)\right)}.$$

Число шагов приближенного алгоритма

Сначала нужно определить что понимается под шагом алгоритма. Для этого кратко опишем алгоритм Метрополиса для генерирования выборки из распределения $\mathrm{P}\left(\cdot \mid s\left(X\right) > l_i\right)$.

1. Устанавливаем счётчик шагов равным нулю
2. Начинаем с выборки $X_1, X_2, \ldots X_Z$ такой, что $s(X_j) > l_i$, $j \in \{1, 2, \ldots, Z\}$
3. Поочередно берем каждый набор генов X_j, выбираем случайный ген, входящий в набор, и заменяем случайным геном не из набора. Если замена сохраняет условие $s(X_j) > l_i,\ j \in \{1, 2, \ldots, Z\}$, то увеличиваем счётчик шагов на единицу, в противном случае отменяем изменения.

Таким образом, под шагом понимается изменение набора генов, которое сохраняет условие для статистики представленности.

Определение необходимого числа шагов произведено при помощи сравнения результатов работы точного и приближенного алгоритмов при запусках на целочисленных рангах. Запуски проводились для различных размеров наборов генов и значений статистики представленности. Число шагов принималось равным $\text{steps} = a \cdot k \cdot Z$, здесь Z – размер выборки, k – размер набора генов, a – параметр, который варьировался. На практике достаточно положить $a = 1$. Так, на рисунке 1 приведены результаты запусков приближенного алгоритма и сравнение полученных результатов с логарифмом точного P-значения.

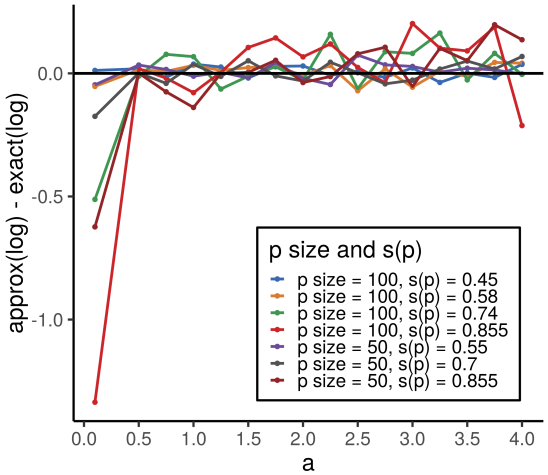

Рис. 1: Сравнение приближенного и точного алгоритмов. По оси OY откладывается разница логарифмов приближенного и точного значений, по оси OX отложен параметр a. Здесь каждой точке соответствует усреднение по 100 запускам приближенного алгоритма.

Заключение

1. Определена верхняя оценка ошибки для приближенного алгоритма.
2. Определено число шагов для эффективной работы приближенного алгоритма.

3. Результаты работы успешно внедрены в пакет FGSEA, написанный на языке программирования R.

Литература

[1] Сергушичев А.А. Алгоритм для быстрого анализа перепредставленности генов / Список-2016. С. 238-244.

[2] Короткевич Г.В. Алгоритмы для эффективного анализа представленности функциональных наборов генов : выпускная квалификационная работа. Университет ИТМО. Санкт-Петербург. 2018.

[3] Subramanian A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles / Proceedings of the National Academy of Sciences of the United States of America. 2005. V.102. №43. P.15545-15550

[4] Botev Z.I., Kroese D.P., An efficient algorithm for rare-event probability estimation, combinatorial optimization, and counting / Methodology and Computing in Applied Probability. 2008. V.10. P. 471-505