Индекс УДК: 004.021

Адаптивный многоуровневый подход Монте-Карло для задачи функционального анализа представленности генов

Сухов В.Д.[1], Короткевич Г.В.[1], Сергушичев А.А.[1]

[1]- Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики (Университет ИТМО), Санкт-Петербург

Научный руководитель: Сергушичев А.А., Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики (Университет ИТМО), Санкт-Петербург

Введение

В настоящее время актуальным и широко распространенным подходом для исследования данных экспрессии генов является метод анализа представленности функциональных наборов генов. Среди существующих, распространенных методов по решению данной задачи можно выделить те, что базируются на простом сэмплировании. К их недостаткам относится ограниченность в вычислении достаточно малых P-значений.

Нами предложен метод, который лишен вышеуказанного недостатка. В его основе лежит адаптивный многоуровневый подход Монте-Карло. Этот подход применяется в случаях, когда необходимо оценивать вероятность редких событий.

Для задачи определения статистической значимости наборов генов входными данными являются: вектор рангов генов $S = (S_1, S_2, ..., S_N)$ размера N и список функциональных наборов генов P. При проведении сравнения наборов генов применяется статистика представленности (GSEA - статистика). Для произвольного набора $p = (p_1, p_2, ..., p_k) \in P$ размера k, статистика представленности определяется следующим образом:

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

где значения вектора ES в свою очередь определяются так:

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

Описание алгоритма

Пусть p - набор генов размера k и $s(p)$ - соответствующее значение статистики представленности. Нас будет интересовать вероятность $P\left(s(q) > s(p)\right)$ того, что случайно выбранный набор генов q размера k имеет значение статистики представленности больше значения $s(p)$. Будем считать, что $s(p) > 0$, так как выполняется следующее:

$$s'(p') = -s(p),$$

где $s'(p')$ - статистика представленности для следующих входных данных: $S' = (S_N, S_{N-1}, ..., S_1)$ и $p' = (N - p_1 + 1, N - P_2 + 1, ..., N - p_k + 1)$. Тогда:

$$P\left(s(q) > s(p)\right) = P\left(s^+(q) > s^+(p)\right) \cdot P\left(s(q) > 0 \;\middle|\; s^+(q) > s^+(p)\right),$$

где $s^+(p) = ES_{i^+}$ и $i^+ = \operatorname*{argmax}_i ES_i$. Для вычисления $P\left(s^+(q) > s^+(p)\right)$ предложен следующий алгоритм:

1. Генерирование сэмплов в соответствии с размером входного набора генов, в количестве Z.

2. Вычисление значений статистики представленности для каждого сэмпла.
3. Определение медианного значения статистик представленности, полученных в пункте 2.
4. Сэмплы имеющие значение статистики представленности меньше медианного отбрасываются и заменяются новыми таким образом, чтобы новое значение статистики было больше медианы из предыдущего пункта.

Вышеуказанные процедуры повторяются до тех пор, пока на некотором шаге j не выполнится следующее условие:

$$m_j \leqslant s(p) < m_{j+1},$$

где m_j - медианное значение статистики представленности для Z сэмплов на j - ом шаге. Тогда искомая вероятность $P\left(s^+(q) > s^+(p)\right)$ лежит в диапазоне $\left(2^{-j-1}, 2^{-j}\right)$.

Описанный алгоритм успешно внедрен в пакет FGSEA, написанном на языке программирования R. Сравнение нового подхода со старым продемонстрировало, что результаты работы алгоритмов хорошо согласуются. Однако, нижняя граница P-значений, получаемых при помощи классического алгоритма FGSEA ограничена параметром $nperm$. Увеличение точности классического алгоритма на порядок требует на порядок большего количества времени работы. Новый же подход лишен этого недостатка и позволяет вычислять сколь угодно малые P-значения.

Заключение

1. Приведен новый алгоритм для проведения функционального анализа представленности генов.
2. Новый подход позволяет вычислять произвольные P-значения.
3. Описанный алгоритм продемонстрировал свою эффективность. Так для вычисления в четыре потока P-значений в интервале от $\left(10^{-6}, 1\right)$ алгоритму FGSEA потребовалось около 4 минут, в то время как новому алгоритму потребовалось около 30 секунд (при Z=500).