Индекс УДК: 004.021

Определение статистически значимых наборов генов методом FGSEA при множественной проверке гипотез

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

Современный уровень развития технологий позволяет проводить сложнейшие биологические эксперименты по анализу работы генов. Актуальной задачей в данной области является определение наборов генов с общим свойством, которые демонстрируют неслучайное поведение в эксперименте. Один из основных подходов по решению данной задачи опирается на применение методов математической статистики. Так, для определения уровня значимости набора генов применяется статистика представленности и выборка, составленная из случайных наборов генов. Это позволяет оценить вероятность (P-значение) того, что случайный набор генов имеет значение статистики представленности больше или равное значению статистики представленности исследуемого набора генов.

В действительности, при анализе экспериментальных данных требуется проводить одновременное тестирование множества гипотез, что приводит к необходимости введения поправки на множественное сравнение. Для нахождения значимых наборов генов после соответствующей поправки требуется уметь вычислять близкие к нулю P-значения.

Цель работы

1. Разработка метода, который позволяет вычислять сколь угодно малые P-значения
2. Анализ и сравнение результатов работы метода FGSEA с оригинальным методом GSEA.

Результаты

В ходе работы была обновлена реализация метода FGSEA, что сделало возможным вычисление сколь угодно малых P-значений. В основу обновленного метода легло применение адаптивного многоуровневого подхода Монте-Карло и цепи Маркова.

Для сравнения методов использовались данные экспрессии генов из базы данных Gene Expression Omnibus. Получение "поправленных" P-значений в методе FGSEA осуществлялось за счёт процедуры Бенджамини-Хохберга, а для P-значений в методе GSEA использовалась оригинальная процедура. Важно отметить, что процедура метода GSEA чувствительна к количеству исходно значимых наборов генов. Это делает результаты работы метода GSEA более консервативными в тех случаях, когда исходно имеется небольшое количество значимых наборов генов. Вследствие этого, в большинстве случаев количество значимых наборов генов, полученных методом GSEA, в среднем в два раза меньше, чем в методе FGSEA. Была установлена зависимость в результатах работы метода GSEA от размера тестируемого набора

генов. Всё это говорит о том, что применение метода FGSEA является перспективным для генерирования новых гипотез. Данный факт выражается в более высокой мощности применяемого статистического критерия в методе FGSEA и позволяет рассмотреть большее число наборов генов-кандидатов, которые потенциально могут отвечать за наблюдаемые различия в биологическом эксперименте.