Создание модуля для эффективной обработки генетических данных с использованием теста хи-квадрат

Выполнил: Соковикова С., 344 гр.

Научный руководитель: к.ф.-м. н., доц. Малов С.В.

Введение

- Тест хи-квадрат очень широко применяется в биостатистике
- В генетике
- Язык R
- R низкопроизводителен

Мотивация

- Тесты проводятся долго
- Очевидно, можно в разы быстрее

Цель работы

- Создать модуль на языке C, который будет реализовывать тест хи-квадрат быстрее, чем стандартные функции языка R. Модуль должен вызываться из R-среды.

Постановка задачи

- Имеется файл gds
- Из файла читается вектор и матрица
- Длина вектора = числу строк матрицы
- Провести хи-квадрат тест, сопоставляя вектор с КАЖДЫМ столбцом матрицы
- Вычислить значения pValue и записать их в новый gds-файл

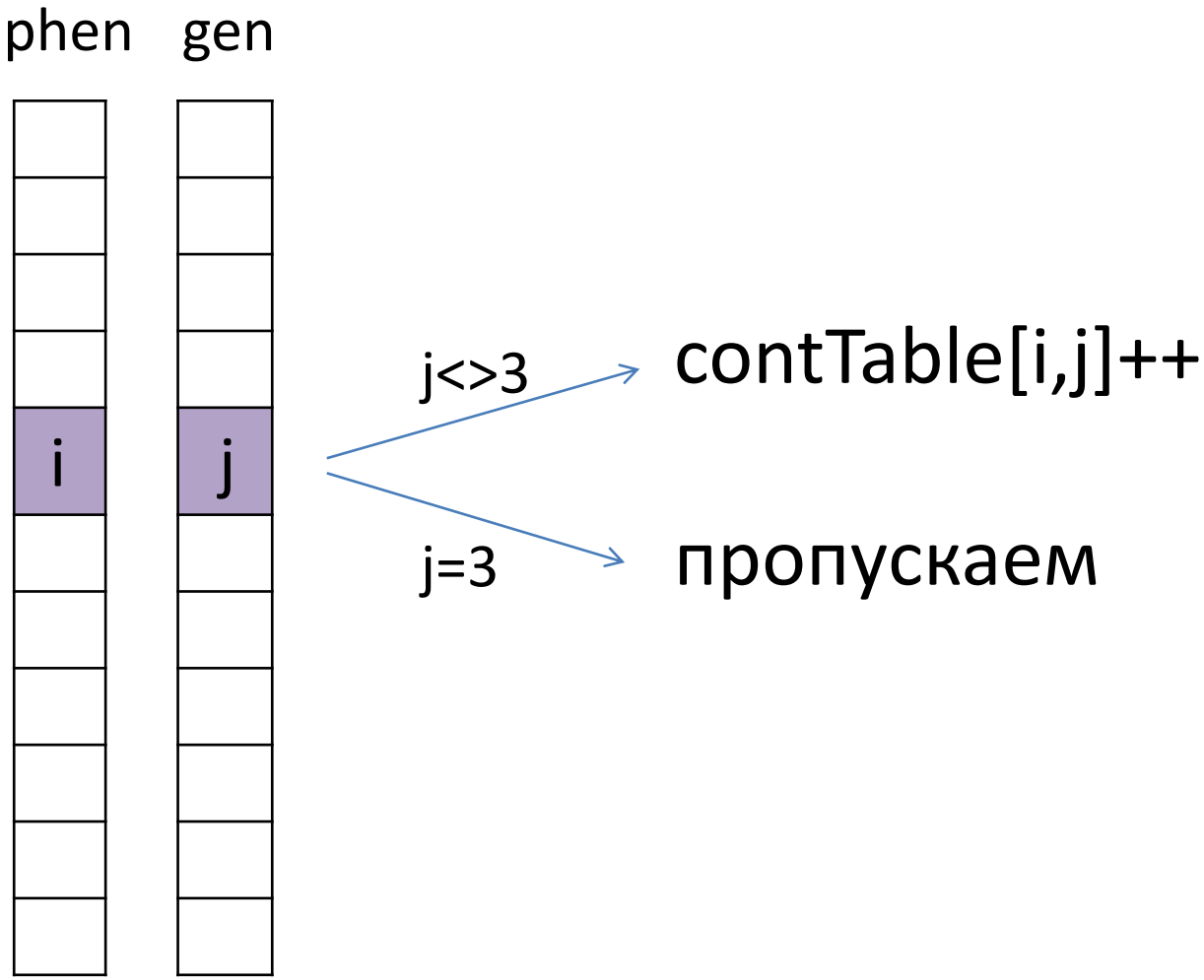
Алгоритм решения (1)

- Читается gds-файл средствами R
- .C() – вызов модуля с передачей ему данных

Алгоритм решения (2)

- Вектор – из чисел 1:k (k<=20)
- Матрица – из чисел 0:3.
- Длина вектора = числу строк в матрице
- Дальнейшие шаги – для каждого столбца матрицы
- Составить таблицу сопряженности вектора со столбцом

Алгоритм решения (3)



Алгоритм решения (4)

- Получена contTable[1:k,0:2]
- Для contTable считаются суммы:
- по строкам sumrow
- по столбцам sumcol
- сумма всех элементов S
- Вычисляется $$\chi^2 = \sum_{i=1}^{k}\sum_{j=0}^{2} \frac{(contTable[i,j] - E[i,j])^2}{E[i,j]}$$

где $$E[i,j] = \frac{sumrow[i] * sumcol[j]}{S}$$

Алгоритм решения (5)

- Значения χ^2 передаются в R
- Средствами R вычисляется

  $pValue(\chi^2)=1-F_{2(k-1)}(\chi^2)$
- Результат записывается в новый gds-файл

Сравнение производительности

Средство	Время выполнения, с
Стандартные функции R	1944.494
Моя реализация	6.277

Результат

- Существенно сокращено время работы