Домашнее задание №3

Д.А. Першин

14 ноября 2014 г.

1 Словесное описание алгоритма

При решении данной задачи будем использовать алгоритм быстрого преобразования Фурье (временная слжность - $O(n \log n)$, память - $O(n \log n)$).
Входные массивы a и b ($len(a) = n, len(b) = m, n \geq m$) разложим на 4 массива каждый следующим образом: a_A будет содержать единицы там, где в исходном массиве a находилось значение A, остальные нули; a_C будет содержать единицы там, где в исходном массиве a находилось значение C, остальные нули; аналогично для a_G и a_T, а также для массива b. Дополним каждый из 8 массиво до одинаковой длины следующим образом: к массывам a_A, a_C, a_G и a_G справа допишим m нулей, к массивам b_A, b_C, b_G и b_G - n нулей (для верного расчета циклической корреляции). В итоге получаем 8 массивов $m + n$ длины каждый. Затем дополним длины массивов до ближайшей степени двойки (для выполнения быстрого преобразоваия Фурье).

Далее для каждой пары векторов, полученных из a и b с одинаковым индексом ищем циклическую корреляцию. По теореме о циклической корреляции (спектр циклической корреляции есть произведение спектров сигналов) для функций f, f' и f'', где

$$f = \frac{1}{N} \sum_{m=0}^{n-1} f'_m f''_{m-n}, n = 0, 1, ..., N-1$$

$$D_n = D'^*_n D''_n, n = 0, 1, ..., N-1$$

Вычисление над функциями производится за $O(n^2)$, но для ДПФ последовательностей данных функций D', D'' циклическая корреляция вычисляется за $O(n)$, а алгоритм быстрого преобразования Фурье позволяет находить ДПФ последовательнсти за $O(n \log n)$ (обратное пробразование также за $O(n \log n)$).

Далее поэлементно складываем массивы для найденных циклических корреляций. В итоге получаем массив c, каждый элемант c_i которого показывает кол-во совпадающих элементов массива b и массива a начиная с индекса i. Поиск максимального элемента дает искомый индекс сдвига в массиве a при наложении на него массива b (также следует проверить не выходит ли массив b за границу массива a, так как изначально мы

увеличили его длину на m и дополнили до ближайшей степени 2). Если таких индексов несколько, то выбираем минимальный.

Алгоритм:

1. 2 входных массива разложем на 8 битовых массива, как описано выше.
2. Дополним каждый из массивов до одинаковой длины $l_a = l_b = m + n$, а затем до ближайшей степени двойки, остаток заполнив 0.
3. Найдем циклическую корреляцию для каждой пары массивов a и b с одинаковым индексом A, C, G или T используя алгоритм БПФ, а затем ОБПФ, назовем их c_A, c_C, c_G, c_T.
4. Сложим получившиеся массивы поэлементно $C[i] = c_A[i] + c_C[i] + c_G[i] + c_T[i]$.
5. Найдем максимум в массиве C, где $C[j] = \max\limits_{i \le n-m} C[i]$, j выберем минимальным из всех, удовлетворяющих данному условию.
6. j - искомый индекс.

2 Доказательство корректнсти

Предположим, что найденный индекс j не является индексом для наложения с максимальным совпадением, но в таком случае существует другой индекс j', такой что при наложении b на $a[j']$ получается большее количество совпадений, чем для j, но это противоречит теореме о циклической корреляции. Таким образом индекс j является индексом, таким что при наложении последовательности b на $a[j]$ получается максимальное количество поэлементных совпадений.

3 Асимптотические оценки

В результате получаем сложность по памяти $O((n+m)\log(n+m))$, так как мы используем 2 массива длины m и n, 8 массивов длиной $n + m$, рекурсивный алгоритм БПФ требует $O((n + m)\log(n + m))$ дополнительной памяти. Сложность по времени равна $O((n + m)\log(n + m))$, так как мы используем БПФ и ОБПФ - $O((n + m)\log(n + m))$, поиск циклической корреляции для ДПФ последовательностей за $O((n + m))$, поиск максимального элемента в результирующем массиве за $O(n - m)$.