0.1. *Шехова А.А., Карепова Е.Д., Шанько Ю.В.* Решение задачи локализации гомологичных участков при сравнении символьных последовательностей

В работе [1] нами был предложен новый метод сравнения символьных последовательностей $\mathbb{P}$ и $\mathbb{Q}$ разной длины n_P и n_Q, соответственно, составленных из символов конечного алфавита $\aleph$.

Метод рассматривает все возможные расположения последовательностей друг относительно друга (для краткости далее — наложения последовательностей $\mathbb{P}$ и $\mathbb{Q}$) и подсчитывает количество совпадающих пар символов, расположенных друг под другом в каждом наложении. Метод основан на вычислении свёртки индикаторных последовательностей каждого символа алфавита $\aleph$, сконструированных из сравниваемых символьных. По теореме о свёртке свёртка двух числовых последовательностей может быть получена как обратное преобразование поточечного произведения отдельных дискретных преобразований Фурье каждой последовательности. Поскольку в предложенном алгоритме вычисляется свертка индикаторных последовательностей для каждого символа алфавита, то результатом работы метода является целочисленная последовательность $\mathbf{C} = \{c_1, \ldots, c_N\}$, где каждое значение отражает число попарных совпадений символов в соответствующем наложении, причем нулевым считается наложение, в котором первый символ последовательности $\mathbb{P}$ накладывается на последний символом последовательности $\mathbb{Q}$. Каждое последующее наложение получается путем сдвига последовательности $\mathbb{Q}$ на один символ вправо относительно фиксированной последовательности $\mathbb{P}$. Таким образом, $N = n_P + n_Q - 1$. Эффективность метода основана на применении быстрого дискретного преобразования Фурье для вычисления сразу всей $\mathbf{C}$. Метод позволяет быстро найти самое выгодное с точки зрения совпадения символов наложение $c^* = \max\limits_{1 \le i \le N} \{c_i\}$. При этом допускаются мутации символов в совпадающих участках (замены, вставки и выпадения). Однако, метод ничего не говорит о расположении сходных подпоследовательностей непосредственно внутри последовательностей $\mathbb{P}$ и $\mathbb{Q}$. Следовательно, возникает задача локализации сходных подпоследовательностей.

В настоящей работе предложен метод локализации сходных участков в сравниваемых последовательностях. Показано, что с помощью c^* можно оценить длину совпадающего участка. Если затем тем же методом из [1] построить еще одну свертку $\mathbf{C}_1$ индикаторных последовательностей со специальным весом, то можно оценить еще один параметр совпадающего участка, а именно, его середину относительно начала наложения. Это позволяет хорошо локализовать совпадающий участок.

В работе продемонстрирована хорошая точность локализации на реальных геномных последовательностях со вставками.

Работа поддержана Красноярским математическим центром, финансируемым Минобрнауки РФ в рамках мероприятий по созданию и развитию региональных НОМЦ (Соглашение 075-02-2025-1606).

Список литературы

[1] Shaidurov V., Karepova E., Sadovsky M. et al. Highly Parallel Convolution Method to Compare DNA Sequences with Enforced In/Del and Mutation Tolerance // LNBI - Part of the Lecture Notes in Computer Science. 2020. Vol. 12108. P. 472–481.