О. В. Савик

АЛГОРИТМЫ ДЛЯ АНАЛИЗА ДАННЫХ ДНК

Рассматриваются алгоритмы анализа данных, применяемых к строковым представлениям ДНК для определения их свойств и функций.

Введение

В настоящее время было создано достаточно много алгоритмов, направленных на изучение свойств и структуры последовательностей ДНК. Большинство этих алгоритмов применяют стандартную технику теорий вероятности и математической статистики для исследования свойств в строении ДНК. Основная задача, которая встает перед исследователем, получившим новую последовательность – определить, что это за последовательность, каковы ее функции и свойства. Наиболее простым способом является сравнение новой последовательности с уже известными последовательностями.

I. Статистический алгоритм поиска подобий между последовательностями

Для последовательностей S_1 и S_1 длины которых соответственно равны N_1 и N_2 строится прямоугольная матрица размером N_1 x N_2. В верхней строке и левом столбце которой записываются сами последовательности. В (i,j) позиции матрицы ставится точка, если i и j символы последовательностей S_1 и S_2 совпадают. Затем выделяются группы точек, расположенных на линии, параллельной диагонали. Полученные таким образом отрезки определяют схожие участки двух последовательностей, причем число точек на отрезке равно длине фрагментов. Этот метод очень прост и нагляден, однако его существенным недостатком является слишком большая избыточность, что затрудняет анализ полученной картины и увеличивает время обработки.

II. Алгоритм скользящей рамки

На каждом шаге алгоритма последовательность S_1 сдвигается относительно S_2 на 1 основание(символ). Общая часть этих последовательностей сканируется окном длиной W и определяется число совпадающих символов. Если это число больше некоторого K, считается, что совпадение найдено. Число операций, выполняемых алгоритмом – N_1xN_2xW достаточно велико. Поэтому на практике применяют его модификацию, позволяющую сократить трудоемкость в W раз. При этом учитывается тот факт, что при сдвиге на одно основание количество совпадающих букв изменяется только за счет граничных символов. Таким образом, полный пересчет совпадений на каждой итерации не осуществляется.

III. Алгоритм нахождения оптимального выравнивания

Существенным недостатком статистических алгоритмов является то, что они не учитывают наличие погрешностей в сравниваемых последовательностях. Поэтому их точность в некоторых случаях весьма невысока.

Для учета погрешностей применяется алгоритм нахождения оптимального выравнивания. Определяется функция сходства F, которая учитывала не только число совпадений, но также и ошибки преобразования ДНК в строковую последовательность. В качестве параметров функции вводятся веса, увеличивающие функцию F при обнаружении совпадения и штрафы за погрешности, такие как замены и вставки.

$$F = K_m * V_m - K_d * V_d - K_c * V_c,$$

где K_m, K_d, K_c- количество совпадений, вставок и замен, V_m, V_d, V_c - параметры, характеризующие веса совпадений, вставок и замен соответственно.

Поиск оптимального значения функции сходства осуществляется в два этапа. На первом - для двух сравниваемых последовательностей строится точечная матрица совпадений. Движение по матрице вдоль диагонали через пустые клетки соответствует заменам, через заполненные - совпадениям, движения вверх или вправо - вставкам. Выравнивание представляет из себя путь из правого верхнего угла точечной матрицы совпадений в левый нижний. При этом необходимо, чтобы этот путь соответствовал максимуму функции сходства. Задача о поиске оптимального пути решается методами динамического программирования. Для построения оптимального пути заполняется матрица наилучших значений функции, по которой затем восстанавливается весь путь. Наиболее удачными реализациями алгоритма оптимального выравнивания являются прогрммные реализации FASTA и BLAST.

1. Франк-Коменецкий, М. Д. Компьютерный анализ генетических текстов // Москва, Наука. – 1990.

Савик Олег Владимирович, магистрант кафедры информационных технологий автоматизированных систем БГУИР, oleg.savik1996@gmail.com.

Научный руководитель: Курулев Александр Петрович, профессор, кандидат технических наук, kaftoe@bsuir.by.