## Секция «Биоинженерия и биоинформатика»

### De novo поиск повторов в геноме, основанный на данных о количестве встречаемых пар коротких слов на различных расстояниях

***Яшина Ксения Игоревна***
*Студент*
*Московский государственный университет имени М.В. Ломоносова, Факультет биоинженерии и биоинформатики, Москва, Россия*
*E-mail: ks.yashina@gmail.com*

Значительная часть генетического материала живых организмов представлена повторяющимися последовательностями, многие из которых могут перемещаться и встраиваться в различные участки генома, в частности в регуляторные и кодирующие области. Определение положения повторов в геноме, а также построение библиотеки их последовательностей играет важную роль в изучении закономерностей развития и эволюции генома. Для решения этой важной и непростой биоинформатической задачи было разработано несколько подходов. Однако ни один из них полностью не отвечает поставленной задаче из-за разнообразия длин и сложностей повторяющихся последовательностей [1]. В данной работе предлагается новый метод для поиска длинных повторов de novo.

Алгоритм реализован в программах на языках Pascal и Python. Принимая во входных данных последовательность генома, он создает список групп, каждая из которых состоит из фрагментов генома с общим мотивом в последовательностях. *На первом этапе* работы алгоритма подсчитывается число всех пар коротких – длины 5 – слов, расположенных в геноме на нескольких фиксированных расстояниях друг от друга. Производится отбор перепредставленных пар на основе списка для двух слов на нескольких разных расстояниях: число пар слов на расстоянии их взаимного расположения в повторе будет значительно превосходить число пар тех же слов, но на других расстояниях. *На втором этапе* по отобранным парам строятся паттерны, и с их помощью осуществляется поиск соответствующих фрагментов генома. Для каждой пары находок одного и того же паттерна строится парное выравнивание, и с использованием критериев сходства находятся точные границы парного повтора. *На третьем этапе* из парных повторов строятся группы. Парные повторы объединяются в группу, если образующие их фрагменты пересекаются в геноме.

Благодаря наличию промежутка из случайных нуклеотидов между словами из подсчитываемых пар, алгоритм обладает преимуществом в нахождении повторов, в которых в ходе эволюции появлялись точечные мутации, вставки, делеции.

Алгоритм проверялся двумя способами. Во-первых, показано, что по паттернам, отобранным по результатам анализа генома человека, находятся все (кроме трех) из известных длинных повторов, представленных консенсусной последовательностью. Во-вторых, алгоритм применен для генома бактерии *Brucella melitensis,* для которой известна повышенная активность мобильных элементов. Найдены 18 повторов длины более 50. Кроме того, найдены 49 коротких – длины от 15 до 39 – повторов, что неожиданно, так как основное направление алгоритма – поиск длинных повторов. Проведен анализ положения найденных повторов относительно положения генов *B.melitensis;* обнаружено, что некоторые из них совпадают или пересекаются с генами бактерии.

Таким образом, показана перспективность разработанного алгоритма для поиска de novo длинных повторов в геномах.

## Литература

1. Saha S, Bridges S, Magbanua ZV, Peterson DG. Empirical comparison of ab initio repeat finding programs. Nucleic Acids Res. 2008 April; 36(7): 2284–2294

## Слова благодарности

Благодарю Нагаева Бориса Эдуардовича и Алексеевского Андрея Владимировича за помощь и поддержку в работе