Определение численных метрик нуклеотидных последовательностей геномов, полученных в процессе секвенирования

Подвиг Макар Васильевич, Злобин Георгий Максимович
ГБОУ СОШ № 77, Санкт-Петербург
Научный руководитель - Логинов Константин Викторович
ГБОУ СОШ № 77, Санкт-Петербург

За последние 40 лет секвенирование ДНК стало одним из основных способов изучения живых организмов. При этом за последние 20 лет в биологии сформировалось новое направление – метагеномика, областью изучения которой является получение и исследование геномных последовательностей непосредственно из образцов среды (почв, океанов, кишечника человека). Под метагеномом понимают совокупность геномных последовательностей из таких сред, а набор организмов, присутствующих в среде, называют сообществом микроорганизмов из данной среды обитания.

Основной задачей анализа метагеномов является задача определения видового разнообразия организмов в данной среде. Секвенирование метагеномных последовательностей позволяет получить информацию об организмах, которые невозможно культивировать синтетически.

При секвенировании метагенома существуют наборы ридов, относящиеся как к разным организмам, так и относящиеся к одному. Для возможности решения задачи определения видового состава необходимо классифицировать последовательности относительно видов организмов, участками ДНК которых они являются.

Цель данного исследования - изучение различий метрик, получаемых на основе анализа геномов организмов различных таксономических единиц.

В ходе проведения исследования были разработаны алгоритмы определения GC состава геномов, вычисления частоты встречаемости нуклеотидных последовательностей длины k и вычисления расстояний между двумя геномами в пространстве тетрануклеотидов, количественно характеризующие различие состава ДНК двух организмов. Также были написаны программы на языке Python, реализующие данные алгоритмы. С помощью разработанных программ были проанализированы геномы вирусов, выявленных в клетке человека SARS-CoV-2/WH-09/human/2020/CHN (коронавирус, страна обнаружения: Китай), SARS-CoV-2/01/human/2020/SWE (коронавирус, страна обнаружения: Швеция), 2019-nCoV/USA-CA9/2020 (коронавирус, страна обнаружения: США); бактерий: Escherichia coli O25b:H4 chromosome (класс гаммапротеобактерий), Rhodoplanes sp. Z2-YC6860 (класс альфапротеобактерий).

Было установлено, что разница между GC составом представителей одного вида незначительна и составляет менее 0,5%. В свою очередь разница между GC составом коронавируса и бактерии кишечной палочки (Escherichia) составляет более 25%, а между составом коронавируса и бактерии Rhodoplanes - более 40%. Различие GC составов двух бактерий разных классов составляет порядка 20%.

В ходе анализа распределения частот встречаемости ди-, три- и тетрануклеотидных последовательностей, а также вычисления расстояний между рассматриваемыми геномами в 256-мерном пространстве тетрануклеотидов, было установлено, что данные метрики являются характерными для особей одного вида и различается у организмов разных таксономических единиц.

Результаты данного исследования подтверждают, что такие метрики как GC состав и распределение частоты нуклеотидных последовательностей длины k характеризуют геном определенного вида и могут быть использованы в качестве критериев для классификации таксономических единиц внутри метагеномных последовательностей в рамках решения задачи биннинга.