КОМПЬЮТЕРНЫЕ ОЦЕНКИ ТЕКСТОВОЙ СЛОЖНОСТИ ГЕНОМОВ ВИРУСОВ И БАКТЕРИЙ

Дергилев А.И., Сафрыгина А.А.[1], Орлов Ю.Л.[1]

Новосибирский госуниверситет, Новосибирск
[1]Первый МГМУ им. Сеченова Минздрава России (Сеченовский Университет), Москва
[2]orlov@bionet.nsc.ru

Развитие технологий высокопроизводительного секвенирования ДНК служит основой роста объема накопленных геномных данных для все более широкого набора организмов. Компьютерное исследование геномов как текста позволяет выявить закономерности и взаимозависимости между символами для классификации геномов вирусов и бактерий, их эволюционного происхождения. Разработка компьютерных алгоритмов анализа генетических текстов представляет собой важную задачу биоинформатики, реализуемую в студенческих дипломных работах в Сеченовском Университете, и в НГУ [1]. Развитие баз данных, таких как NCBI, EMBL, NGDC (National Genomics Data Center of China), насчитывающих уже миллионы полностью секвенированных геномов прокариот, позволяет рассматривать фундаментальные задачи анализа закономерностей передачи генетической информации в ходе эволюции.

Математические методы оценки сложности текста были реализованы в компьютерных программах расчета энтропии, сложности сжатия текста и лингвистической сложности нуклеотидных последовательностей. Их применение выявило различие в уровне сложности последовательностей функциональных участков генов – экзонов, интронов и промоторных районов [2]. Присутствие однонуклеотидных политрактов и коротких тандемных повторов, связанных с повышенной частотой мутаций, понижает сложность текста [2].

Пандемия коронавирусной инфекции вызвала интерес к анализу генома коронавируса и его подвидов [1]. Компьютерное исследование текста геномов прокариот дает основу для классификации, построения филогенетических деревьев. В геноме коронавируса обнаружены фрагменты с низкой сложностью, совпадающие с мононуклеотидными повторами. Анализ лингвистической сложности генома выявил что участок с наименьшей сложностью кодирует S-белок, который является мишенью при разработке противовирусных средств. Мы рассмотрели набор геномов патогенных вирусов и бактерий, определили участки низкой сложности и длинные повторяющиеся фрагменты ДНК.

Благодарности. Работа поддержана грантом Фонда Потанина для преподавателей магистратуры 2025 года (ГСГК-144/25).

Литература.

1. Orlov Y.L., Orlova N.G. Bioinformatics tools for the sequence complexity estimates. Biophysical Reviews. 2023; 15, 1367–1378. doi: 10.1007/s12551-023-01140-y
2. Сафронова Н.С., Пономаренко М.П., Абнизова И.И., Орлова Г.В., Чадаева И.В., Орлов Ю.Л. Фланкирующие повторы мономеров определяют пониженную контекстную сложность сайтов однонуклеотидных полиморфизмов в геноме человека. Вавиловский журнал генетики и селекции. 2015;19(6): 668-674. doi: 10.18699/VJ15.092