КОМПЬЮТЕРНЫЕ ПРОГРАММЫ ОЦЕНКИ СЛОЖНОСТИ ТЕКСТА ДНК ДЛЯ АНАЛИЗА СТРУКТУРЫ ГЕНОМОВ МИКРООРГАНИЗМОВ

Митина А.В.[1], Орлова Н.Г.[2], Дергилев А.И.[3,4], Орлов Ю.Л.[1,3,4]

[1] Первый МГМУ им. И.М. Сеченова Минздрава России (Сеченовский Университет)
ул. Трубецкая, д.8 стр.2, г. Москва, 119991, РФ
[2] Финансовый Университет при Правительстве РФ
Ленинградский проспект, 49/2, г. Москва, 125167, РФ
[3] Новосибирский государственный университет
ул. Пирогова, 1, г. Новосибирск, 630090, РФ
[4] Институт цитологии и генетики СО РАН
просп. ак. Лаврентьева, 10, г. Новосибирск, 630090, РФ; e-mail: orlov@d-health.institute
Поступила в редакцию 02.08.2023. DOI: 10.29039/rusjbpc.2023.0640

Аннотация. Одна из классических задач биоинформатики - поиск повторов и статистически неоднородных участков последовательностей ДНК и полных геномов микроорганизмов. Теоретические подходы к исследованию сложности текста последовательностей макромолекул - ДНК, РНК и белков – развивались до появления полных геномных последовательностей и получили новый импульс в связи с распространением технологий массового параллельного секвенирования и бурным ростом доступных данных. Рассматриваются современные компьютерные методы и существующие программы оценки сложности текста ДНК и построения профиля свойств для анализа структуры геномов микроорганизмов. Дан обзор доступных онлайн-программ для поиска и визуализации повторов текста. Представлена собственная компьютерная реализация метода оценки лингвистической сложности текста и сжатия по Лемпелю-Зиву для выявления структурных особенностей и аномалий геномов микроорганизмов. Представлены примеры профилей анализа сложности текста. Рассмотрено применение оценок сложности к анализу последовательности генома коронавируса SARS-CoV2, последовательности вируса эндемического паротита *Mumps Orthorubulavirus*. Выявлены участки низкой сложности текста.
Ключевые слова: *биоинформатика, биофизические модели, сложность текста, геномы микроорганизмов.*

ВВЕДЕНИЕ

Одной из важнейших задач биоинформатики - исследование сложности символьных последовательностей (ДНК, белков и полных генов), поиск статистически неоднородных участков последовательностей [1]. Для изучения сложности генетических текстов применяют алгоритмы сжатия данных, анализ частоты повторов определенных участков в последовательности, таких как тандемные повторы, которые могут быть связаны с функциональными элементами генома [2-4]. Оценка энтропии (энтропия Шеннона) может быть применена для анализа вариабельности участков последовательностей [5].

Для понимания генетической информации и ее связи с фенотипическими свойствами организмов необходимо проведение анализа последовательностей ДНК как текста, основанных на вычислении числовых значений свойств данной молекулы, таких как содержание GC, аминокислотного состава, вторичной структуры. Эти значения могут быть использованы для создания профиля свойств ДНК [6], что позволяет сравнивать различные последовательности и выявлять свойства, способные влиять на их функциональность [7-9].

Методы профилирования свойств ДНК с использованием компьютерных подходов включают в себя различные алгоритмы и программные инструменты, которые обрабатывают и анализируют биологические данные ДНК для извлечения полезной информации. Примеры применения профилей свойств включают:

1. Сравнение последовательностей ДНК, в том числе выравнивание с использованием физико-химических свойств (классические алгоритмы BLAST [10,11] и последующие модификации). Такие методы позволяют определить гомологию с другими известными последовательностями, предсказать функциональные элементы (например, гены, регуляторные регионы) и проводить анализ мутаций.
2. Предсказание неканонических структур ДНК на основе последовательности [12]. Это позволяет идентифицировать области, обладающие особыми структурными свойствами, такими как G-квадруплексы или двойные спиральные участки [13].
3. Анализ участков, мотивов и сайтов связывания транскрипционных факторов. Эти методы ищут мотивы (повторяющиеся паттерны символов) в последовательностях ДНК, которые могут быть связаны с функциональными элементами, такими как участки связывания белков или регуляторные мотивы промоторов генов [14]. Ранее были даны статистические оценки различия энтропии регуляторных последовательностей генов (сайтов связывания белков – транскрипционных факторов) и фланкирующих последовательностей [15]. Получающие все большее развитие алгоритмы машинного обучения и искусственного интеллекта широко используются для обнаружения и анализа таких мотивов.

4. Предсказание участков эпигенетических модификаций генома, к которым относятся участки метилирования ДНК и модификации гистонов, по статистическим свойствам последовательности [16]. Эти участки связаны с поиском CpG островов [17,18] (участков, обогащенных нуклеотидами GC, которые могут быть метилированы), и позиционированием нуклеосом на последовательности (такие участки на ДНК связаны с модификаций гистонов в составе нуклеосомы) [19,20]. Компьютерные методы могут быть использованы для теоретического предсказания мест эпигенетических модификаций на основе последовательности ДНК, что определяется в настоящее время в основном экспериментально, с помощью технологий серии ChIP-seq (Chromatin immunoprecipitation and sequencing) [21]. Это позволяет понять, как эпигенетические модификации влияют на функциональность генома [22].

5. Моделирование и симуляция структуры. С использованием компьютерных моделей и алгоритмов можно моделировать физические свойства и процессы, связанные с ДНК [6], такие как упаковка, взаимодействие с белками или лекарствами. Это позволяет предсказывать и понимать различные аспекты функционирования ДНК.

Цели данной работы – рассмотреть существующие методы и интернет-доступные программы оценки сложности и визуализации генетических текстов и анализа структурных особенностей полных геномов микроорганизмов, и разработать собственную программу оценки лингвистической сложности текста. В ходе исследований разработан оригинальный программный код для методов оценки сложности генетических текстов [23]. Программа применена для анализа сложности текста коронавируса и полных геномов микроорганизмов.

В целом, оценки сложности генетического текста [24] дали толчок развитию методов предсказания кодирующих участков ДНК, сайтов позиционирования нуклеосом [19], некодирующих РНК [2]. Использование оценок сложности улучшает предсказание регуляторных элементов генов [19,25], хотя является скорее дополнительным компьютерным инструментом для интерпретации найденных закономерностей.

Выявление участков низкой сложности в геномных последовательностях микроорганизмов позволяет определять быстро эволюционирующие части геномов прокариот [26], в том числе в связи с эволюцией окружающей среды [27], исследовать контекстные особенности участков мутаций генов эукариот [15,28], геномных участков сайтов связывания транскрипционных факторов [29,30]. Применительно к вирусам, оценки сложности позволяют рассмотреть связь таких участков с изменениями патогенности [31,32].

МАТЕРИАЛЫ И МЕТОДЫ

Материалами исследования являлись геномные последовательности и их функциональная разметка. Использовались данные из открытых баз данных NCBI (https://www.ncbi.nlm.nih.gov/) и UCSC Genome Browser (https://genome.ucsc.edu/).

Рассмотрены существующие методы оценки сложности последовательности символов и компьютерные реализации применительно к ДНК [1,2]. Общий, наиболее фундаментальный подход к определению сложности символьных последовательностей (текстов) был предложен академиком А.Н. Колмогоровым в 1960-х. Число операций кодирования называется алгоритмической сложностью (сложностью по Колмогорову) [33]. Комбинаторная сложность позволяет находить участки повышенной сложности и простые участки, зачастую соответствующие коротким тандемным повторам. Операционной сложностью называют число операций, необходимых для сжатия текста алгоритмом Лемпеля-Зива [34]. В основе данного алгоритма лежит последовательное сканирование текста и добавление каждой последовательности символов, которую алгоритм встречает впервые, в словарь. Если алгоритм находит последовательность символов, которая уже есть в словаре, то он заменяет эту последовательность на ссылку в соответствующую запись в словаре. Применительно к ДНК могут быть использованы операции инвертирования (учет комплементарных) участков, поиск повторов текста.

Основным доступным инструментом является программный комплекс Complexity (http://wwwmgs.bionet.nsc.ru/mgs/programs/lzcomposer/) [24], разработанный в Институте цитологии и генетики СО РАН. Он позволяет определить сложность генетического текста по Лемпелю-Зиву [2] (операционная сложность) вместе с энтропийными оценками. Методика получила свое развитие в серии применений для анализа регуляторных районов ДНК эукариот [35], сравнения сложности белок-кодирующих и некодирующих последовательностей [36], фланкирующих районов участков мутаций [15,28].

Энтропия Шеннона может быть применена как для анализа коротких последовательностей олигонуклеотидов, так и для протяженных последовательностей. Используется только частный состав символов (из алфавита A, T, C, G для ДНК). Вычисляется данная величина по следующей формуле (адаптировано из [2]):

$$H(x) = -\sum_{i=0}^{n} p(i) log_n p(i), \quad (1)$$

где $p(i) = n_i/L, n_i -$ число встретившихся символов типа i, L – длина последовательности (окна, в котором идет расчет), n – мощность алфавита (для алфавита ДНК равна 4).

Высокая энтропия генетического текста указывает на наличие разнообразных последовательностей (отсутствие повторов). Кроме того, энтропия Шеннона может быть использована для определения потенциальных границ функциональных областей в геноме.

Операционной, или компрессионной, сложностью называют число операций, необходимых для сжатия текста алгоритмом Лемпеля-Зива [24]. Для применения данного алгоритма к нуклеотидным последовательностям

необходимо учитывать все возможные виды повторов (помимо прямых повторов учитываются также инвертированные и комплементарные).

Лингвистической, или комбинаторной сложностью генетических текстов называют отношение числа встретившихся слов к числу всех возможных слов в последовательности заданной длины [37,38]. Формула расчета лингвистической сложности *C* для последовательности символов длины *L*:

$$C = (\sum_{i=1}^{m} V_i)/(\sum_{i=1}^{m} V_{i\,max}), \quad (2)$$

где V_i – число слов длины *i*, *m* – максимальная длина слова $1 \le m \le L$. $V_{i\,max}$ – максимально возможное число слов длины *i* в последовательности длины *L*. $V_{i\,max}$ равен наименьшему из двух чисел: K^i, равное числу всех возможных слов длины *i*, или числу слов, которые можно разместить в последовательности – (*N-i+1*).

Примеры расчета лингвистической сложности даны в работах [2,36].

Разработанная программа осуществляет расчет и визуализацию лингвистической сложности генетического текста, следуя формулам представленным в [2] и [38]. Разработка велась в среде Jupiter Notebook на языке Python (https://jupyter.org/). Программа рассчитывает лингвистическую сложность для любой последовательности, состоящей из четырехбуквенного алфавита. Программный код доступен по ссылке https://github.com/Alinabio/complexity.git.

РЕЗУЛЬТАТЫ

1. Разработка программы. Техническим результатом работы является разработка компьютерного кода. Программа рассчитывает лингвистическую сложность для любой последовательности, состоящей из четырехбуквенного алфавита.

Программа представляет собой простой и понятный инструмент для расчета и визуализации сложности геномов (в форме профиля). Работа с программой осуществляется через python-код, который является одним из основных инструментов биоинформатики (https://www.python.org/downloads/). Программа позволяет эффективно рассчитывать значения лингвистической сложности текста [38], в том числе и для длинных последовательностей (полных хромосом и геномов прокариот). Также программа предоставляет возможность настройки параметров (длины окна, шага, алфавита), что способствует получению результатов, соответствующих потребностям пользователя.

Программный код linguistic complexity состоит из трех смысловых блоков: расчет сложности всей последовательности (с функциями входа), расчет сложности в окне (count_ling_complexity), выдача результатов и визуализация (построение профиля).

Функция count_ling_complexity принимает на вход последовательность, равную длине окна и рассчитывает значение сложности для данной последовательности. Для расчета знаменателя используется цикл for и условные конструкции. В зависимости от величины i к значению знаменателя (denominator), которое по умолчанию равно 4 (количество возможных нуклеотидов длины 1) прибавляется число *i**alphabet* (если *i**alphabet* меньше длины окна) или число len_window - (i-1), если *i**alphabet* больше длины окна.

Далее рассчитывается числитель для последовательности, который равен числу различных встретившихся слов в последовательности. Все встретившиеся слова записываются во множество set_numerator, которое оставляет только уникальные значения. Длина множества равна числителю. В переменной result рассчитывается значение сложности для конкретной последовательности, оно равно отношению числителя на знаменатель. Программа возвращает результат переменной result.

Функция linguistic_complexity принимает на вход уже четыре параметра: на вход передается файл в формате FASTA, длина окна – len_window (по умолчанию - 50), величина шага - step (по умолчанию - 100), размер алфавита – alphabet (по умолчанию - 4). Функция читает FASTA-файл с одной последовательностью и переводит её в строку. Далее создается словарь, ключами которого будут являться серединные позиции окна последовательности, а значениями – лингвистические сложности. Далее в цикле происходит заполнение словаря, значения рассчитываются на основании расчета функции count_ling_complexity. В итоге функция выводит словарь, где ключом служит позиция нуклеотида в середине окна, а значением - рассчитанное значение сложности окна.

Для визуализации полученных результатов была написана функция visualization. Для работы функции необходима установка библиотеки Matplotlib (https://matplotlib.org/). Для получения графика необходим ее модуль Pyplot (https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.plot.html). С помощью данного модуля строится линейный график, где можно увидеть участки низкой и высокой сложности. График может быть конвертирован в текстовый формат для визуализации сторонними программами (Excel).

Компьютерную модель порождения последовательности ДНК с помощью повторов текста (дупликаций и инверсий) [24] можно рассматривать как информационную биофизическую модель. Мы исследовали применения этой модели для анализа последовательности генома коронавируса и поиска ассоциаций с участками патогенности [31,32].

2. Анализ сложности генома коронавируса. Пандемия COVID-19 придала импульс исследованиям в данной области [9], в том числе развитию баз данных биоинформатики для вирусов, масштабному сбору

статистики о каждой мутации в известных штаммах, детальному статистическому анализу и поиску новых моделей связи мутаций в геноме коронавируса и патогенности.

С использованием разработанной программы в геноме коронавируса были найдены участки низкой сложности текста. Их расположение соответствовало участкам повторяющихся мононуклеотидов, находившиеся между белок кодирующими последовательностями. На вход программа принимала последовательность всего генома SARS-CoV-2, полученную из базы данных NCBI (https://www.ncbi.nlm.nih.gov/datasets/taxonomy/694009/).

SARS-CoV-2 относится к группе одноцепочечных РНК-вирусов. Геном данного вируса один из самых крупных вирусных геномов и составляет 29,9 Кб. Структурная организация генома SARS-CoV-2 приведена на рисунке 1.

Наиболее важными генами коронавируса, кодирующими белки, являются:

1. Ген, кодирующий S-белок. S-белок представляет собой гликопротеин оболочки, который играет наиболее важную роль в прикреплении вируса, слиянии и проникновении в клетки-хозяева и служит основной мишенью для разработки нейтрализующих антител, ингибиторов проникновения вируса и вакцин.
2. Ген, кодирующий M-белок. Этот белок играет ключевую роль в формировании вирусной оболочки.
3. Ген, кодирующий E-белок. Данный белок участвует в сборке вирусной частицы.
4. Ген, кодирующий N-белок. Белок принимает участие в защите вирусной РНК от воздействия организма хозяина.
5. Ген, кодирующий белки, связанные с репликацией, процессингом новых вирусных частиц. Этот процесс происходит внутри инфицированной клетки, где вирус использует клеточные механизмы для создания копий своего генома и новых вирусных частиц.
6. Гены, кодирующие белки, которые вызывают иммунный ответ организма [39].

Гены в геноме коронавируса и участки полиморфизмов являются объектами детального изучения с точки зрения патогенности, и функциональной аннотации [40].

В ходе предварительных расчетов нами показано, что наиболее приемлемым вариантом для оценки является длина окна 50 нт, так как данная длина наиболее соответствует участкам низкой сложности текста, находящимся на стыке генов. Далее был построен отдельный график расчета профиля сложности последовательности коронавируса с длиной окна, равной 50 нт (рис. 2).

Наименьшая сложность наблюдается на участке позиции 24625. При расчете лингвистической сложности были получены наиболее интересные результаты, так как участок низкой сложности был найден в гене, кодирующем белок S. Предполагается, что участки низкой сложности наиболее подвержены мутациям, что как раз соотносится с тем, что S-белок служит основной мишенью для разработки нейтрализующих антител, ингибиторов проникновения вируса и вакцин.

3. Анализ генома вируса эндемического паротита. Был проведен анализ последовательности вируса эндемического паротита *Mumps Orthorubulavirus*, вызывающего одноименное заболевание. Паротит известен как опасное вирусное заболевание детского возраста, которое можно предотвратить с помощью вакцины [41]. Однако в последние годы в развитых странах было зарегистрировано несколько крупных вспышек инфекции, вызванной вирусом паротита [42]. При эпидемическом паротите имеет место не только поражение железистых органов (паротит, субмандибулит, сублингвит, панкреатит, орхит, простатит, оофорит – в 5% случаев у девушек и девочек, мастит, тиреоидит, дакриоаденит), но и длительная циркуляция возбудителя в крови. При осложненном эндемическом паротите возможно развитие серозного менингита и менингоэнцефалита, миелита и энцефаломиелита, поражение черепных нервов. В исходе паротит нередко приводит к поражению центральной нервной системы, формирует бесплодие у 50% мужчин [43].

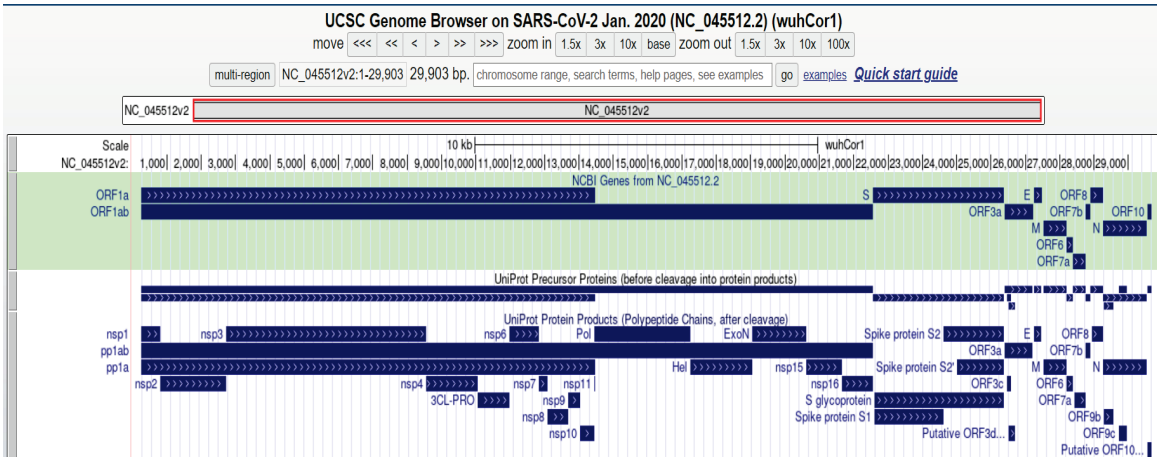

Рисунок 1. Структурная организация генома SARS-CoV-2. Рисунок построен с помощью геномного браузера UCSC (https://genome.ucsc.edu/)

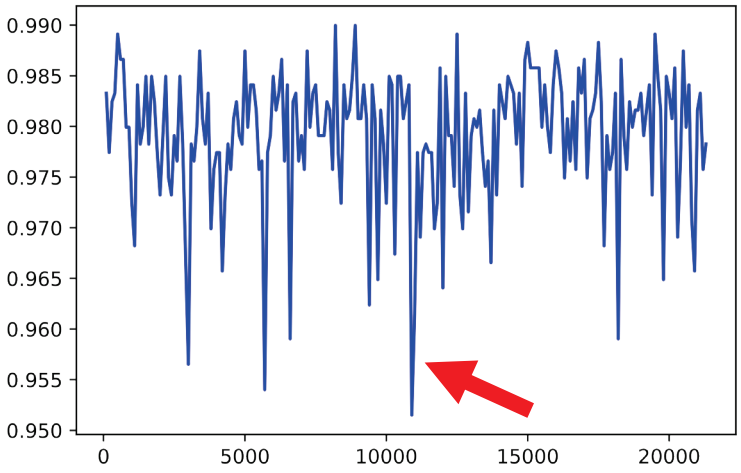

Рисунок 2. Расчет лингвистической сложности для SARS-CoV-2 при шаге, равном 200, и длине окна, равной 50 нуклеотидам (нт). По оси X – позиция в геноме, нт; по оси Y – значение сложности участка текста, значения в интервале [0;1]

Геном вируса составляет 15384 нт и содержит семь генов, кодирующих девять белков (рис. 3). Для данного РНК-вируса был построен профиль лингвистической сложности (рис. 4).

Расчет лингвистической сложности для *Mumps orthorubulavirus* был выполнен при шаге, равном 200, и длине окна, равной 50. Найден участок пониженной сложности в области 1625 нт. Данная позиция соответствует гену NP, который кодирует белок, участвующий в образовании нуклеокапсида. Тот факт, что участок низкой сложности был найден для гена NP, свидетельствует о том, что данный ген наиболее подвержен различного рода мутациям. Этот участок может быть мишенью для лекарственных воздействий. Разработаны противовирусные препараты, направленные на нарушение сборки нуклеокапсида.

ЗАКЛЮЧЕНИЕ

Представлены ключевые компоненты разработанной программы, ее возможности и ограничения, а также примеры ее использования для анализа последовательностей полных генов микроорганизмов, их эволюции [23,26]. Разработка программных конвейеров и инструментов биоинформатики проводится на Цифровой кафедре Сеченовского университета и имеет как исследовательский, так и образовательный характер (https://dk.sechenov.ru/). В данной работе, защищенной в качестве диплома на Цифровой кафедре, рассмотрен процесс разработки программы визуализации лингвистической сложности генома. Представлено общее описание работы программы, описаны ее основные компоненты и алгоритмы.

На примере генома SARS-CoV-2 и вируса эндемического паротита было показано, что программа может быть использована для анализа различных нуклеотидных последовательностей и широкого круга задач. Ранее проводились похожие исследования по оценке сложности текста, которые позволили выявить горячие точки мутации для генома коронавируса [32], представляющие интерес как эволюционирующая нуклеотидная последовательность минимального размера.

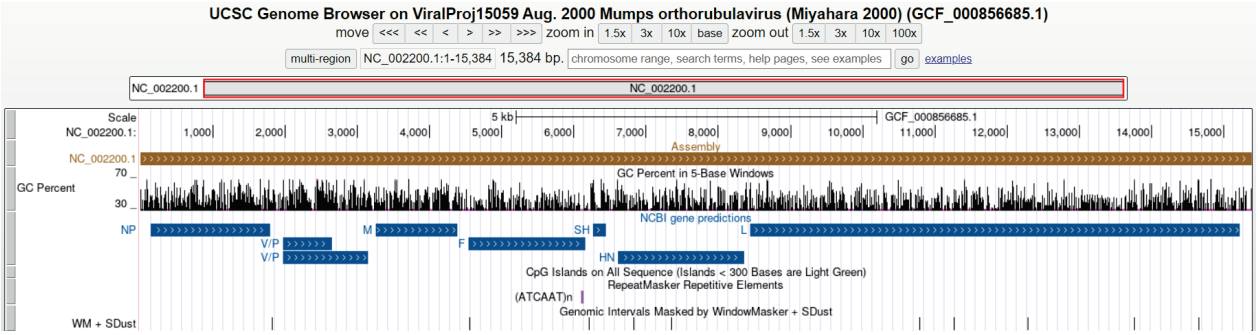

Рисунок 3. Структурная организация генома *Mumps orthorubulavirus*. Рисунок построен с помощью геномного браузера (https://genome.ucsc.edu/)

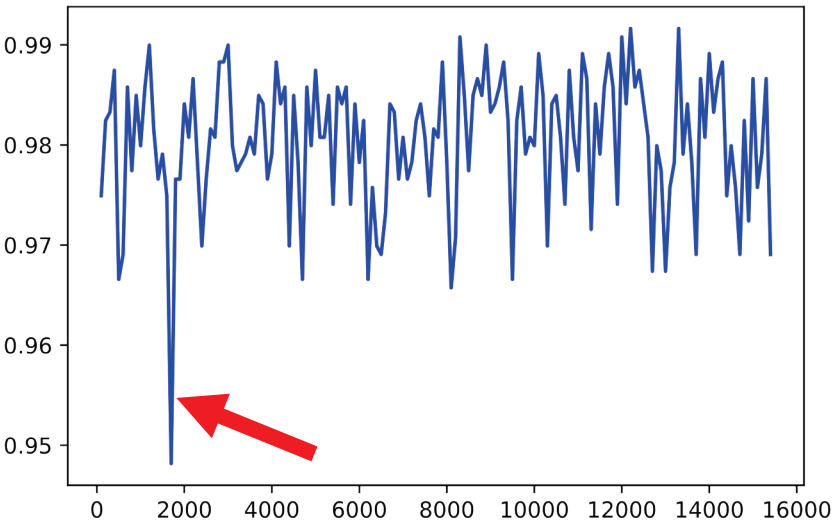

Рисунок 4. Расчет лингвистической сложности для генома *Mumps orthorubulavirus* при шаге, равном 200, и длине окна, равной 50. Обозначения шкал соответствуют рис. 2. Отмечен участок пониженной сложности в области 1625 нт

В биологических системах показано различие в уровнях сложности между белок-кодирующими и некодирующими последовательностями ДНК [36]. Таким образом, можно сказать, что лингвистическая сложность отличается в различных участках ДНК. Белок-кодирующие последовательности обычно более консервативны и менее изменчивы, поскольку они несут ключевую информацию о синтезе белков, который является важным для выживания организма [4]. Некодирующие последовательности, с другой стороны, обладают большей изменчивостью и гибкостью, поскольку они выполняют разнообразные функции, связанные с регуляцией генов и структурой генома. Понимание этих различий помогает нам лучше осознать сложность генома и его эволюционные механизмы. Оценка лингвистической сложности нуклеотидных последовательностей служит основой статического анализа структуры геномов микроорганизмов и патогенов растений [44], важна для задач биотехнологии. Продолжение работы требует интеграции с существующими базами отечественными программными ресурсами биоинформатики, такими как ICGenomics [45], и будет развиваться для анализа геномных последовательностей патогенов растений.

Программный код доступен по ссылке https://github.com/Alinabio/complexity.git.

Благодарности. Работа поддержана грантом РНФ (23-44-00030). Авторы выражают благодарность А.Ю. Потаповой, П.А. Иванову-Ростовцеву и Е.А. Савиной за техническую помощь в работе.

Список литературы / References:

1. Simoes R.P., Wolf I.R., Correa B.A., Valente G.T. Uncovering patterns of the evolution of genomic sequence entropy and complexity. *Mol Genet Genomics*, 2021, vol. 296, no. 2, pp. 289-298, doi: 10.1007/s00438-020-01729-y.
2. Orlov Y.L., Potapov V.N. Complexity: an internet resource for analysis of DNA sequence complexity. *Nucleic Acids Res.*, 2004, vol. 32, pp. W628-W633, doi: 10.1093/nar/gkh466.
3. Bartal A., Jagodnik K.M. Progress in and Opportunities for Applying Information Theory to Computational Biology and Bioinformatics. *Entropy (Basel)*, 2022, vol. 24, no. 7, pp. 925, doi: 10.3390/e24070925.
4. Bernaola-Galvan P., Carpena P., Gomez-Martin C., Oliver J.L. Compositional Structure of the Genome: A Review. *Biology (Basel)*, 2023, vol. 12, no. 6, p. 849, doi: 10.3390/biology12060849.
5. Chang C.H., Hsieh L.C., Chen T.Y., Chen H.D., Luo L., Lee H.C. Shannon information in complete genomes. *J. Bioinform. Comput. Biol.*, 2005, vol. 3, no. 3, pp. 587-608, doi: 10.1142/s0219720005001181.
6. Olson W.K., Zhurkin V.B. Modeling DNA deformations. *Curr Opin Struct Biol.*, 2000, vol. 10, no. 3, pp. 286-297, doi: 10.1016/s0959-440x(00)00086-5.
7. Orlov Y.L., Filippov V.P., Potapov V.N., Kolchanov N.A. Construction of stochastic context trees for genetic texts. *In Silico Biol.*, 2002, vol. 2, no. 3, pp. 233-247.
8. Chanda P., Costa E., Hu J., Sukumar S., Van Hemert J., Walia R. Information Theory in Computational Biology: Where We Stand Today. *Entropy*, 2020, vol. 22, no. 6, p. 627, doi: 10.3390/e22060627.
9. Akbari Rokn Abadi S., Mohammadi A., Koohi S. A new profiling approach for DNA sequences based on the nucleotides' physicochemical features for accurate analysis of SARS-CoV-2 genomes. *BMC Genomics*, 2023, vol. 24, no. 1, p. 266, doi: 10.1186/s12864-023-09373-7.
10. Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W., Lipman D.J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, 1997, vol. 25, no. 17, pp. 3389-3402, doi: 10.1093/nar/25.17.3389.

11. Berselli M., Lavezzo E., Toppo S. NeSSie: a tool for the identification of approximate DNA sequence symmetries. *Bioinformatics*, 2018, vol. 34, no. 14, pp. 2503-2505, doi: 10.1093/bioinformatics/bty142.

12. Andersen E.S. Prediction and design of DNA and RNA structures. *New Biotechnology*, 2010, vol. 27, no. 3, pp. 184-193, doi: 10.1016/j.nbt.2010.02.012.

13. Shi X., Teng H., Sun Z. An updated overview of experimental and computational approaches to identify non-canonical DNA/RNA structures with emphasis on G-quadruplexes and R-loops. *Brief Bioinform.*, 2022, vol. 23, no. 6, pp. bbac441, doi: 10.1093/bib/bbac441.

14. Narad P., Kumar A., Chakraborty A., Patni P., Sengupta A., Wadhwa G., Upadhyaya K.C. Transcription Factor Information System (TFIS): A Tool for Detection of Transcription Factor Binding Sites. *Interdiscip Sci.*, 2017, vol. 9, no. 3, pp. 378-391, doi: 10.1007/s12539-016-0168-5.

15. 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, т. 19, № 6, с. 668-674, doi: 10.18699/VJ15.092 [Safronova N.S., Ponomarenko M.P., Abnizova I.I., Orlova G.V., Chadaeva I.V., Orlov Y.L. Flanking monomer repeats determine decreased context complexity of single nucleotide polymorphism sites in the human genome. *Russian Journal of Genetics: Applied Research*, 2016, vol. 6, no. 8, pp. 809-815 (In Russ.)].

16. Vityaev E.E., Orlov Y.L., Vishnevsky O.V., Pozdnyakov M.A., Kolchanov N.A. Computer system "Gene Discovery" for promoter structure analysis. *In Silico Biol*., 2002, vol. 2, pp. 257-262.

17. Babenko V., Chadaeva I., Orlov Y. Genomic landscape of CpG rich elements in human genome. *BMC evolutionary biology*, 2017, vol. 17, suppl. 1, pp. 19, doi: 10.1186/s12862-016-0864-0.

18. Babenko V.N., Bogomolov A.G., Babenko R.O., Galieva E.R., Orlov Y.L. CpG islands' clustering uncovers early development genes in the human genome. *Computer Science and Information Systems*, 2018, vol. 15, no. 2, pp. 473-485, doi: 10.2298/CSIS170523004B.

19. Орлов Ю.Л., Левицкий В.Г., Смирнова О.Г., Подколодная О.А., Хлебодарова Т.М., Колчанов Н.А. Статистический анализ последовательностей ДНК, содержащих сайты формирования нуклеосом. *Биофизика*, 2006, т. 51, с. 608-14 [Orlov Y.L., Levitskii V.G., Smirnova O.G., Podkolodnaya O.A., Khlebodarova T.M., Kolchanov N.A. Statistical analysis of DNA sequences containing nucleosome positioning sites. *Biophysics*, 2006, vol. 51, no. 4, pp. 541-546 (In Russ.)].

20. Goh W.S., Orlov Y., Li J., Clarke N.D. Blurring of high-resolution data shows that the effect of intrinsic nucleosome occupancy on transcription factor binding is mostly regional, not local. *PLoS Comput Biol.*, 2010, vol. 6, no. 1, e1000649, doi: 10.1371/journal.pcbi.1000649.

21. Дергилев А.И., Спицина А.М., Чадаева И.В., Свичкарев А.В., Науменко Ф.М., Кулакова Е.В., Витяев Е.Е., Чен М., Орлов Ю.Л. Компьютерный анализ совместной локализации сайтов связывания транскрипционных факторов по данным ChIP-seq. *Вавиловский журнал генетики и селекции*, 2016, т. 20, № 6, с. 770-778, doi: 10.18699/VJ16.194 [Dergilev A.I., Spitsina A.M., Chadaeva I.V., Svichkarev A.V., NAumenko F.M., Kulakova E.V., Vityaev E.E., Chen M., Orlov Y.L. Computer analysis of colocalization of the TFs' binding sites in the genome according to the ChIP-seq data. *Russian Journal of Genetics: Applied Research*, 2017, vol. 7, no. 5, pp. 513-522 (In Russ.)].

22. Alipanahi B., Delong A., Weirauch M.T., Frey B.J. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotechnol.*, 2015, vol. 33, no. 8, pp. 831-838, doi: 10.1038/nbt.3300.

23. Митина А.В., Орлов Ю.Л. Оценка лингвистической сложности генетических последовательностей штаммов SARS-CoV-2. *Сборник научных трудов VII Съезда Биофизиков России*: в 2 томах, том 1 - Краснодар: Типография ФГБОУ ВО «КубГТУ», 2023, с. 330, doi: 10.26297/SbR6.2023.001 [Mitina A.V., Orlov Y.L. The estimates of linguistic complexity of genetic sequences of SARS-CoV-2 stamms. *Collection of scientific papers of the VII Congress of Biophysicists of Russia*: in 2 volumes, vol.1 - Krasnodar: Printing house of FGBOU VO "KubGTU", 2023, p. 330 (In Russ.)].

24. Orlov Y.L., Gusev V.D., Miroshnichenko L.A. LZcomposer: Decomposition of Genomic Sequences by Repeat Fragments. *Biofizika*, 2003, vol. 48, suppl. 1, pp. S7-S16.

25. Wu C., Chen J., Liu Y., Hu X. Improved Prediction of Regulatory Element Using Hybrid Abelian Complexity Features with DNA Sequences. *International Journal of Molecular Sciences*, 2019, vol. 20, no. 7, p. 1704, doi: 10.3390/ijms20071704.

26. Орлов Ю.Л., Митина А.В., Суслов В.В., Дергилев А.И. Компьютерные оценки информационной сложности геномов прокариот. *Тезисы докладов 4-й Всероссийской конференции по астробиологии «Геологические, биологические и биогеохимические процессы в решении астробиологических задач»* 27 февраля - 2 марта 2023 г., г.Пущино. Институт физико-химических и биологических проблем почвоведения РАН, с. 20-22 [Orlov Y.L., Mitina A.V., Suslov V.V., Dergilev A.I. Computer estimates of the information complexity of prokaryotic genomes. *Abstracts of the 4th All-Russian Conference on Astrobiology "Geological, biological and biogeochemical processes in solving astrobiological problems"* February 27 - March 2, 2023, Pushchino. Institute of Physicochemical and Biological Problems of Soil Science RAS, pp. 20-22 (In Russ.)].

27. Суслов В.В., Афонников Д.А., Подколодный Н.Л., Орлов Ю.Л. Особенности геномного контекста и GC состав генома прокариот в связи с эволюцией среды обитания. *Палеонтологический журнал*, 2013, т. 47, № 9, с. 1056-1060, doi: 10.1134/S0031030113090220 [Suslov V.V., Afonnikov D.A., Podkolodny N.L., Orlov Y.L. Genome

features and GC content in prokaryotic genomes in connection with environmental evolution. *Paleontological Journal*, 2013, vol. 47, no. 9, pp. 1056-1060 (In Russ.)].

28. Safronova N.S., Babenko V.N., Orlov Y.L. 117 Analysis of SNP containing sites in human genome using text complexity estimates. *Journal of Biomolecular Structure and Dynamics*, 2015, vol. 33, suppl. 1, pp. 73-74, doi: 10.1080/07391102.2015.1032750.

29. Дергилев А.И., Орлова Н.Г., Митина А.В., Орлов Ю.Л. Применение методов оценки сложности текста к анализу геномных кластеров сайтов связывания транскрипционных факторов. *Сборник научных трудов VII Съезда Биофизиков России:* в 2 томах, том 1 - Краснодар: Типография ФГБОУ ВО «КубГТУ», 2023, с. 335-336, doi: 10.26297/SbR6.2023.001 [Dergilev A.I., Orlova N.G., Mitina A.V., Orlov Y.L. Application of methods for assessing text complexity to the analysis of genomic clusters of transcription factor binding sites. *Collection of scientific papers of the VII Congress of Biophysicists of Russia*: in 2 volumes, vol.1 - Krasnodar: Printing house of FGBOU VO "KubGTU", 2023, pp. 335-336 (In Russ.)].

30. Dergilev A.I., Orlova N.G., Dobrovolskaya O.B., Orlov Y.L. Statistical estimates of multiple transcription factors binding in the model plant genomes based on ChIP-seq data. *J Integr Bioinform.*, 2021, vol. 19, no. 1, p. 20200036, doi: 10.1515/jib-2020-0036.

31. Принглаева А.М., Дергилев А.И., Панова А.Д., Орлов Ю.Л. Сложность текста и структура повторов генома на примере коронавируса. *Марчуковские научные чтения 2020*: Тезисы Междунар. конф., посв. 95-летию со дня рождения акад. Г. И. Марчука Новосибирск, 19-23 октября 2020 г. Ин-т вычислит. математики и матем. геофизики СО РАН, Новосибирск: ИПЦ НГУ, 2020, с. 167, doi: 10.24411/9999-017A-2020-10295 [Pringlaeva A.M., Dergilev A.I., Panova A.D., Orlov Y.L. The complexity of the text and the structure of genome repeats on the example of coronavirus. *Marchuk Scientific Readings 2020*: Abstracts of the Intern. conf., dedicated 95th anniversary of the birth of Acad. G. I. Marchuk Novosibirsk, October 19-23, 2020. Inst. Comput. mathematics and math. geophysics SB RAS, Novosibirsk: CPI NSU, 2020, p. 167 (In Russ.)].

32. Галиева А.Г., Лузин А.Н., Орлова Н.Г., Куликова Д.К., Дергилев А.И., Орлов Ю.Л. Биоинформационные подходы для анализа точек мутации генома коронавируса. В сборнике: *Молекулярная диагностика и биобезопасность-2021*. COVID-19: эпидемиология, диагностика, профилактика: сборник тезисов Онлайн-конгресса с международным участием (28-29 апреля 2021 г., Москва). М.: ФБУН ЦНИИ Эпидемиологии Роспотребнадзора, 2021, 144 с. [Galieva A.G., Luzin A.N., Orlova N.G., Kulikova D.K., Dergilev A.I., Orlov Y.L. Bioinformatics approaches to analyze the mutation points of the coronavirus genome. In the collection: *Molecular Diagnostics and Biosafety-2021*. COVID-19: epidemiology, diagnosis, prevention: collection of abstracts of the Online Congress with international participation (April 28-29, 2021, Moscow). M.: Central Research Institute of Epidemiology of Rospotrebnadzor, 2021, 144 p. (In Russ.)].

33. Antao R., Mota A., Machado J.A.T. Kolmogorov complexity as a data similarity metric: application in mitochondrial DNA. *Nonlinear Dyn.*, 2018, vol. 93, no. 3, pp. 1059-1071.

34. Dheemanth H.N. LZW Data Compression. *American Journal of Engineering Research (AJER)*, 2014, vol. 3, no. 2, pp. 22-26.

35. Putta P., Orlov Y.L., Podkolodnyy N.L., Mitra C.K. Relatively conserved common short sequences in transcription factor binding sites and miRNA. *Вавиловский журнал генетики и селекции*, 2011, т. 15, № 4, с. 750-756 [Putta P., Orlov Y.L., Podkolodnyy N.L., Mitra C.K. Relatively conserved common short sequences in transcription factor binding sites and miRNA. *Vavilov Journal of Genetics and Breeding*, 2011, vol. 15, no. 4, pp. 750-756 (In Russ.)].

36. Orlov Y.L., te Boekhorst R., Abnizova I.I. Statistical measures of the structure of genomic sequences: entropy, complexity, and position information. *J Bioinform Comput Biol.*, 2006, vol. 4, pp. 523-536.

37. Popov O., Segal D.M., Trifonov E.N. Linguistic complexity of protein sequences as compared to texts of human languages. *Biosystems*, 1996, vol. 38, no. 1, pp. 65-74, doi: 10.1016/0303-2647(95)01568-x.

38. Troyanskaya O.G., Arbell O., Koren Y., Landau G.M., Bolshoy A. Sequence complexity profiles of prokaryotic genomic sequences: a fast algorithm for calculating linguistic complexity. *Bioinformatics*, 2002, vol. 18, no. 5, pp. 679-688.

39. Lu R., Zhao X., Li J. et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet*, 2020, vol. 395, no. 10224, pp. 565-574, doi: 10.1016/S0140-6736(20)30251-8.

40. Hu B., Guo H., Zhou P. et al. Characteristics of SARS-CoV-2 and COVID-19. *Nat Rev Microbiol*., 2021, vol. 19, pp. 141-154, doi: 10.1038/s41579-020-00459-7.

41. Рубальская Т.С., Ерохов Д.В., Жердева П.Е., Милихина А.В., Гаджиева А.А., Тихонова Н.Т. Генотипирование вируса эпидемического паротита (*Paramyxoviridae: Orthorubulavirus: Mumps Orthorubulavirus*) как элемент лабораторного подтверждения инфекции. *Вопросы вирусологии*, 2023, т. 68, № 1, с. 59-65 [Rubalskaya T.S., Erokhov D.V., Zherdeva P.E., Milikhina A.V., Gadzhieva A.A., Tikhonova N.T. Genotyping of mumps virus (*Paramyxoviridae: Orthorubulavirus: Mumps Orthorubulavirus*) as element of laboratory confirmation of infection. *Questions of virology*, 2023, vol. 68, no. 1, pp. 59-65 (In Russ.)].

42. Su S.B., Chang H.L., Chen A.K. Current Status of Mumps Virus Infection: Epidemiology, Pathogenesis, and Vaccine. *Int J Environ Res Public Health*, 2020, vol. 17, no. 5, p. 1686, doi: 10.3390/ijerph17051686.

43. Юминова Н.В., Контарова Е.О., Балаев Н.В., Артюшенко С.В., Контаров Н.А., Россошанская Н.В., Сидоренко Е.С., Гафаров Р.Р., Зверев В.В. Вакцинопрофилактика кори, эпидемического паротита и краснухи: задачи, проблемы и реалии. *Эпидемиология и Вакцинопрофилактика*, 2011, т. 4, № 59, с. 40-44 [Yuminova N.V.,

Kontarova E.O., Balaev N.V., Artyushenko S.V., Kontarov N.A., Rossoshanskaya N.V., Sidorenko E.S., Gafarov R.R., Zverev V.V. Measles, mumps and rubella vaccination: tasks, problems and realities. *Epidemiology and Vaccinal Prevention*, 2011, vol. 4, no. 59, pp. 40-44 (In Russ.)].

44. Chao H., Zhang S., Hu Y., Ni Q., Xin S., Zhao L., Ivanisenko V.A., Orlov Y.L., Chen M. Integrating omics databases for enhanced crop breeding. *J Integr Bioinform.*, 2023, doi: 10.1515/jib-2023-0012.

45. Orlov Y.L., Bragin A.O., Babenko R.O., Dresvyannikova A.E., Kovalev S.S., Shaderkin I.A., Orlova N.G., Naumenko F.M. Integrated Computer Analysis of Genomic Sequencing Data Based on ICGenomics Tool. In: Advances in Intelligent Systems, Computer Science and Digital Economics. CSDEIS 2019, AISC 1127, *International Journal of Intelligent Systems and Applications (IJISA)*, 2020, pp. 154-164, doi: 10.1007/978-3-030-39216-1_15.

COMPUTATIONAL TOOLS FOR THE DNA TEXT COMPLEXITY ESTIMATES FOR MICROBIAL GENOMES STRUCTURE ANALYSIS

Mitina A.V.[1], Orlova N.G.[2], Dergilev A.I.[3,4], Orlov Y.L.[1,3,4]

[1] I.M. Sechenov First Moscow State Medical University of the Ministry of Health of Russia (Sechenov University)
Trubetskaya 8-2, Moscow, 119991, Russia; e-mail: alinamitina44@gmail.com
[2] Financial University under the Government of the Russian Federation
Leningradsky Ave, 49.2, Moscow, 125167, Russia
[3] Novosibirsk State University
Pirogova str., 1, Novosibirsk, 630090, Russia
[4] Institute of Cytology and Genetics SB RAS
Lavrentieva str., 10, Novosibirsk, 630090, Russia; e-mail: orlov@d-health.institute
Received 02.08.2023. DOI: 10.29039/rusjbpc.2023.0640

Abstract. One of the fundamental tasks in bioinformatics involves searching for repeats, which are statistically heterogeneous segments within DNA sequences and complete genomes of microorganisms. Theoretical approaches to analyzing the complexity of macromolecule sequences (DNA, RNA, and proteins) were established prior to the availability of complete genomic sequences. These approaches have experienced a resurgence due to the proliferation of mass parallel sequencing technologies and the exponential growth of accessible data. This article explores contemporary computer methods and existing programs designed to assess DNA text complexity as well as construct profiles of properties for analysing the genomic structures of microorganisms. The article offers a comprehensive overview of available online programs designed for detecting and visualising repeats within genetic text. Furthermore, the paper introduces a novel computer-based implementation of a method to evaluate the linguistic complexity of text and its compression using Lempel-Ziv. This approach aims to identify structural features and anomalies within the genomes of microorganisms. The article also provides examples of profiles generated through the analysis of text complexity. Application of these complexity estimates in the analysis of genome sequences, such as those of the SARS-CoV-2 coronavirus and the Mumps Orthorubulavirus, is discussed. Specific areas of low complexity within the genetic text have been successfully identified in this research.
***Key words**: bioinformatics, biophysical models, text complexity, microbial genomes.*