Вавиловский журнал генетики и селекции
Vavilov Journal of Genetics and Breeding. 2024;28(8):843-853

doi 10.18699/vjgb-24-92

Новый подход к анализу эволюции SARS-CoV-2, основанный на визуализации и кластеризации больших объемов генетических данных, компактно представленных в оперативной памяти

А.Ю. Пальянов [1, 2, 3], Н.В. Пальянова [2]

[1] Институт систем информатики им. А.П. Ершова Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Научно-исследовательский институт вирусологии, Федеральный исследовательский центр фундаментальной и трансляционной медицины, Новосибирск, Россия
[3] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
palyanov@iis.nsk.su

Аннотация. Коронавирус SARS-CoV-2 – это вирус, для которого было собрано, секвенировано и сохранено рекордное количество вариантов генома из источников по всему миру. Нуклеотидные последовательности в формате FASTA включают 16.8 млн геномов, каждый длиной ≈29 900 нт (нуклеотидов), общим размером $\approx 500 \cdot 10^9$ нт, или 466 Гб. Мы предлагаем способ представления данных, позволяющий разместить без потерь всю эту информацию в оперативной памяти (RAM) обычного персонального компьютера. Более того, будет достаточно всего ≈330 Мб. Выравнивание их всех относительно исходной референсной последовательности Wunah-Hu-1 позволяет представить каждый геном как структуру данных, содержащую списки точечных мутаций, делеций и вставок. Наша реализация такого представления данных привела к коэффициенту сжатия 1:1500 (для сравнения, упаковка данных с помощью популярного архиватора WinRAR дает степень сжатия только 1:62) и обеспечила возможность быстрого вычисления редакционного расстояния между различными вариантами генома. С помощью этого подхода, реализованного в виде программы на C++, мы провели анализ различных свойств набора геномов SARS-CoV-2, содержащихся в NCBI Genbank, собранных за 4.5 года (с 24.12.2019 по 24.06.2024). Были рассчитаны распределение числа геномов от числа неопределенных нуклеотидов "N" в них, число уникальных геномов и кластеров из идентичных геномов, а также распределение кластеров по размеру (числу идентичных геномов) и продолжительности (длине временного интервала между первым и последним геномом каждого кластера). Наконец, эволюция распределений числа изменений (редакционное расстояние между каждым геномом и референсной последовательностью), вызванных заменами, делециями и вставками, была визуализирована в виде 3D поверхностей, наглядно изображающих процесс вирусной эволюции в течение 4.5 лет, с интервалом в одну неделю. Такая визуализация хорошо соотносится с филогенетическими деревьями (обычно рассчитываемыми по 3–4 тыс. представителей вариантов генома), но строится на основе миллионов геномов, отображает больше деталей и не зависит от типа классификации линий/клад.

Ключевые слова: коронавирус; SARS-CoV-2; геном; варианты; эволюция; программная система; большие данные; компактизация; анализ; визуализация.

Для цитирования: Пальянов А.Ю., Пальянова Н.В. Новый подход к анализу эволюции SARS-CoV-2, основанный на визуализации и кластеризации больших объемов генетических данных, компактно представленных в оперативной памяти. *Вавиловский журнал генетики и селекции*. 2024;28(8):843-853. doi 10.18699/vjgb-24-92

Финансирование. Исследование выполнено за счет гранта Российского научного фонда (проект № 23-64-00005).

A novel approach to analyzing the evolution of SARS-CoV-2 based on visualization and clustering of large genetic data compactly represented in operative memory

A.Yu. Palyanov [1, 2, 3], N.V. Palyanova [2]

[1] A.P. Ershov Institute of Informatics Systems of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Research Institute of Virology, Federal Research Center of Fundamental and Translational Medicine, Novosibirsk, Russia
[3] Novosibirsk State University, Novosibirsk, Russia
palyanov@iis.nsk.su

Abstract. SARS-CoV-2 is a virus for which an outstanding number of genome variants were collected, sequenced and stored from sources all around the world. Raw data in FASTA format include 16.8 million genomes, each ≈29,900 nt (nucleotides), with a total size of $\approx 500 \cdot 10^9$ nt, or 465 Gb. We suggest an approach to data representation and organization,

with which all this can be stored losslessly in the operative memory (RAM) of a common PC. Moreover, just ≈330 Mb will be enough. Aligning all genomes versus the initial Wuhan-Hu-1 reference sequence allows each to be represented as a data structure containing lists of point mutations, deletions and insertions. Our implementation of such data representation resulted in a 1:1500 compression ratio (for comparison, compression of the same data with the popular WinRAR archiver gives only 1:62) and fast access to genomes (and their metadata) and comparisons between different genome variants. With this approach implemented as a C++ program, we performed an analysis of various properties of the set of SARS-CoV-2 genomes available in NCBI Genbank (within a period from 24.12.2019 to 24.06.2024). We calculated the distribution of the number of genomes with undetermined nucleotides, 'N's, vs the number of such nucleotides in them, the number of unique genomes and clusters of identical genomes, and the distribution of clusters by size (the number of identical genomes) and duration (the time interval between each cluster's first and last genome). Finally, the evolution of distributions of the number of changes (editing distance between each genome and reference sequence) caused by substitutions, deletions and insertions was visualized as 3D surfaces, which clearly show the process of viral evolution over 4.5 years, with a time step = 1 week. It is in good correspondence with phylogenetic trees (usually based on 3–4 thousand of genome variant representatives), but is built over millions of genomes, shows more details and is independent of the type of lineage/clade classification.

Key words: coronavirus; SARS-CoV-2; genome; variants; evolution; software system; big data; compact representation of data; analysis; visualization.

For citation: Palyanov A.Yu., Palyanova N.V. A novel approach to analyzing the evolution of SARS-CoV-2 based on visualization and clustering of large genetic data compactly represented in operative memory. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2024;28(8):843-853. doi 10.18699/vjgb-24-92

Введение

Коронавирус SARS-CoV-2 (самый первый образец которого, названный Wuhan/Hu-1/2019, был получен 24 декабря 2019 г.) (Wu et al., 2020) вызвал крупнейшую пандемию за последние 100 лет (со времен испанского гриппа 1918–1920 гг.). Спустя 4.5 года он по-прежнему продолжает свое существование, эволюционирует и выявляется у людей по всему миру, хотя и уже далеко не в тех объемах, которые были во время пика пандемии, и не с такими тяжелыми последствиями. Впрочем, как правило, с приходом осени показатели числа заражений снова возрастают, и 2024 г. не исключение. По данным Всемирной ассоциации здравоохранения (https://data.who.int/dashboards/covid19/cases, раздел "COVID-19 cases, country level trends"), к середине сентября 2024 г. во многих странах уже начался рост заболеваемости. Так, например, в России за июль 2024 г. было зарегистрировано 26.7 тыс. случаев заражения SARS-CoV-2, за август – 24.7 тыс., а за первую половину сентября – уже 62.2 тыс. В разных странах наблюдаются те или иные особенности динамики количества заражений, зависящие от множества факторов, анализом взаимосвязей между которыми, в частности, занимаемся и мы (Palyanova et al., 2022, 2023).

Получаемые по всему миру образцы вируса SARS-CoV-2 секвенируют и вносят в базы данных, крупнейшими из которых являются GISAID (gisaid.org) и NCBI Genbank (www.ncbi.nlm.nih.gov/sars-cov-2/): по состоянию на 06.2024 в них содержится более $16.7 \cdot 10^6$ и более $8.6 \cdot 10^6$ образцов геномов SARS-CoV-2 соответственно. Для сравнения, вирус гриппа человека, самые ранние образцы которого в GISAID датируются 1905 г., за более чем столетний период представлен в ней примерно $5.22 \cdot 10^5$ геномами. Типичный размер генома SARS-CoV-2 составляет 29.9 тыс. нт, поэтому полный объем геномов этого вируса, содержащихся в GISAID, составляет около $500 \cdot 10^9$ нт (или 465 Гб), а в Genbank – около $258 \cdot 10^9$ нт (241 Гб). В объем оперативной памяти среднего современного ПК (16…64 Гб) все эти данные одновременно не поместятся, тогда как работа с ними непосредственно из файлов, расположенных на жестком диске (HDD) или твердотельном накопителе (SSD), будет происходить существенно медленнее, чем из оперативной памяти (RAM). Скорости чтения данных с современных HDD/SDD/RAM имеют характерные значения порядка 0.2, 3 и 50 Гб/с соответственно, так что при значительных объемах данных и вычислительных нагрузках работа именно с оперативной памятью крайне желательна.

Несмотря на вакцинацию и медикаментозное лечение, в настоящее время не существует способа полностью устранить SARS-CoV-2 (Cui et al., 2023), так что, по-видимому, он надолго останется с человечеством, пополнив многочисленный перечень возбудителей ОРВИ, насчитывающий более 200 наименований, включая грипп, респираторно-синцитиальную, риновирусную, коронавирусную, аденовирусную и другие инфекции, вызывающие катаральные воспаления дыхательных путей.

Чем дольше существует вирус, тем больше изменений накапливается в его геноме, каждая новая генерация получается в результате репликации вирусов предыдущего поколения, в процессе которой возможно возникновение ошибок/изменений. В результате генных мутаций могут происходить замены, делеции и вставки одного или нескольких нуклеотидов, транслокации, дупликации и инверсии различных частей гена. Так, например, точечные мутации возникают самопроизвольно с частотами 10^{-8}–10^{-6} для ДНК-вирусов и 10^{-6}–10^{-4} – для РНК-вирусов (Sanjuán, Domingo-Calap, 2016), у которых собственная молекулярная машина для репликации (РНК-полимераза) лишена корректирующего ошибки механизма (экзонуклеазы). Исключение составляют коронавирусы и торовирусы, у которых она все-таки имеется (Campagnola et al., 2022), поскольку они обладают одними из наибольших для РНК-вирусов геномами, слишком быстрое накопление ошибок в которых, по-видимому, не является желательным и не способствует выживанию вируса.

Частота возникновения ошибок при репликации SARS-CoV-2 составляет, согласно (Amicone et al., 2022), $1.3 \cdot 10^{-6} \pm 0.2 \cdot 10^{-6}$ замен на позицию за один инфекцион-

ный цикл заражения клетки (т. е. от входа вируса в нее до выхода новых вирионов наружу). При этом скорость эволюционных изменений в геноме SARS-CoV-2 оценивается как $8.9 \cdot 10^{-4}$ замен на позицию в год (Sonnleitner et al., 2022).

Помимо упомянутых механизмов, которые могут воздействовать на отдельный (одиночный) геном, имеются и такие, благодаря которым могут возникать новые комбинации на основе генетического материала разных вариантов генома. Если два разных варианта одного и того же вируса заражают один и тот же организм одновременно (например, заражение штаммами SARS-CoV-2 «Дельта» и «Омикрон» (Bolze et al., 2022)), у них появляется возможность взаимодействовать во время репликации (Simon-Loriere, Holmes, 2011), в результате чего могут возникать рекомбинанты или реассортанты (для вирусов с сегментированным геномом).

Безотносительно того, какой именно механизм вызвал то или иное изменение, для любой пары геномов рассматриваемого вируса может быть рассчитано расстояние Левенштейна (также называемое редакционным расстоянием или дистанцией редактирования), определяемое как минимальное количество односимвольных операций (замены, делеции, вставки), которые нужно внести в первый геном, чтобы получить из него второй (или во второй, чтобы получить первый, – результат получится тот же самый). Другими словами, расстояние Левенштейна задает метрику, определяющую разность между двумя последовательностями символов. Таким образом, каждый вариант генома SARS-CoV-2 из имеющихся миллионов можно сравнить с исходным референсным геномом Wuhan-Hu-1. Для этого необходимо осуществить глобальное выравнивание всех последовательностей относительно референса, что было выполнено нами с помощью программы NextAlign/NextClade (https://github.com/nextstrain/nextclade) (Aksamentov et al., 2021). В результате для каждой последовательности был рассчитан перечень изменений (делеций, вставок или точечных замен), отличающих ее от референса.

Для вируса с размером генома в 30000 нт одна точечная замена может произойти в каждой из 30 тыс. позиций и привести к изменению имеющегося нуклеотида (А, Т, Г или Ц) на один из трех других, что порождает $30000 \cdot 3 = 90000$ различных вариантов. Одиночная вставка может быть сделана в 30001 позиции – добавлена как в начало или конец последовательности, так и в любой из 29999 промежутков между имеющимися нуклеотидами. Она может содержать любую из четырех букв алфавита, т. е. имеется 120004 различных варианта таких вставок. И наконец, делеция может произойти в любой из 30000 позиций, порождая число вариантов, равное числу позиций. Впрочем, делеции и вставки, оставляющие вирус жизнеспособным, чаще всего происходят блоками, размер которых кратен трем, поскольку иначе такое изменение привело бы к сдвигу рамки считывания, что в подавляющем большинстве случаев делает геном нежизнеспособным. Таким образом, даже одно одиночное изменение может быть осуществлено более чем 240 тыс. различных способов, хотя значительная их часть (особенно те, что соответствуют делециям и вставкам) сделает геном нежизнеспособным.

Комбинация из двух произвольных точечных замен – это уже $(240000)^2 = 5.8 \cdot 10^{10}$, а трех – $(240000)^3 = 1.4 \cdot 10^{16}$ вариантов, причем на этот раз среди них будут и те, у которых с рамкой считывания все будет в порядке (результат изменений – удаление или вставка одного триплета, т. е. трех нуклеотидов подряд). При этом число различий между некоторыми современными вариантами SARS-CoV-2 и референсным геномом уже превышает 200, а (для масштаба) редакционное расстояние между SARS-CoV-2 и ближайшим к нему геномом другого вируса – коронавируса летучих мышей, RaTG13, – составляет 1136 (совпадают 96.1 % нуклеотидов) (Zhou et al., 2020; Temmam et al., 2022). Подробнее ряд вопросов о пространстве вариантов генетических последовательностей SARS-CoV-2 рассмотрен в работе (Palyanov, Palyanova, 2023), в которой, в частности, показано, что количество уже реализованных вариантов вируса составляет ничтожно малую долю относительно тех, что являются потенциально возможными. Таким образом, как продолжение мониторинга новых вариантов SARS-CoV-2, так и анализ уже накопленных за прошедшие 4.5 года миллионов геномов представляют интерес как с практической точки зрения, так и для получения новых фундаментальных знаний в области вирусологии и эпидемиологии.

Материалы и методы

Приведенные в работе результаты получены с помощью программного комплекса, созданного нами для осуществления анализа эволюции вирусов. Для разработки использован язык программирования C++, среда разработки – Microsoft Visual Studio Community 2019. Задействован один сторонний программный модуль, необходимый для осуществления глобального выравнивания вирусных геномов, – NextAlign от NextClade (https://github.com/nextstrain/nextclade/releases). Аппаратное обеспечение – компьютер на базе процессора Intel Core i7-10700K, 3.8 ГГц, 8 ядер, 32 Гб оперативной памяти.

Данные для анализа – генетические последовательности SARS-CoV-2. Данные, использованные в настоящей работе, – это полный набор геномов SARS-CoV-2, из БД Genbank (www.ncbi.nlm.nih.gov/sars-cov-2/) за период с 24.12.2019 по 24.06.2024 (4.5 года с момента сбора первого образца этого вируса, Wuhan-Hu-1, 24.12.2019). Их количество составило 8 641 740 шт., объем – 242 Гб. Референсный геном SARS-CoV-2, имеющий длину 29903 нт, состоит из 5′ UTR длиной 265 нт, CDS (в котором закодированы 29 белков (Bai et al., 2022)) длиной 29409 нт и 3′ UTR длиной 229 нт (UTR – нетранслируемая область, CDS – кодирующая последовательность). Этот набор данных, который продолжает пополняться с течением времени, и есть тот фундамент, на котором строятся исследования в области эволюционных изменений SARS-CoV-2. Еще одним источником данных является БД GISAID (в которую, вероятно, данные из Genbank входят практически полностью), геномы из которой еще предстоит проанализировать и сравнить с результатами для геномов из Genbank.

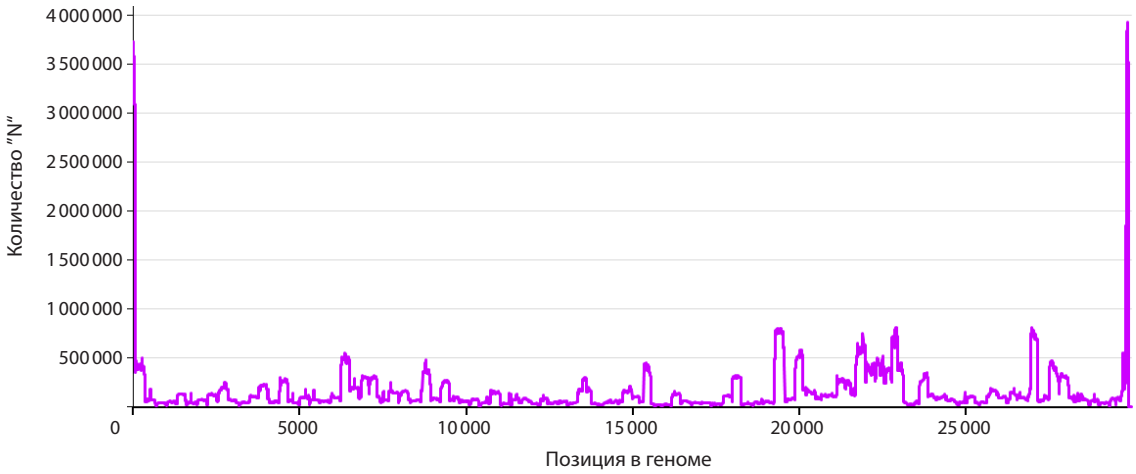

Рис. 1. Зависимость количества "N" от позиции в геноме (по оси абсцисс), полученная в результате суммирования по полному набору выровненных генетических последовательностей SARS-CoV-2 из Genbank в интервале с 24.12.2019 по 24.06.2024.

Качество данных, их предварительный анализ и фильтрация. Один из первых вопросов, возникающих при работе с набором нуклеотидных последовательностей вирусных геномов, – это их качество. В частности, в последовательностях могут присутствовать не только буквы, кодирующие нуклеотиды (A, T, G, C), но и "N", обозначающие неидентифицированные, неизвестные нуклеотиды в соответствующих позициях. Чем больше "N", тем больше неопределенности, и тем хуже для результатов анализа и их достоверности. В связи с этим, конечно, полезно представлять, насколько много таких последовательностей в исследуемом наборе данных, и насколько много "N" может встречаться в том или ином секвенированном геноме.

Расчеты показали, что из полного набора последовательностей (8641740) неидентифицированные нуклеотиды "N" встречаются в 6609933 геномах (76.5 %), а отсутствуют только в оставшихся 2031807 (23.5 %). Однако если рассматривать только CDS, то число геномов без "N" возрастает почти вдвое – до 3742117 (43.3 %). Помимо этого, мы построили зависимость частоты встречаемости "N" от позиции в геноме – на основе полного набора последовательностей, для которых было произведено глобальное выравнивание (рис. 1).

Как видно, имеются два наиболее значительных пика, в начале и в конце генома, соответствующие некодирующим участкам 5′ UTR и 3′ UTR, суммарная длина которых составляет 1.65 % от длины всего генома. Известно также, что в генетических последовательностях SARS-CoV-2 из GISAID и Genbank нетранслируемые области имеют высокий разброс по длинам 5′ UTR и 3′ UTR (Palyanov, Palyanova, 2023). С учетом того, что число геномов, в которых "N" встречаются в UTR и не встречаются в CDS, составляет 22.5 % от числа всех геномов, исключение UTR-участков из рассмотрения увеличит набор пригодных для анализа данных практически вдвое (к 23.5 % последовательностей, в которых "N" вообще не встречаются, ни в CDS, ни в UTR, добавятся еще 22.5 %, в которых "N" есть только в UTR).

В зависимости от того, каким является распределение геномов по числу "N", содержащихся в их CDS, можно либо использовать те из них, где всего несколько "N" (на фоне различий порядка 100 точечных замен это незначительная величина, хотя их наличие и вносит некоторую неопределенность), либо использовать только те геномы, в которых "N" в CDS отсутствуют. Построив упомянутое распределение (рис. 2), мы выяснили, что геномов с одной "N" – 1.8 %, с двумя и тремя – 0.8 и 0.9 % соответственно, а с "N" в пределах от 1 до 10 шт. на геном – 5.4 %. В результате на данном этапе было принято решение работать только с геномами, в которых "N" отсутствуют в CDS, и использовать в расчетах только CDS, исключив 5′ UTR и 3′ UTR.

Методы, алгоритмы и структуры данных. Для построения глобальных выравниваний всех геномов (с геномом Wuhan-Hu-1 в качестве референсной последовательности) мы использовали консольную версию NextAlign (работающую в многопоточном режиме), вызываемую с необходимыми параметрами из нашей программной системы. Это происходит при ее первом запуске или при необходимости пересчета выравниваний (например, в случае задействования другой выборки геномов). На полной выборке, упомянутой выше, состоящей из 8.6 млн геномов SARS-CoV-2, расчет выравниваний занимает около суток на рабочей станции с процессором Intel Core i7-10700K @ 3.8 ГГц (8 ядер, 16 потоков) и 32 Гб оперативной памяти (DDR4, 3600 ГГц). Результатом работы программы на данном этапе является сохранение на жестком диске в рабочей директории программы файлов с результатами всех рассчитанных выравниваний, которые затем используются нашей системой в качестве входных данных, взятых для анализа. Файлы представляют собой таблицы с несколькими десятками колонок, включающих различные характеристики геномов, метаданные, а также списки мутаций, делеций и вставок, отличающих рассматриваемый геном от референсного.

По мере считывания данных в оперативной памяти компьютера динамически формируется список структур,

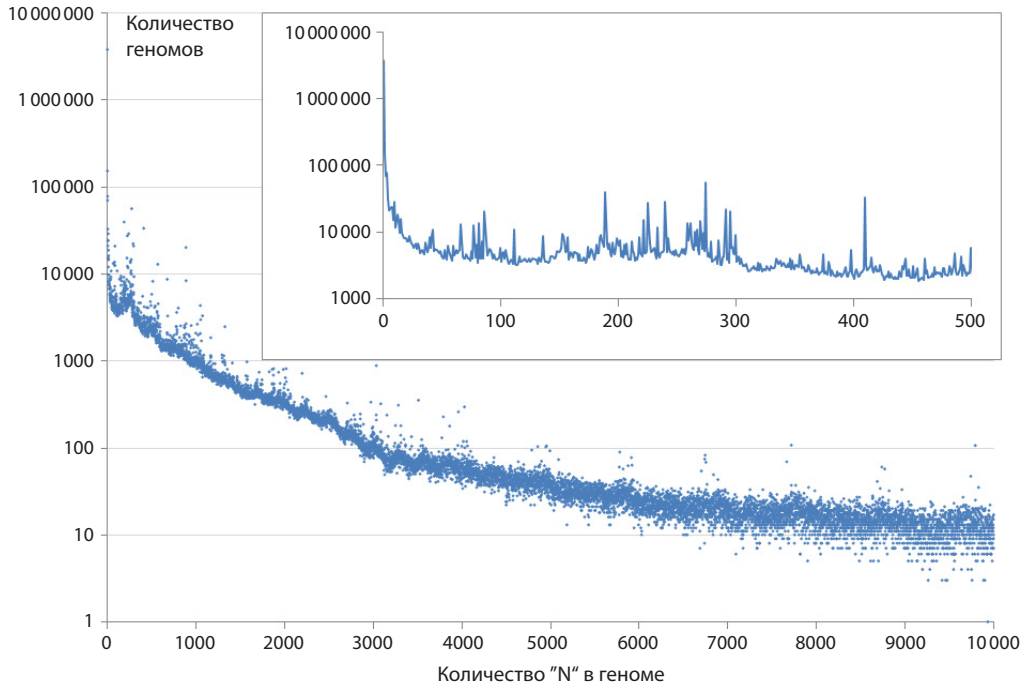

Рис. 2. Распределение геномов по количеству неидентифицированных нуклеотидов "N" в них.

Рассчитано по полному набору генетических последовательностей SARS-CoV-2 из Genbank в интервале с 24.12.2019 по 24.06.2024. Врезка показывает ту же зависимость, но с большим разрешением, для количества "N" в геноме в пределах от 0 до 500.

каждая из которых включает имя вируса, дату получения образца (collection date), географические данные, а также полный набор изменений, отличающих данный вариант от референсного генома:

- список точечных мутаций (однопозиционных замен), каждый элемент которого содержит номер позиции в геноме, соответствующий данной мутации, и букву, кодирующую нуклеотид, появившийся в данной позиции в результате замены (предыдущий нуклеотид, который был до мутации, не храним, при необходимости его всегда можно прочитать из соответствующей позиции референсного генома);
- список делеций, каждая из которых определяется двумя числами – позициями начала и конца делеции;
- список вставок, каждая из которых задается позицией в геноме, сразу после которой произошла вставка, а также вставленной последовательностью.

Такая организация данных позволяет легко и быстро сравнивать два произвольных генома. Особенно быстро получается определить, идентичны они или нет. Вместо сравнения каждой из 29 409 позиций первого и второго геномов достаточно просто сравнить число элементов в их списках точечных мутаций, делеций и вставок, хотя бы при одном отличии уже понятно, что геномы различаются. Впрочем, таким образом можно не только получить результат сравнения геномов, но и вычислить редакционное расстояние между ними. Совпадающие элементы списков различий не дают вклада в разницу между геномами, тогда как каждый элемент отличия от референса, присутствующий в одном геноме и отсутствующий в другом, добавляет соответствующее количество отличий. Каждая замена, произошедшая в одной и той же позиции обоих геномов, но приведшая к заменам на разные нуклеотиды, тоже, конечно, добавляет +1 к редакционному расстоянию. С учетом того что размеры списков небольшие, сравнение осуществляется значительно быстрее, чем сравнение двух геномов без предварительного выравнивания.

Предложенный нами способ компактного представления нуклеотидных последовательностей родственных геномов в оперативной памяти компьютера имеет много общего со способом сжатия, представляющим последовательности в виде филогенетического дерева с заменами на ребрах. Более того, само представление каждого генома как совокупности изменений, которые необходимо осуществить, чтобы перейти от референса к рассматриваемому геному, основано на тех же данных о структуре филогенетического дерева, выстраиваемого на базе множественного выравнивания рассматриваемых последовательностей.

Особенность нашей реализации состоит в том, что структура данных в оперативной памяти, представляющая множество рассматриваемых последовательностей, не является филогенетическим деревом, а вместо этого представляет собой список его «листьев», сортированный в хронологическом порядке, по дате получения образцов. Для таких задач, как анализ не просто имеющегося спектра вариантов вируса, а его эволюционных изменений с учетом времени их возникновения, наш подход обеспечивает значительное преимущество в скорости доступа к данным. Дело в том, что он позволяет двигаться по оси времени, просто увеличивая или уменьшая индекс элемента массива, состоящего из упорядоченных по времени геномов. В древовидном же представлении поиск всех геномов, соответствующих определенному году, месяцу

и дню, в общем случае сведится к обходу всего дерева, и так для каждого нового варианта. При этом каждый «лист» в нашем подходе содержит всю информацию о своей «ветви» дерева, что дает возможность легко и быстро вычислять редакционное расстояние для любой пары вариантов геномов.

Результаты

Кластерная структура множества геномов

В процессе исследования мы обратили внимание на то, что среди рассматриваемых геномов довольно часто оказываются такие, у которых CDS полностью, на 100 %, идентичны между собой, при том что дата сбора образца, географические данные и прочие метаданные чаще всего различаются. Добавив в нашу программную систему функцию выявления всех геномов с идентичными CDS (и объединения их в «кластеры»), мы разбили весь имеющийся у нас набор геномов на группы. Статистика по ним приведена в таблице.

Также мы рассчитали зависимость между размером кластера и числом кластеров того или иного размера (рис. 3). При этом между размерами кластеров и их временем жизни какой-либо явно выраженной зависимости не прослеживается, распределение представляет собой облако точек, большая часть которого сосредоточена в области от 1 до 1000 по оси «размер кластера» и от 1 до 500 – по оси «время существования кластера» (рис. 4).

Статистические данные о кластерах, объединяющих геномы с полностью одинаковыми CDS, включая размер и протяженность во времени

Полное число геномов SARS-CoV-2 из БД Genbank за период с 24.12.2019 по 24.06.2024	8 641 740
Число геномов из полного набора, в которых CDS не содержит неидентифицированных нуклеотидов, “N”	3 742 117
В выборке с CDS без “N” – число геномов с уникальной, больше нигде не встречающейся последовательностью нуклеотидов в CDS	1 690 699
В выборке с CDS без “N” – число геномов, образующих кластеры размером 2 и более	2 051 448
Число кластеров с размером ≥ 2	461 511
Число кластеров, для которых время существования (интервал между самой ранней и самой поздней датой сбора образца генома из числа входящих в кластер) составляет более одного дня	366 427
Максимальный размер кластера (число входящих в него геномов)	12 824
Максимальное время существования кластера (lineage 19A)	1539 дней
Средний размер кластера	4.4
Среднее время существования кластера	14.8 дня
Среднее время существования кластера (исключая те, у которых время существования один день)	18.4 дня

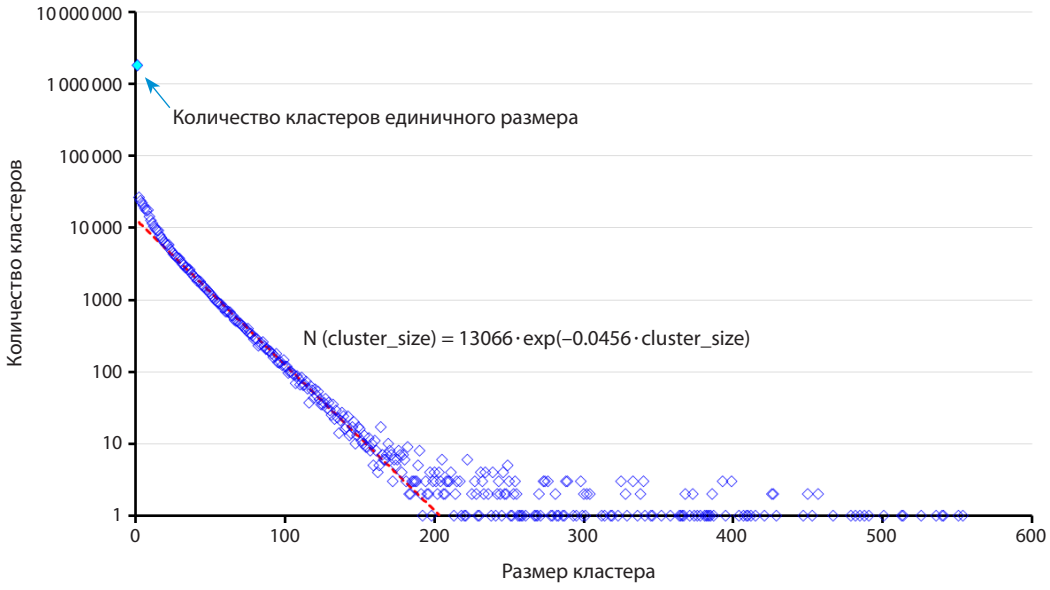

Рис. 3. Зависимость между числом кластеров и их размером для множества геномов SARS-CoV-2 из Genbank в интервале с 24.12.2019 по 24.06.2024.

В интервале значений размера кластера от 20 до 200 она хорошо аппроксимируется экспонентой с параметрами, указанными на рисунке.

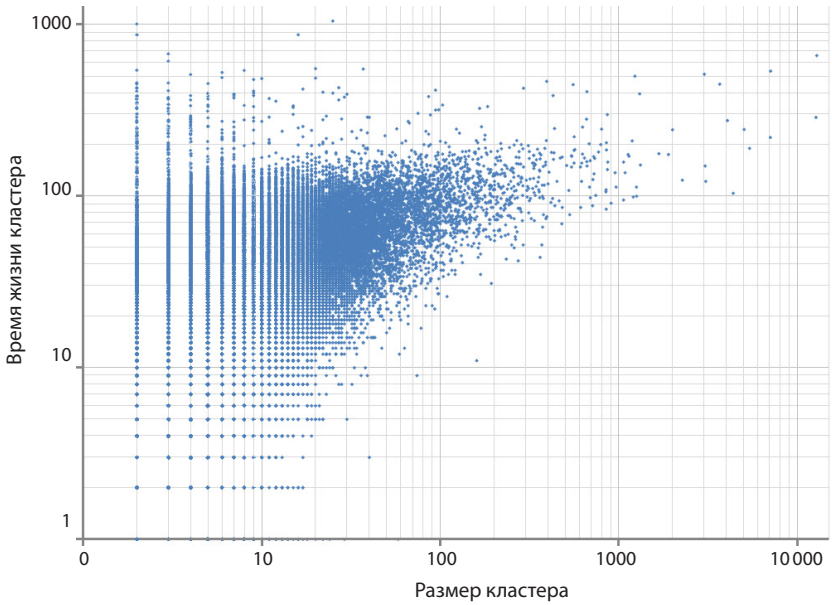

Рис. 4. Облако точек, представляющее множество геномов SARS-CoV-2 из Genbank (в интервале с 24.12.2019 по 24.06.2024) при использовании характеристик «размер кластера» и «время жизни кластера».

Мы отобразили все кластеры размером ≥ 200 на временной оси в виде линий, начало и конец которых соответствуют интервалам существования кластеров. Помимо этого, добавлены все кластеры размером 100–199, конец интервала существования у которых имеет значение ≥ 1000 дней с момента получения первого генома SARS-CoV-2. Это множество кластеров охватывает всю временную шкалу, хотя в интервале между 3 и 3.5 года явно есть и другие кластеры, но все они меньшего размера, чем представленные на рис. 5. Отдельная линия 19A, самая длинная на рис. 5, соответствует кластеру с самым долгим временем жизни (1539 дней или 4.2 года), упомянутому в таблице. В связи с этим генетическая линия 19A, просуществовавшая так долго, представляется довольно интересной. Данный вариант генома обнаруживался достаточно стабильно как в начале пандемии, так и в 2023–2024 гг.

Новый подход к визуализации эволюционных изменений SARS-CoV-2

Получив возможность быстро вычислять редакционное расстояние между парой любых вариантов нуклеотидных последовательностей, мы сделали это для всего множества геномов SARS-CoV-2 из Genbank за 4.5 года. Таким образом, для каждого генома имеется пара чисел – дата получения образца генома (collection date) и редакционное расстояние между ним и референсом. Бывает, что разные варианты геномов оказываются обладателями одинаковой пары значений «дата + редакционное расстояние», поскольку одно и то же значение редакционного расстояния может быть обеспечено как отличием, к примеру, на делецию длиной 30, вставку такой же

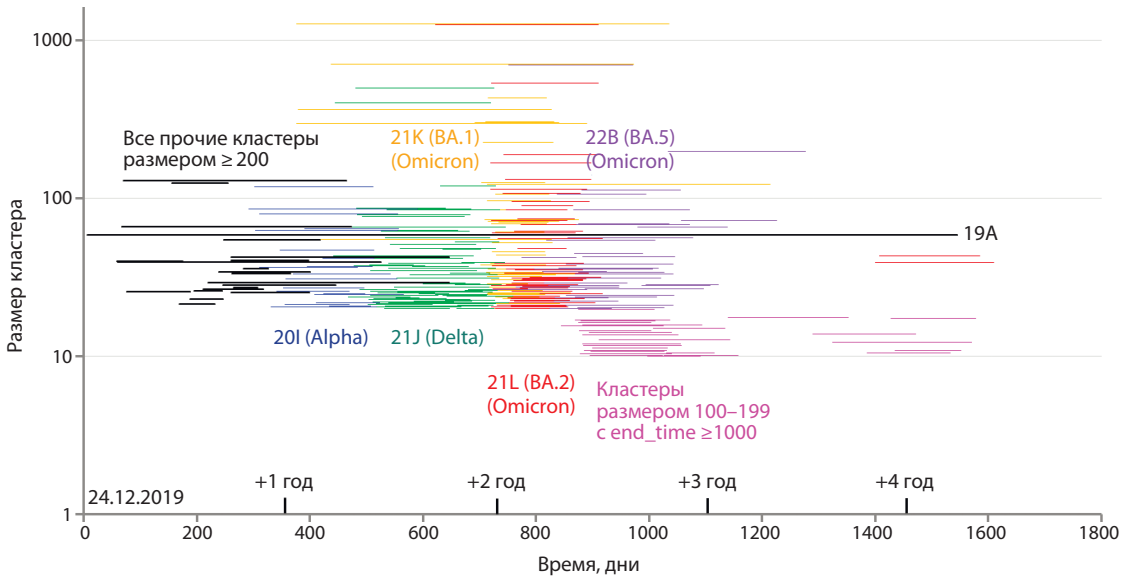

Рис. 5. Наиболее крупные кластеры (размером ≥ 200) и их интервалы существования (линии, соединяющие день первого появления последовательности нуклеотидов, представляющих данный кластер, и день, когда был взят последний образец).

Все кластеры размером ≥ 200, обозначенные одним и тем же цветом, имеют разные последовательности, но относятся к одной и той же генетической линии, название которой показано соответствующим цветом. Исключение составляют «все прочие кластеры размером ≥ 200», обозначенные черным цветом, которые представляют собой набор различных генетических линий (19A, 20A, 20B, 20C, 20E и 20F), а также кластеры размером 100–199, отмеченные пурпурным цветом (время завершения существования у которых составляет ≥ 1000 дней с даты получения первого генома SARS-CoV-2, 24.12.2019 г.).

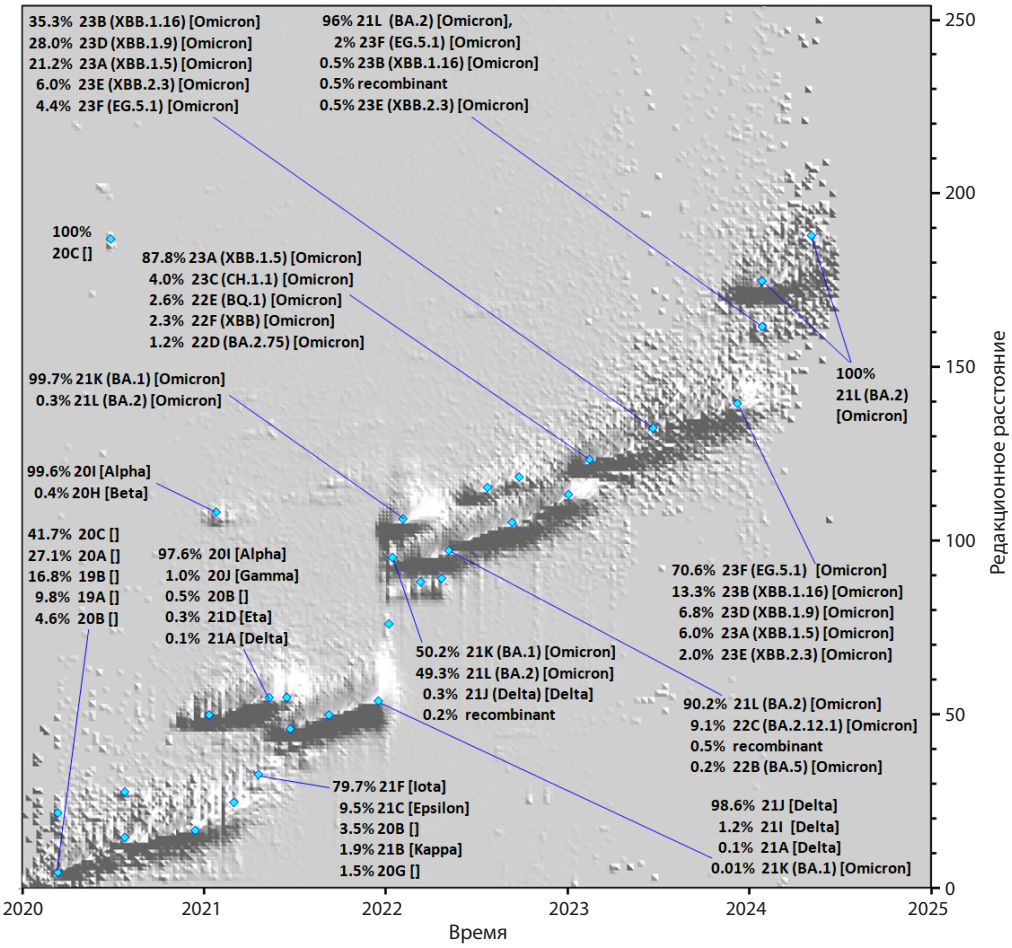

Рис. 6. Ландшафт пространства вариантов SARS-CoV-2, «посещенных» вирусом за период с 24.12.2019 до 24.06.2024, спроецированный на три оси: OX – время (с дискретизацией в 1 неделю), OY – различие (редакционное расстояние) между точкой на ландшафте и референсным геномом, OZ – доля геномов, соответствующих точке с определенными значениями X и Y, отнесенная к общему числу геномов, попавших в базу в неделю X. Такая нормировка необходима в связи с тем, что зависимость числа образцов геномов, собранных в ту или иную неделю по всему миру, имеет существенные изменения с течением времени, и без предложенной нормировки участки, для которых за день было собрано, к примеру, не 10 000, а всего 100 геномов, будут совершенно незаметны на рисунке, тогда как даже с этим числом геномов они вполне информативны.

длины, или же на 30 точечных мутаций, рассредоточенных по всему геному. Если ввести третью величину – количество случаев, когда вариант генома имеет определенное редакционное расстояние до референса и определенную дату получения образца, то можно рассчитать тройки этих величин на основе полного набора геномов SARS-CoV-2 и отобразить их в виде поверхности, что мы и сделали (рис. 6).

На рис. 6 мы отметили голубыми точками ряд интересных элементов ландшафта, для каждого из которых был рассчитан спектр соответствующих им вариантов. Для многих из них эту информацию удалось разместить на рисунке. На ландшафте видны области с различными особенностями – узкие протяженные «горные хребты», имеющие начало, конец и характерный угол наклона с близкими значениями для большинства их них, по-видимому, связанный со скоростью накопления изменений в геноме, появляющихся за счет точечных замен нуклеотидов.

Есть также области, в которых редакционное расстояние для всего множества вариантов, существующих в определенный момент времени, быстро и значительно изменяет среднее значение или испытывает ветвление, разделяясь на несколько параллельных, визуально различимых путей. Предположив, что такие резкие и значительные изменения могут быть связаны с делециями, вставками или рекомбинационными событиями, мы построили еще три рисунка, подобных рис. 6, для которых в качестве редакционного расстояния использовали не полную его величину, а три раздельных вклада – от совокупности точечных замен (рис. 7), делеций (рис. 8) и вставок (рис. 9). При этом из полного числа геномов (в CDS которых нет “N”), 3 742 117 шт., число геномов с мутациями относительно референса составило 3 741 518 шт., число геномов с делециями – 3 520 077 шт., а число геномов со вставками – 528 414 шт.

Как видно на рис. 7–9, особенности динамики эволюционных изменений, вносимых в геномы SARS-CoV-2

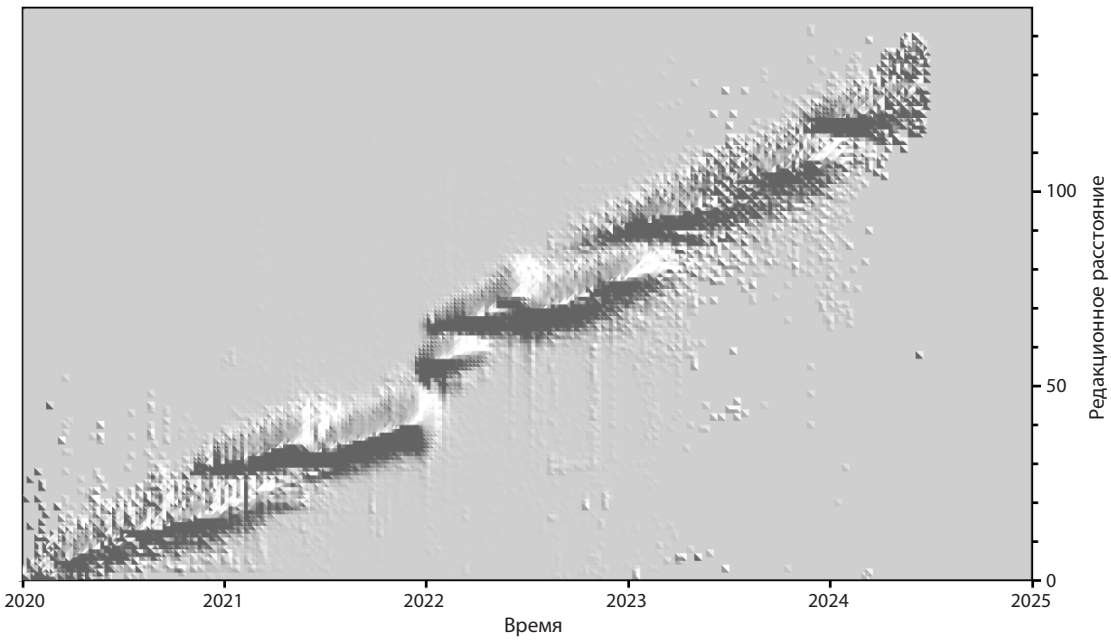

Рис. 7. Ландшафт эволюционных изменений на основе вклада только от точечных замен.

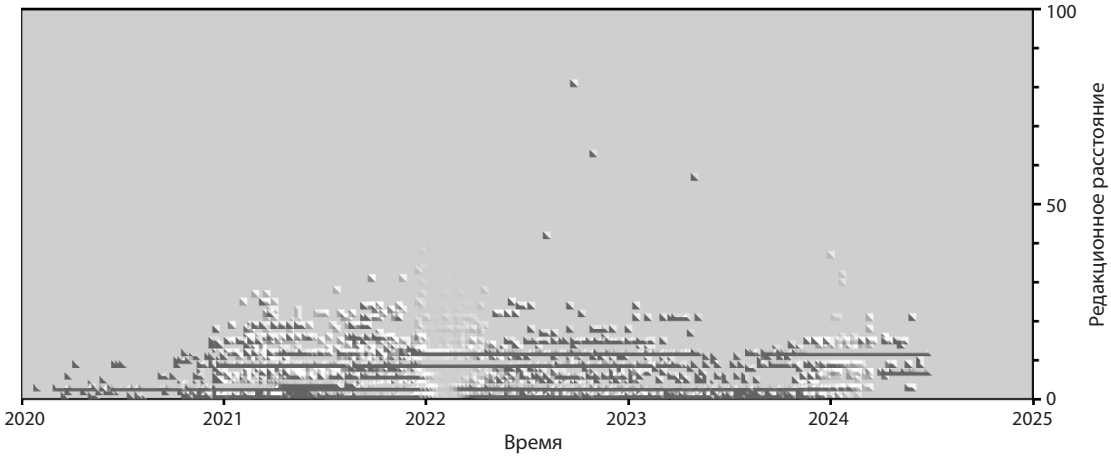

Рис. 8. Ландшафт эволюционных изменений на основе вклада только от вставок.

различными эволюционными механизмами, существенно различаются для точечных замен, делеций и вставок. Из рис. 7 можно заключить, что накопление количества мутаций на большом масштабе времени возрастает линейно (особенно если судить по верхней границе области эволюционного пути). За 4.5 года накопилось около 130 точечных мутаций, т. е. скорость накопления составила примерно 29 шт. в год (2.4 в месяц или 0.08 в день). Влияние делеций тоже ощутимо, и они тоже накапливаются по закону, близкому к линейному, но с меньшей скоростью – около 50 шт. (по совокупной длине) за 4.5 года, т. е. около 11 в год или чуть менее 1 в месяц. Вставки же, как видно из рис. 8, дают заметно меньший вклад, чем делеции, величина которого, за исключением первого года, практически не растет со временем, – он держится на уровне 20 шт. относительно референса (хотя содержание этих вставок, в принципе, может меняться на протяжении рассматриваемого интервала времени).

Обсуждение

Мы выполнили ряд оценок и расчетов, в основном с помощью разработанных нами программных средств, для улучшения понимания того, какие особенности и закономерности обнаруживаются у эволюции генетических последовательностей коронавируса SARS-CoV-2. Предложенный метод визуализации ландшафтов эволюционных изменений позволил наглядно увидеть множество деталей и особенностей, которые не видны, например, на филогенетическом древе. При этом быстрые изменения эволюционной траектории, сопровождающиеся ступенчатыми изменениями величины редакционного расстояния, как видно из рис. 6, как правило, сопровождаются сменой доминирующего в популяции варианта вируса. Так, например, для одного из таких «скачков» на эволюционном ландшафте удалось зарегистрировать генетические линии «Йота», «Эпсилон» и «Каппа» (примерно в первой четверти 2022 г.).

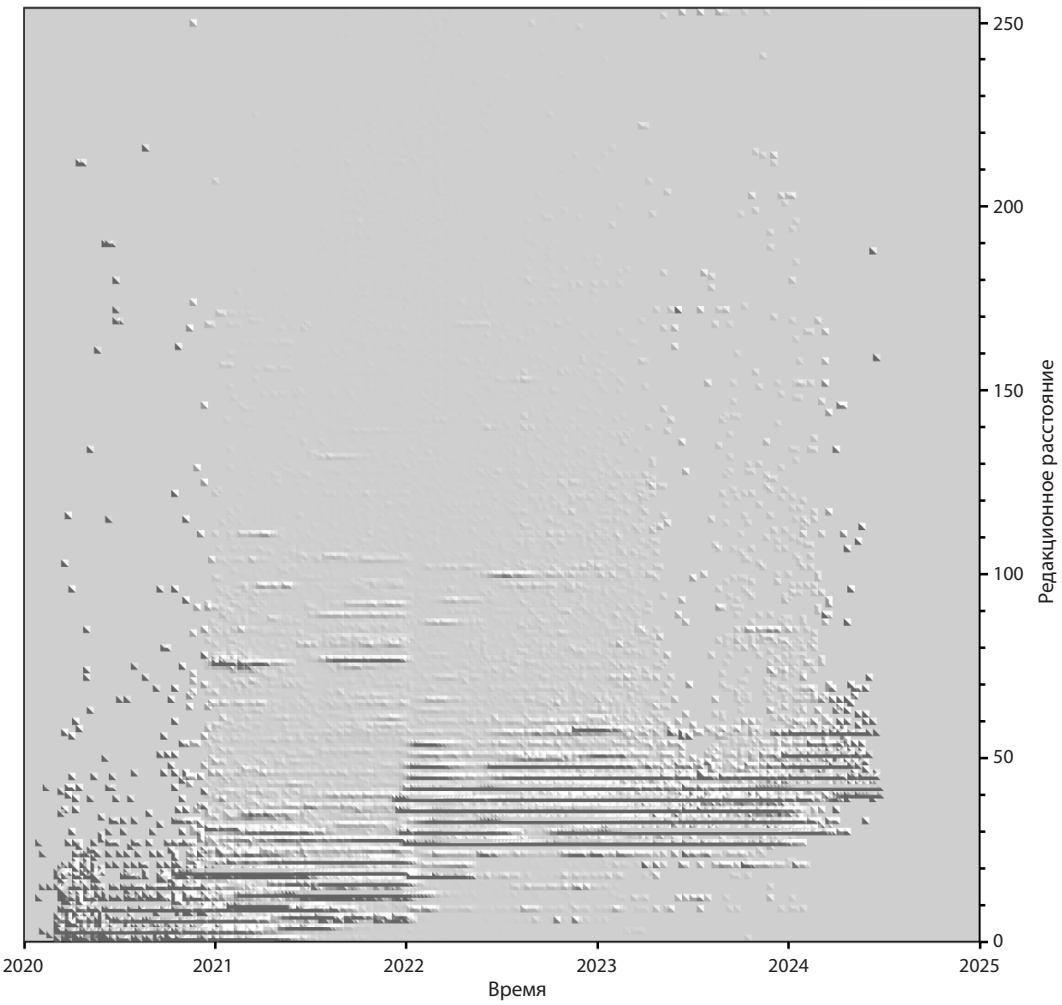

Рис. 9. Ландшафт эволюционных изменений на основе вклада только от делеций.

Наблюдаемое отсутствие роста величины вклада в редакционное расстояние, обеспечиваемого вставками, ранее упомянутое для результатов, связанных с рис. 8, может быть обусловлено тем, что слишком большое число вставок способно нарушать стабильность вируса. Увеличение числа вставок увеличивает физический размер генома и тем самым может ухудшать его способность помещаться внутрь белковой оболочки, которая, предположительно, рассчитана на нахождение внутри нее объекта вполне определенного объема. Таким образом, в ходе эволюции число вставок относительно референсного генома может расти, но не превышать 20–30 нт. Если какие-то из новых вариантов оказываются более приспособленными, чем их предшественники, то очень скоро они могут их вытеснить. Видно, что только в начале 2024 г. исчезают варианты, в которых вообще нет вставок, вероятно, в связи с тем, что за четыре года эволюции все-таки нашлись такие вставки, которые оказались заметно более приспособленными, чем варианты вовсе без вставок, и закрепились в популяции. Также из рис. 8 и 9 понятно, что большинство делеций и вставок имеют длину, кратную 3, что имеет вполне очевидное объяснение, связанное с тем, что иные варианты длин будут приводить к сдвигу рамки считывания при синтезе закодированных в геноме белков и в большинстве случаев – к нежизнеспособным экземплярам геномов, вирионы которых не могут быть собраны.

Заключение

В результате проделанной работы получены следующие основные результаты:

- предложен и программно реализован способ представления нуклеотидных последовательностей геномов вирусов, обеспечивающий исключительно компактное размещение их в оперативной памяти компьютера. На примере коронавируса SARS-CoV-2 показано, что обеспечивается сжатие в 1500 раз. Использование его для пересылки генетических данных по сети могло бы снизить нагрузку на серверы и снизить сетевой трафик в соответствующее число раз (особенно при передаче больших массивов данных);
- для полного набора геномов SARS-CoV-2 (без “N” в CDS) исследовано наличие кластеров полностью идентичных геномов. Выяснено, что их размер может превышать 10 000 шт., а интервал времени, в пределах которого они встречаются, может охватывать до нескольких сотен дней и более;

- предложен новый способ отображения эволюционной динамики вирусов в виде ландшафта, визуализирующего проекцию пространства вариантов геномов вируса на три оси – время (T), редакционное расстояние до референсного генома (D) и долю геномов (P) в каждой точке (T, D) в полном количестве геномов, соответствующих данному T;
- показано, что ландшафт, построенный для D (вычисляемого как сумма вкладов от точечных мутаций, делеций и вставок), можно разделить на три отдельных ландшафта, рассчитываемых отдельно для каждого из вкладов. Каждый из них имеет различный характер, позволяющий оценить вклад и влияние каждого из упомянутых механизмов на эволюцию вируса. Рассчитаны константы, характеризующие каждый из механизмов и скорость приобретаемых благодаря ему изменений;
- обнаруженный нами факт того, что линия 19A существовала наиболее долго по сравнению с остальными кластерами, охватывая весь период пандемии, позволяет нам предложить создавать новые вакцины против SARS-CoV-2 на основе именно этой линии как сохраняющей наибольшую конкурентоспособность по сравнению с остальными вариантами, а значит, содержащей наиболее характерные для данного вируса особенности, которые могут быть распознаны иммунной системой.

В будущем мы планируем исследовать возможности предложенного метода визуализации эволюции более детально, однако уже сейчас можно сказать, что он представляется полезным, имеет потенциал для дальнейшего использования и развития и может быть применен не только к SARS-CoV-2, но и к другим вирусам. Это же можно сказать о предложенном методе компактного представления вирусных геномов, применение которого во всех областях, связанных с хранением, передачей по сети, обработкой и анализом большого количества вариантов родственных геномов (как вирусов, так и живых организмов), обеспечит значительные преимущества.

Список литературы / References

Aksamentov I., Roemer C., Hodcroft B., Neher R.A. Nextclade: clade assignment, mutation calling and quality control for viral genomes. *J. Open Source Software*. 2021;6(67):3773. doi 10.21105/joss.03773

Amicone M., Borges V., Alves M.J., Isidro J., Zé-Zé L., Duarte S., Vieira L., Guiomar R., Gomes J.P., Gordo I. Mutation rate of SARS-CoV-2 and emergence of mutators during experimental evolution. *Evol. Med. Public Health.* 2022;10(1):142-155. doi 10.1093/emph/eoac010

Bai C., Zhong Q., Gao G.F. Overview of SARS-CoV-2 genome-encoded proteins. *Sci. China Life Sci*. 2022;65(2):280-294. doi 10.1007/s11427-021-1964-4

Bolze A., Basler T., White S., Rossi A.D., Wyman D., Dai H., Roychoudhury P., Greninger A.L., Hayashibara K., Beatty M., Shah S., Stous S., McCrone J.T., Kil E., Cassens T., Tsan K., Nguyen J., Ramirez J., Carter S., Cirulli E.T., Barrett K.S., Washington N.L., Belda-Ferre P., Jacobs S., Sandoval E., Becker D., Lu J.T., Isaksson M., Lee W., Luo S. Evidence for SARS-CoV-2 Delta and Omicron co-infections and recombination. *Med.* 2022;3(12):848-859. doi 10.1016/j.medj.2022.10.002

Campagnola G., Govindarajan V., Pelletier A., Canard B., Peersen O.B. The SARS-CoV-2 nsp12 polymerase active site is tuned for large-genome replication. *J. Virol.* 2022;96(16):e0067122. doi 10.1128/jvi.00671-22

Cui X., Wang Y., Zhai J., Xue M., Zheng C., Yu L. Future trajectory of SARS-CoV-2: Constant spillover back and forth between humans and animals. *Virus Res.* 2023;328:199075. doi 10.1016/j.virusres.2023.199075

Palyanov A.Yu., Palyanova N.V. On the space of SARS-CoV-2 genetic sequence variants. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2023;27(7):839-850. doi 10.18699/VJGB-23-97

Palyanova N.V., Sobolev I.A., Alekseev A., Glushenko A., Kazachkova E., Markhaev A., Kononova Y., Gulyaeva M., Adamenko L., Kurskaya O., Bi Y., Xin Y., Sharshov K., Shestopalov A. Genomic and epidemiological features of COVID-19 in the Novosibirsk region during the beginning of the pandemic. *Viruses*. 2022;14(9):2036. doi 10.3390/v14092036

Palyanova N.V., Sobolev I.A., Palyanov A.Yu., Kurskaya O.G., Komissarov A.B., Danilenko D.M., Fadeev A.V., Shestopalov A.M. The development of the SARS-CoV-2 epidemic in different regions of Siberia in the 2020–2022 period. *Viruses*. 2023;15(10):2014. doi 10.3390/v15102014

Sanjuán R., Domingo-Calap P. Mechanisms of viral mutation. *Cell. Mol. Life Sci*. 2016;73(23):4433-4448. doi 10.1007/s00018-016-2299-6

Simon-Loriere E., Holmes E.C. Why do RNA viruses recombine? *Nat. Rev. Microbiol.* 2011;9(8):617-626. doi 10.1038/nrmicro2614

Sonnleitner S.T., Prelog M., Sonnleitner S., Hinterbichler E., Halbfurter H., Kopecky D.B.C., Almanzar G., Koblmüller S., Sturmbauer C., Feist L., Horres R., Posch W., Walde G. Cumulative SARS-CoV-2 mutations and corresponding changes in immunity in an immunocompromised patient indicate viral evolution within the host. *Nat. Commun.* 2022;13(1):2560. doi 10.1038/s41467-022-30163-4

Temmam S., Vongphayloth K., Baquero E., Munier S., Bonomi M., Regnault B., Douangboubpha B., Karami Y., Chrétien D., Sanamxay D., Xayaphet V., Paphaphanh P., Lacoste V., Somlor S., Lakeomany K., Phommavanh N., Pérot P., Dehan O., Amara F., Donati F., Bigot T., Nilges M., Rey F.A., van der Werf S., Brey P.T., Eloit M. Bat coronaviruses related to SARS-CoV-2 and infectious for human cells. *Nature.* 2022;604(7905):330-336. doi 10.1038/s41586-022-04532-4

Wu F., Zhao S., Yu B., Chen Y.M., Wang W., Song Z.-G., Hu Y., Tao Z.-W., Tian J.-H., Pei Y.-Y., Yuan M.-L., Zhang Y.-L., Dai F.-H., Liu Y., Wang Q.-M., Zheng J.-J., Xu L., Holmes E.C., Zhang Y.-Z. A new coronavirus associated with human respiratory disease in China. *Nature*. 2020;579(7798):265-269. doi 10.1038/s41586-020-2008-3

Zhou P., Yang X.L., Wang X.G., Hu B., Zhang L., Zhang W., Si H.-R., Zhu Y., Li B., Huang C.-L., Chen H.-D., Chen J., Luo Y., Guo H., Jiang R.-D., Liu M.-Q., Chen Y., Shen X.-R., Wang X., Zheng X.-S., Zhao K., Chen Q.-J., Deng F., Liu L.-L., Shi Z.-L. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*. 2020;579(7798):270-273. doi 10.1038/s41586-020-2012-7

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Поступила в редакцию 15.09.2024. После доработки 23.10.2024. Принята к публикации 24.10.2024.