

Проект эффективной программной платформы для работы с генетическими данными респираторных вирусов*

А. В. Мордвинов¹, А. В. Стучинский^{1,II}, А. П. Девятериков¹, С. С. Хайрулин¹,
Н. В. Пальянова^{III}, А. Ю. Пальянов^{I,III,III}

¹Институт систем информатики им. А.П. Ершова Сибирского отделения Российской академии наук, Новосибирск, Россия

^{II}Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия

^{III}Научно-исследовательский институт вирусологии Федерального исследовательского центра фундаментальной и трансляционной медицины, Новосибирск, Россия

Аннотация. Статья посвящена разработке отечественной веб-платформы с необходимыми возможностями получения доступа к банкам генетической информации. Основная цель – предоставить исследователям возможность эффективно решать задачи биоинформатики, вирусологии и эпидемиологии, при необходимости расширяя набор доступных на сервере программ для анализа и моделирования. Проект основан на современных, эффективных, обоснованно выбранных программных решениях, обеспечивающих высокую производительность и предоставляющих множество полезных функциональных возможностей. Реализуемая веб-платформа позволяет загружать, хранить, искать и анализировать геномные последовательности вирусов, таких как грипп и SARS-CoV-2, а в перспективе и другие вирусные патогены. Планируется сделать ее доступной для исследователей и периодически обновлять из открытых источников, чтобы повысить удобство и эффективность работы ученых, ведущих исследования в соответствующих областях науки.

Ключевые слова: веб-платформа, биоинформатика, вирусология, эпидемиология, хранение данных, анализ данных, алгоритмы, эффективность.

DOI 10.14357/20718632240406 EDN EUCRNU

Введение

Вирусы и живые организмы непрерывно эволюционируют – со времён появления первой жизни на Земле до наших дней. В основе эволюции – передача генетической информации от предков к потомкам и возникновение изменений в геноме в процессе его репликации за счет ошибок, которые могут при этом возникать время от времени (точечные замены нуклеотидов, делеции, вставки и другие изменения). Большинство

ошибок приводит к появлению особей с приспособленностью хуже, чем у их предков, вплоть до неспособности выживать и оставлять потомство, но иногда результат может быть и иным – удачные изменения тоже случаются, и если они дают их обладателю ощутимые преимущества, то закрепляются в популяции за счет роста числа потомков – носителей таких изменений. Вирусы эволюционируют особенно быстро по сравнению с живыми организмами, поэтому изучение и отслеживание этих процессов для

* Работа выполнена при поддержке Российского научного фонда (проект № 23-64-00005).

тех из них, которые оказывают значительное влияние на жизнь большого числа людей, представляет собой важное направление в вирусологии и эпидемиологии. Таковыми являются респираторные вирусы – возбудители ОРВИ, перечень которых включает более 200 наименований, включая грипп, респираторно-синцитиальную, риновирусную, коронавирусную, аденовирусную и другие инфекции, вызывающие катаральные воспаления дыхательных путей.

Современная наука обладает широким спектром инструментов для анализа вирусов и изменений, происходящих в их геноме в процессе эволюции. Из образца вируса с помощью специального оборудования, секвенатора (от англ. sequence – последовательность), можно извлечь и прочесть (секвенировать) содержащуюся в нем последовательность нуклеотидов ДНК или РНК, а затем сравнить ее с имеющимися образцами, полученными ранее. Это происходит на регулярной основе во множестве диагностических лабораторий по всему миру, особенно теперь, после недавней пандемии COVID-19, вызванной вирусом SARS-CoV-2, и накопленного опыта противодействия этой угрозе. Вирусы состоят из белковой оболочки, внутри которой находится их геном – одна или более цепочек РНК или ДНК – нуклеотидные последовательности, в которых закодированы белковые молекулы, необходимые вирусу для функционирования в организме и для сборки новых вирусных частиц следующего поколения. Эти нуклеотидные последовательности могут быть сохранены в памяти компьютера или на его жестком диске в виде текстовых файлов (как правило, в формате FASTA), содержащих последовательность букв, кодирующих нуклеотиды – А (аденин), G (гуанин), С (цитозин) и Т (тимин). Также в файле указываются метаданные этого экземпляра вируса, которые могут включать дату получения образца, географические данные, данные пациента-носителя и др.

Отслеживание эволюционных изменений распространенных респираторных вирусов позволяет оперативно выявлять новые, ранее не существовавшие варианты вирусных геномов, представляющие потенциальную угрозу, чтобы в случае опасности как можно скорее принять необходимые меры для противодействия ей.

Получаемые при этом данные, годами накапливающиеся день за днем, их необходимо хранить и предоставлять к ним доступ для специалистов, заинтересованных в работе с ними.

1. Платформа GISAID

Лидером в области накопления, хранения и предоставления онлайн-доступа к вирусным геномам является платформа GISAID (gisaid.org, Global Initiative on Sharing All Influenza Data – Глобальная инициатива по предоставлению доступа ко всем данным по вирусу гриппа) – проект, инициированный в 2006-м году [1] и начавший функционировать в 2008-м (в Германии). На момент создания основной целью проекта являлось хранение и распространение генетических данных по птичьему гриппу (avian flu). Однако пандемия гриппа H1N1 в 2009 г. [2], эпидемия гриппа H7N9 в 2013-м [3] и пандемия SARS-CoV-2 в 2019-2023 [4; 5] привели к тому, что уже существующее, предложенное и реализованное GISAID решение было расширено и на эти вирусы. К настоящему времени в регулярно пополняемой базе данных GISAID хранятся вирусы гриппа, коронавирусы, респираторно-синцитиальные вирусы (RSV), оспа обезьян, лихорадка Денге, Зика, Западного Нила, чикунгунья и желтая лихорадка.

Масштабы объемов данных, хранящихся в GISAID, можно проиллюстрировать на примере SARS-CoV-2: для этого вируса к октябрю 2024 г. накоплено уже 17 миллионов геномов, каждый – размером около 30 кб, т.е. суммарно – более 500 Гб. Загружать последовательности в GISAID может любой зарегистрированный пользователь, однако для всех новых данных проводится тщательная проверка. Помимо веб-интерфейса для поиска последовательностей в базах данных и их загрузки на веб-сайте GISAID также имеются встроенные инструменты для работы с ними – BLAST для поиска последовательностей, наиболее близких к образцу, по которому осуществляется поиск; построение филогенетических деревьев; визуализация географических данных о представленных в базе образцах вирусов; расчет и визуализация структуры спайк-гликопротеинов разных вариантов SARS-CoV-2 и многое другое. Кроме того

разработчики регулярно публикуют результаты анализа поступившей к ним информации и статистические данные по ним.

Хотя у пользователя и имеется возможность скачивать геномы из базы данных GISAID, размер одной зачатки ограничен 10000 последовательностей, из-за чего скачивание и анализ действительно больших массивов данных оказывается просто невозможным ни для кого, кроме пользователей с максимальным уровнем доступа к системе. Кроме того, программное обеспечение, обеспечивающее работу веб-компонентов платформы GISAID и базы данных для хранения генетической информации, основано на технологиях 2006–2008 г., что негативно сказывается на быстродействии и функциональных возможностях.

В 2023 году GISAID подверглась серьезной критике, приведшей даже к такой форме выражения мнения по важным для научного сообщества вопросам, как публикация в журнале Nature [6] с заголовком «GISAID in crisis: can the controversial COVID genome database survive?» («GISAID в кризисе: сможет ли выжить спорная база данных геномов ковида?»). Одна из ключевых причин критики связана с проблемой доступности содержащихся в GISAID данных для мирового научного сообщества: «Создатели GISAID обещали, что последовательности, представленные в ней, будут депонированы в полностью открытых хранилищах данных, таких как GenBank, «с максимальной задержкой в шесть месяцев». Однако это обещание так и не было выполнено. В GISAID отсутствует механизм переноса данных в базы данных с открытым доступом и это, по мнению критиков, один из его главных недостатков. В своем ответе журналу Nature GISAID заявил, что он уже доступен для общественности и поэтому не планирует предоставлять механизм для передачи данных. Теоретически исследователи могут загрузить свои последовательности в открытую базу данных после депонирования их в GISAID, но это означает двойную работу. На практике данные оказываются в ловушке жестких ограничений на неопределенный срок. Это побудило группу ученых в 2021 году написать открытое письмо, призывающее других ученых публиковать свои данные о геноме SARS-CoV-2 на платформах с полностью открытым доступом» [7].

Впрочем, несмотря на критику и недовольство некоторыми аспектами GISAID, многие считают, что реальной альтернативы ей нет, а польза имеется, и весьма ощутимая.

2. Платформа NCBI Genbank

Еще один мировой лидер в области хранения генетических данных – NCBI Genbank (<https://www.ncbi.nlm.nih.gov/genbank/>) [8] – база данных Национального центра биотехнологической информации в США, основанного в 1988 г. В отличие от GISAID, открытая политика Genbank предполагает, что любой пользователь может скачать оттуда данные и распоряжаться ими на свое усмотрение. Genbank хранит геномы известных науке живых организмов, от одноклеточных до человека, однако и вирусные геномы в нем также представлены. Размер генома SARS-CoV-2 составляет около 29,9 кб, тогда как у Homo sapiens – 3,1 Гб, а количество разных человеческих геномов в Genbank на данный момент составляет 1582 шт., т.е. суммарный объем данных только по человеку – 4,9 Тб, а всего в Genbank, по данным на январь 2024 г., «содержится 3,7 миллиардов нуклеотидных последовательностей для 557000 формально описанных биологических видов, суммарный объем которых составляет 25 триллионов пар оснований» [9], т.е. 25 Тб. Однако геномов SARS-CoV-2 в Genbank содержится почти вдвое меньше, чем в GISAID – 8,9 миллионов против 17. Для некоторых стран соотношение загруженных геномов в GISAID к загруженным в Genbank оказывается существенно большим, превышая значение 11 тысяч (Табл. 1).

Для проведения крупномасштабных статистических, эволюционных и иных расчетов, требующих как можно более полного набора данных, единственный выход – сочетать данные из обеих баз, исключив дубликаты (не просто по последовательностям, но и с учетом метаданных – даты получения образца, географических данных, данных о носителе, у которого был взят образец и т.д.). Однако поскольку данные GISAID не являются общедоступными, а загруженных последовательностей из многих стран там содержится в сотни, тысячи и даже, в одном случае, в 11,5 тыс. раз больше, чем в Genbank, необходимо формирование и использование

Табл. 1. Выборка из стран мира, суммарно загрузивших в Genbank 98,7% всех содержащихся там нуклеотидных последовательностей вируса SARS-CoV-2. Последовательности из этих же стран в GISAID составляют 82,4%. Порядок расположения стран в таблице – по последней колонке (отношению числа геномов, загруженных в GISAID, к числу загруженных в Genbank)

Страна	GISAID	Genbank	GISAID/Genbank
Бахрейн	11523	13490	0.85
Великобритания	3170661	3421024	0.93
Исландия	24096	25459	0.95
Швейцария	168717	158912	1.06
Германия	952683	803328	1.19
Дания	678881	496689	1.37
США	5214156	3725665	1.40
Аргентина	27449	12584	2.18
Новая Зеландия	50061	16636	3.01
Египет	5146	1198	4.30
Монголия	3071	565	5.44
Мексика	147993	25846	5.73
Южная Африка	56963	9611	5.93
Франция	681798	80708	8.45
Китай	86974	7572	11.49
Казахстан	4401	378	11.64
Бразилия	258842	17892	14.47
Австралия	239214	13378	17.88
Япония	689106	18394	37.46
Россия	89223	1435	62.18
Перу	56712	883	64.23
Индия	352951	4826	73.14
Италия	206163	1889	109.14
Турция	103429	207	499.66
Канада	653213	249	2623.35
Норвегия	80450	7	11492.86
Суммарно (по таблице)	14013876	8858825	1.58
Всего (по всему миру)	16997465	8966561	1.90

совокупной выборки, включающей все данные из Genbank, дополненные выборками по отдельным странам из GISAID для тех из них, которые существенны но недопредставлены в Genbank.

3. Об альтернативах и перспективах

Зарубежные базы данных наполняются всеми исследователями мира, что позволяет им вести учет и анализ последовательностей на мировом уровне. В России пока нет базы вирусных геномов, которая наполнялась бы мировым сообществом и была доступна широкому кругу российских исследователей, в то время как в нашей стране есть ряд довольно сильных групп вирусологов, которые могли бы внести существенный

вклад в анализ вирусных геномов, если бы имели к ним доступ.

В настоящее время существует довольно обширный программный инструментарий, используемый учеными для решения задач в области вирусологии, эпидемиологии и биоинформатики. Случается, однако, что нужные для решения той или иной конкретной задачи программы находятся на веб-серверах, принадлежащих разным организациям из разных стран, в результате чего работа существенно замедляется и осложняется необходимостью обмена данными между ними, который приходится осуществлять “вручную”, скачивая результаты работы одной программы и затем загружая их в следующую. Иногда речь идет о весьма значительных объемах,

исчисляемых десятками гигабайт. Нужные для расчетов программы в отдельных случаях тоже можно скачать и производить все расчеты на локальной машине. А как поступить в случае, если необходимо обеспечить доступ к данным и программам - с удобным интерфейсом и возможностью обмена данными - для целого коллектива ученых? Причем с возможностью, при необходимости, добавлять новые программы, включая собственные разработки, а также настраивать и дополнять возможности веб-интерфейса?

Целью проекта, которому посвящена данная статья, является создание отечественной веб-платформы, ориентированной на задачи биоинформатики, вирусологии и эпидемиологии, основанной на современных, эффективных, основанно выбранных программных решениях, обеспечивающих высокую скорость работы веб-интерфейса и веб-сервера в целом, предоставляющего множество полезных функциональных возможностей. Благодаря сотрудничеству между Научно-исследовательским институтом вирусологии ФИЦ ФТМ и Институтом систем информатики им. А.П. Ершова СО РАН и совместному проекту, поддержанному грантом РНФ, имеется достаточно уникальная в своем роде возможность сделать по-настоящему удобный, эффективный и гибко настраиваемый под нужды ученых многофункциональный инструмент для оперирования генетическими данными распространенных респираторных вирусов для их изучения и противодействия создаваемых ими угроз. Базовый набор возможностей должен включать быстрый доступ к содержимому базы данных и эффективный поиск по ней, возможность загружать собственные генетические данные, а также возможность осуществлять широкий спектр научных расчетов на основе этих данных, включая расчет белковых структур и взаимодействия между ними. Также актуальными задачами, решаемыми при исследовании геномов, являются расчеты их выравниваний (локальных или глобальных, парных или множественных), поиск по геномам, расчет филогенетических деревьев и многое другое, включая разработку новых программных решений для анализа и моделирования.

Таким образом, мы предлагаем и уже реализуем проект, направленный на создание эффективной веб-платформы для загрузки,

хранения, поиска и анализа геномных последовательностей вирусов на примере гриппа и других патогенов, вызывающих наиболее распространенные респираторные заболевания. В работе участвует большая команда, включающая как группу исследователей, работающих непосредственно с вирусами и секвенирующих вирусные геномы, так и биоинформатиков, осуществляющих сборку геномов и анализирующих полученные данные. В связи с этим весьма актуальной представляется задача обеспечения возможности хранения и предоставления быстрого доступа к геномам, полученным нашей командой, как минимум внутри коллектива. Также необходима возможность анализировать эти данные в мировом контексте. Имеющийся функционал в международных и российских базах не позволяет эффективно решать наши задачи, в связи с чем было принято решение начать разработку собственной веб-платформы. Мы также хотим сделать ее доступной для российских исследователей и периодически обновлять из открытых источников, чтобы повысить удобство и эффективность работы российских ученых, работающих в соответствующих областях науки.

4. Концепция и функциональная структура веб-платформы bioprojects.iis.nsk.su

Ядро проекта представляет собой клиент-серверное web-приложение. Для реализации был выбран TypeScript – язык программирования, представленный Microsoft в 2012 году, позиционируемый как средство разработки веб-приложений, расширяющее возможности JavaScript. Использование TypeScript позволяет явно определить ожидаемые типы данных, что улучшает понимание кода и предотвращает ошибки на этапе разработки. TypeScript может быть использован как общий язык для серверной и клиентской части приложения. Благодаря этому удалось создать проект в виде монорепозитария с тремя выделенными модулями (`yarn workspaces`, <https://classic.yarnpkg.com/lang/en/docs/workspaces/>) – frontend, backend и commons. Монорепозитарий является стратегией разработки программного обеспечения, когда код множества подпроектов хранится в одном и том же репозитории.

Среди преимуществ подхода на основе монорепозитория называют следующие:

- упрощение совместной работы и совместного использования кода;
- единая сборка, упрощенное управление зависимостями (поскольку все зависимые компоненты находятся в одной и той же базе кода);
- легкость рефакторинга кода, атомарность коммитов (изменений в программный код);
- простота повторного использования кода: схожая функциональность или коммуникационные протоколы могут быть выделены в общие библиотеки и напрямую подключаться к проектам без необходимости в менеджере пакетов для подключения зависимостей.

Фронтенд (англ. frontend - “передний торец”, “лицевая сторона”) – пользовательский интерфейс и связанные с ним компоненты; применяется в соотношении с базисной частью системы, ее внутренней реализацией, называемой в этом случае бэкендом (англ. backend – “задняя часть”). Разделение программных систем на фронтенд и бэкенд – одно из стандартных решений для архитектуры программного обеспечения, связанное в программной инженерии с принципом разделения ответственности между внешним представлением и внутренней реализацией. В commons собрана дублирующаяся логика (валидация, типизация, DTO, утилиты), которая может быть использована/используется как на пользовательском интерфейсе, так и на web-сервере. Переиспользование логики ускоряет разработку и повышает надежность, в результате чего тестирование существенно упрощается для переиспользуемого кода.

API (сокр. англ. Application Program Interface - интерфейс программирования приложения) – описание способов взаимодействия одной программы с другими. Реализован с помощью библиотеки trpc (<https://trpc.io>), представляющей собой полностью типобезопасный (англ. typesafe – концепция, в которой язык программирования препятствует или предотвращает возможности возникновения ошибок типов) TypeScript-клиент для запросов на сервер. Запросы генерируется автоматически, что исключает потенциальные ошибки в параметрах запросов (в англоязычной литературе это называется end-to-end typesafe API).

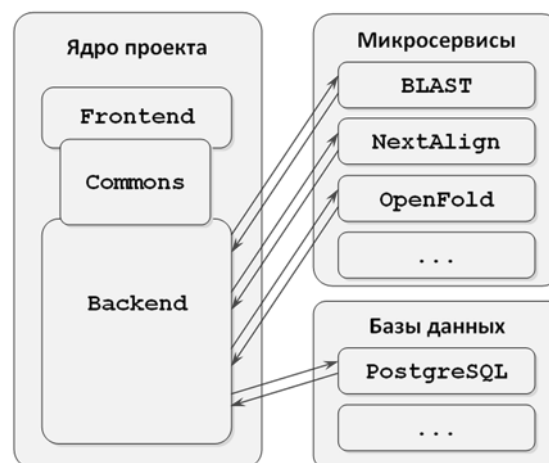


Рис. 1. Структура проекта – основные функциональные блоки, их компоненты и взаимодействие между ними

Интерфейс пользователя реализован на актуальном на сегодняшний день стеке технологий – React+Tailwind CSS (<https://tailwindcss.com>). Tailwind CSS представляет собой CSS-фреймворк с открытым исходным кодом, созданный Адамом Уэтенем с коллегами в 2017 г. и поддерживаемый Tailwind Labs, предназначенный для упрощения верстки, скорости разработки и исключения максимально возможного числа ошибок. Особенность этой библиотеки в том, что вместо предопределения CSS-классов отдельных элементов она предоставляет служебные классы, которые можно объединять для стилизации каждого элемента. В состав платформы запланировано включить как стандартные утилиты и алгоритмы биоинформатики, так и добавлять новые перспективные программные разработки по мере их появления. Каждый из таких компонентов выделяется в отдельный микросервис, запросы к которому управляет серверная часть приложения (Рис. 1).

5. Веб-интерфейс для поиска в базе данных геномов по пользовательским запросам

Минимально необходимый набор программных инструментов для работы с нуклеотидными последовательностями должен включать возможность поиска по различным полям метаданных, связанных с образцами вирусных геномов. Например, выбрать все последовательности по

их географической принадлежности (страна, область, город) или по интервалу дат получения образцов, а также по комбинации одновременно примененных фильтров. Более продвинутой функциональностью предполагает возможность составления выборок вариантов вирусов по возрасту и полу пациента (носителя), по тяжести протекания болезни и ее исходу (хотя такие данные имеются далеко не всегда), по качеству нуклеотидной последовательности (бывает, что геном не полный или в нем имеются неидентифицированные нуклеотиды, обозначаемых буквой N в дополнение к четырем буквам нуклеотидного алфавита – A, T, G, C). Также востребована возможность выбрать из базы все геномы, в которых имеются те или иные точечные мутации, делеции, вставки или иные изменения относительно исходной, референсной последовательности (для SARS-CoV-2 обычно это Wuhan-Hu-1/2019, полученная 24.12.2019 г. [10]). Помимо этого, весьма удобными представляются функция сохранения выборки (в виде массива уникальных идентификаторов образцов геномов) в личном кабинете пользователя, функция объединения нескольких выборок в одну, добавление новых последовательностей, исключение из выборки указанных пользователем последовательностей и т.д. Еще одна возможность из числа продвинутых опций – исключение дубликатов (для которых совпадает и геном, и метаданные), которые иногда возникают. Например, новые последовательности были загружены одновременно в GISAID и Genbank, а позже GISAID, загрузив и добавив к своей базе новые последовательности из Genbank, создает дубликаты, не осуществляя соответствующую проверку.

Вышеописанные возможности мы планируем реализовывать по мере развития проекта. Хорошая реализация, помимо прочего, должна быть еще и быстрой. Помимо составления выборки образцов геномов, и GISAID, и Genbank позволяют эту выборку скачать в виде FASTA-файла. К примеру, при необходимости скачать из GISAID выборку всех образцов SARS-CoV-2, полученных в России, указываем в поле Location значение Russia и получаем 90091 шт. Для скачивания нужно сначала их все выбрать, поставить галочку

в квадратике (checkbox), отмеченном красной желтой стрелкой (Рис. 2).

Нажав упомянутый checkbox, наблюдаем систему в состоянии ожидания, и это продолжается от 4.3 до 17.1 секунд для данной конкретной выборки (среднее значение – 5.5). Осуществив данное действие на нескольких различных компьютерах и браузерах, мы выяснили, что разница средних значений незначительна, однако во всех сериях запусков наблюдается довольно значительный разброс значений времени ожидания.

Если же осуществить фильтрацию по тому же полю, но выбрать Индию, то получим значительно большую выборку – 352986 образцов. Нажатие галочки “выбрать все” для этой выборки приведет к ожиданию в пределах от 51 до 67 секунд. Осуществив измерения также для Франции (682004 шт.), Великобритании (3170661 шт.) и США (5215132 шт.), получим следующую зависимость (Рис. 3).

Для выборки Location: USA время ожидания составляет более 7 минут(!). И все это время на экране можно наблюдать надпись Selecting 5,215,132 submissions..., а вовсе не индикатор прогресса, который показал бы пользователю, сколь долгим будет ожидание и каков прогресс на данный момент.

Можно было бы продолжать довольствоваться тем, что есть, полагая, что таковы ограничения современных технологий работы с данными. К счастью, имеется контрпример – реализация этой же функции в Genbank. Для этой платформы та же операция “выбрать весь набор образцов, получившийся в результате фильтрации”, как оказалось, осуществляется практически мгновенно, за малую долю секунды, и это время не зависит от количества выбираемых последовательностей – проверено на тех же данных. Это – результат правильного выбора программного обеспечения для хранения данных и доступа к ним. Выбранная нами бесплатная объектно-реляционная СУБД (система управления базами данных) с открытым исходным кодом PostgreSQL (<https://www.postgresql.org>) позволит работать с данными так же быстро, как и в Genbank, база данных в основе которого тоже является реляционной [11]. Какая именно база данных используется в GISAID, остается неизвестным.

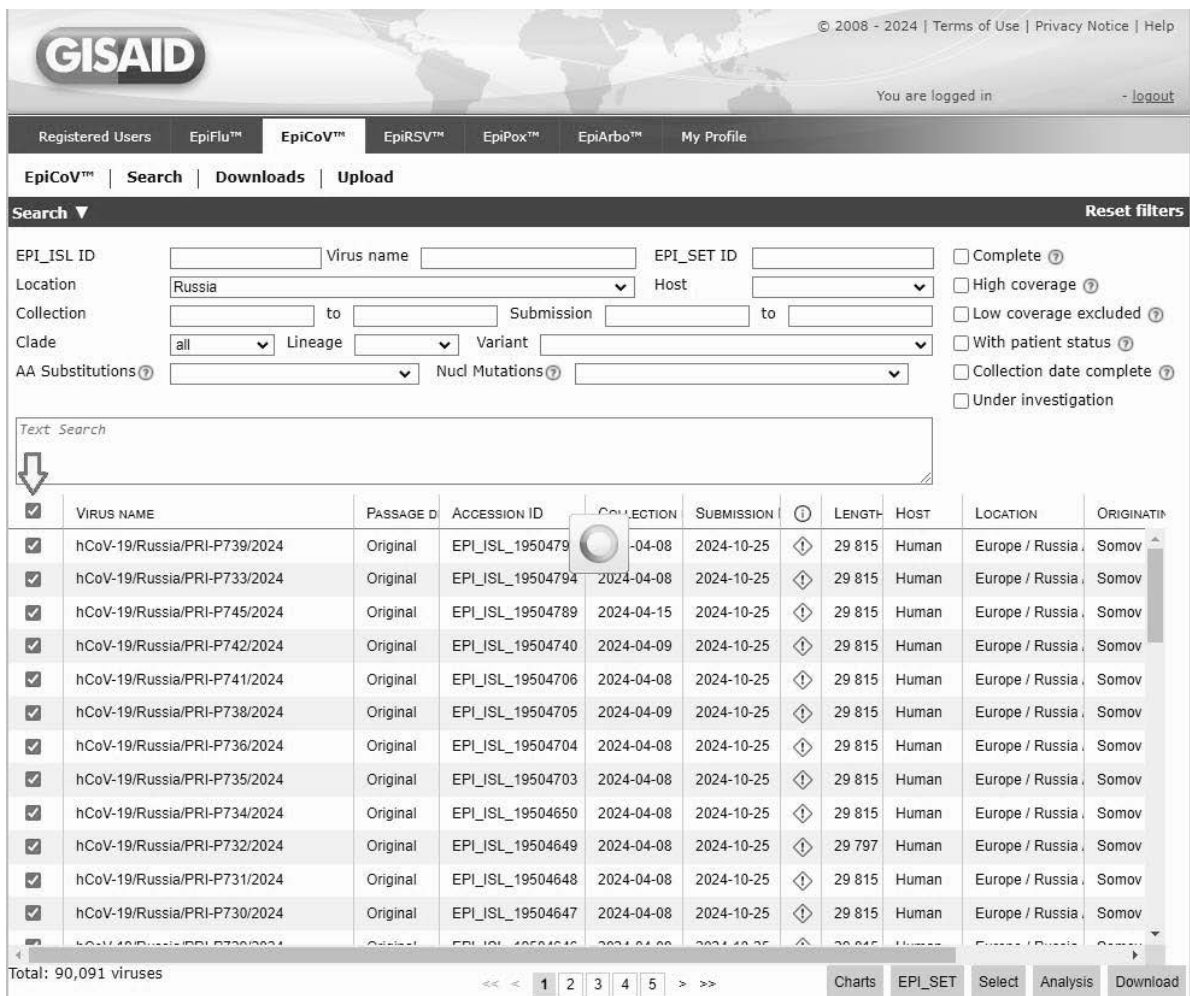


Рис. 2. Скриншот, иллюстрирующий внешний вид веб-интерфейса GISAID для поиска последовательностей в базе данных SARS-CoV-2. Результат фильтрации по полю Location: Russia и нажатия галочки “выбрать все” для дальнейшего скачивания или анализа полученной выборки

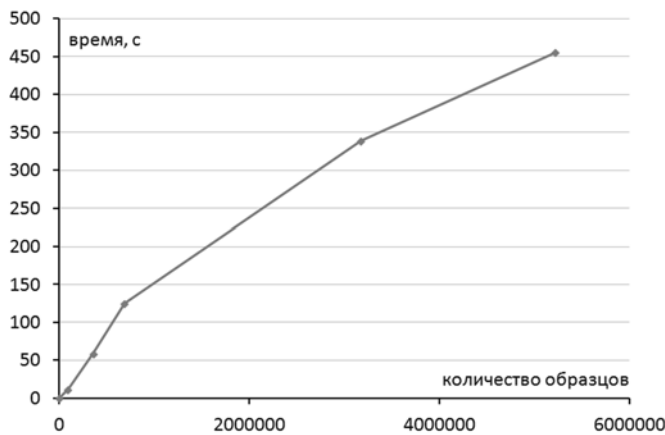


Рис. 3. Время (в секундах), необходимое для выбора всех геномов из пользовательской выборки (для осуществления дальнейших действий с ними) в зависимости от размера этой выборки

После выделения выборки достаточно нажать “скачать” и начнется загрузка (предварительно можно выбрать набор полей метаданных, которые будут сохранены в заголовках FASTA файлов). Если попробовать сделать то же самое в GISAID, то, во-первых, мы не сможем скачать ни одну из выборок, поскольку размер каждой из них превышает 10000. Чтобы сделать выборку подходящего размера, придется подобрать, например, интервал времени, допустим, от первого появления SARS-CoV-2 в конце 2019 г. до 26.05.2021 г, и тогда получим 9967 образцов из России в пределах этого интервала. Нажав Download, получаем диалоговое окно с возможностью выбора, что именно скачать для выбранных последовательностей – “даты и локаций”, “метаданные о статусе пациента”, “сами последовательности в формате FASTA” и еще несколько вариантов. Возможности скачать одновременно последовательности и метаданные для них, как в Genbank, не предлагается. На следующем диалоговом окне необходимо согласиться с условиями использования скачиваемых последовательностей, и только тогда начнется загрузка. После этого придется повторить процесс уже для метаданных, поскольку в некоторых исследованиях они не менее важны, чем сами последовательности. Затем, вручную подобрать еще один интервал времени для следующих 10000 последовательностей, и так пока не будет скачан весь необходимый объем данных. Чтобы скачать всю базу, необходимо 17 миллионов зачек в ручную (с затратами времени – не менее 10 минут на каждую), что, очевидно, за разумное время сделать совершенно нереально, тогда как скачать весь набор данных по SARS-CoV-2 из Genbank можно за сутки. В этом и заключается та самая упомянутая выше основная претензия к GISAID, состоящая в том, что формально к любой последовательности в их базе данных доступ имеется, но получить всю выборку на практике возможности нет, тогда как для некоторых исследований это необходимо.

6. Возможные роли и уровни доступа к системе

Свободный доступ без регистрации в системе не позволяет сохранять пользовательские настройки, создавать коллекции, хранить

результаты расчетов в личном кабинете. С другой стороны, слишком ограниченный доступ оставляет мало возможностей. Для удобства и эффективности работы мы запланировали следующие уровни доступа к нашей системе:

Системные администраторы. Могут вносить любые изменения в любые базы данных. Для страховки (для защиты от случайных или ошибочных необратимых изменений), предыдущие данные должны помещаться в корзину и храниться там на протяжении месяца.

Исследователи. Имеют доступ ко всем данным без ограничений, могут загружать данные, могут вносить изменения (только в те данные, которые ранее загрузили сами), могут пользоваться коллекциями и встроенными программами для анализа данных.

Зарегистрированные пользователи. Могут пользоваться поиском и видеть содержимое базы данных генетических последовательностей, использовать BLAST, выполнять расчет выравниваний и запуск других программ для анализа, если это не повлечет ощутимой потери производительности при большом количестве одновременно работающих с системой пользователей.

Пользователи без регистрации (Гости). Могут пользоваться поиском и видеть только базу генетических последовательностей.

7. Программный инструментарий для работы с нуклеотидными последовательностями

Помимо проблемы наполнения БД нуклеотидными последовательностями необходимы также и программные инструменты, предоставляющие различные возможности для их анализа, от выравнивания, поиска близкородственных вариантов и построения филогенетических деревьев до расчета трехмерных структур вирусных белков и собственных прикладных программ. Эти возможности могут быть обеспечены следующими широко известными, признанными в биоинформатике программами с открытым исходным кодом, которые можно встроить в состав веб-сервера:

1) BLAST (сокр. англ. Basic Local Alignment Search Tool – средство поиска основного локального выравнивания) – семейство компьютерных программ для поиска сходных аминокислотных

или нуклеотидных последовательностей [12]. Используя BLAST, можно сравнить изучаемый образец с последовательностями из базы данных и найти предполагаемые гомологи. Является важнейшим инструментом для молекулярных биологов, биоинформатиков и систематиков.

2) NextAlign – алгоритм выравнивания последовательностей вирусных геномов, используемый в Nextclade [13]. Он осуществляет попарное локальное выравнивание поданных на вход последовательностей относительно заданного референсного генома.

3) Программы поиска (выявления) рекомбинантных вирусных последовательностей. Появление рекомбинантных вирусных геномов представляет собой довольно редкое, но при этом значимое событие для анализа эволюции вирусов. Оно получается в результате обмена частями геномов двух разных вариантов вирусов одного вида (что возможно при заражении одновременно обоими вариантами), в результате чего может возникнуть новый, ранее не существовавший

вариант генома с новыми свойствами или особенностями.

4) OpenFold [14] – система глубокого обучения, выполняющая предсказания пространственной структуры белков по их аминокислотным последовательностям. OpenFold является близким аналогом AlphaFold2 компании DeepMind, реализованным на основе PyTorch, с открытым исходным кодом. Одно из самых явных различий между AlphaFold2 и OpenFold заключается в том, что OpenFold поддается обучению.

Все перечисленные программы входят в список запланированных для встраивания в систему и в настоящее время находятся на разных стадиях готовности (Рис. 4).

Внешний вид веб-интерфейса в режиме поиска нуклеотидных последовательностей в базе данных по совокупности примененных фильтров и в режиме поиска по нуклеотидной последовательности (с использованием программы BLAST) в нашей реализации выглядят следующим образом (Рис. 5. и Рис. 6, соответственно):

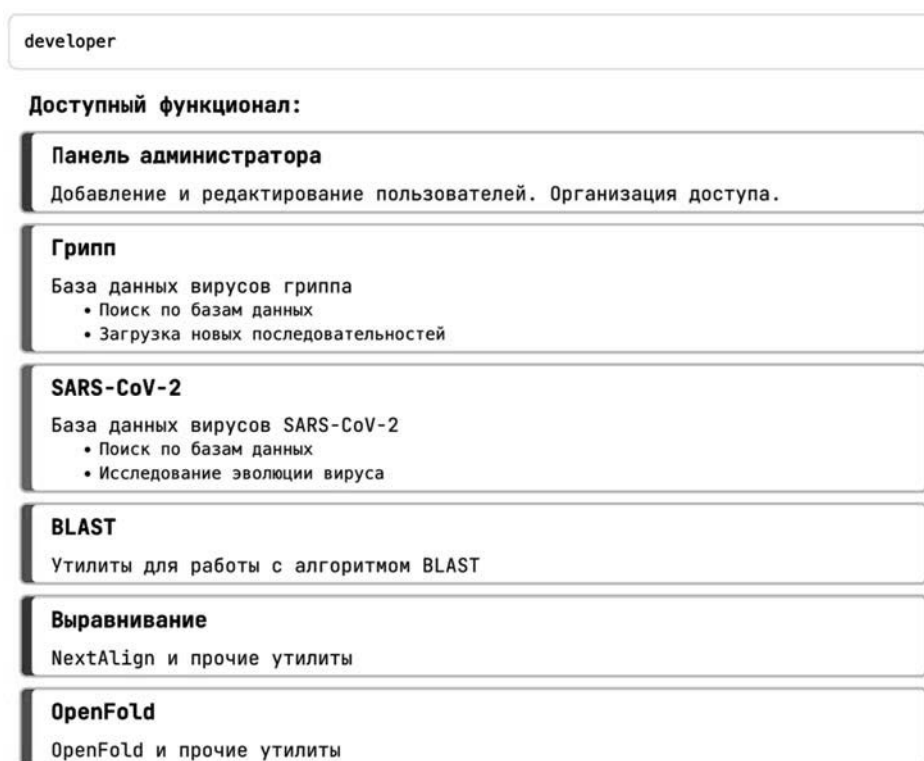


Рис. 4. Скриншот основного экрана, появляющегося при начале работы с системой, предоставляющего перечень доступных возможностей и предлагающего воспользоваться ими. Возможность загружать новые последовательности SARS-CoV-2 находится в процессе реализации и будет доступна в последующих версиях

Войти

Вирус гриппа

Имя вируса

Стрейн

Носитель

Пол

Локация

Дата забора

Найти

1

2

3

4

5

...

22

>

50 / стр.

Имя вируса	Стрейн	Носитель	Возраст	Пол	Локация	Дата забора
Influenza A virus (A/duck/Siberia/145/96(H11N9))	H11N9	Avian			Russia	1996
Influenza A virus (A/duck/Siberia/151/96(H4N6))	H4N6	Avian			Russia	1996
Influenza A virus (A/duck/Siberia/1566/1997(H3N8))	H3N8	Avian			Russia	1997
Influenza B virus (B/Russia/69)		Human			Russia	1969
Influenza A virus (A/duck/Siberia/272/1998(H13N6))	H13N6	Avian			Russia	1998
Influenza A virus (A/duck/Siberia/272PF/1998(H13N6))	H13N6	Avian			Russia	1998

Имя сегмента

Сегмент

Полнота

Длина

Influenza A virus (A/duck/Siberia/272PF/1998(H13N6)) PA gene for polymerase acidic protein, complete cds	3	с	2208
Influenza A virus (A/duck/Siberia/272PF/1998(H13N6)) PB2 gene for polymerase basic protein 2, complete cds	1	с	2316
Influenza A virus (A/duck/Siberia/272PF/1998(H13N6)) PB1, PB1-F2 genes for polymerase basic protein 1, polymerase basic protein 1-F2, complete cds	2	с	2316
Influenza A virus (A/duck/Siberia/272PF/1998(H13N6)) HA gene for haemagglutinin, complete cds	4	с	1746
Influenza A virus (A/duck/Siberia/272PF/1998(H13N6)) NP gene for nucleoprotein, complete cds	5	с	1540

В базе: 105092

В выборке: 1092

Рис. 5. Внешний вид веб-интерфейса в режиме поиска нуклеотидных последовательностей в базе данных вирусов гриппа. На данном этапе разработки все данные в нашей БД – искусственно сгенерированные значения и нуклеотидные последовательности. До конца 2024 года база будет наполнена реальными данными из GenBank и собственными данными команды, участвующей в проекте

BLAST

Изменить способ ввода

Выбор файла

input_example.fasta

File details:

Name: input_example.fasta

Type:

Size: 2557 bytes

База данных Blast:

genbankBlastDb

Запустить BLAST

qseqid: A/Cambodia/KSH230332/2023|EPI_ISL_18543355|A/_H5N1|||2023-11-23|Chin_Savuth|KSH230332|KSH230332|2023-11-25|National_Institute_of_Public_Healt

sseqid: MN602508

pident: 98.197

evalue: 0

bitscore: 3971

length: 2274

qcovus: 100

mismatch: 39

gapopen: 2

gaps: 2

qseqid: A/Cambodia/KSH230332/2023|EPI_ISL_18543355|A/_H5N1|||2023-11-23|Chin_Savuth|KSH230332|KSH230332|2023-11-25|National_Institute_of_Public_Healt

Рис. 6. Внешний вид веб-интерфейса в режиме поиска в БД по заданной нуклеотидной последовательности (с использованием программы BLAST)

Заключение

В статье представлены основные результаты работы по созданию отечественной веб-платформы для загрузки, хранения, поиска и анализа информации о распространенных респираторных вирусах, включая их генетические последовательности и метаданные, являющиеся итогом начального этапа разработки системы. Представлен работающий прототип с веб-интерфейсом, базой данных, системной панелью администратора и начальным набором функций для работы с нуклеотидными последовательностями, который в следующем году будет пополняться новыми возможностями, включая наши собственные разработки – ускоренную многопоточную версию [15] алгоритма 3SEQ [16], используемого для выявления рекомбинантных вирусных последовательностей, и программную систему, в которой реализован новый подход к анализу эволюционных изменений SARS-CoV-2 за весь период его существования на основе всех доступных образцов.

Литература

1. Bogner P., Capua I., Lipman D.J., Cox N.J. A global initiative on sharing avian flu data // *Nature*. 2006. № 442(7106), C. 981. <https://doi.org/10.1038/442981a>.
2. Farley M.M. 2009 H1N1 influenza: a twenty-first century pandemic with roots in the early twentieth century // *Am. J. Med. Sci.* 2010. № 340(3), C. 202-208. <https://doi.org/10.1097/MAJ.0b013e3181e937b0>.
3. Tanner W.D., Toth D.J., Gundlapalli A.V. The pandemic potential of avian influenza A(H7N9) virus: a review // *Epidemiol. Infect.* 2015. № 143(16), C. 3359-3374. <https://doi.org/10.1017/S0950268815001570>.
4. Martellucci C.A., Flacco M.E., Cappadona R., Bravi F., Mantovani L., Manzoli L. SARS-CoV-2 pandemic: An overview // *Advances in Biological Regulation*. 2020. № 77, C. 100736. <https://doi.org/10.1016/j.jbior.2020.100736>.
5. Si Y., Wu W., Xue X., Sun X., Qin Y., Li Y., Qiu C., Li Y., Zhuo Z., Mi Y., Zheng P. The evolution of SARS-CoV-2 and the COVID-19 pandemic // *PeerJ*. 2023. № 11, C. e15990. <https://doi.org/10.7717/peerj.15990>.
6. Lenharo M. GISAID in crisis: can the controversial COVID genome database survive? // *Nature*. 2023. № 617(7961), C. 455-457. <https://doi.org/10.1038/d41586-023-01517-9>.
7. <https://microbius.ru/news/gisaid-v-krizise-smozhet-li-vyzhit-vyzyvayuschaya-spory-baza-dannyh-genomov-covid>.
8. Sayers E.W., Cavanaugh M., Clark K., Pruitt K.D., Schoch C.L., Sherry S.T., et al. GenBank // *Nucleic Acids Res.* 2022. №50(D1), C. D161-D164. <https://doi.org/10.1093/nar/gkab1135>.
9. Sayers E.W., Cavanaugh M., Clark K., Pruitt K.D., et al. GenBank 2024 Update // *Nucleic Acids Research*. 2024. № 52(D1), C. D134-D137. <https://doi.org/10.1093/nar/gkad903>.
10. Wu F., Zhao S., Yu B., Chen Y.M., Wang W., et al. A new coronavirus associated with human respiratory disease in China // *Nature*. 2020. № 579(7798), C. 265-269. <https://doi.org/10.1038/s41586-020-2008-3>.
11. <http://courseware.cutm.ac.in/wp-content/uploads/2020/10/Gen-Bank.pdf>.
12. Pertsemlidis A., Fondon J.W. Having a BLAST with bioinformatics (and avoiding BLASTphemy) // *Genome Biology*. 2001. № 2(10), C. reviews2002.1. <https://doi.org/10.1186/gb-2001-2-10-reviews2002>.
13. Aksamentov I., Roemer C. et al. Nextclade: clade assignment, mutation calling and quality control for viral genomes // *J. Open Source Software*. 2021. № 6(67), C. 3773. <https://doi.org/10.21105/joss.03773>.
14. Ahdritz G., Bouatta N., Floristan C., Kadyan S., et al. OpenFold: retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization // *Nature Methods*. 2024. №21(8), C. 1514-1524. <https://doi.org/10.1038/s41592-024-02272-z>.
15. Девятериков А.П., Пальянов А.Ю. Ускорение поиска рекомбинантных вирусных последовательностей для алгоритма 3SEQ за счет многопоточности и учета дат сбора образцов // *Математическая биология и биоинформатика*. 2024. № 19(2), C. 338-353. <https://doi.org/10.17537/2024.19.338>, <https://github.com/NotNa19/RecombinantDetector>.
16. Lam H.M., Ratmann O. and Boni M.F. Improved algorithmic complexity for the 3SEQ recombination detection algorithm // *Mol. Biol. Evol.* 2018. №35, C. 247-251. <https://doi.org/10.1093/molbev/msx263>.

Мордвинов Александр Валентинович. ИСИ СО РАН, Новосибирск, Россия. Аспирант. НГУ, Новосибирск, Россия. Ассистент кафедры систем информатики. Область научных интересов: математическое и программное обеспечение вычислительных машин, информационные технологии, проектирование и разработка программных систем, веб-программирование, базы данных. E-mail: a.mordvinov@g.nsu.ru

Стучинский Арсений Викторович. ИСИ СО РАН, Новосибирск, Россия. Программист. НГУ, Новосибирск, Россия. Магистрант. Область научных интересов: программирование, математика, информационные технологии, компьютерное моделирование, разработка мобильных приложений для Android. E-mail: a.stuchinskii@g.nsu.ru

Девятериков Антон Павлович. ИСИ СО РАН, Новосибирск, Россия. Аспирант. Область научных интересов: математика, информационные технологии, проектирование и разработка программного обеспечения для научных исследований, компьютерное моделирование. E-mail: a.devyaterikov@g.nsu.ru

Хайрулин Сергей Сергеевич. ИСИ СО РАН, Новосибирск, Россия. Младший научный сотрудник. Область научных интересов: компьютерное моделирование, численные методы, информационные технологии, проектирование и разработка программного обеспечения для научных исследований; параллельные вычисления на GPU, высокопроизводительные, распределенные и суперкомпьютерные вычисления. E-mail: s.khayrulin@gmail.com

Пальянова Наталья Валерьевна. НИИ вирусологии ФИЦ ФТМ, Новосибирск, Россия. Младший научный сотрудник. Область научных интересов: вирусология, клеточная биология, эпидемиология, биоинформатика. E-mail: natalia.palyanova@gmail.com

Пальянов Андрей Юрьевич ИСИ СО РАН, Новосибирск, Россия. Директор, доктор физико-математических наук. НИИ вирусологии ФИЦ ФТМ, Новосибирск, Россия. Научный сотрудник. НГУ, Новосибирск, Россия. Старший преподаватель. Область научных интересов: алгоритмы, вычислительные методы и программные системы для компьютерного моделирования и анализа данных в нейробиологии, кибернетике, биомеханике, гидродинамике, молекулярной биологии и биоинформатике; высокопроизводительные параллельные вычисления на GPU; трехмерная компьютерная графика и визуализация; физические симуляторы; когнитивные системы; теория сознания; теория эволюции; виртуальные организмы. E-mail: palyanov@iis.nsk.su

On the Project of an Effective Software Platform for Working with Genetic Data of Respiratory Viruses

A.V. Mordvinov¹, A.V. Stuchinsky^{1,II}, A.P. Devyaterikov¹, S.S. Khayrulin¹,
N.V. Palyanova^{III}, A.Yu. Palyanov^{I,II,III}

¹A.P. Ershov Institute of Informatics Systems, Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia

^{II}Novosibirsk State University, Novosibirsk, Russia

^{III}Research Institute of Virology, Federal Research Center for Fundamental and Translational Medicine, Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia

Abstract. Progress in sequencing technologies, i.e. reading the nucleotide sequences of living organisms, has led to a rapid growth of the amount of genetic data. The largest global projects that accumulate this information and provide online access to it are Genbank and GISAID. Also they provide basic capabilities for analyzing this data online, but they are quite limited. This significantly limits our abilities to effectively solve a number of scientific problems and tasks, so we decided to develop our own domestic (Russian) web platform with capabilities which we need. The main goal of this project is to provide a team of researchers with the opportunity to effectively solve problems in bioinformatics, virology and epidemiology, based on modern, effective, reasonably selected software solutions operating with high performance and providing many useful functionalities which can be extended by adding new necessary programs for analyzing and modeling. The web platform we are implementing will allow to download, store, search and analyze genomic sequences of viruses, such as influenza, SARS-CoV-2 and, in perspective, other viral pathogens. In addition, the project will develop and advance through efforts of IT part of our team taking into account actual needs of bioinformaticians and virologists. We plan to make it available to researchers around the world and periodically update both the software and the data (from open sources) to improve the convenience and efficiency for scientists working in the relevant areas.

Keywords: web platform, bioinformatics, virology, epidemiology, data storage, data analysis, algorithms, efficiency.

DOI 10.14357/20718632240406 EDN EUCRNU

References

1. Bogner P., Capua I., Lipman D.J., Cox N.J. A global initiative on sharing avian flu data // *Nature*. 2006. № 442(7106), C. 981. <https://doi.org/10.1038/442981a>
2. Farley M.M. 2009 H1N1 influenza: a twenty-first century pandemic with roots in the early twentieth century // *Am. J. Med. Sci.* 2010. № 340(3), C. 202-208. <https://doi.org/10.1097/MAJ.0b013e3181e937b0>

3. Tanner W.D., Toth D.J., Gundlapalli A.V. The pandemic potential of avian influenza A(H7N9) virus: a review // *Epidemiol. Infect.* 2015. № 143(16), C. 3359-3374. <https://doi.org/10.1017/S0950268815001570>.
4. Martellucci C.A., Flacco M.E., Cappadona R., Bravi F., Mantovani L., Manzoli L. SARS-CoV-2 pandemic: An overview // *Advances in Biological Regulation*. 2020. № 77, C. 100736. <https://doi.org/10.1016/j.jbior.2020.100736>
5. Si Y., Wu W., Xue X., Sun X., Qin Y., Li Y., Qiu C., Li Y., Zhuo Z., Mi Y., Zheng P. The evolution of SARS-CoV-2 and the COVID-19 pandemic // *PeerJ*. 2023. № 11, C. e15990. <https://doi.org/10.7717/peerj.15990>.
6. Lenharo M. GISAID in crisis: can the controversial COVID genome database survive? // *Nature*. 2023. № 617(7961), C. 455-457. <https://doi.org/10.1038/d41586-023-01517-9>
7. <https://microbius.ru/news/gisaid-v-krizise-smozhet-li-vyzhit-vyzyvayuschaya-spory-baza-dannyh-genomov-covid>
8. Sayers E.W., Cavanaugh M., Clark K., Pruitt K.D., Schoch C.L., Sherry S.T., et al. GenBank // *Nucleic Acids Res.* 2022. №50(D1), C. D161-D164. <https://doi.org/10.1093/nar/gkab1135>.
9. Sayers E.W., Cavanaugh M., Clark K., Pruitt K.D., et al. GenBank 2024 Update // *Nucleic Acids Research*. 2024. № 52(D1), C. D134–D137. <https://doi.org/10.1093/nar/gkad903>.
10. Wu F., Zhao S., Yu B., Chen Y.M., Wang W., et al. A new coronavirus associated with human respiratory disease in China // *Nature*. 2020. № 579(7798), C. 265-269. <https://doi.org/10.1038/s41586-020-2008-3>.
11. <http://courseware.cutm.ac.in/wp-content/uploads/2020/10/Gen-Bank.pdf>.
12. Pertsemlidis A., Fondon J.W. Having a BLAST with bioinformatics (and avoiding BLASTphemy) // *Genome Biology*. 2001. № 2(10), C. reviews2002.1. <https://doi.org/10.1186/gb-2001-2-10-reviews2002>
13. Aksamentov I., Roemer C. et al. Nextclade: clade assignment, mutation calling and quality control for viral genomes // *J. Open Source Software*. 2021. № 6(67), C. 3773. <https://doi.org/10.21105/joss.03773>.
14. Ahdritz G., Bouatta N., Floristean C., Kadyan S., et al. OpenFold: retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization // *Nature Methods*. 2024. №21(8), C. 1514-1524. <https://doi.org/10.1038/s41592-024-02272-z>.
15. Devyaterikov A.P., Palyanov A.Yu. Acceleration of recombinant viral sequences search by 3SEQ algorithm via adding support of multi-threaded calculations and considering sample collection dates // *Mathematical Biology and Bioinformatics*. 2024. № 19(2), C. 338-353. <https://doi.org/10.17537/2024.19.338>. <https://github.com/NotNa19/RecombinantDetector>.
16. Lam H.M., Ratmann O. and Boni M.F. Improved algorithmic complexity for the 3SEQ recombination detection algorithm // *Mol. Biol. Evol.* 2018. №35, C. 247–251. <https://doi.org/10.1093/molbev/msx263>.

Mordvinov Alexander V. A.P. Ershov Institute of Informatics Systems, Novosibirsk, Russia – Ph.D. student. Novosibirsk State University, Novosibirsk, Russia – assistant at the Department of Informatics Systems, Faculty of Information Technologies. Area of scientific interests: information technologies, mathematics and software engineering for design and development of software and computational systems; web programming; databases. E-mail: a.mordvinov@g.nsu.ru

Stuchinsky Arseny V. A.P. Ershov Institute of Informatics Systems, Novosibirsk, Russia – software engineer. Novosibirsk State University, Novosibirsk, Russia – postgraduate student of the Department of Programming, the Faculty of Mechanics and Mathematics. Area of scientific interests: software engineering, information technologies, computational modelling, development of Android mobile applications. E-mail: a.stuchinskii@g.nsu.ru

Devyaterikov Anton P. A.P. Ershov Institute of Informatics Systems, Novosibirsk, Russia. Ph.D. student, software engineer. Area of scientific interests: mathematics, information technologies, design and development of software for scientific research, computational modeling. E-mail: a.devyaterikov@g.nsu.ru

Khayrulin Sergey S. A.P. Ershov Institute of Informatics Systems, Novosibirsk, Russia – junior researcher. Area of scientific interests: computational modeling, numerical methods, information technologies, design and development of software for scientific research; GPU, high-performance, parallel and distributed computing. E-mail: s.khayrulin@gmail.com

Palyanova Natalia V. Research Institute of Virology, Federal Research Center for Fundamental and Translational Medicine, Novosibirsk, Russia – junior researcher. Area of scientific interests: virology, cell biology, epidemiology, bioinformatics. E-mail: natalia.palyanova@gmail.com

Palyanov Andrey Yu. Ph.D., Doctor of physics and mathematics. A.P. Ershov Institute of Informatics Systems, Novosibirsk, Russia – director. Research Institute of Virology, Federal Research Center for Fundamental and Translational Medicine, Novosibirsk, Russia – researcher. Novosibirsk State University, Novosibirsk, Russia – senior lecturer at the Department of Programming, Faculty of Mechanics and Mathematics. Research interests: algorithms, computational methods and software systems for computational modeling and data analysis in neurobiology, cybernetics, biomechanics, hydrodynamics, molecular biology and bioinformatics; high-performance parallel computing on GPU; 3D graphics and visualization; physical simulators; cognitive systems; theory of consciousness; theory of evolution; virtual organisms. E-mail: palyanov@iis.nsk.su