УДК 004.9

Построение онлайн системы с Web интерфейсом для хранения, обработки и анализа генетических последовательностей вируса SARS-CoV-2.

Старцев П.А. (Институт систем информатики СО РАН)

Работа с геномами, поиск в известной математически формализованной последовательности как можно более стабильной или максимальной по длине подцепочке, поиск, как математическая задача для программного обеспечения по различным другим критериям являются одними из самых актуальных задач при работе с современным инструментарием исследователя в области вирусологии. Спектр прикладных задач сегодняшнего времени включает в себя создание новых праймеров для диагностики вируса по стабильным участкам генома и выявления последних мутаций вируса, классификации и кластеризации накопленного материала для более точных исследований дальнейшей мутации и тому подобное. Построение вычислительных комплексов на основе базы данных и расширение их функциональности при помощи как онлайн технологий, так и запуска серверных приложений, является сложной практической задачей в сфере программирования для более эффективной работы вирусологов и специалистов из смежных областей в сфере диагностики и лечения вирусных заболеваний.

Ключевые слова: *SARS-CoV-2, коронавирус, онлайн система, генетическая последовательность, геном, ИСИ СО РАН.*

1. Введение.

Несмотря на стабилизацию ситуации с коронавирусом в России, актуальность работ по этому типу вирусов остается достаточно высокой: наблюдается дальнейшее распространение вируса на фоне снижения смертности [4], выявляются мутации, проверяются новые механизмы работы с математическим инструментарием [5] и подходы для диагностики и лечения остающегося опасным вируса и сопутствующих заболеваний [3]. Поскольку SARS-CoV-2 обладает одним из самых больших геномов среди всех выявленных РНК-вирусов (см. [1], с.19), очень подробно и на больших объемах данных разделен на белки по типам (S, E, M, и т.п. [1]), то с каждым днем становится все более актуальной задача поиска по фрагменту генома, возможность найти часть последовательности нуклеотидов, начиная с какой-либо позиции или внутри одного белка. Кластеризация участков генома и работа с конкретными

мутациями внутри кластера также требуют более тонкой настройки поиска последовательностей нуклеотидов внутри БД (см. [4], с.4). Длина последовательностей одного генома SARS-CoV-2 в базе данных – около 30000 нуклеотидов. Повышенный интерес с точки зрения создания вакцин и диагностических систем представляет из себя S (Spike) белок [2, с. 13]. Перечисленные выше возможности поиска - это несколько примеров функциональности, не представленной непосредственно средствами GENBANK, которые необходимо разрабатывать под нужды специалистов самостоятельно. Более того, к собственной базе данных можно строить обращения помимо систем поиска и программ, в этом случае гибкость запросов поиска ограничена лишь средствами языка обращения к БД SQL.

Целью работы является построение программной системы на основе собственной базы данных для работы как с геномом SARS-CoV-2, так и с его отдельными участками, а также с метаинформацией (сведения о возрасте пациента, лаборатории, в которой получена последовательность, дате вакцинации, и т.п.). В текущей, описанной в настоящей работе, конфигурации, реализована работа как с онлайн интерфейсом (скриншоты, см. рис.1 и далее), так и с дополнениями, обладающей следующими качествами:

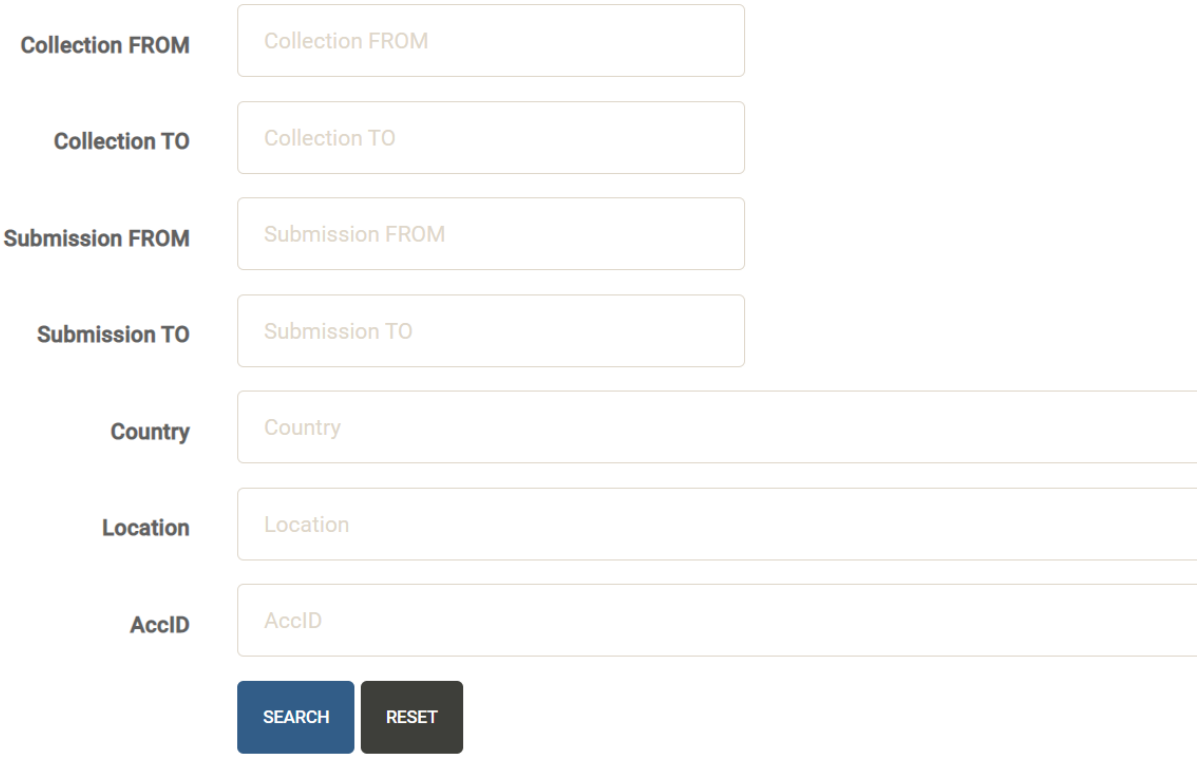

Рис.1 Внешний вид страницы поиска по параметрам системы. Язык интерфейса – английский.

- Возможность загрузки, первичной валидации/оценки и кластеризации/классификации генетических последовательностей из различных баз данных в виде больших файлов в формате fasta.
- Возможность запуска подпрограмм, необходимых для обработки как промежуточных результатов (например, результатов поиска), так и работы с единичными последовательностями.
- Развитие специфических задач программного обеспечения (далее ПО) системы для специалиста в области вирусологии.
- Накопление статистических данных и данных производительности ПО.

2. Построение и свойства системы.

2.1. Описание текущей конфигурации системы.

В настоящее время система включает в себя максимально возможную, сравнимую по объему базу данных с GENBANK по коронавирусу SARS-CoV-2 (оценки и сравнения приведены на середину 2023 года, см. п. 3.1). Больше о базе данных GENBANK (далее БДG) можно найти, например, здесь [8].

База GENBANK постоянно пополняется и является одним из доступных хранилищ данных, но в силу последних обстоятельств ограничительного свойства работы с международными хранилищами, постепенно возникает необходимость локализовать хранилище секвенированных последовательностей нуклеотидов с сохранением метаинформации. Также в БДG декларируются ограничения по валидации и правам собственности на последовательности геномов, что не позволяет научно использовать этот материал без локальной верификации, ссылки на ограничения проприетарного свойства и уточнения классификационных признаков[8]. Последовательности в БДG зачастую имеют процент нуклеотидов, которые явно не определены (помечены, как правило, символом N). Кроме того, извлечение данных в этом конкретном случае сопряжено с неудобством технического свойства при скачивании файлов большого объема. Поисковые механизмы сайта достаточно глубоко разработаны, но при некоторых критериях поиска, например по неточной дате секвенирования, есть достаточно серьезное несоответствие данных при составлении загрузочных файлов. Кроме того, согласно [8], в явном виде не разделяется информация о количестве поступающих новых цепочек нуклеотидов именно по SARS-CoV-2, а собирается более общая статистика, что не позволяет точно оценить динамику обновления данных именно в интересующем исследователей сегменте данных по гранту и

работе с SARS-CoV-2 в целом. Также есть список ошибок на официальном сайте в разделе validations [8].

Была создана локальная база данных (далее БД) Postgres SQL версии 14. Выбор версии и типа БД производился по следующим критериям:

- Потенциальный переход на отечественный продукт
- Минимизация проприетарных рисков
- Открытость кода, лицензирование, допускающее гибкую работу с ядром ПО и обслуживанием версий кода
- Возможность работы с большими объемами данных
- Бесплатная обновляемость и доступная система сопровождения ПО
- Безопасность, в том числе с точки зрения современных вызовов санкционного характера

В качестве серверного решения используется пакет открытого ПО Spring, язык программирования JAVA[6]. В качестве Web слоя используются библиотеки с открытым исходным кодом для современной линейки браузеров на базе Javascript [8,9]. Конфигурация системы управления БД, комплекс серверных технологий, конкретных слоев серверного ПО и подходов к разработке будут рассмотрены в отдельной статье.

2.2. Запуск подпрограмм: BLASTN.

В системе есть возможность из Web интерфейса запускать подпрограммы уровня операционной системы. Для этого используется механизм потоков JAVA [6]: Создается объект-процесс, указываются его параметры запуска, и, после выполнения, проверяется результат, возвращаемый операционной системой. Есть возможность логирования как работы самой программы, так и контекста запуска для отладки и контроля. (см. рис. 5 раздела 2.4)

Семейство программ BLASTN [5] представляет из себя набор библиотек и файлов запуска с консольным интерфейсом. На вход программе в системе передаются:

- Файл с последовательностями в формате fasta (одной или более) нуклеотидов для сравнения
- Файл, сформированный в формате fasta [7] с последовательностями нуклеотидов, отобранными по критериям поиска из системы. Результаты текущей конфигурации поиска можно сохранить в файл локальной системы в формате fasta и передать на обработку как параметр запуска
- Файл результата. В этот файл консольная программа выведет результат работы.

- Дополнительные параметры, которые можно настраивать в системе (формат вывода данных, столбцы таблиц информации, ограничения вывода и т.п.)

Возможен запуск других аналогичных программ с консольным выводом.

BLAST (Basic Local Alignment Search Tool) как семейство программ, создано для работы с нуклеотидными последовательностями, с базами данных секвенированных геномов и их участков [7]. В нашем случае подпрограмма BLASTN является одним из важнейших инструментов семейства для поиска относительно коротких участков последовательности или одного участка последовательности с небольшими отклонениями относительно изучаемой последовательности. Формат вывода результата может быть настроен при вызове и может содержать как сравниваемые участки последовательности, так и общую информацию в виде списка дополнительной метаинформации исследуемых сравнений. Исследователя может интересовать, например, страна происхождения генома, тип вируса и т.п.[5,7].

До создания системы приходилось подготавливать файлы, которые передаются на вход работе программы BLASTN вручную. Особое неудобство вызывает при таком методе работы формирование файла сравнения (аналога данных из БД), который может быть слишком большим по объему для работы с основными используемыми текстовыми редакторами. Вместо такого подхода можно использовать раздел поиска WEB интерфейса (см. рис. 1) и возможности системы сериализации непосредственно в файл текстового формата.

2.3. Web интерфейс

Web интерфейс (см. рис. 1-6) вместо локального приложения был выбран по следующим критериям:

- Возможность работы не только локально, но и удаленно.
- Поиск через Web интерфейс на сегодняшний день включает в себя поля дат (секвенирования и передачи в исходную базу данных), локации (как страны, так и более узкой области внутри страны), идентификатора записи и подстроки последовательности.
- Использование локальной программы сужает круг разработчиков самой специфической части интерфейса, возникают риски развитию системы – для дальнейшей разработки собственного специфического интерфейса для данного языка программирования и библиотек может не найтись разработчика в дальнейшем.

- Отдельная часть интерфейса локальной программы так или иначе дублирует соответствующие уже готовые части браузера, что повышает время разработки и вероятность ошибок в коде, который и не нужно было создавать.

Для создания Web интерфейса (см. рис 2), используются технологии Javascript библиотек с открытым кодом и библиотека шаблонов THYMELEAF[9].

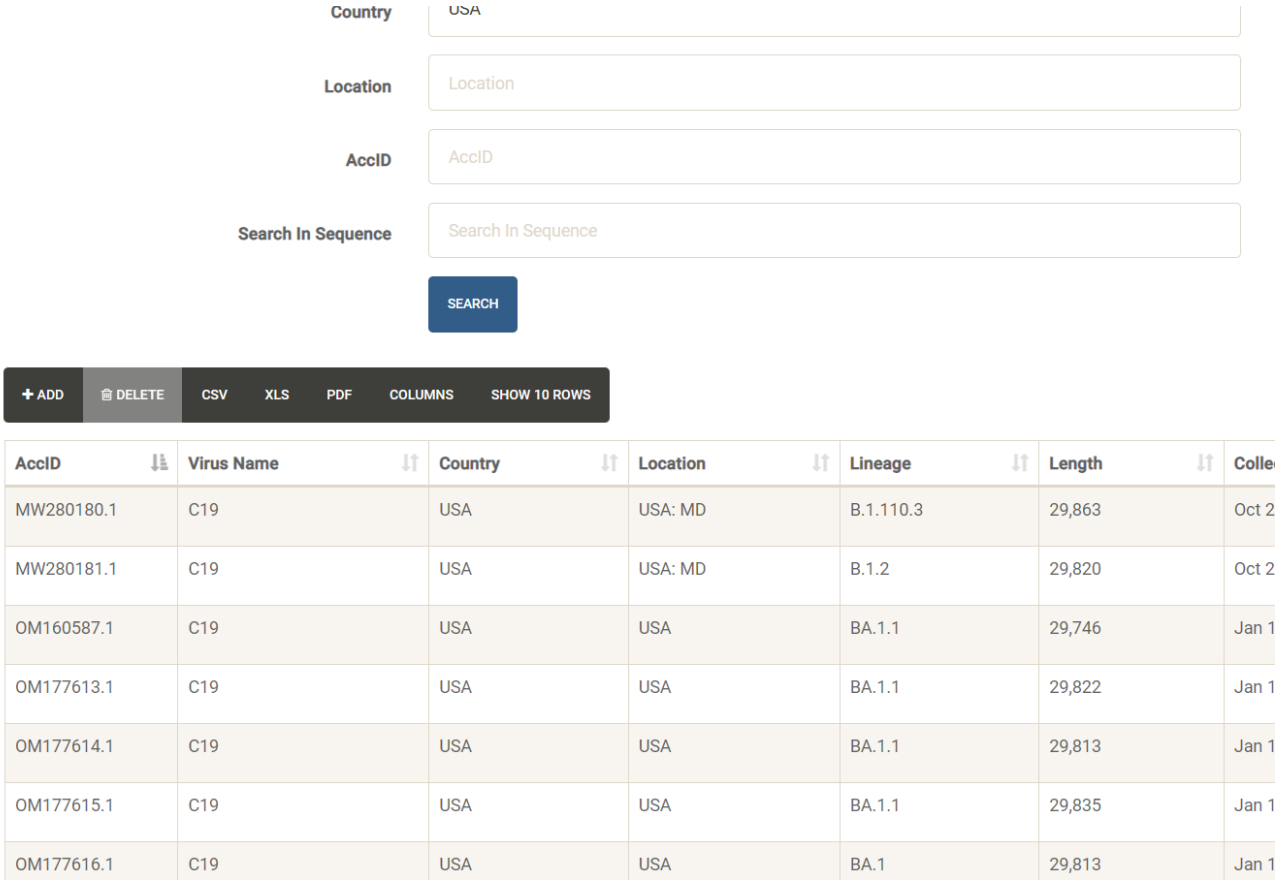

Рис.2 Результаты поиска по стране и более узкой локации внутри страны. Кнопки в столбце Tools позволяют перейти на страницу нуклеотидной последовательности (см. рис 3,4). Скриншот с браузера Опера.

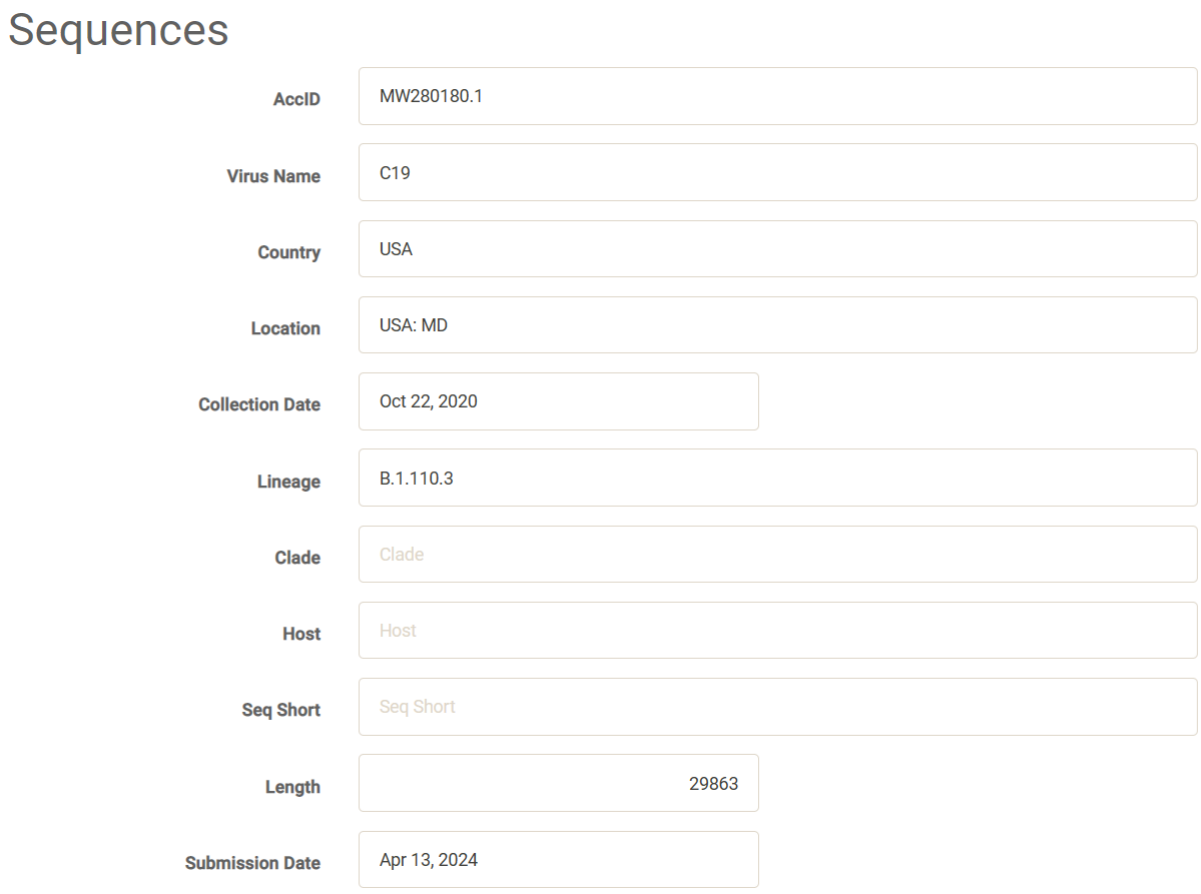

Рис.3 Страница с полной информацией о генетической последовательности из базы данных (Начало).

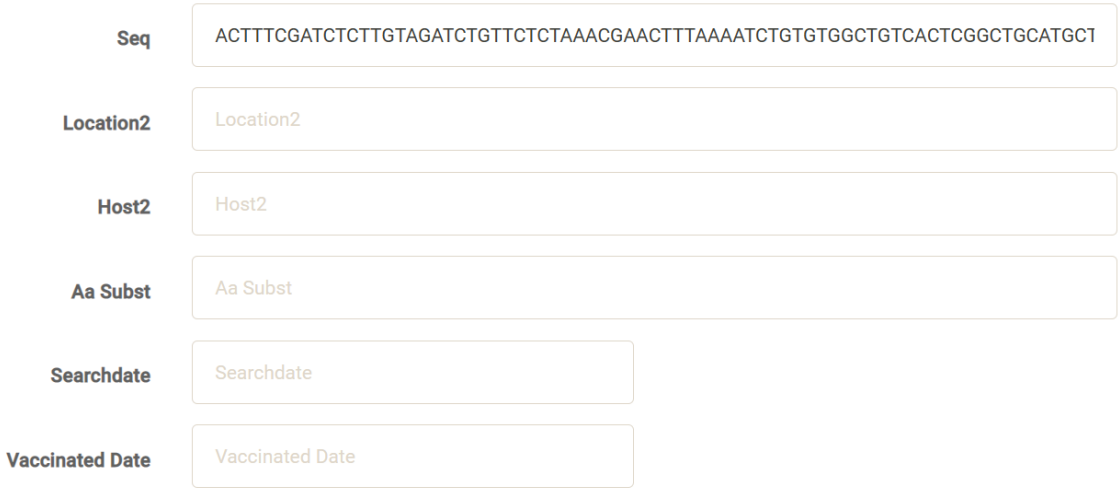

Рис.4 Страница с полной информацией о генетической последовательности из базы данных(Окончание). Из поля «Последовательность» (Seq) можно скопировать саму последовательность для дальнейшей работы.

2.4. Процесс и результаты запуска подпрограммы BLASTN

Для запуска программ в системе удобно пользоваться сериализацией результатов поиска в файлы, удобные для чтения и работы с текстовыми редакторами. Специально для этого существует возможность сериализации, в том числе, и в формат fasta[5].

После запуска через Web интерфейс (см. раздел 2.2) программа blastn начинает работу при помощи одного из исполняемых файлов (возможен запуск через файлы операционных систем Windows или Linux, настраивается при сборке системы).

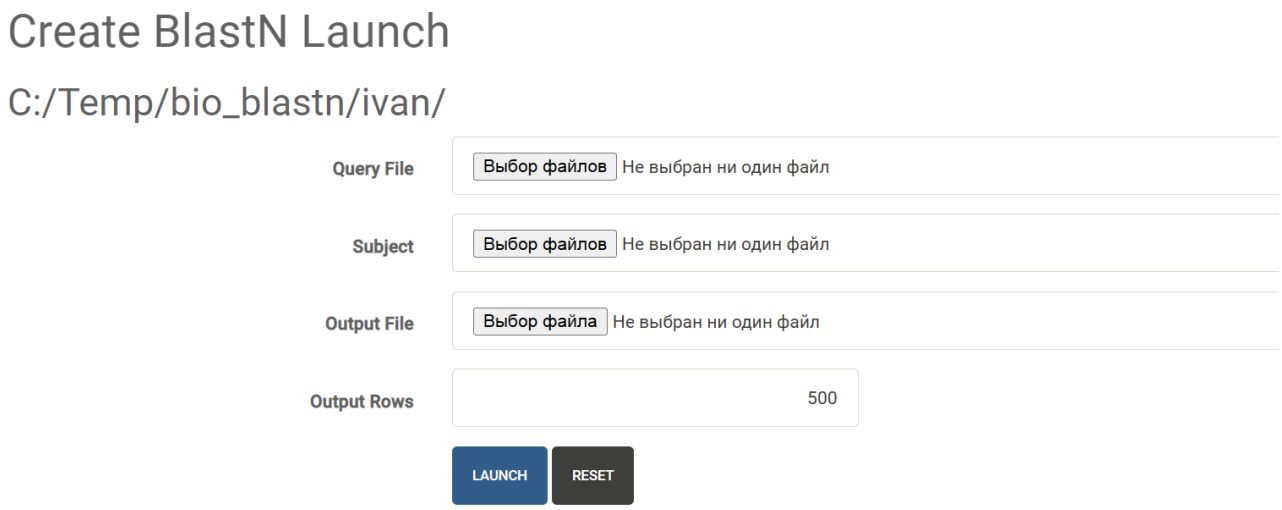

Рис.5 Страница запуска подпрограммы BLASTN. Есть возможность работать в собственной директории, указываемой в настройках.

Работа с blastn возможна и вне Web интерфейса.

После запуска программа переходит на страницу статистики (результатов) запусков (рис.6). Есть возможность поиска по дате, контекстный поиск, например, по году или части названия файла. Формат выходящих данных программы blastn является одной из настроек (0-11) и может содержать, например, таблицу данных без комментариев (значение 6, настроено в системе). Настройки столбцов могут содержать статистику совпадений/несовпадений участков генома, идентификаторы последовательностей, специальные функции и т.п., всего несколько десятков опциональных выходных данных, разделенных пробелом.

В качестве одного из направлений развития системы может быть рассмотрена база специального формата, непосредственно предназначенная для работы с blastn и создаваемая при помощи этой же командной строки. Это может быть и временный объект, наполняемый данными под один поиск, а также специальное хранилище для постоянных обращений под

конкретную задачу, наполняемый из Web интерфейса или по настраиваемой задаче (скрипту). Создается как файл в файловой системе[5].

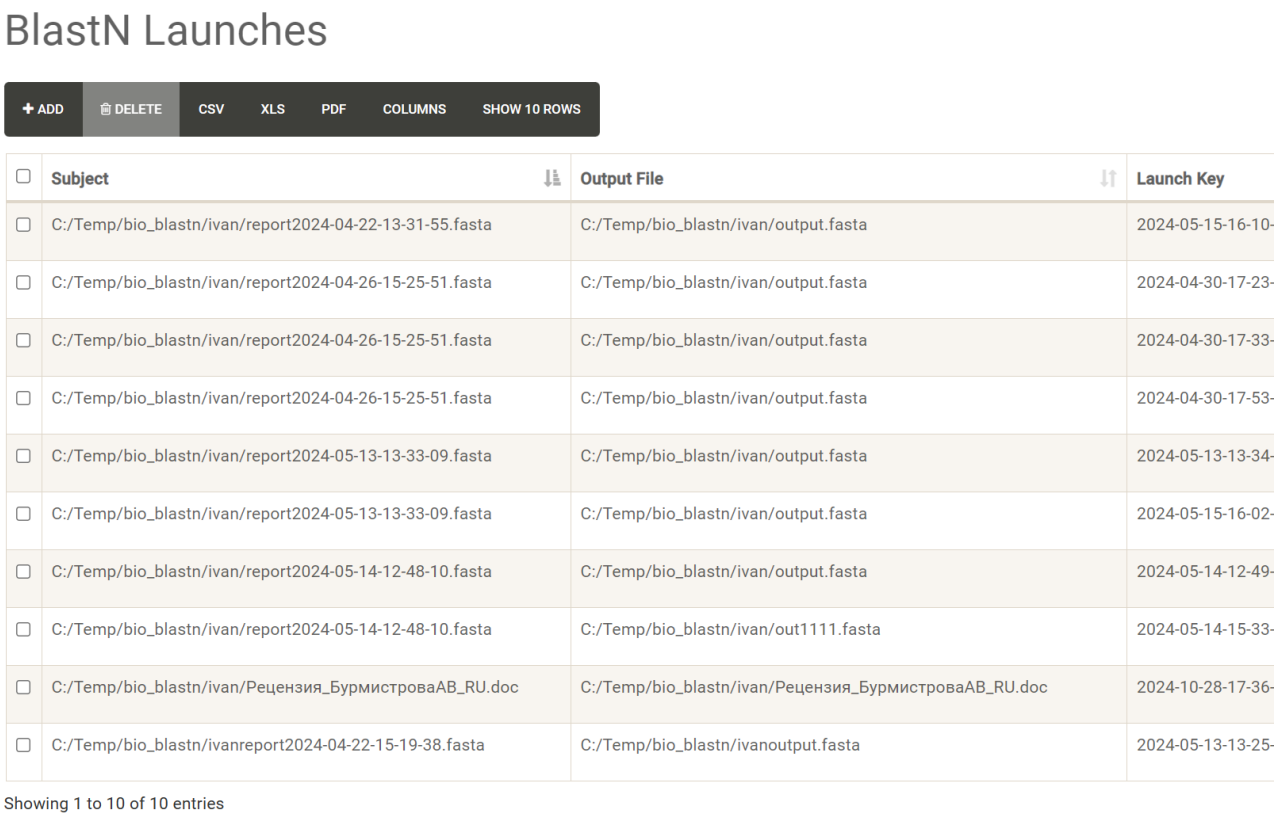

Рис.6 Страница статистики запусков подпрограммы BLASTN.

3. Актуальность, статистика, планы по развитию системы.

3.1. Актуальность проекта в цифрах и преимущества работы.

Создание базы данных и наполнение ее последовательностями генома и метаинформацией за несколько лет наблюдения позволило собирать статистику и обращаться к этим данным через удобные инструменты работы с языком запросов SQL вместо работы с большими файлами. Скорость поиска по базе данных (на точное соответствие по идентификатору, по датам периода, по частичному соответствию локации) не превышает нескольких секунд, что сравнимо с поиском по данным GENBANK на сайте, однако пока уступает по функциональности. При этом преимущество системы в том, что можно сразу использовать результаты выборки вместо скачивания файла с удаленного ресурса. Происходит сохранение набора последовательностей в fasta файл, готовый к работе

с другим инструментарием, причем приемлемого размера, например, легко редактируемый в привычных для работы редакторах.

Сохранены механизмы загрузки данных из GENBANK в систему для обновления данных со временем. Доступна статистика по загруженным данным, в том числе она может выдаваться по запросам к БД: как агрегатных функций, так и аналитических. Данные проверены на соответствие полученной информации по полям БД и прочим ошибкам (могут быть допущены некоторые неточности при выборе загрузки с сайта GENBANK) на тестах.

	records	country
1	3570727	USA
2	1857744	United Kingdom
3	800717	Germany
4	102676	Denmark
5	86051	Switzerland
6	46398	France
7	26885	Bahrain
8	23807	Slovakia
9	10469	Japan
10	9462	Iceland

Рис.7 Общее количество записей в Базе Данных по странам в порядке убывания, 10 первых позиций.

3.2. Текущее состояние и планы по развитию системы.

Как сказано выше, на данный момент в системе реализован функционал по поиску последовательностей в БД, сохранению результатов поиска в файл для дальнейшей работы в удобном текстовом формате, ряд настроек, запуск консольных приложений, логирование и обработка ошибок. Данные, подгруженные из GENBANK, проверяются на соответствие требованиям работы, доступны дальнейшие загрузки и ведется работа по улучшению тестирования этих процессов.

Планы на развитие включают в себя:

- Дополнение механизмов поиска
- Разделение функционала системы между пользователями, локальные настройки, например, последовательности, доступные только одному исследователю.

- Разделение функционала на открытый (публичный) и закрытый (доступный только локальным пользователям). Работы по безопасности контента от внешних угроз и источников.
- Развитие БД: Возможности загрузки данных из других источников. Загрузка метаинформации отдельно от последовательностей.

4. Благодарности.

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

Список литературы

1. Бруякин С.Д., Макаревич Д.А. Структурные белки коронавируса SARS-COV-2: роль, иммуногенность, суперантигенные свойства и возможности использования для терапевтических целей // Вестник ВолгГМУ. 2021. Вып. 2(78). С. 18-27.
2. Воробьев П. О., Тиллиб С. В. Однодоменное антитело для связывания консервативного эпитопа рецептор-связывающего домена белка Spike коронавируса SARS-COV-2 // Вестник РГМУ. 2023. №1. С.12-21.
3. Гарафутдинов Р.Р., Мавзютов А.Р., Никоноров Ю.М., Чубукова О.В., Матниязов Р.Т., Баймиев Ан.Х., Максимов И.В., Мифтахов И.Ю., Халикова Е.Ю., Кулуев Б.Р., Баймиев Ал.Х., Чемерис А.В. Бетакоронавирус SARS-CoV-2, его геном, разнообразие генотипов и молекулярно-биологические типы борьбы с ним. // Биомика. 2020. Т.12(2). С. 242-271.
4. Краснов Я.М., Попова А.Ю., Сафронов В.А., Федоров А.В., Баданин Д.В., Щербакова С.А., Кутырев В.В. Анализ геномного разнообразия SARS-Cov-2 и эпидемиологических признаков адаптации возбудителя COVID-19 к человеческой популяции // Проблемы особо опасных инфекций. 2020. №3. С.70.
5. Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J. Basic Local Alignment Search Tool // Journal Of Molecular Biology. 1990. Vol. 215. N 3. P. 403-410.
6. Get Java for desktop applications. Официальная информация [Электронный ресурс]. URL: https://www.java.com/ru/ (дата обращения: 19.09.2024).
7. Hu G., Kurgan L., Sequence Similarity Searching // Current Protocols In Protein Science. 2019. Vol.95. N 1. P. e71.
8. What is GenBank? Официальная информация [Электронный ресурс]. URL: https://www.ncbi.nlm.nih.gov/genbank/ (дата обращения: 20.09.2024).
9. What Spring can do. Официальная информация [Электронный ресурс]. URL: https://spring.io/ (дата обращения: 20.09.2024).