УДК 004.89

АНАЛИЗ БИОИНФОРМАЦИИ ДЛЯ МОНИТОРИНГА ОПАСНОЙ КОРОНАВИРУСНОЙ ИНФЕКЦИИ

М.В. Спринджук
Доцент кафедры экономической информатики БГУИР, кандидат технических наук, лектор
stepanenkomatvei@yandex.ru

М.В. Спринджук
Окончил аспирантуру БГУИР. Область научных интересов связана с разработкой алгоритмов, моделей и программного обеспечения для анализа биоинформационных данных геномной природы.

Аннотация. Выполнен анализ биоинформационных данных Монголии и Беларуси. Показаны волны смертности и выздоровления в Монголии по эпидемиологическим данным временных рядов. Проанализирован профиль геномных вариаций коронавируса в Беларуси и Монголии.

Ключевые слова: медицинская кибернетика, коронавирус, эпидемиология, системы медицинского назначения, прикладная математика, мониторинг опасных инфекций.

Введение.

Современное геномное секвенирование позволяет выделить и амплифицировать крупный геном РНК коронавируса [1-6]. Анализ полученных данных (рисунок 1) может идентифицировать и классифицировать элементарные структуры генома. С усовершенствованиями и доступностью компьютерной техники появились новые возможности более полноценно изучать биоинформационные данные, под которыми понимают геномный текст и соответствующие метаданные.

Актуальность.

Внезапный очаг новой коронавирусной инфекции COVID-19 появился в середине декабря 2019 года в Китае, в городе Ухань, инфекция распространилась на многие города Китая, Юго-Восточной Азии, и по всему миру.

За период с декабря 2019 года по апрель 2023 года в мире умерло около 68370864 больных с подтвержденным диагнозом новой коронавирусной инфекции. За это же время в Беларуси умерло 7118 пациента, в России – 397604, в Монголии – 2136 [https://www.worldometers.info/coronavirus/].

Материалы и методы исследования.

В течение интервала времени 2019-2023 гг. нами была выполнена загрузка SARS-CoV-2 геномов из общедоступной базы данных GISAID (*Global Initiative on sharing all influenza data = Глобальная инициатива по обмену всеми данными о гриппе*) [https://www.epicov.org/epi3/frontend#275474].

Для анализа данных применялось разработанное программное обеспечение на основе платформы Galaxy, в конвейере анализа данных были использованы модули Pangolin и Nextclade. Дополнительно для визуализации данных применялось программное обеспечение Stata [7] (рисунки 2,3), JMP SAS [8] (рисунки 4-6, 9). Также было разработано программное обеспечение

для визуализации прогноза COVID-19 смертности и выздоровления на основе анализа временных рядов. Для его реализации применялись языки программирования R (рисунки 7, 8) [9] и Python [10].

Pangolin [11,12] был разработан для реализации динамической номенклатуры линий и кластеров-кладов передачи SARS-CoV-2, известной как номенклатура Pango. Pangolin присваивает распознанную линию и имя кластера по принципу, опубликованному A. Rambaut др., 2020 [14].

Nextclade [13] – это инструмент, который определяет различия между загруженными пользователем геномными текстами и эталонной последовательностью и использует эти различия для идентификации, распознавания и присвоения линий передачи и кластеров-кладов, а также сообщает о потенциальных проблемах качества последовательностей в представляемых данных. Руководство пользователя доступно по адресу docs.nextstrain.org/projects/nextclade.

Филогенез нового коронавируса человека.

Филогенез коронавирусных геномов был исследован нами по разработанным ранее методикам анализа биоинформационных данных, опубликованным в источнике [14] и как продолжение вычислительных экспериментов.

В основе филогенетических исследований лежат методы множественного выравнивания нуклеотидных или аминокислотных последовательностей и вычислительные методы построения из результатов выравнивания визуализации, отображающей эволюционные отношения геномов, количественной характеристики эволюционной трансформации (рисунки 8,9).

Филогенетичекие вычислительные эксперименты ограничены техническими возможностями доступной компьютерной техники, отсутствием в нашем случае метаданных имеющихся геномов, что не позволяет в полной мере выполнить пространственный анализ и подробную кластеризацию. Также результаты такого исследования зависят от выбранных алгоритмов и моделей пред- и постобработки данных, методик фильтрации результатов выравнивания и обрезки ветвей кладограмм. Также очевидно, что имеющиеся в базе данных ошибки секвенирования, повлекут неизбежно и ошибки в результатах филогенетических вычислительных экспериментов.

Результаты исследования филогенеза коронавируса в Беларуси и Монголии.

Результаты филогенетического исследования показывают разнообразие мутационных кладов коронавируса. Отмечается также трансмиссионная активность различных кладов внутри страны и выход их из Беларуси за рубеж. По данным филогенетического исследования имеются основания предполагать, что занос новых вариантов на территорию Беларуси продолжается, однако данных о патогенности различных местных штаммов и вариантов недостаточно, чтобы сформировать строгие научные выводы. Биоинформационный анализ показывает вездесущность коронавируса, его уникальные возможности приспосабливаться мутационной изменчивостью. По сравнению с предыдущими вычислительными экспериментами по данным Беларуси, России и Европы нами наблюдается начальная тенденция к доминированию в Беларуси линий передачи Омикрон над Дельтой и другими кластерами и обнаруживается большое мутационное разнообразие изолятов, которые были идентифицированы как варианты Омикрона. В Монголии, согласно нашим вычислительным экспериментам, доминирует *Альфа штамм* коронавируса (рисунки 4, 5).

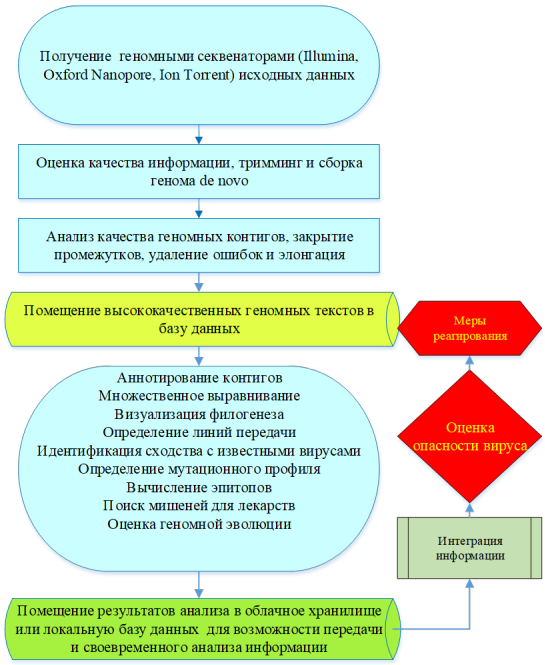

Рисунок 1. Конвейер получения и анализа информации антивирусного мониторинга. Разработка автора доклада

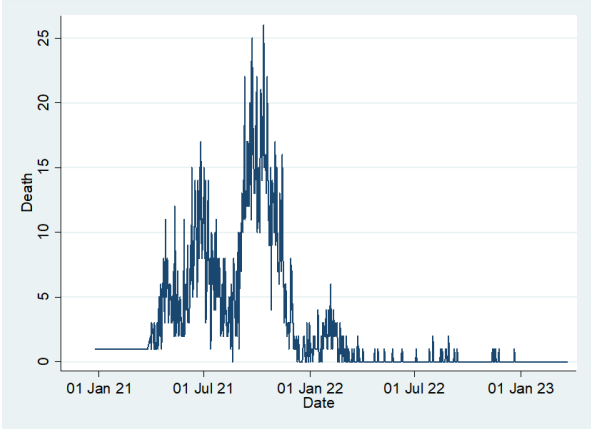

Рисунок 2. Визузализация динамики COVID-19 смертности в Монголии

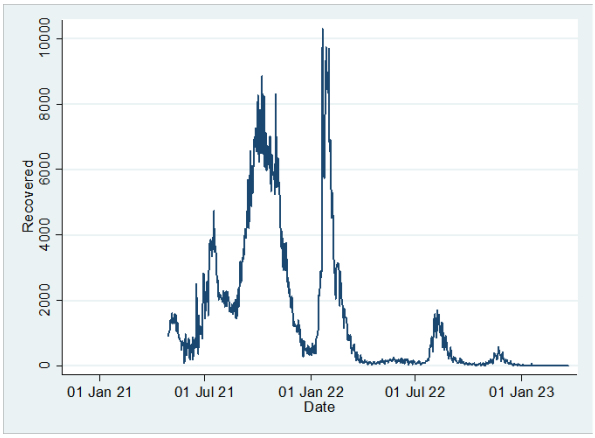

Рисунок 3. Визуализация динамики COVID-19 выздоровления в Монголии

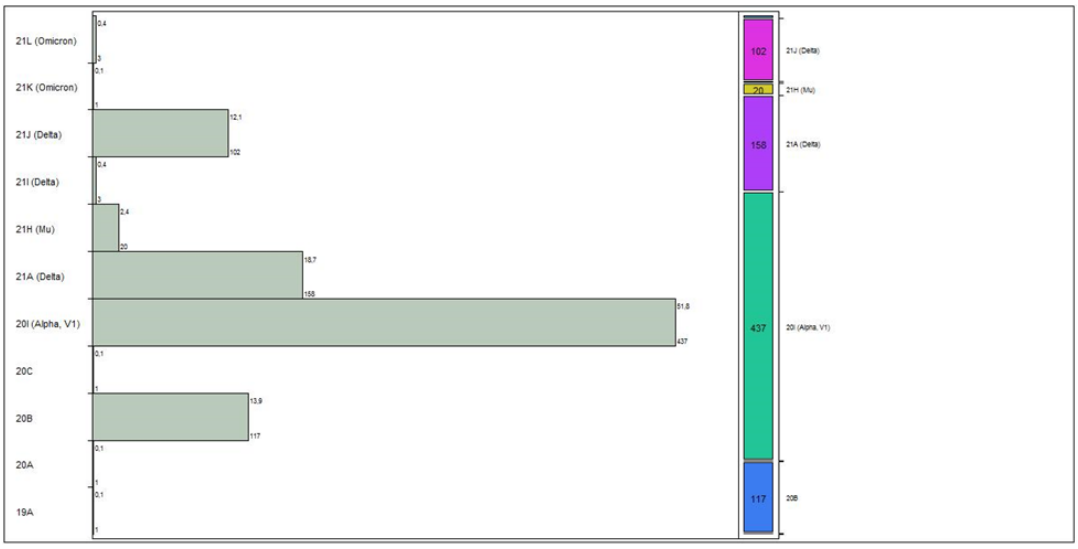

Рисунок 4. Данные Монголии, 844 образца. Показано доминирование *Альфа штамма* коронавируса

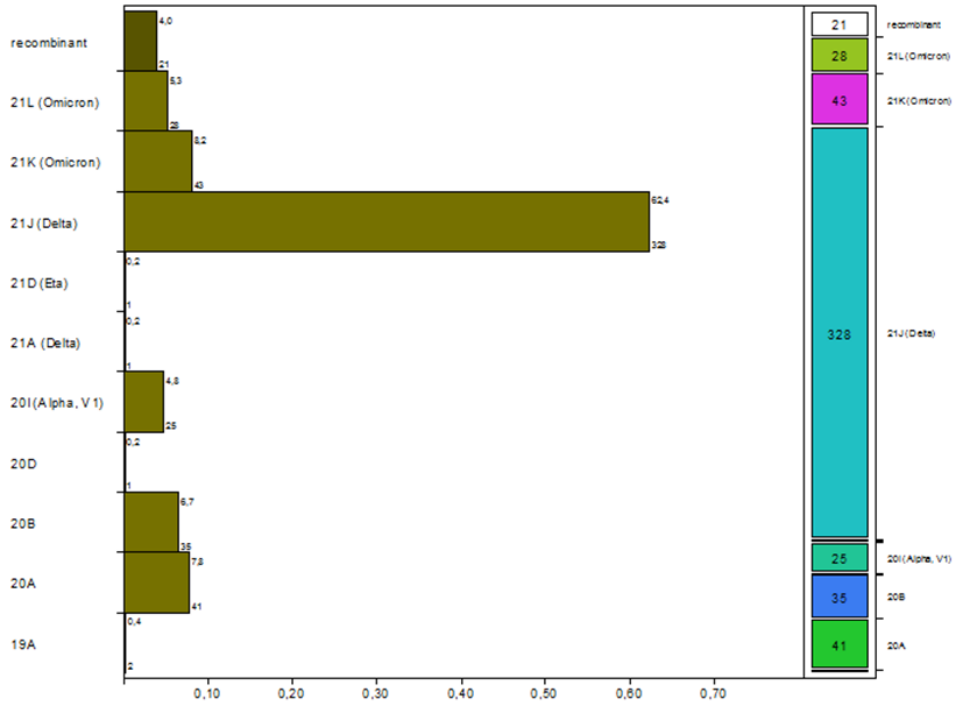

Рисунок 5. Столбчатая диаграмма численного представления результатов идентификации и классификации образцов 526 геномов SARS-CoV-2 из Беларуси. Показано доминирование *Дельта штамма коронавируса*

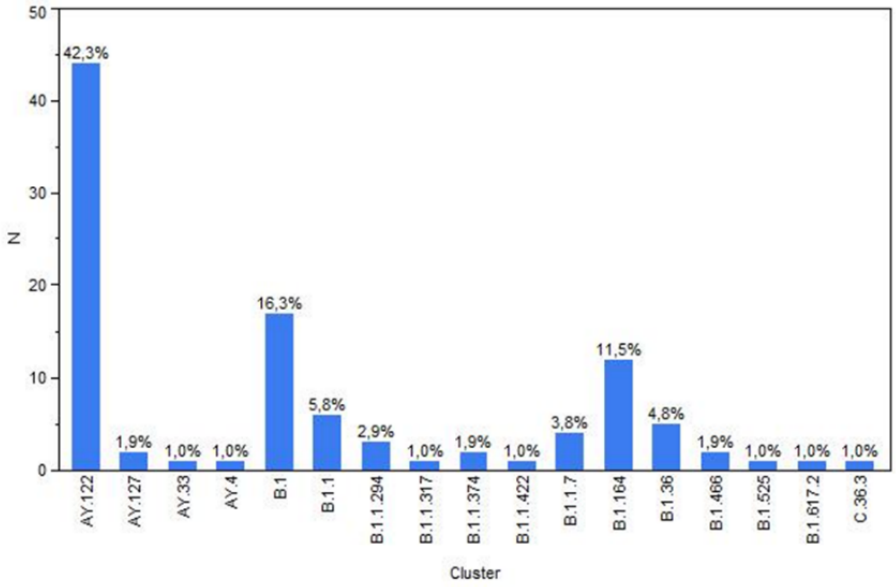

Рисунок 6. Pango статистика линий передачи коронавируса (линия AY.122 названа как «Русский» вариант Дельта-штамма коронавируса, причины распространения точно неизвестны)

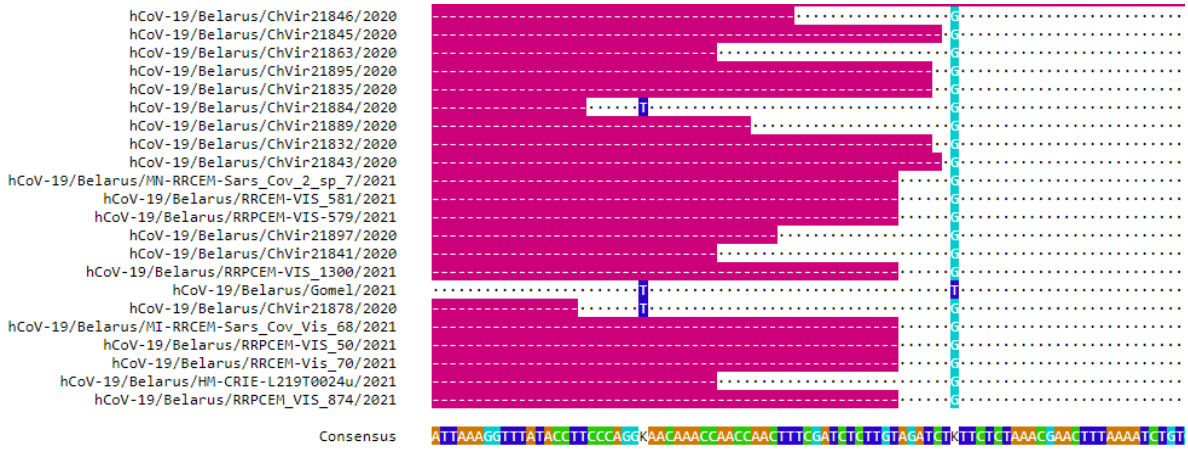

Рисунок 7 – Фрагмент динамической визуализации множественного выравнивания (модуль языка R Decipher)

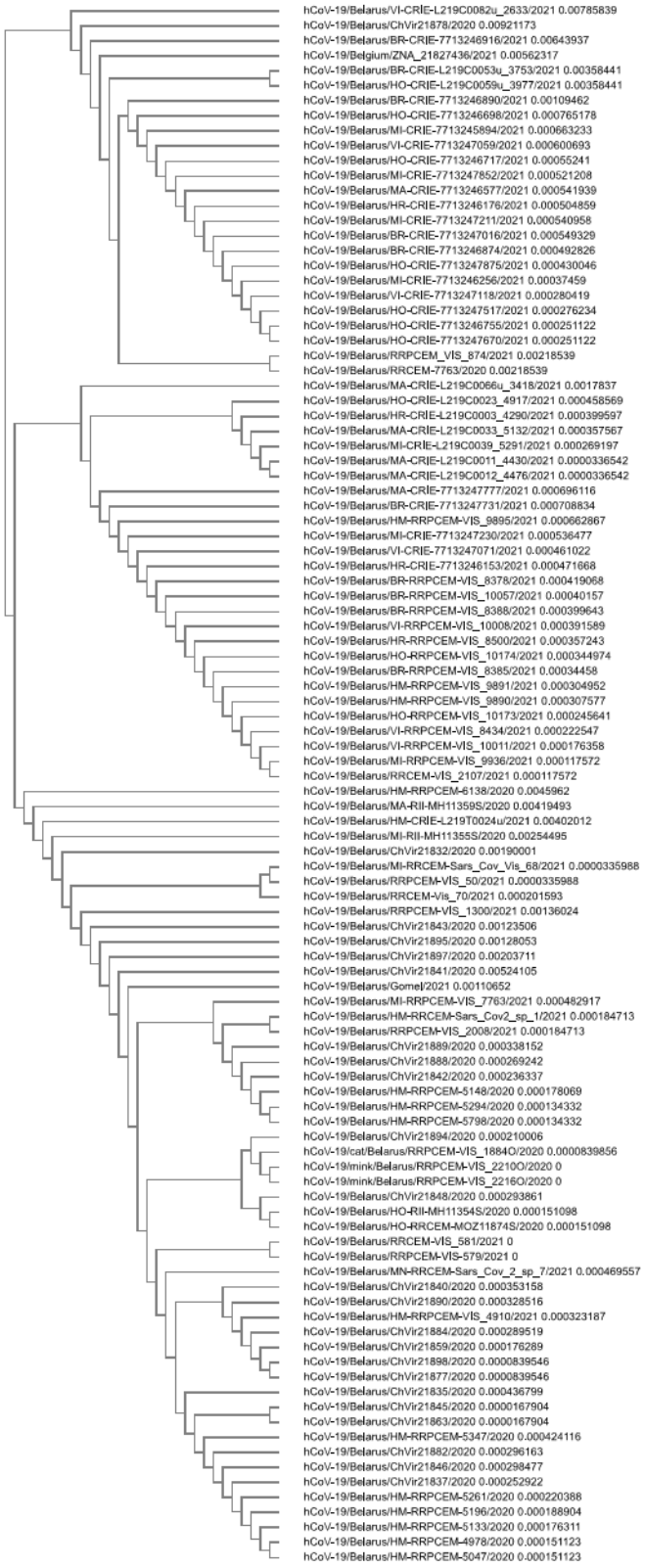

Рисунок 8 Дерево филогенеза 104 геномов коронавируса из Беларуси (модуль Clustal Omega)

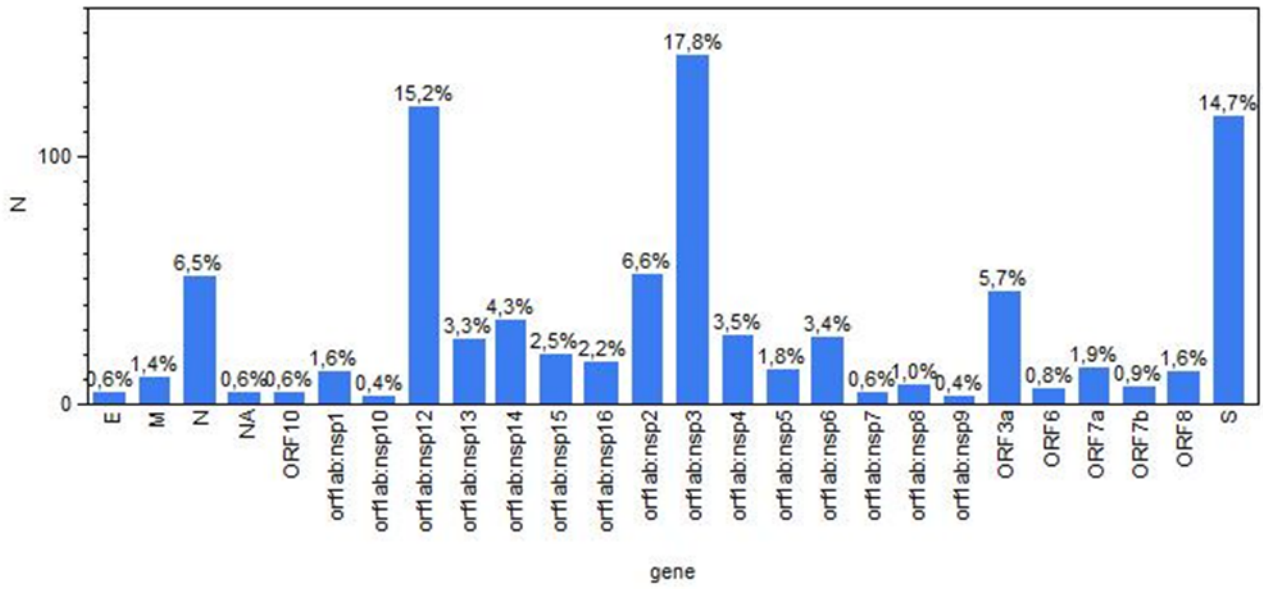

Рисунок 9 Процентное представление частоты мутаций в генах 104 образцов геномов белорусского коронавируса

Заключение.

Вычислительные эксперименты показывают, что геномный профиль коронавируса в Монголии и Беларуси различен, как и эпидемиологические характеристики этой опасной инфекции. Необходимы более крупномасштабные синхронизированные исследования в области геоинформационного изучения эволюции новых опасных вирусов.

Работа выполнена при поддержке проектов БРФФИ: «Математическое моделирование передачи и распространения COVID-19 инфекции на основе систем дифференциальных уравнений и алгоритмов обработки данных с применением технологии машинного обучения» Ф21МН-001, № ГР 20213518 от 27.09.2021; «Ретроспективный анализ клинического и иммунологического статуса групп COVID-19 пациентов с сопутствующим туберкулезом и ВИЧ инфекцией по данным РНПЦ Пульмонологии и фтизиатрии г. Минска», № ГР 20210456 от 31.03.2021; «Разработка и скрининг мукозной вакцины против COVID-19 на основе векторной платформы кишечного аденовируса», № ГР 20210889 от 26.04.2021.

Список литературы

[1] Coronavirus genomics and bioinformatics analysis / P.C. Woo [et al.] // Viruses. – 2010. – V. 2, № 8. – P. 1804-20.

[2] Bioinformatics and evolutionary insight on the spike glycoprotein gene of QX-like and Massachusetts strains of infectious bronchitis virus / S.H. Abro [et al.] // Virol J. – 2012. – V. 9. – P. 211.

[3] Bioinformatics analysis of the factors controlling type I IFN gene expression in autoimmune disease and virus-induced immunity / D. Feng [et al.] // Front Immunol. – 2013. – V. 4. – P. 291.

[4] Proficiency Testing of Virus Diagnostics Based on Bioinformatics Analysis of Simulated In Silico High-Throughput Sequencing Data Sets / A. Brinkmann [et al.] // J Clin Microbiol. – 2019. – V. 57, № 8. – P. e00466-19. doi: 10.1128/JCM.00466-19. Print 2019 Aug.

[5] Virus bioinformatics: databases and recent applications / P. Kellam [et al.] // Appl Bioinformatics. – 2002. – V. 1, № 1. – P. 37-42.

[6] [Sequence analysis for genes encoding nucleoprotein and envelope protein of a new human coronavirus NL63 identified from a pediatric patient in Beijing by bioinformatics] / J.F. Xing [et al.] // Bing Du Xue Bao. – 2007. – V. 23, № 4. – P. 245-51.

[7] Stata time-series : reference manual : release 10 / StataCorp LP. // – Stata Press, 2007. – 343 p.

[8] JMP statistical discovery software / B. Jones [et al.] // Wiley Interdisciplinary Reviews: Computational Statistics. – 2011. – V. 3, № 3. – P. 188-194.

[9] R programming for bioinformatics / R. Gentleman // – CRC, 2009. – 326 p.

[10] Python data analysis / I. Idris // – Place Published: Packt Publishing Ltd, 2014. – 348 p.

[11] Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool / Á. O'Toole [et al.] // Virus Evolution. – 2021. – V. 7, № 2. – P. veab064.

[12] Assessing uncertainty in the rooting of the SARS-CoV-2 phylogeny / L. Pipes [et al.] // Molecular biology and evolution. – 2021. – V. 38, № 4. – P. 1537-1543.

[13] Nextclade: clade assignment, mutation calling and quality control for viral genomes / I. Aksamentov [et al.] // Journal of Open SourceSoftware. – 2021. – V. 6, № 65. – P. 3773.

[14] Алгоритмы обработки геномов коронавируса для целей и задач современной иммуноинформатики, вакциномики и вирусологии / М.В. Спринджук, Владыко, А.С., Титов, Л.П., Чжочжуан, Лу, Берник, В.И. // Цифровая трансформация. – 2022. – V. 22, № 8. – P. 71-81.

BIOINFORMATION ANALYSIS FOR DANGEROUS CORONAVIRUS INFECTION MONITORING

Sprindzuk M.V.
PhD of Technical Sciences
MD (medical doctor)
Senior research scientist
Computer science lecturer
Working on position of associate professor

Department of Economical Informatics
Faculty of Computer Engineering
Belarusian State University of computer science and Radio Electronics, Republic of Belarus
Institute of Mathematics of the NASB
United Institute of Informatics Problems of the NASB
E-mail: stepanenkomatvei@yandex.ru

Abstract. The analysis of bioinformatic data of Mongolia and Belarus has been performed. Waves of mortality and recovery in Mongolia are visualized according to the time series epidemiological data. The profile of genomic variations of coronavirus in Belarus and Mongolia has been investigated.

Keywords: medical cybernetics, coronavirus, epidemiology, medical systems, applied mathematics, monitoring of dangerous infections