

Метод главных компонент в таргетном подходе к определению рода коронавирусов

Чалей М.Б.^{*1}, Кутыркин В.А.^{**2}

¹*Институт математических проблем биологии – филиал ИПМ им. М.В. Келдыша РАН, Пушкино, Московская область, Россия*

²*Московский государственный технический университет им. Н.Э. Баумана, Москва, Россия*

Аннотация. Предложен оригинальный подход к классификации коронавирусов, основанный на представлении анализируемого гена (N-гена белка нуклеокапсида) соответствующим вектором частот кодонов аминокислот и его последующего сравнения с вектором усредненных частот кодонов для аналогичных известных генов вирусного таксона (одного из четырех родов коронавируса). Для определения принадлежности анализируемого вектора частот к каждому рассматриваемому таксону нестандартным образом применяется метод главных компонент. Метод протестирован на 5769 N-генах коронавирусов четырех родов и показал надежность распознавания рода выше 95 %. Предлагаемый подход к классификации коронавирусов позволяет сократить размерность вектора частот кодонов до 28 компонент без снижения надежности, ограничиваясь рассмотрением наиболее значимых частот встречаемости кодонов аминокислот в N-гене. Подход относится к методам без выравнивания, которые в последнее десятилетие завоевывают все большую популярность для классификации вирусов.

Ключевые слова: N-ген, коронавирусы, классификация, методы без выравнивания, метод главных компонент

ВВЕДЕНИЕ

Распространение инфекций, обусловленных вирусами, появление их новых штаммов (вариантов), также как и повторяющиеся вспышки ранее известных вирусных заболеваний (emerging and reemerging diseases) придают особое значение развитию таксономии вирусов, как одному из средств изучения зоонозных очагов вирусов, предсказания свойств и оценки опасности распространения вирусов в окружающей человека среде. Таксономия вирусов способствует осмысленному поиску и разработке препаратов направленного действия против вирусов, проявляющих патогенные свойства. Знание о таксономической группе вируса может быть полезно и при разработке новых вакцин.

В сущности, вирусы являются мобильными генетическими элементами, кодирующими хотя бы один белок капсида (вириона), в котором упакована РНК или ДНК вирусного генома [1]. Единая система классификации и номенклатуры вирусов разрабатывается Международным комитетом по таксономии вирусов (ICTV, International Committee on Taxonomy of Viruses). Исчерпывающий обзор по истории, структуре, принципах работы комитета, достижениях и перспективах представлен в работах [1, 2]. Интересно, что свое сегодняшнее название комитет получил в 1975 году,

*maramaria@yandex.ru

**vkutyarkin@yandex.ru

после переименования Международного комитета по номенклатуре вирусов (ICNV, International Committee on Nomenclature of Viruses), который был основан в 1966 году в Москве во время 9-го Конгресса Международной ассоциации микробиологических сообществ [1]. К настоящему времени для классификации вирусов приняты 15 таксономических рангов, основными среди которых являются: надцарство (Realm), царство (Kingdom), тип (Phylum), класс (Class), отряд (Order), семейство (Family), род (Genus) и вид (Species) [1–3]. Вид является наименьшей таксономической единицей и определяется как монофилетическая группа вирусов с единым защищенным генофондом [4].

Первые методы и подходы к классификации вирусов строились на серологических реакциях. Антигенная группа (серокомплекс) долгое время являлась базовой таксономической единицей для внутриродовой классификации вирусов. Эта классификация и сейчас не утратила своего значения. С развитием молекулярно-генетических методов и электронной микроскопии, таксономия вирусов стала опираться на их физико-химические характеристики (морфология вириона, характеристики ДНК или РНК генома, способ репликации, наличие липидной оболочки и др.) и эволюционную филогению [1].

По мере совершенствования методов секвенирования и биоинформатики в качестве критериев разделения видов, родов и семейств было предложено использовать математические методы, учитывающие количественные генетические различия между вирусами или их группами, определяемые в попарном сравнении геномов [5–8]. Такие методы особенно актуальны в эпоху метагеномики, когда необходимо классифицировать новые вирусные штаммы, о которых нет иной информации, кроме последовательности генома. Для разделения видов, родов и семейств анализируются гены и аминокислотные последовательности консервативных вирусных белков, например РНК-зависимой РНК-полимеразы (RdRp) или белка нуклеокапсида [9, 10].

В работе [11] выделены главные подходы к таксономическому разделению анализируемых вирусных геномов, и рассмотрены различные компьютерные программы для их классификации. Первый подход учитывает изменения в известных паттернах нуклеотидных или аминокислотных последовательностей вирусов (VConTACT [12]). Второй рассматривает гомологию соответствующих генов или белков (GRAViTy [13], SDT [14]). Третий подход опирается на филогению (например, APCluster [15], VICTOR [16]). И, наконец, четвертый подход использует методы выравнивания последовательностей и определение попарного генетического расстояния (DEmARC [17], MyCoV [18], и др.). Как правило, существующие компьютерные программы используют комбинации этих подходов. Программы также различаются и по возможности выделения иерархических уровней таксонов вирусов.

В связи с огромным разнообразием вирусов и стремительным ростом метагеномных данных, некоторые исследователи полагают [19], что для разграничения таксономических уровней вирусов следует рассматривать и 3D структуры вирусных белков, которые генерируются искусственным интеллектом, например такими программами, как AlphaFold [20, 21] и ESMFold [22].

На сегодняшний день общепринятой практикой в построении таксономии вирусов является предварительное множественное выравнивание последовательностей, на основе которого выполняется филогенетический анализ (MAFFT [23], IQ-TREE [24], ClustalOmega [25]).

Сканирование баз данных с целью выявления наибольшей близости с известными референсными последовательностями, множественное выравнивание и построение филогении для определения таксонов, к которым относятся новые вирусы, в эпоху метагеномных исследований становится все сложнее по причине стремительно растущего числа секвенируемых геномов. Поэтому все больше исследователей обращаются к созданию методов выведения филогении без предварительного

выравнивания нуклеотидных или аминокислотных последовательностей. Это так называемые alignment-free методы (свободные от выравнивания, или методы без выравнивания). Среди значительных преимуществ этих методов – их устойчивость к рекомбинациям и существенно более высокая скорость исполнения программы классификации в сравнении с классическим путем выведения филогении методом максимального правдоподобия [26]. Методы без выравнивания в последнее десятилетие завоевывают все большую популярность. Сравнительный обзор этих методов и характеристик, которые они используют для оценки попарного расстояния, можно найти в [27]. Например, разработчики компьютерной программы SAM [28] вводят попарное расстояние между вирусными геномами на основе частоты пересечения списков кодонов, которые не используются в генах двух сравниваемых вирусов. Основная идея методов без выравнивания заключается в том, чтобы каждый анализируемый геном представить отдельным вектором характеристик в n -мерном векторном пространстве и затем использовать алгоритмы классификации или модели нейронных сетей [29, 30].

Метод, предлагаемый в настоящей работе для классификации коронавирусов по родам, также можно отнести к методам без выравнивания, использующим предварительно известные данные для обучения.

Ранее для распознавания рода коронавирусов нами был предложен типологический подход [31], основанный на усреднении частот кодонов в генах прототипных штаммов рода. Прототипные штаммы – это своего рода образцы или типы видов вирусов, которые выделены и поддерживаются в лабораториях, и обладают неизменными свойствами. Такие типы служат для сравнения с другими вирусами. В отличие от прототипных штаммов, вирусы в природе быстро изменяются и приобретают новые свойства, как, например, это происходило с вирусом SARS-CoV-2, вызвавшего эпидемию ковида, сопровождавшуюся быстрой сменой штаммов (вариантов): альфа, бета, гамма, дельта и омикрон [32–35].

Типологический подход дал хорошие результаты при использовании различных комбинаций генов коронавирусов в определении рода [31]. Однако он показал существенные ошибки при использовании только одного, достаточно консервативного и относительно короткого, N-гена белка нуклеокапсида. Хотя, ранее N-ген успешно использовался и для построения филогенетических дендрограмм при изучении географии распространения коронавирусов [36], и в молекулярно-эпидемиологическом анализе вариантов вируса SARS-CoV-2 [37].

Чтобы исправить значительные погрешности в распознавании рода на основе N-гена, в дальнейшем мы использовали типологический подход, опирающийся на 67 индивидуальных прототипных штаммов подродов без их усреднения по родам [38]. В результате достоверность распознавания подрода и, соответственно, рода, составляла 99 %. Заметим, что в рамках типологического подхода при распознавании использовался один вид статистики, применение которой для усредненных характеристик рода оказалось неэффективным.

В настоящей работе для распознавания рода применяется новый таргетный (на основе усредненных характеристик N-гена) подход и предлагаются другие статистики, особым образом использующие известный метод главных компонент. Обычно, целью этого метода является выявления главных компонент со значительными дисперсиями в выборке многомерных данных одного типа. С помощью главных компонент можно выделить факторы, которые вносят наибольший разброс в экспериментальные данные.

Кратко опишем метод главных компонент для выборки данных одного типа (определяемого родом коронавируса). Для каждой компоненты многомерных векторов (данных одинаковой размерности) вычисляется ее среднее значение. Затем из каждой компоненты многомерного вектора рассматриваемой выборки сначала вычитается её среднее значение, и полученный результат делится на это среднее значение. В

результате такого преобразования создается выборка трансформированных векторов, с нулевым средним значением каждой компоненты. После чего для трансформированной многомерной выборки стандартным образом вычисляется ковариационная матрица C .

Для получения главных компонент матрица C представляется в виде $C = Q \cdot D \cdot Q^T$, где Q – ортогональная матрица, столбцы которой являются собственными векторами матрицы C , D – диагональная матрица, на диагонали которой стоят собственные значения матрицы C , являющиеся дисперсиями соответствующих главных компонент. Кроме того, матрица Q^T – матрица преобразований к главным компонентам трансформированных многомерных векторов выборки.

В предлагаемом нами подходе к распознаванию рода коронавируса метод главных компонент применяется к выборкам многомерных данных (векторов одинаковой размерности), но относящихся к различным типам, соответствующих четырем родам коронавирусов: *Alphacoronavirus* (α -CoV), *Betacoronavirus* (β -CoV), *Deltacoronavirus* (δ -CoV) и *Gammacoronavirus* (γ -CoV). Анализируемый вектор представляет N-ген отдельного коронавируса. Компонентами вектора служат частоты кодонов аминокислот, встречающихся в рассматриваемом N-гене. Из всех 64 кодонов генетического кода учитываются только 59 кодонов. Кодоны с минимальными частотами встречаемости, в частности стоп-кодоны, не рассматриваются. Для каждой выборки векторов частот, соответствующей одному роду, определяются свои главные компоненты и их дисперсии. Кроме того, для рода фиксируется свой способ их вычисления. Для вычисления главных компонент и дисперсий по каждому роду создавались обучающие выборки векторов частот, общим числом 2113. Для тестирования и демонстрации эффективности нового подхода к распознаванию рода коронавируса из базы GenBank [39] были выбраны 4057 N-генов коронавирусов с известным родом. При этом обучающая и тестируемая выборки пересекались не более чем на 15 %.

Распознавание типа (рода коронавируса) анализируемого многомерного вектора частот использует следующую процедуру. Для каждого предполагаемого типа данных по анализируемому вектору вычисляются (с помощью четырех фиксированных способов) возможные значения главных компонент и их дисперсии. Кроме того, квадрат вычисленного значения каждой компоненты (в фиксированном способе вычислений) делится на соответствующую ей дисперсию. После чего, полученные величины усредняются. Таким образом, для анализируемого вектора частот получаются четыре числа, соответствующие каждому из четырех способов вычислений. Минимальное число, относящееся к одному из способов, определяет тип данных. В работе показана значительная эффективность (практически 100 %) предлагаемого метода.

Поясним эффективность предлагаемого метода распознавания в предположении, что выборка одного типа представлена реализациями некоторого своего нормального распределения многомерной случайной величины. Тогда для выборки данных одного типа усредненное значение нормированных (на соответствующие дисперсии) главных компонент близко к единице. Для разных типов данных способы вычисления их усредненных значений отличаются. Если для векторов частот одного типа данных применить способ вычисления усредненных значений данных другого типа, то при существенном различии нормальных распределений, характеризующих тип данных, возможно значительное превышение теоретически ожидаемых единичных значений. В работе продемонстрировано, что такое предположение о существенном различии нормальных распределений имеет место. На этом факте различия и основан предлагаемый метод распознавания рода коронавирусов.

Предлагаемый в настоящей работе таргетный подход разделения коронавирусов по родам с помощью метода главных компонент позволяет существенно, практически в 2

раза сократить размерность анализируемых векторов частот кодонов с 59 до 28 или 30, при условии, что частоты ранжированы по убыванию. При сокращении размерности векторов подход гарантирует достоверность определения рода не менее 95 %.

МЕТОД

Рассматриваются нуклеотидные последовательности N-гена белка нуклеокапсида из геномов коронавирусов четырех родов (α -CoV, β -CoV, δ -CoV и γ -CoV), полученные из базы данных GenBank [39]. Гены каждого рода представлены соответствующей отдельной выборкой.

Для каждого гена вычисляются частоты 64 кодонов аминокислот. Редко встречающиеся у всех генов кодоны (одни и те же в четырех выборках) исключаются из рассмотрения. Таким образом, при распознавании каждый ген характеризуется вектором частот рассматриваемых кодонов в виде $\mathbf{P} = (P_1, P_2, \dots, P_n)^T$, где P_i – частота встречаемости i -го кодона ($i = \overline{1, n}$), n – количество рассматриваемых кодонов.

Пусть $X \in \{\alpha, \beta, \delta, \gamma\}$ – индекс рода коронавируса. Для каждого рода по всем генам его выборки рассчитываются средние частоты встречаемости каждого рассматриваемого кодона. На их основе создается вектор средних частот $\overline{\mathbf{P}}^X = (\overline{P}_1^X, \overline{P}_2^X, \dots, \overline{P}_n^X)^T$. Для каждого вектора частот $\mathbf{P} = (P_1, P_2, \dots, P_n)^T$ из выборки всех векторов частот рода X вычисляется трансформированный вектор $\mathbf{p} = \left(\frac{P_1 - \overline{P}_1^X}{\overline{P}_1^X}, \frac{P_2 - \overline{P}_2^X}{\overline{P}_2^X}, \dots, \frac{P_n - \overline{P}_n^X}{\overline{P}_n^X} \right)^T = \mathbf{F}_X(\mathbf{P})$, где \mathbf{F}_X – операция X -трансформирования вектора частот. Следует подчеркнуть, что такая операция X - трансформирования может применяться и к векторам частот другого рода.

Для выборки векторов частот из рода X с помощью операции \mathbf{F}_X создается соответствующая выборка трансформированных векторов. На основе этой выборки вычисляется ковариационная матрица \mathbf{C}_X для рода X . Для получения главных компонент рода X матрица \mathbf{C}_X представляется в виде $\mathbf{C}_X = \mathbf{Q}_X \cdot \mathbf{D}_X \cdot \mathbf{Q}_X^T$, где $\mathbf{Q}_X = (q_1^X, q_2^X, \dots, q_n^X)$ – ортогональная матрица, столбцы которой являются собственными векторами матрицы \mathbf{C}_X , т.е. $\mathbf{C}_X \cdot \mathbf{q}_i^X = d_i^X \mathbf{q}_i^X$ для $i = \overline{1, n}$;

$$\mathbf{D}_X = \begin{pmatrix} d_1^X & 0 & \dots & 0 \\ 0 & d_2^X & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & d_n^X \end{pmatrix} \text{ – диагональная матрица.}$$

На диагонали матрицы \mathbf{D}_X стоят собственные значения матрицы \mathbf{C}_X , являющиеся дисперсиями соответствующих главных компонент. Кроме того, \mathbf{Q}_X^T – транспонированная матрица (для \mathbf{Q}_X), являющаяся матрицей преобразований к главным компонентам трансформированных многомерных векторов рода X .

Опишем процедуру распознавания рода анализируемого вектора частот $\mathbf{P} = (P_1, P_2, \dots, P_n)^T$ на основе предлагаемого метода главных компонент. Для каждого анализируемого вектора частот вычисляется вектор $\mathbf{y}^X = (y_1^X, y_2^X, \dots, y_n^X)^T = \mathbf{Q}_X^T \cdot \mathbf{F}_X(\mathbf{P})$ в предположении, что он является выборочным вектором главных компонент рода X ($X \in \{\alpha, \beta, \delta, \gamma\}$). Для этого вектора вычисляется значение $z^X = \frac{1}{n} \sum_{i=1}^n \frac{(y_i^X)^2}{d_i^X}$.

Минимальное число среди $z^\alpha, z^\beta, z^\gamma$ и z^δ указывает на род анализируемого вектора частот. Способ такого выбора объясним с помощью следующих рассуждений.

Если многомерный вектор частот $P = (P_1, P_2, \dots, P_n)^T$ принадлежит к выборке из рода X , представляющей реализации нормального распределения, то рассчитанные значения $\frac{y_1^x}{\sqrt{d_1^x}}, \frac{y_2^x}{\sqrt{d_2^x}}, \dots, \frac{y_n^x}{\sqrt{d_n^x}}$ являются реализациями одномерной случайной величины с нулевым матожиданием и единичной дисперсией. Поэтому среднее значение их квадратов z^x близко к единице. В противном случае, когда вектор $P = (P_1, P_2, \dots, P_n)^T$ не принадлежит к выборке из рода X , такие средние значения, рассчитанные с помощью трансформации F_x и матрицы ковариации C_x , как показало практическое исследование, значительно превышают единицу. Таким образом, проделанная работа показала эффективность предложенного метода распознавания.

МАТЕРИАЛЫ

Обучающая и тестируемая выборки N-генов коронавируса

Из базы данных GenBank [39] были получены две выборки полных геномов коронавируса четырех родов (*Alphacoronavirus* (α -CoV), *Betacoronavirus* (β -CoV), *Deltacoronavirus* (δ -CoV) и *Gammacoronavirus* (γ -CoV)), накопленных в базе к концу 2025 года. Число индивидуальных геномов в обеих выборках составило 5769. Из этих геномов были выделены кодирующие последовательности N-генов, из которых формировались обучающая и тестируемая выборки, таким образом, чтобы для каждого рода коронавируса имелись представители всех подродов. В обучающую выборку вошли только те N-гены, для геномов которых был известен как род, так и подрод коронавируса. Тестируемая выборка содержала N-гены коронавируса известного рода, но необязательно с определенным подродом. В таблице 1 представлено количество N-генов в обучающей и тестируемой выборках для каждого из четырех родов, также как и количество генов общих для обеих выборок. Как можно видеть из таблицы 1, тестируемая выборка, в среднем, на 90 % отличалась от обучающей выборки.

Таблица 1. Количественные и качественные характеристики обучающей и тестируемой выборки N-генов коронавируса

| Род коронавируса | Обучающая выборка | Тестируемая выборка | Количество общих N-генов между выборками | Доля новых N-генов в тестируемой выборке |
|------------------|-------------------|---------------------|--|--|
| α -CoV | 251 | 1281 | 89 | 93 % |
| β -CoV | 1647 | 2126 | 251 | 88 % |
| δ -CoV | 80 | 244 | 8 | 97 % |
| γ -CoV | 135 | 406 | 53 | 87 % |

Формирование векторов частот кодонов различной размерности

Как было описано в разделах Введение и Метод, каждому анализируемому N-гену сопоставлялся вектор частот рассматриваемых кодонов аминокислот. Максимальное число рассматриваемых кодонов $n = 59$, так как кодоны с самыми низкими частотами исключались из рассмотрения. Ранее в работе [38] были рассчитаны средние частоты встречаемости кодонов в N-генах из 67 прототипных штаммов подродов для всех четырех родов коронавируса. Список кодонов, упорядоченных по убыванию частоты встречаемости в N-генах этих прототипных штаммов, представлен в таблице 2. Три стоп-кодона (taa, tga и tga) и два кодона цистеина (tgc, tgt) не вошли в этот список по

причине слишком малой частоты. В настоящей работе для каждого анализируемого N-гена все компоненты векторов частот кодонов, рассматриваемых при распознавании, были упорядочены согласно списку кодонов в таблице 2. Кроме того, при фиксированной размерности $n = \overline{7,59}$ пространства рассматриваемых векторов частот вида $P = (P_1, P_2, \dots, P_n)^T$, номера компонент соответствуют первым n кодонам в таблице 2.

Таблица 2. Список кодонов аминокислот в соответствии с убыванием их частоты встречаемости в N-генах 67 прототипных штаммов подродов коронавируса (см. текст)

| | | | | | | | | | | | | | | |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
| aag | gat | aaa | gct | ggt | aat | cct | caa | cca | tct | cag | act | gga | gaa | aga |
| 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 |
| gca | gtt | gac | ttt | cgt | tca | aac | att | ctt | gag | aca | ggc | tgg | agt | gcc |
| 31 | 32 | 33 | 34 | 35 | 36 | 37 | 38 | 39 | 40 | 41 | 42 | 43 | 44 | 45 |
| ttc | atg | tat | ccc | cgc | acc | tac | gtg | ttg | gtc | cat | agc | agg | tcc | gta |
| 46 | 47 | 48 | 49 | 50 | 51 | 52 | 53 | 54 | 55 | 56 | 57 | 58 | 59 | |
| ctg | ggg | atc | ctc | cta | gcg | ccg | ata | cga | cac | tta | tcg | acg | cgg | |

В настоящей работе проводилось исследование возможности оптимизации («сокращения») размерности пространства рассматриваемых векторов частот кодонов для достоверного распознавания рода коронавируса с помощью предлагаемого метода главных компонент. Проведенные испытания позволили выявить размерность пространства рассматриваемых векторов частот кодонов значительно меньшее 59 для достаточной надёжности (на уровне более 95 %) распознавания рода коронавируса.

Предлагаемый подход использовал обучающие выборки N-генов четырех родов X ($X \in \{\alpha, \beta, \delta, \gamma\}$) для создания матриц Q_X^T преобразований к главным компонентам трансформированных многомерных векторов рода X . Эти матрицы использовались для распознавания рода коронавируса как в обучающей, так и в тестируемой выборках.

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Результаты распознавания рода коронавируса в обучающей выборке

В качестве обучающей выборки были выбраны N-гены коронавируса с известным родом и подродом. Количественный состав этой выборки по родам показан в таблице 1. Обучающая выборка для каждого рода формировалась таким образом, чтобы в ней были представители всех подродов. Число N-генов с известным подродом в выборке определялось наличием геномов коронавируса этих подродов в базе GenBank [39] и, зачастую значительно отличалось между подродами.

Согласно описанному выше методу, для каждого рода X , где $X \in \{\alpha, \beta, \delta, \gamma\}$ по многомерным векторам $P = (P_1, P_2, \dots, P_n)^T$ частот кодонов (одинаковой размерности n) в N-генах обучающей выборки строились оператор трансформации F_X и матрица ковариации C_X , затем рассчитывались матрица D_X , на диагонали которой стоят дисперсии главных компонент, и ортогональная матрица Q_X^T преобразований к главным компонентам трансформированных многомерных векторов рода X . Отметим, что построение операторов трансформации векторов и вычисление указанных матриц выполнялось отдельно для каждой размерности n ($n = \overline{7,59}$) пространства рассматриваемых векторов частот кодонов.

Для каждого гена (представленного n -мерным вектором частот кодонов) из общей обучающей выборки N -генов всех родов были вычислены, в соответствии с методом, величины $z^\alpha, z^\beta, z^\gamma$ и z^δ . Среди них выбиралась минимальная величина, определяющая род. Для обучающей выборки результаты надежного (на уровне более 95 %) применения предлагаемого метода распознавания рода в зависимости от использования размерности $n = \overline{7,59}$ рассматриваемых векторов частот кодонов представлены на рисунке 1. Количество N -генов для каждого из четырех родов коронавирусов в обучающей выборке указано в таблице 1. Рисунок 1 показывает долю N -генов с правильно распознанным родом в зависимости от размерности $n = \overline{7,59}$ рассматриваемых векторов частот кодонов. Как можно видеть из рисунка 1, уровень надежности 95 % в распознавании всех четырех родов коронавируса достигался, начиная с использования размерности $n = 12$ рассматриваемых векторов частот. Напомним, что все компоненты векторов частот, соответствующие кодонам, были упорядочены так, как показано в таблице 2.

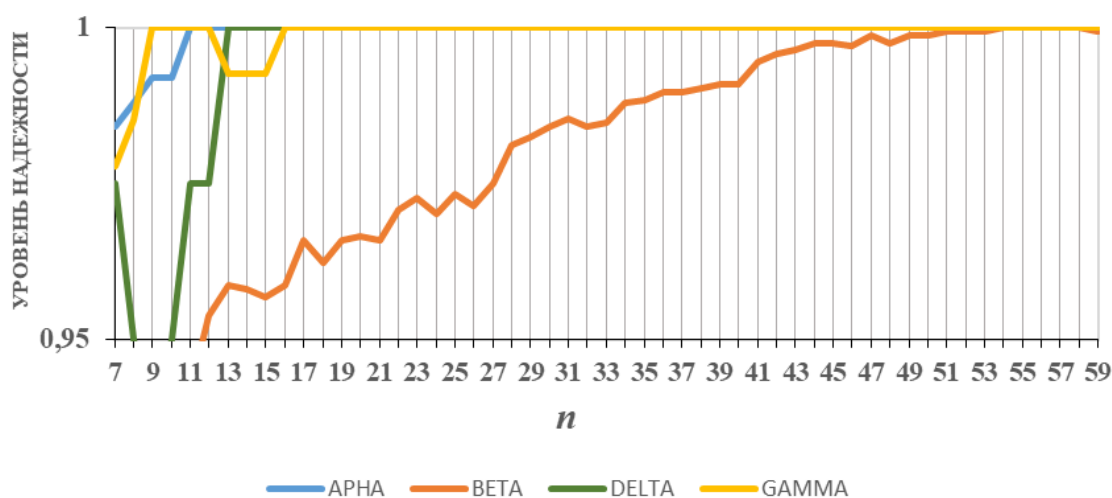


Рис. 1. Доля (на уровне выше 0.95) N -генов из обучающей выборки с правильно распознанным родом коронавирусов (*Alphacoronavirus*, *Betacoronavirus*, *Deltacoronavirus*, *Gammacoronavirus*) показана в зависимости от размерности n пространства рассматриваемых векторов частот кодонов, используемых при распознавании. Количество N -генов в обучающей выборки по родам представлено в таблице 1.

Результаты распознавания рода коронавирусов в тестируемой выборке

Тестируемая выборка N -генов коронавирусов с известным родом в два раза превышала по численности обучающую выборку (см. таблицу 1) и, в среднем по родам, на 91 % состояла из новых генов.

Вектор частот кодонов $\mathbf{P} = (P_1, P_2, \dots, P_n)^T$ для каждого N -гена тестируемой выборки проходил испытание в предположении, что N -ген (как представитель генома коронавируса) принадлежит роду X -CoV, где $X \in \{\alpha, \beta, \delta, \gamma\}$. В соответствии с предполагаемым родом коронавируса, по многомерным векторам частот кодонов в N -генах обучающей выборки этого рода строился оператор трансформации \mathbf{F}_X (см. Метод). Далее для каждого анализируемого вектора частот тестируемой выборки вычислялся вектор $\mathbf{y}^X = (y_1^X, y_2^X, \dots, y_n^X)^T = \mathbf{Q}_X^T \cdot \mathbf{F}_X(\mathbf{P})$ в предположении, что он является выборочным вектором главных компонент рода X -CoV ($X \in \{\alpha, \beta, \delta, \gamma\}$), где \mathbf{Q}_X^T – ортогональная матрица преобразований к главным компонентам трансформированных

многомерных векторов рода X-CoV, рассчитанная по обучающей выборке согласно методу. При вычислении значений $z^X = \frac{1}{n} \sum_{i=1}^n \frac{(y_i^X)^2}{d_i^X}$ также использовались элементы d_i^X ($i = \overline{1, n}$) соответствующей диагональной матрицы D_X , вычисленной по обучающей выборке. Минимальная из величин $z^\alpha, z^\beta, z^\gamma$ и z^δ определяла род анализируемого N-гена (представленного n -мерным вектором частот кодонов) тестируемой выборки. Таким образом, распознавание рода X-CoV для каждого отдельного вектора частот N-гена тестируемой выборки происходило путем последовательных предположений о его принадлежности к одному из четырех родов коронавирусов (α -CoV, β -CoV, δ -CoV и γ -CoV). Такое распознавание использовало соответствующие операторы трансформации, диагональную матрицу для дисперсий главных компонент и ортогональную матрицу преобразований к главным компонентам трансформированных многомерных векторов, рассчитанных для рода X-CoV по обучающей выборке.

Рисунок 2 показывает долю N-генов (соответствующих рассматриваемым векторам частот кодонов в пространстве размерности n) тестируемой выборки с правильно распознанным родом в зависимости от размерности векторов частот, используемых при распознавании. Напомним, что все n ($n = \overline{7, 59}$) компонент в рассматриваемых векторах частот кодонов были упорядочены в соответствии с таблицей 2. Как можно видеть из рисунка 2, уровень надежности 95 % в распознавании всех четырех родов коронавируса достигался при использовании размерности $n = 28$ и более для пространства рассматриваемых векторов частот кодонов.

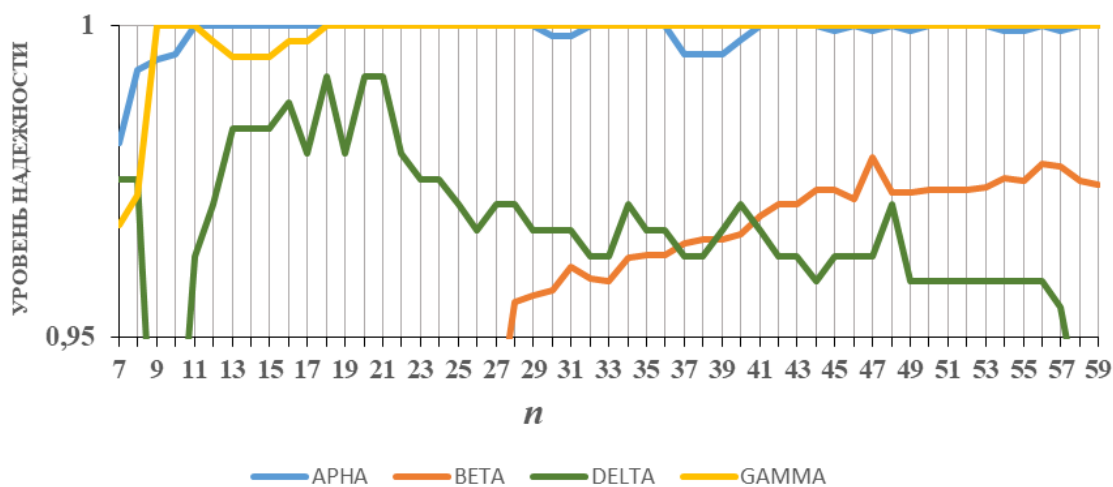


Рис. 2. Доля (на уровне выше 0,95) N-генов из тестируемой выборки с правильно распознанным родом коронавируса (*Alphacoronavirus*, *Betacoronavirus*, *Deltacoronavirus*, *Gammacoronavirus*) показана в зависимости от размерности n пространства рассматриваемых векторов частот кодонов, используемых при распознавании. Количество N-генов тестируемой выборки по родам представлено в таблице 1.

Ранее в работе [38] для распознавания подрода (и рода) с помощью векторов частот кодонов, соответствующих N-генам 67 прототипных штаммов для 23 подродов, применение статистики отклонения анализируемого вектора частот от вектора частот кодонов, усредненных по подкладам, не дало достоверных результатов распознавания. Поэтому для распознавания подрода коронавирусов использовался типологический подход, в котором рассматривалась статистика отклонения анализируемого вектора частот кодонов от вектора частот каждого отдельного прототипного штамма. Это позволило получить практически 100 % результат распознавания подрода (и рода) и

также сократить размерность пространства рассматриваемых векторов частот кодонов с 59 до 38.

В настоящей работе использование нестандартным образом метода главных компонент позволило отказаться от типологического подхода к определению рода коронавирусов и выполнить классификацию вирусных N-генов по родам более простым путем усреднения уже известных таксономических данных. Такой подход к таксономической классификации вирусов представляется нам весьма перспективным в силу своей простоты, эффективности и скорости выполнения. Он не требует никаких преобразований геномной последовательности или переходу к последовательности аминокислотной, свободен от процедуры выравнивания, опирающейся на специальные модели эволюции генома. Если авторы работы [40] полагают, что РНК-вирусы из-за высокой скорости мутаций лучше классифицируются по аминокислотным последовательностям, чем по нуклеотидным, то наша работа на примере коронавирусов, которые также относятся к этой группе, показывает, что вопрос эффективной классификации остается вопросом методологии.

Помимо преимуществ в классификации вирусов, которые имеют методы без выравнивания, применяемый нами таргетный подход на основе N-гена коронавирусов, представленного вектором частот кодонов, позволяет сокращение числа рассматриваемых кодонов до 28 (см. рис. 2) или 38 (как было получено ранее при типологическом подходе). В этом случае достоверность распознавания для всех родов коронавирусов составляет не менее 95 %. Можно полагать, что список из 38 кодонов (см. табл. 2), встречающихся в N-гене коронавирусов, является надежным для классификации коронавирусов по родам, как получивший подтверждение в двух различных методах.

ЗАКЛЮЧЕНИЕ

Использование N-гена в качестве таргета вместе с нестандартным применением метода главных компонент для классификации на его основе, способствовало получению достоверных результатов в определении рода коронавируса. Кроме того, благодаря такому подходу, размерность вектора частот кодонов N-гена в процедуре распознавания рода может быть сокращена до 28 компонент. При этом надежность распознавания не менее 95 % достигается для всех четырех родов коронавирусов. Предлагаемый в работе метод не требует выравнивания последовательностей и опирается на усредненные характеристики (частоты встречаемости кодонов в генах) известных таксономических данных, что делает его весьма простым и эффективным для классификации вирусов.

СПИСОК ЛИТЕРАТУРЫ

1. Львов Д.К., Акимкин В.Г., Забережный А.Д., Борисевич С.В., Альховский С.В. Таксономия и мегатаксономия вирусов (домен *Vira*) – текущий статус. *Вопросы вирусологии*. 2025. Т. 70. № 5. С. 401–416. doi: [10.36233/0507-4088-344](https://doi.org/10.36233/0507-4088-344)
2. Siddell S.G., Smith D.B., Adriaenssens E., Alfenas-Zerbini P., Dutilh B.E., Garcia M.L., Junglen S., Krupovic M., Kuhn J.H., Lambert A.J., et al. Virus taxonomy and the role of the International Committee on Taxonomy of Viruses (ICTV). *J. Gen. Virol.* 2023. V. 104. No. 5. Article No. 001840. doi: [10.1099/jgv.0.001840](https://doi.org/10.1099/jgv.0.001840)
3. Gorbalenya A.E., Krupovic M., Mushegian A., Kropinski A.M., Siddell S.G., Varsani A., Adams M.J., Davison A.J., Dutilh B.E., Harrach B. et al. The new scope of virus taxonomy: partitioning the virosphere into 15 hierarchical ranks. *Nat. Microbiol.* 2020. V. 5. P. 668–674. doi: [10.1038/s41564-020-0709-x](https://doi.org/10.1038/s41564-020-0709-x)
4. Львов Д.К., Гулюкин М.Ю., Забережный А.Д., Гулюкин А.М. Формирование популяционного генофонда потенциально угрожающих биобезопасности

- зоонозных вирусов. *Вопросы вирусологии*. 2020. Т. 65. № 5. С. 243–258. doi: [10.36233/0507-4088-2020-65-5-1](https://doi.org/10.36233/0507-4088-2020-65-5-1)
5. Lauber C., Gorbalenya A.E. Genetics-based classification of filoviruses calls for expanded sampling of genomic sequences. *Viruses*. 2012. V. 4. No. 9. P. 1425–1437. doi: [10.3390/v4091425](https://doi.org/10.3390/v4091425)
 6. Bao Y., Chetvernin V., Tatusova T. Improvements to pairwise sequence comparison (PASC): a genome-based web tool for virus classification. *Arch. Virol.* 2014. V. 159. P. 3293–3304. doi: [10.1007/s00705-014-2197-x](https://doi.org/10.1007/s00705-014-2197-x)
 7. Simmonds P. Methods for virus classification and the challenge of incorporating metagenomic sequence data. *J. Gen. Virol.* 2015. V. 96. No. 6. P. 1193–1206. doi: [10.1099/jgv.0.000016](https://doi.org/10.1099/jgv.0.000016)
 8. Aiewsakun P., Simmonds P. The genomic underpinnings of eukaryotic virus taxonomy: creating a sequence-based framework for family-level virus classification. *Microbiome*. 2018. V. 6. Article No. 38. doi: [10.1186/s40168-018-0422-7](https://doi.org/10.1186/s40168-018-0422-7)
 9. Wilkinson D.A., Joffrin L., Lebarbenchon C., Mavingui P. Analysis of partial sequences of the RNA-dependent RNA polymerase gene as a tool for genus and subgenus classification of coronaviruses. *J. Gen. Virol.* 2020. V. 101. No. 12. P. 1261–1269. doi: [10.1099/jgv.0.001494](https://doi.org/10.1099/jgv.0.001494)
 10. Tang X., Shang J., Sun Y. RdRp-based sensitive taxonomic classification of RNA viruses for metagenomic data. *Brief. Bioinform.* 2022. V. 23. No. 2. Article No. bbac011. doi: [10.1093/bib/bbac011](https://doi.org/10.1093/bib/bbac011)
 11. Gorbalenya A.E., Lauber C. Bioinformatics of virus taxonomy: foundations and tools for developing sequence-based hierarchical classification. *Curr. Opin. Virol.* 2022. V. 52. P. 48–56. doi: [10.1016/j.coviro.2021.11.003](https://doi.org/10.1016/j.coviro.2021.11.003)
 12. Bin Jang H., Bolduc B., Zablocki O., Kuhn J.H., Roux S., Adriaenssens E.M., Brister J.R., Kropinski A.M., Krupovic M., Lavigne R., et al. Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nat. Biotechnol.* 2019. V. 37. P. 632–639. doi: [10.1038/s41587-019-0100-8](https://doi.org/10.1038/s41587-019-0100-8)
 13. Aiewsakun P., Adriaenssens E.M., Lavigne R., Kropinski A.M., Simmonds P. Evaluation of the genomic diversity of viruses infecting bacteria, archaea and eukaryotes using a common bioinformatic platform: steps towards a unified taxonomy. *J. Gen. Virol.* 2018. V. 99. No. 9. P. 1331–1343. doi: [10.1099/jgv.0.001110](https://doi.org/10.1099/jgv.0.001110)
 14. Muhire B.M., Varsani A., Martin D.P. SDT: a virus classification tool based on pairwise sequence alignment and identity calculation. *PLoS One*. 2014. V. 9. No. 9. Article No. e108277. doi: [10.1371/journal.pone.0108277](https://doi.org/10.1371/journal.pone.0108277)
 15. Fischer S., Freuling C.M., Muller T., Pfaff F., Bodenhofer U., Hoper D., Fischer M., Marston D.A., Fooks A.R., Mettenleiter T.C., et al. Defining objective clusters for rabies virus sequences using affinity propagation clustering. *PLoS Negl. Trop. Dis.* 2018. V. 12. No. 1. Article No. e0006182. doi: [10.1371/journal.pntd.0006182](https://doi.org/10.1371/journal.pntd.0006182)
 16. Meier-Kolthoff J.P., Göker M. VICTOR: genome-based phylogeny and classification of prokaryotic viruses. *Bioinformatics*. 2017. V. 33. No. 21. P. 3396–3404. doi: [10.1093/bioinformatics/btx440](https://doi.org/10.1093/bioinformatics/btx440)
 17. Lauber C., Gorbalenya A.E. Genetics-based classification of filoviruses calls for expanded sampling of genomic sequences. *Viruses*. 2012. V. 4. No. 9. P. 1425–1437. doi: [10.3390/v4091425](https://doi.org/10.3390/v4091425)
 18. Wilkinson D.A., Joffrin L., Lebarbenchon C., Mavingui P. Analysis of partial sequences of the RNA-dependent RNA polymerase gene as a tool for genus and subgenus classification of coronaviruses. *J. Gen. Virol.* 2020. V. 101. No. 12. P. 1261–1269. doi: [10.1099/jgv.0.001494](https://doi.org/10.1099/jgv.0.001494)
 19. Dos Santos I.C., de Souza R.D.S., Tolstoy I., Oliveira L.S., Gruber A. Integrating Sequence- and Structure-Based Similarity Metrics for the Demarcation of Multiple

- Viral Taxonomic Levels. *Viruses*. 2025. V. 17. No. 5. Article No. 642. doi: [10.3390/v17050642](https://doi.org/10.3390/v17050642)
20. Jumper J., Evans R., Pritzel A., Green T., Figurnov M., Ronneberger O., Tunyasuvunakool K., Bates R., Žídek A., Potapenko A., et al. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature*. 2021. V. 596. P. 583–589. doi: [10.1038/s41586-021-03819-2](https://doi.org/10.1038/s41586-021-03819-2)
 21. Varadi M., Velankar S. The impact of AlphaFold Protein Structure Database on the fields of life sciences. *Proteomics*. 2023. V. 23. No. 17. Article No. e2200128. doi: [10.1002/pmic.202200128](https://doi.org/10.1002/pmic.202200128)
 22. Lin Z., Akin H., Rao R., Hie B., Zhu Z., Lu W., Smetanin N., Verkuil R., Kabeli O., Shmueli Y., et al. Evolutionary-Scale Prediction of Atomic-Level Protein Structure with a Language Model. *Science*. 2023. V. 379. No. 6637. P. 1123–1130. doi: [10.1126/science.ade2574](https://doi.org/10.1126/science.ade2574)
 23. Katoh K., Standley D.M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 2013. V. 30. No. 4. P. 772–780. doi: [10.1093/molbev/mst010](https://doi.org/10.1093/molbev/mst010)
 24. Minh B.Q., Schmidt H.A., Chernomor O., Schrempf D., Woodhams M.D., von Haeseler A., Lanfear R. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* 2020. V. 37. No. 5. P. 1530–1534. doi: [10.1093/molbev/msaa015](https://doi.org/10.1093/molbev/msaa015)
 25. Sievers F., Higgins D.G. The Clustal Omega Multiple Alignment Package. *Methods Mol. Biol.* 2021. V. 2231. P. 3–16. doi: [10.1007/978-1-0716-1036-7_1](https://doi.org/10.1007/978-1-0716-1036-7_1)
 26. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 2007. V. 24. No. 8. P. 1586–1591. doi: [10.1093/molbev/msm088](https://doi.org/10.1093/molbev/msm088)
 27. Shaikat M.A., Nguyen T.T., Hsu E.B., Yang S., Bhatti A. Comparative study of encoded and alignment-based methods for virus taxonomy classification. *Sci. Rep.* 2023. V. 13. Article No. 18662. doi: [10.1038/s41598-023-45461-0](https://doi.org/10.1038/s41598-023-45461-0)
 28. Miller J.B., McKinnon L.M., Whiting M.F., Ridge P.G. CAM: an alignment-free method to recover phylogenies using codon aversion motifs. *PeerJ*. 2019. V. 7. Article No. e6984. doi: [10.7717/peerj.6984](https://doi.org/10.7717/peerj.6984)
 29. Chen J., Yang L., Li L., Goodison S., Sun Y. Alignment-free comparison of metagenomics sequences via approximate string matching. *Bioinform. Adv.* 2022. V. 2. No. 1 Article No. vbac077. doi: [10.1093/bioadv/vbac077](https://doi.org/10.1093/bioadv/vbac077)
 30. Yu H., Yau S.S. The optimal metric for viral genome space. *Comput. Struct. Biotechnol. J.* 2024. V. 23. P. 2083–2096. doi: [10.1016/j.csbj.2024.05.005](https://doi.org/10.1016/j.csbj.2024.05.005)
 31. Чалей М.Б., Кутыркин В.А. Распознавание рода коронавируса на основе прототипных штаммов. *Мат. биол. биоинф.* 2022. Т. 17. № 1. С. 10–27. doi: [10.17537/2022.17.10](https://doi.org/10.17537/2022.17.10)
 32. Parums D.V. Editorial: Revised World Health Organization (WHO) terminology for variants of concern and variants of interest of SARS-CoV-2. *Med. Sci. Monit.* 2021. V. 27. Article No. e933622. doi: [10.12659/MSM.933622](https://doi.org/10.12659/MSM.933622)
 33. Karim S.S.A., Karim Q.A. Omicron SARS-CoV-2 variant: a new chapter in the COVID-19 pandemic. *Lancet*. 2021. V. 398. No. 10317. P. 2126–2128. doi: [10.1016/S0140-6736\(21\)02758-6](https://doi.org/10.1016/S0140-6736(21)02758-6)
 34. Allen H., Tessier E., Turner C., Anderson C., Blomquist P., Simons D., Løchen A., Jarvis C.I., Groves N., Capelastegui F., et al. Comparative transmission of SARS-CoV-2 Omicron (B.1.1.529) and Delta (B.1.617.2) variants and the impact of vaccination: national cohort study, England. *Epidemiol. Infect.* 2023. V. 151. Article No. e58. doi: [10.1017/S0950268823000420](https://doi.org/10.1017/S0950268823000420)
 35. Lambrou A.S., Shirk P., Steele M.K., Paul P., Paden C.R., Cadwell B., Reese H.E., Aoki Y., Hassell N., Zheng X.Y., et al. Genomic surveillance for SARS-CoV-2 variants: Predominance of the Delta (B.1.617.2) and omicron (B.1.1.529) variants –

- United States, June 2021 – January 2022. *MMWR Morb. Mortal. Wkly Rep.* 2022. V. 71. No. 6.: P. 206–211. doi: [10.15585/mmwr.mm7106a4](https://doi.org/10.15585/mmwr.mm7106a4)
36. Глотов А.Г., Нефедченко А.В., Южаков А.Г., Котенева С.В., Глотова Т.И., Комина А.К., Красников Н.Ю. Генетический полиморфизм сибирских изолятов коронавируса крупного рогатого скота (Coronaviridae: Betacoronavirus-1: Bovine-Like coronaviruses). *Вопросы вирусологии.* 2022. Т. 67. № 6. С. 465–474. doi: [10.36233/0507-4088-141](https://doi.org/10.36233/0507-4088-141)
 37. Ожмегова Е.Н., Савочкина Т.Е., Прилипов А.Г., Тихомиров Е.Е., Ларичев В.Ф., Сайфуллин М.А., Гребенникова Т.В. Молекулярно-эпидемиологический анализ геновариантов SARS-CoV-2 на территории Москвы и Московской области. *Вопросы вирусологии.* 2022. Т. 67. № 6. С. 496–505. doi: [10.36233/0507-4088-146](https://doi.org/10.36233/0507-4088-146)
 38. Чалей М.Б., Кутыркин В.А. Выбор таргета в геномах прототипных штаммов для распознавания подрода коронавирусов. *Мат. биол. биоинф.* 2023. Т. 18. № 2. С. 267–281. doi: [10.17537/2023.18.267](https://doi.org/10.17537/2023.18.267)
 39. GenBank. URL: <https://www.ncbi.nlm.nih.gov/genbank> (accessed 20.12.2025). Sayers E.W., Cavanaugh M., Frisse L., Pruitt K.D., Schneider V.A., Underwood B.A., Yankie L., Karsch-Mizrachi I. GenBank 2025 update. *Nucl. Acids Res.* 2025. V. 53. No. D1. P. D56–D61. doi: [10.1093/nar/gkae1114](https://doi.org/10.1093/nar/gkae1114)
 40. Yuan W.G., Liu G.F., Shi Y.H., Xie K.M., Jiang J.Z., Yuan L.H. A discussion of RNA virus taxonomy based on the 2020 International Committee on Taxonomy of Viruses report. *Front. Microbiol.* 2022. V. 13. Article No. 960465. doi: [10.3389/fmicb.2022.960465](https://doi.org/10.3389/fmicb.2022.960465)

Рукопись поступила в редакцию 21.11.2025, переработанный вариант поступил 27.12.2025.
Дата опубликования 22.01.2026.

===== BIOINFORMATICS =====

Principal Component Analysis in Targeted Approach to Coronavirus Genus Recognition

Chaley M.B.¹, Kutyrkin V.A.²

¹*Institute of Mathematical Problems of Biology RAS, Keldysh Institute of Applied Mathematics of Russian Academy of Sciences, Pushchino, Russia*

²*Moscow State Technical University n.a. N.E. Bauman, Moscow, Russia*

Abstract. An original approach to coronavirus classification is proposed which is basing on presentation of gene analyzed (N-gene of nucleocapsid protein) by corresponding vector of amino acid codon frequencies and its subsequent comparison with vector of averaged codon frequencies for the known N-genes of viral taxon (one of the four coronavirus genera). Principal component analysis is used in non-standard way to determine whether frequency vector analyzed belongs to one of the taxons under consideration. Method was tested on 5769 N-genes of the four coronavirus genera and showed reliability of genus recognition above 95 %. Approach proposed for classification of the coronaviruses allows reducing dimension of codon frequency vector to 28 components without decrease of reliability, by considering the most significant amino acid codon frequencies in N-gene. The approach refers to alignment free methods which become increasingly popular in the last decade for virus classification.

Key words: *N-gene, coronaviruses, classification, alignment free methods, principal component analysis*