 

**МАТЕМАТИЧЕСКАЯ БИОЛОГИЯ, БИОИНФОРМАТИКА/MATHEMATICAL BIOLOGY, BIOINFORMATICS**

**DOI: https://doi.org/10.60797/jbg.2026.32.7 EDN: NSUVYK**

# ПРИМЕР ЦЕЛЕВОГО ПОДХОДА К РАСПОЗНАВАНИЮ РОДА КОРОНАВИРУСОВ С ПОМОЩЬЮ МЕТОДА ГЛАВНЫХ КОМПОНЕНТ

Научная статья

**Чалей М.Б.[1,] *, Кутыркин В.А.[2]**
[1] ORCID : 0000-0003-0236-2673;
[2] ORCID : 0000-0001-6815-002X;
[1] Институт математематических проблем биологии - филиал Института прикладной математики им. М.В.Келдыша РАН, Пущино, Российская Федерация
[2] Московский государственный технический университет им. Н.Э. Баумана, Москва, Российская Федерация

* Корреспондирующий автор (maramaria[at]yandex.ru)

Предложена: 13.05.2026; Принята: 25.06.2026; Опубликована: 26.06.2026

**Аннотация**

На примере распознавания рода коронавирусов представлен оригинальный подход к определению таксономии вирусов в метагеномных исследованиях. Подход относится к целевым (таргетным) методам по распознаванию вирусов на основе частот кодонов в N-гене белка нуклеокапсида. Используемый подход основан на нестандартном применении метода главных компонент к обучающим выборкам векторов частот кодонов для N-генов коронавирусов разных родов. Показана репрезентативность предложенной обучающей выборки. Продемонстрирована возможность значительного уменьшения параметров распознавания путем сокращения размерности пространства векторов частот кодонов N-гена. Предлагаемый подход к определению таксономической принадлежности вирусов относится к группе методов без выравнивания, развиваемых в последние десятилетия для выявления родства и эволюции видов вирусов по их секвенированным геномам.

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

# AN EXAMPLE OF TARGETED APPROACH TO CORONAVIRUS GENUSRECOGNITION IN APPLYING PRINCIPAL COMPONENT ANALYSIS

Research article

**Chaley M.B.[1,] *, Kutyrkin V.A.[2]**
[1] ORCID : 0000-0003-0236-2673;
[2] ORCID : 0000-0001-6815-002X;
[1] Pushchino, Moscow Oblast, Pushchino, Russian Federation
[2] Moscow State Technical University n.a. N.E. Bauman, Moscow, Russian Federation

* Corresponding author (maramaria[at]yandex.ru)

Suggested: 13.05.2026; Accepted: 25.06.2026; Published: 26.06.2026

**Abstract**

An original approach to determining virus taxonomy inmetagenomic studies is presented with an example of coronavirus genusrecognition. This approach may be referred to targeted methods for classifyingvirus, as it considers codon frequencies only in the N-gene of nucleocapsidprotein. The approach used is based on non-standard application of principalcomponent analysis to the training samples of codon frequency vectors for the N-genesof coronaviruses from different genera. Representativeness of the trainingsamples proposed is shown in the work. Possibility of significant reduction in numberof recognition parameters is demonstrated by reducing dimension of codon frequencyvector space for the N-genes. The approach, proposed for determining taxonomicaffiliation of the viruses, belongs to a group of alignment free methods whichare developing in recent decades to identify the relationships and evolution ofvirus species basing on their sequenced genomes.

**Keywords:** N-gene, coronaviruses, virustaxonomy classification, alignment free methods, principal component analysis.

## Введение

Таксономия вирусов, установленная и развиваемая Международным комитетом по таксономии вирусов ICTV [1], не только описывает и систематизирует наши знания об известных вирусах, но также способствует разработке вакцин и методов лечения вирусных инфекций. В случае обнаружения нового патогенного вируса определение его вида, рода и семейства ускоряет выбор возможной терапии и вакцинации, особенно если средства и способы борьбы против родственных ему вирусов уже существуют.

Благодаря современным технологиям классифицировать вирусы и определять эволюционные связи между ними возможно в результате сравнительного анализа их геномов. Быстрое развитие технологий секвенирования и рост метагеномных данных создают запрос на создание новых вычислительных методов, программ для анализа и классификации вирусов. С переходом к революционной нанопоровой технологии (Oxford Nanopore) секвенирования третьего поколения с длинным прочтением (LRS — long-read sequencing) открываются новые возможности вирусной

 

метагеномики [2], [3]. В отличие от высоко производительной технологии NGS (Illumina), известной как секвенирование с коротким прочтением (SRS — short-read sequencing), LRS технология позволяет значительно быстрее и надежнее определять области генома, содержащие тандемные повторы, и реконструировать полные геномы за одно прочтение [4]. Технология LRS способствует классификации вирусов с высоким разрешением и точному определению важных генетических элементов. Кроме того, её существенным преимуществом является возможность идентифицировать химические модификации нуклеотидов, возникающие на уровне эпигенетических процессов.

Основные подходы к классификации метагеномных данных (идентификации вирусов) можно условно разделить на четыре основные группы [5]: выравнивание; статистические и скрытые марковские модели (HMM); идентификация вирусов на основе машинного обучения на основе частот k-меров (небольших последовательностей из k нуклеотидных или аминокислотных остатков); гибридные методологии, которые сочетают в себе несколько стратегий.

Одними из первых, и широко используемых в настоящее время в метагеномике, подходов к таксономии являются методы определения сходства путем выравнивания с референсными геномами из баз данных, например, RefSeq [6]. Программа BLAST [7] является самым известным инструментом в этой группе. Специализированный портал NCBI Virus [8] предоставляет BLAST-интерфейс для анализа вирусных последовательностей, сравниваемых с тщательно отобранными метаданными о вирусах, в том числе и с коронавирусами рода *Betacoronavirus*. В настоящее время широко используются программы множественного выравнивания такие, как MUSCLE [9], MAFFT [10] и др. По результатам работы этих программ методами филогении [11] (наиболее часто методом правдоподобия [12]) строятся дендрограммы вирусных таксонов. В случае, когда денрограммы строятся на основе выравненных последовательностей, количественные различия определяют меру расстояния между ними.

Выравнивание последовательностей и анализ k-меров, требуют очень больших вычислительных затрат, что осложняет их применение для оценки огромных массивов данных. Потому в последние два десятилетия разрабатываются различные методы без выравнивания, обладающие достаточной эффективностью [13], [14], [15]. Методы без выравнивания для сравнения и классификации последовательностей извлекают и используют их характерные признаки или шаблоны.

Наши работы по распознаванию рода коронавирусов (*Alphacoronavirus*, *Betacoronavirus*, *Deltacoronavirus*, *Gammacoronavirus*) также направлены на создание новых подходов к распознаванию таксономической принадлежности вирусов без сравнительного выравнивания их геномных последовательностей относительно известных геномов вирусов. С этой целью в настоящей работе анализируемый вирус характеризуется вектором частот кодонов одного из его генов, кодирующего вирусный белок нуклеокапсида.

Ранее для распознавания рода коронавирусов на основе частот кодонов в генах нами рассматривался вариантный подход, который использовал как отдельные структурные (S-, M-, N- гены) и неструктурные гены (объединенные в единой ORF1ab), так и их различные комбинации-варианты [16]. Всего было рассмотрено шесть вариантов. Отметим, что ORF1ab содержит, в частности, ген РНК-зависимой РНК-полимеразы (RdRp), обычно используемый для разделения видов, родов и семейств коронавирусов [1].

Анализируемому геному коронавируса при вариантном подходе соответствовала вариантная строка, компоненты которой содержали результат распознавания рода коронавируса по каждому варианту. Мозаичность компонент вариантных строк отражала рекомбинантные процессы в геномах. Наиболее эффективное распознавание рода продемонстрировал вариант, использующий только ORF1ab. Распознавание рода коронавируса при рассмотрении частот кодонов структурных генов оказалось чуть менее эффективно, с уровнем достоверности (чувствительности) не менее 95%.

Вариантный подход использовал сравнение частот кодонов в генах анализируемого вирусного генома с усредненными частотами кодонов в соответствующих генах прототипных штаммов каждого рода коронавирусов, то есть вирусов, рассматриваемых Международным комитетом по таксономии вирусов ICTV [1] в качестве типичных представителей вирусного таксона. При этом среди структурных генов распознавание на основе N-гена показало недостаточно высокую чувствительность только для одного рода *Alphacoronavirus*.

Хотя N-ген имеет достаточно малую длину — 1200 нукл. (для сравнения S-ген имеет длину порядка 3700 нукл.), ранее он успешно использовался для построения филогенетических дендрограмм при изучении географии распространения коронавирусов [17] и в молекулярно-эпидемиологическом анализе вариантов вируса SARS-CoV-2 [18]. В работе [19] по распознаванию рода и подрода коронавируса также использовался N-ген белка нуклеопротеина, частоты кодонов которого из анализируемого генома сравнивались с частотами кодонов в N-генах индивидуальных прототипных штаммов подродов коронавирусов. Такой подход был назван типологическим [19], [20].

В работе [19] было показано, что при типологическом подходе распознавание на основе N-гена наиболее эффективно (в сравнении с S-геном и ORF1ab). Поэтому N-ген белка нуклеокапсида был выбран в качестве целевого (таргетного) гена в дальнейшем анализе.

Сложность типологического подхода обусловлена необходимостью сравнения с большим числом прототипных штаммов. В этом его сходство с методами выравнивания, которые также требуют сравнения с множеством эталонов.

Для упрощения задачи распознавания рода коронавирусов нами было выполнено сравнение вектора частот кодонов в анализируемом N-гене с усредненными по родам векторами частот кодонов в N-генах из обучающей выборки. Однако такой подход показал слишком низкую эффективность распознавания рода коронавируса. Поэтому в работе [21] для распознавания рода коронавирусов был предложен другой подход, в котором особым образом использовался метод главных компонент для векторов частот кодонов в анализируемых N-генах. В этом подходе на основе обучающей выборки для каждого рода коронавирусов создаются процедуры преобразования векторов частот кодонов в N-генах. Для фиксированного рода процедура преобразования выполняется следующим образом. Сначала на основе векторов частот кодонов в N-генах из обучающей выборки рассматриваемого рода коронавирусов вычисляется усредненный вектор частот кодонов в этом роде. На его основе осуществляется трансформация каждого вектора

 

частот кодонов рода, и для каждого рода коронавирусов создается соответствующая выборка трансформированных векторов частот кодонов. Далее к выборке трансформированных векторов частот кодонов N-генов каждого рода применяется метод главных компонент, для которых найдены их дисперсии. Таким образом, на основе обучающей выборки для анализируемого вектора частот кодонов N-гена созданы фиксированные процедуры его преобразования, зависящие от рода коронавирусов.

Согласно процедуре предполагаемого рода, анализируемый вектор частот кодонов N-гена коронавируса сначала трансформируется и, затем, для трансформированного вектора вычисляются главные компоненты. В результате для предполагаемого рода коронавирусов вычисляется сумма квадратов значений главных компонент, нормированных на соответствующие им дисперсии. Таким образом, для анализируемого коронавируса вычисляются такие суммы для возможных распознаваемых родов. На основе минимальной суммы выбирается род анализируемого коронавируса.

Описываемый выше способ распознавания рода коронавирусов основан на предположении о том, что для каждого рода векторы частот кодонов в N-генах коронавирусов могут быть достаточно близки к различным нормальным многомерным распределениям типичным для каждого рода коронавирусов. Эксперименты показали, что, начиная с некоторой размерности пространства векторов частот кодонов N-гена, достигался уровень чувствительности 95% (и выше) в распознавании рода коронавируса для обучающей и тестируемых выборок. Однако для тестируемой выборки в работе [21] наблюдались следующие явления. Тестируемая выборка показывала приемлемый уровень чувствительности (95%) распознавания рода *Betacoronavirus* только для достаточно высокой размерности ($n$=28) пространства векторов частот кодонов в N-генах. Кроме того, для рода *Deltacoronavirus*, начиная с размерности 16 происходит существенное снижение уровня чувствительности к распознаванию этого рода. По-видимому, такие явления обусловлены недостаточной репрезентативностью обучающей выборки.

В настоящей работе увеличен размер обучающей выборки. Показано, что приемлемый уровень эффективности распознавания рода коронавирусов достигается, начиная с достаточно низкой размерности пространства анализируемых векторов частот кодонов в N-генах и стабильно возрастает с ростом этой размерности.

Таким образом, в настоящей работе на примере распознавания рода коронавируса по N-гену белка нуклеокапсида демонстрируется возможность успешного применения метода главных компонент для классификации вирусов без выравнивания их полных геномов, с опорой только на выделенный целевой ген. Предлагаемый подход может существенно упростить распознавание последовательностей геномов, и избежать ошибок идентификации, обусловленных частыми рекомбинациями между вирусными геномами. Кроме того, дополнительно упростить анализ возможно путем значительного уменьшения размерности векторов частот кодонов, характеризующих целевой ген.

**Методы и принципы исследования**

В предыдущей работе [21], посвященной распознаванию рода коронавирусов, основанному на методе главных компонент, объем обучающей выборки N-генов был существенно меньше объема тестируемой. В настоящей работе объем обучающей выборки значительно увеличен, особенно для родов *Deltacoronavirus* и *Gammacoronavirus*. Численный состав по родам для обучающей и тестируемой выборок N-генов коронавирусов в настоящей работе, соответственно, составлял: для *Alphacoronavirus* (α-CoV) — 1281 и 251, для *Betacoronavirus* (β-CoV) — 2126 и 1647, для *Deltacoronavirus* (δ-CoV) — 244 и 80, и для *Gammacoronavirus* (γ-CoV) — 406 и 135. Обучающая и тестируемая выборки различались между собой в среднем на 91%.

Исходные последовательности N-генов были получены из базы данных GenBank [22]. Порядок компонент в векторах частот кодонов соответствовал нумерации кодонов, упорядоченных по убыванию частот встречаемости в N-генах прототипных штаммов для родов коронавирусов, GenBank коды доступа которых приведены в работе [23]. Таблица 1 представляет список упорядоченных таким образом кодонов, за которыми закреплены соответствующие номера. Последние пять кодонов, включая кодоны терминации, с самыми низкими частотами были исключены из анализа. Убывание частоты встречаемости кодона происходит в соответствии с ростом его порядкового номера в таблице 1.

Таблица 1 - Нумерованный список кодонов аминокислот в соответствии с убыванием их частоты встречаемости в N-генах прототипных штаммов подродов коронавирусов

DOI: https://doi.org/10.60797/jbg.2026.32.7.1

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

*Примечание: по ист. [23]*

Кратко опишем метод распознавания рода коронавирусов на основании метода главных компонент.

Каждый N-ген характеризуется вектором частот кодонов в виде $\boldsymbol{P} = \left(P_1, P_2, \ldots, P_n\right)^{\mathrm{T}}$, где $P_i$ — частота встречаемости $i$-го кодона $(i = \overline{1,n})$, $n$ — количество первых упорядоченных кодонов. Наименьшее число

 

рассматриваемых кодонов — два, наибольшее — 59. Пять кодонов с минимальными частотами встречаемости не рассматривались.

Для анализируемого вектора частот коронавирусов опишем его трансформацию, соответствующую каждому роду коронавирусов, где $X \in \{\alpha, \beta, \delta, \gamma\}$ — индекс рода коронавируса.

Для каждого рода по всем генам его обучающей выборки рассчитываются средние частоты встречаемости каждого рассматриваемого кодона. Тем самым для рода *X* создается вектор средних частот $\overline{\boldsymbol{P}}^X = \left(\bar{P}_1^X, \bar{P}_2^X, \ldots, \bar{P}_n^X\right)^{\mathrm{T}}$ и для каждого вектора частот обучающей выборки из рода *X* вычисляется трансформированный вектор $\boldsymbol{p} = \left(\frac{P_1 - \bar{P}_1^X}{\bar{P}_1^X}, \frac{P_2 - \bar{P}_2^X}{\bar{P}_2^X}, \ldots, \frac{P_n - \bar{P}_n^X}{\bar{P}_n^X}\right)^{\mathrm{T}} = \boldsymbol{F}_X(\boldsymbol{P})$ где $\boldsymbol{F}_X$ — операция *X*-трансформирования вектора частот. Следовательно, для обучающей выборки векторов частот из рода *X* с помощью операции $\boldsymbol{F}_X$ создается соответствующая выборка трансформированных векторов. На основе этой выборки вычисляется ковариационная матрица $\boldsymbol{C}_X$ для рода *X*. Для получения главных компонент рода *X* матрица $\boldsymbol{C}_X$ представляется в виде $\boldsymbol{C}_X = \boldsymbol{Q}_X \cdot \boldsymbol{D}_X \cdot \boldsymbol{Q}_X^{\mathrm{T}}$, где $\boldsymbol{Q}_X = \left(\boldsymbol{q}_1^X, \boldsymbol{q}_2^X, \ldots, \boldsymbol{q}_n^X\right)$ — ортогональная матрица, столбцы которой являются собственными векторами матрицы $\boldsymbol{C}_X$ т.е. $\boldsymbol{C}_X \cdot \boldsymbol{q}_i^X = d_i^X \boldsymbol{q}_i^X$ для $i = \overline{1, n}$ и

$$\boldsymbol{D}_X = \begin{pmatrix} d_1^X & 0 & \ldots & 0 \\ 0 & d_2^X & \mathrm{O} & \mathrm{M} \\ \mathrm{M} & \mathrm{O} & \mathrm{O} & 0 \\ 0 & \mathrm{L} & 0 & d_n^X \end{pmatrix}$$ – диагональная матрица.

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

Опишем процедуру распознавания рода анализируемого вектора частот $\boldsymbol{P} = \left(P_1, P_2, \ldots, P_n\right)^{\mathrm{T}}$ на основе предлагаемого метода главных компонент. Для каждого рода коронавирусов осуществляется следующая процедура преобразования этого вектора. Сначала этот вектор преобразуется с помощью операций трансформации $\boldsymbol{F}_X (X \in \{\alpha, \beta, \delta, \gamma\})$ и, затем, вычисляются главные вектора $\boldsymbol{y}^X = \left(y_1^X, y_2^X, \ldots, y_n^X\right)^{\mathrm{T}} = \boldsymbol{Q}_X^{\mathrm{T}} \cdot \boldsymbol{F}_X(\boldsymbol{P})$ и суммы $z^X = \frac{1}{n} \sum_{i=1}^{n} \frac{\left(y_i^X\right)^2}{d_i^X}$. Минимальное число среди $z^\alpha$, $z^\beta$, $z^\gamma$ и $z^\delta$ указывает на род анализируемого вектора частот.

**Результаты и обсуждение**

Эффективность распознавания рода коронавирусов в предлагаемом подходе оценивалась с помощью двух параметров: чувствительности и специфичности распознавания. Согласно определению [24], чувствительность — отношение количества истинно положительных результатов к общему количеству объектов с признаком, умноженное на 100%; специфичность — отношение количества истинно отрицательных результатов к общему количеству объектов без признака, умноженное на 100%.

**3.1. Распознавания рода коронавирусов в зависимости от объема обучающей выборки**

Рассмотрим особенности в распознавании рода коронавирусов, проявившиеся на тестируемой выборке N-генов в работе [21], которые показывает рисунок 1. Как отмечалось во Введении, приемлемый уровень чувствительности (95%) в распознавании рода *Betacoronavirus* достигается, начиная с достаточно высокой размерности (*n*=28) пространства векторов частот кодонов N-гена. Одновременно с улучшением распознавания рода *Betacoronavirus*, для рода *Deltacoronavirus*, начиная с размерности *n*=21, происходит существенное снижение уровня чувствительности в распознавании этого рода. Такие явления говорят о недостаточной репрезентативности обучающей выборки в работе [21]. Поэтому в настоящей работе объём обучающей выборки N-генов был значительно увеличен для всех родов коронавирусов.

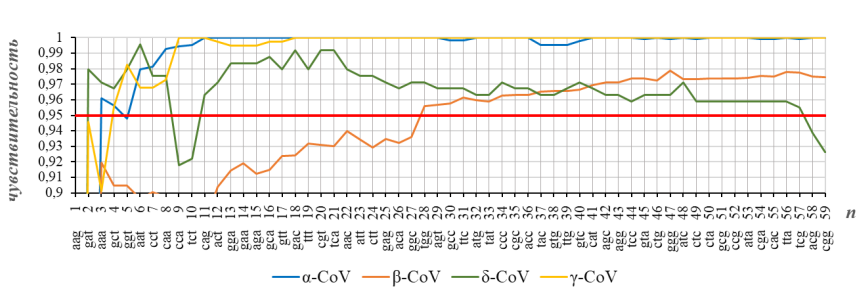


Рисунок 1 - Чувствительность распознавания родов коронавирусов (*Alphacoronavirus*, *Betacoronavirus*, *Deltacoronavirus*, *Gammacoronavirus*) для N-генов (представленных векторами частот кодонов), соответствующая обучающей выборке в работе [21]

DOI: https://doi.org/10.60797/jbg.2026.32.7.2

 

*Примечание: доля N-генов с правильно распознанным родом показана в зависимости от размерности n рассматриваемого пространства векторов частот кодонов; нумерация компонент вектора сопровождается указанием кодона, частота которого соответствует компоненте; порядок следования кодонов соответствует их списку в таблице 1*

На рисунке 2 показаны результаты распознавания тестируемой выборки при увеличении численности обучающей выборки, в среднем, в три раза для каждого рода коронавирусов (см. Материалы). Как можно видеть из рисунка 2, чувствительность в распознавании каждого рода (доля N-генов с правильно распознанным родом) непрерывно возрастает с ростом размерности рассматриваемого пространства векторов *n*, начиная с достаточно малой размерности *n*=12. С этой размерности специфичность распознавания рода *Alphacoronavirus* растет, начиная с 97.7%, и специфичность распознавания остальных трех родов составляет практически 100%. Следовательно, можно говорить о репрезентативности обучающей выборки.

Таким образом, увеличение объема обучающей выборки привело к отсутствию негативных особенностей в чувствительности предлагаемого подхода и его высокой специфичности при распознавании.

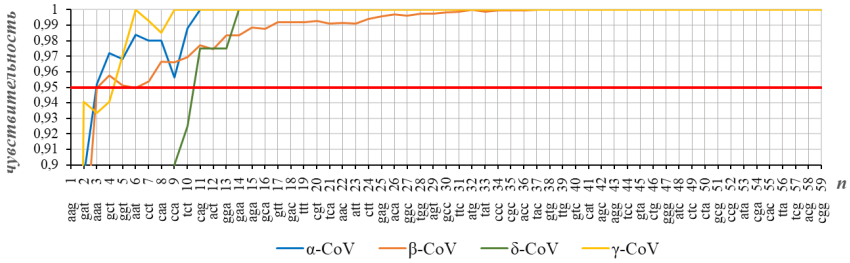


Рисунок 2 - Чувствительность распознавания родов коронавирусов (*Alphacoronavirus*, *Betacoronavirus*, *Deltacoronavirus*, *Gammacoronavirus*) для N-генов (представленных векторами частот кодонов), достигаемая с помощью обучающей выборки, используемой в настоящей работе
DOI: https://doi.org/10.60797/jbg.2026.32.7.3

*Примечание: доля N-генов с правильно распознанным родом показана в зависимости от размерности n рассматриваемого пространства векторов частот кодонов; рядом с номером компоненты вектора указан кодон, частота которого определяет эту компоненту; порядок следования кодонов соответствует их списку в таблице 1*

**3.2. Распознавание обучающей выборки**

Если обучающая выборка может считаться репрезентативной, то мы вправе ожидать, что результат распознавания с ее помощью любой тестируемой выборки не должен кардинально отличаться от результата, который достигается, когда обучающая выборка распознает саму себя. Рисунок 3 показывает, как обучающая выборка, используемая в настоящей работе, распознает себя в качестве тестируемой.

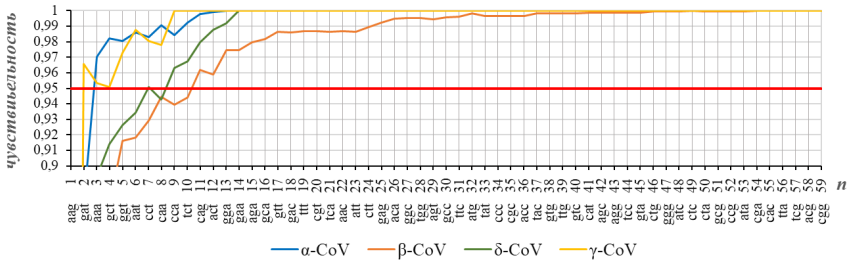


Рисунок 3 - Чувствительность распознавания родов коронавирусов (*Alphacoronavirus*, *Betacoronavirus*, *Deltacoronavirus*, *Gammacoronavirus*) для N-генов (представленных векторами частот кодонов), полученные для обучающей выборки, используемой в настоящей работе, в случае, когда она одновременно является и тестируемой
DOI: https://doi.org/10.60797/jbg.2026.32.7.4

*Примечание: доля N-генов с правильно распознанным родом показана в зависимости от размерности n рассматриваемого пространства векторов частот кодонов; рядом с номером компоненты вектора указан кодон, частота которого определяет эту компоненту; порядок следования кодонов соответствует их списку в таблице 1*

Как можно видеть из рисунка 3, начиная с размерности векторного пространства *n*=12, чувствительность в распознавании каждого рода коронавирусов становится выше 0.95 и продолжает стабильно расти с увеличением размерности *n* пространства векторов частот кодонов, с помощью которых анализируются N-гены коронавирусов.

 

Такой же результат наблюдался и на рисунке 2 при распознавании рода на тестируемой выборке. Сравнение рис. 2 и рис. 3 дополнительно подтверждает достоверность в распознавании рода коронавирусов с помощью используемой в настоящей работе обучающей выборки.

Таким образом, из рисунка 2 и рисунка 3 следует, что для эффективного распознавания рода коронавирусов можно ограничиться достаточно низкой размерностью ($n$=12) пространства векторов частот кодонов в N-гене.

**Заключение**

В настоящей работе представлен пример использования оригинального подхода к распознаванию геномов коронавирусов, относящийся к методам целевого (таргетного) распознавания без выравнивания. Подход исходит из оценки частот встречаемости кодонов в таргетном гене (в настоящей работе это N-ген белка нуклеокапсида). Подход основан на использовании метода главных компонент и требует репрезентативной обучающей выборки, наличие которой показано в настоящей работе. Кроме того, отмечена возможность значительного уменьшения параметров распознавания путем сокращения размерности пространства векторов частот кодонов N-гена.

**Конфликт интересов**

Не указан.

**Рецензия**

Все статьи проходят рецензирование. Но рецензент или автор статьи предпочли не публиковать рецензию к этой статье в открытом доступе. Рецензия может быть предоставлена компетентным органам по запросу.

**Conflict of Interest**

None declared.

**Review**

All articles are peer-reviewed. But the reviewer or the author of the article chose not to publish a review of this article in the public domain. The review can be provided to the competent authorities upon request.

**Список литературы / References**

1. Львов Д.К. Таксономия и мегатаксономия вирусов (домен Vira) — текущий статус / Д.К. Львов, В.Г. Акимкин, А.Д. Забережный [и др.] // Вопросы вирусологии. — 2025. — Т. 70. — № 5. — С. 401–416. — DOI: 10.36233/0507-4088-344.

2. Kim C. Unraveling metagenomics through long-read sequencing: a comprehensive review / C. Kim, M. Pongpanich, T. Porntaveetus // Journal of Translational Medicine. — 2024. — Vol. 22. — Art. 111. — DOI: 10.1186/s12967-024-04917-1.

3. Le B. A review of computational approaches for metagenomics by long-read sequencing / B. Le, L. Jia, T. Pang [et al.] // Science China Life Sciences. — 2026. — DOI: 10.1007/s11427-025-3133-3.

4. Ерёмин А.А. Нанопоровое секвенирование в геномике метагеномике и эпигеномике: инструменты и алгоритмы анализа / А.А. Ерёмин, А.В. Сергеев, М.Э. Зверева [и др.] // Математическая биология и биоинформатика. — 2025. — Т. 20. — № 2. — С. 588–624. — DOI: 10.17537/2025.20.588.

5. Galeeva J. Bioinformatics tools and approaches for virus discovery in genomic data: a systematic review / J. Galeeva, P. Kuzmichenko, A. Manolov [et al.] // Viruses. — 2025. — Vol. 17. — № 12. — Art. 1538. — DOI: 10.3390/v17121538.

6. Pruitt K.D. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy / K.D. Pruitt, T. Tatusova, G.R. Brown [et al.] // Nucleic Acids Research. — 2011. — Vol. 40. — № D1. — P. D130–D135. — DOI: 10.1093/nar/gkr1079.

7. Johnson M. NCBI BLAST: a better web interface / M. Johnson, I. Zaretskaya, Y. Raytselis [et al.] // Nucleic Acids Research. — 2008. — Vol. 36. — Ssuppl. 2. — P. W5–W9. — DOI: 10.1093/nar/gkn201.

8. NCBI Virus. — URL: https://www.ncbi.nlm.nih.gov/labs/virus/vssi/ (accessed: 06.04.2026).

9. Edgar R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput / R.C. Edgar // Nucleic Acids Research. — 2004. — Vol. 32. — № 5. — P. 1792–1797. — DOI: 10.1093/nar/gkh340.

10. Katoh K. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization / K. Katoh, J. Rozewicki, K.D.Yamada // Briefings in Bioinformatics. — 2019. — Vol. 20. — № 4. — P. 1160–1166. — DOI: 10.1093/bib/bbx108.

11. Kapli P. Phylogenetic tree building in the genomic age / P. Kapli, Z. Yang, M.J. Telford // Nature Reviews. Genetics. — 2020. — Vol. 21. — № 7. — P. 428–444. — DOI: 10.1038/s41576-020-0233-0.

12. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood / Z. Yang // Molecular Biology and Evolution. — 2007. — Vol. 24. — № 8. — P. 1586–1591. — DOI: 10.1093/molbev/msm088.

13. Chen J. Alignment-free comparison of metagenomics sequences via approximate string matching / J. Chen, L. Yang, L. Li [et al.] // Bioinformatics Advances. — 2022. — Vol. 2. — № 1. — Art. vbac077. — DOI: 10.1093/bioadv/vbac077.

14. Shaukat M.A. Comparative study of encoded and alignment-based methods for virus taxonomy classification / M.A. Shaukat, T.T. Nguyen, E.B. Hsu [et al.] // Scientific Reports. — 2023. — Vol. 13. — № 1. — Art. 18662. — DOI: 10.1038/s41598-023-45461-0.

15. Yu H. The optimal metric for viral genome space / H. Yu, S.S. Yau // Computational and Structural Biotechnology Journal. — 2024. — Vol. 23. — P. 2083–2096. — DOI: 10.1016/j.csbj.2024.05.005.

16. Чалей М.Б. Распознавание рода коронавируса на основе прототипных штаммов / М.Б. Чалей, В.А. Кутыркин // Математическая биология и биоинформатика. — 2022. — Т. 17. — № 1. — С. 10–27. — DOI: 10.17537/2022.17.10.

17. Глотов А.Г. Генетический полиморфизм сибирских изолятов коронавируса крупного рогатого скота (*Coronaviridae*: *Betacoronavirus*-1: Bovine-Like coronaviruses) / А.Г. Глотов, А.В. Нефедченко, А.Г. Южаков [и др.] // Вопросы вирусологии. — 2022. —Т. 67. — № 6. — С. 465–474. — DOI: 10.36233/0507-4088-141.

 

18. Ожмегова Е.Н. Молекулярно-эпидемиологический анализ геновариантов SARS-CoV-2 на территории Москвы и Московской области / Е.Н. Ожмегова, Т.Е. Савочкина, А.Г. Прилипов [и др.] // Вопросы вирусологии. — 2022. — Т. 67. — № 6. — С. 496–505. — DOI: 10.36233/0507-4088-146.

19. Чалей М.Б. Выбор таргета в геномах прототипных штаммов для распознавания подрода коронавирусов / М.Б. Чалей, В.А. Кутыркин // Математическая биология и биоинформатика. — 2023. — Т. 18. — № 2. — С. 267–281. — DOI: 10.17537/2023.18.267.

20. Чалей М.Б. Типологические подходы к распознаванию рода и подрода коронавирусов по структурным и неструктурным генам / М.Б. Чалей, В.А. Кутыркин // Математическая биология и биоинформатика. — 2024. — Т. 19. — № 2. — С. 593–606. — DOI: 10.17537/2024.19.593.

21. Чалей М.Б. Метод главных компонент в таргетном подходе к определению рода коронавирусов / М.Б. Чалей, В.А. Кутыркин // Математическая биология и биоинформатика. — 2026. — Т. 21. — № 1. — С. 1–13. — DOI: 10.17537/2026.21.1.

22. Benson D.A. GenBank / D.A. Benson, M. Cavanaugh, K. Clark [et al.] // Nucleic Acids Research. — 2013. — Vol. 41 (Database issue). — P. D36–D42. — DOI: 10.1093/nar/gks1195.

23. Щелканов М.Ю. История изучения и современная классификация коронавирусов (Nidovirales: Coronaviridae) / М.Ю. Щелканов, А.Ю. Попова, В.Г. Дедков [и др.] // Инфекция и иммунитет. — 2020. — Т. 10. — № 2. — С. 221–246. — DOI: 10.15789/2220-7619-HOI-1412.

24. Корнеенков А.А. Вычисление и интерпретация показателей информативности диагностических медицинских технологий / А.А. Корнеенков, С.В. Рязанцев, Е.Э. Вяземская // Медицинский Совет. — 2019. — Т. 20. — С. 45–51. — DOI: 10.21518/2079-701X-2019-20-45-51.

## Список литературы на английском языке / References in English

1. Lvov D.K. Taksonomiya i megataksonomiya virusov (domen Vira) — tekushchii status [Virus taxonomy and megataxonomy (Vira domain) – current status] / D.K. Lvov, V.G. Akimkin, A.D. Zaberezhnii [et al.] // Voprosi virusologii [Virology issues]. — 2025. — Vol. 70. — № 5. — P. 401–416. — DOI: 10.36233/0507-4088-344. [in Russian]

2. Kim C. Unraveling metagenomics through long-read sequencing: a comprehensive review / C. Kim, M. Pongpanich, T. Porntaveetus // Journal of Translational Medicine. — 2024. — Vol. 22. — Art. 111. — DOI: 10.1186/s12967-024-04917-1.

3. Le B. A review of computational approaches for metagenomics by long-read sequencing / B. Le, L. Jia, T. Pang [et al.] // Science China Life Sciences. — 2026. — DOI: 10.1007/s11427-025-3133-3.

4. Eremin A.A. Nanoporovoe sekvenirovanie v genomike metagenomike i epigenomike: instrumenti i algoritmi analiza [Tools and algorithms for nanopore sequencing data analysis in genomics, metagenomics, and epigenomics] / A.A. Eremin, A.V. Sergeev, M.E. Zvereva [et al.] // Matematicheskaya biologiya i bioinformatika [Mathematical biology and bioinformatics]. — 2025. — Vol. 20. — № 2. — P. 588–624. — DOI: 10.17537/2025.20.588. [in Russian]

5. Galeeva J. Bioinformatics tools and approaches for virus discovery in genomic data: a systematic review / J. Galeeva, P. Kuzmichenko, A. Manolov [et al.] // Viruses. — 2025. — Vol. 17. — № 12. — Art. 1538. — DOI: 10.3390/v17121538.

6. Pruitt K.D. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy / K.D. Pruitt, T. Tatusova, G.R. Brown [et al.] // Nucleic Acids Research. — 2011. — Vol. 40. — № D1. — P. D130–D135. — DOI: 10.1093/nar/gkr1079.

7. Johnson M. NCBI BLAST: a better web interface / M. Johnson, I. Zaretskaya, Y. Raytselis [et al.] // Nucleic Acids Research. — 2008. — Vol. 36. — Ssuppl. 2. — P. W5–W9. — DOI: 10.1093/nar/gkn201.

8. NCBI Virus. — URL: https://www.ncbi.nlm.nih.gov/labs/virus/vssi/ (accessed: 06.04.2026).

9. Edgar R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput / R.C. Edgar // Nucleic Acids Research. — 2004. — Vol. 32. — № 5. — P. 1792–1797. — DOI: 10.1093/nar/gkh340.

10. Katoh K. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization / K. Katoh, J. Rozewicki, K.D.Yamada // Briefings in Bioinformatics. — 2019. — Vol. 20. — № 4. — P. 1160–1166. — DOI: 10.1093/bib/bbx108.

11. Kapli P. Phylogenetic tree building in the genomic age / P. Kapli, Z. Yang, M.J. Telford // Nature Reviews. Genetics. — 2020. — Vol. 21. — № 7. — P. 428–444. — DOI: 10.1038/s41576-020-0233-0.

12. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood / Z. Yang // Molecular Biology and Evolution. — 2007. — Vol. 24. — № 8. — P. 1586–1591. — DOI: 10.1093/molbev/msm088.

13. Chen J. Alignment-free comparison of metagenomics sequences via approximate string matching / J. Chen, L. Yang, L. Li [et al.] // Bioinformatics Advances. — 2022. — Vol. 2. — № 1. — Art. vbac077. — DOI: 10.1093/bioadv/vbac077.

14. Shaukat M.A. Comparative study of encoded and alignment-based methods for virus taxonomy classification / M.A. Shaukat, T.T. Nguyen, E.B. Hsu [et al.] // Scientific Reports. — 2023. — Vol. 13. — № 1. — Art. 18662. — DOI: 10.1038/s41598-023-45461-0.

15. Yu H. The optimal metric for viral genome space / H. Yu, S.S. Yau // Computational and Structural Biotechnology Journal. — 2024. — Vol. 23. — P. 2083–2096. — DOI: 10.1016/j.csbj.2024.05.005.

16. Chaley M.B. Raspoznavanie roda koronavirusa na osnove prototipnikh shtammov [Coronavirus genus recognition based on prototype virus variants] / M.B. Chaley, V.A. Kutirkin // Matematicheskaya biologiya i bioinformatika [Mathematical biology and bioinformatics]. — 2022. — Vol. 17. — № 1. — P. 10–27. — DOI: 10.17537/2022.17.10. [in Russian]

17. Glotov A.G. Geneticheskii polimorfizm sibirskikh izolyatov koronavirusa krupnogo rogatogo skota (Coronaviridae: Betacoronavirus-1: Bovine-Like coronaviruses) [Genetic diversity of Siberian bovine coronavirus isolates (Coronaviridae: Coronavirinae: Betacoronavirus-1: Bovine-Like coronaviruses)] / A.G. Glotov, A.V. Nefedchenko, A.G. Yuzhakov [et al.] //

 

Voprosi virusologii [Problems of Virology]. — 2022. —Vol. 67. — № 6. — P. 465–474. — DOI: 10.36233/0507-4088-141. [in Russian]

18. Ozhmegova Ye.N. Molekulyarno-epidemiologicheskii analiz genovariantov SARS-CoV-2 na territorii Moskvi i Moskovskoi oblasti [Molecular epidemiological analysis of SARS-CoV-2 genovariants in Moscow and Moscow region] / Ye.N. Ozhmegova, T.E. Savochkina, A.G. Prilipov [et al.] // Voprosi virusologii [Problems of Virology]. — 2022. — Vol. 67. — № 6. — P. 496–505. — DOI: 10.36233/0507-4088-146. [in Russian]

19. Chaley M.B. Vibor targeta v genomakh prototipnikh shtammov dlya raspoznavaniya podroda koronavirusov [Choice of target in the genomes of prototypic strains to recognize subgenus of coronaviruses] / M.B. Chaley, V.A. Kutirkin // Matematicheskaya biologiya i bioinformatika [Mathematical biology and bioinformatics]. — 2023. — Vol. 18. — № 2. — P. 267–281. — DOI: 10.17537/2023.18.267. [in Russian]

20. Chaley M.B. Tipologicheskie podkhodi k raspoznavaniyu roda i podroda koronavirusov po strukturnim i nestrukturnim genam [Typological approaches to recognizing genus and subgenus of coronaviruses by structural and non-structural genes] / M.B. Chaley, V.A. Kutirkin // Matematicheskaya biologiya i bioinformatika [Mathematical biology and bioinformatics]. — 2024. — Vol. 19. — № 2. — P. 593–606. — DOI: 10.17537/2024.19.593. [in Russian]

21. Chaley M.B. Metod glavnikh komponent v targetnom podkhode k opredeleniyu roda koronavirusov [Principal component analysis in targeted approach to coronavirus genus recognition] / M.B. Chaley, V.A. Kutirkin // Matematicheskaya biologiya i bioinformatika [Mathematical biology and bioinformatics]. — 2026. — Vol. 21. — № 1. — P. 1–13. — DOI: 10.17537/2026.21.1. [in Russian]

22. Benson D.A. GenBank / D.A. Benson, M. Cavanaugh, K. Clark [et al.] // Nucleic Acids Research. — 2013. — Vol. 41 (Database issue). — P. D36–D42. — DOI: 10.1093/nar/gks1195.

23. Shchelkanov M.Yu. Istoriya izucheniya i sovremennaya klassifikatsiya koronavirusov (Nidovirales: Coronaviridae) [The history of investigation and modern classification of coronaviruses (Nidovirales: Coronaviridae)] / M.Yu. Shchelkanov, A.Yu. Popova, V.G. Dedkov [et al.] // Infektsiya i immunitet [Russian Journal of Infection and Immunity]. — 2020. — Vol. 10. — № 2. — P. 221–246. — DOI: 10.15789/2220-7619-HOI-1412. [in Russian]

24. Korneenkov A.A. Vichislenie i interpretatsiya pokazatelei informativnosti diagnosticheskikh meditsinskikh tekhnologii [Symptom dynamics assessment of the disease by methods of survival analysis] / A.A. Korneenkov, S.V. Ryazantsev, Ye.E. Vyazemskaya // Meditsinskii Sovet [Medical Council]. — 2019. — Vol. 20. — P. 45–51. — DOI: 10.21518/2079-701X-2019-20-45-51. [in Russian]