Математическая биология и биоинформатика
2023. Т. 18. № 2. С. 267–281. doi: 10.17537/2023.18.267

========== **БИОИНФОРМАТИКА** ==========

Выбор таргета в геномах прототипных штаммов для распознавания подрода коронавирусов

Чалей М.Б.[*1], Кутыркин В.А.[2]**

[1]*Институт математических проблем биологии – филиал ИПМ им. М.В. Келдыша РАН, Пущино, Московская область, Россия*
[2]*Московский государственный технический университет им. Н.Э. Баумана, Москва, Россия*

Аннотация. В работе предложен таргетный подход к распознаванию подрода коронавируса на основе распределения частот кодонов N-гена белка нуклеокапсида. В предложенном подходе на основе статистики вычисляется отклонение распределения частот кодонов в N-гене анализируемого генома коронавируса от такого же распределения в каждом из 67 прототипных штаммов, характеризующих 23 подрода в четырех родах коронавирусов. Наименьшее отклонение от распределения в одном из прототипных штаммов указывает на подрод, к которому принадлежит этот штамм. Такой подход оказался эффективным и обеспечивает достоверность распознавания подрода коронавируса не менее 99 %. Среди всех кодонов генетического кода в соответствии с распределением их частот в N-гене коронавирусов выделены совокупности из 38 и 7 кодонов, обеспечивающих требуемую эффективность распознавания. Выделенные в этих совокупностях кодоны фиксируют таксономическую структуру подрода коронавируса.

Ключевые слова: *подрод коронавируса, таргетный подход, прототипные штаммы коронавирусов, N-ген коронавируса, распределение частот кодонов в N-гене.*

ВВЕДЕНИЕ

Коронавирусная пандемия ускорила развитие информационных технологий для изучения эволюции вирусов и для борьбы с эпидемиями и вспышками новых инфекций неясного происхождения [1]. Широко известные мировому научному сообществу базы нуклеотидных данных такие, как GISAID (Global Initiative on Sharing Avian Influenza Data) [2], NCBI GenBank (National Center for Biotechnology Information GenBank) [3], ENA (European Nucleotide Archive) [4] и CNGBdb (China National GeneBank DataBase) [5], пополнились огромным числом секвенированных геномов коронавирусов, в частности, вызвавшем пандемию SARS-CoV-2 (Severe Acute Respiratory Syndrome-related coronavirus 2) [6]. Были созданы новые веб-сервисы для обработки геномных данных [6, 7] и изучены пути запросов системной информации в существующих базах данных по геномике, эпидемиологии и клиническим исследованиям SARS-CoV-2 [8]. Для обнаружения новых вирусов в общедоступных 5.7 миллионах экологически разнообразных баз данных, объемы нуклеотидных последовательностей в которых оцениваются как 10.2×10^{15} нукл., разработана инфраструктура облачных вычислений Serratus [9]. В качестве одного из таргетов для поиска РНК-содержащих вирусов, использовался ген РНК-зависимой РНК-полимеразы (RdRP). В результате было

[*]maramaria@yandex.ru
[**]vkutyrkin@yandex.ru

обнаружено около 132000 новых РНК-вирусов, в том числе девять новых коронавирусов [9].

Со времени открытия вирусов в конце 19 века доминирующим фактором в изучении и классификации являлась их патогенность [10]. В начале 21 века с развитием технологий секвенирования и наступления эры метагеномных исследований [11–13] такой подход следовало изменить, что было выполнено Международным комитетом по таксономии вирусов (ICTV) [14]. Поскольку метагеномные исследования выделяют вирус только как нуклеотидную последовательность из случайной выборки соответствующего природного разнообразия при отсутствии какой-либо дополнительной биологической информации, следовательно, таксономия вирусов должна опираться на филогенетический анализ последовательностей их геномов [10, 14].

Таксономическая структура вирусов и, в частности, коронавирусов была существенно пересмотрена в 2018 г. после введения понятие подрода и математической оценки некоторых рангов (подрод, род, семейство) [15]. Семейство *Coronaviridae* разделилось на два подсемейства *Orthocoronavirinae* и *Letovirinae*, в котором были введены новые род *Alphaletovirus* и подрод *Milecovirus*, включающий вид *Microhyla letovirus* 1, выявленный в результате метагеномных исследований. С выделением отдельного подсемейства *Orthocoronavirinae* была сформирована его новая структура, разделившая четыре рода *Alphacoronavirus* (α-CoV), *Betacoronavirus* (β-CoV), *Deltacoronavirus* (δ-CoV) и *Gammacoronavirus* (γ-CoV) на 23 подрода. Тогда же был введен и подрод *Sarbecovirus* рода *Betacoronavirus*, хотя вид *Severe acute respiratory syndrome coronavirus*, отнесенный к этому подроду, был известен уже более 10 лет [16].

Современная структура таксономии вирусов, подобная таксономии Линнея в биологии, представлена 15 рангами [17]. Наименьшим таксоном является вид, который относится к некоторому подроду, подрод – к роду, род – к подсемейству, подсемейство входит в семейство, и так далее происходит переход к таксонам все более высокого ранга. В 2021 г. Международным комитетом по таксономии вирусов было объявлено о введении единого бинарного формата именования вирусов, с указанием рода и вида [18], как это принято в биологии. Таким образом, в настоящее время таксономия вирусов приводится в соответствие с общими биологическими правилами.

Для построения филогенетических дендрограмм вирусов (также и коронавирусов) получил распространение метод максимального правдоподобия [19–21], который применяется для оценки родственной близости анализируемых геномов после выравнивания их нуклеотидных последовательностей [22]. Установление родства и общего предка в филогении и классификации вирусов имеет большое значение для выявления источника эпидемии или пандемии. Поэтому недавно было высказано предложение [23] дополнительно использовать непараметрические методы и подходы кладистического анализа (такие, как максимальная парсимония [24], анализ трех таксонов [25], алгоритм среднего консенсуса [26]) для подтверждения надежности результатов филогенетического анализа и построения таксономии вирусов.

Процедуры построения филогенетических дендрограмм при предварительном выравнивании генетических последовательностей весьма трудоемки. Принимая во внимание экспоненциальный рост данных о новых секвенированных вирусных последовательностях, их филогенетический анализ со временем будет усложняться и несомненно потребует передовых информационных технологий. В то время как для быстрого аннотирования выявляемых вирусов, определения их положения в таксономической системе можно использовать типологический подход к распознаванию подрода и рода, основываясь на характерных особенностях нуклеотидных последовательностей, свойственных классифицированным ранее вирусам.

Ранее [27] рассматривались различные подходы к распознаванию рода коронавирусов (CoVs) на основе геномов прототипных штаммов. Геном коронавируса характеризовался распределением частот кодонов его отдельных структурных и неструктурных генов: M-гена мембранного белка, S-гена спайк белка, N-гена белка нуклеокапсида и гена ORF1ab, кодирующего ряд неструктурных белков. Каждый из четырех родов (α-CoV, β-CoV, δ-CoV, γ-CoV) подразделяется на подроды, которые также характеризуются несколькими прототипными штаммами [28]. Поэтому род коронавируса характеризовался усредненным (аналитическим) распределением частот кодонов всех его прототипных штаммов. В работе [27] вводился вариантный подход к распознаванию рода. Каждый вариант подхода к распознаванию основывался на различных комбинациях структурных и неструктурных генов. Для первого варианта V_1 использовалась комбинация генов S+N+M+ORF1ab. Дополнительно исследовались еще пять других вариантов сочетания этих генов: V_2) S+N+M, V_3) S+N, V_4) ORF1ab, V_5) S и V_6) N. Как было показано, наиболее эффективное распознавание соответствовало варианту V_4, использующему только ген ORF1ab. Однако, наилучший результат распознавания родов β-Cov, δ-Cov и γ-Cov показал вариант V_6, основанный на N-гене белка нуклеокапсида. Недостаточно высокий уровень распознавания α-CoV с помощью N-гена, как показали дальнейшие исследования, был вызван усреднением распределений частот кодонов прототипных штаммов значительного числа подродов (12 подродов) в этом роде коронавирусов.

Для отслеживания мутантных штаммов SARS-CoV-2 и других видов коронавирусов был предложен достаточно быстрый и экономичный подход таргетного секвенирования [29]. В этом случае определялась последовательность только одного наиболее вариабельного фрагмента гена S-белка, ответственного за связывание с клеточными рецепторами при инфицировании. Аналогично, таргетный подход можно предложить и для быстрой идентификации рода, подрода и вида коронавируса, основываясь на отдельно выделенных фрагментах вирусного генома.

В настоящей работе была поставлена задача быстрого и надежного распознавания подрода по фрагменту (таргету) генома коронавируса. Среди выделенных ранее таргетов (различных комбинаций генов) вариантного подхода был выбран N-ген, как имеющий наименьшую длину ~1200 нукл. среди остальных рассматриваемых генов (для сравнения ген ORF1ab имеет длину ~20000 нукл.). Кодируемый N-геном высоко консервативный белок нуклеокапсида выполняет функции, связанные с вирусным патогенезом, транскрипцией и репликацией. N-ген часто используют для молекулярной диагностики CoVs, [30, 31].

От распознавания подрода коронавируса на основе усредненных распределений (аналитических средних распределений) частот кодонов в N-гене прототипных штаммов подродов мы перешли к распознаванию на основе отдельных распределений частот кодонов N-гена для каждого прототипного штамма. Такой переход способствовал более эффективному распознаванию как подрода, так и рода коронавирусов. Кроме того, исследовалась зависимость эффективности распознавания от использования различных групп кодонов из N-гена. Группы кодонов формировались согласно выбранным уровням частот кодонов в N-гене, усредненных по 67 прототипным штаммам всех четырех родов коронавирусов. В результате исследования были выявлены две наиболее эффективные группы кодонов. Сокращение числа кодонов аминокислот, используемых в процедуре распознавания, способствовало её оптимизации. Проведенная оптимизация позволила значительно улучшить эффективность распознавания рода коронавирусов по сравнению с распознаванием на основе аналитических средних распределений в работе [27].

МАТЕРИАЛЫ

В работе использовались геномы коронавирусов четырех родов, полученные из базы данных нуклеотидных последовательностей GenBank [32, 33]. Вместе их количество составляло 3242. Из этого числа для 2051 генома вместе с аннотированием рода был указан и подрод. В качестве обучающих выборок в каждом роде и подроде коронавирусов для создания усредненного распределения частот кодонов N-гена использовались соответствующие гены прототипных штаммов GenBank, коды доступа которых приведены в таблице 1. Для рода α-CoV рассматривались 22 прототипных штамма, для рода β-CoV – 28, для родов δ-CoV и γ-CoV – 10 и 7 прототипных штаммов, соответственно [28]. Все прототипные штаммы четырех родов коронавирусов разделялись на 23 подрода (см. табл. 1).

Таблица 1. Коды доступа прототипных (п/т) штаммов коронавирусов в GenBank. Обозначение подродов и индексация их прототипных штаммов, принятые в работе

Род	Подрод	Вид	GenBank ID п/т штамма	Индекс п/т штамма подрода *k*
Alphacoronavirus α-Cov	*Colacovirus* α01	Коронавирус летучих мышей CDPHE15 (Bat coronavirus CDPHE15)	NC_022103	1
	Decacovirus α02	Альфакоронавирус больших подковоносов HuB2013 (*Rhinolophus ferrumequinum* alphacoronavirus HuB2013)	NC_028814	2
		Коронавирус летучих мышей HKU10 (Bat coronavirus HKU10)	NC_018871	3
	Duvinacovirus α03	Коронавирус человека 229E (Human coronavirus 229E)	NC_002645	4
	Luchacovirus α04	Коронавирус крыс Лунцюань Rn (Lucheng Rn rat coronavirus)	NC_032730	5
	Minacovirus α05	Коронавирус норок 1-го типа (Mink coronavirus 1)	NC_023760	6
		Коронавирус хорьков (Ferret coronavirus)	KX512809	7
			KX512810	8
	Minunacovirus α06	Коронавирус длиннокрылов 1-го типа (*Miniopterus* bat coronavirus 1)	EU420138	9
		Коронавирус длиннокрылов HKU8 (*Miniopterus* bat coronavirus HKU8)	NC_010438	10
	Myotacovirus α07	Альфакоронавирус азиатских рыбоядных ночниц Sax-2011 (*Myotis ricketti* alphacoronavirus Sax2011)	NC_028811	11
	Nyctacovirus α08	Альфакоронавирус китайских вечерниц SC2013 (*Nyctacus velutinus* alhacoronavirus SC2013)	NC_028833	12
	Pedacovirus α09	Вирус эпизоотической диареи свиней (Porcine epidemic diarrhea virus)	KT323979	13
		Коронавирус домовых гладконосов 512 (*Scotophilus*	NC_009657	14

Род	Подрод	Вид	GenBank ID п/т штамма	Индекс п/т штамма подрода *k*
		bat coronavirus 512)		
	Rhinacovirus α10	Коронавирус подковоносов HKU2 (*Rhinolopus* bat coronavirus HKU2)	NC_009988	15
	Setracovirus α11	Коронавирус человека NL63 (Human coronavirus NL63)	AY567487	16
		NL63-подобный коронавирус BtKYNL63-9b (NL63-related bat coronavirus BtKYNL63-9b)	KY073745	17
	Tegacovirus α12	Альфакоронавирус 1-го типа (Alphacoronavirus 1)	NC_038861	18
			KP981644	19
			FJ938051	20
			AY994055	21
			KR270796	22
Betacoronavirus β-Cov	*Embecovirus* β01	Бетакоронавирус 1-го типа (Betacoronavirus 1)	KF294357	1
			BCU00735	2
			KX432213	3
			EF446615	4
			AY391777	5
			NC_017083	6
			MF083115	7
		Коронавирус крыс Китая HKU24 (China Rattus coronavirus HKU24)	NC_026011	8
		Коронавирус мышей (Murine coronavirus)	AC_000192	9
			KF294371	10
			NC_012936	11
		Коронавирус человека HKU1 (Human coronavirus HKU1)	NC_006577	12
	Hibecovirus β02	Бетакоронавирус листоносов Пратта Zhejiang2013 (Bat Hp-betacoronavirus Zhejiang2013)	NC_025217	13
	Merbecovirus β03	Коронавирус Ближневосточного респираторного синдрома (Middle East respiratory syndrome-related coronavirus)	KF917527	14
			JX869059	15
			MG596803	16
		Коронавирус ежей 1-го типа (Hedgehog coronavirus 1)	MK679660	17
		Коронавирус косолапых кожанов HKU4 (*Tylonycteris* bat coronavirus HKU4)	NC_009019	18
		Коронавирус нетопырей HKU5 (*Pipustrellus* bat coronavirus HKU5)	NC_009020	19
	Nobecovirus β04	Коронавирус ночных крыланов GCCDC1 (*Rousettus* bat coronavirus GCCDC1)	NC_030886	20
		Коронавирус ночных крыланов HKU9 (Rousettus bat coronavirus HKU9)	NC_009021	21
	Sarbecovirus β05	Коронавирус китайских подковоносов (*Rhinolophus sinicus* coronavirus)	MG772933	22
			MG772934	23

Род	Подрод	Вид	GenBank ID п/т штамма	Индекс п/т штамма подрода *k*
		Коронавирус тяжелого острого респираторного синдрома (Severe acute respiratory syndrome-related coronavirus)	AY278489	24
			FJ588686	25
		Коронавирус тяжелого острого респираторного синдрома 2-го типа (Severe acute respiratory syndrome-related coronavirus 2)	NC_045512	26
			MT121216	27
			MN996532	28
Deltacoronovirus δ-Cov	*Andecovirus* δ01	Коронавирус свиязей HKU20 (Wigeon coronavirus KKU20)	NC_016995	1
	Buldecovirus δ02	Дельтакоронавирус свиней (Porcine deltacoronavirus)	JQ065042	2
			KJ569769	3
			NC_016992	4
		Коронавирус белоглазок HKU16 (White eye coronavirus HKU16)	NC_016991	5
		Коронавирус бюльбюлей HKU11 (Bulbul coronavirus HKU11)	FJ376620	6
		Коронавирус муний HKU13 (Munia coronavirus HKU13)	NC_011550	7
			NC_016993	8
	Herdecovirus δ03	Коронавирус квакв HKU19 (Night heron coronavirus HKU19)	NC_016994	9
	Moordecovirus δ04	Коронавирус камышниц HKU21 (Common morgen coronavirus HKU21)	NC_016996	10
Gammacoronavirus γ-Cov	*Cegacovirus* γ01	Коронавирус китообразных (Cetacean coronavirus)	EU111742	1
			KF793826	2
	Igacovirus γ02	Коронавирус птиц (Avian coronavirus)	KF696629	3
			GQ504724	4
			NC_010800	5
			AY641576	6
			MK423877	7

При распознавании рода с помощью вариантного подхода [27] было показано, что распознавание на основе N-гена белка нуклеокапсида дает наилучший результат по трем родам (β-, δ- и γ- CoVs). Вследствие такого результата и достаточно небольшой длины N-гена (~1200 нукл.) в настоящей работе распознавание подрода коронавируса основывается только на этом гене. Как и ранее, при таком распознавании из рассмотрения исключались пять кодонов с самыми низкими частотами, три из которых являлись кодонами терминации. Таким образом, каждый рассматриваемый геном коронавируса характеризовался распределением частот 59 кодонов в своем N-гене.

Дополнительным аргументом в пользу выбора N-гена в качестве таргета для распознавания подрода послужило сравнение результатов вариантного подхода на основе аналитических распределений подродов, то есть усредненных (по подроду) распределений частот кодонов прототипных штаммов коронавирусов. Таблица 2 показывает ошибки в определении рода, в результате распознавания подрода у 67 прототипных штаммов всех четырех родов коронавирусов (см. табл. 1), в зависимости от используемого варианта.

Таблица 2. Количество ошибок в определении рода для шести вариантов таргетного анализа, полученное на основе распознавания подрода прототипных штаммов

Варианты (таргеты)	α–CoV	β–CoV	δ–CoV	γ–CoV	итого
V_1) S+N+M+ORF1ab	1	5	1	0	7
V_2) S+N+M	1	1	1	0	3
V_3) S+N	1	0	0	0	1
V_4) ORF1ab	1	5	1	0	7
V_5) S	0	1	0	3	4
V_6) N	1	0	0	0	1

Как можно видеть, минимальное число ошибок достигается при использовании третьего (S + N – объединения генов) и шестого (N-гена белка нуклеокапсида) вариантов. Шестой вариант является более предпочтительным вследствие использования фрагмента меньшей длины из генома коронавируса.

МЕТОДЫ

При распознавании подрода коронавируса использовались два подхода.

Распознавание подрода коронавируса на основе аналитических распределений частот кодонов N-гена

Поскольку 67 прототипных штаммов распределены по 23-м подродам, то каждый подрод будем характеризовать усредненным распределением частот 59 кодонов N-гена в подроде. Такое распределение будем называть аналитическим распределением частот кодонов подрода.

Пусть $S \in \{\alpha, \beta, \delta, \gamma\}$ – символ, используемый для обозначения рода коронавируса и j – двузначный номер подрода в своем роде, согласно табл. 1. Например, $\alpha 08$ – обозначение подрода *Nyctacovirus* рода α-CoV. Каждое аналитическое распределение частот кодонов N-гена в подроде Sj будет описываться строкой вида:

$$\boldsymbol{P}^{Sj} = (P_1^{Sj}, P_2^{Sj}, ..., P_{59}^{Sj}), \quad (1)$$

где P_i^{Sj} – частота i-го кодона в аналитическом распределении подрода Sj для $i = \overline{1,59}$. Рассматриваемый геном коронавируса будет описываться распределением кодонов своего N-гена, которое определяется строкой:

$$\boldsymbol{p} = (p_1, p_2, ..., p_{59}), \quad (2)$$

где p_i – частота i-го кодона в распределении частот кодонов N-гена этого генома.

Отклонение рассматриваемого генома коронавируса с распределением частот (2) от подрода Sj определяется числом $\mathrm{D}^{Sj}(\boldsymbol{p})$, вычисляемым, согласно обозначению (1), по формуле:

$$\mathrm{D}^{Sj}(\boldsymbol{p}) = \frac{1}{7}\sum_{i=1}^{59}\frac{|P_i^{Sj} - p_i|}{P_i^{Sj}}, \quad (3)$$

где дробь 1/7 введена для удобства визуального анализа результатов.

Получив значения отклонений распределения (2) от всех распределений подродов, выбирают тот подрод, для которого реализуется минимальное отклонение. В результате принимается гипотеза о принадлежности рассматриваемого генома коронавируса к такому подроду.

Математическая биология и биоинформатика. 2023. Т. 18. № 2. doi: 10.17537/2023.18.267

Распознавание подрода коронавируса на основе распределений частот кодонов N-гена прототипных штаммов

Рассмотрим другой подход к распознаванию подрода коронавируса. В этом подходе выявляется один из 67 прототипов коронавируса, к которому наиболее близок рассматриваемый коронавирус. Подрод этого прототипа признается подродом анализируемого коронавируса.

Как и в первом подходе, $S \in \{\alpha, \beta, \delta, \gamma\}$ – символ, используемый для обозначения рода коронавируса и j – двузначный номер подрода в своем роде. Обозначение Sj,k будет использоваться, согласно таблице 1, для k-го прототипа подрода Sj. Каждое распределение частот кодонов N-гена в подроде Sj будет описываться строкой вида:

$$\boldsymbol{p}^{Sj,k} = (p_1^{Sj,k}, p_2^{Sj,k}, ..., p_{59}^{Sj,k}), \quad (4)$$

где $p_i^{Sj,k}$ – частота i-го кодона в распределении прототипного штамма Sj,k для $i = \overline{1,59}$. Как и ранее, рассматриваемый геном коронавируса будет описываться распределением кодонов своего N-гена, которое определяется строкой:

$$\boldsymbol{p} = (p_1, p_2, ..., p_{59}), \quad (5)$$

где p_i – частота i-го кодона в распределении частот кодонов N-гена этого генома.

Отклонение рассматриваемого генома коронавируса с распределением частот (5) от прототипного штамма Sj,k определяется числом $d^{Sj,k}(\boldsymbol{p})$, вычисляемым, согласно обозначению (4), по формуле:

$$d^{Sj,k}(\boldsymbol{p}) = \frac{1}{7}\sum_{i=1}^{59}\frac{|p_i^{Sj,k} - p_i|}{p_i^{Sj,k}}. \quad (6)$$

Получив значения отклонений распределения (6) от всех распределений прототипных штаммов, выбирают тот штамм Sj,k, для которого реализуется минимальное отклонение. В результате принимается гипотеза о принадлежности рассматриваемого генома коронавируса к подроду Sj.

В работе, при использовании указанных выше двух подходов для распознавания подрода на основании N-генов, из аннотированных по подроду 2051 генома в первом случае (с усреднением) были отмечено 43 ошибки (из них 19 – ошибочно определен род коронавируса), во втором (без усреднения) – 36 ошибок, из которых только три ошибки неверно определенного рода. Этот результат показывает, что использование аналитических (усредненных) распределений частот кодонов прототипных штаммов менее эффективно для распознавания как подрода, так и рода коронавирусов.

При высокой скорости мутации вирусов и коронавирусов, в частности, было бы интересно оценить, какие кодоны N-гена существенны для определения принадлежности коронавируса к подроду. В связи с этим встает вопрос о возможном уменьшении количества кодонов в N-гене коронавируса для распознавания его подрода. Для изучения такой возможности в работе осуществлялся отбор кодонов N-гена на основании их средней частоты встречаемости в прототипных штаммах всех подродов. На рисунке 1 показан график такой средней частоты встречаемости кодонов в N-генах всех прототипных штаммов коронавирусов (см. табл. 1). При отборе кодонов задавался порог частоты встречаемости кодона. Выбор порога частоты встречаемости кодона означает, что для распознавания используются только те кодоны, частота встречаемости которых превышает величину порога. На рисунке 1 отдельные пороги представлены горизонтальными линиями разного цвета с указанием числа кодонов, частоты которых превышают порог.

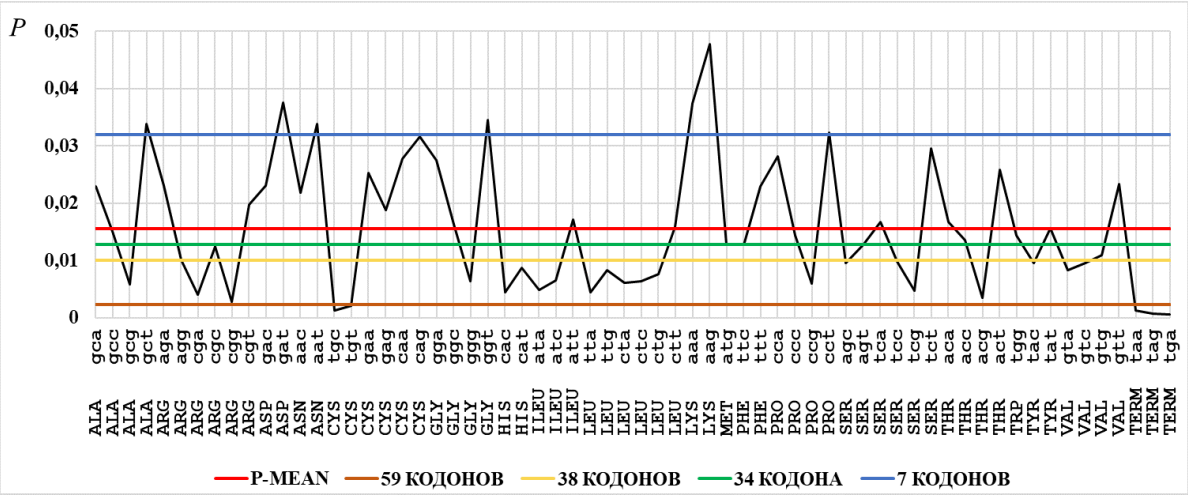

Рис. 1. Средняя частота встречаемости кодонов генетического кода в N-генах прототипных штаммов подродов коронавирусов (черная линия). Цветными горизонтальными линиями показаны различные пороги частоты с указанием количества кодонов, частоты которых выше соответствующего порога. P-MEAN – средняя частота по всем кодонам в N-генах прототипных штаммов.

Пусть выбранный порог оставил в рассмотрении совокупность $C_r = \{c_1, c_2, ..., c_r\}$ кодонов N-гена, где $|C_r| = r$ – количество кодонов в этой совокупности. Кроме того, ρ_i – частота встречаемости кодона c_i в N-гене рассматриваемого генома коронавируса для $i = \overline{1,r}$. Как и ранее (см. формулу (6)), $\rho_i^{Sj,k}$ – частота встречаемости кодона c_i в N-гене прототипного штамма Sj,k, где Sj – подрод штамма и k – номер штамма в этом подроде. Таким образом, прототипный штамм Sj,k определяется строкой $\boldsymbol{\rho}^{Sj,k} = (\rho_1^{Sj,k}, \rho_2^{Sj,k}, ..., \rho_r^{Sj,k})$ и анализируемый геном коронавируса – строкой $\boldsymbol{\rho} = (\rho_1, \rho_2, ..., \rho_r)$. Тогда отклонение рассматриваемого генома коронавируса с распределением частот $\boldsymbol{\rho}$ от прототипного штамма Sj,k определяется числом $\Delta^{Sj,k}(\boldsymbol{\rho})$, вычисляемым по формуле:

$$\Delta^{Sj,k}(\boldsymbol{\rho}) = \frac{1}{7}\sum_{i=1}^{r}\frac{|\rho_i^{Sj,k} - \rho_i|}{\rho_i^{Sj,k}}. \qquad (7)$$

Получив значения отклонений распределения (7) от всех распределений прототипных штаммов, выбирают тот штамм Sj,k, для которого реализуется минимальное отклонение. В результате принимается гипотеза о принадлежности рассматриваемого генома коронавируса к подроду Sj и, тем самым, к роду S (см. табл. 1).

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Рассмотрим результаты распознавания подрода коронавирусов на основе совокупности кодонов из N-гена, частота встречаемости которых, определялась выбранным порогом.

Пусть $M(r)$ – это количество ошибок при распознавании подрода коронавирусов с помощью совокупности кодонов C_r и формулы (7) для аннотированного по подроду 2051 генома коронавирусов. На рисунке 2 показана зависимость такого количества ошибок от числа кодонов в совокупности C_r.

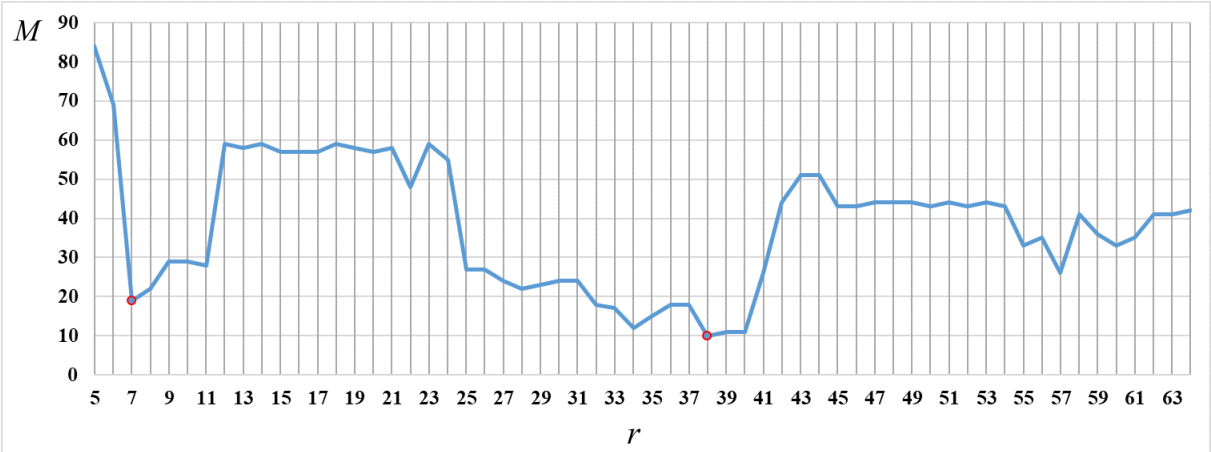

Рис. 2. Количество ошибок M при распознавании подрода на аннотированной выборке из 2051 генома коронавирусов в зависимости от размера r используемой для этого совокупности кодонов C_r.

Таблица 3. Распределение количества ошибок при распознавании подрода в выборке аннотированных 2051 генома коронавирусов с использований совокупностей кодонов C_7 и C_{38}. Обозначения подродов коронавирусов соответствуют таблице 1

Подрод	Размер выборки	Ошибки на совокупности C_7	Ошибки на совокупности C_{38}
α01	2	0	0
α02	14	2	2
α03	67	3	3
α04	2	0	0
α05	6	0	0
α06	5	0	0
α07	2	0	0
α08	2	0	0
α09	6	2	2
α10	6	1	0
α11	64	1	1
α12	75	1	0
β01	263	0	0
β02	2	0	0
β03	336	3	0
β04	11	4	0
β05	1035	1	2
δ01	2	0	0
δ02	12	1	0
δ03	2	0	0
δ04	2	0	0
γ01	4	0	0
γ02	131	0	0
Итого:	2051	19	10

Анализ графика на рисунке 2 позволяет выделить два особых минимума на совокупностях C_{38} из 38-ми и C_7 из семи кодонов. На совокупности C_{38} достигается минимальное количество ошибок ($M = 10$). Особо выделяется резкий минимум ($M = 19$) на совокупности C_7. Таблица 3 показывает детали результатов распознавания подрода в выборке геномов коронавирусов, аннотированных по подроду. Строки с подродами, для которых количество ошибок при распознавании на совокупностях

кодонов C_7 и C_{38} различаются, выделены в таблице 3. Можно сказать, что в случаях, когда подрод представлен достаточно репрезентативной выборкой, такие различия минимальны.

Отметим, что при распознавании на основе совокупности C_{38} из десяти ошибок при распознавании подрода нет ни одной ошибки распознавания рода. При аналогичной процедуре распознавания подрода коронавируса с использованием аналитических (усредненных) распределений частот кодонов для подродов произошло 45 ошибок распознавания подрода на основе совокупности C_{38} и 123 ошибки на основе совокупности C_7. Тем самым подтверждается эффективность распознавания с использованием индивидуальных распределений частот кодонов N-гена в прототипных штаммах. Таким образом, предложенные методы распознавания подрода коронавирусов дают не менее 99 % достоверности.

Полученная эффективность распознавания подрода коронавирусов на основе сокращенных совокупностей кодонов в N-гене коронавируса позволяет предположить, что выделенные в совокупностях C_{38} и C_7 кодоны имеют существенную роль для фиксации таксономической структуры подрода.

	C_{38}	C_7		C_{38}	C_7		C_{38}	C_7		C_{38}	C_7
LYS aaa	1	1	THR aca	1	0	ALA gct	1	1	LEU ctt	1	0
LYS aag	1	1	THR acc	1	0	ALA gca	1	0	LEU cta	0	0
ASN aat	1	1	THR act	1	0	ALA gcc	1	0	LEU ctc	0	0
ASN aac	1	0	THR acg	0	0	ALA gcg	0	0	LEU ctg	0	0
ARG cgt	1	0	MET atg	1	0	VAL gtt	1	0	LEU tta	0	0
ARG cgc	1	0	ILEU att	1	0	VAL gtg	1	0	LEU ttg	0	0
ARG cga	0	0	ILEU ata	0	0	VAL gta	0	0	PHE ttt	1	0
ARG cgg	0	0	ILEU atc	0	0	VAL gtc	0	0	PHE ttc	1	0
ARG aga	1	0	ASP gat	1	1	CYS caa	1	0	TYR tat	1	0
ARG agg	1	0	ASP gac	1	0	CYS cag	1	0	TYR tac	0	0
SER agt	1	0	CYS gaa	1	0	HIS cac	0	0	TERM taa	0	0
SER agc	0	0	CYS gag	1	0	HIS cat	0	0	TERM tag	0	0
SER tca	1	0	GLY ggt	1	1	PRO cct	1	1	TERM tga	0	0
SER tct	1	0	GLY gga	1	0	PRO cca	1	0	TRP tgg	1	0
SER tcc	0	0	GLY ggc	1	0	PRO ccc	1	0	CYS tgc	0	0
SER tcg	0	0	GLY ggg	0	0	PRO ccg	0	0	CYS tgt	0	0

Рис. 3. Совокупности кодонов C_{38} и C_7, на которых достигается наименьшее количество ошибок распознавания подрода по N-гену коронавируса, представлены в структуре таблицы универсального генетического кода. Цифра 1 показывает вхождение кодона в совокупность, 0 – отсутствие кодона в совокупности. Слева от кодона приведено трехбуквенное обозначение кодируемой аминокислоты. Стоп-кодоны обозначены как TERM.

На рисунке 3 показаны кодоны генетического кода, входящие в совокупности C_{38} и C_7, где цифра 1 указывает на вхождение кодона в совокупность, и цифра 0 указывает на отсутствие кодона в рассматриваемой совокупности. Совокупности кодонов C_{38} и C_7 рассматриваются на рисунке 3 с точки зрения их расположения в структуре кодонов универсального генетического кода. Как можно видеть, структура генетического кода представлена блоками (или боксами) из четырех кодонов, различающихся только по третьему нуклеотиду. Блоки, которые содержат синонимичные кодоны одной аминокислоты, принято называть фамильными боксами. Из 16 боксов генетического кода восемь являются фамильными боксами. Отметим, что в совокупность C_{38} входит, как правило, не менее половины кодонов из фамильных боксов. В целом, совокупность C_{38} имеет не менее одного представителя в каждом боксе. В отличие от C_{38},

Математическая биология и биоинформатика. 2023. Т. 18. № 2. doi: 10.17537/2023.18.267

совокупность C_7 вообще не имеет своих представителей в 11 боксах из 16. Среди синонимичных кодонов в оставшихся пяти боксах совокупность C_7, фактически, выделяет единственный кодон, что, возможно, вместе указывает на особую роль таких кодонов в структуре N-гена и их значимость для идентичности подрода коронавируса.

ЗАКЛЮЧЕНИЕ

В работе реализован таргетный подход для распознавания подрода коронавирусов на основе распределения частот кодонов в N-гене белка нуклеокапсида. Ранее для распознавания рода коронавирусов использовался вариантный подход, основанный на целом наборе таргетов (отдельных генов коронавирусов и их различных объединений). Предложенный в настоящей работе подход оказался намного более простым и позволяет обеспечить достоверность распознавания не менее 99 %.

Распознавание подрода коронавируса опирается на сравнение распределений частот кодонов в его N-гене и в N-генах прототипных штаммов, определяющих соответствующие подроды коронавирусов. Для сравнения применяется статистика, аналогичная статистике, предложенной ранее для распознавания рода коронавируса. Отличие статистики, используемой в настоящей работе, состоит в том, что она учитывает отклонение распределения частот кодонов в N-гене анализируемого коронавируса от аналогичного распределения каждого прототипного штамма. В то время, как ранее при распознавании рода использовались усредненные (аналитические) распределения частот кодонов по всем прототипным штаммам рода.

В работе предпринята попытка выделения в N-гене кодонов, наиболее значимых для распознавания подрода. Для этого был построен график зависимости количества ошибок распознавания от выбранных совокупностей кодонов, упорядоченных по мере возрастания частоты их встречаемости. В результате, наиболее эффективной оказалась совокупность из 38 кодонов. Эта совокупность включает не менее половины кодонов в каждом из 16 боксов генетического кода, содержащих по четыре кодона, отличающихся только в третьей позиции. Отметим, что выбор N-гена в качестве таргета для распознавания рода коронавирусов, позволил безошибочное определение рода на выборке из 2051 генома коронавируса, аннотированной по подроду, на основании совокупности из 38 кодонов.

Наименьшее число кодонов, обеспечивающее требуемую эффективность (99 %) распознавания, оказалось в совокупности из семи кодонов. В эту совокупность вошли кодоны из пяти боксов генетического кода, причем в трех боксах синонимичных кодонов выбрано по одному кодону. Возможно, что совокупность из этих семи кодонов является опорной для фиксации таксономической структуры подрода в N-гене коронавируса.

СПИСОК ЛИТЕРАТУРЫ

1. Спринджук М.В., Берник В.И., Калоша Н.И., Батгэрел Б., Автоматизация и математический аппарат анализа биоинформационных данных геномной природы. *Системный анализ и управление в биомедицинских системах*. 2022. Т. 21. № 4. С. 129–139. doi: 10.36622/VSTU.2022.21.4.018
2. *GISAID*. URL: https://gisaid.org (accessed 14.06.2023).
3. *GenBank*. URL: https://www.ncbi.nlm.nih.gov/genbank (accessed 14.06.2023).
4. *ENA*. URL: https://www.ebi.ac.uk/ena/browser/home (accessed 14.06.2023).
5. *CNGBdb*. URL: https://db.cngb.org (accessed 14.06.2023).
6. Liu B., Liu K., Zhang H., Zhang L., Bian Y., Huang L. CoV-Seq, a new tool for SARS-CoV-2 genome analysis and visualization: development and usability study. *J. Med. Internet Res*. 2020. V. 22. No. 10. Article No. e22299. doi: 10.2196/22299

7. Cleemput S., Dumon W., Fonseca V., Abdool Karim W., Giovanetti M., Alcantara L.C., Deforche K., de Oliveira T. Genome detective coronavirus typing tool for rapid identification and characterization of novel coronavirus genomes. *Bioinformatics*. 2020. V. 36. No. 11. P. 3552–3555. doi: 10.1093/bioinformatics/btaa145
8. Seong D.Y., Park J., Yi K., Hong D. Systematic guidelines for effective utilization of COVID-19 databases in genomic, epidemiologic, and clinical research. *Viruses*. 2023. V. 15. No. 3. Article No. 692. doi: 10.3390/v15030692
9. Edgar R.C. Taylor J., Lin V., Altman T., Barbera P., Meleshko D., Lohr D., Novakovsky G., Buchfink B., Al-Shayeb B. et al. Petabase-scale sequence alignment catalyses viral discovery. *Nature*. 2022. V. 602. P. 142–147. doi: 10.1038/s41586-021-04332-2
10. Gorbalenya A.E., Siddell S.G. Recognizing species as a new focus of virus research. *PLoS Pathog*. 2021. V. 17. No. 3. Article No. e1009318. doi: 10.1371/journal.ppat.1009318
11. Höper D., Wylezich C., Beer M. Loeffler 4.0: diagnostic metagenomics. *Adv. Virus Res.* 2017. V. 99. P. 17–37. doi: 10.1016/bs.aivir.2017.08.001
12. Greninger A.L. A decade of RNA virus metagenomics is (not) enough. *Virus Res*. 2018. V. 244. P. 218–229. doi: 10.1016/j.virusres.2017.10.014
13. Zhang Y.Z., Shi M., Holmes E.C. Using metagenomics to characterize an expanding virosphere. *Cell*. 2018. V. 172. No. 6. P. 1168–1172. doi: 10.1016/j.cell.2018.02.043
14. Adams M.J., Lefkowitz E.J., King A.M.Q., Harrach B., Harrison R.L., Knowles N.J., Kropinski A.M., Krupovic M., Kuhn J.H., Mushegian A.R. et al. 50 years of the International Committee on Taxonomy of Viruses: progress and prospects. *Arch. Virol*. 2017. V. 162. P. 1441–1446. doi: 10.1007/s00705-016-3215-y
15. Siddell S.G., Walker P.J., Lefkowitz E.J., Mushegian A.R., Adams M.J., Dutilh B.E., Gorbalenya A.E., Harrach B., Harrison R.L., Junglen S. et al. Additional changes to taxonomy ratified in a special vote by the International Committee on Taxonomy of Viruses (October 2018). *Arch. Virol*. 2019. V. 164. P. 943–946. doi: 10.1007/s00705-018-04136-2
16. Spaan W.J.M., Brian D., Cavanagh D., de Groot R.J., Enjuanes L., Gorbalenya A.E., Holmes K.V., Masters P., Rottier P., Taguchi F. et al. Coronaviridae. In: *Virus taxonomy. Eighth report of the International Committee on Taxonomy of Viruses*. Eds. Fauquet C.M., et al. Elsevier, Academic Press., 2005. P. 947–964. doi: https://doi.org/10.1016/B978-0-12-249951-7.50015-8
17. 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. No. 5. P. 668–674.
18. Walker P.J., Siddell S.G., Lefkowitz E.J., Mushegian A.R., Adriaenssens E.M., Alfenas-Zerbini P., Davison A.J., Dempsey D.M., Dutilh B.E., García M.L., et al. Changes to virus taxonomy and to the international code of virus classification and nomenclature ratified by the International Committee on Taxonomy of Viruses (2021). *Arch. Virol*. 2021. V. 166. No. 9. P. 2633–2648. doi: 10.1007/s00705-021-05156-1
19. Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol*. 1981. V. 17. P. 368–376. doi: 10.1007/BF01734359
20. Felsenstein J. *Inferring Phylogenies*. Sunderland, MA: Sinauer Associates, 2003. 664 p.
21. Nguyen L.T., Schmidt H.A., von Haeseler A., Minh B.Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum likelihood phylogenies. *Mol. Biol. Evol*. 2015. V. 32. No. 1. P. 268–274. doi: 10.1093/molbev/msu300
22. Katoh K., Rozewicki J., Yamada K.D. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Brief. Bioinform*. 2019. V. 20. No. 4. P. 1160–1166. doi: 10.1093/bib/bbx108

23. Mavrodiev E.V., Tursky M.L., Mavrodiev N.E., Schroder L., Laktionov A.P., Ebach M.C., Williams D.M. On classification and taxonomy of coronaviruses (*Riboviria*, *Nidovirales*, *Coronaviridae*) with special focus on severe acute respiratory syndrome-related coronavirus 2 (SARS-CoV-2). *Math. Biol. Bioinf*. 2022. V. 17. No. 2. P. 289–311. doi: 10.17537/2022.17.289
24. Kitching I.J., Forey P., Forey P.L., Humphries C., Williams D.M. *Cladistics, the Theory and Practice of Parsimony Analysis*. Oxford and New York: Oxford University Press, 1998. 228 p.
25. Nelson G., Platnick N. Three-taxon statements, a more precise use of parsimony? *Cladistics*. 1991. V. 7. No. 4. P. 351–366. doi: 10.1111/j.1096-0031.1991.tb00044.x
26. Creevey C.J., McInerney J.O. Trees from trees: construction of phylogenetic supertrees using Clann. In: *Bioinformatics for DNA sequence analysis*. Ed. Posada D. New York: Springer Humana Press, 2009. P. 139–161.
27. 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
28. Щелканов М.Ю., Попова А.Ю., Дедков В.Г., Акимкин В.Г., Малеев В.В. История изучения и современная классификация коронавирусов (*Nidovirales*: *Coronaviridae*). *Инфекция и иммунитет*. 2020. Т. 10. № 2. С. 221–246. doi: 10.15789/2220-7619-HOI-1412
29. Борисова Н.И., Котов И.А., Колесников А.А., Каптелова В.В., Сперанская А.С., Кондрашева Л.Ю., Тиванова Е.В., Хафизов К.Ф., Акимкин В.Г. Мониторинг распространения вариантов SARS-CoV-2 (*Coronaviridae*: *Coronavirinae*: *Betacoronavirus*; *Sarbecovirus*) на территории Московского региона с помощью таргетного высокопроизводительного секвенирования. *Вопросы вирусологии*. 2021. Т. 66. № 4. С. 269–278. doi: 10.36233/0507-4088-72
30. Vlasova A.N., Saif L.J. Bovine coronavirus and the associated diseases. *Front. Vet. Sci*. 2021. V. 8. Article No. 643220. doi: 10.3389/fvets.2021.643220
31. 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*: *Betacoronavirus*-1). *Вопросы вирусологии*. 2022. Т. 67. № 5. С. 465–474. doi: 10.36233/0507-4088-141
32. Sayers E.W., Cavanaugh M., Clark K., Ostell J., Pruitt K.D., Karsch-Mizrachi I. GenBank. *Nucleic Acids Res*. 2019. V. 47. No. D1. P. D94–D99. doi: 10.1093/nar/gky989
33. Sayers E.W., Beck J., Bolton E.E., Bourexis D., Brister J.R., Canese K., Comeau D.C., Funk K., Kim S., Klimke W., et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*. 2021. V. 49. No. D1. P. D10-D17. doi: 10.1093/nar/gkaa892

Рукопись поступила в редакцию 19.06.2023, переработанный вариант поступил 29.06.2023.
Дата опубликования 15.07.2023.

Choice of Target in the Genomes of Prototypic Strains to Recognize Subgenus of Coronaviruses

Chaley M.B.[1], Kutyrkin V.A.[2]

[1]*Institute of Mathematical Problems of Biology RAS, – Branch of Keldysh Institute of Applied Mathematics RAS, Pushchino, Russia*

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

Abstract. Targeted approach to recognition of coronavirus subgenus on the base of codon frequency distribution in the N-gene of nucleocapsid protein was proposed in the work. Deviation of codon frequency distribution in the N-gene of coronavirus genome analyzed from the same distributions for the 67 prototypic strains, which characterize the 23 subgenera in the four coronavirus genera, is calculated on the base of statistics in the approach proposed. The smallest value of such a deviation from certain prototypic strain points at subgenus to which this strain belongs. The approach proposed appeared to be effective and supports significance for recognizing coronavirus subgenus at least 99 %. Populations of the 38 and 7 codons providing for needed efficiency level were selected out of all codons of the genetic code in accordance with their frequency distribution. The codons from the populations outlined fix taxonomic structure of coronavirus subgenus.

Key words: *coronavirus subgenus, targeted approach, prototypic strains of coronavirus, coronavirus N-gene, N-gene codon frequency distribution.*