Математическая биология и биоинформатика
2022. Т. 17. № 1. С. 10–27. doi: 10.17537/2022.17.10

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

Распознавание рода коронавируса на основе прототипных штаммов

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. Предложен вариантный подход к распознаванию рода коронавируса на основе распределения частот кодонов в ORF1ab и генах структурных белков (S, M и N). Метод основан на модифицированной статистике, ранее показавшей свою эффективность при распознавании видов флавивирусов. Вариантный подход использует для распознавания рода коронавирусов как различные комбинации нескольких генов коронавируса, так и отдельные гены. Род коронавируса окончательно определяется по результатам анализа всех рассматриваемых вариантов. Предлагаемый метод разработан с помощью обучающей выборки геномов прототипных штаммов коронавирусов родов *Alphacoronavirus*, *Betacoronavirus*, *Deltacoronavirus* и *Gammacoronavirus*. Использование вариантного подхода для распознавания рода коронавируса показало его высокую достоверность на уровне 95 %. Среди всех вариантов совместного анализа наибольшую надежность (98 %) показало использование распознавания рода коронавирусов на основе частот кодонов ORF1ab. Вариантный подход выявил явление мозаичности в геномах коронавирусов при распознавании рода, когда результаты распознавания рода по отдельным генам расходились с окончательным определением рода коронавируса. Такое явление, по-видимому, отражает гомологичные рекомбинации генов между коронавирусами различных видов и пластичность генома коронавирусов в эволюционных процессах.

Ключевые слова: *геном коронавируса, ORF1ab, S-ген, M-ген, N-ген, статистический анализ, вариантный подход к распознаванию рода коронавируса.*

ВВЕДЕНИЕ

Коронавирусы (CoV, семейство *Coronaviridae*) представляют собой обширное семейство РНК-содержащих вирусов, инфицирующих млекопитающих, в том числе, человека, птиц и земноводных. Впервые коронавирус был выделен в 1931 г. как вирус, вызывающий инфекционный бронхит у цыплят (*Infectious bronchitis virus* – IBV), который в настоящее время носит название коронавирус птиц (*Avian coronavirus* – ACoV) [1]. Впоследствии оказалось, что коронавирусы вызывают заболевания различной тяжести у многих других видов птиц, как домашних, так и диких. Коронавирусы также вызывают гастроэнтериты и поражают респираторный тракт у домашних свиней, крупного рогатого скота, у кошек и собак, верблюдов, землероек, крыс и мышевидных грызунов, ежей, летучих мышей, китов, и др. [2]. В настоящее время известны семь коронавирусов, вызывающих заболевания у человека. Обычно их разделяют на две группы: HCoV 229E, HCoV NL63, HCoV HKU1 и HCoV OC43, при которых заболевание протекает в основном по типу ОРВИ; и особо опасные, такие как

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

SARS-CoV-1 (*severe acute respiratory syndrome-related coronavirus*), MERS-CoV (*middle east respiratory syndrome coronavirus*) и SARS-CoV-2 (*severe acute respiratory syndrome 2 coronavirus*), вызвавший пандемию COVID-19 [3]. Более подробную информацию об этих коронавирусах и их природных источниках можно найти, например, в [4, 5].

Таксономическая классификация коронавирусов сложилась к 2011 г., когда в каталоге Международного комитета по таксономии вирусов (International Committee on Taxonomy of Viruses – ICTV) род Coronavirus перешел в статус подсемейства *Coronavirinae*, включающего четыре рода: *Alphacoronavirus*, *Betacoronavirus*, *Deltacoronavirus* и *Gammacoronavirus* [6]. Таксономическая структура коронавирусов была существенно пересмотрена в 2018 г. в результате введения понятие подрода и математической оценки некоторых рангов (подрод, род, семейство) по филогенетическому дереву, построенному методом наибольшего правдоподобия после множественного выравнивания полноразмерных геномов всех известных представителей семейства. Исходя из суммы различий по совокупности признаков между двумя узлами филогенетического дерева, измеренной вдоль соединяющих их ребер, были установлены соответствующие пороги: для подрода – 0.186, для рода – 0.789 и для подсемейства – 1.583. С расширением множества известных коронавирусов, эти значения могут постепенно меняться. Кроме того, обнаруженный в 2018 г. методом метагеномного анализа новый коронавирус [7], получивший название летовирус узкоротых квакш (*Microhyla fissipes*) 1-го типа был выделен ICTV в отдельное подсемейство *Letovirinae*, в то время как все ранее известные коронавирусы составили подсемейство *Orthocoronavirinae* [8].

В настоящей работе рассматриваются коронавирусы только подсемейства *Orthocoronavirinae*. Установлено, что это подсемейство связано с крылатыми животными, являющихся резервуаром коронавирусов: рукокрылые – для родов *Alphacoronavirus* (α–CoV) и *Betacoronavirus* (β–CoV), птицы – для родов *Deltacoronavirus* (δ–CoV) и *Gammacoronavirus* (γ–CoV) [4].

Свое название коронавирусы получили после того, как с помощью электронной микроскопии была выявлена их форма, напоминающая солнечную корону [9]. Такая форма вызвана тем, что вирион коронавируса обрамляют булавовидные шипы поверхностного гликопротеина S (spike protein, или S-белка). S-белок отвечает за проникновение вируса в клетку, связываясь с её определенными трансмембранными рецепторами. В липидную оболочку вириона заякорены еще два структурных белка: каналообразующий оболочечный белок E (envelope protein) и трансмембранный белок M (membrane protein). У коронавируса есть нуклеокапсид, внутри которого находятся вирионная геномная (+)РНК (РНК позитивной полярности, на которой сразу может идти синтез белка на рибосомах инфицированной клетки), и связанный с нуклеокапсидом N-белок (nucleocapsid protein) [10, 11].

Геномная РНК вируса выступает в качестве мРНК для синтеза двух протяженных полипротеинов pp1a и pp1ab длиной около 4000 и 7000 аминокислотных остатков соответственно. Полипротеин pp1ab образуется в результате игнорирования рибосомой стоп-сигнала из-за шпильки, смещающей рамку считывания, как правило на –1 нуклеотид. Рамки считывания таких полипротеинов в геноме соответственно обозначаются как ORF1a и ORF1ab. Далее из полипротеинов pp1a и pp1ab высвобождаются две протеазы: главная протеаза Mpro (major protease) и папаин-подобная протеаза PLpro (papain-like protease), которые затем расщепляют сплошной полипротеин pp1a на 11 и полипротеин pp1ab на 16 отдельных неструктурных белков (nsp – non-structural protein). Эти неструктурные белки выполняют различные важные функции в жизненном цикле вируса [2, 11].

Например, nsp12 – РНК-зависимая РНК-полимераза RdRp (RNA-dependent RNA-polymerase) синтезирует (–)РНК, комплементарную геномной (+)РНК, которая, в свою очередь, выступает в качестве матрицы для синтеза геномных РНК для дочерних

вирионов. Помимо этого, RdRp на матрице вирионной геномной (+)РНК синтезирует серию субгеномных (–)РНК, далее используемых этой же полимеразой для синтеза субгеномных матричных (+)РНК, с которых считываются некоторые неструктурные и структурные белки, в частности, S-, M- и N-белок. Коронавирусы имеют самый длинный геном из всех геномов РНК-вирусов, который составляет порядка 30000 нукл., из которых около 2/3 занимает ORF1ab.

Выравнивание последовательности генома коронавирусов показывает гомологию 58 % в области, кодирующей неструктурные белки, 43 % – в области, кодирующей структурные белки и 54 % – на уровне всего генома, предполагая, что неструктурные белки более консервативны, а структурные белки более разнообразны и поддерживают адаптацию вируса к новым хозяевам [11].

Естественная изменчивость генома коронавирусов обеспечивается спонтанными мутациями в ходе его репликации и гомологичной рекомбинацией с другими геномами коронавирусов. Коронавирусы накапливают меньше мутаций, чем большинство РНК-вирусов, потому что они кодируют фермент nsp14, исправляющий ошибки репликации [12]. Наиболее вариабельным у коронавирусов является ген S-белка, отвечающий за их межвидовую передачу, вирулентность и контагиозность. Скорость его мутаций у разных видов коронавирусов различается. Например, у коронавирусов IBV, вызывающих инфекции респираторного тракта у птиц, скорость мутаций в S-гене на порядок выше ($0.3-0.6\times10^{-2}$ замен в год/сайт [13]), чем у коронавирусов HCoV 229E, вызывающих относительно легко протекающие респираторные заболевания человека (3×10^{-4} замен в год/сайт [14]). При такой скорости мутирования вид коронавируса HCoV NL63, провоцирующий развитие кашля у детей, по оценкам, разошелся со своим предком HCoV 229E в 11 веке [14]. Однако, следует понимать, что мутации охватывают гораздо больше генов, чем только ген белка S.

Наибольшую пластичность геному коронавируса создает возможность обширной гомологичной рекомбинации с геномами других коронавирусов, циркулирующих в одних и тех же хозяевах. По этой причине геном коронавирусов в природных резервуарах всегда является мозаичным. В основе такой рекомбинации лежит сложный жизненный цикл коронавирусов, включающий в себя образование набора субгеномных РНК. Точный механизм генетической рекомбинации у коронавирусов остается неясным: рекомбинационные сайты кажутся расположенными по геному случайным образом [15, 16].

В связи с продолжающейся пандемией COVID-19 полногеномное секвенирование изолятов вируса, хотя и рассматривается как наилучший способ генетического анализа по распространению новых патогенных штаммов мутирующего вируса SARS-CoV-2 [17], однако для отслеживания мутантных штаммов SARS-CoV-2 и других видов коронавирусов возможен более быстрый и экономичный подход таргетного секвенирования [18]. При таком подходе секвенированию подвергается только один геномный фрагмент коронавируса, например, фрагмент гена шиповидного S-белка, учитывая его способность к связыванию с клеточными рецепторами при инфицировании.

Вот уже два с лишним года не затухает пандемия, вызванная коронавирусом SARS-CoV-2, переживая все новые волны распространения мутантных штаммов. Как никогда остро встал вопрос оценки рисков возникновения возможных новых эпидемических источников, инициированных коронавирусами. Чем разнообразнее и точнее инструментарий идентификации и анализа коронавирусов, тем скорее и успешнее может быть выполнена такая оценка. Исходя из предыдущего опыта создания статистического метода для быстрого распознавания вида флавивируса (среди которых вирус желтой лихорадки, лихорадки денге, лихорадки Западного Нила и др.) по известной последовательности его генома [19], в настоящей работе предлагается надежный метод

распознавания рода коронавируса. Этот метод получил название вариантного подхода к распознаванию рода коронавируса. Вариантный подход использует для распознавания как комбинации нескольких структурных генов коронавируса, так и отдельные гены, секвенирование которых займет меньше времени, чем определение последовательности всего генома.

Вариантный подход к распознаванию рода коронавируса

По результатам исследования структуры кодирования ORF1ab, S, M и N генов коронавирусов, было показано, что в последовательностях этих генов наблюдается скрытая триплетная профильная периодичность [20]. Такая периодичность характеризуется 3-профильной матрицей, в которой в каждом из тех столбцов стоит распределение букв четырехбуквенного алфавита ДНК. Для каждой структурной единицы из одного рода создавалась своя усредненная 3-профильная матрица. Было отмечено, что для всех родов коронавирусов (α–CoV, β–CoV, δ–CoV и γ–CoV) в гене мембранного белка M скрытая 3-профильная периодичность слабо выражена. Поэтому для этого гена во всех родах коронавирусов скрытые 3-профильные матрицы не вычислялись. Для каждой из трех остальных структурных единиц сравнение усредненных 3-профильных матриц между родами не выявило отличий, которые можно было бы использовать для создания методов или процедур идентификации рода коронавируса.

Ранее, при создании метода распознавания видов флавивирусов использовалась статистика, основанная на сравнении распределений кодонов в гене полипротеина флавивируса [19]. Если геном флавивируса обычно представлен непрерывной цепочкой кодирующих последовательностей вирусных генов, то в геноме коронавируса более сложная структурная организация генов. Кроме того, для создания метода идентификации из генома коронавируса были отобраны отдельные, особо значимые гены: ген ORF1ab; ген нуклеопротеина N – белка, формирующего комплекс с РНК-геномом вируса; ген мембранного M-белка; ген S-белка, образующего шипы на поверхности вирусного вириона. Эти гены не содержат интронов и из них можно сформировать псевдо-ген полипротеина, аналогичный гену полипротеина флавивируса. Такой псевдо-ген состоит из существенно отличающихся по длине структурных единиц (генов), которые вносят неравноценный вклад в общее распределение частот кодонов. Поэтому при создании статистики для распознавания рода коронавируса учитывалась длина каждой из структурных единиц псевдо-гена.

Целью настоящей работы было создание достаточно простого, быстрого и надежного метода распознавания рода коронавируса. При разработке метода мы опирались на опыт предыдущей работы по эффективному распознаванию вида флавивируса на основе распределения частот кодонов в кодирующей части генома, транслируемой единым полипротеином. По аналогии с геном полипротеина флавивируса можно рассмотреть псевдо-ген, составленный из отдельных генов структурных белков коронавируса. В качестве таких генов в работе были выбраны ген полипротеина pp1ab – ORF1ab, кодирующий 16 неструктурных белков, и три гена структурных белков: S-ген, M-ген и N-ген. Совместное распределение частот кодонов этих генов, учитывающее длину каждого в псевдогене, использовалось в первом варианте распознавания рода коронавируса. Для этого первого варианта V_1 использовалось обозначение S+N+M+ORF1ab. Поскольку каждый из генов, входящих в вариант V_1, имеет свою степень пластичности и консервативности, для распознавания рода коронавируса дополнительно были исследованы 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.

Результат распознавания коронавируса в вариантном подходе представляется в виде вариантной строки $\boldsymbol{v} = (v_1, v_2, v_3, v_4, v_5, v_6)$, где v_i – результат распознавания рода коронавируса с помощью варианта V_i для $i = \overline{1,6}$, т.е. $v_i \in \{\alpha, \beta, \delta, \gamma\} - \mathrm{CoV}$.

Однозначный результат вариантного подхода для анализируемого генома коронавируса определяется в случае четырех одинаковых компонент в вариантной строке. Такой подход позволяет сравнить надежность всех рассматриваемых вариантов распознавания и степень консервативности рассматриваемых генов. Мозаичность вариантной строки, т.е. наличие в ней неодинаковых компонент, может иметь различное биологическое объяснение, в частности, с помощью предположения о возможной гомологичной рекомбинации между генами коронавирусами различных видов [21–23].

МАТЕРИАЛЫ

Для создания в каждом роде коронавирусов усредненного распределения кодонов использовались соответствующие гены прототипных штаммов рода [2]. Данные о кодах доступа GenBank для набора последовательностей геномов прототипных штаммов, из которых выделены анализируемые в работе гены ORF1ab, S, M и N, представлены в таблице 1.

В настоящее время род *Alphacoronavirus* подразделяется на 12 подродов (22 прототипных штамма), род *Betacoronavirus* – на пять подродов (28 прототипных штаммов), род *Deltacoronavius* – на четыре подрода (10 прототипных штаммов) и род *Gammacoronavirus* – на два подрода (7 прототипных штаммов). В настоящей работе для удобства анализа и описания результатов вводится отдельная нумерация прототипных штаммов каждого рода (см. табл. 1).

Таблица 1. Коды доступа прототипных (п/т) штаммов коронавирусов в GenBank и их нумерация, принятая в работе

Род (Genus)	Подрод (Subgenus)	Вид (Species)	GenBank ID п/т штамма	N п/т штамма рода
Alphacoronavirus	*Colacovirus*	Коронавирус летучих мышей CDPHE15 (Bat coronavirus CDPHE15)	NC_022103	1
	Decacovirus	Альфакоронавирус больших подковоносов HuB2013 (*Rhinolophus ferrumequinum* alphacoronavirus HuB2013)	NC_028814	2
		Коронавирус летучих мышей HKU10 (Bat coronavirus HKU10)	NC_018871	3
	Duvinacovirus	Коронавирус человека 229E (Human coronavirus 229E)	NC_002645	4
	Luchacovirus	Коронавирус крыс Лунцюань Rn (Lucheng Rn rat coronavirus)	NC_032730	5
	Minacovirus	Коронавирус норок 1-го типа (Mink coronavirus 1)	NC_023760	6
		Коронавирус хорьков (Ferret coronavirus)	KX512809	7
			KX512810	8
	Minunacovirus	Коронавирус длиннокрылов 1-го типа (Miniopterus bat coronavirus 1)	EU420138	9
		Коронавирус длиннокрылов HKU8 (Miniopterus bat coronavirus HKU8)	NC_010438	10
	Miotacovirus	Альфакоронавирус азиатских рыбоядных ночниц Sax-2011 (*Myotis ricketti* alphacoronavirus Sax2011)	NC_28811	11

Род (Genus)	Подрод (Subgenus)	Вид (Species)	GenBank ID п/т штамма	N п/т штамма рода
	Nyctacovirus	Альфакоронавирус китайских вечерниц SC2013 (*Nyctacus velutinus* alhacoronavirus SC2013)	NC_028833	12
	Pedacovirus	Вирус эпизоотической диареи свиней (Porcine epidemic diarrhea virus)	KT323979	13
		Коронавирус домовых гладконосов 512 (Scotophilus bat coronavirus 512)	NC_009657	14
	Rhinacovirus	Коронавирус подковоносов HKU2 (Rhinolopus bat coronavirus HKU2)	NC_009988	15
	Setracovirus	Коронавирус человека NL63 (Human coronavirus NL63)	AY567487	16
		NL63-подобный коронавирус BtKYNL63-9b (NL63-related bat coronavirus BtKYNL63-9b)	KY073745	17
	Tegacovirus	Альфакоронавирус 1-го типа (Alphacoronavirus 1)	KP981644	18
			FJ938051	19
			AY994055	20
			KR270796	21
			NC_038861	22
Betacoronavirus	*Embecovirus*	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	Бетакоронавирус листоносов Пратта Zhejiang2013 (Bat Hp-betacoronavirus Zhejiang2013)	NC_025217	13
	Merbecovirus	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	Коронавирус ночных крыланов GCCDC1 (Rousettus bat coronavirus GCCDC1)	NC_030886	20
		Коронавирус ночных крыланов HKU9 (Rousettus bat coronavirus HKU9)	NC_009021	21
	Sarbecovirus		MG772933	22

Род (Genus)	Подрод (Subgenus)	Вид (Species)	GenBank ID п/т штамма	N п/т штамма рода
		Коронавирус китайских подковоносов (*Rhinolophus sinicus* coronavirus)	MG772934	23
		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	*Andecovirus*	Коронавирус свиязей HKU20 (Wigeon coronavirus KKU20)	NC_016995	1
	Buldecovirus	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	Короновирус квакв HKU19 (Night heron coronavirus HKU19)	NC_016994	9
	Moordecovirus	Короновирус камышниц HKU21 (Common morgen coronavirus HKU21)	NC_016996	10
Gammacoronavirus	*Cegacovirus*	Коронавирус китообразных (Cetacean coronavirus)	EU111742	1
			KF793826	2
	Igacovirus	Коронавирус птиц (Avian coronavirus)	KF696629	3
			GQ504724	4
			NC_010800	5
			AY641576	6
			MK423877	7

Для каждого из четырех генов ORF1ab, S, M и N последовательности генома прототипного штамма коронавируса рассматривалось распределение частот кодонов. Количество анализируемых геномов прототипных штаммов каждого рода составляло: 22, 28, 10 и 7 для родов α–CoV, β–CoV, δ–CoV и γ–CoV, соответственно (см. табл. 1).

Помимо геномов прототипных штаммов в работе использовались геномы коронавирусов четырех родов, полученные из базы данных нуклеотидных последовательностей GenBank выпуска 237 [24]. Количество всех анализируемых последовательностей геномов (вместе с прототипными штаммами) для родов α–CoV, β–CoV, δ–CoV и γ–CoV составляло: 924, 1954, 19 и 345, соответственно.

МЕТОДЫ

Введем количественные характеристики, необходимые для предлагаемого в работе метода распознавания рода коронавируса, к которому относится анализируемая последовательность генома. Средняя длина по всем анализируемым в работе генам (их кодирующим районам – CDS) для каждого рода коронавирусов показана в таблице 2.

Таблица 2. Средняя длина CDS анализируемых генов коронавирусов в каждом роде

ген	**α–CoV**, нукл	**β–CoV**, нукл	**δ–CoV**, нукл	**γ–CoV**, нукл
ORF1ab	$L^{\alpha}_{ORF} = 20284$	$L^{\beta}_{ORF} = 21275$	$L^{\delta}_{ORF} = 18794$	$L^{\gamma}_{ORF} = 19861$
S	$L^{\alpha}_{S} = 4105$	$L^{\beta}_{S} = 3926$	$L^{\delta}_{S} = 3590$	$L^{\gamma}_{S} = 3591$
N	$L^{\alpha}_{N} = 1270$	$L^{\beta}_{N} = 1267$	$L^{\delta}_{N} = 1044$	$L^{\gamma}_{N} = 1229$
M	$L^{\alpha}_{M} = 692$	$L^{\beta}_{M} = 669$	$L^{\delta}_{M} = 679$	$L^{\gamma}_{M} = 678$
S+N+M+ORF1ab	$L^{\alpha} = 26351$	$L^{\beta} = 27137$	$L^{\delta} = 24107$	$L^{\gamma} = 25359$

В качестве обучающей выборки для распознавания рода использовались известные прототипные штаммы короновирусов, коды доступа которых в GenBank исходно были даны в работе [2] и представлены в таблице 1. Для рода *Alphacoronavirus* количественные характеристики имеют вид:

$n = \overline{1,22}$ – номер прототипного штамма для рода α–CoV;

$\boldsymbol{p}^{\alpha}_{ORF}(n)$, $\boldsymbol{p}^{\alpha}_{S}(n)$, $\boldsymbol{p}^{\alpha}_{M}(n)$, $\boldsymbol{p}^{\alpha}_{N}(n)$ – распределение частот кодонов в CDS генов ORF1ab, S, M, N для прототипа с номером *n*;

$l^{\alpha}_{ORF}(n)$, $l^{\alpha}_{S}(n)$, $l^{\alpha}_{M}(n)$, $l^{\alpha}_{N}(n)$ – длины CDS генов ORF1ab, S, M, N в прототипе с номером *n*;

$l^{\alpha}(n) = l^{\alpha}_{ORF}(n) + l^{\alpha}_{S}(n) + l^{\alpha}_{M}(n) + l^{\alpha}_{N}(n)$ – сумма длин соответствующих генов в прототипе с номером *n*.

В рамках указанных выше обозначений для усредненного взвешенного распределения частот кодонов $\boldsymbol{P}^{\alpha}$ рода α–CoV используется формула:

$$\boldsymbol{P}^{\alpha} = \frac{1}{22}\sum_{n=1}^{22}\left(\frac{l^{\alpha}_{ORF}(n)}{l^{\alpha}(n)}\boldsymbol{p}^{\alpha}_{ORF}(n) + \frac{l^{\alpha}_{S}(n)}{l^{\alpha}(n)}\boldsymbol{p}^{\alpha}_{S}(n) + \frac{l^{\alpha}_{M}(n)}{l^{\alpha}(n)}\boldsymbol{p}^{\alpha}_{M}(n) + \frac{l^{\alpha}_{N}(n)}{l^{\alpha}(n)}\boldsymbol{p}^{\alpha}_{N}(n)\right). \quad (1)$$

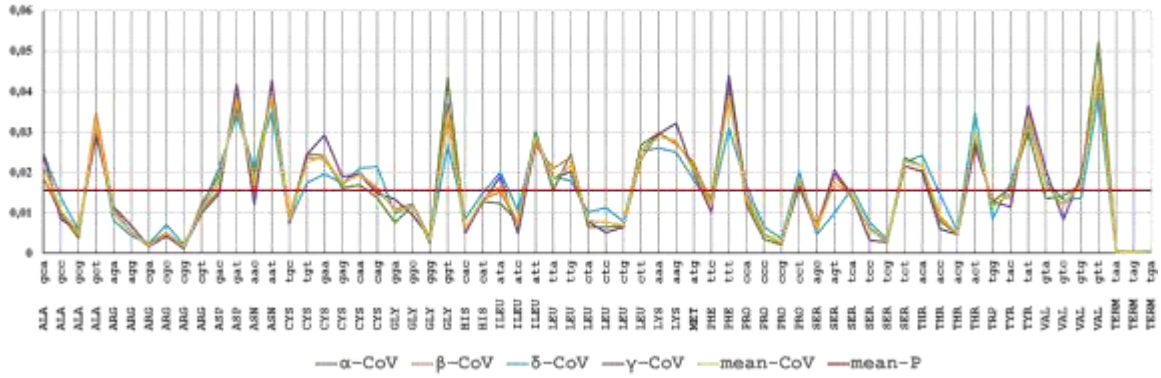

Рис. 1. Усредненные взвешенные распределения $\boldsymbol{P}^{x}$ ($x \in \{\alpha,\beta,\delta,\gamma\}$) частот кодонов для первого варианта объединения генов S+N+M+ORF1ab из геномов прототипных штаммов коронавирусов для родов α–CoV, β–CoV, δ–CoV и γ–CoV (см. табл. 1). Среднее арифметическое распределение частот кодонов для всех родов показано желтой линией (mean-CoV). Красной линией (mean-P) показано среднее значение частот кодонов в таком среднем арифметическом распределении. Триплеты кодонов указаны в паре с названием кодируемой аминокислоты (нижняя линия подписей по горизонтальной оси).

Аналогично формуле (1) рассчитываются усредненные взвешенные распределения частот кодонов $\boldsymbol{P}^{\beta}$, $\boldsymbol{P}^{\delta}$ и $\boldsymbol{P}^{\gamma}$ для родов β–CoV, δ–CoV и γ–CoV, соответственно. На

рисунке 1 представлены графики усредненных взвешенных распределений частот кодонов *P* для рассматриваемых родов коронавирусов.

Результаты численных экспериментов показали, что распознавание рода коронавируса существенно улучшается, если из распределений кодонов исключить самые низкие частоты. Такими частотами оказались частоты всех трех стоп-кодонов (TERM) и двух (cga и cgg) из шести синонимичных кодонов аргинина (ARG), что учитывалось при вычислении усредненных распределений частот кодонов по формуле (1). В частности, для рода α–CoV формат усредненного взвешенного распределения $\boldsymbol{P}^{\alpha}$ имеет вид: $\boldsymbol{P}^{\alpha} = (P_1^{\alpha}, P_2^{\alpha}, \ldots, P_{59}^{\alpha})$.

Если рассматривается геном коронавируса неизвестного рода, то для его распределений частот кодонов (после исключения указанных выше кодонов) в генах ORF1ab, S, M и N используются обозначения: $\boldsymbol{p}_{\mathrm{ORF}}$, $\boldsymbol{p}_{\mathrm{S}}$, $\boldsymbol{p}_{\mathrm{M}}$ и $\boldsymbol{p}_{\mathrm{N}}$. Следовательно, аналогично формуле (1), взвешенное распределение $\boldsymbol{p}$ частот кодонов в этом геноме вычисляется согласно формуле:

$$\boldsymbol{p} = \frac{l_{\mathrm{ORF}}}{l}\boldsymbol{p}_{\mathrm{ORF}} + \frac{l_{\mathrm{S}}}{l}\boldsymbol{p}_{\mathrm{S}} + \frac{l_{\mathrm{M}}}{l}\boldsymbol{p}_{\mathrm{M}} + \frac{l_{\mathrm{N}}}{l}\boldsymbol{p}_{\mathrm{N}}, \qquad (2)$$

где l_{ORF}, l_{S}, l_{M}, l_{N} – длины соответствующих генов в рассматриваемом геноме коронавируса и $l = l_{\mathrm{ORF}} + l_{\mathrm{S}} + l_{\mathrm{M}} + l_{\mathrm{N}}$. Кроме того, формат распределения $\boldsymbol{p}$ имеет вид: $\boldsymbol{p} = (p_1, p_2, \ldots, p_{59})$.

Пусть $x \in \{\alpha,\beta,\delta,\gamma\}$, т.е. x – символ, обозначающий род коронавируса. Используя это обозначение, отклонение $D(\boldsymbol{P}^x, \boldsymbol{p})$ взвешенного распределения $\boldsymbol{p}$ частот кодонов в анализируемом геноме отдельного прототипного штамма или генома коронавируса неизвестного рода от усредненного взвешенного распределения $\boldsymbol{P}^x$ частот кодонов рода *x*–CoV вычисляется по формуле:

$$D(\boldsymbol{P}^x, \boldsymbol{p}) = \frac{1}{7}\sum_{i=1}^{59}\frac{|P_i^x - p_i|}{P_i^x}. \qquad (3)$$

Формулы (1)–(3) введены для случая, когда совместно рассматривают четыре гена ORF1ab, S, M и N, для чего вводится обозначение: V_1) S+N+M+ORF1ab. Аналогичные формулы также используются для вариантов объединения генов: V_2) S+N+M, V_3) S+N и для отдельных генов V_4) ORF1ab, V_5) S и V_6) N. Отметим, что при анализе отдельного гена N из 64 кодонов исключаются пять кодонов, среди которых три кодона терминации (TERM) и два кодона (tgt и tgc) цистеина (CYS), имеющие наименьшую частоту встречаемости только для N-генов коронавирусов. Среди отклонений $D(\boldsymbol{P}^{\alpha}, \boldsymbol{p})$, $D(\boldsymbol{P}^{\beta}, \boldsymbol{p})$, $D(\boldsymbol{P}^{\delta}, \boldsymbol{p})$ и $D(\boldsymbol{P}^{\gamma}, \boldsymbol{p})$ выбирается минимальное, которое предлагается как результат распознавания рода анализируемого генома коронавируса с помощью выбранного варианта объединения генов.

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Вначале рассмотрим результаты применения предлагаемого вариантного подхода к распознаванию рода коронавирусов на основе обучающей выборки из прототипных штаммов.

Распознавание рода коронавируса для прототипных штаммов

Результаты применения статистики (3) к обучающим выборкам для различных вариантов объединения генов прототипных штаммов показаны в таблице 3а и таблице 3б.

Таблица 3а. Количественные результаты правильного распознавания рода прототипных (п/т) штаммов коронавирусов с помощью различных вариантов объединения генов

Варианты объединения генов		**α–CoV** Всего 22 п/т штамма	**β–CoV** Всего 28 п/т штаммов	**δ–CoV** Всего 10 п/т штаммов	**γ–CoV** Всего 7 п/т штаммов
V_1	S+N+M+ORF1ab	20	25	9	7
V_2	S+N+M	22	24	9	7
V_3	S+N	22	25	9	7
V_4	ORF1ab	20	25	9	7
V_5	S	22	23	9	7
V_6	N	19	28	10	7

Из таблицы 3а следует, что применение предложенного метода распознавания рода коронавируса по обучающей выборке показало результат в рамках допустимой статистической погрешности. В среднем, надежность распознавания по всем шести вариантам составила 93 %.

Конкретный вид ошибок в распознавании каждого из четырех родов коронавирусов показан в таблице 3б. Для каждого рода указывается номер прототипного штамма по таблице 1 и ошибочная принадлежность к другому роду отмечается прописной буквой α, β, δ, γ в соответствии с названием рода *Alphacoronavirus* (α–CoV), *Betacoronavirus* (β–CoV), *Deltacoronavirus* (δ–CoV), *Gammacoronavirus* (γ–CoV). Так, например, пятый прототипный штамм рода *Alphacoronavirus* (см. табл. 1) распознан усредненным взвешенным распределением кодонов первого варианта объединения генов V_1, как относящийся к роду *Betacoronavirus*, что в соответствующей строке обозначено 5β.

Таблица 3б. Результаты ошибочного распознавания рода у прототипных штаммов четырех родов коронавирусов (α–CoV, β–CoV, δ–CoV и γ–CoV)

Варианты объединения генов	**α–CoV**					**β–CoV**					**δ–CoV**	**γ–CoV**
V_1	5β				18β	12γ		19δ	20δ		10γ	-
V_2						12α	13δ		20δ	21α	10β	-
V_3						12α	13δ			21α	10α	-
V_4	5β				18β	12γ		19δ	20δ		10β	-
V_5						12α	13δ	19δ	20α	21α	10α	-
V_6		11β	12β	17β							-	-

Распознавание рода среди последовательностей геномов коронавирусов в GenBank

В силу достаточно высокой надежности распознавания рода коронавирусов, полученной по обучающей выборке из прототипных штаммов, применим статистику (3) для распознавания рода последовательностей вирусных геномов из базы GenBank,

используя те же усредненные взвешенные распределения частот кодонов, полученные для прототипов каждого рода коронавирусов. Результаты такого распознавания рода коронавирусов для шести вариантов объединения генов показаны в таблице 4. В среднем, надежность распознавания по всем шести вариантам составила 95.6%, т.е. полученный результат находится в рамках стандартной статистической погрешности.

Таблица 4. Количественные результаты правильного распознавания рода последовательностей геномов коронавирусов (вместе с геномами рассматриваемых прототипных штаммов) из базы GenBank с помощью различных вариантов объединения генов

Варианты объединения генов		**α–CoV** Всего геномов: 924	**β–CoV** Всего геномов: 1954	**δ–CoV** Всего геномов: 19	**γ–CoV** Всего геномов: 345
V_1	S+N+M+ORF1ab	911	1883	16	342
V_2	S+N+M	903	1887	17	342
V_3	S+N	831	1898	17	342
V_4	ORF1ab	910	1906	16	343
V_5	S	898	1898	17	192
V_6	N	743	1924	19	343

Вариантный подход к распознаванию рода коронавирусов

Рассмотрим вариантный подход к распознаванию рода коронавируса, основанного на анализе результатов распознавания рода с помощью нескольких вариантов объединения генов, которые перечислены в таблице 3а и таблице 4. В вариантном подходе результат распознавания рода для последовательности вирусного генома выглядит в виде вариантной строки $\boldsymbol{v} = (v_1, v_2, v_3, v_4, v_5, v_6)$, где $v_i \in \{\alpha, \beta, \delta, \gamma\}$ – есть результат распознавания рода (*Alphacoronavirus*, *Betacoronavirus*, *Deltacoronavirus*, *Gammacoronavirus*) коронавируса с использованием *i*-го $(i = \overline{1,6})$ варианта объединения генов.

В подавляющем большинстве случаев компоненты строки вариантного распознавания представлены одной буквой. Однако, встречаются случаи, когда в вариантной строке распознавания встречаются разные компоненты, то есть вариантный подход проявляет некоторую мозаичность в распознавании рода по геному коронавируса.

В таблице 5 представлены количественные результаты распознавания рода с помощью вариантного подхода для анализируемых последовательностей коронавирусных геномов из GenBank. Сначала для каждого рода указано число правильных распознаваний по всем шести вариантам объединения генов. Явление мозаичности представлено в следующих двух строках таблицы, где указаны числа правильных распознаваний рода по пяти и четырем компонентам вариантной строки распознавания.

Для вариантного подхода можно рассмотреть следующее правило распознавания рода коронавируса по анализируемой геномной последовательности. Если среди шести компонент вариантной строки не менее четырех указывают на один и тот же род коронавируса, то этот род и признается в качестве результата распознавания. Последняя строка в таблице 5 суммирует количественные результаты распознавания четырех родов коронавирусов для последовательностей геномов из GenBank с использованием такого правила вариантного подхода.

Таблица 5. Результаты распознавания рода для отдельных последовательностей геномов коронавирусов из базы GenBank с помощью правила вариантного подхода

Число одинаковых компонент в вариантной строке, корректно указывающих род коронавируса	**α–CoV** Всего геномов 924	**β–CoV** Всего геномов 1954	**δ–CoV** Всего геномов 19	**γ–CoV** Всего геномов 345
6	687	1867	16	192
5	168	11	-	149
4	50	9	1	-
	распознано: 905	распознано: 1887	распознано: 17	распознано: 341

Как видно из таблицы 5, применение правила вариантного подхода гарантирует надежность распознавания рода коронавируса в 97 %. Следует отметить, что, согласно таблице 4, распознавание рода коронавируса только на основе гена ORF1ab обеспечивает надежность в 97.9 %. Однако, вариантный поход может способствовать выявлению структурных особенностей в генах коронавирусов.

Явление мозаичности генов в геномах коронавирусов при использовании вариантного подхода

В вариантном подходе при распознавании рода коронавируса отдельного генома число одинаковых компонент в вариантной строке может варьироваться от двух до шести. Однако, в случаях, когда не все компоненты одинаковы, пять одинаковых компонент встречаются наиболее часто. В качестве примера таких случаев рассмотрим результаты применения вариантного подхода к распознаванию рода *Gammacoronavirus* в специальной выборке 149 геномов коронавирусов этого рода из базы GenBank (см. табл. 6). Полная выборка анализируемых в работе геномов коронавирусов рода *Gammacoronavirus* из базы GenBank составляет 345 последовательностей (см. табл. 5).

Таблица 6 представляет вариантные строки, полученные при распознавании рода коронавируса для 149 последовательностей геномов из GenBank в результате применения вариантного подхода. Из таблицы 6 следует, что среди всех 149 геномов вариантные строки имеют пять одинаковых компонент (с первой по четвёртую и шестую), правильно определяющих род *Gammacoronavirus*. Ошибка в определении рода во всех этих вариантных строках находится в пятой компоненте, соответствующей S-гену спайк-белка. При этом пятая компонента указывает только на род *Alphacoronavirus* (15.4 %) или род *Betacoronavirus* (84.6 %). Аналогичное явление мозаичности можно наблюдать при распознавании рода *Alphacoronavirus* (см. таблицу 7).

Пример явления мозаичности в геномах коронавирусов из рода *Alphacoronavirus*, связанный, главным образом, с N-геном белка нуклеокапсида, представлен в таблице 7. Несмотря на то, что для родов *Betacoronavirus*, *Deltacoronavirus* и *Gammacoronavirus* достоверность распознавания на основе только одного N-гена белка нуклеокапсида имеет, практически, 100%-ю надежность (см. табл. 4), при распознавании рода *Alphacoronavirus* выявляется 131 вариантная строка, где шестая компонента, соответствующая N-гену, ошибочно указывает на род *Betacoronavirus*. Кроме того, в 39 вариантных строках содержится такое же указание на *Betacoronavirus* только в двух компонентах: в шестой и в третьей, отвечающей за вариант объединения S+N (см. табл. 7). Таким образом, ген нуклеокапсида не позволяет однозначно идентифицировать роды α–CoV и β–CoV.

Таблица 6. Явление мозаичности в распознавании коронавирусных геномов рода *Gammacoronavirus* с помощью вариантного подхода (см. табл. 3а, табл. 3б)

GenBank ID		$v_1 v_2 v_3 v_4 v_5 v_6$	**GenBank ID**		$v_1 v_2 v_3 v_4 v_5 v_6$	**GenBank ID**		$v_1 v_2 v_3 v_4 v_5 v_6$
1	MK071267	γ γ γ γ β γ	51	KX348115	γ γ γ γ β γ	101	KP118892	γ γ γ γ β γ
2	KX258195	γ γ γ γ β γ	52	KP118882	γ γ γ γ β γ	102	KP118888	γ γ γ γ β γ
3	KY933089	γ γ γ γ β γ	53	KP118881	γ γ γ γ β γ	103	KX252787	γ γ γ γ α γ
4	KY626044	γ γ γ γ β γ	54	KX364294	γ γ γ γ α γ	104	KX236001	γ γ γ γ β γ
5	KY626045	γ γ γ γ α γ	55	JX195175	γ γ γ γ β γ	105	EU637854	γ γ γ γ β γ
6	AY338732	γ γ γ γ β γ	56	KJ425486	γ γ γ γ α γ	106	KX364291	γ γ γ γ β γ
7	AY319651	γ γ γ γ β γ	57	KX236008	γ γ γ γ β γ	107	KP118893	γ γ γ γ β γ
8	AY646283	γ γ γ γ β γ	58	KP790143	γ γ γ γ β γ	108	KJ425510	γ γ γ γ β γ
9	LN610099	γ γ γ γ β γ	59	KX236015	γ γ γ γ β γ	109	KP118884	γ γ γ γ β γ
10	KX266757	γ γ γ γ β γ	60	KX077987	γ γ γ γ α γ	110	KP118885	γ γ γ γ β γ
11	MH779860	γ γ γ γ β γ	61	JX195178	γ γ γ γ α γ	111	KX236003	γ γ γ γ β γ
12	MH779856	γ γ γ γ β γ	62	JX195177	γ γ γ γ α γ	112	KX236002	γ γ γ γ α γ
13	MH779857	γ γ γ γ β γ	63	KX302874	γ γ γ γ α γ	113	KX236013	γ γ γ γ β γ
14	MH779858	γ γ γ γ β γ	64	KX640829	γ γ γ γ β γ	114	KX219798	γ γ γ γ α γ
15	MH779859	γ γ γ γ β γ	65	KP343691	γ γ γ γ β γ	115	KX372250	γ γ γ γ β γ
16	KU556805	γ γ γ γ β γ	66	KX400753	γ γ γ γ β γ	116	KX219799	γ γ γ γ β γ
17	KU556806	γ γ γ γ β γ	67	MK574042	γ γ γ γ β γ	117	KX219796	γ γ γ γ β γ
18	KU556807	γ γ γ γ β γ	68	MK574043	γ γ γ γ β γ	118	JQ088078	γ γ γ γ β γ
19	KU556804	γ γ γ γ β γ	69	KX219793	γ γ γ γ β γ	119	KY776701	γ γ γ γ β γ
20	KF663559	γ γ γ γ β γ	70	KX247127	γ γ γ γ α γ	120	KY407557	γ γ γ γ β γ
21	KP662631	γ γ γ γ β γ	71	KX236014	γ γ γ γ α γ	121	MH924835	γ γ γ γ α γ
22	KT736032	γ γ γ γ β γ	72	KP036503	γ γ γ γ β γ	122	KY776700	γ γ γ γ β γ
23	MN566147	γ γ γ γ β γ	73	KP118883	γ γ γ γ β γ	123	MK032180	γ γ γ γ β γ
24	MN599049	γ γ γ γ β γ	74	MG448607	γ γ γ γ β γ	124	MK581203	γ γ γ γ β γ
25	MK878536	γ γ γ γ β γ	75	KX247128	γ γ γ γ β γ	125	MK581206	γ γ γ γ β γ
26	MN512434	γ γ γ γ β γ	76	KP118890	γ γ γ γ β γ	126	MK581205	γ γ γ γ β γ
27	MN512435	γ γ γ γ β γ	77	KX252774	γ γ γ γ β γ	127	MK581200	γ γ γ γ α γ
28	MN512436	γ γ γ γ β γ	78	KP036504	γ γ γ γ β γ	128	MK581201	γ γ γ γ α γ
29	MN512437	γ γ γ γ β γ	79	KP118880	γ γ γ γ β γ	129	MK581202	γ γ γ γ β γ
30	MN512438	γ γ γ γ β γ	80	KX252776	γ γ γ γ β γ	130	MK581207	γ γ γ γ β γ
31	KY805846	γ γ γ γ α γ	81	KP118887	γ γ γ γ β γ	131	MK581208	γ γ γ γ β γ
32	MG233398	γ γ γ γ α γ	82	KX252773	γ γ γ γ β γ	132	GQ504722	γ γ γ γ β γ
33	KY273667	γ γ γ γ α γ	83	JF274479	γ γ γ γ β γ	133	GQ504723	γ γ γ γ β γ
34	KR608272	γ γ γ γ β γ	84	KP868572	γ γ γ γ β γ	134	FJ888351	γ γ γ γ α γ
35	KY588135	γ γ γ γ α γ	85	KP118891	γ γ γ γ β γ	135	MK217372	γ γ γ γ β γ
36	KT852992	γ γ γ γ β γ	86	KX434788	γ γ γ γ β γ	136	MK937829	γ γ γ γ α γ
37	KF931628	γ γ γ γ β γ	87	KX302870	γ γ γ γ β γ	137	MK217373	γ γ γ γ β γ
38	KF460437	γ γ γ γ β γ	88	KP790144	γ γ γ γ β γ	138	MK217374	γ γ γ γ β γ
39	GU393331	γ γ γ γ β γ	89	KP790146	γ γ γ γ β γ	139	MK217375	γ γ γ γ β γ
40	AY514485	γ γ γ γ β γ	90	KP036502	γ γ γ γ β γ	140	KX185056	γ γ γ γ β γ
41	GU393332	γ γ γ γ β γ	91	KP790145	γ γ γ γ β γ	141	KX185059	γ γ γ γ β γ
42	GU393334	γ γ γ γ α γ	92	KJ425508	γ γ γ γ α γ	142	KU900739	γ γ γ γ β γ
43	GU393338	γ γ γ γ β γ	93	KX302866	γ γ γ γ β γ	143	KU900740	γ γ γ γ β γ
44	KF377577	γ γ γ γ β γ	94	KJ425509	γ γ γ γ β γ	144	KU900743	γ γ γ γ β γ
45	GQ504720	γ γ γ γ β γ	95	KP036505	γ γ γ γ β γ	145	MN128086	γ γ γ γ β γ
46	GQ504721	γ γ γ γ β γ	96	KX364297	γ γ γ γ β γ	146	MN128088	γ γ γ γ β γ
47	EU418975	γ γ γ γ β γ	97	MK937833	γ γ γ γ α γ	147	MN128087	γ γ γ γ β γ
48	EU418976	γ γ γ γ β γ	98	KX434790	γ γ γ γ β γ	148	KC008600	γ γ γ γ α γ
49	FJ904714	γ γ γ γ β γ	99	KX302861	γ γ γ γ β γ	149	GQ427174	γ γ γ γ β γ
50	FJ904715	γ γ γ γ β γ	100	KX302865	γ γ γ γ β γ			

Таблица 7. Явление мозаичности в распознавании коронавирусных геномов рода *Alphacoronavirus* с помощью вариантного подхода (см. табл. 3а, табл. 3б)

GenBank ID		$v_1 v_2 v_3 v_4 v_5 v_6$	GenBank ID		$v_1 v_2 v_3 v_4 v_5 v_6$	GenBank ID		$v_1 v_2 v_3 v_4 v_5 v_6$
1	MH938448	α α α α α β	58	LT900499	α α α α α β	115	KM609213	α α α α α β
2	MH938450	α α α α α β	59	LT900500	α α α α α β	116	MH593900	α α α α α β
3	MH938449	α α α α α β	60	LT900501	α α α α α β	117	KJ196348	α α α α α β
4	MK472068	α α α α α β	61	LT900502	α α α α α β	118	KU297956	α α α α α β
5	MK472069	α α α α α β	62	KR003452	α α α α α β	119	KP403954	α α α α α β
6	MK472070	α α α α α β	63	JN825712	α α α α α β	120	KJ645682	α α α α α β
7	MK472071	α α α α α β	64	KX016034	α α α α α β	121	MK138516	α α α α α β
8	MH687935	α α α α α β	65	KR809885	α α α α α β	122	KX883635	α α α α α β
9	MH687936	α α α α α β	66	KT199103	α α α α α β	123	KR818832	α α α α α β
10	MH687939	α α α α α β	67	MF346935	α α α α α β	124	MK250953	α α α α α β
11	MH687940	α α α α α β	68	MK606369	α α α α α β	125	MK720945	α α α α α β
12	MH687949	α α α α α β	69	MG837012	α α α α α β	126	MK720946	α α α α α β
13	MH687950	α α α α α β	70	MN114121	α α α α α β	127	MN611517	α α α α α β
14	MH687951	α α α α α β	71	MK644601	α α α α α β	128	MN611521	α α α α α β
15	MH687959	α α α α α β	72	MH726408	α α α α α β	129	MK720944	α α α α α β
16	MH687962	α α α α α β	73	MH726374	α α α α α β	130	NC_028811	α α α α α β
17	KJ473799	α α α α α β	74	MH726393	α α α α α β	131	NC_028833	α α α α α β
18	KJ473800	α α α α α β	75	MH726368	α α α α α β	132	LT906582	α α β α α β
19	KJ473798	α α α α α β	76	MH726372	α α α α α β	133	KX981440	α α β α α β
20	KJ473806	α α α α α β	77	MH726369	α α α α α β	134	MF375374	α α β α α β
21	KJ473809	α α α α α β	78	MH726394	α α α α α β	135	KM887144	α α β α α β
22	MK211373	α α α α α β	79	MH726395	α α α α α β	136	LT905450	α α β α α β
23	KY799179	α α α α α β	80	MH726381	α α α α α β	137	LT905451	α α β α α β
24	KY073746	α α α α α β	81	MH726382	α α α α α β	138	LT906620	α α β α α β
25	KY073745	α α α α α β	82	MH726405	α α α α α β	139	MH726366	α α β α α β
26	LM645058	α α α α α β	83	MH726383	α α α α α β	140	MH726367	α α β α α β
27	LT898408	α α α α α β	84	MK690502	α α α α α β	141	MH726380	α α β α α β
28	LT898409	α α α α α β	85	KX534206	α α α α α β	142	KY070587	α α β α α β
29	LT898410	α α α α α β	86	MK482396	α α α α α β	143	MK841494	α α β α α β
30	LT898411	α α α α α β	87	MK482397	α α α α α β	144	LT897799	α α β α α β
31	LT898412	α α α α α β	88	LT898447	α α α α α β	145	KY007140	α α β α α β
32	LT898413	α α α α α β	89	KX812523	α α α α α β	146	GU937797	α α β α α β
33	LT898414	α α α α α β	90	KX812524	α α α α α β	147	GU937797	α α β α α β
34	LT898415	α α α α α β	91	KU558701	α α α α α β	148	MK702008	α α β α α β
35	LT898416	α α α α α β	92	LM645057	α α α α α β	149	MK409659	α α β α α β
36	LT898417	α α α α α β	93	MF807951	α α α α α β	150	MK409657	α α β α α β
37	LT898418	α α α α α β	94	KR153325	α α α α α β	151	MK409658	α α β α α β
38	LT898421	α α α α α β	95	KR153326	α α α α α β	152	KU664503	α α β α α β
39	LT898423	α α α α α β	96	KY793536	α α α α α β	153	KX839246	α α β α α β
40	LT898425	α α α α α β	97	KR095279	α α α α α β	154	KY486713	α α β α α β
41	LT898426	α α α α α β	98	KF760557	α α α α α β	155	KY486714	α α β α α β
42	LT898427	α α α α α β	99	MH061339	α α α α α β	156	KX839247	α α β α α β
43	LT898430	α α α α α β	100	MH061336	α α α α α β	157	KX839248	α α β α α β
44	LT898431	α α α α α β	101	MH061343	α α α α α β	158	KX839249	α α β α α β
45	LT898432	α α α α α β	102	MH061337	α α α α α β	159	KX839250	α α β α α β
46	LT898433	α α α α α β	103	MH061338	α α α α α β	160	KX839251	α α β α α β
47	LT898435	α α α α α β	104	KC196276	α α α α α β	161	KP890336	α α β α α β
48	LT898436	α α α α α β	105	KR011756	α α α α α β	162	MH061341	α α β α α β
49	LT898438	α α α α α β	106	MK138353	α α α α α β	163	MH061340	α α β α α β
50	LT898439	α α α α α β	107	JX112709	α α α α α β	164	AF353511	α α β α α β
51	LT898440	α α α α α β	108	MH708243	α α α α α β	165	EF185992	α α β α α β
52	LT898441	α α α α α β	109	MK862249	α α α α α β	166	MF782687	α α β α α β
53	LT898443	α α α α α β	110	KT941120	α α α α α β	167	KM609208	α α β α α β
54	LT898444	α α α α α β	111	MF782686	α α α α α β	168	KU252649	α α β α α β
55	LT898445	α α α α α β	112	KY111278	α α α α α β	169	KX550281	α α β α α β

GenBank ID		$v_1 v_2 v_3 v_4 v_5 v_6$	GenBank ID		$v_1 v_2 v_3 v_4 v_5 v_6$	GenBank ID		$v_1 v_2 v_3 v_4 v_5 v_6$
56	LT898446	α α α α α β	113	KM609211	α α α α α β	170	KU847996	α α β α α β
57	LT900498	α α α α α β	114	KM609212	α α α α α β			

Биологическую значимость наблюдаемого явления мозаичности можно будет оценить только при проведении дальнейших исследований. По-видимому, это явление определяется высокой генетической изменчивостью коронавирусов и их способностью к рекомбинации в процессе межпопуляционных взаимоотношений вирусов и их природных резервуаров.

ЗАКЛЮЧЕНИЕ

В работе предложены методы распознавания рода коронавируса на основе модифицированной статистики, ранее показавшей свою эффективность при распознавании видов флавивирусов. В основе этой статистики лежит использование распределений кодонов в ORF1ab, кодирующей неструктурные белки, и в генах структурных белков (S, M и N) коронавируса. Предлагаемые методы разработаны с помощью обучающей выборки геномов прототипных штаммов.

В отличие от однозначной трансляции единого полипротеина в геноме флавивируса, механизм трансляции неструктурных и структурных белков коронавирусов многостадийный, поэтому учитывались как длина генов, так и различные варианты их совместного анализа. Рассматривались шесть вариантов совместного анализа генов: V_1) S+N+M+ORF1ab, V_2) S+N+M, V_3) S+N, V_4) ORF1ab, V_5) S и V_6) N.

Использование предложенных статистик для распознавания рода коронавируса показало их высокую достоверность на уровне 95 %. Среди всех вариантов совместного анализа наибольшую надежность (98 %) показало использование распознавания рода коронавирусов на основе частот кодонов ORF1ab.

В работе был рассмотрен вариантный подход, в котором результат распознавания рода коронавирусов представлялся в виде строки из шести компонент. В этой строке каждая компонента представляла результат распознавания рода анализируемого генома коронавируса на основе одного из шести вариантов совместного анализа четырех генов коронавируса. При таком подходе однозначный результат распознавания признавался при наличии не менее четырех одинаковых компонент в вариантной строке. Однако надежность такого подхода оказалась немного ниже (97 %.) надежности только на основе гена ORF1ab. С другой стороны, вариантный подход выявил явление мозаичности при распознавании рода коронавируса. Так, например, среди распознанных 341 геномов (из всех 345) рода *Gammacoronavirus* в пятой компоненте вариантной строки, отвечающей за S-ген, «распознавался» род *Alphacoronavirus* (22) или *Betacoronavirus* (127). Также среди 905 распознанных геномов (из всех 924) рода *Alphacoronavirus* у 170 геномов был «распознан» род *Betacoronavirus* в шестой компоненте, соответствующей N-гену белка нуклеокапсида. Можно предположить, что явление мозаичности отражает пластичность генома коронавирусов и, своего рода, степень готовности к освоению не только новых ареалов существования, но и к поиску новых хозяев.

СПИСОК ЛИТЕРАТУРЫ

1. Schalk A.F., Hawn M.C. An apparently new respiratory disease of baby chicks. *J. Am. Vet. Med. Assoc*. 1931. V. 78. P. 413–423.
2. 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

3. Хайтович А.Б. Коронавирусы (таксономия, структура вируса). *Крымский журнал экспериментальной и клинической медицины*. 2020. Т. 10. № 3. С. 69–81. doi: 10.37279/2224-6444-2020-10-3-69-81
4. Львов Д.К., Альховский С.В. Истоки пандемии COVID-19: экология и генетика коронавирусов (*Betacoronavirus*: *Coronaviridae*) SARS-CoV, SARS-CoV-2 (подрод *Sarbecovirus*), MERS-CoV (подрод *Merbecovirus*). *Вопросы вирусологии*. 2020. Т. 65. № 2. С. 62–70. doi: 10.36233/0507-4088-2020-65-2-62-70
5. Львов Д.К., Альховский С.В., Колобухина Л.В., Бурцева Е.И. Этиология эпидемической вспышки COVID-19 в г. Ухань (провинция Хубэй, Китайская Народная Республика), ассоциированной с вирусом 2019-nCoV (*Nidovirales*, *Coronaviridae*, *Coronavirinae*, *Betacoronavirus*, подрод *Sarbecovirus*): уроки эпидемии SARS-CoV. *Вопросы вирусологии*. 2020. Т. 65. № 1. С. 6–15. doi: 10.36233/0507-4088-2020-65-1-6-15
6. *Virus Taxonomy. Classification and Nomenclature of Viruses. Ninth Report of the International Committee on Taxonomy of Viruses*. Eds. King A.M.Q., Adams M.J., Carstens E.B., Lefkowitz E.J. Elsevier Academic Press, 2011. 1338 p.
7. Bukhari K., Mulley G., Gulyaeva A.A., Zhao L., Shu G., Jiang J., Neuman B.W. Description and initial characterization of metatranscriptomic nidovirus-like genomes from the proposed new family *Abyssoviridae*, and from a sister group to the *Coronavirinae*, the proposed genus *Alphaletovirus*. *Virology*. 2018. V. 524. P. 160–171. doi: 10.1016/j.virol.2018.08.010
8. Ziebuhr J., Baric R.S., Baker S., de Groot R.J., Drosten C., Gulyaeva A., Haagmans B.L., Neuman B.W., Perlman S., Poon L.L.M., Sola I., Gorbalenya A.E. *Reorganization of the family Coronaviridae into two families, Coronaviridae (including the current subfamily Coronavirinae and the new subfamily Letovirinae) and the new family Tobaniviridae (accommodating the current subfamily Torovirinae and three other subfamilies), revision of the genus rank structure and introduction of a new subgenus rank*: Proposal 2017.013S (08.08.2018) for International Committee on Taxonomy of Viruses.
9. Neuman B.W., Adair B.D., Yoshioka C., Quispe J.D., Kuhn G.O.P., Milligan R.A., Yeager M., Buchmeier M.J. Supramolecular architecture of severe acute respiratory syndrome coronavirus revealed by electron cryomicroscopy. *J. Virol*. 2006. V. 80. No. 16. P. 7918–7928. doi: 10.1128/JVI.00645-06
10. *Virology: principles and applications*. Eds. J. Carter, V. Saunders. Chichester, England: John Wiley & Sons Ltd, 2007. 358 p. doi: 10.1093/tropej/fmn001
11. Chen Y., Liu Q., Guo D. Emerging coronaviruses: Genome structure, replication, and pathogenesis. *J. Med. Virol*. 2020. V. 92. P. 418–423. doi: 10.1002/jmv.25681
12. Ma Y., Wu L., Shaw N., Gao Y., Wang J., Sun Y., Lou Z., Yan L., Zhang R., Rao Z. Structural basis and functional analysis of the SARS coronavirus nspl4-nspl0 complex. *PNAS*. 2015. V. 112. No. 30. P. 9436–9441. doi: 10.1073/pnas.1508686112
13. Cavanagh D., Mawditt K., Adzharet A., Gough R.E., Picault J.P., Naylor C.J., Haydon D., Shaw K., Britton P. Does IBV change slowly despite the capacity of the spike protein to vary greatly? *Adv. Exp. Med. Biol*. 1998. V. 440. P. 729–734. doi: 10.1007/978-1-4615-5331-1_94
14. Pyrc K., Dijkman R., Deng L., Jebbink M.F., Ross H.A., Berkhout B., der Hoek L. Mosaic structure of human coronavirus NL63, one thousand years of evolution. *J. Mol. Biol*. 2006. V. 364. P. 964–973. doi: 10.1016/j.jmb.2006.09.074
15. Su S., Wong G., Shi W., Liu J., Lai A.C.K., Zhou J., Liu W., Bi Y., Gao G.F. Epidemiology, Genetic Recombination, and Pathogenesis of Coronaviruses. *Trends in Microbiology*. 2016. V. 24. No. 6. P. 490–502. doi: 10.1016/j.tim.2016.03.003

16. Edara V-V., Pinsky B.A., Suthar M.S., Lai L., Davis-Gardner M.E., Floyd K., Flowers M.W., Wrammert J., Hussaini L., Rose Ciric C. et al. Infection and vaccine-induced neutralizing-antibody responses to the SARS-CoV-2 B.1.617 Variants. *N. Engl. J. Med*. 2021. V. 385. No. 7. P. 664–666. doi: 10.1056/NEJMc2107799
17. Long S.W., Olsen R.J., Christensen P.A., Subedi S., Olson R., Davis J.J., Ojeda Saavedra M., Yerramilli P., Pruitt L., Reppond K. Sequence Analysis of 20,453 Severe Acute Respiratory Syndrome coronavirus 2 genomes from the Houston metropolitan area identifies the emergence and widespread distribution of multiple isolates of all major variants of concern. *Am. J. Pathol*. 2021. V. 191. No. 6. P. 983–992. doi: 10.1016/j.ajpath.2021.03.004
18. 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 SARSCoV-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
19. Чалей М.Б., Тюлько Ж.С., Кутыркин В.А. Распознавание видов флавивирусов на основе кодирующих последовательностей полипротеинов. *Математическая биология и биоинформатика*. 2019. Т. 14. № 2. С. 533–542. doi: 10.17537/2019.14.533
20. Чалей М.Б., Тюлько Ж.С., Кутыркин В.А. Исследование структуры кодирования ORF1ab, S, M и N генов коронавирусов. *Математическая биология и биоинформатика*. 2020. Т. 15. № 2. С. 441–454. doi: 10.17537/2020.15.441
21. Lai M.M.C. Recombination in large RNA viruses: Coronaviruses. *Seminars in Virology*. 1996. V. 7. No. 6. P. 381–388. doi: 10.1006/smvy.1996.0046
22. Tao Y., Shi M., Chommanard C., Queen K., Zhang J., Markotter W., Kuzmin I.V., Holmes E.C., Tong S. Surveillance of bat coronaviruses in Kenya identifies relatives of human coronaviruses NL63 and 229E and their recombination history. *Journal of Virology*. 2017. V. 91. No. 5. P. e01953–16. doi: 10.1128/JVI.01953-16
23. Luk H.K.H., Li X., Fung J., Lau S.K.P., Woo P.C.Y. Molecular epidemiology, evolution and phylogeny of SARS coronavirus. *Infection, Genetics and Evolution*. 2019. V. 71. P. 21–30. doi: 10.1016/j.meegid.2019.03.001
24. *GenBank*. URL: https://ftp.ncbi.nlm.nih.gov/genbank/ (accessed 20.12.2021).

Рукопись поступила в редакцию 28.01.2022, переработанный вариант поступил 09.03.2022.
Дата опубликования 15.03.2022.

Coronavirus Genus Recognition Based on Prototype Virus Variants

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

[1]*Institute of Mathematical Problems of Biology RAS, Pushchino, Russia*
[2]*Moscow State Technical University n.a. N.E. Bauman, Moscow, Russia*

***Abstract*.** Method named as variant approach to recognizing genus of coronavirus that is based on frequency of codon distribution in viral ORF1ab and genes of structural proteins (S, M and N) was proposed in the work. This method uses modified statistics whose efficiency was demonstrated earlier for flavivirus species recognition. To recognize genus of coronavirus the variant approach considers both various combinations of several structural coronavirus genes and individual structural genes. Finally, coronavirus genus is determined in the result of analysis of all variants considered. The method proposed was developed with the help of learning sample from prototype viral variants of *Alphacoronavirus*, *Betacoronavirus*, *Deltacoronavirus* and *Gammacoronavirus* genus. Application of the variant approach to recognizing genus of coronavirus has demonstrated the approach high assurance at level of 95 %. Among all variants of joint analysis, the most reliability (98 %) in recognizing genus has been achieved if codon frequency of the ORF1ab was used. Variant approach has revealed a phenomenon of mosaic structure in coronavirus genomes, i.e., when the results of genus recognition for a few genes differ from final conclusion about coronavirus genus. It seems that such phenomenon reflects homologous recombinations of the genes between various species of the coronaviruses and plasticity of their genomes in evolutionary processes.

Key words: *coronavirus genome, ORF1ab, S-gene, M-gene, N-gene, statistical analysis, variant approach to recognizing coronavirus genus.*