*Математическая биология и биоинформатика*
*2020. Т. 15. № 2. С. 441–454. doi: 10.17537/2020.15.441*

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

# Исследование структуры кодирования ORF1ab, S, M и N генов коронавирусов

**Чалей М.Б.**[*1], **Тюлько Ж.С.**[†2,3], **Кутыркин В.А.**[‡4]

[1]*Институт математических проблем биологии – филиал Института прикладной математики им. М.В. Келдыша РАН, Пущино, Московская область, Россия*
[2]*Омский государственный медицинский университет Минздрава России, Омск, Россия*
[3]*ФБУН «Омский НИИ природно-очаговых инфекций» Роспотребнадзора, Омск, Россия*
[4]*Московский государственный технический университет им. Н.Э. Баумана, Москва, Россия*

***Аннотация.*** В работе спектрально-статистический подход применен к сравнительному анализу геномов коронавирусов четырех родов *Alphacoronavirus*, *Betacoronavirus* (включая новый SARS-CoV-2 вирус), *Gammacoronavirus* и *Deltacoronavirus*, который выполнялся с точки зрения наличия 3-регулярности и скрытой триплетной профильной периодичности в кодирующих последовательностях четырех структурных генов: ORF1ab, кодирующего транскриптазу; S-гена гликопротеина, формирующего шипы; M-гена мембранного белка; N-гена нуклеопротеина. Общее число анализируемых геномов составило 3410. Соответственно, оно определяло и численность каждой выделенной группы генов. В результате, практически во всех CDS анализируемых генов ORF1ab, S и N была выявлена скрытая профильная триплетная периодичность и высокое значение индекса 3-регулярности, как показателя качества сохранности триплетной структуры кодирования. Для M-генов, напротив, была выявлена тенденция к размытию их структуры кодирования вплоть до однородности 60 % этих генов в анализируемых геномах альфакоронавирусов и 67 % в геномах гаммакоронавирусов. Тенденция размытия такой структуры, сопровождаемая снижением среднего значения индекса 3-регулярности в сравнении с остальными генами, при сохранении триплетной профильной периодичности, была отмечена и для M-генов SARS-CoV-2 вируса. Возможно, отмеченная тенденция отражает значение изменчивости M-генов при адаптации коронавируса к новым хозяевам рода. Анализ матриц 3-профильной периодичности для четырех, анализируемых в работе генов вируса SARS-CoV-2, выделенного в Европе, Азии и США, не выявил их значимого различия, что предполагает единый источник распространения этого вируса.

***Ключевые слова:*** *свойство 3-регулярности CDS, скрытая профильная триплетная периодичность, геном коронавируса, геном SARS-Cov-2 вируса, ORF1ab, S-ген, M-ген, N-ген.*

## ВВЕДЕНИЕ

30 января 2020 года Всемирная организация здравоохранения объявила текущую вспышку нового коронавируса, который был впервые обнаружен в китайском городе Ухань в декабре 2019 года (обозначаемого сейчас, как SARS-CoV-2, где SARS – Severe

[*] maramaria@yandex.ru
[†] tjs@omsk-osma.ru
[‡] vkutyrkin@yandex.ru

Acute Respiratory Syndrome Coronavirus), «чрезвычайной ситуацией в области общественного здравоохранения, имеющей международное значение» [1].

Коронавирусы включены в отряд *Nidovirales*, семейство *Coronaviridae*, подсемейство *Coronavirinae*. К *Nidovirales* относят оболочечные вирусы с инфекционной односегментной линейной одноцепочечной РНК позитивной полярности, которые имеют общий план строения генома и схемы репликации, их геномы являются самыми крупными РНК-геномами среди вирусов. Кроме *Coronaviridae* в отряд *Nidovirales* входят еще два семейства: *Arteriviridae* и *Roniviridae*. Первое объединяет вирусы млекопитающих, но не человека (включая вирус артериита лошадей и вирус репродуктивно-респираторного синдрома свиней), второе – исключительно вирусы членистоногих (ракообразные, насекомые) [2]. Подсемейство *Coronavirinae* разделяют на роды *Alphacoronavirus, Betacoronavirus, Gammacoronavirus, Deltacoronavirus*. Например, в род *Betacoronavirus* входят виды вирусов, выделенных от рукокрылых, парнокопытных и, в частности, новый коронавирус человека SARS-CoV-2. Коронавирусы распространены среди широкого круга позвоночных хозяев [2] и способны к сравнительно быстрой адаптации к новому хозяину. Механизмы такой адаптации, по-видимому, включают в себя рекомбинации и накопление точечных замен в геноме [3–6]. Ранее уже были исследованы случаи, когда коронавирусы из природных резервуаров адаптировались к новым хозяевам при попадании в их организм от промежуточного хозяина. В качестве примера можно привести еще два бетакоронавируса человека: считается, что природным резервуаром SARS-CoV служат летучие мыши, а промежуточными хозяевами – циветы [7]. В другом случае природным резервуаром MERS (Middle East Respiratory Syndrome) были летучие мыши, а промежуточными хозяевами – верблюды [8]. По некоторым данным, предполагается также возможность прямого инфицирования человека коронавирусами от летучих мышей в Китае [9], что делает изучение закономерностей изменения их вирусных геномов необходимым ввиду возможности возникновения новых вариантов, опасных для человека и животных. В настоящее время выявлены семь человеческих коронавирусов. Причиной легких заболеваний являются вирусы 229E, OC43, NL63 и HKU1, а тяжелые патологии вызываются видами SARS-CoV, MERS-CoV и SARS-CoV-2. Геном SARS-CoV-2 оказался гомологичен MERS-CoV на 50 %, SARS-CoV – на 79 %. Вирус SARS-CoV-2 рассматривается как природно-очаговый, с происхождением из популяции рукокрылых летучих мышей, поскольку для его генома с геномом штамма RaTG13 коронавируса азиатского подковоноса (*Rhinolophus affinis*) установлена гомология 96 % [10]. Промежуточными хозяевами могут являться мелкие млекопитающие.

Геном коронавирусов содержит несколько открытых рамок считывания (ORF), кодирующих как структурные, так и неструктурные белки [10–13]. При входе в клетку геномная РНК транслируется с образованием неструктурных белков из двух открытых рамок считывания ORF1a и ORF1b. При считывании последовательности ORF1a продуцируется полипептид pp1a (440–500 кДа), который расщепляется на 11 неструктурных белков. В некоторых случаях рибосомы при трансляции игнорируют стоп-кодон ORF1a из-за шпильки, смещающей рамку считывания на -1 нуклеотид непосредственно перед стоп-кодоном, что позволяет провести трансляцию ORF1b, производя полипептид pp1ab (740–810 кДа), который расщепляется на 15 неструктурных белков. В таких случаях ORF полипептида pp1ab обозначают как ORF1ab.

В дополнение к геномной РНК с вирусной последовательности РНК продуцируются основные субгеномные РНК, кодирующие структурные белки (белок шипа – S, белок оболочки – E, мембранный белок – M и нуклеокапсидный белок – N), и также несколько дополнительных белков. Четыре основных структурных белка выполняют

свои роли при сборке вирусных частиц и влияют на инфекционность коронавируса. Интегральный мембранный белок M является наиболее распространенным структурным белком и определяет форму вирусной оболочки. Белок M также изменяет участок мембраны клетки для сборки вируса и захватывает другие структурные белки, обеспечивая их взаимодействие в месте сборки вирусной частицы. При делеции гена M-белка вирусные частицы не формируются. Белок нуклеокапсида N защищает вирусную РНК, формируя нуклеокапсид, и увеличивает продукцию вирусных частиц. Шипы, сформированные тремя S-белками на поверхности коронавируса, способствует связыванию его с рецепторами и слиянию между мембранами вируса и клетки-хозяина, чтобы облегчить проникновение вируса в клетку. Исследования показали, что шипы необязательны для сборки вируса, но необходимы для инфицирования клетки. Мембранный белок E вместе с M-белком включен в вирусную оболочку, и их взаимодействие необходимо для производства и высвобождения вирусных частиц. Белок E также способен формировать ионные каналы и их наличие может влиять на вирулентность коронавирусов [13].

В настоящей работе для исследования геномов коронавирусов самых различных видов, с точки зрения наличия в них скрытой триплетной профильной периодичности и 3-регулярности, рассматривается структура кодирования в нуклеотидных последовательностях (CDS) четырех генов, занимающих порядка 80 % всей длины вирусного генома: ORF1ab-гена, S-гена, M-гена и N-гена.

Ранее в работах [14, 15] был предложен особый тип скрытой периодичности в последовательностях ДНК и РНК, который получил название профильной периодичности. Также было показано, что для кодирующих районов характерны периоды кратные или равные трем. В частности, для генов полипротеинов исследованных 15 видов флавивирусов распознавалась скрытая триплетная профильная периодичность [16]. Вместе со скрытой триплетной периодичностью всегда наблюдается свойство 3-регулярности спектров, описывающих структурно-статистический свойства анализируемых последовательностей. Следует отметить, что свойство 3-регулярности присуще кодирующим районам даже в том случае, когда скрытая триплетная периодичность в них не распознаётся. Грубо говоря, можно было бы назвать свойство 3-регулярности слабой триплетной периодичностью. В качестве интегральной оценки паттерна скрытой триплетной профильной периодичности можно рассматривать матрицу частот встречаемости нуклеотидов в позициях триплетов. Количественная оценка степени выраженности свойства 3-регулярности даётся в зависимости от значений индекса 3-регулярности от 0.7 до 1.0, которые присущи именно кодирующим районам [15, 16].

В настоящей работе на основе анализа скрытой триплетной профильной периодичности и регулярности рассматривались четыре структурных гена из геномов известных родов коронавирусов (*Alphacoronavirus, Betacoronavirus, Deltacoronavirus, Gammacoronavirus*), кодирующих РНК-полимеразу (ORF1ab), спайк-белки (S), составляющие коронавирусные шипы, а также ген, кодирующий мембранный белок (M) и ген нуклеопротеина (N) – белка образующего нуклеопротеиновый комплекс с РНК-геномом внутри вирусного вириона. Отдельно такие гены рассматривались для SARS-CoV-2 бетакоронавируса, вызвавшего пандемию COVID-19. Гены для анализа SARS-CoV-2 были отобраны из геномов коронавирусов в соответствии с их координатами, аннотированными в базе данных GenBank на специальной странице, посвященной COVID-19 [17]. Геномы SARS-CoV-2 были взяты в равных количествах по 69 геномов, выделенных из изолятов в Европе, Азии и Соединенных штатах Америки. Кроме того, для анализа были отобраны геномы коронавирусов из базы данных вирусов в GenBank в следующих количествах: для родов *Alphacoronavirus* – 909, *Betacoronavirus* – 1940, *Deltacoronavirus* – 10 и *Gammacoronavirus* – 344.

Таким образом, исследование структурных свойств генов нового SARS-CoV-2 вируса позволило показать его отличие и сходство как с коронавирусами других родов, так и с родом *Betacoronavirus*, к которому он относится.

## МАТЕРИАЛЫ И МЕТОДЫ

### Геномы и гены коронавирусов

Геномы коронавирусов четырех родов, анализируемых в работе, в основном были получены из базы данных нуклеотидных последовательностей GenBank выпуска 237 от 15 апреля 2020 г. [18]. К этим геномам была добавлена выборка полностью секвенированных геномов коронавируса SARS-CoV-2 (рода *Betacoronavirus*), вызывающих у человека тяжелое вирусное заболевание ковид-19. Выборка была сделана из доступных нуклеотидных данных [17] для трех географических ареалов: Азии, Европы и США, по 69 геномов для каждого ареала. Отбор геномов производился таким образом, чтобы представлять данные различных стран. Так, например, в азиатской выборке находились геномы изолятов вируса из разных провинций Китая, из Гонконга, Непала, Индии, Пакистана, Южной Кореи, Ирана, Тайваня и др.

В результате были отобраны вирусные геномы (в количестве 3410) полностью секвенированные, не содержащие гепов и хорошо аннотированные, с тем, чтобы в каждом геноме можно было выделить кодирующие последовательности (CDS) для всех четырех генов: ORF1ab, S, M и N. При выделении CDS для генов ORF1ab учитывался сдвиг рамки считывания при переходе от ORF1a к ORF1b, о чем уже упоминалось выше, во введении. Средняя длина CDS для ORF1ab гена составляла порядка 21000 нуклеотидов (нукл.), CDS для S гена – 4000 нукл., для CDS M- и N-генов – 680 нукл. и 1200 нукл., соответственно.

### Спектрально-статистический подход

Исходно введение понятия профильной периодичности основывалось на теоретических вероятностных моделях случайных строк, составленных из независимых случайных букв в текстовом алфавите ДНК [14, 15]. Такие случайные строки получили название профильных строк. В дальнейшем [19], для распознавания скрытой профильной периодичности в нуклеотидных последовательностях ДНК и РНК были предложены более сложные модели случайных строк. Однако, для наглядности, в настоящей работе демонстрация структурно-статистических свойств случайных строк может быть ограничена профильными строками. В спектрально-статистическом подходе, используемом для распознавания скрытой периодичности и регулярности, нуклеотидные последовательности геномов рассматриваются как реализации соответствующих случайных строк.

Исследование структурно-статистических свойств профильной строки основано на графическом анализе специальных функциональных зависимостей от длины тестируемых периодов (тест-периодов). Графики этих зависимостей получили названия соответствующих им спектров. Первый спектр $\Psi_1$, названный главным, показывает проявление однородности (неоднородности) в профильной строке. Если строка является однородной, то главный спектр $\Psi_1$ является нулевым, как показано на рис. 1,а. В противном случае неоднородности профильной строки этот спектр отличен от нуля. Кроме того, в случае неоднородности, спектр $\Psi_1$ также может показывать наличие свойства 3-регулярности в профильной строке. Это свойство характеризуется периодичностью спектра $\Psi_1$ с наличием максимумов на тест-периодах, кратных трем (см. рис. 1,б). Наличие триплетной профильной периодичности в профильной строке подтверждается с помощью спектра проявления 3-профильности $\Psi_3$. При наличии в



Математическая биология и биоинформатика. 2020. Т. 15. № 2. doi: 10.17537/2020.15.441

профильной строке триплетной профильности этот спектр является нулевым, что показывает рисунок 1,в. В противном случае, когда спектр $\Psi_3$ отличен от нулевого, в профильной строке отсутствует триплетная профильность (профильная периодичность периода 3). Кроме того, при наличии в строке триплетной профильной периодичности, в её главном спектре $\Psi_1$ проявляется свойство 3-регулярности.

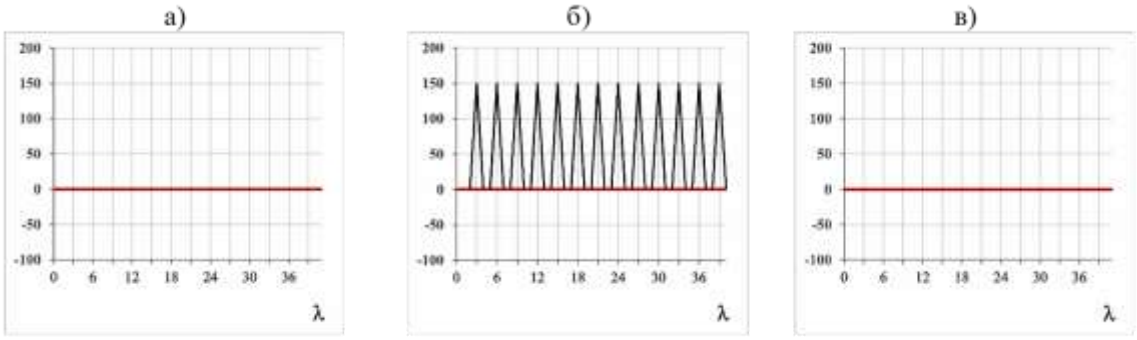


**Рис. 1. а**) Схема $\Psi_1$ спектра проявления однородности профильной строки. **б**) Вид $\Psi_1$ спектра при проявлении неоднородности и наличия свойства 3-регулярности в профильной строке. **в**) Вид $\Psi_3$ спектра при наличии триплетной периодичности в профильной строке.

Согласно спектрально-статистическому подходу, нуклеотидные последовательности рассматриваются как реализации соответствующих профильных строк. Спектры, характеризующие структурно-статистические свойства нуклеотидных последовательностей, рассматриваются как выборочные спектры. Поэтому главный выборочный спектр нуклеотидной последовательности назван спектром отклонения от однородности и для него используется обозначение $\Psi_1^*$. Для нуклеотидной последовательности аналогом спектра $\Psi_3$ является спектр отклонения от 3-профильности, для которого используется обозначение $\Psi_3^*$. Для выборочных спектров гипотеза об их отклонении от нуля принимается на уровне значимости 5 %. Например, для нуклеотидной последовательности гипотеза об её однородности принимается в том случае, когда доля тест-периодов, на которых достигается значение спектра $\Psi_1^* > 0$, не превосходит критического значения, выбираемого из одностороннего доверительного интервала, индуцированного уровнем значимости 5 % и количеством тест-периодов. Аналогично принимается гипотеза о наличии скрытой 3-профильности (скрытой триплетной профильной периодичности) на основе выборочного спектра $\Psi_3^*$.

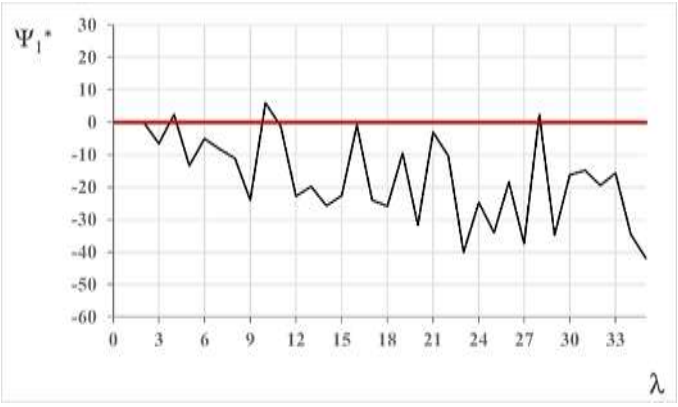


**Рис. 2.** Спектр отклонения от однородности для M-гена из изолята альфа коронавируса верблюда (Camel alphacoronavirus isolate camel/Jeddah/N68(a)/2014, GenBank код доступа KT368901).

На рисунке 2 показан спектр нуклеотидной последовательности гена мембранного белка альфа коронавируса, которая, согласно указанному выше правилу, является однородной.

На основе исследования кодирующих районов целого ряда организмов, в работе [15] для выявления наличия в спектре $\Psi_1^*$ нуклеотидной последовательности свойства 3-регулярности было предложено пороговое значение 0.7 для индекса 3-регулярности $I_3$. Если для спектра $\Psi_1^*$ значение $I_3 \geq 0.7$, то в этом спектре признается наличие свойства 3-регулярности.

В спектре отклонения от однородности, показанном на рисунке 2, свойство 3-регулярности отсутствует. На рисунке 3 спектр $\Psi_1^*$ последовательности, признаваемой неоднородной, обладает свойством 3-регулярности ($I_3 = 0.99$).

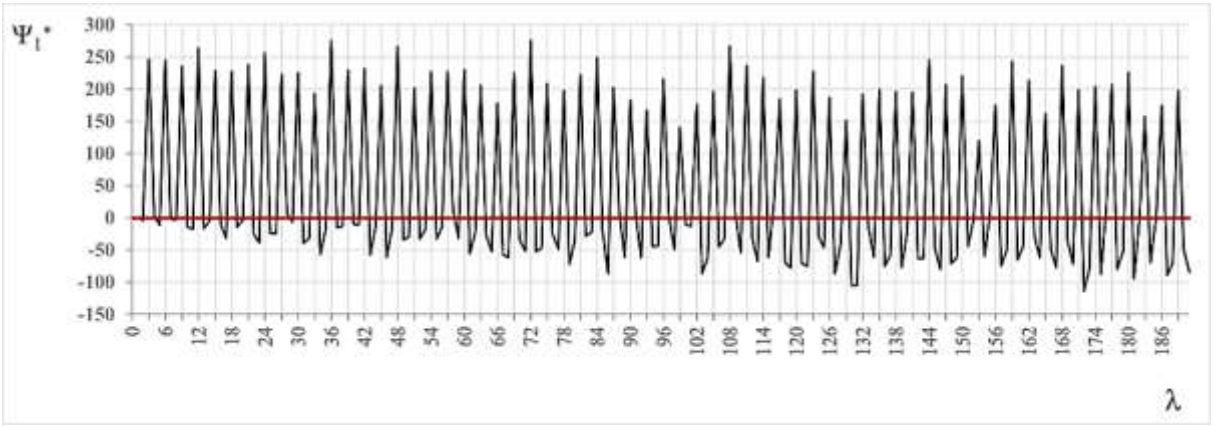


**Рис. 3.** Спектр отклонения от однородности для S-гена из европейского изолята SARS-CoV-2 вируса человека (GenBank код доступа MT534285).

На рисунке 4,а показан 3-регулярный спектр $\Psi_1^*$, со значением индекса $I_3 = 1.0$, для нуклеотидной последовательности N-гена изолята SARS-CoV-2 вируса человека, выявляющий неоднородность этого гена. Согласно выборочному спектру $\Psi_3^*$, показанному на рисунке 4,б, в этой последовательности распознается скрытая 3-профильная периодичность. Аналогичный анализ, проведенный для последовательности S-гена (см. рис. 3), также распознаёт скрытую триплетную периодичность.

Поскольку модель триплетной профильной периодичности является вероятностной (стохастической), оценка паттерна такой периодичности в нуклеотидной последовательности представляется в виде стохастической 3-профильной матрицы размера $4 \times 3$. В первом столбце этой матрицы стоит вектор частот букв алфавита ДНК в первой позиции паттерна. То есть, в первой позиции стохастического паттерна стоит случайная буква. Аналогично, во второй и в третьей позиции этого паттерна стоят соответствующие случайные буквы с векторами частот во втором и третьем столбце стохастической 3-профильной матрицы. В качестве примера таких 3-профильных матриц паттернов скрытой триплетной периодичности в таблице 1 приведены усредненные матрицы триплетной профильной периодичности для четырех структурных генов SARS-CoV-2, полученные в результате анализа геномов изолятов этого вируса в Азии, Европе и США. Выборки для каждого географического ареала насчитывали по 69 геномов и отбирались из базы GenBank на специальной странице, посвященной COVID-19 [17], таким образом, чтобы охватить представителей различных стран и штатов США.

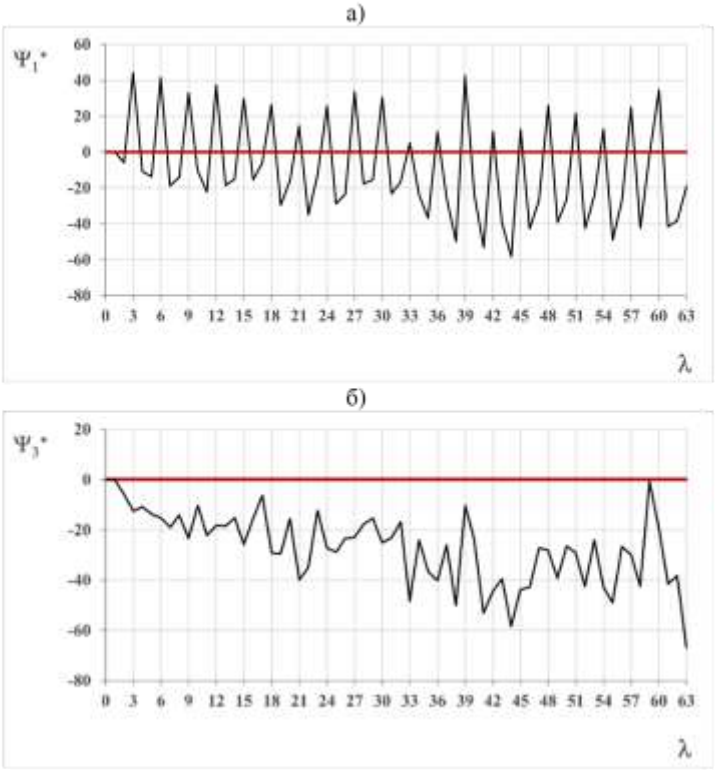


**Рис. 4. а**) Спектр отклонения от однородности и **б**) спектр отклонения от 3-профильности для N-гена из европейского изолята SARS-CoV-2 вируса человека (код доступа GenBank MT534285).

**Таблица 1.** Усредненные матрицы скрытой 3-профильной периодичности для анализируемых ORF1ab-, S-, M- и N- генов в выборках геномов из изолятов SARS-CoV-2 вируса в Азии, Европе и США

| | **SARS-CoV-2 ASIA** | | | **SARS-CoV-2 EU** | | | **SARS-CoV-2 USA** | | |
|---|---|---|---|---|---|---|---|---|---|
| ORF1ab | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 |
| A | 0.30 | 0.32 | 0.28 | 0.30 | 0.32 | 0.28 | 0.30 | 0.32 | 0.28 |
| T | 0.23 | 0.30 | 0.44 | 0.23 | 0.30 | 0.44 | 0.23 | 0.30 | 0.44 |
| G | 0.31 | 0.16 | 0.13 | 0.31 | 0.16 | 0.13 | 0.31 | 0.16 | 0.13 |
| C | 0.16 | 0.23 | 0.15 | 0.16 | 0.23 | 0.15 | 0.16 | 0.23 | 0.15 |
| S | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 |
| A | 0.30 | 0.31 | 0.27 | 0.30 | 0.31 | 0.27 | 0.30 | 0.31 | 0.27 |
| T | 0.24 | 0.29 | 0.46 | 0.24 | 0.29 | 0.46 | 0.24 | 0.29 | 0.46 |
| G | 0.29 | 0.16 | 0.11 | 0.29 | 0.16 | 0.11 | 0.29 | 0.16 | 0.11 |
| C | 0.16 | 0.24 | 0.16 | 0.16 | 0.24 | 0.16 | 0.16 | 0.24 | 0.16 |
| M | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 |
| A | 0.30 | 0.22 | 0.24 | 0.30 | 0.22 | 0.24 | 0.30 | 0.22 | 0.24 |
| T | 0.22 | 0.37 | 0.37 | 0.22 | 0.37 | 0.37 | 0.22 | 0.37 | 0.37 |
| G | 0.26 | 0.20 | 0.16 | 0.26 | 0.20 | 0.16 | 0.26 | 0.20 | 0.16 |
| C | 0.22 | 0.21 | 0.23 | 0.22 | 0.21 | 0.23 | 0.22 | 0.21 | 0.23 |
| N | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 |
| A | 0.31 | 0.33 | 0.30 | 0.31 | 0.33 | 0.30 | 0.31 | 0.33 | 0.30 |
| T | 0.15 | 0.16 | 0.32 | 0.15 | 0.16 | 0.32 | 0.15 | 0.16 | 0.32 |
| G | 0.30 | 0.22 | 0.15 | 0.30 | 0.22 | 0.15 | 0.30 | 0.22 | 0.15 |
| C | 0.24 | 0.28 | 0.23 | 0.24 | 0.28 | 0.23 | 0.24 | 0.28 | 0.23 |

Как видно из таблицы 1, 3-профильные матрицы генов соответствующих белков практически совпадают для всех трех географических ареалов. Эту же особенность можно отметить и между 3-профильной матрицей для ORF1ab – открытой рамки считывания, кодирующей транскриптазу, и соответствующей матрицей гена спайк-гликопротеина (S-белка).

## РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

В настоящей работе были проанализированы кодирующие последовательности четырех генов (ORF1ab, S, M и N) из более чем трех тысяч геномов коронавирусов, относящихся к четырём основным родам (альфа; бета, включая SARS-CoV-2; гамма и дельта). Предметом анализа являлось получение количественных показателей наличия свойства 3-регулярности и скрытой триплетной периодичности в нуклеотидных последовательностях этих генов. Полученные результаты представлены в таблице 2. Как можно видеть из таблицы 2, для родов бета и дельта практически во всех анализируемых генах распознавалась скрытая 3-профильная периодичность, сопровождаемая свойством 3-регулярности. Такое же явление наблюдается и для родов альфа и гамма в генах ORF1ab, S и N. Резкий контраст с этим явлением составляют гены мембранных белков (M-гены) альфа- и гаммакоронавирусов. Так среди генов M-белков альфакоронавирусов 3-профильность (3-регулярность) распознавалось только для 40 % этих генов. Еще меньший процент (23 %) генов со скрытой триплетной профильностью выявляется для M-генов гаммакоронавирусов.

**Таблица 2.** Анализ сохранности структуры триплетного кодирования для неоднородных генов ORF1ab, S, M и N в анализируемых геномах коронавирусов

| **Род коронавируса** | | **Количество 3-регулярных генов / среднее значение индекса 3-регулярности** | | | | **Количество генов с 3-профильностью** | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | **ORF1ab** | **S** | **M** | **N** | **ORF1ab** | **S** | **M** | **N** |
| Alphacoronavirus | | 909/ 1.0 | 909/ 0.99 | 367/ 0.87 | 908/ 0.99 | 908 | 909 | 365 | 906 |
| Betacoronavirus (кроме SARS-CoV-2) | | 1940/ 1.0 | 1912/ 0.99 | 1818/ 0.85 | 1940/ 0.99 | 1924 | 1925 | 1809 | 1938 |
| SARS-CoV-2 Betacoronavirus | ASIA | 69/ 1.0 | 69/ 0.99 | 69/ 0.76 | 69/ 1.00 | 69 | 69 | 69 | 69 |
| | EU | 69/ 1.0 | 69/ 0.99 | 69/ 0.77 | 69/ 1.00 | 69 | 69 | 69 | 69 |
| | USA | 69/ 1.0 | 69/ 0.99 | 69/ 0.76 | 69/ 1.00 | 69 | 69 | 69 | 69 |
| Deltacoronavirus | | 10/ 1.0 | 10/ 0.99 | 9/ 0.94 | 10/ 0.99 | 10 | 10 | 9 | 10 |
| Gammacoronavirus | | 344/ 1.0 | 344/ 1.0 | 80/ 0.82 | 344/ 0.98 | 344 | 344 | 80 | 344 |

Отдельный анализ оставшихся M-генов из этих родов показал, что в них наблюдается свойство структурной однородности, как показано в таблице 3. Кроме

того, несмотря на то, что М-гены SARS-CoV-2 вируса распознаются как 3-регулярные и обладающие на сегодняшний момент свойством скрытой триплетной периодичности, для них характерно более низкое среднее значение (0.76) индекса 3-регулярности, чем для соответствующих генов коронавирусов альфа, бета (без SARS-CoV-2) и гамма родов. Следовательно, триплетная структура кодирования в этих генах также показывает тенденцию к размытию.

**Таблица 3.** Количество однородных структурных генов ORF1ab, S, M и N в выборке анализируемых геномов коронавирусов. Каждый анализируемый геном содержит все четыре рассматриваемых гена

| **Род коронавируса** | | **Количество анализируемых геномов** | **Количество выявленных однородных генов** | | | |
|---|---|---|---|---|---|---|
| | | | **ORF1ab** | **S** | **M** | **N** |
| Alphacoronavirus | | 909 | 0 | 0 | 543 | 1 |
| Betacoronavirus (кроме SARS-CoV-2) | | 1940 | 0 | 0 | 11 | 0 |
| SARS-CoV-2* Betacoronavirus | ASIA | 69 | 0 | 0 | 0 | 0 |
| | EU | 69 | 0 | 0 | 0 | 0 |
| | USA | 69 | 0 | 0 | 0 | 0 |
| Deltacoronavirus | | 10 | 0 | 0 | 1 | 0 |
| Gammacoronavirus | | 344 | 0 | 0 | 231 | 0 |

*Отдельно рассматриваются геномы SARS-CoV-2 вируса, выделенного у пациентов с ковид-19 в Азии, Европе и Соединенных Штатах Америки.

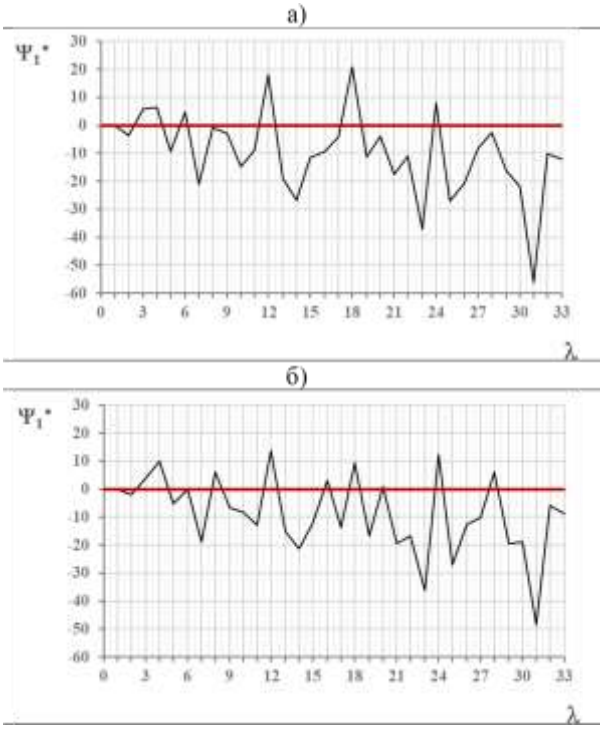


**Рис. 5.** Спектры отклонения от однородности для последовательностей М-генов из синонимичных геномов коронавирусов. **а**) Геном SARS-CoV-2 вируса, выделенного у пациента в г. Ухань в декабре 2019 г. (код доступа GenBank MN908947). Индекс 3-регулярности спектра $I_3 = 0.76$. **б**) Геном штамма RaTG13 коронавируса азиатского подковоноса (*Rhinolophus affinis*) (код доступа GenBank MN996532). Индекс 3-регулярности спектра $I_3 = 0.70$.

В качестве примера такого анализа триплетной структуры кодирования для М-гена приведем спектры отклонения от однородности на рисунке 5 и спектры отклонения от 3-профильности (скрытой триплетной профильной периодичности) на рисунке 6.

Графики на рисунке 5,а и рисунке 6,а получены для M-гена вирусного генома SARS-CoV-2, изолированного у первого пациента, погибшего от атипичной пневмонии в г. Ухань (штамм Wuhan-Hu-1) [20]. Графики на рисунке 5,б и рисунке 6,б получены для M-гена штамма RaTG13 вирусного генома SARS-CoV-2, изолированного у азиатского подковоноса из отряда рукокрылых, популяции которых предполагаются в качестве природного очага этого вируса [10].

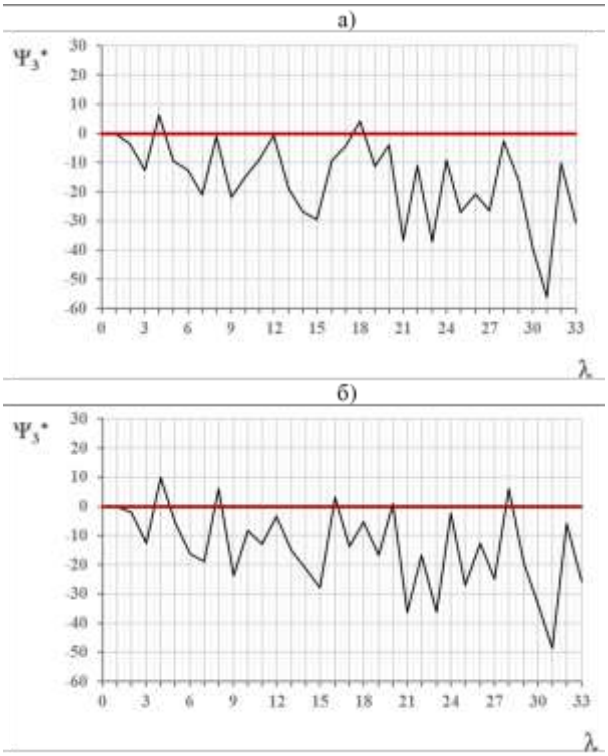


**Рис. 6.** Спектры отклонения от 3-профильности для последовательностей M-генов из геномов коронавирусов. **а)** Геном SARS-CoV-2, тот же что и на рис. 5,а (код доступа GenBank MN908947). **б)** Геном штамма RaTG13, тот же что и на рис. 5,б (код доступа GenBank MN996532).

Как можно видеть из рисунка 5,а и рисунка 5,б, последовательности M-генов обоих штаммов являются неоднородными и 3-регулярными с характерным именно для этих генов SARS-CoV-2 значением индекса 3-регулярности $I_3 < 0.8$. График спектра отклонения от 3-профильности на рисунке 6,а показывает, что для M-гена штамма Wuhan-Hu-1 выявляется еще и свойство скрытой 3-профильной периодичности, которое не распознается для M-гена штамма RaTG13 (рис. 6,б).

Из полученных результатов проведенного анализа можно предположить, что, несмотря на представление о консервативности гена мембранного белка коронавируса [21], M-ген мутирует достаточно активно, так как свойства мембранного белка для перехода вируса к новому хозяину также весьма существенны и отбираются эволюционным путем. Частицы разных видов коронавирусов значительно различаются по размеру и являются плеоморфными. M-белок способствует сборке оболочки вируса, взаимодействуя с вирусным рибонуклеопротеином и S-гликопротеинами и такие взаимодействия, а также взаимодействия M белков друг с другом определяют как форму вируса, так и частоту расположения шипов, формируемых на поверхности вируса S-белками. Белок M в разных конформациях способен менять кривизну мембранной оболочки вируса и его размер [22]. Кроме того известно, что белок M многих коронавирусов и, в частности, SARS-CoV-2 является ингибитором пути передачи сигналов интерферона I типа, ключевого компонента противовирусного

ответа врожденного иммунитета хозяина [23] и, следовательно, может подвергаться интенсивному отбору при смене хозяина.

### Оценка стохастического паттерна скрытой триплетной периодичности

Рассмотрим таблицу 4, содержащую 3-профильные матрицы для генов ORF1ab, S и N трех родов коронавирусов, включая соответствующие гены SARS-CoV-2, который относится к роду бета. Эти матрицы служат для оценки стохастического паттерна скрытой триплетной периодичности в последовательностях генов. Из таблицы 4 следует, что отличие этих матриц между родами прослеживается в основном в третьей позиции паттерна. В частности, альфакоронавирус по всем генам в этой позиции отличается частотой встречаемости нуклеотида A. Гаммакоронавирус аналогичным образом отличается от других родов по нуклеотиду C. Следовательно, в разделении коронавирусов по родам, существенную роль играют синонимичные замены кодонов в генах.

**Таблица 4.** Усредненные матрицы скрытой 3-профильной периодичности для анализируемых ORF1ab, S и N генов из геномов трех родов коронавирусов*

| | ***Alphacoronavirus*** | | | | ***Betacoronavirus*** **(кроме SARS-CoV-2)** | | | | **SARS-CoV-2 (ASIA, EU, USA)** | | | | ***Gammacoronavirus*** | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ORF1ab | | 1 | 2 | 3 | | 1 | 2 | 3 | | 1 | 2 | 3 | | 1 | 2 | 3 |
| | A | 0.28 | 0.30 | 0.18 | A | 0.29 | 0.32 | 0.25 | A | 0.30 | 0.32 | 0.28 | A | 0.28 | 0.32 | 0.25 |
| | T | 0.24 | 0.31 | 0.47 | T | 0.24 | 0.31 | 0.45 | T | 0.23 | 0.30 | 0.44 | T | 0.23 | 0.31 | 0.46 |
| | G | 0.33 | 0.17 | 0.17 | G | 0.32 | 0.16 | 0.15 | G | 0.31 | 0.16 | 0.13 | G | 0.33 | 0.16 | 0.16 |
| | C | 0.15 | 0.21 | 0.17 | C | 0.16 | 0.22 | 0.16 | C | 0.16 | 0.23 | 0.15 | C | 0.15 | 0.20 | 0.13 |
| S | | 1 | 2 | 3 | | 1 | 2 | 3 | | 1 | 2 | 3 | | 1 | 2 | 3 |
| | A | 0.29 | 0.29 | 0.18 | A | 0.29 | 0.31 | 0.23 | A | 0.30 | 0.31 | 0.27 | A | 0.32 | 0.30 | 0.24 |
| | T | 0.24 | 0.31 | 0.50 | T | 0.25 | 0.29 | 0.48 | T | 0.24 | 0.29 | 0.46 | T | 0.25 | 0.29 | 0.52 |
| | G | 0.30 | 0.17 | 0.15 | G | 0.29 | 0.16 | 0.12 | G | 0.29 | 0.16 | 0.11 | G | 0.29 | 0.18 | 0.13 |
| | C | 0.17 | 0.24 | 0.18 | C | 0.17 | 0.24 | 0.16 | C | 0.16 | 0.24 | 0.16 | C | 0.15 | 0.23 | 0.11 |
| N | | 1 | 2 | 3 | | 1 | 2 | 3 | | 1 | 2 | 3 | | 1 | 2 | 3 |
| | A | 0.32 | 0.37 | 0.23 | A | 0.31 | 0.33 | 0.28 | A | 0.31 | 0.33 | 0.30 | A | 0.29 | 0.34 | 0.30 |
| | T | 0.15 | 0.19 | 0.37 | T | 0.15 | 0.17 | 0.34 | T | 0.15 | 0.16 | 0.32 | T | 0.15 | 0.17 | 0.38 |
| | G | 0.32 | 0.20 | 0.19 | G | 0.30 | 0.21 | 0.16 | G | 0.30 | 0.22 | 0.15 | G | 0.35 | 0.23 | 0.20 |
| | C | 0.21 | 0.24 | 0.21 | C | 0.24 | 0.28 | 0.22 | C | 0.24 | 0.28 | 0.23 | C | 0.21 | 0.25 | 0.12 |

*Количество генов с выявленной 3-профильностью представлено в Таблице 2. Матрицы триплетной периодичности соответствующих генов для рода *Deltacoronavirus* не рассматривались вследствие недостаточного количества представителей. Также не рассматривались матрицы для M генов, так как для родов *Alphacoronavirus* и *Gammacoronavirus* была выявлена тенденция к однородности их нуклеотидной последовательности.

Оценки триплетных паттернов рассматриваемых генов для SARS-CoV-2 от остальных бета коронавирусов (см. табл. 4) различаются не столь существенно. К тому же, выше отмечалось, что матрицы скрытой 3-профильной периодичности для анализируемых в работе генов SARS-CoV-2 по трем географическим ареалам (Азия, Европа и США) оказались одинаковыми (см. табл. 1). Это наблюдение свидетельствует в поддержку гипотезы о едином источнике вируса SARS-CoV-2, вызвавшего пандемию COVID-19.

## ЗАКЛЮЧЕНИЕ

В работе были проанализированы более трех тысяч полных геномов коронавирусов из четырех родов, включая SARS-CoV-2, и проведено исследование структуры кодирования для четырёх основных генов (ORF1ab, S, M и N).

Для M-генов мембранных белков была выявлена тенденция к размытию их триплетной структуры кодирования. Возможно, эта тенденция есть следствие их быстрого мутирования, способствующего адаптации коронавируса к новым хозяевам рода.

В работе было показано совпадение усредненной структуры скрытой триплетной периодичности генов SARS-CoV-2 для трех географических ареалов (Азии, Европы и США). Такое совпадение позволяет предположить единый источник распространения вируса SARS-CoV-2.

## СПИСОК ЛИТЕРАТУРЫ


1. Guo Y.R., Cao Q.D., Hong Z.S., Tan Y.Y., Chen S.D., Jin H.J., Tan K.S., Wang D.Y., Yan Y. The origin, transmission and clinical therapies on coronavirus disease 2019 (COVID-19) outbreak – an update on the status. *Military Medical Research*. 2020. V. 7. Article No. 11. doi: 10.1186/s40779-020-00240-0
2. *Руководство по вирусологии: Вирусы и вирусные инфекции человека и животных*. Под. ред. Львова Д.К. М.: ООО Издательство «Медицинское информационное агентство», 2013.
3. Lai M.M.C., Brayton P.R., Armen R.C., Patton C.D., Pugh C., Stohlman S.A. Mouse hepatitis virus A59: mRNA structure and genetic localization of the sequence divergence from hepatotropic strain MHV-3. *Journal of virology*. 1981. V. 39. No. 3. P. 823–834.
4. Lai M.M.C., Baric R.S., Makino S., Keck J.G., Egbert J., Leibowitz J., Stohlman S.A. Recombination between non segmented RNA genomes of murine coronaviruses. *Journal of virology*. 1985. V. 56. No. 2. P. 449–456.
5. Yu W.-B., Tang G.-D., Zhang L., Corlett R.T. Decoding the evolution and transmissions of the novel pneumonia coronavirus (SARS-CoV-2 / HCoV-19) using whole genomic data. *Zoological Research*. 2020. V. 41. No. 3. P. 247−257. doi: 10.24272/j.issn.2095-8137.2020.022
6. Zhao Z., Li H., Wu X., Zhong Y., Zhang K., Zhang Y.-P., Boerwinkle E., Fu Y.-X. Moderate mutation rate in the SARS coronavirus genome and its implications. *BMC Evolutionary Biology*. 2004. V. 4. Article No. 21. doi: 10.1186/1471-2148-4-21
7. Liu L., Fang Q., Deng F., Wang H., Yi C.E., Ba L., Yu W., Lin R.D., Li T., Hu Z., et al. Natural mutations in the receptor binding domain of spike glycoprotein determine the reactivity of cross-neutralization between palm civet coronavirus and severe acute respiratory syndrome coronavirus. *Journal of Virology*. 2007. V. 81. No. 9. P. 4694–4700. doi: 10.1128/JVI.02389-06
8. Cotton M., Watson S.J., Kellam P., Al-Rabeeah A., Makhdoom F.R.C.S. H.Q., Assiri A., Al-Tawfiq J., Alhakeem R., Madani H., AlRabiah F., et al. Transmission and evolution of the Middle East respiratory syndrome coronavirus in Saudi Arabia: a descriptive genomic study. *Lancet*. 2013. V. 382. P. 1993–2002. doi: 10.1016/S0140-6736(13)61887-5
9. Wang N., Li S.Y., Yang X.L., Huang H.M., Zhang Y.J., Guo H., Luo C.M., Miller M., Zhu G., Chmura A.A., et al. Serological evidence of bat SARS-related coronavirus infection in humans. *Virologica Sinica*. 2018. V. 33. No. 1. P. 104-107. doi: 10.1007/s12250-018-0012-7

10. 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
11. Hogue B.G., Machamer C.E. Coronavirus structural proteins and virus assembly. In: *Nidoviruses*. Ed. Perlman S. ASM Press, 2007. P. 179–200.
12. Kim D., Lee J.-Y., Yang J.-S., Kim J.W., Kim V.N., Chang H. The architecture of SARS-CoV-2 transcriptome. *Cell*. 2020. V. 181. No. 4. P. 914–921.e10. doi: 10.1016/j.cell.2020.04.011
13. Schoeman, D., Fielding, B.C. Coronavirus envelope protein: current knowledge. *Virology Journal*. 2019. V. 16. Article No. 69. doi: 10.1186/s12985-019-1182-0
14. Chaley M., Kutyrkin V. Profile-Statistical Periodicity of DNA Coding Regions. *DNA Research*. 2011. V. 18. P. 353–362. doi: 10.1093/dnares/dsr023
15. Кутыркин В.А., Чалей М.Б. Модель организации кодирования в прокариотических организмах. *Математическая биология и биоинформатика*. 2016. Т. 11. № 1. С. 24–45. doi: 10.17537/2016.11.24
16. 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
17. *NCBI SARS-CoV-2 Resources*. URL: https://www.ncbi.nlm.nih.gov/sars-cov-2/ (accessed 30.10.2020).
18. *GenBank*. URL: https://ftp.ncbi.nlm.nih.gov/genbank/ (accessed 30.04.2020).
19. Chaley M., Kutyrkin V. Stochastic Models for Description of Structural-Statistical Properties in DNA Sequences. *Journal of Theoretical Biology*. 2020. V. 496. Article No. 110126. doi: 10.1016/j.jtbi.2019.110126
20. Wu F., Zhao S., Yu B., Chen Y.M., Wang W., Song Z.G., Hu Y., Tao Z.W., Tian J.H., Pei Y.Y., Yuan M.L., et al. A new coronavirus associated with human respiratory disease in China. *Nature*. 2020. V. 579. No. 7798. P. 265–269. doi: 10.1038/s41586-020-2008-3
21. Stadler K., Masignani V., Eickmann M., Becker S., Abrignani S., Klenk H.D., Rappuoli R. SARS – beginning to understand a new virus. *Nature Reviews. Microbiology*. 2003. V. 1. No. 3. P. 209–218. doi: 10.1038/nrmicro775
22. Neuman B.W., Kiss G., Kunding A.H., Bhella D., Baksh M.F., Connelly, S., Droese B., Klaus, J.P., Makino S., Sawicki S.G., Siddell S.G., et al. 2011. A structural analysis of M protein in coronavirus assembly and morphology. *Journal of Structural Biology*. V. 174. No. 1. P. 11–22. doi: 10.1016/j.jsb.2010.11.021
23. Li J.Y., Liao C.H., Wang Q., Tan Y.J., Luo R., Qiu Y., Ge X.Y. The ORF6, ORF8 and nucleocapsid proteins of SARS-CoV-2 inhibit type I interferon signaling pathway. *Virus Research*. 2020. V. 286. Artice No. 198074. doi: 10.1016/j.virusres.2020.198074

Рукопись поступила в редакцию 21.11.2020, переработанный вариант поступил 14.12.2020.
Дата опубликования 25.12.2020.

# Coding Structure for the ORF1ab, S, M and N Coronavirus Genes

**Chaley M.B.**[1], **Tyulko Zh.S.**[2,3], **Kutyrkin V.A.**[4]

[1]*Institute of Mathematical Problems of Biology, Keldysh Institute of Applied Mathematics of RAS, Pushchino, Russia*
[2]*Omsk State Medical University of Ministry of Healthcare of the Russian Federation,*
[3]*Federal Service for Surveillance on Consumer Rights Protection and Human Wellbeing, Omsk Research Institute of Natural Focal Infections, Omsk, Russia*
[4]*Moscow State Technical University n.a. N.E. Bauman, Moscow, Russia*

***Abstract.*** Spectral-statistical approach was applied to comparative analysis of coronavirus genomes from the four genus *Alphacoronavirus*, *Betacoronavirus* (including new SARS-CoV-2 virus), *Gammacoronavirus* and *Deltacoronavirus*. This analysis was done from the point of view of 3-regularity and latent triplet profile periodicity existence in the coding sequences of four structural genes: ORF1ab encoding transcriptase; S-gene of glycoprotein forming spikes; M-gene of membrane protein; N-gene of nucleoprotein. A whole number of the genomes analyzed was equal to 3410. Gene numbers in each of the four groups in the study respectively were the same. In the result, practically, in the CDSs of all analyzed genes of ORF1ab, S and N the latent profile triplet periodicity was revealed and high value of 3-regularity index, being a quality estimate of coding triplet structure conservation, was determined. On the contrary, for coding structure of M-genes a tendency was revealed to diffuse up to homogeneity for 60 % of the genes in the genomes of alphacoronaviruses analyzed and for 67 % of the genes of the gammacoronaviruses. Tendency of the such structure diffusion, being accompanied by decrease of 3-regularity index average value in comparison with other genes, while the triplet profile periodicity remains saved, was also noted for M-genes of SARS-CoV-2 viruses. Probably, this tendency reflects a significance of M-genes variability in coronavirus adaptation to the novel hosts of genus. Analysis of 3-profile periodicity matrices of the four groups of SARS-CoV-2 genes considered in the work, for the viruses isolated in Europe, Asia and USA, did not revealed their significant difference, that is allowing to propose a single source of this virus propagation.

***Key words:*** *property of 3-regularity in CDS, latent profile triplet periodicity, coronavirus genome, SARS-Cov-2 virus genome, ORF1ab, S-gene, M-gene, N-gene.*