

Типологические подходы к распознаванию рода и подрода коронавирусов по структурным и неструктурным генам

Чалей М.Б.^{*1}, Кутыркин В.А.^{**2}

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

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

Аннотация. В работе предлагаются типологические подходы к распознаванию рода и подрода коронавирусов на основе частот кодонов в кодирующих последовательностях их генов. Распознавание осуществляется на основе статистики, выявляющей отклонение рассматриваемого гена коронавируса от соответствующего гена из генома известного рода или подрода. В работе рассматривались структурные S- и N-гены, и также неструктурные гены, объединенные единой рамкой считывания ORF1ab. Три из четырех подходов выявили высокую эффективность распознавания рода и подрода коронавирусов на основании N-гена. В последовательности N-гена были выделены кодоны наиболее значимые для распознавания. Предложенный статистический подход для определения таксономической принадлежности вирусов является менее трудоемким в сравнении с методами филогенетического анализа и может быть использован для аннотирования вирусных геномов.

Ключевые слова: распознавание рода и подрода коронавирусов, геном коронавируса, прототипные штаммы коронавирусов, S-, N-гены коронавируса, ORF1ab коронавируса.

ВВЕДЕНИЕ

Вирусы являются одними из главных возбудителей инфекционных и онкологических заболеваний человека, животных и растений [1–4]. Но они также служат и моделями для работы над актуальными задачами генетики и молекулярной биологии. Например, технология рекомбинантных РНК [5], созданная на основе вируса SV40 способствовала многим достижениям биотехнологии, в частности, промышленному получению генноинженерного инсулина, гормона роста и интерферона [6, 7].

Часто вирусные заболевания протекают со сходной симптоматикой. Но для лечения и описания инфекции нужно выяснить, какой именно вирус вызвал заболевание. С этой целью используются, в том числе, и биоинформационные методы. Секвенированный вирусный геном сканируют по базам данных для выявления наилучшего сходства с известными геномами вирусов. При этом используются методы выравнивания последовательностей геномов, с последующим построением их филогенетических деревьев [8]. Подобным образом было установлено вероятное происхождение коронавируса SARS-CoV-2, вызвавшего пандемию COVID-19, от коронавируса летучих мышей [9, 10].

*maramaria@yandex.ru

**vkutyrrkin@yandex.ru

Метагеномные исследования, основанные на технологии секвенирования нового поколения и дальнейшем биоинформационном анализе, предоставляют новые возможности для быстрой идентификации уже изолированных вирусов и для поиска новых вирусов непосредственно в пробах тканей животных или человека. Изучаются вирусы, содержащиеся в пробах воды, воздуха, почвы. Так в исследованиях морской воды были обнаружены 5.5 тысяч новых видов РНК-вирусов, принадлежащих как к основным пяти типам РНК-вирусов, так и к новым типам [11].

В связи с бурным ростом данных о вирусных геномах была отмечена необходимость более тесного взаимодействия биоинформатики и вирусологии, так как биоинформатика может предоставить вирусологам как базы, так и методы анализа данных [12]. В частности, среди задач биоинформатики в вирусологии указывалось на аннотирование вирусных геномов. Вирусная информатика подразумевает существование целого комплекса баз данных, баз знаний о вирусах и программные инструменты для работы с ними [13]. За время пандемии появилось много специальных программных инструментов для работы с геномами SARS-CoV-2 [14, 15]. В настоящее время вирусная информатика – это новая активно развивающаяся ветвь биоинформатики, в которой соединяются биоинформационные модели, компьютерные программы множественного выравнивания последовательностей, обширные исследования геномов вирусов и их эволюционный анализ [13, 16].

Вирусы сложно классифицировать, так как у них не выделяется каких-то консервативных генов, как у бактерий и эукариот, для которых классификация строится, соответственно, на основании 16S и 18S РНК. Исторически сложилось, что до недавнего времени доминирующим фактором в изучении и классификации вирусов являлась их патогенность [17]. С развитием технологий секвенирования и наступлением эры метагеномных исследований [18–20] Международный комитет по таксономии вирусов (ICTV) [21] начал применять математические подходы к системному описанию вирусов. Также было объявлено о введении единого бинарного формата именования вирусов, с указанием рода и вида, как наименьшего ранга вирусной классификации [22]. Вид относится к некоторому подроду, подрод – к роду, род – к подсемейству, подсемейство входит в семейство, и так далее происходит переход к таксонам более высокого ранга, до последнего 15-го [23]. Построение таксономической структуры вирусов опирается на филогенетические дендрограммы при предварительном выравнивании последовательностей вирусных геномов [8]. Это весьма трудоемкий метод, особенно при быстром росте количества вирусных геномов.

В настоящей работе на примере распознавания подрода и рода коронавирусов предлагается более простой и достаточно эффективный типологический подход к аннотированию вирусов, использующий частотные характеристики кодонов отдельных генов. Для этого в работе проводится сравнение распознавания рода и подрода коронавирусов на основе структурных генов, кодирующих белки вириона (белок нуклеокапсида N и спайк белок S), и генов неструктурных белков, объединенных в единую рамку считывания ORF1ab. Для типологического подхода характерно усреднение известных данных для определения сходства или подобия с ними рассматриваемого объекта.

Ранее для распознавания вида и подтипа флавивируса по его геному, кодирующему полипротеин, был предложен метод, основанный на использовании усредненных распределений частот кодонов в выборках геномов флавивирусов из GenBank [24] по каждому из рассматриваемых видов и подтипов [25]. Распознавание осуществлялось на основе статистики, измеряющей отклонение частот кодонов в анализируемой последовательности от средних частот в каждой группе видов и подтипов флавивирусов. Метод показал практически 100 %-ю надежность. Аналогично распознаванию флавивирусов по геному, кодирующему единый полипротеин, для распознавания рода коронавирусов моделировались варианты искусственных геномов

для квази-протеинов. На их основе также была достигнута высокая достоверность при распознавании рода (подрода) коронавируса [26].

В настоящей работе для распознавания рода и подрода коронавируса на основе распределения частот кодонов (векторов частот) в структурных и неструктурных генах используются следующие типологические подходы.

Сначала рассматриваются подходы, направленные на распознавание рода анализируемого коронавируса.

В первом подходе создаются усредненные вектора частот по четырем родам коронавируса. Анализируемый коронавирус относится к тому роду, усредненный частотный вектор которого наиболее близок к рассматриваемому вектору частот.

Во втором подходе для каждого рода создаются усредненные вектора частот его прототипных штаммов. На основе сравнения с ними вектора частот анализируемого коронавируса выбирается его род.

Следующие два подхода направлены на распознавание подрода (тем самым и рода). Для этого рассматриваются коронавирусы, для которых выполнено аннотирование по подкладам.

В третьем подходе для каждого подрода создаются усредненные вектора частот кодонов прототипных штаммов подрода. На их основе, аналогично первому подходу, выбирается подрод (и тем самым и род) анализируемого коронавируса.

В четвертом подходе для каждого подрода создаются вектора частот индивидуальных прототипных штаммов, на основе которых выбирается подрод анализируемого генома коронавируса.

При анализе результатов четырех подходов выявляется, что наиболее эффективным (в сравнении с S-геном и ORF1ab) оказывается распознавание подрода (и тем самым и рода) на основе N-гена при использовании четвертого подхода.

МАТЕРИАЛЫ И МЕТОДЫ

Рассматриваемый типологический подход к распознаванию рода (подрода) коронавируса основан на статистическом анализе векторов частот кодонов (распределения частот кодонов) в генах коронавируса. Согласно современной классификации, коронавирусы подразделяются на два подсемейства, одно из которых содержит только один сравнительно недавно открытый летовирус узкоротых квакш (*Microhyla fissipes*). Все остальные коронавирусы объединены в подсемейство *Orthocoronavirinae*, подразделяемое на четыре рода: *Alphacoronavirus* (α -CoV) и *Betacoronavirus* (β -CoV), *Deltacoronavirus* (δ -CoV) и *Gammacoronavirus* (γ -CoV) [27]. В настоящей работе рассматривается распознавание этих четырех родов, которые подразделяются на 23 подрода, как показано в таблице 1.

Таблица 1. Названия подродов коронавируса и их обозначения, принятые в работе. Количество прототипных штаммов (п/т) подрода и число анализируемых, аннотированных по подроду, геномов из GenBank

Род	Подрод	Обозначение	Количество п/т штаммов	Количество аннотированных по подроду геномов без п/т штаммов
<i>Alphacoronavirus</i> α -Cov	<i>Colacovirus</i>	$\alpha 01$	1	1
	<i>Decacovirus</i>	$\alpha 02$	2	12
	<i>Duvinacovirus</i>	$\alpha 03$	1	66
	<i>Luchacovirus</i>	$\alpha 04$	1	1
	<i>Minacovirus</i>	$\alpha 05$	3	3

Род	Подрод	Обозначение	Количество п/т штаммов	Количество аннотированных по подроду геномов без п/т штаммов
	<i>Minunacovirus</i>	$\alpha 06$	2	3
	<i>Myotacovirus</i>	$\alpha 07$	1	1
	<i>Nyctacovirus</i>	$\alpha 08$	1	1
	<i>Pedacovirus</i>	$\alpha 09$	2	4
	<i>Rhinacovirus</i>	$\alpha 10$	1	5
	<i>Setracovirus</i>	$\alpha 11$	2	62
	<i>Tegacovirus</i>	$\alpha 12$	5	70
<i>Betacoronavirus</i> β -Cov	<i>Embecovirus</i>	$\beta 01$	12	251
	<i>Hibecovirus</i>	$\beta 02$	1	1
	<i>Merbecovirus</i>	$\beta 03$	6	330
	<i>Nobecovirus</i>	$\beta 04$	2	9
	<i>Sarbecovirus</i>	$\beta 05$	7	1028
<i>Deltacoronavirus</i> δ -Cov	<i>Andecovirus</i>	$\delta 01$	1	1
	<i>Buldecovirus</i>	$\delta 02$	7	5
	<i>Herdecovirus</i>	$\delta 03$	1	1
	<i>Moordecovirus</i>	$\delta 04$	1	1
<i>Gammacoronavirus</i> γ -Cov	<i>Cegacovirus</i>	$\gamma 01$	2	2
	<i>Igacovirus</i>	$\gamma 02$	5	126
ИТОГО:			67	1984

Для создания метода распознавания из генома коронавируса были отобраны отдельные, особо значимые гены: ген нуклеопротеина N – белка, формирующего комплекс с РНК-геномом коронавируса; ген S-белка, образующего шипы на поверхности вирусного вириона, и рамка считывания ORF1ab, содержащая неструктурные гены коронавируса. Средняя длина этих генов в нуклеотидах (нукл.) представлена в таблице 2.

Таблица 2. Средняя длина CDS анализируемых генов коронавирусов

ген	<i>l</i> , нукл.
ORF1ab	20000
S	3700
N	1200

Для создания в каждом роде (и подроде) коронавирусов усредненных векторов частот кодонов использовались аннотированные геномы из базы данных GenBank [24]. Данные о кодах доступа GenBank для набора последовательностей геномов прототипных штаммов, представлены в [27].

Для каждого из трех генов ORF1ab, S и N последовательности генома прототипного штамма коронавируса рассматривался соответствующий вектор 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 коронавируса четырех родов, полученные из базы данных нуклеотидных последовательностей GenBank [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.

Ранее, при создании метода распознавания видов флавивирусов использовалась статистика, основанная на сравнении векторов частот кодонов в гене полипротеина флавивируса [25]. Аналогичная статистика используется при распознавании рода и подрода коронавируса для каждого гена (ORF1ab, S- и N-генов).

В работе использовались 3242 генома коронавируса четырех родов, полученные из базы данных GenBank. Из этого числа для 2051 генома вместе с аннотированием рода был указан и подрод.

При распознавании из рассмотрения исключались пять кодонов с самыми низкими частотами, три из которых являлись кодонами терминации, и два кодона (tgt и tgc) цистеина (CYS). Таким образом, каждый рассматриваемый геном коронавируса характеризовался распределением частот 59 кодонов в своем гене.

Распознавание рода коронавируса на основе первого и второго типологических подходов

I) Для анализируемого гена рассмотрим вектор p частот кодонов $p = (p_1, p_2, \dots, p_{59})$. Пусть $X \in \{\alpha, \beta, \delta, \gamma\}$, где X – символ, обозначающий род коронавируса. Используя это обозначение, отклонение $D(P^X, p)$ вектора p частот кодонов в анализируемом гене коронавируса неизвестного рода от усредненного вектора $P^X = (P_1^X, P_2^X, \dots, P_{59}^X)$ частот кодонов рода X -CoV в рассматриваемом гене вычисляется по формуле:

$$D_1(P^X, p) = \sum_{i=1}^{59} \frac{|P_i^X - p_i|}{P_i^X}. \quad (1)$$

Среди отклонений $D_1(P^\alpha, p)$, $D_1(P^\beta, p)$, $D_1(P^\delta, p)$ и $D_1(P^\gamma, p)$ выбирается минимальное, которое определяет результат распознавания рода анализируемого генома коронавируса с помощью выбранного гена. Таким образом, согласно первому подходу, выбирается род анализируемого коронавируса рассматриваемого гена.

II) Пусть $Q^X = (Q_1^X, Q_2^X, \dots, Q_{59}^X)$ – усредненный вектор частот по всем прототипным штаммам рода X -CoV в рассматриваемом гене. Тогда отклонение $D_2(Q^X, p)$ вектора p частот кодонов в анализируемом гене коронавируса неизвестного рода от усредненного вектора Q^X частот кодонов прототипных штаммов рода X -CoV вычисляется по формуле:

$$D_2(Q^X, p) = \sum_{i=1}^{59} \frac{|Q_i^X - p_i|}{Q_i^X}. \quad (2)$$

Среди отклонений $D_2(Q^\alpha, p)$, $D_2(Q^\beta, p)$, $D_2(Q^\delta, p)$ и $D_2(Q^\gamma, p)$ выбирается минимальное, которое определяет результат распознавания рода анализируемого генома коронавируса с помощью выбранного гена. Таким образом, согласно второму подходу, выбирается род анализируемого коронавируса рассматриваемого гена.

Распознавание подрода коронавируса на основе третьего и четвертого типологических подходов

III) Поскольку 67 прототипных штаммов распределены по 23-м подкладам (см. табл. 1), то каждый подклад в рассматриваемом гене будем характеризовать усредненным распределением частот 59 кодонов гена в подкладе. Такое распределение будем называть усредненным вектором частот кодонов подклада рассматриваемого гена (ORF1ab, S- или N-гена).

Как и ранее, $X \in \{\alpha, \beta, \delta, \gamma\}$ – символ, используемый для обозначения рода коронавируса и j – двузначный номер подклада в своем роде, согласно таблице 1. Например, $\alpha 12$ – обозначение подклада *Tegacovirus* рода α -CoV. Каждый усредненный вектор частот кодонов гена в подкладе Xj будет описываться строкой вида: $\mathbf{P}^{Xj} = (P_1^{Xj}, P_2^{Xj}, \dots, P_{59}^{Xj})$, где P_i^{Xj} – частота i -го кодона в усредненном векторе частот кодонов подклада Xj для $i = \overline{1, 59}$. Рассматриваемый геном коронавируса будет описываться вектором частот кодонов своего гена, который определяется строкой: $\mathbf{p} = (p_1, p_2, \dots, p_{59})$, где p_i – частота i -го кодона в векторе частот кодонов гена этого генома.

Отклонение вектора частот $\mathbf{p}$ рассматриваемого генома коронавируса от усредненного вектора $\mathbf{P}^{Xj}$ частот подклада Xj определяется числом $D_3^{Xj}(\mathbf{p})$, вычисляемым по формуле:

$$D_3^{Xj}(\mathbf{P}^{Xj}, \mathbf{p}) = \sum_{i=1}^{59} \frac{|P_i^{Xj} - p_i|}{P_i^{Xj}}. \quad (3)$$

Получив значения отклонений (3) от каждого из векторов частот кодонов подкладов, выбирают тот подклад, для которого отклонение минимально. Таким образом, согласно третьему подходу, выбирается подклад анализируемого коронавируса для рассматриваемого гена (ORF1ab, S- или N-гена).

IV) Рассмотрим другой подход к распознаванию подклада коронавируса. В этом подходе для рассматриваемого гена (ORF1ab, S- или N-гена) выявляется один из 67 его индивидуальных прототипов, к которому наиболее близок ген анализируемого коронавируса. Подклад этого прототипа признается подкладом анализируемого коронавируса для рассматриваемого гена.

Как и в первом подходе, $X \in \{\alpha, \beta, \delta, \gamma\}$ – символ, используемый для обозначения рода коронавируса и j – двузначный номер подклада в своем роде. Обозначение Xj, k будет использоваться для k -го индивидуального прототипа подклада Xj рассматриваемого гена (ORF1ab, S- или N-). Каждый вектор частот кодонов k -го $k = (\overline{1, 67})$ индивидуального прототипа в подкладе Xj этого гена будет описываться вектором вида: $\mathbf{q}^{Xj, k} = (q_1^{Xj, k}, q_2^{Xj, k}, \dots, q_{59}^{Xj, k})$, где $q_i^{Xj, k}$ – частота i -го кодона в векторе частот прототипного штамма Xj, k для $i = \overline{1, 59}$. Как и ранее, рассматриваемый геном коронавируса будет описываться вектором частот кодонов своего гена, которое определяется вектором вида: $\mathbf{p} = (p_1, p_2, \dots, p_{59})$, где p_i – частота i -го кодона в векторе частот кодонов гена.

Отклонение рассматриваемого гена коронавируса с вектором частот $\mathbf{p}$ от такого же гена прототипного штамма Xj, k определяется числом $D_4^{Xj, k}(\mathbf{q}^{Xj, k}, \mathbf{p})$, вычисляемым по формуле:

$$D_4^{Xj, k}(\mathbf{q}^{Xj, k}, \mathbf{p}) = \sum_{i=1}^{59} \frac{|q_i^{Xj, k} - p_i|}{q_i^{Xj, k}}. \quad (4)$$

Получив значения отклонений (4) от всех векторов частот индивидуальных прототипных штаммов, выбирают тот штамм $X_{j,k}$, для которого отклонение минимально. В результате выбирается принадлежность рассматриваемого гена коронавируса к подроду X_j этого гена. В результате определения подрода считается определенным и род X .

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Рассмотрим результаты применения статистики (1) первого типологического подхода к распознаванию рода геномов коронавирусов из базы GenBank. Число случаев неверного распознавания каждого рода коронавирусов с помощью структурных генов S, N и неструктурных генов в единой рамке считывания ORF1ab показано в таблице 3.

Таблица 3. Количество ошибок распознавания рода коронавирусов из базы GenBank при использовании усредненных (по родам) всех векторов частот кодонов рассматриваемых генов (вместе с прототипными штаммами)

Род коронавируса	ORF1ab	S-ген	N-ген
α -CoV Всего: 924	190	36	29
β -CoV Всего: 1954	648	232	72
δ -CoV Всего: 19	0	2	0
γ -CoV Всего: 345	2	5	3
Итого: 3242	840	275	104

Из таблицы 3 видно, что применение первого типологического подхода при распознавании рода коронавирусов дает неудовлетворительные результаты на основе ORF1ab и S-гена. В то время как использование N-гена ведет к высокой эффективности распознавания рода.

Применим к распознаванию рода коронавирусов статистику (2) второго типологического подхода. Число ошибочных распознаваний с помощью генов S, N и ORF1ab приведены в таблице 4.

Таблица 4. Количество ошибок распознавания рода коронавирусов из базы GenBank при использовании усредненных (по родам) векторов частот кодонов для генов прототипных штаммов

Род коронавируса	ORF1ab	S-ген	N-ген
α -CoV Всего: 924	14	26	181
β -CoV Всего: 1954	48	56	30
δ -CoV Всего: 19	3	2	0
γ -CoV Всего: 345	2	153	2
Итого: 3242	67	237	213

Из таблицы 4 следует, что наилучший результат в распознавании на основе второго типологического подхода показывают неструктурные гены, объединенные рамкой считывания ORF1ab. Отметим, что менее значимый результат распознавания рода на основе N-гена обусловлен большим количеством ошибок в α -роде, разделенным на 12 подродов, что значительно превосходит число подродов во всех остальных родах (см. табл. 1). Этот факт говорит о том, что второй типологический подход мало пригоден для таксонов, выделяющихся значительным расщеплением на таксоны более низкого ранга.

От определения рода коронавируса перейдем к распознаванию подрода, что позволяет автоматически идентифицировать и род. Для этого сначала будем использовать третий типологический подход на основе анализа 2051 аннотированных по подроду генома из числа 3025 геномов известного рода. В таблице 5 показано количественное распределение геномов коронавируса с известным подродом (см. табл. 1). Правее приводятся данные о числе ошибок при распознавании подрода коронавируса для трех рассматриваемых генов. В последней строке таблицы 5 приведено общее число ошибочных распознаваний подрода коронавируса по каждому из генов S, N и ORF1ab при использовании третьего типологического подхода.

Таблица 5. Количество ошибок в распознавании подрода для 2051 генома коронавируса, аннотированных по подкладам (включая прототипные штаммы), с использованием усредненных по подкладам частот кодонов в ORF1ab, и в S- и N- генах (обозначения подродов коронавируса соответствуют таблице 3)

Подрод	Размер выборки	Ошибки в определении подрода		
		ORF1ab	S-ген	N-ген
α 01	2	0	0	0
α 02	14	3	2	4
α 03	67	1	3	4
α 04	2	0	0	0
α 05	6	0	0	0
α 06	5	2	0	1
α 07	2	0	0	0
α 08	2	0	0	0
α 09	6	2	0	0
α 10	6	0	0	0
α 11	64	3	3	3
α 12	75	1	0	0
β 01	263	52	17	7
β 02	2	0	0	0
β 03	336	18	31	30
β 04	11	0	0	0
β 05	1035	120	114	2
δ 01	2	0	0	0
δ 02	12	1	0	0
δ 03	2	0	0	0
δ 04	2	0	0	0
γ 01	4	0	0	0
γ 02	131	2	2	1
Итого:	2051	205	172	52

Как видно из таблицы 5, наибольшая эффективность распознавания подрода достигается с помощью N-гена, который обеспечивает высокую надежность (порядка 97 %).

Из результатов применения первого (табл. 3) и третьего (табл. 5) типологических подходов можно сделать вывод о наибольшей эффективности распознавания подрода (рода) коронавируса на основе N-гена. Кроме того, из таблицы 2 видно, что к

преимуществу его использования для распознавания относится его малая, в сравнении с другими генами, длина (1200 нукл.). Поэтому далее для распознавания подрода коронавируса используется четвертый типологический подход на основе только N-гена. Подчеркнем, что далее в четвертом типологическом подходе распознавание осуществляется с помощью сравнения с векторами частот кодонов N-гена в индивидуальных прототипных штаммах подродов коронавируса.

При применении типологических подходов можно сокращать число используемых кодонов, выделяя особые совокупности кодонов для оптимизации распознавания. Ранее при применении четвертого типологического подхода на основе N-гена была рассмотрена совокупность из 38 кодонов [28], использование которой повысило эффективность распознавания подрода и рода коронавируса. В таблице 6 показано количественное распределение геномов коронавируса с известным подродом (см. табл. 1). Правее приводятся данные о числе ошибок при распознавании подрода коронавируса для N-гена на основе использования исходных 59 кодонов и выделенных 38 кодонов. В последней строке таблицы 6 приведено общее число ошибочных распознаваний подрода коронавируса по N-гену при использовании 59 и 38 кодонов.

Таблица 6. Количество случаев неверно распознанного подрода по N-генам из 2051 генома коронавируса, аннотированных по подкладам (включая прототипные штаммы), с использованием индивидуальных распределений частот кодонов в N-генах прототипных штаммов подрода (обозначения подродов коронавируса соответствуют табл. 1)

Подрод	Размер выборки	Ошибки в определении подрода с помощью 59 кодонов N-гена	Ошибки в определении подрода с помощью 38 кодонов N-гена
α 01	2	0	0
α 02	14	1	2
α 03	67	30	3
α 04	2	0	0
α 05	6	0	0
α 06	5	0	0
α 07	2	0	0
α 08	2	0	0
α 09	6	2	2
α 10	6	1	0
α 11	64	1	1
α 12	75	0	0
β 01	263	0	0
β 02	2	0	0
β 03	336	0	0
β 04	11	0	0
β 05	1035	1	2
δ 01	2	0	0
δ 02	12	0	0
δ 03	2	0	0
δ 04	2	0	0
γ 01	4	0	0
γ 02	131	0	0
Итого:	2051	36	10

Как видно из сравнения таблицы 6 и таблицы 5, четвертый типологический подход на основе исходных 59 кодонов N-гена эффективнее третьего для распознавания подрода коронавируса. Последняя строка таблицы 6 показывает, что сокращение числа кодонов с 59 до 38 в четвертом типологическом подходе ведет к повышению

эффективности распознавания подрода коронавирусов. Вместе с тем ошибки распознавания подрода не выходят за границы своего рода.

Рисунок 1 показывает 38 кодонов N-гена, выделенных в четвертом типологическом подходе для распознавания подрода коронавирусов, в структуре разбиения генетического кода на «фамильные боксы» из четырех кодонов. Трехбуквенные обозначения аминокислот показаны слева от соответствующих им кодонов. Как можно видеть из рисунка, в каждом боксе находится хотя бы один кодон из выделенных 38.

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

Рис. 1. Размещение в структуре генетического кода 38 кодонов (выделены голубым цветом), повышающих эффективность распознавания подрода на основе N-гена в четвертом типологическом подходе.

ЗАКЛЮЧЕНИЕ

В работе рассматривалась возможность распознавания рода и подрода коронавируса отдельно на основе его структурных S- и N- генов и неструктурных генов, объединенных в единую рамку считывания ORF1ab. Исследование производилось с помощью распределения частот (векторов частот) кодонов этих генов. Для этого использовалась выборка геномов коронавирусов из базы GenBank с известным родом, две трети из которых были также аннотированы по подроду.

Для распознавания были рассмотрены четыре типологических подхода. В первом подходе для каждого гена в качестве типа рода коронавирусов рассматривался усредненный вектор частот кодонов для всех генов этого рода. Принадлежность рассматриваемого генома к тому или иному роду коронавируса для каждого гена определялась по степени отклонения его вектора частот от типов для каждого рода. Наименьшее отклонение указывало на род коронавируса.

Во втором типологическом подходе для каждого гена в качестве типа рода коронавирусов рассматривался усредненный вектор частот кодонов генов прототипных штаммов этого рода. Аналогично первому подходу, принадлежность рассматриваемого генома на основе анализируемого гена определялась согласно наименьшему отклонению его вектора частот кодонов от типа каждого рода.

В третьем и четвертом подходах распознавался подрод рассматриваемого генома коронавируса на основе каждого из анализируемых генов. В третьем подходе для каждого гена в качестве типа подрода коронавирусов рассматривался усредненный

вектор частот кодонов всех генов этого подрода. Подрод генома для каждого гена определялся аналогично первому подходу.

В четвертом подходе для каждого гена в качестве типа коронавируса рассматривается вектор частот гена индивидуального прототипного штамма. Таким образом рассматривались 67 типов коронавирусов. Поскольку каждый тип однозначно принадлежит некоторому подроду коронавирусов, распознавание типа определяет и подрод. Принадлежность гена к типу определяется минимальным отклонением вектора частот коронавируса от соответствующего типа.

Применение первого и третьего подхода показало их высокую эффективность (порядка 97 %). Эффективность второго подхода существенно снижена расщеплением рода α -CoV на большое количество подродов. Применение первого и третьего подхода выявили наибольшую эффективность распознавания на основе N-гена. Поэтому четвертый подход для распознавания подрода коронавируса основывался только на N-гене. Этот подход оказался наиболее эффективным для выявления рода и подрода коронавируса. Кроме того, этот подход позволил сократить количество кодонов в N-гене коронавирусов для повышения эффективности распознавания их рода и подрода до практически 100 % распознавания.

Рассмотренные примеры распознавания показали, что идентификация вирусного таксона может быть более надёжной на основе таксонов более низкого ранга.

Предложенные в настоящей работе типологические подходы могут быть использованы для классификации других вирусов и выявления наиболее информативных генов в вирусных геномах.

СПИСОК ЛИТЕРАТУРЫ

1. Greber U.F., Bartenschlager R. Editorial: An expanded view of viruses. *FEMS Microbiol Rev.* 2017. V. 41. No. 1. P. 1–4. doi: [10.1093/femsre/fuw044](https://doi.org/10.1093/femsre/fuw044)
2. Львов Д.К. Рождение и развитие вирусологии – история изучения новых и возвращающихся вирусных инфекций. *Вопросы вирусологии.* 2012. Т. 57. № 1S. С. 5–20.
3. Koyuncu O.O., Hogue I.B., Enquist L.W. Virus infections in the nervous system. *Cell Host Microbe.* 2013. V. 13. No. 4. P. 379–393. doi: [10.1016/j.chom.2013.03.010](https://doi.org/10.1016/j.chom.2013.03.010)
4. Schiller J.T., Lowy D.R. An introduction to virus infections and human cancer. *Recent Results Cancer Res.* 2021. V. 217. P. 1–11. doi: [10.1007/978-3-030-57362-1_1](https://doi.org/10.1007/978-3-030-57362-1_1)
5. Jackson D.A., Symons R.H., Berg P. Biochemical method for inserting new genetic information into DNA of Simian Virus 40: Circular SV40 DNA molecules containing Lambda phage genes and the galactose operon of *Escherichia coli*. *Proc. Natl. Acad. Sci. USA.* 1972. V. 69. No. 10. P. 2904–2909. doi: [10.1073/pnas.69.10.2904](https://doi.org/10.1073/pnas.69.10.2904)
6. Nagata S., Taira H., Hall A., Johnsrud L., Streuli M., Ecsödi J., Boll W., Cantell K., Weissmann C. Synthesis in *E. coli* of a polypeptide with human leukocyte interferon activity. *Nature.* 1980. V. 284. No. 5754. P. 316–320. doi: [10.1038/284316a0](https://doi.org/10.1038/284316a0)
7. Aggarwal S.R. What's fueling the biotech engine-2011 to 2012. *Nature Biotech.* 2012. V. 30. No. 12. P.1191–1197. doi: [10.1038/nbt.2437](https://doi.org/10.1038/nbt.2437)
8. Katoh K., Rozewicki J., Yamada K.D. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Brief. Bioinform.* 2019. V. 20. No. 4. P. 1160–1166. doi: [10.1093/bib/bbx108](https://doi.org/10.1093/bib/bbx108)
9. Zhou P., Yang X.L., Wang X.G., Hu B., Zhang L., Zhang W., Si H.R., Zhu Y., Li B., Huang C.L., et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature.* 2020. V. 579. No. 7798. P. 270–273. doi: [10.1038/s41586-020-2012-7](https://doi.org/10.1038/s41586-020-2012-7)

10. Boni M.F., Lemey P., Jiang X., Lam T.T., Perry B.W., Castoe T.A., Rambaut A., Robertson D.L. Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. *Nat. Microbiol.* 2020. V. 5. No. 11. P. 1408–1417. doi: [10.1038/s41564-020-0771-4](https://doi.org/10.1038/s41564-020-0771-4)
11. Zayed A.A., Wainaina J.M., Dominguez-Huerta G., Pelletier E., Guo J., Mohssen M., Tian F., Pratama A.A., Bolduc B., Zablocki O., et al. Cryptic and abundant marine viruses at the evolutionary origins of Earth's RNA virome. *Science*. 2022. V. 376. No. 6589. P. 156–376162. doi: [10.1126/science.abm5847](https://doi.org/10.1126/science.abm5847)
12. Ibrahim B., McMahon D.P., Hufsky F., Beer M., Deng L., Mercier P.L., Palmarini M., Thiel V., Marz M. A new era of virus bioinformatics. *Virus Res.* 2018. V. 251. P. 86–90. doi: [10.1016/j.virusres.2018.05.009](https://doi.org/10.1016/j.virusres.2018.05.009)
13. Lin Y., Qian Y., Qi X., Shen B. Databases, knowledgebases, and software tools for virus informatics. *Adv. Exp. Med. Biol.* 2022. V. 1368. P. 1–19. doi: [10.1007/978-981-16-8969-7_1](https://doi.org/10.1007/978-981-16-8969-7_1)
14. Tan M., Xia J., Luo H., Meng G., Zhu Z. Applying the digital data and the bioinformatics tools in SARS-CoV-2 research. *Comput. Struct. Biotechnol. J.* 2023. V. 21. P. 4697–4705. doi: [10.1016/j.csbj.2023.09.044](https://doi.org/10.1016/j.csbj.2023.09.044)
15. Hu T., Li J., Zhou H., Li C., Holmes E.C., Shi W. Bioinformatics resources for SARS-CoV-2 discovery and surveillance. *Brief. Bioinform.* 2021. V. 22. No. 2. P. 631–641. doi: [10.1093/bib/bba386](https://doi.org/10.1093/bib/bba386)
16. Vello F., Filippini F., Righetto I. Bioinformatics goes viral: I. Databases, phylogenetics and phylodynamics tools for boosting virus research. *Viruses*. 2024. V. 16. No. 9. Article No. 1425. doi: [10.3390/v16091425](https://doi.org/10.3390/v16091425)
17. Gorbalenya A.E., Siddell S.G. Recognizing species as a new focus of virus research. *PLoS Pathog.* 2021. V. 17. No. 3. Article No. e1009318. doi: [10.1371/journal.ppat.1009318](https://doi.org/10.1371/journal.ppat.1009318)
18. Höper D., Wylezich C., Beer M. Loeffler 4.0: diagnostic metagenomics. *Adv. Virus Res.* 2017. V. 99. P. 17–37. doi: [10.1016/bs.aivir.2017.08.001](https://doi.org/10.1016/bs.aivir.2017.08.001)
19. Greninger A.L. A decade of RNA virus metagenomics is (not) enough. *Virus Res.* 2018. V. 244. P. 218–229. doi: [10.1016/j.virusres.2017.10.014](https://doi.org/10.1016/j.virusres.2017.10.014)
20. Zhang Y.Z., Shi M., Holmes E.C. Using metagenomics to characterize an expanding virosphere. *Cell*. 2018. V. 172. No. 6. P. 1168–1172. doi: [10.1016/j.cell.2018.02.043](https://doi.org/10.1016/j.cell.2018.02.043)
21. Adams M.J., Lefkowitz E.J., King A.M., Harrach B., Harrison R.L., Knowles N.J., Kropinski A.M., Krupovic M., Kuhn J.H., Mushegian A.R., et al. 50 years of the International Committee on Taxonomy of Viruses: progress and prospects. *Arch. Virol.* 2017. V. 162. No. 5. P. 1441–1446. doi: [10.1007/s00705-016-3215-y](https://doi.org/10.1007/s00705-016-3215-y)
22. Walker P.J., Siddell S.G., Lefkowitz E.J., Mushegian A.R., Adriaenssens E.M., Alfenas-Zerbini P., Davison A.J., Dempsey D.M., Dutilh B.E., García M.L., Harrach B., et al. Changes to virus taxonomy and to the international code of virus classification and nomenclature ratified by the International Committee on Taxonomy of Viruses (2021). *Arch. Virol.* 2021. V. 166. No. 9. P. 2633–2648. doi: [10.1007/s00705-021-05156-1](https://doi.org/10.1007/s00705-021-05156-1)
23. Gorbalenya A.E., Krupovic M., Mushegian A., Kropinski A.M., Siddell S.G., Varsani A., Adams M.J., Davison A.J., Dutilh B.E., Harrach B. et al. The new scope of virus taxonomy: partitioning the virosphere into 15 hierarchical ranks. *Nat. Microbiol.* 2020. V. 5. No. 5. P. 668–674. doi: [10.1038/s41564-020-0709-x](https://doi.org/10.1038/s41564-020-0709-x)
24. GenBank. URL: <https://www.ncbi.nlm.nih.gov/genbank> (accessed 29.11.2024). Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. *Nucleic Acids Res.* 2013. V. 41 (Database issue). Article No. D36–42. doi: [10.1093/nar/gks1195](https://doi.org/10.1093/nar/gks1195)

25. Чалей М.Б., Тюлько Ж.С., Кутыркин В.А. Распознавание видов флавивирусов на основе кодирующих последовательностей полипротеинов. *Мат. Биол. Биоинф.* 2019. Т. 14. № 2. С. 533–542. doi: [10.17537/2019.14.533](https://doi.org/10.17537/2019.14.533)
26. Чалей М.Б., Кутыркин В.А. Распознавание рода коронавируса на основе прототипных штаммов. *Мат. Биол. Биоинф.* 2022. Т. 17. № 1. С. 10–27. doi: [10.17537/2022.17.10](https://doi.org/10.17537/2022.17.10)
27. 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](https://doi.org/10.15789/2220-7619-HOI-1412)
28. Chaley M., Kutyrkin V. Optimization of a coronavirus genus recognition procedure based on the N-gene of prototypic strains. *E3S Web of Conferences*. 2023. V. 419. Article No. 02010. doi: [10.1051/e3sconf/202341902010](https://doi.org/10.1051/e3sconf/202341902010)

Рукопись поступила в редакцию 21.11.2024, переработанный вариант поступил 24.12.2024.
Дата опубликования 08.01.2025.

===== BIOINFORMATICS =====

Typological approaches to recognizing genus and subgenus of coronaviruses by structural and non-structural genes

Chaley M.B.¹, Kutyrkin V.A.²

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

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

Abstract. Owing to rapid growth of data on viral genomes in the result of metagenomic researches, bioinformatics and virology are increasingly interacting. There is even the term viral informatics, implying the existence of a whole complex of the databases, knowledge databases about the viruses and software tools for working with them. Among the problems of bioinformatics in virology, it was earlier pointed out to annotation of viral genomes. In the present work on the example of recognizing of subgenus and genus of the coronaviruses a fairly simple and effective typological approach to virus annotation is proposed which uses frequency characteristics of the codons in individual genes. Typological approach is characterized by averaging known data, in particular, such codon frequency characteristics, to determine the similarity or resemblance with them of analogical characteristics for object under consideration. Recognition of subgenus and genus is based on statistics that reveals deviation of coronavirus gene considered from corresponding gene of viral genome with known genus or subgenus. The work compares recognition based on structural genes encoding virion proteins (nucleocapsid protein N and spike protein S) and genes of non-structural proteins combined into a single reading frame ORF1ab. Four typological approaches were discussed in the article. In the first two averaging of all available data and data on prototypical strains only was done over the genera. In the third approach original

data on prototype strains were averaged over the subgenera. The fourth approach was based on individual frequency characteristics of prototype strains of the subgenera. Three of the four typological approaches revealed high efficiency in recognizing genus and subgenus of the coronaviruses while using N-gene. The fourth approach proved to be the most effective for identifying genus and subgenus of the coronaviruses. In addition, it has made it possible to reduce the number of codons considered in N-gene of the coronaviruses and to increase recognition efficiency to almost 100 %.

Key words: *recognizing genus and subgenus of coronaviruses, the genome of coronavirus, prototype strains of the coronaviruses, S-, N- genes of coronavirus, ORF1ab of coronavirus.*