Математическая биология и биоинформатика
2019. Т. 14. № 2. С. 533–542. doi: 10.17537/2019.14.533

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

Распознавание видов флавивирусов на основе кодирующих последовательностей полипротеинов

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

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

Аннотация. Предлагается метод распознавания вида флавивируса, включая распознавание подтипа, по последовательности его генома. Метод основан на использовании частотных характеристик кодонов аминокислот в полных кодирующих последовательностях полипротеинов. Высокая надежность этого метода подтверждается при распознавании 15 групп геномов флавивирусов различных видов и подтипов, имеющих достаточное количество представителей в GenBank. Рассматриваются десять различных видов флавивирусов, четыре подтипа вируса лихорадки денге и вирус Кунджин, как подтип вируса лихорадки Западного Нила.

Ключевые слова: *геном флавивируса, скрытая профильная триплетная периодичность, частоты кодонов аминокислот, распознавание вида флавируса.*

ВВЕДЕНИЕ

Флавивирусы – род арбовирусов из семейства *Flaviridae*, которые распространяются при укусах комаров или клещей. Флавивирусы могут вызывать тяжелые заболевания человека, домашних и диких животных и птиц [1–5]. Они поражают нервную систему, вызывая парезы, параличи и энцефалиты и, кроме того, часто проявляются симптомами геморрагических лихорадок. Наиболее известны сезонный клещевой энцефалит [6], лихорадка денге [7], желтая лихорадка [8], лихорадка Западного Нила [9] и др. Последнее время в связи с развитием туризма, миграционными процессами и активизацией деловых связей возникла проблема завозных случаев вирусных заболеваний из тропических и субтропических стран в неэндемичные районы, в том числе и в Россию [10, 11].

Для идентификации вида или подтипа флавивируса, вызвавшего заболевание, наряду с традиционными методами иммуноферментного анализа применяется метод обратной транскрипции и полимеразной цепной реакции с дальнейшей процедурой секвенирования cDNA генома вируса [11]. Полученную в результате нуклеотидную последовательность сравнивают с известными геномными последовательностями вирусов в GenBank [12]. Однако, такое сравнение требует дополнительного выравнивания последовательностей для выявления сходства. При сравнении и изучении эволюции геномов вирусов также используются частотные характеристики их нуклеотидного, динуклеотидного состава, предпочтительного использования кодонов [13–15].

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

Геном флавивируса представлен одноцепочечной РНК положительной полярности длиной около 11 тыс. нуклеотидов. РНК кодирует три структурных (C, prM/M, E) и семь неструктурных (NS1, NS2A, NS2B, NS3, NS4A, NS4B, NS5) белков [16]. Последовательно расположенные друг за другом и считываемые в единой рамке считывания, гены вирусных белков образуют полную кодирующую последовательность (CDS) полипротеина.

Ранее, в работе [17] исследовались CDS полипротеинов для 13 видов флавивирусов. Было показано, что во всех представителях CDS полипротеинов этих видов распознавалась скрытая профильная триплетная периодичность. Всего было проанализировано 62 представителя, т.е. на каждый исследуемый вид флавивируса в среднем приходилось по пять CDS полипротеинов. В настоящей работе исследовались 15 групп CDS полипротеинов флавивирусов с суммарным количеством представителей 7060. Рассматривались только те виды флавивирусов или отдельные группы подтипов, для которых в базе GenBank [12] находилось не менее 14 CDS полипротеинов. Анализировались полные кодирующие последовательности десяти видов флавивирусов, таких, как вирус Багаза (Bagaza), вирус японского энцефалита (Japanese encephalitis), вирус энцефалита долины Мюррея (Murray Valley encephalitis), вирус Повассан (Powassan), вирус энцефалита Сент-Луис (Saint Louis encephalitis), утиный вирус Тембусу (Duck Tembusu), вирус клещевого энцефалита (Tick-borne encephalitis), вирус Усуту (Usutu), вирусы лихорадки Западного Нила и желтой лихорадки. Кроме того, отдельно анализировались группы CDS полипротеинов для четырех подтипов вируса лихорадки денге (Dengue serotype 1, Dengue serotype 2, Dengue serotype 3, Dengue serotype 4) и вируса Кунджин (Kunjin), который в настоящий момент считается подтипом вируса лихорадки Западного Нила, распространенным в Австралии. Во всех анализируемых CDS полипротеинов флавивирусов распознавалась скрытая профильная триплетная периодичность.

Ранее в работе [18] была предложена стохастическая модель организации кодирования в последовательностях ДНК (текстовых строках), которая объясняет проявление скрытой профильной триплетной периодичности. Эта модель имеет вид случайной строки, полученной на основе последовательных и независимых реализаций случайного кодона размера три. В настоящей работе такой случайный кодон является случайной величиной, имеющей в качестве значений 61 кодон аминокислот и три кодона терминации. Каждое из 64 значений случайного кодона характеризуется соответствующей вероятностью реализации этого значения, так что сумма этих вероятностей равна единице. Таким образом, случайный кодон определяется вероятностным распределением его значений. Стохастическая модель на основе случайного кодона получила название SHOC-модели (Stochastic Homogeneous Organization of Coding). Фактически эта модель эквивалентна полиномиальной схеме с 64-мя исходами из заданного числа независимых испытаний, в каждом из которых реализуется один из кодонов генетического кода. Формально SHOC-модель можно описать случайной строкой $\underbrace{Cdn\,Cdn\ldots Cdn}_{m\ раз}$ с длиной $n=3m$, где символ Cdn обозначает упомянутый выше случайный кодон. Этот кодон однозначно характеризуется своим вероятностным распределением $P=(p_1, p_2, ..., p_{64})$ 64-х значений, где p_i – вероятность реализации *i*-го кодона из списка кодонов генетического кода. С точки зрения SHOC-модели CDS полипротеина рассматривается как реализация случайной строки, представляющей эту модель. Следовательно, CDS полипротеина определяет выборочное распределение $\hat{P}=(\hat{p}_1, \hat{p}_2, ..., \hat{p}_{64})$ кодонов генетического кода, являющееся оценкой вероятностного распределения P случайного кодона Cdn.

Математическая биология и биоинформатика. 2019. Т. 14. № 2. doi: 10.17537/2019.14.533

Таким образом, все CDS полипротеинов флавивирусов, анализируемых в работе 15 групп, индуцируют множество соответствующих выборочных вероятностных распределений, образующих множество точек в 64-мерном пространстве. В настоящей работе рассматривается задача кластеризации этого множества точек, где отдельный кластер характеризуется только одним из рассматриваемых видов или подтипов флавирусов. В работе предложен алгоритм, которой позволяет по выборочному вероятностному распределению кодонов CDS полипротеина однозначно отнести её к кластеру соответствующего вида иди подтипа флавивируса. Следует отметить, что предложенный алгоритм позволяет корректно решить следующую задачу. При фиксированном виде или подтипе флавивируса для рассматриваемой конкретной CDS полипротеина можно корректно определить принадлежит ли ему эта CDS.

Разработанный в настоящей работе алгоритм по распознаванию вида и подтипа флавивирусов основывается на предварительной обучающей процедуре. Каждая выборка CDS полипротеинов, относящаяся к флавивирусам одной группы, делилась на две численно равные выборки, одна из которых служила обучающей, а вторая рассматривалась как тестируемая. Предложенный в работе алгоритм распознавания вида флавивируса создан на основе этой обучающей процедуры. Этот алгоритм показал практически 100 % эффективность. Работа этого алгоритма будет продемонстрирована в работе на основе полной выборки.

МАТЕРИАЛЫ И МЕТОДЫ

В настоящей работе для анализа были выбраны геномы флавивирусов из GenBank выпуска 231 от 15 апреля 2019 г. Все отобранные геномы прошли дополнительную фильтрацию, чтобы исключить последовательности, секвенированные неполностью или с большим числом не идентифицированных нуклеотидов. Ранее, в работе [17] гены структурных белков флавивирусов рассматривались вместе с CDS полипротеинов. В настоящей работе мы ограничились анализом только CDS полипротеинов, которые имеют длину около 10^4 нуклеотидов. Следует отметить, что интроны в геномах флавивирусов практически не встречаются, так, что CDS полипротеина флавивируса представлена непрерывной кодирующей последовательностью. Таблица 1 отражает количественный и качественный состав анализируемых в работе CDS полипротеинов. В рассматриваемой версии GenBank содержались геномы порядка сотни различных видов флавивирусов, из которых нами были отобраны геномы для десяти видов и пяти подтипов вирусов, имеющих наибольшее количество представителей в GenBank.

Для выявления структурно-статистических свойств последовательностей ДНК ранее был предложен спектрально-статистический подход [19]. Согласно этому подходу, последовательность ДНК рассматривается как реализация некоторой случайной строки. В рамках этого подхода были предложены модели случайных строк, в реализациях которых выявляется скрытая профильная периодичность [18]. Кроме того, этот подход позволяет ввести понятие профильной эквивалентности для широкого класса случайных строк. В реализациях профильно-эквивалентных строк выявляется один и тот же тип скрытой профильной периодичности. В частности, было показано, что в кодирующих районах геномов различных прокариот и эукариот в большинстве случаев распознается скрытая профильная триплетная периодичность [19]. В работе в качестве наиболее общей модели, объясняющей наличие такой скрытой периодичности, была предложена, так называемая SHOC-модель [18]. Как отмечалось во введении, эту модель однозначно характеризует случайный кодон, являющийся случайной величиной со значениями в текстовых строках из трех букв заданного алфавита. Для кодирующих районов ДНК в качестве значений такого случайного кодона выступают триплеты генетического кода. Следовательно, в рамках SHOC-модели для кодирующей последовательности ДНК, где наблюдается скрытая

триплетная периодичность, можно получить оценку вероятностного распределения случайного кодона в виде набора (вектора) частот кодонов генетического кода, встречающихся в этой последовательности. Поэтому такие последовательности ДНК можно классифицировать по типам распределений случайных кодонов.

Проведенный анализ полной выборки CDS полипротеинов показал, что в них распознается скрытая профильная триплетная периодичность. Поэтому в настоящей работе предлагается процедура распознавания видов флавивирусов на основе анализа распределений частот кодонов аминокислот в CDS их полипротеинов.

В работе рассматривались 15 групп CDS полипротеинов флавивирусов. Виды и подтипы флавивирусов и количество CDS полипротеинов N_j (в полной и обучающей выборках) каждой группы представлены в таблице 1, где за каждой группой закрепляется её номер j ($j = \overline{1,15}$).

Таблица 1. Количественный состав 15 групп анализируемых CDS полипротеинов флавивирусов

j	Виды и подтипы флавивирусов	Полная выборка N_j	Обучающая выборка N_j
1	Bagaza virus	18	9
2	Dengue serotype 1 virus	1895	947
3	Dengue serotype 2 virus	1412	706
4	Dengue serotype 3 virus	922	461
5	Dengue serotype 4 virus	190	95
6	Japanese encephalitis virus	300	150
7	Kunjin virus	43	21
8	Murray Valley encephalitis virus	14	7
9	Powassan virus	21	10
10	Saint Louis encephalitis virus	38	19
11	Duck Tembusu virus	110	55
12	Tick-borne encephalitis virus	175	87
13	Usutu virus	147	73
14	West Nile virus	1652	826
15	Yellow fever virus	123	61

Пусть $n = \overline{1, N_i}$ – порядковый номер гена полипротеина в *i*-том типе флавивируса. Тогда $\boldsymbol{P}^{(i)}(n)$ – вектор распределения частот кодонов аминокислот этого гена имеет вид:

$$\boldsymbol{P}^{(i)}(n) = (p_1^{(i)}(n), p_2^{(i)}(n), ..., p_k^{(i)}(n), ..., p_{64}^{(i)}(n)). \quad (1)$$

Обозначим $e_k^{(j)}$ среднюю частоту *k*-го кодона ($k = \overline{1,64}$) в *j*-том типе флавивируса, т.е.

$$e_k^{(j)} = \frac{1}{N_j} \sum_{n=1}^{N_j} p_k^{(j)}(n). \quad (2)$$

Тогда, согласно формулам (2), вектор средних частот кодонов вирусов *j*-го типа примет вид:

$$\boldsymbol{E}^{(j)} = (e_1^{(j)}, e_2^{(j)}, ..., e_k^{(j)}, ..., e_{64}^{(j)}). \quad (3)$$

На рисунке 1 показаны графические иллюстрации векторов средних частот кодонов аминокислот для полной выборки 15 групп (10 видов и пять подтипов), анализируемых

в работе флавивирусов. Рисунок 1 отражает качественное сходство векторов средних частот кодонов CDS полипротеинов рассматриваемых групп вирусов.

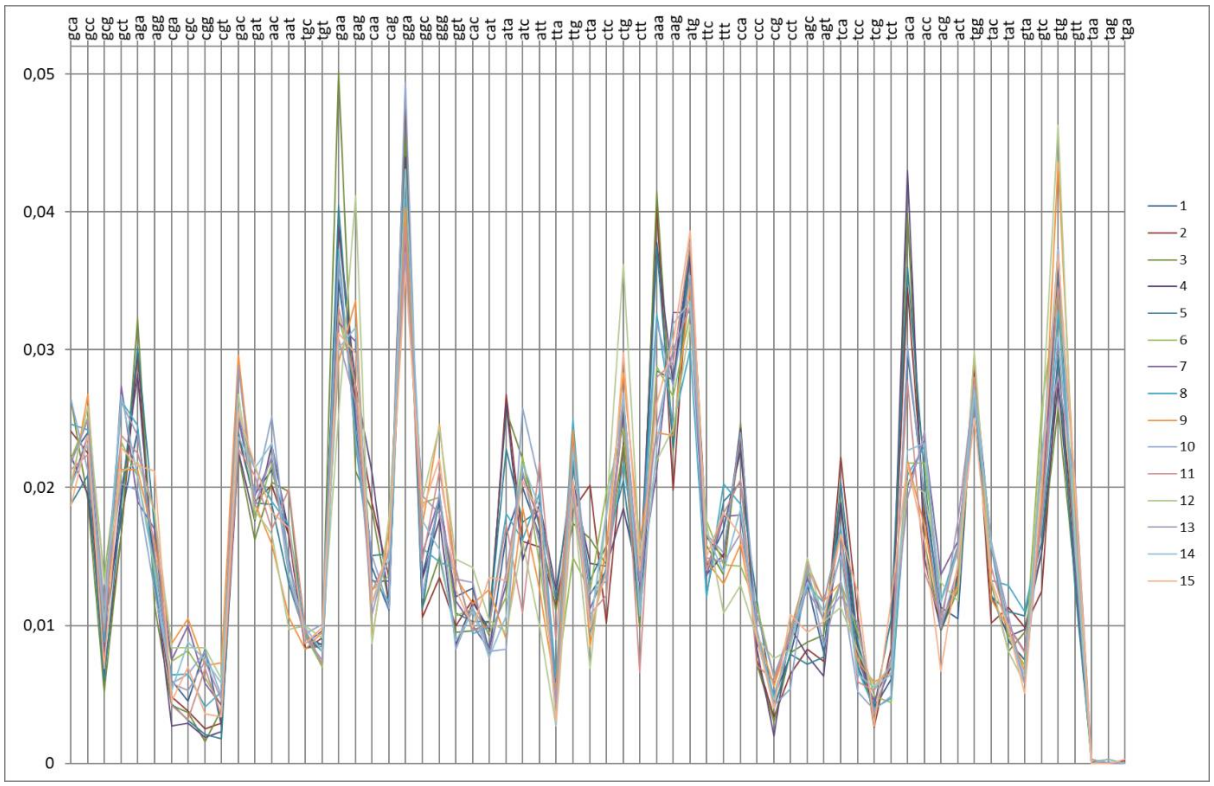

Рис. 1. Распределения средних частот кодонов аминокислот для полной выборки 15-ти групп CDS полипротеинов флавивирусов.

На рисунках 2 и 3 показаны средние частоты синонимичных кодонов аминокислот (по полной выборке) аланина и аргинина для анализируемых в работе 15 групп флавивирусов. Из рисунков следует, что для надежного решения задачи распознавания среди анализируемых 15 групп вирусов, такой информации недостаточно.

В работе для целей распознавания вводятся величины $r_{(j)}^{(i)}(n) = r(\boldsymbol{P}^{(i)}(n), \boldsymbol{E}^{(j)})$ отклонения («расстояния») вектора частот кодонов *n*-й CDS полипротеина вируса *i*-го вида (или подтипа) от вектора средних частот кодонов в CDS вируса *j*-го вида (или подтипа). Для расчета этого отклонения используется формула:

$$r_{(j)}^{(i)}(n) = r(\boldsymbol{P}^{(i)}(n), \boldsymbol{E}^{(j)}) = \gamma \sum_{k=1}^{64} \frac{| p_k^{(i)}(n) - e_k^{(j)} |}{e_k^{(j)}}, \qquad (4)$$

где для удобства анализа результатов вычислений выбрано значение $\gamma = \frac{1}{7}$.

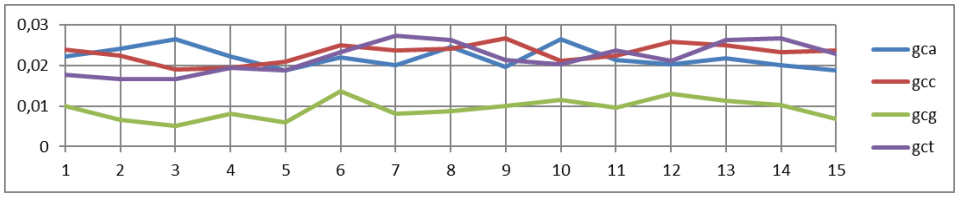

Рис. 2. Средние частоты синонимичных кодонов аланина для анализируемых CDS полипротеинов 15 групп флавивирусов (полная выборка).

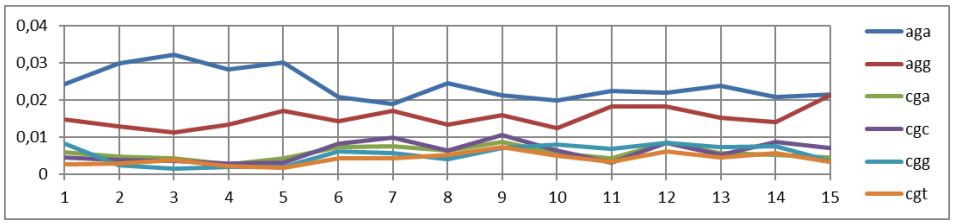

Рис. 3. Средние частоты синонимичных кодонов аргинина для анализируемых CDS полипротеинов 15 групп флавивирусов (полная выборка).

Для демонстрации принципа распознавания вида (подтипа) вируса по вектору распределения частот кодонов в CDS полипротеина в работе вводится величина $r^{(i)}_{(j)}$, представляющая среднее отклонение векторов частот кодонов в CDS вирусов *i*-го вида (подтипа) от вектора средних частот кодонов в вирусах *j*-го вида (подтипа):

$$r^{(i)}_{(j)} = \frac{1}{N_i}\sum_{n=1}^{N_i} r^{(i)}_{(j)}(n). \qquad (5)$$

Матрица $R = (r^{(i)}_{(j)})^{15}_{15}$ таких величин (верхний индекс – строка, нижний – столбец) для обучающей выборки представлена в таблице 2, где диагональная компонента $r^{(i)}_{(i)}$ является минимальной в *i*-ой строке для $i = \overline{1,15}$. Аналогичная матрица для полной выборки, показанная в таблице 3, обладает тем же свойством.

Таблица 2. Средние отклонения векторов частот кодонов в CDS полипротеинов вирусов *i*-го вида (подтипа) $(i = \overline{1,15})$ от вектора средних частот кодонов в вирусах *j*-го вида (подтипа) $(j = \overline{1,15})$ для обучающей выборки

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	0.463	2.403	2.734	2.279	2.030	1.658	1.270	1.407	1.902	1.499	1.458	2.452	1.510	1.681	2.157
2	2.015	0.708	1.548	1.591	1.528	2.849	2.681	2.434	3.362	2.698	2.494	4.084	2.892	3.165	2.803
3	2.027	1.583	0.781	1.499	1.668	2.674	2.612	2.351	3.332	2.683	2.381	4.027	2.906	2.960	2.998
4	1.844	1.845	1.763	0.474	1.333	2.814	2.398	2.192	3.102	2.734	2.143	3.928	2.771	3.110	3.014
5	1.799	1.817	1.874	1.440	0.469	2.632	2.171	1.982	2.789	2.707	2.021	3.584	2.491	2.922	2.642
6	1.816	3.350	3.338	3.456	3.081	0.543	1.444	1.880	1.791	2.088	2.347	2.023	1.374	1.400	2.327
7	1.428	3.180	3.282	2.947	2.563	1.368	0.271	1.404	1.525	1.887	1.846	2.090	1.347	1.274	1.838
8	1.622	2.867	2.999	2.735	2.344	1.805	1.390	0.609	1.961	2.103	1.738	2.617	1.680	1.929	2.231
9	2.199	4.063	4.351	3.999	3.579	1.860	1.648	2.116	0.616	2.325	2.472	1.199	1.722	1.743	2.265
10	1.424	2.833	2.973	2.928	2.779	1.716	1.698	1.839	2.065	0.726	2.119	2.636	1.789	1.633	2.380
11	1.492	2.997	3.129	2.569	2.310	2.002	1.477	1.583	1.943	2.222	0.286	2.718	1.751	1.995	2.136
12	2.464	4.355	4.682	4.392	3.908	1.936	2.031	2.550	1.214	2.699	2.699	0.630	1.842	1.958	2.652
13	1.645	3.320	3.601	3.196	2.815	1.274	1.306	1.684	1.534	1.974	1.946	1.792	0.179	1.352	1.859
14	1.737	3.354	3.317	3.353	3.101	1.261	1.176	1.715	1.570	1.747	2.032	1.827	1.309	0.242	1.968
15	2.226	2.822	3.151	3.047	2.610	2.185	1.705	2.213	1.964	2.571	2.190	2.369	1.763	2.022	0.645

Из указанных выше свойств таблицы 2 для распознавания вида (подтипа) вируса по гену полипротеина в работе предлагается следующий принцип. Если $\boldsymbol{P} = (p_1, p_2, ..., p_{64})$ – вектор частот кодонов аминокислот анализируемой CDS полипротеина, то номер *m* распознаваемого вида (подтипа) вируса удовлетворяет условию:

$$r(\boldsymbol{P}, \boldsymbol{E}^{(m)}) = \min\{r(\boldsymbol{P}, \boldsymbol{E}^{(j)}) : j = \overline{1,15}\}, \qquad (6)$$

где, согласно формуле (4):

$$r(\boldsymbol{P}, \boldsymbol{E}^{(j)}) = \gamma \sum_{k=1}^{64} \frac{| p_k - e_k^{(j)} |}{e_k^{(j)}} \text{ для } j = \overline{1,15}. \quad (7)$$

Следовательно, для анализируемой CDS полипротеина выбирается тот вид (или подтип) вируса, на котором достигается минимальное отклонение вектора частот кодонов аминокислот CDS полипротеина от среднего вектора частот кодонов аминокислот этого вида (подтипа). Тестирование этого принципа для обучающей выборки показало его практически 100 % эффективность. Поэтому этот же принцип использовался для распознавания по полной выборки.

Таблица 3. Средние отклонения векторов частот кодонов в CDS полипротеинов вирусов *i*-го вида (подтипа) $(i = \overline{1,15})$ от вектора средних частот кодонов в вирусах *j*-го вида (подтипа) $(j = \overline{1,15})$ для полной выборки

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	0.374	2.442	2.738	2.294	2.052	1.654	1.288	1.422	1.926	1.436	1.458	2.451	1.535	1.672	2.177
2	1.973	0.692	1.532	1.563	1.483	2.811	2.655	2.401	3.456	2.670	2.450	4.046	2.885	3.162	2.809
3	2.018	1.580	0.817	1.525	1.653	2.659	2.628	2.356	3.465	2.673	2.384	4.041	2.943	2.982	3.029
4	1.817	1.833	1.708	0.437	1.319	2.768	2.381	2.171	3.236	2.653	2.118	3.909	2.765	3.085	3.011
5	1.769	1.836	1.892	1.419	0.426	2.618	2.163	1.977	2.919	2.634	2.009	3.557	2.509	2.913	2.647
6	1.794	3.351	3.377	3.483	3.069	0.521	1.444	1.818	1.796	2.023	2.338	1.998	1.380	1.380	2.307
7	1.370	3.184	3.275	2.970	2.553	1.346	0.238	1.337	1.588	1.828	1.828	2.099	1.354	1.282	1.822
8	1.548	2.831	2.951	2.698	2.297	1.807	1.394	0.514	2.043	2.021	1.744	2.629	1.706	1.912	2.212
9	2.124	4.005	4.218	3.953	3.519	1.814	1.591	1.990	0.462	2.268	2.389	1.201	1.714	1.734	2.197
10	1.406	2.810	2.913	2.939	2.830	1.702	1.739	1.837	2.104	0.699	2.148	2.633	1.812	1.592	2.381
11	1.452	2.995	3.115	2.586	2.310	2.009	1.502	1.529	2.094	2.158	0.283	2.708	1.792	2.021	2.127
12	2.446	4.365	4.657	4.428	3.912	1.945	2.036	2.499	1.193	2.696	2.704	0.634	1.880	1.982	2.620
13	1.610	3.328	3.586	3.222	2.823	1.268	1.319	1.646	1.549	1.960	1.929	1.781	0.188	1.357	1.809
14	1.702	3.352	3.331	3.376	3.108	1.250	1.166	1.684	1.567	1.732	2.020	1.833	1.321	0.228	1.977
15	2.179	2.793	3.093	2.972	2.556	2.178	1.708	2.162	2.005	2.550	2.180	2.400	1.806	2.029	0.608

Таблица 4. Значения максимального отклонения $M^{(j)}$ среди векторов частот кодонов аминокислот для CDS полипротеинов вирусов одного вида (подтипа) от соответствующего им вектора средних частот по каждому из рассматриваемых в работе виду или подтипу (15 групп вирусов)

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0.558	1.261	1.714	0.882	0.785	1.529	0.429	0.768	0.552	1.241	0.941	0.881	0.946	1.655	1.402

Кроме того, для подтверждения правильности выбранного вида (подтипа) вируса в работе используются величины:

$$M^{(j)} = \max\{r_{(j)}^{(j)}(n) : n = \overline{1, N_j}\},\ j = \overline{1,15}, \quad (8)$$

где $M^{(j)}$ – максимальное отклонение («расстояние») среди векторов частот кодонов для CDS полипротеинов вирусов *j*-го типа от вектора средних частот кодонов вируса этого вида (подтипа). Значения величин $M^{(j)}$ $(j = \overline{1,15})$ для полной выборки приведены в таблице 4.

Указанный выше принцип, согласно величинам из формулы (8), будет дополнен следующим правилом. Если, согласно принципу, для CDS полипротеина с вектором частот кодонов аминокислот $\boldsymbol{P}$ выбран *j*-й вид (подтип) вируса, то этот выбор остается в силе, при выполнении условия (см. формулу (7)):

$$r(\boldsymbol{P}, \boldsymbol{E}^{(j)}) \le M^{(j)}. \quad (9)$$

Если условие (9) нарушено, то считаем, что вид (или подтип) вируса для анализируемой CDS полипротеина не определен.

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Применение предложенного метода распознавания вида или подтипа флавивируса по анализируемому вектору частот кодонов аминокислот в CDS полипротеина показало следующий результат. Для 12 групп вирусов (см. Табл. 1), а именно: Bagaza (№ 1), Dengue подтип 1 (№ 2), Dengue подтип 4 (№ 5), Kunjin (№ 7), Murray Valley encephalitis (№ 8), Powassan (№ 9), Saint Louis encephalitis (№ 10), Duck Tembusu (№ 11), Tick-borne encephalitis (№ 12), Usutu (№ 13), Yellow fever (№ 15), ошибок в распознавании вируса не было.

Таблица 5. Коды доступа GenBank для CDS полипротеинов с неопределенным видом флавивируса

№	GenBank код доступа	Номер группы вируса	Предлагаемый номер группы вируса j	$r(\boldsymbol{P},\boldsymbol{E}^{(j)})$	$M^{(j)}$
1	EF105389	3	4	1.352	0.882
2	EF105390	3	4	1.378	0.882
3	EF105382	3	4	1.317	0.882
4	EF105381	3	4	1.297	0.882
5	EF105380	3	4	1.303	0.882
6	EF105383	3	4	1.266	0.882
7	EF105386	3	4	1.345	0.882
8	EF105385	3	4	1.350	0.882
9	EF457904	3	4	1.343	0.882
10	EF105384	3	4	1.404	0.882
11	EF105387	3	4	1.465	0.882
12	EU003591	3	4	1.413	0.882
13	EF105388	3	4	1.442	0.882
14	EF105378	3	4	1.346	0.882
15	FJ467493	3	4	1.571	0.882
16	KX274130	3	2	1.572	1.261
17	KM677246	6	8	1.067	0.768
18	HM596272	6	8	1.034	0.768
19	JF915894	6	7	1.187	0.429
20	KY703856	14	7	1.260	0.429
21	KY703855	14	13	1.214	0.946
22	GQ851604	14	13	1.241	0.946
23	GQ851605	14	13	1.223	0.946

При распознавании остальных трех групп флавивирусов ошибки составляли менее 0.5 % от общего количества CDS полипротеинов этих типов. При этом было всего два случая неправильного определения вида вируса. CDS полипротеина вируса West Nile (GenBank код доступа JN887352) была признана за CDS вируса Kunjin, и CDS полипротеина вируса West Nile (GenBank код доступа FJ159130) была признан за CDS вируса Japanese encephalitis. При распознавании CDS полипротеинов вируса Dengue подтипа 2 вид вируса не был определен в 16 случаях. Аналогично вид вируса не был определен для трех CDS полипротеинов вируса Japanese encephalitis и четырех CDS полипротеинов вируса West Nile. В Таблице 5 приведены коды доступа GenBank для CDS полипротеинов, вид вируса которых не был определен в силу того, что условие (9)

не выполняется. Соответствие номера группы вируса его названию такое же, как в таблице 1.

ЗАКЛЮЧЕНИЕ

В работе предложен метод, который по кодирующей последовательности полипротеина флавивируса определяет его вид. Этот метод протестирован в распознавании 10 видов и пяти подтипов флавивирусов, геномы которых на сегодняшний день имеют значительное представительство в GenBank. Показано, что для 10 видов и двух подтипов флавивирусов метод показал 100 % надежности. Ошибки распознавания вида вируса не превысили 0.5 %.

В общем, применение предлагаемого метода не ограничивается только идентификацией видов РНК-содержащих флавивирусов. Возможно его использование и для подтверждения вида вирусов других родов и семейств, безотносительно того, представлен геном вируса последовательностью ДНК или РНК.

Поскольку профильная триплетная периодичность является неотъемлемой характеристикой любой достаточно длинной CDS, частотные характеристики кодонов аминокислот, на которые опирается предлагаемый метод, для CDS, выбранной в качестве идентификатора, также могут использоваться с целью определения или подтверждения вида вируса.

В целом, предлагаемый метод упрощает процедуру идентификации вида вируса по последовательности его секвенированного генома, так как для применения метода не требуется знания о наличии и расположении консервативных участков вирусного генома, которые существенны при сравнении анализируемой последовательности с известными нуклеотидными последовательностями генотипов вирусов, хранящихся в международных базах данных.

СПИСОК ЛИТЕРАТУРЫ

1. Fernández-Pinero J., Davidson I., Elizalde M., Perk S., Khinich Y., Jiménez-Clavero M.A. Bagaza virus and Israel turkey meningoencephalomyelitis virus are a single virus species. *Journal of General Virology*. 2014. V. 95. P. 883–887. doi: 10.1099/vir.0.061465-0.
2. Zhang W., Chen S., Mahalingam S., Wang M., Cheng A. An updated review of avian-origin Tembusu virus: a newly emerging avian Flavivirus. *Journal of General Virology*. 2017. V. 98. P. 2413–2420. doi: 10.1099/jgv.0.000908.
3. Benzarti E., Linden A., Desmecht D., Garigliany M. Mosquito-borne epornitic flaviviruses: an update and review. *Journal of General Virology.* 2019. V. 100. P. 119–132. doi: 10.1099/jgv.0.001203.
4. Diaz A., Coffey L.L., Burkett-Cadena N., Day J.F. Reemergence of St. Louis encephalitis virus in the Americas. *Emerging Infectious Diseases*. 2018. V. 24. P. 2150–2157. doi: 10.3201/eid2412.180372.
5. Clé M., Beck C., Salinas S., Lecollinet S., Gutierrez S., Van de Perre P., Baldet T., Foulongne V., Simonin Y. Usutu virus: A new threat? *Epidemiology and Infection.* 2019. V. 147. Article No. e232. doi: 10.1017/S0950268819001213.
6. Ruzek D., Avšič Županc T., Borde J., Chrdle A., Eyer L., Karganova G., Kholodilov I., Knap N., Kozlovskaya L., Matveev A., Miller A.D., Osolodkin D.I., Överby A.K., Tikunova N., Tkachev S., Zajkowska J. Tick-borne encephalitis in Europe and Russia: Review of pathogenesis, clinical features, therapy, and vaccines. *Antiviral Research.* 2019. V. 164. P. 23–51. doi: 10.1016/j.antiviral.2019.01.014.
7. Khetarpal N., Khanna I. Dengue Fever: Causes, Complications, and Vaccine Strategies. *Journal of Immunology Research.* 2016. Article ID 6803098. doi: 10.1155/2016/6803098.

8. Kleinert R.D.V., Montoya-Diaz E., Khera T., Welsch K., Tegtmeyer B., Hoehl S., Ciesek S., Brown R.J.P. Yellow fever: Integrating current knowledge with technological innovations to identify strategies for controlling a re-emerging virus. *Viruses*. 2019. V. 11. Article No. 960. doi: 10.3390/v11100960.
9. Petersen L.R., Brault A.C., Nasci R.S. West Nile virus: review of the literature. *JAMA*. 2013. V. 310. P. 308–315. doi: 10.1001/jama.2013.8042.
10. Путинцева Е.В., Смелянский В.П., Бородай Н.В., Алексейчик И.О., Шахов Л.О., Ткаченко Г.А., Шпак И.М., Казорина Е.В., Викторов Д.В., Топорков А.В. Лихорадка Западного Нила в 2016 г. в мире и на территории Российской Федерации, прогноз развития ситуации в 2017 г. *Проблемы особо опасных инфекций*. 2017. № 1. С. 29–36. doi: 10.21055/0370-1069-2017-1-29-36.
11. *Бюллетень нормативных и методических документов госсанэпиднадзора*. 2016. Т. 3. № 65. С. 25–38.
12. Benson D.A., Cavanaugh M., Clark K., Karsch-Mizrachi I., Lipman D.J., Ostell J., Sayers E.W. GenBank. *Nucleic Acids Res*. 2013. V. 41(Database issue). P. D36–D42. doi: 10.1093/nar/gks1195.
13. Jenkins G., Pagel M., Gould E., de A. Zanotto P.M., Holmes E.C. Evolution of base composition and codon usage bias in the genus Flavivirus. *Journal of Molecular Evolution*. 2001. V. 52. P. 383–390. doi: 10.1007/s002390010168.
14. Belalov I.S., Lukashev A.N. Causes and implications of codon usage bias in RNA viruses. *PLoS One*. 2013. V. 8. № 2. Article No. e56642. doi: 10.1371/journal.pone.0056642.
15. Di Giallonardo F., Schlub T.E., Shi M., Holmes E.C. Dinucleotide composition in animal RNA viruses is shaped more by virus family than by host species. *Journal of Virology*. 2017. V. 91. Article No. e02381-16. doi: 10.1128/jvi.02381-16.
16. *Руководство по вирусологии: Вирусы и вирусные инфекции человека и животных*. Под. ред. Львова Д.К. М.: Медицинское информационное агентство, 2013.
17. Тюлько Ж.С., Кутыркин В.А., Чалей М.Б. Структурно-статистические свойства геномов флавивирусов. *Математическая биология и биоинформатика*. 2017. Т. 12. № 2. С. 343–353. doi: 10.17537/2017.12.343.
18. Chaley M., Kutyrkin V. Stochastic model of homogeneous coding and latent periodicity in DNA sequences. *Journal of Theoretical Biology*. 2016. V. 390. P. 106–116. doi: 10.1016/j.jtbi.2015.11.014.
19. Chaley M., Kutyrkin V. Spectral-statistical approach for revealing latent regular structures in DNA sequences. In: *Data Mining Techniques for the Life Sciences*. Eds. Carugo O., Eisenhaber F. New York: Springer Science+Business Media, 2016. P. 315–340. doi: 10.1007/978-1-4939-3572-7.

Рукопись поступила в редакцию 28.08.2019, переработанный вариант поступил 13.11.2019.
Дата опубликования 19.11 .2019.