***Доклады Международной конференции "Математическая биология и биоинформатика"**. Под ред. В.Д. Лахно. Том 7. Пущино: ИМПБ РАН, 2018. Статья № e53. doi: 10.17537/icmbb18.10*

Особенности кодирующих последовательностей геномов флавивирусов

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

[1]*ИМПБ РАН– филиал ИПМ им. М.В. Келдыша РАН, Пущино, Россия*
[2]*ОмГМУ Минздрава России, Омск, Россия*
[3]*МГТУ им. Н.Э. Баумана, Москва, Россия*

maramaria@yandex.ru

Структурно-статистические свойства одноцепочечных РНК-геномов флавивирусов исследуются с помощью спектрально-статистического подхода. Полноразмерные последовательности, кодирующие полипротеины, рассматриваются вместе с сегментами, кодирующими вирусные белки. Отдельно анализируются две группы: геномы вирусов патогенных для человека и теплокровных животных, и геномы вирусов специфичных только для комаров. В целом, все анализируемые кодирующие последовательности геномов флавивирусов проявляют свойства 3-регулярности и скрытой триплетной профильной периодичности, аналогично кодирующим районам геномов прокариот и эукариот. Однако в сегментах, кодирующих вирусные белки, полностью отсутствует двухуровневая организация кодирования. Кроме того, для значительной части этих сегментов выявляется однородность их последовательностей. Причем среди геномов вирусов специфичных для комаров свойство однородности выражено сильнее. Такие особенности кодирующих последовательностей геномов флавивирусов объясняются высокой скоростью их мутации и простой структурой. В работе приводятся примеры того, как вставки и делеции отдельных нуклеотидов могут приводить к однородности последовательности, в которой исходно определялась скрытая триплетная периодичность.

Ключевые слова: *геном флавивирусов, спектрально-статистический подход, скрытая профильная периодичность, свойство 3-регулярности.*

Specifics of Coding Sequences in the Flavivirus Genomes

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

[1]*IMPB RAS – Branch of KIAM RAS, Pushchino, Moscow Region, Russia*
[2]*Omsk State Medical University of Ministry of Healthcare of the Russian Federation, Omsk, Russia*
[3]*Moscow State Technical University n. a. N.E. Bauman, Moscow, Russia*

Structural-statistical properties of the flavivirus single-strand RNA genomes are investigated with the help of spectral-statistical approach. Full-length sequences, encoding the polyproteins, are considered along with the segments coding virus proteins. Two groups are analyzed separately: the genomes of viruses being pathogenic for human and warm-blooded animals and the genomes of viruses that are mosquitos specific only. In general, all analyzed coding sequences of the flavivirus genomes display the properties of 3-regularity and latent triplet profile periodicity analogous to the coding regions of prokaryotic and eukaryotic genomes. However, in the segments coding virus proteins two-level organization of encoding is absolutely missing. Moreover, sequence homogeneity is revealed in significant part of these segments. At that, the property of homogeneity is revealed more frequently among the genomes of viruses which are mosquitos specific. Such particularities of the coding sequences in flavivirus genomes are explained by their simple structure and high rate of the mutations. The examples demonstrating as indels of few nucleotides may induce homogeneity of sequence, where latent triplet profile periodicity has been originally recognized, are given in the work.

Key words: flavivirus genome, spectral-statistical approach, latent profile periodicity, property of 3-regularity.

1. Введение

Применение спектрально-статистического подхода (2S-подхода) к исследованию структурно-статистических свойств кодирующих районов (CDSs) ДНК-геномов прокариот и эукариот показало, что более 95 % CDSs являются неоднородными последовательностями, обладающими свойствами 3-регулярности и скрытой профильной периодичности (профильности) с периодом равным или кратным трем [1]. Причем, свойство 3-регулярности в отсутствие триплетной профильности может рассматриваться как ослабленная (в результате мутационных изменений в ДНК) триплетная профильность. Таким образом, свойство 3-регулярности есть проявление первого уровня организации кодирования, связанного с размером кодонов генетического кода. Скрытая профильность с периодом кратным, но не равным трём отражает наличие второго уровня в организации кодирования, который, как правило, обусловлен размером повторяющихся доменов в структуре белков. В настоящей работе исследуются структурно-статистические свойства нуклеотидных последовательностей одноцепочечных РНК-геномов арбовирусов рода *Flavivirus* из семейства *Flaviviridae*. Рассматриваются геномы, в которых CDSs специфичных белков вируса последовательно, без перекрывания, располагаются в CDS общего полипептида. Для каждого генома анализируются структурно-статистические свойства всех CDSs. Поскольку геномы РНК-вирусов быстро мутируют [2], в них можно ожидать отклонение от закономерностей, выявленных в геномах прокариот и эукариот. Например, возможна потеря свойства профильной периодичности в отдельных кодирующих сегментах вирусного генома.

Геном флавивирусов представлен одноцепочечной РНК позитивной полярности (≈ 11 000 нукл.), которая является инфекционной. Нуклеокапсид флавивирусного вириона содержит капсидный белок C и геномную одноцепочечную РНК. Нуклеокапсид окружен липидной мембраной, в которую включены мембранный белок M и оболочечный белок E, взаимодействующие при сборке вириона. Вирусные белки C, M, E и семь неструктурных белков (NS1, NS2A, NS2B, NS3, NS4A, NS4B, NS5), необходимые для размножения вируса в клетках хозяина [3], последовательно считываются в единой рамке считывания при синтезе полипротеина, процессинг которого происходит ко- и пост-трансляционно.

Многие флавивирусы, являются опаснейшими патогенами для человека и при укусе насекомых (клещи, комары) могут вызывать парезы, параличи, энцефалиты и геморрагические лихорадки с высокой летальностью [3]. Наибольшую тревогу вызывает распространение по всему миру опасных болезней тропического и субтропического пояса, вызываемых такими флавивирусами, как вирусы желтой лихорадки, лихорадки Дэнге, Зика, Западного Нила, японского энцефалита [4]. Для России особенно актуальна проблема эпидемического распространения вирусов клещевого энцефалита, Западного Нила, Повассан [5] и омской геморрагической лихорадки [6].

Большинство флавивирусов передается теплокровным хозяевам членистоногими переносчиками [3, 7]. Однако некоторые флавивирусы, открытые в последние годы, специфичны только для насекомых и не передаются теплокровным [8, 9].

В работе с помощью спектрально-статистического подхода анализируются 62 последовательности полных геномов флавивирусов вместе с 415 отдельными, неперекрывающимися кодирующими последовательностями из этих геномов. Исходные файлы с данными были получены из базы GenBank [10]. В основном, в работе рассматриваются вирусы, представляющие опасность для теплокровных и, в частности, для человека, переносчиками которых являются комары и клещи. В отдельную группу выделены вирусы, найденные только у комаров.

Результаты анализа структурно-статистических свойств CDSs, полученные для геномов флавивирусов сравниваются с аналогичными результатами, полученными ранее для CDSs из геномов бактерий и многоклеточных организмов.

2. Материалы и методы

2.1. Спектрально-статистический подход

Для исследования структурно-статистических свойств, как отдельных кодирующих сегментов, так и полных геномов вирусов, исследуемых в настоящей работе, применялся спектрально-статистический подход (2S-подход) [1].

При распознавании скрытой профильной периодичности в нуклеотидной последовательности этот подход опирается на анализ статистических спектров, аналитический вид которых подробно описан ранее [1]. Поэтому в настоящей работе используется только анализ графических представлений статистических спектров 2S-подхода.

Согласно 2S-подходу, оценка длины скрытого профильного периода в анализируемой нуклеотидной последовательности осуществляется на основе анализа её характеристического спектра, обозначаемого символом **C**. Однако такой анализ проводится только в том случае, когда на основе соответствующего статистического критерия последовательность признаётся неоднородной. Этот критерий использует анализ спектра отклонения этой последовательности от однородности, обозначаемого как $\mathbf{D}_1$. Пример такого спектра для кодирующей последовательности (CDS) тетрапирольной метилазы из генома бактерии *Mycoplasma mycoides* из базы данных KEGG [11]

показан на рисунке 1,**а**. Последовательность признаётся неоднородной, если значения спектра $\mathbf{D}_1$ превышают единицу не менее чем на 5 % тест-периодов в анализируемом диапазоне. Так анализируемая последовательность, спектр $\mathbf{D}_1$ которой показан на рисунке 1,**а**, признаётся неоднородной.

Если анализируемая последовательность признана неоднородной, то в качестве оценки длины скрытого профильного периода в этой последовательности выбирается тот тест-период, на котором впервые (с учётом статистической погрешности) достигается максимальное значение её характеристического спектра **C**. На рисунке 1,**б** показан характеристический спектр для CDS тетрапирольной метилазы *M. mycoides*. Согласно этому спектру в качестве оценки длины скрытого профильного периода следует выбрать тест-период в 12 нукл.

Для подтверждения выбранной оценки L длины скрытого профильного периода в 2S-подходе используется спектр $\mathbf{D}_L$ отклонения анализируемой последовательности от L-профильности (скрытой L-профильной периодичности). В последовательности признаётся наличие скрытой L-профильности, если значения спектра $\mathbf{D}_L$ не превышают единицу не менее чем на 95 % тест-периодов в анализируемом диапазоне. На рисунке 1,в показан спектр $\mathbf{D}_{12}$ отклонения от 12-профильности анализируемой кодирующей последовательности из генома *M. mycoides*. Согласно этому спектру в ней распознаётся 12-профильность. Вместе с тем, согласно спектру $\mathbf{D}_3$ скрытая 3-профильность в этой последовательности отсутствует (см. рис. 1,**г**).

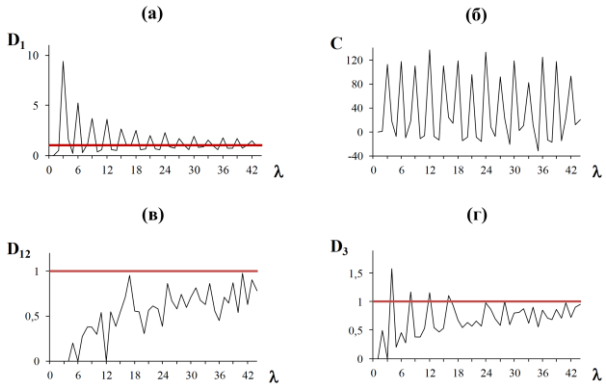

Рис. 1. Анализ CDS тетрапирольной метилазы (KEGG ENTRY MSC_0564, 888 нукл.) из генома бактерии *M. mycoides* с помощью 2S-подхода. (**а**) Спектр отклонения от однородности. (**б**) Характеристический спектр. (**в**) Спектр отклонения от скрытой 12-профильной периодичности. (**г**) Спектр отклонения от скрытой триплетной профильной периодичности.

Характеристический спектр CDS тетрапирольной метилазы *M. mycoides* обладает ещё одним свойством: практически, все его максимумы (пики) наблюдаются на тест-периодах, кратных трём. Такая особенность была названа в работе [1] свойством 3-регулярности анализируемой последовательности. Было показано, что свойство 3-регулярности в кодирующих районах обусловлено размером кодона генетического кода в 3 нукл. Этот вывод был подтверждён на основе численных экспериментов с бинарно перекодированными абзацами литературных текстов. Кроме того, в кодирующих районах прокариотических и эукариотических организмов практически все длины скрытых профильных периодов кратны трём [1]. Ранее наличие двухуровневой организации кодирования в CDS отмечалось, если в последовательности наблюдается свойство 3-регулярности и длина скрытого профильного периода кратна, но не равна трём. Отметим, что это явление наблюдалось в значительной части CDSs геномов прокариот и эукариот [1]. Согласно такому подходу, в последовательности CDS тетрапирольной метилазы *M. mycoides* обнаружена двухуровневая организация кодирования (см. рис. 1).

Согласно спектрам 2S-подхода, показанным на рисунке 2, в последовательности CDS для агматиновой деиминазы (agmatine deiminase) *M. mycoides* выявляется свойство 3-регулярности и распознаётся скрытая триплетная профильность, т.е. двухуровневая организация кодирования отсутствует.

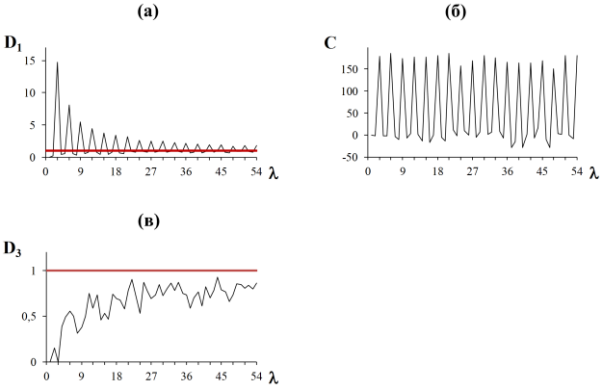

Рис. 2. Анализ спектрально-статистических свойств CDS агматиновой деиминазы *M. mycoides* (KEGG ENTRY MSC_0701, 1095 нукл.) с помощью 2S-подхода. (**а**) Спектр отклонения от однородности. (**б**) Характеристический спектр. (**в**) Спектр отклонения от скрытой 3-профильной периодичности.

2.2. Геномные последовательности флавивирусов

Две выборки кодирующих геномных последовательностей флавивирусов из базы GenBank анализировались в работе. Одна выборка содержала CDSs из геномов вирусов, переносимых комарами и клещами, и передаваемых человеку при укусе насекомого. В неё, в частности, вошли кодирующие последовательности из геномов вирусов жёлтой лихорадки, лихорадки Дэнге, японского энцефалита и энцефалита долины Муррея, энцефалита Сент-Луис и вируса Зика. Общая численность этой выборки составила 384

последовательности. Другая выборка содержала CDSs из геномов вирусов, специфичных только для комаров и не вызывающих заболеваний у теплокровных. В эту выборку вошли кодирующие последовательности геномов вирусов Aedes flavivirus, Culex flavivirus и Cell fusing agent – вируса, первоначально выделенного из линии клеток комаров *Aedes aegypti*. В этой выборке находилось 95 последовательностей.

3. Результаты и обсуждение

Согласно результатам работы [1] практически все кодирующие последовательности (CDSs) из проанализированных пяти геномов эукариот и 10 геномов прокариот обладают свойством 3-регулярности. Кроме того, в более 90 % этих последовательностей распознаётся скрытая профильная периодичность с длиной скрытого профильного периода кратной трём. При этом около 77 % кодирующих районов обладают скрытой триплетной профильной периодичностью и, приблизительно, в 13 % кодирующих последовательностей распознаётся двухуровневая организация кодирования с длиной периода, кратной, но не равной трём. Ранее отмечалось, что в кодирующих последовательностях с двухуровневой организацией кодирования может наблюдаться корреляция длины скрытого периода с длиной повторяющегося домена в кодируемом белке.

По результатам анализа, выполненного в настоящей работе, на рисунке 3 представлена дендрограмма структурно-статистических свойств кодирующих последовательностей из геномов флавивирусов, опасных для теплокровных и, в частности, для человека, которые передаются при укусах комаров и клещей. Из этой дендрограммы следует, что во всех CDSs, кодирующих полипротеин распознаётся скрытая триплетная профильная периодичность. Однако, среди CDSs отдельных структурных единиц полипротеина скрытая триплетная профильная периодичность наблюдается только в 86 % последовательностей. Остальные 14 % CDSs отдельных структурных единиц не обладают свойством 3-регулярности. Около 9 % CDSs отдельных структурных белков вируса оказываются однородными, 4 % CDSs не обладают скрытой профильностью и 0.3 % CDSs обладают скрытой профильностью с длиной периода, некратной трём.

На рисунке 4 приведена аналогичная дендрограмма для CDSs вирусов, специфичных для комаров. Здесь также во всех CDSs полипротеинов распознаётся скрытая триплетная профильная периодичность. Однако, в CDSs отдельных структурных единиц на 20 % снижен уровень последовательностей, в которых распознаётся скрытая триплетная профильная периодичность; кроме того, среди этих CDSs процент однородных последовательностей в три раза выше в сравнении с аналогичными CDSs вирусов, поражающих теплокровных. Сравнивая структурно-статистические свойства CDSs геномов прокариот и эукариот [1] с аналогичными свойствами CDSs белков вирусного генома, можно отметить следующие особенности. В CDSs белков вирусного генома отсутствует двухуровневая организация кодирования, что, вероятнее всего, связано с отсутствием повторяющихся структур в CDSs вирусов. Кроме того, в CDSs генома флавивирусов наблюдается значительная доля однородных последовательностей. Возможно, это связано с высокой мутабильностью вируса [12], что может объяснять и другие особенности в отклонении от 3-профильности в CDSs вирусов (см. рис. 3 и рис. 4). Приведем пример того, как кодирующая последовательность неструктурного белка NS2B со скрытой триплетной профильностью теряет это свойство в результате двух вставок или двух делеций (см. рис. 5), превращаясь в однородную последовательность. Исходная последовательность длиной 375 нукл. получена из GenBank (Accession Number J741266). На рисунке 5,**а** спектр отклонения от однородности $\mathbf{D}_1$ исходной последовательности фиксирует её неоднородность. Спектр $\mathbf{D}_3$ отклонения от триплетной профильности в этой последовательности (см. рис. 5,**б**) показывает наличие в ней скрытой триплетной профильности. Вставки двух нуклеотидов в исходную последовательность, показанные на рисунке 5,**в**, приводят к тому, что последовательность становится однородной, что подтверждает спектр $\mathbf{D}_1$ на рисунке 5,**г**. Аналогичный эффект наблюдается при удалении двух нуклеотидов (см. рис. 5,**д**), что подтверждает спектр $\mathbf{D}_1$ на рисунке 5,**е**.

4. Заключение

Исследование структурно-статистических свойств кодирующих последовательностей геномов флавивирусов и сравнительный анализ с выявленными ранее свойствами кодирующих районов геномов прокариот и эукариот подробно описаны в работе [13]. В большинстве кодирующих районов геномов флавивирусов, как и в кодирующих последовательностях геномов прокариот и эукариот, выявляются свойства 3-регулярности и скрытой триплетной профильной периодичности (3-профильности). Однако в кодирующих последовательностях геномов флавивирусов полностью отсутствует двухуровневая организация кодирования, которая свойственна части кодирующих районов геномов прокариотических и эукариотических организмов и обусловлена повторяющимися структурными доменами в кодируемых белках. Кроме того, для весьма значительной доли кодирующих последовательностей геномов флавивирусов характерна однородность нуклеотидной последовательности. Показано, что эффект однородности в кодирующих последовательностях

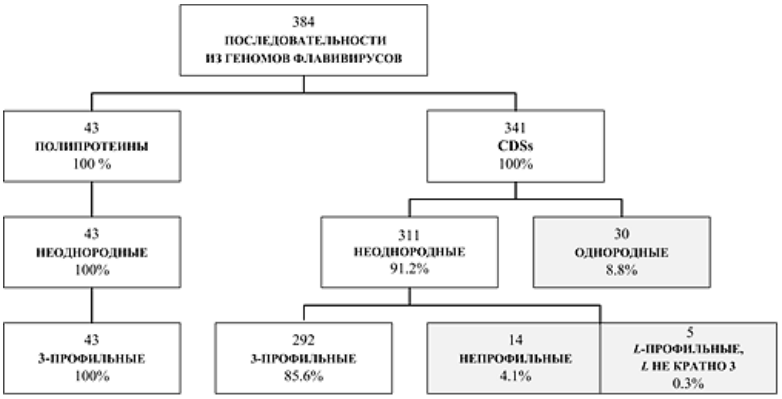

Рис. 3. Дендрограмма структурно-статистических свойств CDSs из геномов флавивирусов, переносимых комарами и клещами и поражающих теплокровных.

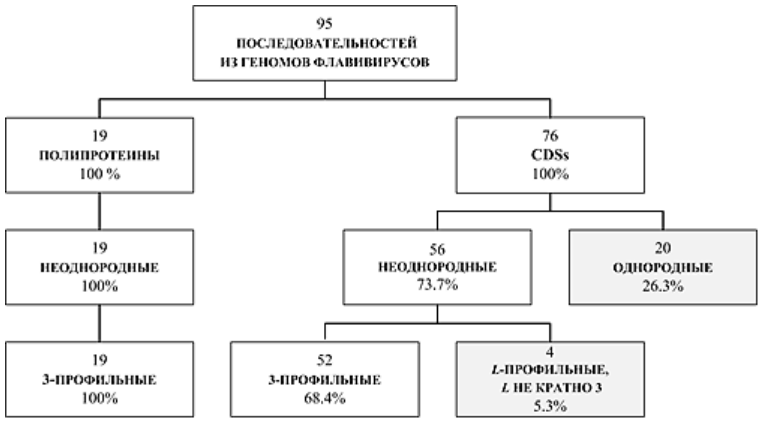

Рис. 4. Дендрограмма структурно-статистических свойств CDSs из геномов флавивирусов, специфичных для комаров.

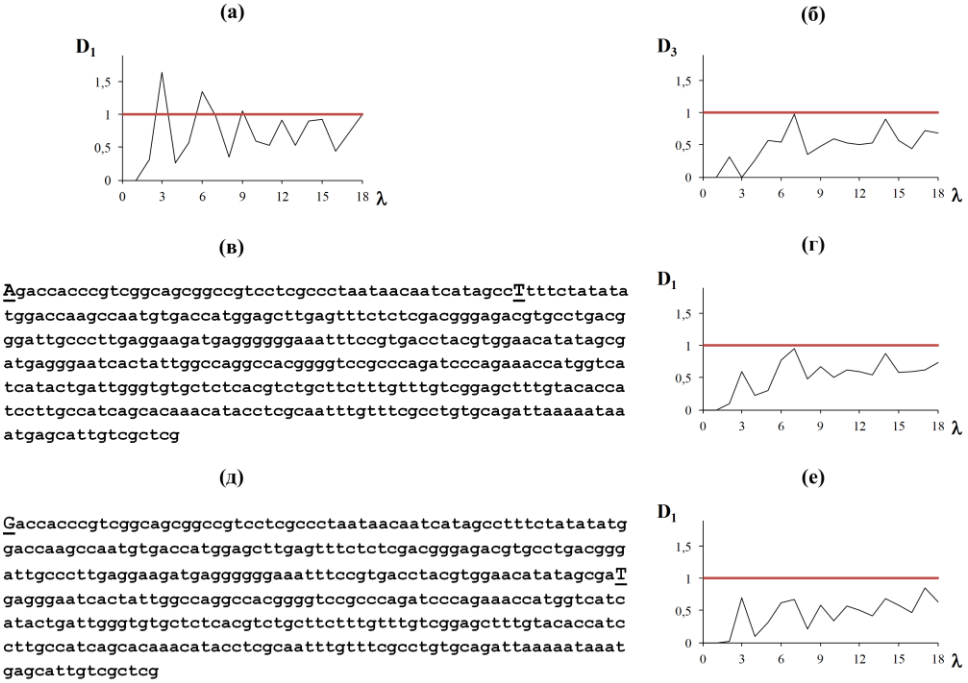

Рис. 5. Иллюстрация искажения скрытой 3-профильности (**а**, **б**) в CDS для неструктурного белка NS2B специфичного вируса комаров *Aedes* до однородности в результате двух вставок (**в**, **г**) и двух делеций (**д**, **е**) в исходной последовательности. Вставленные и делетированные нуклеотиды показаны заглавными буквами.

геномов флавивирусов может быть вызван, как минимум, двумя точечными мутациями в виде вставок и делеций. Выявленные особенности в геномах флавивирусов могут быть обусловлены простой структурой и высокой скоростью мутаций вирусных геномов.

5. Список литературы

1. Кутыркин В.А., Чалей М.Б. Модель организации кодирования в прокариотических организмах. *Мат. биол. биоинф*. 2016. Т. 11. С. 24–45. doi: 10.17537/2016.11.24.
2. Duffy S., Shackelton L.A., Holmes E.C. Rates of evolutionary change in viruses: patterns and determinants. *Nature Reviews Genetics*. 2008. V. 9. P. 267–276. doi: 10.1038/nrg2323.
3. *Руководство по вирусологии: Вирусы и вирусные инфекции человека и животных*. Под. ред. Львова Д.К. М.: Медицинское информационное агентство, 2013.
4. Holbrook M.R. Historical perspectives on flavivirus research. *Viruses*. 2017. V. 9. P. 97. doi: 10.3390/v9050097.
5. Субботина Е.Л., Локтев В.Б. Молекулярная эволюция вируса клещевого энцефалита и вируса Повассан. *Молекулярная биология*. 2012. Т. 46. С. 82–92.
6. Ястребов В.К., Якименко В.В. Омская геморрагическая лихорадка: итоги исследований (1946–2013). *Вопросы вирусологии*. 2014. Т. 59. № 6. С. 5–11.
7. Grard G., Moureau G., Charrel R.N., Lemasson J.J., Gonzalez J.P., Gallian P., Gritsun T.S., Holmes E.C., Gould E.A., de Lamballerie X. Genetic characterization of tick-borne flaviviruses: New insights into evolution, pathogenetic determinants and taxonomy. *Virology*. 2007. V. 361. P. 80–92. doi: 10.1016/j.virol.2006.09.015.
8. Blitvich B.J., Firth A.E. Insect-specific flaviviruses: a systematic review of their discovery, host range, mode of transmission, superinfection exclusion potential and genomic organization. *Viruses*. 2015. V. 7. P. 1927–1959. doi: 10.3390/v7041927.
9. Calzolari M., Zé-Zé L., Vázquez A., Seco M.P.S., Amaro F., Dottori M. Insect-specific flaviviruses, a worldwide widespread group of viruses only detected in insects. *Infection, Genetics and Evolution*. 2016. V. 40. P. 381–388. doi: 10.1016/j.meegid.2015.07.032.
10. *GenBank*. URL: https://www.ncbi.nlm.nih.gov/genbank/ (дата обращения 10.06.2018).
11. *KEGG: Kyoto Encyclopedia of Genes and Genomes*. URL: http://www.kegg.jp (дата обращения: 10.06.2018).
12. Combe M., Sanjuán R. Variation in RNA virus mutation rates across host cells. *PLoS Pathog*. 2014. V. 10. P. e1003855. doi: 10.1371/journal.ppat.1003855.
13. Тюлько Ж.С., Кутыркин В.А., Чалей М.Б. *Мат. биол. биоинф*. 2017. Т. 12. С. 343–353. doi: 10.17537/2017.12.