***Доклады Международной конференции "Математическая биология и биоинформатика"**. Под ред. В.Д. Лахно. Том 7. Пущино: ИМПБ РАН, 2018. Статья № e55. doi: 10.17537/icmbb18.105*

Исследование метагеномов океанической микробиоты

Зимин А.А.[1, 2], Назипова Н.Н.[3]

[1]*ФГБНУ Институт биохимии и физиологии микроорганизмов им. Г. К. Скрябина РАН, Российская Федерация, г. Пущино, Московская область*
[2]*Пущинский государственный естественно-научный институт, Российская Федерация, г. Пущино, Московская область*
[3]*ИМПБ РАН – филиал ИПМ им. М.В. Келдыша РАН, Российская Федерация, г. Пущино, Московская область*

zimin@ibpm.pushchino.ru

ДНК-лигаза фага Т4 – это ключевой инструмент генетической инженерии. Поиск ее гомологов в метагеномах океанической микробиоты – это путь создания нового инструментария для развития методов обратной генетики, включая редактирование генома человека. В метагеномах пелагической и осадочной глубоководной микробиоты было найдено соответственно 476 и 24 гомолога ДНК-лигазы фага Т4. Филогенетический анализ данных аминокислотных последовательностей показал, что большинство из них образуют отдельную ветвь на дереве ДНК-лигаз, близко к которой находится ветвь АТФ-зависимых ДНК-лигаз энтеробактерий. Этот результат говорит о наличие в толще воды открытого океана на его дне новых еще неизвестных бактерий. Ряд обнаруженных гомологов показал сильное сходство с некоторыми фаговыми ДНК-лигазами.

Ключевые слова: *морская метагеномика, глубоководные генетические исследования, микробиота и вирусы мирового океана, ДНК-лигаза бактериофага Т4, полинуклеотидлигазы*.

The study of metagenomes of oceanic microbiota

Zimin A.A.[1, 2], Nazipova N.N.[3]

[1]*Skryabin's Institute of Biochemistry and Physiology of Microorganisms of the Russian Academy of Sciences, Russia, Pushchino, Moscow region*
[2]*Pushchino State Institute of Natural Science, Russia, Pushchino, Moscow region*
[3]*IMPB RAS – Branch of KIAM RAS, Russia, Pushchino, Moscow region*

The DNA ligase of phage T4 is a key tool of genetic engineering. The search of its homologues in metagenomes of marine microbiota is the way of creating a new tools for the development of methods of reverse genetics, including editing the human genome. By the use of the PSI-BLST algorithm, 476 and 24 homologues of T4 DNA ligase, respectively, were found in metagenomes of the pelagic and sedimentary deep-sea microbiota. Phylogenetic analysis of the amino acid sequences showed that most of them form a separate branch on the DNA ligase phylogenetic tree, close to which there is a branch of ATP-dependent DNA-ligase of enterobacteria. This finding releals presence in the ocean, both in the water column and at its bottom, of new yet unknown bacteria. A number of the homologues showed great similarity with some phage DNA-ligases.

Key words: the marine metagenomics, deep-sea genetic studies, world ocean microbiota and viruses, T4 bacteriophage DNA ligase, polynucleotidligase.

Данное исследование направлено на решение двух глобальных задач современной биологии. Во-первых, это изучение микробиоты мирового океана, включая глубоководные сообщества. Во-вторых, это поиск в океанической микробиоте новых аминокислотных последовательностей ДНК-лигаз, гомологичных ДНК-лигазе бактериовируса Т4. Такие находки могут привести к получению новых ферментов с новыми свойствами для их применения, как в генетической инженерии, так и других областях науки. ДНК-лигазы – это перспективные объекты для антибактериальной и противоопухолевой терапии. Изучение многообразия этих ферментов методами биоинформатики, поиск новых представителей данной группы ферментов в базах данных биологических последовательностей формирует

существенный задел для будущих фармакологических исследований.

1. Полинуклеотидлигазы и направления их исследований

Полинуклеотидлигазы – это ферменты, которые восстанавливают фосфодиэфирную связь (сшивают) 5'-PO4 и 3'-OH – концы полинуклеотидов. В природе есть два вида полинуклеотидлигаз: это РНК-лигазы и ДНК-лигазы. Последние являются хранителями целостности генома. Дисфункция по ДНК-лигазе лежит в основе ряда генетических заболеваний человека. ДНК-лигаза бактериофага Т4 является основным инструментом генной инженерии с момента появления этого метода в 1977 году. К настоящему времени поняты основные структурные основы лигирования полинуклеотидов. Развиваются работы по исследованию происхождения этих ключевых для жизни ферментов. Открытие новых генов, кодирующих новые полинуклеотидлигазы в метагеномах различных экологических ниш, создает реальные перспективы развития новых направлений в современной биохимии генома.

2. Вириомика океана

Высокопроизводительное метагеномное секвенирование позволило оценить геномное богатство различных биотопов без выделения отдельных микроорганизмов или их вирусов. Это привело к появлению метагеномики – направлению биологии, которое исследует геномы организмов и вирусов непосредственно из природных проб. Совокупность геномных данных из определенной пробы называется метагеномом. Основной задачей метагеномики на сегодняшний день является изучение истинного биоразнообразия микробиоты и ее экологических взаимодействий в самых различных средах, начиная с внутренностей живых организмов и кончая почвами глубин мирового океана. Первое метагеномное исследование было сделано в 2002 г., когда Breitbart и соавторы изучили вириомы морской воды из четырех океанов вокруг США [1]. Была создана первая морская вириомная база данных размером около 6 Gbp, было найдено много новых вирусных геномов и показано преобладание в океане бактериовирусов, содержащих одноцепочечную ДНК. Очень интересны результаты сравнительного исследования состава двух глубоководных вирусных сообществ, полученных из зоны разлома Романче в Атлантическом океане (пробы собраны на глубине 5200 м) и из юго-западной части Средиземного моря (с глубины 2400 м). В обоих виромах преобладали вирусы архей и бактерий, на которые приходилось 92.3 % относительного разнообразия в Атлантическом океане и 83.6 % в Средиземном море [2]. Вопросы глобальной биогеографии вирусов и взамосвязей в вирусных сообществах были затронуты в работе по сравнительному анализу двух полярных пресноводных вириомов. Оказалось, что в полярных пресноводных вириомах доминируют вирусы, пока еще неизвестные науке и известные уже вирусы, имеющие одноцепочечные ДНК. Эти два уникальных вирусных сообщества в основном связаны именно друг с другом и имеются лишь незначительные генетические сходства с вириомами из других сред, включая вирион Северного Ледовитого океана. [3]. Исследуются вириомы самых различных морских биотопов. Например, метагеномные исследования показали, что вирусные сообщества, связанные с кораллами, очень разнообразны. Эта экологическая ниша содержит бактериовирусы, имеющие как двуцепочечную, так и одноцепочечную ДНК, а также РНК-содержащие бактериофаги. Среди идентифицированных вирусов кораллов преобладали бактериофаги отряда *Caudovirales*, содержащие двуцепочечную ДНК [4, 5]. Специализированные базы данных морской метагеномики также содержат вириомные данные. Например, это база больших данных MAR специально предназначена для исследований в области морской метагеномики (https://s1.sfb.uit.no/public/mar/) [6]. Для анализа вириомов разрабатываются специальные программные средства. Например, VIROME – специализированный пакет программных средств для анализа вирусных метагеномных последовательностей, который может использоваться при подготовке данных для публикации [7]. Другим подобным средством является FastViromeExplorer [8]. Это программный пакет для идентификации вирусов эукариот и бактериофагов и изучения их биоразнообразия в данных, полученных методами метагеномики.

3. Метагеномные базы данных океана

Одной из наиболее известных баз данных морских генетических последовательностей является база данных, метагеномов из Саргассова моря близ Бермудов [9]. Более обширное метагеномное исследование морской планктонической микробиоты было проведено из поверхностных вод в Северной Атлантике и рядом с Панамским каналом в Тихом океане [10]. Было определено 7.7 млн. последовательностей ДНК общей длиной 6.3 Gbp. База данных (БД) GOS содержит более 6 млн. аминокислотных последовательностей, транслированных из ДНК. Базы больших генетических данных о глубоководной морской микробиоте донных отложений Северного Ледовитого океана в районе подводного Срединного арктического хребета имени Гаккеля (Arctic Mid-Ocean Ridge) были созданы следующим образом. Геномная информация была получена методом глубокого метагеномного секвенирование ДНК образца осадка

GC14, в результате чего был получен меньший набор данных (LCGC14, 8.6 Гбит) и после амплификации ДНК (MDA) большой набор метагеномных данных (LCGC14AMP, 56.6 Gbp) (http://opensource.scilifelab.se/) [11]. БД GOS и LCGC14 были использованы для анализа в данной работе.

4. Методы, использованные в работе

Для проведения сравнения аминокислотной последовательностей ДНК-лигазы бактериофага T4, продукта гена 30, с базами данных белковых последовательностей на сервере NCBI использовался алгоритмом PSI-BLAST [12] с уровнем достоверности результатов *E-value* < 3*e*–29. При этом позиционный итерационный поиск аналогов данного белка производился до тех пор, пока каждая последующая итерация обнаруживала в базе данных GenBank новые локальные сходства. Когда картина сходства последовательностей не изменялась, поиск аналогов прекращался. Полученный файл с аминокислотными последовательностями гомологов ДНК-лигазы бактериофага T4 в формате FASTA использовали для обработки в пакете программ Mega6 [13].

На основе полученных данных был осуществлен предварительный филогенетический анализ гомологов ДНК-лигазы фага T4 как из баз данных аминокислотных последовательностей белков пелагической морской микробиоты, белков морской микробиоты осадочных пород, так и белков из различных биотопов суши. Для этого предварительно были проведены выравнивания последовательностей с помощью программы ClustalW [14], филогенетическое дерево строилось с помощью пакета программ Mega6 [13]. Длина ДНК-лигазы бактериофага T4 487 аминокислотных остатков. В связи с этим как программа ClustalW, так и другие средства пакета Mega6, в отличие от программного средства PSI-BLAST, весьма ограничены в своих возможностях при анализе последовательностей значительно более коротких, чем основной объект сравнения. Ряд найденных программным средством PSI-BLAST последовательностей аминокислот гомологов ДНК-лигазы фага T4 пришлось делетировать из исходного файла в формате FASTA. На необходимость такой редакции указывало само программное средство после запуска той или иной обработки. Данная ситуация вполне адекватна для наборов генетических данных полученных средствами метагеномики. В процессе подготовки проб для исходного определения генетических последовательностей происходит выделения ДНК из пробы морской воды или экстракта морского бентоса после соответствующей фильтрации жидкого материала. Уже на этой стадии происходит разрыв длинных геномных ДНК в самых различных местах. Дальнейшая обработка выделенной и очищенной от других веществ ДНК ультразвуком вносит массу дополнительных разрывов. В результате секвенирования большого разнообразия геномов весьма ограниченно находятся перекрытия для ридов, что часто приводит к появлению последовательностей лишь частей генов в руках исследователя. Статистические ограничения, имеющиеся у программ, включенных в любой пакет обработки, не позволяют работать с сильно укороченными последовательностями, которых обычно содержится до 70 % от общего числа последовательностей в метагеноме любой пробы. Длина самой короткой ДНК-лигазы, фермента бактериофага T3, 346 аминокислот, а ее активного центра примерно 230 аминокислот. Минимальный размер последовательности оставленной в файле для данного нашего анализа был 238 аминокислот. С точки зрения этих генетических параметров, можно полагать, что выбранный нами подход использования программных средств был вполне адекватным.

5. Поиск гомологов ДНК-лигазы бактериофага T4 в базах данных океанических генетических последовательностей

После второй итерации программы сравнения PSI-BLAST было найдено 476 гомологов ДНК-лигазы бактериофага T4 в больших генетических данных о морской пеларгической микробиоте и 24 гомолога в данных о глубоководной океанической микробиоте при значении статистической характеристики *E* < 3*e* –29. Для того, что бы понять место найденных аминокислотных последовательностей в биосфере Земли мы предприняли предварительный филогенетический анализ.

6. Филогенетический анализ морских аминокислотных последовательностей, гомологичных ДНК-лигазе бактериофага T4

Для предварительного филогенетического анализа и построения исходного дерева мы оставили в файле для сравнения только аминокислотные последовательности длиной более 238 аминокислот. Это связано с тем, что длина активного ферментативного центра данного фермента также около 230 аминокислот. Тем самым данные последовательности можно условно считать полными по возможному содержанию всех аминокислот, которые кодируют активный центр ферментов подобных исследуемому.

Для реконструкции филогении данной группы аминокислотных последовательностей с помощью пакета программ Mega6 были использованы следующие параметры работы средства.

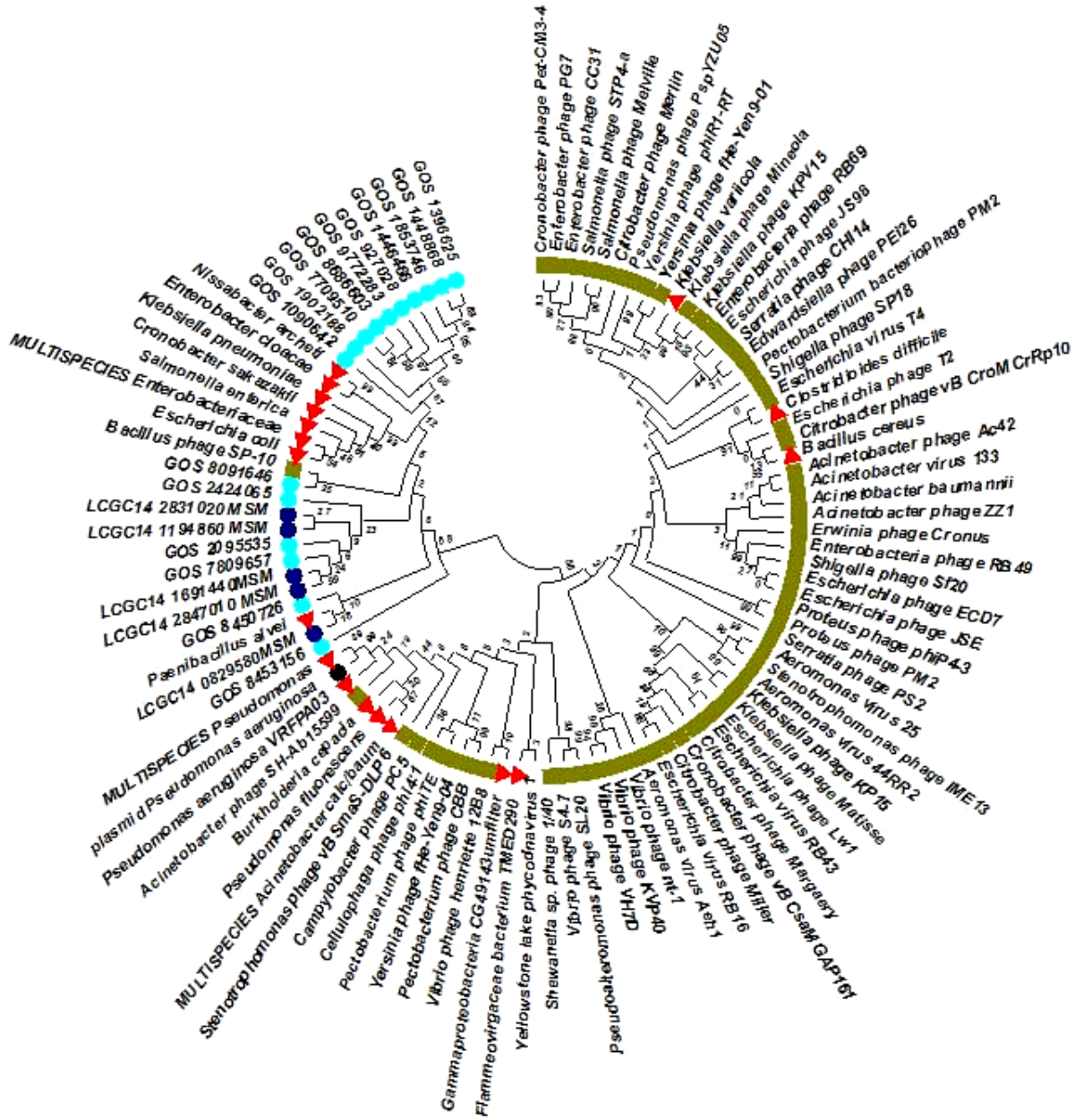

Рис. 1. Филогенетический анализ 100 гомологов ДНК-лигазы бактериофага Т4 из метагеномов пелагической и глубоководной осадочной микробиоты океана. Номера последовательностей и параметры анализа приведены в тексте статьи. Аминокислотные последовательности из базы метагеномных данных планктонической микробиоты (GOS) обозначены кружками голубого цвета, из базы данных глубоководной осадочной микробиоты Северного Ледовитого океана (LCGC) – темно-синими кружками, бактериальные АТФ-зависимые ДНК-лигазы – красными кружками, последовательность из генома плазмиды синегнойной палочки – черным кружком, последовательность из генома вируса водорослей не имеет специального обозначения, а ДНК-лигазы бактериофагов обозначены квадратами зеленым цветом.

Для филогенетического теста использовался «Bootstrap method» с числом итераций 1000 [15]. Также была использована модель аминокислотных замен Jones–Taylor–Thornton (JTT) [16]. Ветви, появлявшиеся менее чем в 50 % повторов, не использовались при построении финального дерева. Все позиции, содержащие бреши или в которых данные отсутствовали, были убраны. Эволюционные расстояния воспроизводятся на дереве в единицах, соответствующих числу аминокислотных замен на белок. Мы использовали на дереве понятные сокращенные названия живых сущностей, в геномах которых были найдены аминокислотные последовательности.

Полученное в результате филогенетического анализа дерево приведено на рисунке 1. Филогенетическое дерево, полученное нами в данной работе, имеет две основные ветви. На одной из них располагаются все морские лигазы, как глубоководные осадочные, так и пелагические, а

также последовательности из ряда бактерий. Это, в первую очередь, подветвь АТФ-зависимых ДНК-лигаз различных энтеробактерий, а также последовательность из *Paenibacillus alvei*. *Paenibacillus alvei* является бациллой, обычно обнаруживаемой в колониях пчел, а также в других средах, включая почву, молоко и человека. Она продуцирует альвеолизин, токсин, активируемый тиолом. Единственным исключением является нахождение на этой ветви лигазы бациллярного бактериофага SP10. Эта единственная вирусная последовательность на данной ветви филогенетического дерева может говорить как о включении гена этой лигазы в геном вируса из генома бактерий, так и о данном гене как прородители данной группы генов лигаз энтеробактерий и морских бактерий, выявленных в метагеномах GOS и LCGC. Другая ветвь содержит ДНК-лигазы бактериофагов, в той или иной мере родственных T4, последовательности ряда АТФ-зависимых ДНК-лигаз бактерий, а также последовательность ДНК-лигазы вируса водорослей. Это фикоднавирус 1. Его геном был собран из метагеномного набора данных Йеллоустонского озера. Геномные анализы показали, что этот фикоднавирус озера Йеллоустоун (YSLPV1) имеет длину генома 178 262 нуклеотидов, и он филогенетически близок к празиновирусами (также *Phycodnaviridae*). Надо заметить, что по длине генома он близок к T4-подобным бактериофагам.

7. В глубоководных осадках гомологов ДНК-дигазы фага T4 значительно меньше, чем в пелагической зоне

В генетических данных о глубоководных осадках около замка Локи Срединного арктического хребта им. Гаккеля достоверно было найдено 24 гомолога ДНК-лигазы бактериофага T4. Подобный поиск в больших данных о генетических последовательностях пеларгической микробиоты вокруг Панамского перешейка привел к находке 476 хитов.

Возможно, малое число гомологов ДНК-лигазы бактериофага T4, найденное нами в больших генетических данных о морской микробиоте глубоководных отложений Северного Ледовитого океана, может быть связано с влиянием гидротермальных источников, находящихся в 15 километрах от места выемки данных проб. В будущем, когда накопятся метагеномные данные, добытые в самых разных районах океана и с различных его глубин, можно будет повторить исследования. Возможно, результаты могут оказаться отличными от полученных в этом исследовании.

8. Почему некоторые морские АТФ-зависимые ДНК-лигазы формируют свой собственный кластер последовательностей?

Однозначно можно сказать, что мы обнаружили новый кластер АТФ-зависимых ДНК-лигаз. Исходя из филогенетического анализа, можно утверждать, что ряд аминокислотных последовательностей новых ферментов принадлежат к некультивируемым морским пелагическим бактериям. Более того, необходимо отметить, что среди наиболее близких к ним бактерий на филогенетическом дереве нет ни метантрофов, ни других архей. Можно с достаточной долей уверенности полагать, что мы обнаружили новый таксономический кластер морских, в том числе и глубоководных, протеобактерий, вероятно, близких к энтеробактериям. В дальнейшем, планируется расширить данный поиск с использование других генетических маркеров для исследования этого подмножества геномов в метагеномах морской микробиоты.

9. Выводы

1. С помощью одной из программ семейства BLAST нами были исследованы базы данных генетических последовательностей пелагической и глубоководной осадочной микробиоты и было найдено 476 и 24 новых гомолога ДНК-лигазы бактериофага T4, соответственно.
2. Предварительный филогенетический анализ океанических аминокислотных последовательностей показал, что они образуют отдельную ветвь на дереве ДНК-лигаз. Эта ветвь наиболее близко располагается к ветви АТФ-зависимых ДНК-лигаз энтеробактерий.
3. На основе предварительного филогенетического анализа можно предположить, что ген ДНК-лигазы бактериофага бацилл SP10 появился в геноме этого фага путем горизонтального переноса и происходит от АТФ-зависимых ДНК-лигаз грамотрицательных бактерий. Возможен также обратный сценарий, и этот ген является прародителем АТФ-зависимых ДНК-лигаз ряда бактерий.

10. Благодарности

Данная работа была частично поддержана проектом РФФИ №16-44-230855р_а.

11. Список литературы

1. Angly F.E., Felts B., Breitbart M., Salamon P., Edwards R.A., Carlson C., Chan A.M., Haynes M., Kelley S, Liu H., et al. The marine viromes of four oceanic regions. *PLoS Biology*. 2006. V. 4. Article No. e368. doi: 10.1371/journal.pbio.0040368.

2. Winter C., Garcia J.A., Weinbauer M.G., DuBow M.S., Herndl G.J. Comparison of deep-water viromes from the atlantic ocean and the mediterranean sea. *PLoS One*. 2014. V. 9. № 6. Article No. e100600. doi: 10.1371/journal.pone.0100600.
3. Aguirre de Carcer D., Lopez-Bueno A., Pearce D.A., Alcami A. Biodiversity and distribution of polar freshwater DNA viruses. *Sci. Adv*. 2015. V. 1. Article No. e1400127. doi: 10.1126/sciadv.1400127.
4. Wood-Charlson E.M., Weynberg K.D., Suttle C.A., Roux S., van Oppen M.J. Metagenomic characterization of viral communities in corals: mining biological signal from methodological noise. *Environ Microbiol*. 2015. V. 17. № 10. P. 3440–3449. doi: 10.1111/1462-2920.12803.
5. Laffy P.W., Wood-Charlson E.M., Turaev D., Jutz S., Pascelli C., Botté E.S., Bell S.C., Peirce T.E., Weynberg K.D., van Oppen M.J.H., et al. Reef invertebrate viromics: diversity, host specificity and functional capacity. *Environ Microbiol*. 2018. doi: 10.1111/1462-2920.14110.
6. Klemetsen T., Raknes I.A., Fu J., Agafonov A., Balasundaram S.V., Tartari G., Robertsen E., Willassen N.P. The MAR databases: development and implementation of databases specific for marine metagenomics. *Nucleic Acids Res*. 2018. V. 46. № D1. P. D692–D699. doi: 10.1093/nar/gkx1036.
7. Wommack K.E., Bhavsar J., Polson S.W., Chen J., Dumas M., Srinivasiah S., Furman M., Jamindar S., Nasko D.J. VIROME: a standard operating procedure for analysis of viral metagenome sequences. *Standards in Genomic Sciences*. 2012. V. 6. P. 427–439. doi: 10.4056/sigs.2945050
8. Tithi S.S., Aylward F.O., Jensen R.V., Zhang L. FastViromeExplorer: a pipeline for virus and phage identification and abundance profiling in metagenomics data. *Peer J*. 2018. V. 6. Article No. e4227. doi: 10.7717/peerj.4227.
9. Venter J.C., Remington K., Heidelberg J.F., Halpern A.L., Rusch D., Eisen J.A., Wu D., Paulsen I., Nelson K.E., Nelson W., et al. Environmental genome shotgun sequencing of the Sargasso Sea. *Science*. 2004. V. 304. № 5667. P. 66–74. doi: 10.1126/science.1093857.
10. Yooseph S., Sutton G., Rusch D.B., Halpern A.L., Williamson S.J., Remington K., Eisen J.A., Heidelberg K.B., Manning G., Li W., Jaroszewski L., et al. The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. *PLoS Biol*. 2007. V. 5. № 3. Article No. e16. doi: 10.1371/journal.pbio.0050016.
11. Spang A., Saw J.H., Jørgensen S.L., Zaremba-Niedzwiedzka K., Martijn J., Lind A.E., van Eijk R., Schleper C., Guy L., Ettema T.J.G. Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature*. 2015 V. 521. № 7551. P. 173–179. doi: 10.1038/nature14447.
12. Altschul S.F., Madden T.L., Schäffer A.A., Zhang J., Zhang Z., Miller W., Lipman D.J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. 1997. V. 25. P. 3389–3402.
13. Tamura K., Stecher G., Peterson D., Filipski A., and Kumar S. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Molecular Biology and Evolution*. 2013. V. 30. P. 2725–2729. doi: 10.1093/molbev/mst197.
14. Jeanmougin F., Thompson J.D., Gouy M., Higgins D.G., Gibson T.J. Multiple sequence alignment with Clustal X. *Trends Biochem. Sci*. 1998. V. 23. P. 403–405.
15. Felsenstein J. Confidence limits on phylogenies: An approach using the bootstrap. *Evolution*. 1985. V. 39. P. 783–791.
16. Jones D.T., Taylor W.R., Thornton J.M. The rapid generation of mutation data matrices from protein sequences. *Computer Applications in the Biosciences*. 1992. V. 8. P. 275–282.