Инфекция и иммунитет
2021, Т. 11, № 1, с. 42–56

Russian Journal of Infection and Immunity = Infektsiya i immunitet
2021, vol. 11, no. 1, pp. 42–56

ИСПОЛЬЗОВАНИЕ МЕТОДОВ СТАТИСТИЧЕСКОЙ ФИЛОГЕНЕТИКИ В ВИРУСОЛОГИИ

Ю.А. Вакуленко[1,2], А.Н. Лукашев[1,3], А.А. Девяткин[3]

[1]Институт медицинской паразитологии, тропических и трансмиссивных заболеваний им. Е.И. Марциновского, Первый Московский государственный медицинский университет имени И.М. Сеченова, Москва, Россия
[2]Московский государственный университет им. М.В. Ломоносова, Москва, Россия
[3]Институт молекулярной медицины, Первый Московский государственный медицинский университет имени И.М. Сеченова, Москва, Россия

Резюме. Молекулярная филогенетика, и в частности статистическая филогенетика, широко применяется для решения фундаментальных и прикладных задач вирусологии. Байесовские, или статистические, филогенетические методы, вошедшие в практику 10–15 лет назад, значительно расширили круг вопросов, на которые можно получить ответы, исходя из анализа нуклеотидных и аминокислотных последовательностей. Возможность использования разных моделей эволюции позволяет восстанавливать хронологию, географию и динамику распространения инфекции. Например, при анализе последовательностей ВИЧ глобально распространенной группы М байесовскими методами филогеографического анализа было показано, что последний общий предок этих вирусов с вероятностью 99% возник в окрестностях города Киншаса (Демократическая Республика Конго) в начале 1920-х гг. В другой работе показали, что серотип вируса гриппа H9N2, скорее всего, перешел к человеку от диких уток в Гонконге в конце 60-х гг. XX в. Кроме того, при помощи байесовского анализа можно оценить влияние определенных событий или принимаемых мер на развитие эпидемического процесса. Так, например, ретроспективно было показано, что число заражений вирусом гепатита С в Египте увеличилось на несколько порядков в середине XX в. Резкий рост новых случаев связывают с началом лечения шистосомоза. Лекарство вводили при помощи уколов, нестерильные шприцы применяли многократно. Набор методов байесовского анализа был использован в десятках тысяч исследований, описывающих разные аспекты возникновения и распространения инфекционных заболеваний человека и животных. Сложность байесовских филогенетических методов определяет строгие требования к анализируемым данным. Корректность результатов филогенетического анализа зависит от ряда факторов. Например, необходим выбор эволюционной модели, наиболее адекватно описывающей исследуемые объекты. Обязательным этапом при формулировании результатов является обоснование выбранной модели. Для вирусов характерно заимствование генетических элементов из других организмов, поэтому геномы даже близкородственных вирусов могут иметь негомологичные участки, непригодные для филогенетического анализа. Другим условием является создание репрезентативной выборки исследуемых объектов. Зачастую в публикациях не указываются все этапы выполнения анализа, из-за чего полученные результаты могут трактоваться неоднозначно. Коррект-

Адрес для переписки:
Девяткин Андрей Андреевич
119048, Россия, Москва, ул. Трубецкая, 8/2,
Институт молекулярной медицины, Первый Московский государственный медицинский университет имени И.М. Сеченова.
Тел.: 8 (495) 609-14-00.
E-mail: andreideviatkin@gmail.com

Contacts:
Andrei A. Deviatkin
119048, Russian Federation, Moscow, Trubetskaya str., 8/2,
Institute of Molecular Medicine, Sechenov First Moscow State Medical University.
Phone: +7 (495) 609-14-00.
E-mail: andreideviatkin@gmail.com

Для цитирования:
Вакуленко Ю.А., Лукашев А.Н., Девяткин А.А. Использование методов статистической филогенетики в вирусологии // Инфекция и иммунитет. 2021. Т. 11, № 1. С. 42–56. doi: 10.15789/2220-7619-TUO-1519

Citation:
Vakulenko Yu.A., Lukashev A.N., Deviatkin A.A. The use of statistical phylogenetics in virology // Russian Journal of Infection and Immunity = Infektsiya i immunitet, 2021, vol. 11, no. 1, pp. 42–56. doi: 10.15789/2220-7619-TUO-1519

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 19-115-50403.

DOI: http://dx.doi.org/10.15789/2220-7619-TUO-1519

ное использование методов статистической филогенетики в вирусологии возможно только при понимании принципов их работы, способов подготовки данных для анализа, критериев выбора эволюционных моделей для исследования.

Ключевые слова: *молекулярная эпидемиология, байесовская филогенетика, вирусные популяции, расследование вспышек инфекционных заболеваний, рекомбинация, модели эволюции.*

THE USE OF STATISTICAL PHYLOGENETICS IN VIROLOGY

Vakulenko Yu.A.[a,b], Lukashev A.N.[b,c], Deviatkin A.A.[c]

[a] *Martsinovsky Institute of Medical Parasitology, Tropical and Vector Borne Diseases, Sechenov First Moscow State Medical University, Moscow, Russian Federation*
[b] *Lomonosov Moscow State University, Moscow, Russian Federation*
[c] *Institute of Molecular Medicine, Sechenov First Moscow State Medical University, Moscow, Russian Federation*

Abstract. Molecular phylogenetics, particularly statistical phylogenetics, is widely used to solve the fundamental and applied problems in virology. Bayesian, or statistical, phylogenetic methods, which came into practice 10–15 years ago, markedly expanded the range of questions that can be answered based on analyzing nucleotide and amino acid sequences. An opportunity of using various evolution models allows inferring the chronology, geography and dynamics of the infection spreading. For example, analysis of globally distributed HIV group M by Bayesian methods demonstrated with a probability of 99% that the most recent common ancestor of these viruses existed in the surroundings of the city of Kinshasa (Democratic Republic of the Congo) in the early 1920s. Another study showed that H9N2 influenza virus most likely passed on to humans from wild ducks in Hong Kong in the late 1960s. In addition, using of the Bayesian analysis allows to evaluating the effect of measures taken on the development of the epidemic process. For example, it was shown retrospectively that the rate of hepatitis C virus infection cases in Egypt increased by several orders of magnitude in the mid-20th century. A sharp rise in new case rate is associated with the treatment for schistosomiasis by using non-sterile repeatedly used syringes. A set of Bayesian analysis methods has been applied in tens of thousands of researches describing various aspects of the occurrence and spread of infectious diseases in humans and animals. This was facilitated by the development and accessibility of software that implements these methods. The complexity of Bayesian phylogenetic methods imposes strict requirements on the data being analyzed. The correctness of the phylogenetic analysis data depends on various factors. For example, it is necessary to choose an evolutionary model that most adequately describes the studied objects. A mandatory step in formulating the results is the justification of the selected model. For viruses, the acquisition of genetic elements from other organisms is typical, therefore, the genomes even from closely related viruses may have non-homologous regions unsuitable for phylogenetic analysis. Another aspect is the creation of a representative dataset. Sometimes, all stages of the analysis are not indicated in publications, so that the data obtained can be interpreted ambiguously. The correct use of statistical phylogenetics methods in virology is possible only upon understanding their principles, proper methods of data preparation and evolutionary model selection criteria.

Key words: *molecular epidemiology, Bayesian phylogenetics, viral populations, investigation of infectious diseases, recombination, evolutionary models.*

Введение

Термин «филогения» был введен Эрнстом Геккелем в 1866 году в труде «Общая морфология организмов» [16]. Это слово состоит из двух корней. В оригинальной интерпретации первый корень, *phylon* (греч.), имеет три разных значения — «стебель», «ветвящаяся группа» и «линейная группа организмов». Второй корень, *genea* (греч.), трактовался как «история эволюции». Новый термин «филогения» служил заменой словосочетанию «эволюционная история групп организмов». Филогенетика — это изучение филогении, то есть изучение эволюционных взаимоотношений как на межвидовом, так и на внутривидовом уровнях. Филогенетический анализ является методом оценки эволюционных отношений организмов [14]. Филогенетическое дерево — это самый распространенный способ иллюстрации результатов филогенетического анализа [16]. Оно представляет собой ветвящуюся диаграмму и состоит из нескольких частей. Корнем является место предполагаемого общего предка всех исследуемых организмов в дереве. Если взаимоотношения устанавливаются без учета предковой формы, то такое дерево называется неукорененным. Место исследуемого организма в дереве называется листом. От каждого листа отходит ветвь. Места соединения ветвей называются узлами.

В настоящее время филогенетические методы анализа широко применяются для решения фундаментальных и прикладных задач биологии. В текущем обзоре рассматриваются возможные применения методов статистической филогенетики в вирусологии. Молекулярная филогенетика, и в частности статистическая

филогенетика вирусов, получили наиболее широкое распространение. Это связано с тем, что большая часть вирусов, вызывающих инфекционные заболевания человека, обладают РНК-содержащим геномом [49]. Механизмы репликации РНК-вирусов характеризуются высокой частотой ошибок репликации, являющихся одним из факторов высокой молекулярной изменчивости вирусов. Для РНК-вирусов было предложено эмпирическое правило, гласящее, что полимераза делает одну ошибку в каждом вновь синтезированном геноме [20]. Другим фактором высокой молекулярной изменчивости является быстрый цикл размножения и, соответственно, высокая частота «бутылочных горлышек», через которые проходит популяция вируса. Для многих вирусов цикл инфекции составляет порядка недели, а каждому следующему инфицированному организму передается ничтожно малая часть популяции вируса. Это ведет к высокой эффективности фиксации мутаций — пожалуй, даже более важной составляющей изменчивости вирусов, чем собственно возникновение мутаций при ошибках репликации генома. В результате, например, геном ВИЧ-1 изменяется за один год в среднем на 0,1–1% [8], вируса полиомиелита — на 1% [41], а коронавирусов — на 0,01–0,5% [78]. Это обуславливает высокую изменчивость, и, как следствие, необычайную адаптивную способность вирусных популяций к факторам среды.

С помощью сравнения нуклеотидных последовательностей можно определить родство между разными группами вирусов как на межвидовом, так и внутривидовом уровне. Исторически первые методы филогенетического анализа основывались на оценке генетических дистанций между последовательностями. Эти методы позволяют выявить родство вирусных последовательностей, грубо оценить скорость накопления замен и время происхождения общего предка. Оценка скорости накопления замен таким способом была бы возможна лишь в том случае, если бы эта скорость была постоянной, чего практически не бывает в вирусных популяциях. Байесовские (статистические) филогенетические методы, вошедшие в практику 10–15 лет назад, значительно расширили круг вопросов, на которые можно получить ответы, исходя из анализа нуклеотидных последовательностей. В рамках данных методов, в дополнение к моделям с заменой нуклеотидов, возможно использование концепции молекулярных часов и динамики численности популяций. Байесовский подход позволяет выбрать статистически наиболее достоверное филогенетическое дерево из всех возможных вариантов на основании расчета вероятности существования такого дерева при условии того, что имеется дополнительная информация о свойствах исследуемой вирусной популяции.

Одним из преимуществ байесовских филогенетических методов является возможность использования не только нуклеотидных последовательностей вирусов, но и другой доступной информации, например данных о географическом происхождении и времени выделения вируса. Это позволяет датировать как события в эволюции вируса, произошедшие в далеком прошлом, так и восстанавливать ход эпидемии во времени при расследовании вспышек вирусных инфекций. Так, для эпидемии лихорадки Эбола в Западной Африке (декабрь 2013 — декабрь 2015) с помощью байесовских филогенетических методов удалось выяснить происхождение и динамику передачи вируса [35]. Благодаря использованию байесовских филогенетических методов было установлено, что самая распространенная группа ВИЧ-1 возникла в 1920-м году в городе Киншаса (ДРК) [30], а группой других исследователей описано появление первого инфицированного ВИЧ человека в США: нулевой пациент был заражен примерно в 1970-м году, с вероятностью 99% жил в городе Нью-Йорке [88]. Кроме того, байесовские филогенетические методы используются для обнаружения «подозрительных» событий. Так, например, было выяснено, что циркуляция не встречавшегося с 80-х годов генотипа А энтеровируса 71 в Китае в 2008–2010 гг. оказалась связанной с попаданием лабораторного штамма в окружающую среду [83]. Оценки времени происхождения вируса, полученные с помощью байесовских филогенетических методов, хорошо согласуются с данными эпидемиологической статистики, как было показано на примере вируса бешенства [18].

Возможность использования разных моделей эволюции позволяет восстанавливать хронологию, географию и динамику распространения инфекции. В связи с этим при выборе метода проведения филогенетических исследований вирусов предпочтение отдается байесовским методам. Статьи, описывающие программное обеспечение, реализующее различные варианты и способы применения этих методов, процитированы десятки тысяч раз. Применение байесовских методов в филогенетике предполагает адекватный подбор моделей, параметров расчетов и исходных данных, что требует проведения тщательного предварительного анализа.

В данном обзоре приведено описание основных этапов проведения филогенетического анализа, принципы работы методов статистической филогенетики, способы подготовки данных для анализа, критерии выбора эволюционных моделей для исследования и примеры применения методов на практике.

Методы филогенетического анализа в вирусологии

Филогенетический анализ позволяет выяснить отношения родства между разными группами вирусов на основании сравнения нуклеотидных последовательностей. Выделяются несколько этапов выполнения филогенетического анализа: создание репрезентативной выборки, выравнивание последовательностей, выявление взаимоотношений между разными вирусами [1].

Установить отношения родства, или филогенетические взаимоотношения, между объектами изучения можно только в том случае, если существует общий предок этих объектов. Для вирусов характерно заимствование генетических элементов из других организмов (негомологичная рекомбинация), поэтому геномы даже близкородственных вирусов могут иметь негомологичные, то есть неродственные участки, непригодные для филогенетического анализа. Если имеются данные о негомологичности участков вирусного генома, выравнивание и дальнейший анализ таких участков будут грубой, но достаточно распространенной ошибкой.

Изменения в вирусном геноме могут носить характер точечных мутаций, вставок и делеций. В изучаемых последовательностях должны существовать незначительно изменившиеся участки, которые позволят установить их общее происхождение. Процесс выравнивания последовательностей заключается в поиске такого взаимного расположения разных фрагментов нуклеотидных или аминокислотных последовательностей, при котором наименее измененные участки находятся в одном и том же положении относительно друг друга. Алгоритмы выравнивания являются отдельным направлением биоинформатики [59, 72]. В практике наиболее широко распространены методы прогрессивного выравнивания (например, Clustal [37], T-Coffee [60]). При использовании этих методов сначала строится матрица сходств последовательностей, затем выполняется попарное выравнивание для самых близких пар последовательностей, после чего происходит последовательное выравнивание новых последовательностей с зафиксированными первыми попарными выравниваниями. Недостатком такого подхода является невозможность исправления ошибок, которые могут возникать на любом этапе выполнения анализа, что особенно значимо при сравнении генетически далеких друг от друга последовательностей. Этого можно избежать, используя итеративные методы прогрессивного выравнивания (например, MAFFT [45], MUSCLE [27]). Сначала методом прогрессивного выравнивания создается первое множественное выравнивание, затем в цикле выполняется несколько итераций улучшения выравнивания, что приводит, с одной стороны, к более достоверному результату, а с другой — к повышению требований к вычислительным мощностям. При сравнении менее близкородственных вирусов более точные результаты выравнивания могут быть получены при анализе менее вариабельных аминокислотных последовательностей, в то время как обычно для анализа используют нуклеотидные последовательности. Поэтому на практике часто прибегают к выравниванию нуклеотидных последовательностей на основе кодируемой ими аминокислотной последовательности (translation-based alignment). Такой способ выравнивания реализован в программе TranslatorX [2].

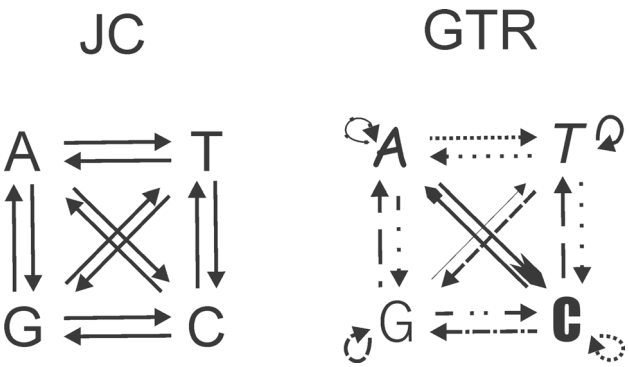

Рисунок 1. Модель Джукса–Кантора (JC) является самой простой (частной) моделью эволюции нуклеиновых кислот — предполагаются равные частоты нуклеотидов и равные вероятности возникновения замен (одинаковые стрелки). Модель GTR является самой сложной (общей) моделью эволюции нуклеиновых кислот — предполагаются разные частоты нуклеотидов (разные шрифты) и разные вероятности возникновения замен (разные прямые стрелки). Кроме того, модель GTR допускает замены типа A-на-A, T-на-T, C-на-C, G-на-G (разные круговые стрелки), которые возможны в результате нескольких последовательных замен в одной и той же позиции генома

Figure 1. The Jukes–Cantor (JC) model is the simplest (particular) model for the evolution of nucleic acids — equal nucleotide frequencies and equal probability of substitution (identical arrows) are assumed. The GTR model is the most complex (general) model for the evolution of nucleic acids — different nucleotide frequencies (different fonts) and different probabilities of substitutions (different straight arrows) are assumed. In addition, the GTR model assumes substitution A-on-A, T-on-T, C-on-C, G-on-G (different circular arrows), which are possible as a result of several consecutive replacements at the same genome position

Подавляющее большинство современных методов филогенетического анализа основаны на сопоставлении нуклеотидных или аминокислотных дистанций между последовательностями. Расчет дистанций имеет смысл только для выравненных последовательностей. В простейшем варианте дистанция рассчитывается для каждой возможной пары последовательностей из выборки как доля различающихся нуклеотидов/аминокислот между этими последовательностями. Существуют разные способы расчета дистанций, что обусловлено использованием разных моделей эволюции нуклеотидных или аминокислотных последовательностей. Так, например, наиболее частная модель Джукса–Кантора (Jukes and Cantor 1969, JC) [42] (рис. 1) подразумевает равную частоту всех нуклеотидов и равную вероятность возникновения всех замен в последовательностях. Другие модели в разной степени учитывают эти два ключевых параметра (4 частоты нуклеотидов, 16 типов замен). На рис. 1 представлены самая простая (JC) и самая сложная (GTR) модели эволюции нуклеиновых кислот. Высокое разнообразие моделей эволюции нуклеотидных последовательностей вызвано тем, что по биологическим причинам для организмов характерны разные варианты отношений частот нуклеотидов и скоростей накопления замен. Применимость модели к конкретному набору последовательностей можно оценить при помощи вычисления соответствующих показателей совместимости, например информационного критерия Акаике (AIC, Akaike Information Criterion) [3] или байесовского информационного критерия (BIC, Bayesian Information Criterion) [68], реализованных во многих пакетах для филогенетического анализа. Для расчета аминокислотных дистанций чаще всего используют эмпирические модели, основанные на информации о наблюдаемых частотах аминокислот и вероятностях каждой возможной замены аминокислоты [6].

Наиболее часто используемым методом построения филогенетических деревьев на основе матрицы дистанций является метод присоединения соседей (neighbor joining, NJ) [67]. Принцип его работы заключается в поэтапном поиске пары соседних последовательностей, которые минимизируют общую длину ветвей дерева. Другими словами, сначала идентифицируют ту пару последовательностей, при объединении которой в один таксон длина всех ветвей дерева минимальна. Затем рассчитывают дистанцию между оставшимися последовательностями и общим узлом объединенной пары. Такой цикл повторяют, пока все последовательности не присоединят поэтапно. Филогенетическая реконструкция при помощи метода присоединения соседей является одним из самых быстрых вариантов выполнения анализа [50].

Другой класс методов филогенетического анализа применяется для расчета наиболее вероятные значения параметров эволюционной модели и топологии филогенетического дерева. К этой категории относятся байесовские методы статистического анализа, которые подробно разбираются в следующем разделе, а также построение филогенетического дерева по принципу максимального правдоподобия (Maximum Likelihood, ML) [31]. Этот метод основан на поиске такого филогенетического дерева, при котором вероятность наблюдать исследуемое выравнивание нуклеотидных или аминокислотных последовательностей максимальна. Применение группы таких методов требует значительных вычислительных мощностей.

В ряде работ проводилось сравнение разных методов филогенетического анализа на одинаковых данных. Топология деревьев, полученных разными методами, сопоставлялась с топологией истинного дерева. В публикации Джозефа Фельзенштейна, первым описавшего применение метода максимального правдоподобия, показано, что в большинстве случаев ML демонстрировал правильную топологию дерева чаще других методов. Тем не менее при анализе коротких последовательностей метод NJ оказался предпочтительнее [50]. В аналогичной работе, выполненной Масатоси Нэи, автором метода NJ, делается вывод, что оба метода имеют сопоставимую точность в контексте определения верной топологии дерева [81].

Основные принципы работы байесовских методов статистической филогенетики

Основу байесовских методов статистической филогенетики составляет теорема, сформулированная Томасом Байесом. Теорема Байеса позволяет оценить неизвестные параметры исследуемого явления на основании наблюдений. Результат такой оценки называется байесовским выводом. Например, известно, что в ящике находится какое-то животное. Необходимо определить, какое животное спрятано. Если из ящика слышно мяуканье, то можно сделать вывод, что в ящике, скорее всего, находится кот. Другими словами, сведения об исследуемом явлении, известные априори (то есть до опыта — «в ящике животное», «чаще всего мяукают коты»), можно уточнить на основании знаний, полученных после проведения опыта (апостериори — «мяукает»). Этот подход отличается от обычной описательной статистики, которая

отвечает на вопросы типа «Если в ящике кот, с какой вероятностью будет слышно мяуканье?» и позволяет объединять в одном результате много итераций опыта.

Применение методов байесовского анализа в филогенетике впервые было предложено примерно 25 лет назад [64]. В случае байесовской филогенетики наблюдаемыми данными являются генетические последовательности исследуемых вирусов, а эволюционная модель для таких последовательностей включает модель замен, модель молекулярных часов, описывающую скорости накопления замен на ветвях дерева, филогенетическое дерево и модель, описывающую его ветвление (демографическая модель). Таким образом, результатом байесовского вывода является наиболее вероятное филогенетическое дерево и совокупность значений параметров эволюционной модели при условии имеющегося выравнивания генетических последовательностей. Помимо генетических последовательностей, в качестве априорных данных, то есть известных независимо от эксперимента, можно использовать географические координаты мест выделения последовательностей, время сбора содержащего последовательности материала, информацию о животном — хозяине вируса и т. п. Благодаря этому можно определять хронологию основных событий в эволюции вируса, воспроизводить историю географического распространения как патогена в целом, так и отдельных групп вируса.

Ключевой особенностью байесовских методов является возможность одновременного изучения целостной модели эволюции, включающей все ее параметры. Данное свойство выгодно отличает этот класс алгоритмов от «классических» методов молекулярной филогении, где используется последовательный расчет эволюционных параметров.

Для расчета апостериорных распределений параметров модели в байесовских методах используются алгоритмы Монте-Карло с цепями Маркова (MCMC, Markov Chain Monte Carlo) [58]. Алгоритм MCMC выполняется в несколько этапов. Для каждой переменной (параметра анализа) исследуется пространство вероятностного распределения. Как правило, первоначальное состояние переменной (например, скорость замен A в G) выбирается произвольно. Затем рассчитывается вероятность того, что это значение справедливо при заданных условиях. После этого случайным образом изменяется состояние переменной. Рассчитывается вероятность справедливости нового состояния, а затем сравнивается с предыдущим значением. Если вероятность справедливости второго состояния выше, чем у первого, будет подтвержден переход переменной из первого состояния во второе, если ниже — переход маловероятен. Такой процесс повторяется миллионы раз, благодаря чему осуществляется поиск наиболее вероятного значения каждой исследуемой переменной. Естественно, вследствие многократного выполнения шагов в пространстве возможных состояний расчет наиболее вероятного значения для всех исследуемых параметров является вычислительно затратным процессом. Следует отметить, что с помощью байесовских методов вычисляется не конкретное значение каждой переменной, а вероятностное распределение значений переменной, для которого возможно определить величину медианы и доверительного интервала.

Выбор модели для филогенетического анализа

Эволюционную модель можно рассматривать как способ формализации наблюдаемых закономерностей возникновения исследуемых последовательностей. Эволюционные модели различаются количеством учитываемых параметров. Более сложные модели лучше описывают данные, но избыточное количество параметров может приводить к большим вычислительным нагрузкам, а также к лишенным биологического смысла результатам. При выборе модели необходимо сохранять баланс между числом параметров и ее способностью описывать биологические данные.

Результаты любого филогенетического анализа зависят от допущений использованной модели. Если модель плохо описывает биологические данные, то результаты могут оказаться ошибочными. Например, филогенетическая модель, предполагающая постоянную скорость накопления замен в разных участках последовательностей, приводит к некорректным результатам, когда скорость накопления замен неравномерна. При этом искаженный результат может получиться, даже если остальные параметры модели верны [33].

На практике только на основании полученных результатов филогенетического анализа невозможно определить, являются они верными или нет. Для этого требуется, в частности, анализ соответствия выбранной эволюционной модели исследуемым объектам. Обязательным этапом при формулировании результатов является обоснование выбранной модели. Если в публикации не приведено описание этапа выбора модели, то приведенные в ней результаты могут трактоваться неоднозначно.

Байесовские филогенетические методы позволяют использовать филогенетические модели разной сложности. Универсальной модели, подходящей для любого случая, не существует.

В байесовской филогенетике для сравнения моделей применяется подход, основанный на вычислении коэффициента Байеса (Bayes Factor, BF) [70, 80]. BF — это отношение вероятностей данных при условии выбора разных моделей,

$$BF = \frac{P(D|M1)}{P(D|M2)},$$

где D — это данные; M1 — первая модель; M2 — вторая модель; P(D|M1) и P(D|M2) — вероятности наблюдения данных при условии использования первой модели и второй модели соответственно [40].

Для выбора модели эволюции принято использовать десятичный логарифм коэффициента Байеса — log BF. Если log(BF) принимает значение от 0 до 1, то значимой разницы между моделями нет. Значения log(BF) более 1 свидетельствуют о превосходстве M1 над M2 [44], однако на практике достаточно значимой разницей между моделями считают значения log(BF) больше 10. Сравнение моделей эволюции при помощи log(BF) является вычислительно чрезвычайно затратным и требует значительного объема «ручной» работы по перебору параметров.

Для приближенного вычисления вероятности P(D|M) достаточной точностью отличаются методы path sampling (PS) [52], stepping stone (SS) [89] и generalized stepping stone (GSS) [29]. Они реализованы в программах BEAST [24] и BEAST2 [11]. Недостатком данных методов является необходимость подбора параметров при каждом исследовании и большая вычислительная сложность. Альтернативой данным методам является алгоритм вложенного выбора (nested sampling, NS) [66, 71], который, в отличие от других, позволяет вычислять стандартное отклонение для P(D|M). Он реализован в программе BEAST2 [66].

Выбор модели замен. Выбор модели замен осуществляется во время подготовки к исследованию при помощи отдельных биоинформатических инструментов (например, jModelTest, ModelGenerator [46], PartitionFinder2 [51]). Следует отметить, что скорость накопления мутаций для нуклеотидов, стоящих в разных позициях кодона, может отличаться. По этой причине при анализе белок-кодирующих последовательностей целесообразно подбирать наиболее подходящую модель замен для каждой позиции кодона по отдельности. В противном случае топология филогенетического дерева может оказаться некорректной [43]. Программа PartitionFinder2 позволяет найти оптимальную модель замен для каждой позиции кодона. Другой подход для выбора модели замен — bModelTest [12], ставший частью программного пакета BEAST2: в этом случае модель замен используется в качестве параметра байесовского вывода. В процессе анализа цепь Маркова может переходить с одной модели на другую. При этом, чем дольше цепь находится на одной из моделей замен, тем такая модель предпочтительнее. Разные модели поддерживаются в разной степени, что учитывают при итоговой оценке всех параметров.

Выбор модели молекулярных часов. Согласно концепции молекулярных часов, предложенной в 1960-х годах [56, 94, 95], замены в родственных нуклеотидных или аминокислотных последовательностях происходят с практически постоянной скоростью. Если раньше знания о времени эволюционных событий основывались только на ископаемых данных, то благодаря этой концепции стало возможно определять время существования общих предков организмов, используя генетические последовательности.

В целом существующие модели молекулярных часов можно классифицировать на основе количества параметров, описывающих скорости накопления замен на ветвях [39]. Самая первая и простейшая модель молекулярных часов предполагала постоянную скорость эволюции для разных таксонов [94]. Это означает, что генетическое расстояние между последовательностями организмов линейно зависит от времени с момента расхождения организмов. Данную модель также называют строгими молекулярными часами (strict clock). В результате увеличения числа генетических последовательностей было получено немало свидетельств того, что скорости накопления замен в разных группах организмов сильно варьируются. В таком случае простая модель, предполагающая постоянную скорость замен, недостаточно точно отражает реальные события [86]. В модели ослабленных молекулярных часов (relaxed molecular clock) предполагается, что на каждой ветви филогенетического дерева свое значение скорости накопления замен, то есть число параметров равно числу ветвей дерева [22]. Третья группа моделей допускает число различных скоростей замен в диапазоне между единицей (то есть строгие часы) и числом ветвей дерева (ослабленные часы). К таким моделям относится модель локальных молекулярных часов [25], предполагающая, что скорости внутри определенного числа клад филогенетического дерева постоянные, но отличаются между кладами. Эту модель целесообразно использовать при изучении нескольких таксонов, предполагая, что внутри таксонов скорость накопления замен постоянна.

Выбор модели молекулярных часов проводят при помощи вычисления коэффициента Байеса. Если используемая модель плохо описывает данные, то это критически сказывается на полученных сведениях о времени расхождения групп вирусов [87]. При сравнении разных

моделей необходимо включать в рассмотрение модель строгих часов. Если гипотезу о том, что скорость накопления замен постоянна, нельзя отвергнуть, строгие молекулярные часы предпочтительнее более сложных моделей молекулярных часов, так как у данной модели самое малое число параметров и, соответственно, меньше время расчетов и доверительные интервалы результатов.

Выбор модели динамики численности популяции. Одной из частей эволюционной модели в байесовских филогенетических методах является модель ветвления дерева (tree prior). Такие модели описывают вероятностные законы образования новых клад в дереве, а также устанавливают априорные распределения на топологию и длины ветвей филогенетического дерева. Ветвление филогенетического дерева косвенно может говорить о динамике численности популяции. Как правило, в программах, реализующих байесовский филогенетический вывод, используются три стохастические модели ветвления: коалесцентные модели, модель рождения–гибели и модель видообразования Юла [92].

Теория слияния, или коалесценции (coalescence), предложенная Кингманом в 1982 г. [48], является способом описания того, как популяционные процессы определяют форму генеалогии исследуемых последовательностей (топологию дерева). При использовании теории коалесценции чаще всего применяют три демографические модели. Самая простая модель предполагает постоянный размер популяции (constant size). Модель экспоненциального роста (exponential growth) [36] подразумевает экспоненциальный рост размера популяции на протяжении всего времени. Модель байесовского графика горизонта (Bayesian skyline plot) [21] позволяет реконструировать апостериорное распределение эффективного размера популяции во времени.

Ключевой величиной, которая характеризует распространение инфекции (и репродукцию любого организма в общем виде), является базовое число репродукции R_0, то есть ожидаемое количество вторичных инфекций, вызванных одним первичным случаем в полностью восприимчивой популяции [4]. Эта величина определяет скорость распространения вируса без вмешательства и то, какие вмешательства необходимы для сдерживания эпидемии и какой размер иммунной прослойки (доля вакцинированных или переболевших) позволит контролировать распространение вируса [47]. При помощи методов статистической филогенетики в вирусологии можно определить скорости изменения размера популяции по наблюдаемому набору последовательностей генов, отобранных из этой популяции [61]. Применение коалесцентных моделей оправдано при значениях R_0, близких к 1 (что редко встречается в микробиологии), и плохо подходит для воспроизводства филодинамики вирусной популяции, если экспоненциальный рост размера популяции происходит стохастично [76]. В таких случаях следует использовать модель «рождение–гибель» (birth–death) [75, 77]. Модель «рождение–гибель» позволяет вычислить скорости появления и вымирания групп в популяции [60]. В этой модели присутствуют два параметра: скорость ветвления и скорость вымирания группы. Событию «рождение» соответствует ветвление филогенетического дерева, а событию «гибель» — лист дерева. В программе BEAST2 доступны расширения коалесцентной модели и модели рождения и гибели, включающие эпидемиологическую модель SIR (Susceptible–Infected–Recovered) [75, 84]. Модель SIR описывает динамику групп восприимчивых, инфицированных и выздоровевших индивидов при эпидемии. Также в BEAST2 доступна модель рождения и смерти BDSKY [74], позволяющая оценить эффективное число репродукции R_e, которое отражает динамику R в течение вспышки заболевания. Изменение эффективного числа репродукции позволяет судить об эффективности противоэпидемических мер: если значение R_e ниже 1, то число случаев заболеваний будет снижаться.

Модель Юла описывает процесс видообразования [92]. Единственный параметр этой модели — скорость возникновения нового вида — определяет скорость, с которой ветвь разделяется на две другие ветви. При филогенетическом анализе набора видов целесообразно включить такую модель в перечень моделей для сравнения в первую очередь, поскольку она не требует сложных вычислений.

С одной стороны, выбор демографической модели зависит от цели исследования. Например, при исследовании видообразования используют модель Юла и модель рождения и смерти. В случае исследования эволюции патогенов во время эпидемии целесообразно использовать коалесцентную модель или модель рождения и смерти, включающую эпидемиологическую модель. С другой стороны, выбранный процесс ветвления определяет топологию дерева и другие его параметры [15]. Например, при использовании коалесцентной модели с постоянной численностью популяции возраст корня дерева будет старше, чем у дерева, построенного при условии экспоненциального роста численности. Таким образом, перед анализом необходимо выбрать несколько моделей, потенциально подходящих для цели исследования, и сравнить их с помощью вычисления коэффициента Байеса.

Данные, используемые в байесовском филогенетическом анализе, и особенности их подготовки

Главным источником данных при филогенетическом анализе являются выравнивания нуклеотидных или аминокислотных последовательностей вирусов. Возможно использовать выравнивания как полных геномов, так и одного или нескольких вирусных белков. Анализ выравниваний длинных геномных фрагментов следует проводить с особенной осторожностью, так как для большинства вирусов характерна рекомбинация, или горизонтальный обмен генетической информацией. В результате негомологичной рекомбинации (то есть с ДНК/РНК других организмов) фрагменты генома одного вируса могут иметь разное происхождение, что исключает возможность корректного выравнивания последовательностей РНК/ДНК. В случае гомологичной рекомбинации фрагменты генома будут иметь разную филогенетическую историю, что также затрудняет проведение филогенетического анализа (рис. 2). Таким образом, оказывается невозможным определить единственное положение рекомбинанта на дереве и другие параметры эволюции — различные участки генома будут иметь разные эволюционные события (рис. 2). По этой причине вклад рекомбинационных событий в эволюцию исследуемых вирусов требует внимательного изучения до начала анализа [82].

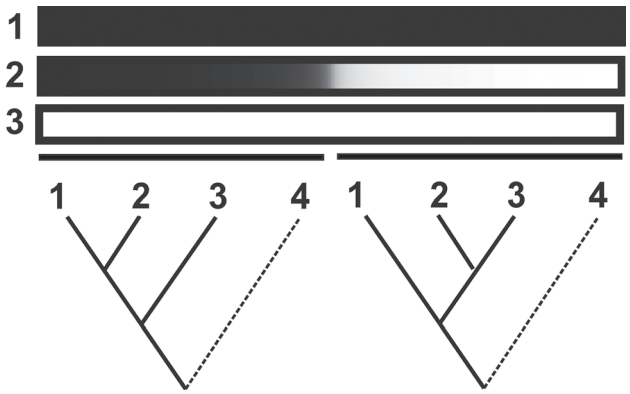

Рисунок 2. Влияние рекомбинации на топологию филогенетического дерева. Последовательность 2 — результат рекомбинации последовательностей 1 и 3. Одинаковым цветом выделены похожие участки последовательностей. Укорененные филогенетические деревья, построенные по разным участкам генома, иллюстрируют отношения между последовательностями. Топология деревьев изменяется из-за использования участков генома разного происхождения [34]

Figure 2. The effect of recombination on the topology of the phylogenetic tree. The sequence 2 is the result of the recombination of sequences 1 and 3. Similar fragments of sequences are highlighted in the same color. Rooted phylogenetic trees built on different parts of the genome illustrate the relationship between sequences. Tree topology is changing due to the use of genome fragments of different origin [34]

В целом, перед филогенетическим анализом необходимо исключить события рекомбинации с помощью набора алгоритмов для ее детекции, реализованных, например, в программе RDP4 [57].

Байесовские филогенетические методы позволяют строить филогенетические деревья во временно́й шкале. Для этого, помимо нуклеотидных последовательностей, используют даты выделения вирусов [22, 77]. Использование дат выделения вируссодержащего материала возможно только в том случае, если исследуемая выборка последовательностей содержит временной сигнал. Это значит, что между двумя последовательностями изолятов, выделенных из вирусной популяции в разное время, должно быть измеримое различие нуклеотидных или аминокислотных замен [23, 63]. Для изучения временной структуры выборки последовательностей используют подход, основанный на регрессионном анализе. Сначала строят укорененное филогенетическое дерево любым из «классических» способов, описанных ранее (например, методом максимального правдоподобия). Длины ветвей такого дерева должны отражать количество генетических различий. Затем строят линейную регрессию между датами выделения вирусов и расстоянием от корня до листьев, соответствующих вирусным последовательностям. При наличии временной структуры наблюдается корреляция между данными величинами, а отсутствие линейного тренда говорит о том, что данные не подходят для байесовского филогенетического анализа с использованием моделей молекулярных часов. Анализ временно́й структуры популяции реализован в программе TempEst [63].

Разрешение вырожденных нуклеотидов. Качество выравнивания нуклеотидных или аминокислотных последовательностей играет решающую роль в байесовском филогенетическом анализе. При секвенировании вирусных последовательностей некоторые позиции могут иметь неоднозначную расшифровку. Неопределенность позиций связана как с техническими ограничениями при секвенировании, так и с тем, что РНК-вирусы существуют в виде квазивидов, когда каждый образец вируса представляет собой облако почти идентичных геномов, которые отличаются несколькими заменами [19]. В резуль-

тате некоторые позиции в вирусных последовательностях, доступных в базах данных нуклеотидов (например, GenBank), имеют неоднозначные значения, обозначенные кодами IUPAC (Y для T/C, R для G/A и т. д.). Иногда такие последовательности существенно искажают результаты филогенетического анализа [82]. Если в последовательности много вырожденных нуклеотидов, то это, скорее всего, является результатом низкого качества секвенирования, и такую последовательность следует исключить из набора данных. Единичные вырожденные нуклеотиды можно разрешать по консенсусу. Алгоритм для автоматизации этого процесса был предложен в статье Vakulenko Yu. и соавт. [82] и доступен по адресу https://github.com/v-julia/resolve_ambiguous. Также желательно вручную проверять выравнивание последовательностей на наличие ошибок секвенирования, например сдвигов открытых рамок считывания, так как такие ошибки влияют на результаты анализа.

Выбор последовательностей при полномасштабном анализе

Для выполнения филогенетического анализа необходимо соблюдение условия репрезентативности исследуемой выборки. Количество доступных последовательностей в базах данных значительно различается для разных видов вирусов. Для вирусов, имеющих большое медицинское значение, доступны десятки тысяч последовательностей. Для менее изученных вирусов число последовательностей на несколько порядков меньше. Чаще всего последовательности получали в результате исследований вспышек вируса или в ходе надзора в конкретном регионе. В результате этого распределение данных по годам и регионам выделения неравномерно даже для хорошо описанных вирусов. Таким образом, доступное число последовательностей редко соответствует реальному разнообразию вирусной популяции.

Анализ тысяч последовательностей вирусов затруднителен из-за вычислительной сложности. Объем выборки для анализа может быть целенаправленно сокращен при условии сохранения генетического разнообразия и проведения контроля влияния сокращения на результат исследований. Одним из способов формирования выборки может быть удаление почти идентичных последовательностей. Такая возможность реализована в различных программах: Jalview [85], CD-HIT [32], UCLUST [28], skipredundant из пакета программ EMBOSS [65]. При этом сохраняется максимальное генетическое разнообразие последовательностей. Кроме того, вычисленные скорости замен и возрасты корней деревьев совпадают чаще, чем при случайном выборе последовательностей. С другой стороны, удаление похожих последовательностей может стать причиной артефактов — например, при расчете динамики численности вирусной популяции [82].

При другом способе уменьшения размера выборки среди последовательностей выделяют группы, которые соответствуют отдельным исследованиям. Затем в итоговую выборку включают все последовательности из небольших исследований и небольшую часть последовательностей из крупных научных исследований, в которых обычно изучается большое количество последовательностей из одного региона. При таком способе вероятность сохранения редких последовательностей выше, и результаты анализа искажаются в меньшей степени [82].

Эффект от использования выборок разного объема, оказываемый на результаты исследования, можно оценить путем симулирования. Для энтеровируса A71 нами была смоделирована ситуация, при которой известна только часть из почти 5000 опубликованных последовательностей [83]. При размере выборки 150–400 последовательностей результаты расчета скорости накопления замен и возраста корней филогенетических деревьев различались в несколько раз. Аналогичные результаты были получены при филогенетическом анализе вируса клещевого энцефалита [17]. Таким образом, неполноценность выборок влияет на оценки ключевых эволюционных параметров в случае анализа небольших наборов данных. При этом размер возможной ошибки в принципе нельзя оценить, если имеется лишь небольшая выборка известных последовательностей вируса, и результат следует интерпретировать с особой осторожностью.

Примеры применения байесовских филогенетических методов на практике

Байесовские филогенетические методы позволяют строить филогенетические деревья во временной шкале. Это делает возможным восстановление хронологии основных событий в эволюции вируса. Следует отметить, что подобные оценки находятся в соответствии с данными эпидемиологической статистики. Например, в Центральной Европе бешенство среди лисиц практически не регистрировалось перед Второй мировой войной. При этом в степной зоне европейской части СССР в середине 1930-х годов были описаны массовые случаи гибели лисиц по неустановленной причине [9]. В 1939-м году на территории Польши произошла вспышка бешенства среди этих животных [5]. Впоследствии зарегистрировали миграцию бешеных лисиц в западном направ-

лении со средней скоростью 30–60 км в год. В 1968 году бешенство среди лисиц появилось во Франции. Из этих наблюдений можно сделать вывод о том, что бешенство среди лисиц в Западной Европе возникло примерно в годы Второй мировой войны. Согласно результатам байесовского филогенетического анализа, данная группа вирусов бешенства возникла в 1939 году [18].

В ряде исследований используют филогенетические деревья для объединения генетических данных и информации о распределении популяции в пространстве [7]. Такой анализ, позволяющий расследовать вспышки инфекционных заболеваний, называется филогеографическим. Разные модели филогеографии реализованы в программах BEAST [79] и BEAST2 [13]. В зависимости от задачи и доступных данных возможно использование как точных координат мест получения вирус-содержащего материала [10, 54], так и названий населенных пунктов [53, 55]. Например, при анализе последовательностей ВИЧ глобально распространенной группы M байесовскими методами филогеографического анализа было показано, что последний общий предок этих вирусов с вероятностью 99% возник в окрестностях города Киншаса (Демократическая Республика Конго) [30].

В байесовский филогенетический анализ можно ввести и другие характеристики исследуемых вирусов. Существуют примеры использования данных о хозяевах вирусов для изучения роли смены хозяев в эволюции вируса птичьего гриппа [91] и MERS-CoV [26]. Так, например, серотип вируса гриппа H9N2, скорее всего, возник у диких уток [91]. Судя по всему, это случилось в Гонконге в конце 60-х годов XX века. После этого происходило несколько независимых заносов вируса в популяцию домашней курицы, от которых заражались люди. Считается, что этот серотип вируса гриппа потенциально может вызвать пандемию [73].

Методы статистической филогенетики могут использоваться и для расследования неестественных событий в эволюции вирусов (естественное или умышленное внесение вируса в циркуляцию, антропогенный перенос вирусов). В 2008–2015 гг. в Китае произошла вспышка энтеровирусного везикулярного стоматита. Всего заболело почти 14 млн человек. По некоторым оценкам, 40% случаев были вызваны энтеровирусом типа A71 (EV-A71) [90]. При этом из 2308 летальных случаев 2136 были вызваны EV-A71. Часть вирусов EV-A71, выделенных в ходе этой эпидемии, принадлежала к генотипу A [93]. Ближайшим родственником таких вирусов оказался культивируемый в лабораториях «прототипный» штамм BrCr, выделенный в США в начале 1970-х гг., который не встречается в циркуляции. При помощи байесовского анализа показали, что скорость возникновения замен в вирусах генотипа A EV-A71 статистически значимо отличается от среднего значения для EV-A71 в целом [83]. Это указывает на реинтродукцию вируса BrCr в окружающую среду из лаборатории.

При изучении эволюции и эпидемиологии РНК-содержащих вирусов необходимо учитывать возможные изменения размера популяции патогена, зависящие от эффективного числа репродукции — Re. Рост размера популяции является маркером «успешного» распространения инфекции и повышения генетического разнообразия вируса. Байесовские методы филогенетического анализа позволяют воспроизвести динамику этого параметра во времени [21, 38, 74]. Так, например, было показано, что число заражений вирусом гепатита С (ВГС) в Египте резко увеличилось на несколько порядков в середине XX века [62]. После этого происходило плавное незначительное снижение эффективного размера популяции ВГС. Резкий рост новых случаев связывают с началом лечения шистосомоза. Лекарство вводили при помощи уколов, нестерильные шприцы использовали многократно [21]. Плавное незначительное снижение размера популяции ВГС объясняется постепенным замещением парентеральных антишистосомальных препаратов на лекарства, употребляемые перорально. Таким образом, изучение динамики размера вирусной популяции позволяет оценить влияние среды обитания патогена на распространение инфекции. Под «средой обитания патогена» мы подразумеваем совокупность большого числа социальных, географических и эпидемиологических факторов. Например, в различных сообществах люди предпочитают разную дистанцию. Это влияет на распространение инфекций, передающихся воздушно-капельным путем. Климатические условия серьезно различаются в разных регионах. Считается, что сезонность в распространении гриппа по большей части обусловлена колебаниями абсолютной влажности [69]. Кроме того, на вероятность передачи инфекции влияют принятые регуляторными организациями меры эпидемиологического контроля. Эффект каждого из подобных факторов на распространение инфекции можно оценить при помощи изучения динамики размера вирусной популяции.

Заключение

Байесовские методы филогенетического анализа нашли широкое применение в изучении эволюции вирусов. Этот набор методов был использован в десятках тысяч исследований,

описывающих разные аспекты возникновения и распространения инфекционных заболеваний человека и животных. Относительно недавно появились программные пакеты, позволяющие не обладающему специальной биоинформатической компетенцией исследователю восстановить хронологию и географию распространения вирусов при помощи методов байесовской филогенетики. Тем не менее корректные результаты можно получить только в случае грамотного выбора всех влияющих на анализ параметров.

Список литературы/References

1. Лукашов В.В. Молекулярная эволюция и филогенетический анализ. М.: БИНОМ. Лаборатория знаний, 2009. 256 с. [Lukashov V.V. Molecular evolution and phylogenetic analysis. *Moscow: BINOM. Laboratoriia Znaniy, 2009. 256 p. (In Russ.)*]
2. Abascal F., Zardoya R., Telford M.J. TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic Acids Res., 2010, vol. 38: W7-13. doi: 10.1093/nar/gkq291*
3. Akaike H. A new look at the statistical model identification. *IEEE Trans. Automat. Contr., 1974, vol. 19, no. 6, pp. 716–723. doi: 10.1109/TAC.1974.1100705*
4. Anderson R.M., May R.M. Population biology of infectious diseases: part I. *Nature, 1979, vol. 280, no. 5721, pp. 361–367. doi: 10.1038/280361a0*
5. Anderson R.M., May R.M., Jackson H.C., Smith A.M. Population dynamics of fox rabies in Europe. *Nature, 1981, vol. 289, no. 5800, pp. 765–771. doi: 10.1038/289765a0*
6. Arenas M. Trends in substitution models of molecular evolution. *Front. Genet, 2015, vol. 6: 319. doi: 10.3389/fgene.2015.00319*
7. Avise J.C. Phylogeography: retrospect and prospect. *J. Biogeogr., 2009, vol. 36, no. 1, pp. 3–15. doi: 10.1111/j.1365-2699.2008.02032.x*
8. Berry I.M., Ribeiro R., Kothari M., Athreya G., Daniels M., Lee H.Y., Bruno W., Leitner T. Unequal evolutionary rates in the human immunodeficiency virus type 1 (HIV-1) pandemic: the evolutionary rate of HIV-1 slows down when the epidemic rate increases. *J. Virol., 2007, vol. 81, no. 19, pp. 10625–10635. doi: 10.1128/jvi.00985-07*
9. Botvinkin A., Kosenko M. Rabies in the european parts of Russia, Belarus and Ukraine. In: Historical perspective of rabies in Europe and the Mediterranean Basin: a testament to rabies. *OIE: Paris, France, 2004, pp. 47–65.*
10. Bouckaert R. Phylogeography by diffusion on a sphere: whole world phylogeography. *Peer J., 2016, vol. 4: e2406. doi: 10.7717/peerj.2406*
11. Bouckaert R., Heled J., Kühnert D., Vaughan T., Wu C.H., Xie D., Suchard M.A., Rambaut A., Drummond A.J. BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol., 2014, vol. 10, no. 4, pp. 1–6. doi: 10.1371/journal.pcbi.1003537*
12. Bouckaert R.R., Drummond A.J. bModelTest: Bayesian phylogenetic site model averaging and model comparison. *BMC Evol. Biol., 2017, vol. 17, no. 1, pp. 1–11. doi: 10.1186/s12862-017-0890-6*
13. Bouckaert R., Vaughan T.G., Barido-Sottani J., Duchêne S., Fourment M., Gavryushkina A., Heled J., Jones G., Kühnert D., Maio De N., Matschiner M., Mendes F.K., Müller N.F., Ogilvie H.A., Plessis du L., Popinga A., Rambaut A., Rasmussen D., Siveroni I., Suchard M.A., Wu C.H., Xie D., Zhang Ch., Stadler T., Drummond A.J. BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLOS Comput. Biol., 2019, vol. 15, no. 4: e1006650. doi: 10.1371/journal.pcbi.1006650*
14. Choudhuri S. Phylogenetic Analysis. In: Bioinformatics for Beginners. *Elsevier, 2014. pp 209–218.*
15. Colijn C., Plazzotta G. A Metric on phylogenetic tree shapes. *Syst. Biol., 2018, vol. 67, no. 1, pp. 113–126. doi: 10.1093/sysbio/syx046*
16. Dayrat B. The roots of phylogeny: how did haeckel build his trees? *Syst. Biol., 2003, vol. 52, no. 4, pp. 515–527. doi: 10.1080/10635150390218277*
17. Deviatkin A.A., Kholodilov I.S., Vakulenko Yu.A., Karganova G.G., Lukashev A.N. Tick-Borne encephalitis virus: an emerging ancient zoonosis? *Viruses, 2020, vol. 12, no. 2: 247. doi: 10.3390/v12020247*
18. Deviatkin A.A., Lukashev A.N., Poleshchuk E.M., Dedkov V.G., Tkachev S.E., Sidorov G.N., Karganova G.G., Galkina I.V., Shchelkanov M.Yu., Shipulin G.A. The phylodynamics of the rabies virus in the Russian Federation. *PLoS One, 2017, vol. 12, no. 2: e0171855. doi: 10.1371/journal.pone.0171855*
19. Domingo E., Sheldon J., Perales C. Viral quasispecies evolution. *Microbiol. Mol. Biol. Rev., 2012, vol. 76, no. 2, pp. 159–216. doi: 10.1128/MMBR.05023-11*
20. Drake J.W., Holland J.J. Mutation rates among RNA viruses. *Proc. Natl. Acad. Sci., 1999, vol. 96, no. 24, pp. 13910–13913. doi: 10.1073/pnas.96.24.13910*
21. Drummond A.J. Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol. Biol. Evol., 2005, vol. 22, no. 5, pp. 1185–1192. doi: 10.1093/molbev/msi103*
22. Drummond A.J., Ho S.Y.W., Phillips M.J., Rambaut A. Relaxed phylogenetics and dating with confidence. *PLoS Biol., 2006, vol. 4, no. 5: e88. doi: 10.1371/journal.pbio.0040088*
23. Drummond A.J., Pybus O.G., Rambaut A., Roald F., Rodrigo A.G. Measurably evolving populations. *Trends Ecol. Evol., 2003, vol. 18, no. 9, pp. 481–488. doi: 10.1016/S0169-5347(03)00216-7*
24. Drummond A.J., Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol., 2007, vol. 7, no. 1: 214. doi: 10.1186/1471-2148-7-214*
25. Drummond A.J., Suchard M.A. Bayesian random local clocks, or one rate to rule them all. *BMC Biol., 2010, vol. 8, no. 1: 114. doi: 10.1186/1741-7007-8-114*
26. Dudas G., Carvalho L.M., Rambaut A., Bedford T. MERS-CoV spillover at the camel-human interface. *Elife, 2018, vol. 7: e31257. doi: 10.7554/eLife.31257*

27. Edgar R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res., 2004, vol. 32, no. 5, pp. 1792–1797. doi: 10.1093/nar/gkh340*
28. Edgar R.C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics, 2010, vol. 26, no. 19, pp. 2460–2461. doi: 10.1093/bioinformatics/btq461*
29. Fan Y., Wu R., Chen M.-H., Kuo L., Lewis P.O. Choosing among partition models in Bayesian phylogenetics. *Mol. Biol. Evol., 2011, vol. 28, no. 1, pp. 523–532. doi: 10.1093/molbev/msq224*
30. Faria N.R., Rambaut A., Suchard M.A., Baele G., Bedford T., Ward M.J., Tatem A.J., Sousa J.D., Arinaminpathy N., Pépin J., Posada D., Peeters M., Pybus O.G., Lemey P. The early spread and epidemic ignition of HIV-1 in human populations. *Science, 2014, vol. 346, no. 6205, pp. 56–61. doi: 10.1126/science.1256739*
31. Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol., 1981, vol. 17, no. 6, pp. 368–376. doi: 10.1007/BF01734359*
32. Fu L., Niu B., Zhu Z., Wu S., Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics, 2012, vol. 28, no. 23, pp. 3150–3152. doi: 10.1093/bioinformatics/bts565*
33. Gaut B.S., Lewis P.O. Success of maximum likelihood phylogeny inference in the four-taxon case. *Mol. Biol. Evol., 1995, vol. 12, no. 1, pp. 152–162. doi: 10.1093/oxfordjournals.molbev.a040183*
34. Gibbs M.J., Armstrong J.S., Gibbs A.J. Sister-scanning: a Monte Carlo procedure for assessing signals in rebombinant sequences. *Bioinformatics, 2000, vol. 16, no. 7, pp. 573–582. doi: 10.1093/bioinformatics/16.7.573*
35. Gire S.K., Goba A., Andersen K.G., Sealfon R.S., Park D.J., Kanneh L., Jalloh S., Momoh M., Fullah M., Dudas G., Wohl S., Moses L.M., Yozwiak N.L., Winnicki S., Matranga C.B., Malboeuf C.M., Qu J., Gladden A.D., Schaffner S.F., Yang X., Jiang P.P., Nekoui M., Colubri A., Coomber M.R., Fonnie M., Moigboi A., Gbakie M., Kamara F.K., Tucker V., Konuwa E., Saffa S., Sellu J., Jalloh A.A., Kovoma A., Koninga J., Mustapha I., Kargbo K., Foday M., Yillah M., Kanneh F., Robert W., Massally J.L., Chapman S.B., Bochicchio J., Murphy C., Nusbaum C., Young S., Birren B.W., Grant D.S., Scheiffelin J.S., Lander E.S., Happi C., Gevao S.M., Gnirke A., Rambaut A., Garry R.F., Khan S.H., Sabeti P.C. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science, 2014, vol. 345, no. 6202, pp. 1369–1372. doi: 10.1126/science.1259657*
36. Griffiths R.C., Tavaré S. Sampling theory for neutral alleles in a varying environment. *Philos. Trans. R. Soc. London Ser. B. Biol. Sci., 1994, vol. 344, no. 1310, pp. 403–410. doi: 10.1098/rstb.1994.0079*
37. Higgins D.G., Sharp P.M. CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. *Gene, 1988, vol. 73, no. 1, pp. 237–244. doi: 10.1016/0378-1119(88)90330-7*
38. Hill V., Baele G. Bayesian estimation of past population dynamics in BEAST 1.10 using the skygrid coalescent model. *Mol. Biol. Evol., 2019, vol. 36, no. 11, pp. 2620–2628. doi: 10.1093/molbev/msz172*
39. Ho S.Y.W., Duchene S. Molecular-clock methods for estimating evolutionary rates and timescales. *Mol. Ecol., 2014, vol. 23, no. 24, pp. 5947–5965. doi: 10.1111/mec.12953*
40. Jeffreys H. Some tests of significance, treated by the theory of probability. *Math Proc. Cambridge Philos. Soc., 1935, vol. 31, no. 2, pp. 203–222. doi: 10.1017/S030500410001330X*
41. Jorba J., Campagnoli R., De L., Kew O. Calibration of multiple poliovirus molecular clocks covering an extended evolutionary range. *J. Virol., 2008, vol. 82, no. 9, pp. 4429–4440. doi: 10.1128/JVI.02354-07*
42. Jukes T., Cantor C. Evolution of protein molecules. In: Mammalian protein metabolism. *New York: Academic Press, 1969, pp. 21–132.*
43. Kainer D., Lanfear R. The effects of partitioning on phylogenetic inference. *Mol. Biol. Evol., 2015, vol. 32, no. 6, pp. 1611–1627. doi: 10.1093/molbev/msv026*
44. Kass R., Raftery A. Bayes factors. *J. Am. Stat. Assoc., 1995, vol. 90, pp. 773–795. doi: 10.2307/2291091*
45. Katoh K., Standley D.M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol., 2013, vol. 30, no. 4, pp. 772–780. doi: 10.1093/molbev/mst010*
46. Keane T.M., Creevey C.J., Pentony M.M., Al E. Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. *BMC Evol. Biol., vol. 6: 29. doi: 10.1186/1471-2148-6-29*
47. Keeling M.J., Rohani P. Modeling infectious diseases in humans and animals. *New Jersey: Princeton University Press, 2007. 408 p.*
48. Kingman J.F.C. The coalescent. *Stoch. Process. Their Appl., 1982, vol. 13, no. 3, pp. 235–248. doi: 10.1016/0304-4149(82)90011-4*
49. Koonin E.V., Dolja V.V., Krupovic M. Origins and evolution of viruses of eukaryotes: the ultimate modularity. *Virology, 2015, vol. 479–480, pp. 2–25. doi: 10.1016/j.virol.2015.02.039*
50. Kuhner M.K., Felsenstein J. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol. Biol. Evol., 1994, vol. 11, no. 3, pp. 459–468. doi: 10.1093/oxfordjournals.molbev.a040126*
51. Lanfear R., Frandsen P.B., Wright A.M., Senfeld T., Calcott B. PartitionFinder 2: new methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. *Mol. Biol. Evol., 2017, vol. 34, no. 3, pp. 772–773. doi: 10.1093/molbev/msw260*
52. Lartillot N., Philippe H. Computing bayes factors using thermodynamic integration. *Syst. Biol., 2006, vol. 55, no. 2, pp. 195–207. doi: 10.1080/10635150500433722*
53. Lemey P., Rambaut A., Drummond A.J., Suchard M.A. Bayesian phylogeography finds its roots. *PLoS Comput. Biol., 2009, vol. 5, no. 9: e1000520. doi: 10.1371/journal.pcbi.1000520*
54. Lemey P., Rambaut A., Welch J.J., Suchard M.A. Phylogeography takes a relaxed random walk in continuous space and time. *Mol. Biol. Evol., 2010, vol. 27, no. 8, pp. 1877–1885. doi: 10.1093/molbev/msq067*
55. Maio De N., Wu C.H., O'Reilly K.M., Wilson D. New routes to phylogeography: a Bayesian structured coalescent approximation. *PLOS Genet., 2015, vol. 11, no. 8: e1005421. doi: 10.1371/journal.pgen.1005421*
56. Margoliash E. Primary structure and evolution of cytochrome C. *Proc. Natl. Acad. Sci., 1963, vol. 50, no. 4, pp. 672–679. doi: 10.1073/pnas.50.4.672*
57. Martin D.P., Murrell B., Golden M. Khoosal A., Muhire B. RDP4: detection and analysis of recombination patterns in virus genomes. *Virus Evol., 2015, vol. 1, no. 1, pp. 1–5. doi: 10.1093/ve/vev003*

58. Nascimento F.F., dos Reis M., Yang Z. A biologist's guide to Bayesian phylogenetic analysis. *Nat. Ecol. Evol., 2017, vol. 1, no. 10, pp. 1446–1454. doi: 10.1038/s41559-017-0280-x*
59. Needleman S.B., Wunsch C.D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol., 1970, vol. 48, no. 3, pp. 443–453. doi: 10.1016/0022-2836(70)90057-4*
60. Notredame C., Higgins D.G., Heringa J. T-coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol., 2000, vol. 302, no. 1, pp. 205–217. doi: 10.1006/jmbi.2000.4042*
61. Parag K.V., Pybus O.G. Exact Bayesian inference for phylogenetic birth-death models. *Bioinformatics, 2018, vol. 34, no. 21, pp. 3638–3645. doi: 10.1093/bioinformatics/bty337*
62. Pybus O.G. The epidemiology and iatrogenic transmission of hepatitis C virus in Egypt: a Bayesian coalescent approach. *Mol. Biol. Evol., 2003, vol. 20, no. 3, pp. 381–387. doi: 10.1093/molbev/msg043*
63. Rambaut A., Lam T., Carvalho L., Pybus O. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol., 2016, vol. 2, no. 1: vew007. doi: 10.1093/ve/vew007*
64. Rannala B., Yang Z. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *J. Mol. Evol., 1996, vol. 43, no. 3, pp. 304–311. doi: 10.1007/PL00006090*
65. Rice P., Longden I., Bleasby A. EMBOSS: the European molecular biology open software suite. *Trends Genet., 2000, vol. 16, no. 6, pp. 276–277. doi: 10.1016/S0168-9525(00)02024-2*
66. Russel P.M., Brewer B.J., Klaere S., Bouckaert R.R. Model selection and parameter inference in phylogenetics using nested sampling. *Syst. Biol., 2019, vol. 68, no. 2, pp. 219–233. doi: 10.1093/sysbio/syy050*
67. Saitou N., Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol., 1987, vol. 4, no. 4, pp. 406–425. doi: 10.1093/oxfordjournals.molbev.a040454*
68. Schwarz G. Estimating the dimension of a model. *Ann. Stat., 1978, vol. 6, no. 2, pp. 461–464. doi: 10.1214/aos/1176344136*
69. Shaman J., Kohn M. Absolute humidity modulates influenza survival, transmission, and seasonality. *Proc. Natl. Acad. Sci. USA, 2009, vol. 106, no. 9, pp. 3243–3248. doi: 10.1073/pnas.0806852106*
70. Sinsheimer J.S., Lake J.A., Little R.J.A. Bayesian hypothesis testing of four-taxon topologies using molecular sequence data. *Biometrics, 1996, vol. 52, no. 1: 193. doi: 10.2307/2533156*
71. Skilling J. Nested sampling for general Bayesian computation. *Bayesian Anal., 2006, vol. 1, no. 4, pp. 833–860. doi: 10.1214/06-BA127*
72. Smith T.F., Waterman M.S. Identification of common molecular subsequences. *J. Mol. Biol., 1981, vol. 147, no. 1, pp. 195–197. doi: 10.1016/0022-2836(81)90087-5*
73. Song W., Qin K. Human-infecting influenza A (H9N2) virus: a forgotten potential pandemic strain? *Zoonoses Public Health, 2020, vol. 67, no. 3, pp. 203–212. doi: 10.1111/zph.12685*
74. Stadler T., Kuhnert D., Bonhoeffer S., Drummond A.J. Birth-death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). *Proc. Natl. Acad. Sci., 2013, vol. 110, no. 1, pp. 228–233. doi: 10.1073/pnas.1207965110*
75. Stadler T., Kouyos R., Wyl V. von, Yerly S., Böni J., Bürgisser P., Klimkait T., Joos B., Rieder P., Xie D., Günthard H.F., Drummond A.J. Estimating the basic reproductive number from viral sequence data. *Mol. Biol. Evol., 2012, vol. 29, no. 1, pp. 347–357. doi: 10.1093/molbev/msr217*
76. Stadler T., Vaughan T.G., Gavryushkin A., Guindon S., Kühnert D., Leventhal G.E., Drummond A.J. How well can the exponential-growth coalescent approximate constant-rate birth-death population dynamics? *Proc. R. Soc. B. Biol. Sci., 2015, vol. 282, no. 1806: 20150420. doi: 10.1098/rspb.2015.0420*
77. Stadler T., Yang Z. Dating phylogenies with sequentially sampled tips. *Syst. Biol., 2013, vol. 62, no. 5, pp. 674–688. doi: 10.1093/sysbio/syt030*
78. Su S., Wong G., Shi W., Liu J., Lai A.C.K., Zhou J., Liu W., Bi Y., Gao G.F. Epidemiology, genetic recombination, and pathogenesis of coronaviruses. *Trends Microbiol., 2016, vol. 24, no. 6, pp. 490–502. doi: 10.1016/j.tim.2016.03.003*
79. Suchard M., Lemey P., Baele G., Ayres D.L., Drummond A.J., Rambaut A. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol., 2018, vol. 4, no. 1: vey016. doi: DOI:10.1093/ve/vey016*
80. Suchard M.A., Weiss R.E., Sinsheimer J.S. Bayesian selection of continuous-time Markov chain evolutionary models. *Mol. Biol. Evol., 2001, vol. 18, no. 6, pp. 1001–1013. doi: 10.1093/oxfordjournals.molbev.a003872*
81. Tateno Y., Takezaki N., Nei M. Relative efficiencies of the maximum-likelihood, neighbor-joining, and maximum-parsimony methods when substitution rate varies with site. *Mol. Biol. Evol., 1994, vol. 11, no. 2, pp. 261–277. doi: 10.1093/oxfordjournals.molbev.a040108*
82. Vakulenko Yu., Deviatkin A., Lukashev A. The effect of sample bias and experimental artefacts on the statistical phylogenetic analysis of picornaviruses. *Viruses, 2019, vol. 11, no. 11: 1032. doi: 10.3390/v11111032*
83. Vakulenko Yu., Deviatkin A., Lukashev A. Using statistical phylogenetics for investigation of enterovirus 71 genotype A reintroduction into circulation. *Viruses, 2019, vol. 11, no. 10: 895. doi: 10.3390/v11100895*
84. Vaughan T.G., Leventhal G.E., Rasmussen D.A., Drummond A.J., Welch D., Stadler T. Estimating epidemic incidence and prevalence from genomic data. *Mol. Biol. Evol., 2019, vol. 36, no. 8, pp. 1804–1816. doi: 10.1093/molbev/msz106*
85. Waterhouse A.M., Procter J.B., Martin D.M., Clamp M., Barton G.J. Jalview Version 2 — a multiple sequence alignment editor and analysis workbench. *Bioinformatics, 2009, vol. 25, no. 9, pp. 1189–1191. doi: 10.1093/bioinformatics/btp033*
86. Welch J., Bromham L. Molecular dating when rates vary. *Trends Ecol. Evol., 2005, vol. 20, no. 6, pp. 320–327. doi: 10.1016/j.tree.2005.02.007*
87. Worobey M., Han G.-Z., Rambaut A. A synchronized global sweep of the internal genes of modern avian influenza virus. *Nature, 2014, vol. 508, no. 7495, pp. 254–257. doi: 10.1038/nature13016*
88. Worobey M., Watts T.D., McKay R.A., Suchard M.A., Granade T., Teuwen D.E., Koblin B.A., Heneine W., Lemey P., Jaffe H.W. 1970s and 'Patient 0' HIV-1 genomes illuminate early HIV/AIDS history in North America. *Nature, 2016, vol. 539, no. 7627, pp. 98–101. doi: 10.1038/nature19827*
89. Xie W., Lewis P.O., Fan Y., Kuo L., Chen M.H. Improving marginal likelihood estimation for Bayesian phylogenetic model selection. *Syst. Biol., 2011, vol. 60, no. 2, pp. 150–160. doi: 10.1093/sysbio/syq085*

90. Yang B., Liu F., Liao Q., Wu P., Chang Z., Huang J., Long L., Luo L., Li Y., Leung G.M., Cowling B.J., Yu H. Epidemiology of hand, foot and mouth disease in China, 2008 to 2015 prior to the introduction of EV-A71 vaccine. *Euro Surveill., 2017, vol. 22, no. 50: 16-00824. doi: 10.2807/1560-7917.ES.2017.22.50.16-00824*
91. Yang J., Xie D., Nie Z., Xu B., Drummond A.J. Inferring host roles in Bayesian phylodynamics of global avian influenza A virus H9N2. *Virology, 2019, vol. 538, pp. 86–96. doi: 10.1016/j.virol.2019.09.011*
92. Yule G.U. Mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F.R.S. *Philos. Trans. R., 1924, vol. B213, pp. 21–87.*
93. Zhu J., Luo Z., Wang J., Xu Z., Chen H., Fan D., Gao N., Ping G., Zhou Z., Zhang Y., An J. Phylogenetic analysis of enterovirus 71 circulating in Beijing, China from 2007 to 2009. *PLoS One, 2013, vol. 8, no. 2: e56318. doi: 10.1371/journal.pone.0056318*
94. Zuckerkandl E., Pauling L. Molecular disease, evolution, and genic heterogeneity. In: Horizons in Biochemistry. *New York: Academic Press, 1962, pp. 189–225.*
95. Zuckerkandl E., Pauling L. Molecules as documents of history. *J. Theor. Biol., 1965, vol. 8, no. 2, pp. 357–366.*

Авторы:
Вакуленко Ю.А., младший научный сотрудник, Институт медицинской паразитологии, тропических и трансмиссивных заболеваний им. Е.И. Марциновского, Первый Московский государственный медицинский университет имени И.М. Сеченова, Москва, Россия; аспирант биологического факультета, МГУ им. М.В. Ломоносова, Москва, Россия;
Лукашев А.Н., д.м.н., член-корреспондент РАН, директор Института медицинской паразитологии, тропических и трансмиссивных заболеваний им. Е.И. Марциновского, Первый Московский государственный медицинский университет имени И.М. Сеченова, Москва, Россия; ведущий научный сотрудник, Институт молекулярной медицины, Первый Московский государственный медицинский университет имени И.М. Сеченова, Москва, Россия;
Девяткин А.А., к.б.н., старший научный сотрудник, Институт молекулярной медицины, Первый Московский государственный медицинский университет имени И.М. Сеченова, Москва, Россия.

Authors:
Vakulenko Yu.A., Junior Researcher, Martsinovsky Institute of Medical Parasitology, Tropical and Vector Borne Diseases, Sechenov First Moscow State Medical University, Moscow, Russian Federation; PhD-student, Faculty of Biology, Lomonosov Moscow State University, Moscow, Russian Federation;
Lukashev A.N., PhD, MD (Medicine), RAS Full Member, Director of Martsinovsky Institute of Medical Parasitology, Tropical and Vector Borne Diseases, Sechenov First Moscow State Medical University, Moscow, Russian Federation; Leading Researcher, Institute of Molecular Medicine, Sechenov First Moscow State Medical University, Moscow, Russian Federation;
Deviatkin A.A., PhD (Biology), Senior Researcher, Institute of Molecular Medicine, Sechenov First Moscow State Medical University, Moscow, Russian Federation.

Поступила в редакцию 25.06.2020
Принята к печати 13.07.2020

Received 25.06.2020
Accepted 13.07.2020