ПРОГРАММИРОВАНИЕ, 2022, № 3, с. 78–91

КОМПЬЮТЕРНАЯ ГРАФИКА И ВИЗУАЛИЗАЦИЯ

УДК 004.9

PHYLOTRAVIS: НОВЫЙ ПОДХОД К ВИЗУАЛИЗАЦИИ ФИЛОГЕНЕТИЧЕСКОГО ДЕРЕВА

© 2022 г. М. Форгани[a,b,*] (ORCID: 0000-0002-9443-3610), П. А. Васёв[a,] (ORCID: 0000-0003-3854-0670), М. А. Болков[c,***] (ORCID: 0000-0003-2763-9907), Э. С. Рэмзи[d,****] (ORCID: 0000-0001-7086-5825), А. Ю. Берсенев[a,*****] (ORCID: 0000-0001-5843-5224)**

[a] *Институт математики и механики им. Н.Н. Красовского УрО РАН, 620108 Екатеринбург, ул. Софьи Ковалевской, д. 16, Россия*

[b] *Уральский федеральный университет им. первого Президента России Б.Н. Ельцина, 620002 Екатеринбург, ул. Мира, д. 19, Россия*

[c] *Институт иммунологии и физиологии УрО РАН, 620049 Екатеринбург, ул. Первомайская, д. 106, Россия*

[d] *Научно-исследовательский институт гриппа имени А.А. Смородинцева, 197376 Санкт-Петербург, ул. Профессора Попова, д. 15/17, Россия*

**E-mail: forghani@imm.uran.ru*

***E-mail: vasev@imm.uran.ru*

****E-mail: mbolkov@iip.uran.ru*

*****E-mail: warmsunnyday@mail.ru*

******E-mail: bay@hackerdom.ru*

Поступила в редакцию 18.12.2021 г.
После доработки 11.01.2022 г.
Принята к публикации 20.01.2022 г.

Изучение эволюции является необходимой задачей при прогнозирования изменчивости вида, и особенно для таких патогенов, как вирусы. Один из основных этапов эволюционного анализа это построение филогенетического дерева. Данная работа посвящена новому подходу для визуализации филогенеза, который основан на построении эволюционной траектории таксона в трехмерном пространстве. Эволюционной траекторией является путь, который соединяет конкретный таксон и корень эволюционного дерева. Реконструируя предковые последовательности и применяя унитарное кодирование, каждый узел дерева представляется в виде многомерного объекта, которые затем через методы вложения выстраиваются в трехмерном пространстве, за счет чего восстанавливаются эволюционные пути от листов до корня дерева. Данный подход позволяет визуализировать резкие изменения направления эволюции в локальном и глобальном масштабах. Результатом работы являются эксперименты по визуализации эволюционной траектории вируса гриппа H3N2 и создание web-платформы PhyloTraVis с публичным доступом. Результаты предполагают применения нашего подхода также для раннего обнаружения изменения направления эволюции, изучения динамики эволюции, оценки появлений новых вариантов вируса, а также моделирования возможного антигенного разнообразия, что является актуальной задачей вычислительной вирусологии.

DOI: 10.31857/S0132347422030049

1. ВВЕДЕНИЕ

Молекулярная эволюция – это процесс, лежащий в основе молекулярных изменений последовательностей ДНК, РНК и/или аминокислот между поколениями. Молекулярную эволюцию можно рассматривать как алгоритмический процесс, включающий три последовательных этапа: на уровне генотипа это формирование поколения за счет случайных мутаций; затем на фенотипическом уровне работает естественный отбор; и на последнем этапе происходит репродукция новых дивергированных вариантов [1]. Накапливая мутации в генетической последовательности, поколение может повысить свою функциональность и получить превосходство над другими вариантами.

Изучение эволюции позволяет предсказывать направление изменчивости живых существ, что становится особенно важной задачей в отношении эпидемически опасных патогеннов. Оно является необходимым этапом при разработке эф-

фективных препаратов и вакцин, особенно, когда необходимо предсказать консервативные участки, не подверженные изменчивости [2]. Кроме того, эволюция таких патогеннов, как вирусы и вироиды связана с эволюцией других живых существ, и расширяет наше представление об эволюции всей жизни на планете Земля [3].

В отношении вирусов изучение направления эволюции имеет большое значение. В основном, изменения в популяции вирусов происходят из-за мутаций, возникающих в процессе репликации. Вирусная полимераза подвержена ошибкам, что приводит к мутациям в вирусном геноме. Это может привести к образованию сложной популяции родственных, но неидентичных геномов, называемых квазивидами [4].

Вирусы – одна из самых простых моделей для изучения эволюции. Небольшие размеры вирусных геномов и белков позволяют учитывать не только значительные изменения вирусов, но и незначительные аминокислотные замены, которые оказались критически значимыми для антигенной эволюции вирусов [5]. Иногда даже одна замена аминокислоты может привести к смене антигенного кластера, как это произошло между двумя антигенными кластерами BE89 и SI87 вируса гриппа [6]. Для большинства вирусов изменчивость происходит двумя классическими путями: антигенный дрейф, то есть постепенное накопление изменений; и антигенная изменчивость (сдвиг) или реассортмент, представляющие собой внезапные и значительные изменения в белках вируса. Процесс эволюции вызывает изменения антигенных характеристик вируса, что приводит к избеганию вирусом иммунного ответа хозяина и снижению эффективности вакцин [7].

Эволюцию можно анализировать на разных уровнях. Например, в случае вируса гриппа существуют модели для описания эволюции, построенные на основе филогенетической и популяционной генетики, антигенных отношений, эпидемиологических данных и данных о структуре белка [8]. Основополагающей и классической моделью для представления эволюционных отношений между различными видами является построение филогенетического дерева. Филогенетическое дерево – это биоинформатическая структура данных, в которой различия и сходства между видами (например, в генетическом пространстве) демонстрируются в компактной форме наподобие дендрограммы. На самом деле, дерево преобразует сложные эволюционные взаимоотношения в графической форме, удобной для восприятия человеком [9], [10].

Задачи визуализации являются неотъемлемой частью большинства научных исследований, в том числе и в области биоинформатики [11]. Вообще говоря, преобразование набора нуклеотидных или аминокислотных последовательностей в визуализированное дерево требует выполнения двух последовательных шагов: оценка филогенетического отношения между видами с помощью филогенетического анализа (методы которого подразделяются на фенетические и кладистические), и визуализация полученных результатов филогенетического анализа. В этом случае, визуализация направлена на содействие лучшему пониманию генетической дивергенции, на мониторинг, обнаружение новых вариантов, проверку, получение и осмысление данных [12].

Ниже представлены три ведущих подхода, разработанных для моделирования эволюции, в которых визуализация играет решающую роль. Ито и соавторы в 2011 году предложили модель для предсказания будущего направления эволюции вируса гриппа [13]. Их модель построена на основе применения известного метода многомерного шкалирования (MDS) [14] к расстояниям между аминокислотными последовательностями выделенных в прошлом штаммов. Представленная ими 3D-визуализация одновременно демонстрирует направление эволюции и филогенез последовательностей, что невозможно обеспечить с помощью традиционного филогенетического анализа. Их подход позволил достичь полноты предсказания (recall) около 0.70, что говорит о способности модели предсказывать аминокислотные замены [14].

Другим примером применения визуализации для графического представления эволюции является антигенная картография, предложенная Смитом и соавторами [6]. Основная идея этого подхода заключается в применении модифицированного метода MDS на антигенное расстояние для определения местоположения антигенов и антисывороток на антигенной карте. Антигенная картография по-прежнему является одним из традиционных инструментов для демонстрации антигенной эволюции патогенов с высокой антигенной изменчивостью. Иногда визуализация является промежуточным шагом для моделирования эволюции. Например, как было предложено в работах [5], [15], некоторые признаки или переменные для математической модели антигенной эволюции могут быть извлечены из филогенетического дерева. Несмотря на все вышеприведенные подходы, моделирование и визуализация эволюции до сих пор являются актуальными задачами в биоинформатике.

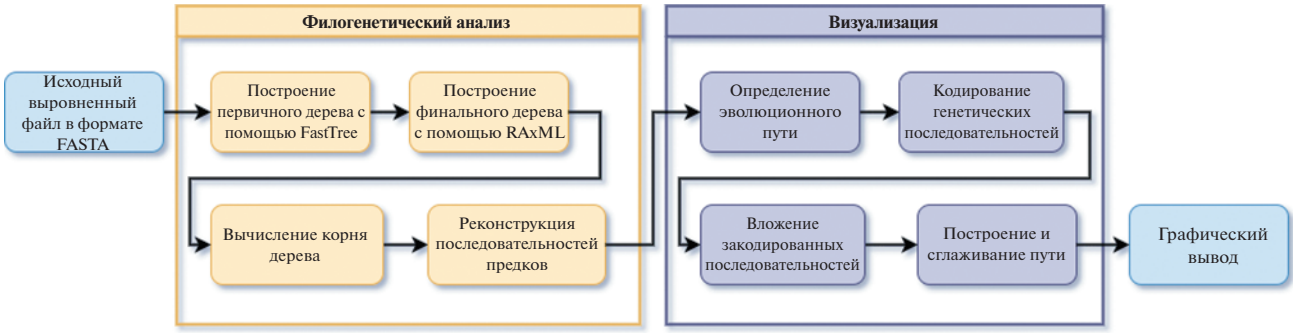

Рис. 1. Общая схема предлагаемого подхода. Она состоит из двух последовательных этапов: филогенетического анализа и визуализации. Следует обратить внимание, что на входе рабочего процесса подается выровненный FASTA-файл.

Филогенетическое дерево можно посторить принимая модель эволюции. Фактически, такая модель является инструментом для оценки эволюционного расстояния, то есть меры генетической дивергенции по наблюдаемым различиям между видами. Существуют различные модели, большинство из которых основаны на модели Маркова для эволюции последовательностей. Эти модели различаются в зависимости от типа генетических данных (ДНК, белок или кодон) и параметров, описывающих скорость генетических замен. Например, известно, что скорость транзиций, то есть замены пуринового азотистого основания пуриновым (A $\leftrightarrow$ G), или пиримидинового – другим пиримидиновым (C $\leftrightarrow$ T), выше, чем скорость трансверсий (замены пуринового основания пиримидиновым, и наоборот). Модель Кимуры, известная как K80, присваивает индивидуальную скорость каждой из транзиций и трансверсий, представляя скорости как параметры модели [16].

Параллельно со стремительным развитием моделирования и подходов к выводу топологии дерева развивались методы визуализации, а также соответствующие пакеты. Примерами могут служить такие пакеты, как TREEVIEW [17], PHYLO_WIN [18], FIGTREE [19], iTOL [20], Phylo.io [21], PhyloExplorer [22], и Treeio [23].

Развивая наш предыдущий опыт визуализации филогенетического дерева [10], [24], в данной работе мы сосредоточились на представлении новой 2D и 3D-визуализации филогенетического дерева с использованием информации о последовательности предковых узлов. Главная идея представленного подхода основана на визуализации алгоритмов решения кубика Рубика [25]. При визуализации решений кубика Рубика путь решения начинается со случайного состояния и заканчивается конечным состоянием (которое является полным решением кубика). Путь решения визуализируется с помощью применения методов унитарного кодирования и стохастического вложения соседей с t-распределением (t-SNE) [26]. Рассматривая таксон, расположенный в листе, как начальное состояние и корень дерева как конечное состояние, мы представляем обратный эволюционный путь аналогичный тому, что представлен в визуализации кубика Рубика. Для построения такого пути необходимо наличие генетических последовательностей внутренних узлов дерева, которые могут быть получены с помощью алгоритмов для реконструкции предковых последовательностей.

Наш вклад в данной работе заключается в создании нового подхода к визуализации филогенетического дерева и реализации данного подхода в виде онлайн-платформы под названием *PhyloTraVis* (Phylogenetic Trajectory Visualization). Платформа находится в открытом доступе по адресу phylotravis.viroinformatics.com. Предложенный подход может быть использован в различных исследованиях, включая моделирование антигенной эволюции вирусов. Остальная часть статьи организована следующим образом. В разделе 2 более подробно описана методология и соответствующие алгоритмы. Раздел 3 включает постановку эксперимента и результаты. Наконец, в разделе 4 приведено заключение.

2. МЕТОДОЛОГИЯ

PhyloTraVis – это платформа, разработанная с использованием FLASK [27], Biopython [28] и Scikit-learn [29]. Она состоит из двух частей: филогенетического анализа и визуализации. На первом этапе выровненный файл FASTA, включающий генетические последовательности таксонов, переносится в матрицу, которая затем встраивается в трехмерное пространство алгоритмами

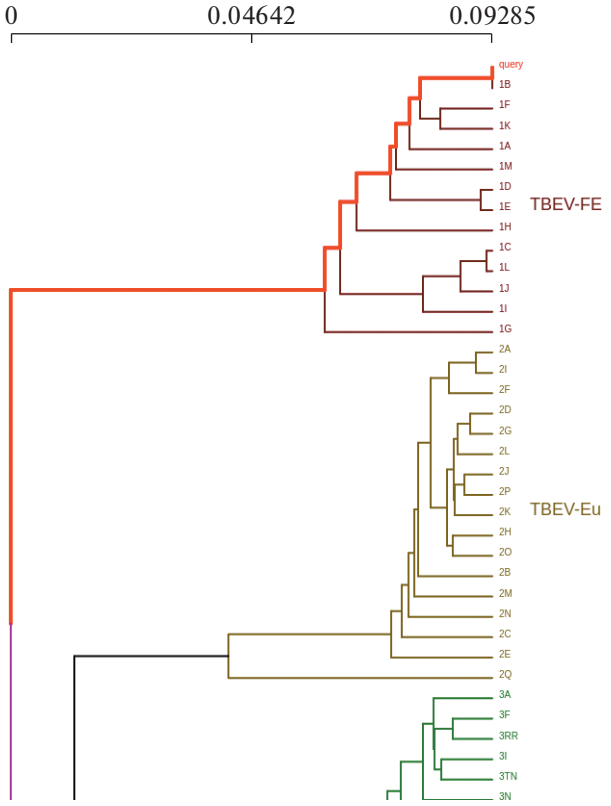

Рис. 2. Филогенетическое дерево обычно представляется в виде дендрограммы. Эволюционная траектория – это путь, начинающийся от таксона и заканчивающийся в корне дерева. Пример такого пути выделен на дереве красным цветом. PhyloTraVis рассматривает каждый путь как отдельный объект для визуализации. Здесь представлена только часть дерева, полученная на платформе TBEV-Analyzer [32], разработанной для вируса клещевого энцефалита.

второго этапа. Общий рабочий процесс нашего подхода показан на рис. 1.

Входной файл должен быть в формате FASTA и выровнен. Более того, поскольку построение филогенетического дерева требует, чтобы все генетические последовательности были уникальными, во входном файле не должно быть дублирующихся последовательностей. В противном случае платформа возвращает исключение, информируя пользователя. Филогенетический анализ проводится с помощью двух пакетов: FastTree [30] и RAxML [31]. Randomized Axelerated Maximum Likelihood (RAxML) – это программа для проведения филогенетического анализа по методу максимального правдоподобия, обеспечивающая высокую точность. Стохастические модели (такие как максимальное правдоподобие) более применимы в биологических исследованиях, обеспечивая множество желаемых статистических свойств, однако они часто страдают от низкой вычислительной эффективности. По этой причине мы используем FastTree, один из самых быстрых методов оценки максимального правдоподобия, для создания начального дерева. Начальное дерево вводится в RAxML для создания окончательного филогенетического дерева. RAxML предоставляет широкий диапазон параметров, включая выбор модели замены. При работе с большими наборами данных RAxML обладает выгодными преимуществами, так как он позволяет проводить вычисления параллельно для определения наилучшей оценки по методу максимального правдоподобия.

Конечный результат работы филогенетического анализа зависит от топологии дерева, а также генетических последовательностей всех внутренних узлов и корня, как наиболее общего предка. При этом, обязательным условием расчета предковых последовательностей в RAxML является наличие корня. Укорененное дерево можно просто получить из неукорененного дерева с помощью RAxML, используя флаг ’-f I’. Корень дерева располагается на той ветви, которая наилучшим образом уравновешивает длины поддеревьев для левого и правого поддеревьев.

Реконструкция предковой последовательности – это подход, использующий максимальное правдоподобие или байесовские методы для статистического вывода генетической последовательности предковых узлов. Вычисление предковых последовательностей сильно зависит от топологии и филогенеза. Чтобы для укорененного дерева получить предковые последовательности, в пакете RAxML используется флаг ’-f A’. В итоге RAxML создаст несколько файлов, включающих вероятности и последовательности предков, которые будут использованы на следующем этапе.

На этапе визуализации происходит объединение генетической информации всех узлов дерева и топологии дерева для его вложения в новое трехмерное представление. Включение информации о предковых узлах в визуализацию осуществляется путем определения филогенетической траектории (или пути). Филогенетическая траектория – это ненаправленный путь графа, начинающийся от таксона и заканчивающийся в корне дерева (см. рис. 2). Такой путь представляет собой эволюционную историю целевого производного таксона от наиболее общего предка (или корня), а длина ветви между двумя узлами определяется генетическим сходством между ними. Путь генерируется из дерева с помощью модуля “phylo” из пакета Biopython [28]. Модуль позволяет извлечь родственные отношения между узлами дерева, что необходимо для реконструкции эволюцион-

ной траектории. Следует обратить внимание, что общее количество путей равно общему количеству таксонов (количеству видов во входном FASTA-файле).

Хотя формирование пути зависит от генетической информации всех таксонов, на уровне визуализации его можно рассматривать как самостоятельный объект. Основной задачей на этапе визуализации является встраивание эволюционного пути в 3D/2D-пространство. Для этого используются генетические последовательности узлов пути, так что задача превращается во вложение каждого узла из генетического пространства в 3D/2D-пространство. Тем не менее, перед вложением необходимо, чтобы генетические данные были представлены в числовом формате, поскольку большинство математических методов вложений работают с числовыми векторами.

Существуют различные методы представления генетической последовательности в числовом пространстве. Вдохновленные визуализацией решения кубика Рубика [25], для кодирования и аминокислот и нуклеотидов мы использовали унитарное кодирование (one hot encoding). При таком кодировании не учитывается приоритет между строительными блоками генетической последовательности. В таблице 1 приведен пример унитарного кодирования для нуклеотидов. Аналогичная таблица может быть составлена для набора аминокислот.

Применяя унитарное кодирование, генетические последовательности для всех узлов (включая внутренние и конечные) представляются в виде бинарной матрицы. Каждая строка матрицы – это бинарная последовательность, связанная с определенным узлом дерева, которая представляет узел в бинарном многомерном пространстве. Для уменьшения размерности матрицы принимаются метод вложения, переходя в двумерное или трехмерное пространство. Отметим, что в отличие от нашего последнего подхода для визуализации филогенетического дерева [24], в котором каждый путь отдельно встраивается в трехмерное пространство, PhyloTraVis встраивает бинарную матрицу в низкоразмерное пространство, рассматривая все узлы вместе.

В настоящее время платформа предоставляет два известных метода вложения: MDS и t-SNE. В то время как метод MDS стремится максимально сохранить расстояния между объектами при вложении, метод t-SNE преобразует объекты в совместную вероятность и пытается минимизировать дивергенцию Кульбака–Лейблера (KL) между вероятностями в высокоразмерном и низкоразмерном пространствах. Другими словами, для n точек $X = \{x_1, x_2, \ldots, x_n\}$, $x_i \in \mathbb{R}^p$, и их матрицы аффинности (сходства) D, MDS стремится минимизировать приведенную ниже объективную функцию

Таблица 1. Унитарное кодирование для представления алфавитной последовательности нуклеотидов в числовом бинарном пространстве

Нуклеотид	Код
A	(1, 0, 0, 0)
C	(0, 1, 0, 0)
G	(0, 0, 1, 0)
T	(0, 0, 0, 1)
'–' Gap	(0, 0, 0, 0)

$$\min_{Y} \sum_{i=1}^{n} \sum_{j=1}^{n} (d_{i,j} - \hat{d}_{i,j})^2,$$

где $Y = \{y_1, y_2, \ldots, y_n\}$, $y_i \in \mathbb{R}^q$ – точки в пространстве низкой размерности ($q < p$), а $\hat{D}$ – матрица их аффинности в новом пространстве.

Метод t-SNE использует расстояние между объектами в высоко- и низкоразмерном пространстве для определения условной вероятности, которая определяет, принадлежат ли две точки к одной группе или нет. Если $p_{j|i}$ и $q_{j|i}$ – условные вероятности для точек i, j, представляющие сходство точки данных x_j с точкой данных x_i в высоко- и низкоразмерном пространстве соответственно, то объективная функция расхождения KL может быть определена как:

$$KL(P \parallel Q) = \sum_{i \neq j} p_{j|i} \log \frac{p_{j|i}}{q_{j|i}},$$

что выражает общую стоимость представления объектов в низкоразмерном пространстве. Расхождение может быть минимизировано алгоритмом градиентного спуска.

После вложения узлов в 3D/2D-пространство и до финальной визуализации, мы применяем технику сглаживания, называемую алгоритмом кривой Безье, чтобы улучшить представление пути в низкоразмерном пространстве. Кривая Безье – это параметрическая кривая, которая определяется набором точек, называемых контрольными. Количество точек минус одна представляет поря-

```
N,X,Y,Z,R,G,B,TEXT
1,-0.4047768712043762,-15.214411735534668,-6.602959156036377,58,0,197,HK/1/68
1,-1.2805455923080444,-15.827953338623047,-6.8607497215271,53,0,202,
1,-1.6616933345794678,-16.126724243164062,-6.948936939239502,51,0,204,
1,-1.8214616775512695,-16.22327995300293,-6.913727283477783,50,0,205,
1,2.3749938011169434,12.389556884765625,2.3549633026123047,129,0,126,
1,2.185904026031494,12.525310516357422,2.4929850101470947,129,0,126,
1,2.014712333679199,12.572115898132324,2.7459797859191895,130,0,125,
1,1.8646876811981201,12.506885528564453,3.1399433612823486,132,0,123,
1,1.738439679145813,12.305143356323242,3.6987006664276123,136,0,119,
1,1.6378767490386963,11.942933082580566,4.440480709075928,142,0,113,
1,1.564407229423523,11.399391174316406,5.373449802398682,149,0,106,
1,1.5194145441055298,10.6599702835083,6.490389347076416,159,0,96,
1,1.5048946142196655,9.720165252685547,7.762887001037598,170,0,85,
1,1.5240517854690552,8.590001106262207,9.136513710021973,183,0,72,
2,-0.9473636746406555,-16.751054763793945,-8.015810012817383,42,0,213,HK/107/71
2,-1.0571842193603516,-16.98721122741699 2,-7.431650638580322,44,0,211,
2,-1.1745434999465942,-16.97091293334961,-7.110812187194824,46,0,209,
2,-1.3054088354110718,-16.821640014648438,-6.894881248474121,48,0,207,
2,-1.4556471109390259,-16.60980224609375,-6.689837455749512,50,0,205,
2,-1.6283318996429443,-16.37251091003418,-6.438336372375488,52,0,203,
2,-1.8214434385299683,-16.12669563293457,-6.109073162078857,55,0,200,
2,-2.027768611907959,-15.878969192504883,-5.691685676574707,59,0,196,
2,-2.2363240718841553,-15.631830215454102,-5.192210674285889,62,0,193,
2,-2.434361696243286,-15.387199401855469,-4.628213405609131,66,0,189,
2,-2.609226703643799,-15.147857666015625,-4.02383279800415,71,0,184,
......
```

Рис. 3. Входные данные для вывода на экран. N – номер ломаной, X, Y, Z – координаты очередного узла ломаной, R, G, B – цвет узла. Узлы с одинаковым значением N описывают одну ломаную. В последнем узле также указывается название таксона в колонке *TEXT*.

док кривой. Например, если даны две различные точки p_0 и p_1, линейная кривая Безье определяется следующим образом [32]:

$$B_{p_0,p_1}(t) = p_0 + (p_1 - p_0)t = (1-t)p_0 + tp_1,$$

где $0 \le t \le 1$.

Вообще говоря, кривая Безье со степенью n может быть рекурсивно выражена как линейная комбинация двух кривых Безье степени $n-1$, как показано ниже:

$$B_{p_0}(t) = p_0$$
$$Bp_0, \ldots, p_n(t) =$$
$$= (1-t)Bp_0, \ldots, p_{n-1}(t) + tBp_1, \ldots, p_n(t),$$

где $0 \le t \le 1$, $Bp_0, \ldots, p_{n-1}(t)$ and $Bp_1, \ldots, p_n(t)$ кривые Безье порядка $n-1$ для множества точек $p_0, \ldots, p_{n-1}$ and $p_1, \ldots, p_n$ соответственно.

После сглаживания все эволюционные пути переводятся в кривые Безье, которые визуализируются в трехмерном пространстве. При применении кривой Безье, в построении конечного пути участвуют все промежуточные узлы. Узлы, создающие путь, включая предков, служат контрольными точками для кривой. Поскольку координаты этих точек получены с помощью вложения этих объектов из многомерного пространства последовательностей в 3D-пути, полученные конечные кривые претендуют на отображение полной истории таксонов на основе представленного филогенетического дерева.

В результате вычислений траекторий формируется набор кривых в трехмерном пространстве. Каждой кривой соответствует название своего таксона, филогенетическую траекторию которого она представляет. Кривые преобразуются в набор ломаных и записываются в текстовый файл в формате CSV с колонками ($N, X, Y, Z, R, G, B,$ *TEXT*), см. рис. 3.

Эти (и только эти) данные являются входными для программы, выводящей их на экран. Программа работает в веб-браузере и с помощью технологий трехмерной графики изображает заданные ломаные и текст. Пользователь в интерактивном режиме может менять направление взгляда, включать и отключать вывод названий таксонов, управлять другими параметрами.

Программа реализована с помощью технологии *Vrungel* [34], ведущим разработчиком которой является один из авторов данной статьи. Технология состоит из языка программирования и его интерпретатора, работающего в веб-браузере. Язык позволяет относительно кратко описывать деревья из объектов, формирующих а) трехмерную сцену, б) двумерный интерфейс, в) дополнительные вычисления. Код программы визуализации представлен на рис. 4.

Каждый объект описывается конструкцией вида “имя: главная-особенность параметры... осо-

```
/// загрузка модулей, содержащих определения используемых особенностей
load files="lib3dv3 csv params io gui render-params df misc scene-explorer-3d";

/// загрузка и подготовка данных
pq:  get_query_param name="csv_file";
dat: load-file file=@pq->output | parse_csv | rescale_rgb;

/// описание 3D сцены и ее рендеринг
render3d target=@view
{
    @dat | linestrips;

    @dat
      | df_filter code="(line) => line.TEXT?.length > Ø"
      | text3d size=0.2 visible=@cb1->value color=@titlecol->value;
};

/// интерфейс пользователя gui
screen auto-activate {

  column padding="1em" {
    cb1: checkbox text="Show titles";
    titlecol: select_color value=[1,1,1];
  };

  view: view3d;
};

/// доп. функции
/// rescale_rgb - делит значения в колонках R,G,B на 255, приводя их значения к 0..1
register_feature name="rescale_rgb" {
  df_div column="R" coef=255 | df_div column="G" coef=255 | df_div column="B" coef=255;
};
```

Рис. 4. Упрощенная версия кода программы визуализации. Полная версия доступна в github.com/viewzavr/vr-flu-evolution/blob/main/main.cl.

бенности... { вложенные_объекты... }". Имя объекта необязательно и его параметры – как в других языках программирования. Особенности – это идентификаторы окружений, которые будут добавлены в создаваемый объект (аналогия – mixin в языке Руби). Главная особенность – это основное окружение объекта (аналогия – "классы" в других языках программирования).

С помощью записей из объектов формируется дерево, узлы которого – созданные объекты. Отношение родитель-дети может затем использоваться в различной семантике. Например в двумерном пользовательском интерфейсе оно используется для расчета положения элементов интерфеса на экране.

Конструкции вида `param1=@obj->param2` означают ссылку. При изменении значения в `@obj->param2` оно будет скопировано в текущий объект в параметр `param1`. При изменении значений параметров вызываются настроенные объектами обработчики событий. Таким образом можно утверждать, что представленный код обладает признаками реактивного программирования. Конструкция вида `object1 | object2 ... | objectN` означает конвейер, где входы и выходы объектов замыкаются с помощью ссылок в цепочку (то есть `object2 input=@object1->input` и так далее). Например, конструкция `@dat | linestrips` создает объект `linestrips`, который отвечает за генерацию трехмерного представления ломаных, а входом для него будет служить значение `@dat->output`. Определения особенностей объектов можно задавать на языке JavaScript или программы визуализации. Например, на рис. 4 в строке `register_feature name="rescale_rgb"`, определяется новая особенность с именем "`rescale_rgb`".

Трехмерное представление во Vrungel реализовано с помощью библиотеки ThreeJS [35]. Особенность `view3d` задает двумерную область вывода. Особенность `render3d` – формирует цикл рендеринга с выводом в указанную область вывода. Перечень объектов, которые будут отрисовываться с помощью `render3d`, задается отношением родитель-дети. Объекты `linestrips` и `text3d` формируют необходимые трехмерные

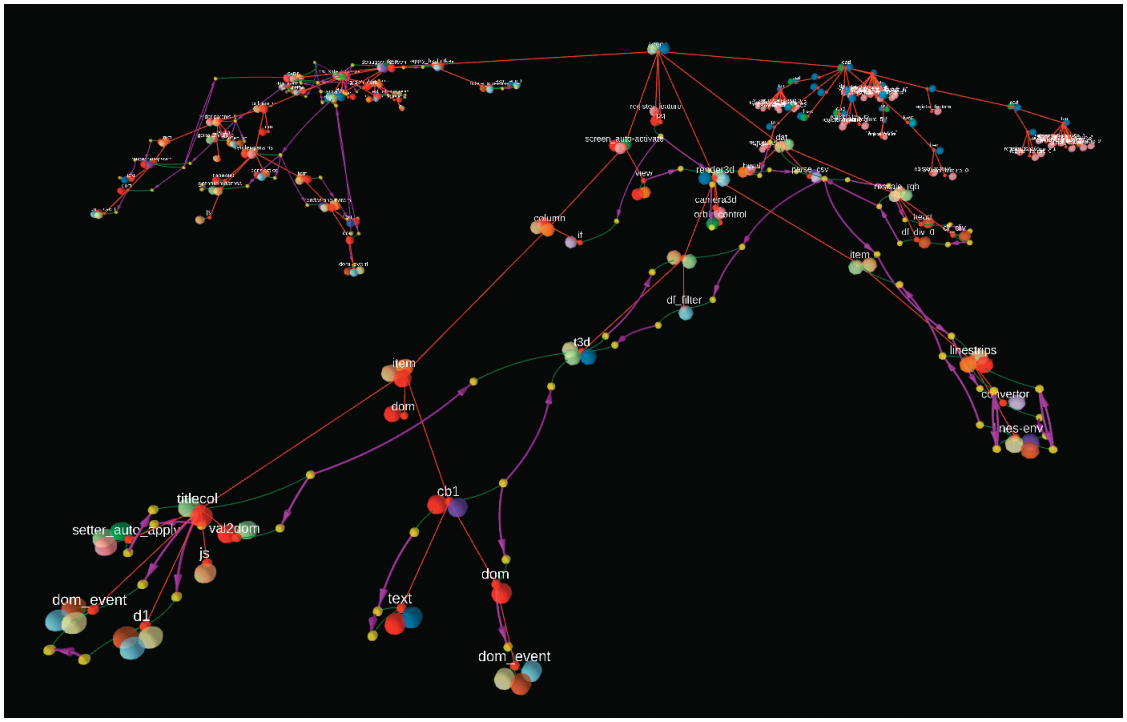

Рис. 5. Визуальный образ дерева объектов, порожденных программой визуализации. Красные шары – объекты, желтые – параметры объектов, большие шары – особенности (mixin) объектов. Показаны связи: красным – отношение родитель-дети, зеленым – объект-параметр, фиолетовые стрелки – передача значений параметров по ссылкам.

образы, ломаные и текст соответственно, и являясь детьми render3d, поступают к нему для рендеринга.

Написанная программа передается на вход интерпретатору, входящему в проект Vrungel. Он работает в веб-браузере, загружает программу. Процесс работы далее определяется созданным деревом объектов. Например, в программе рис. 4 произойдет следующее:

1. Объект `get_query_param` считает query-параметр из URL-строки браузера, имя параметра `csv_file`.

2. Считанное значение параметра `csv_file` будет передано объекту `load-file`.

3. В качестве реакции на изменение значения `load-file` прочитает файл, размещаемый по адресу, указанному в значении.

4. Прочитанное содержимое файла поступит объекту `parse_csv` и затем объекту `rescale_rgb`. Результат записывается в объект `dat` в поле `output`.

5. Данные из `dat` поступают в `linestrips` – и формируется `threejs`-объект для рендеринга ломаных. Также данные из `dat` поступают на фильтр пустых строк и далее в `text3d` – формируется `threejs`-объект рендеринга текста.

6. Сформированные `threejs`-объекты собираются объектом `render3d`, 60 раз в секунду происходит рендеринг с выводом во `view3d`.

7. Пользователь может управлять камерой, а также двумя элементами управления – checkbox для включения-выключения текста и `select_color` для выбора его цвета.

Интересной особенностью Vrungel является экспериментальный отладчик, который изображает сущности загруженной программы в трехмерном пространстве, см рис. 5. В отладчике используется метафора молекулы [35], которая позволяет видеть программу как в целом, так и в частностях.

Информация по проекту Vrungel и его исходные коды публично доступны по адресу https://github.com/viewzavr/vrungel.

3. ЭКСПЕРИМЕНТЫ И РЕЗУЛЬТАТЫ

Для того чтобы продемонстрировать возможности нашего подхода, мы визуализировали филогенетическое дерево для вируса гриппа. Было сделано две визуализации. В первой мы показываем визуализацию построенного филогенетического дерева с относительно большим количеством штаммов. Вторая визуализация показывает

Таблица 2. Параметры, установленные для визуализации белка гемагглютинина (HA) вируса гриппа в наших экспериментах. Здесь мы использовали двухэтапную реконструкцию филогенетического дерева

Процессы	Параметры
Первая модель RAxML	-m PROTCATGTR -p 12345 -e 0.01
Вторая модель RAxML	-m PROTGAMMAGTR -p 12345 -e 0.01
Rooting (создание корня)	-f I -m PROTCATGTR -p 12345 -e 0.01
Ancestral reconstruction (реконструкция предков)	-f A -m PROTCATGTR -p 12345 -e 0.01
Encoding (кодирование)	binary (в настоящее время можно выбрать только двоичный формат)
Embedding (вложение)	t-SNE (perplexity=30.0 early_exaggeration=12.0 learning_rate=200.0 n_iter=1000 random_state=1234)

возможность предложенного метода для математического моделирования фенотипа, например, антигенной эволюции. Более подробно каждый шаг описан в следующих подразделах.

3.1. Подготовка данных

Для первого эксперимента были выбраны два набора данных вируса гриппа подтипа H3N2. Причиной выбора этого подтипа является его высокая изменчивость по сравнению с другими подтипами вируса гриппа. Первый набор штаммов включает 512 последовательностей белка гемагглютинина (HA), собранных в 1968–2007 годах, взятых из [37]. Хотя файл уже был выровнен, однако процесс выравнивания может быть выполнен с помощью таких программ, как широко используемая MUltiple Sequence Comparison by Log-Expectation (также известная как MUSCLE) [38]. MUSCLE имеет более высокую скорость и точность по сравнению с другими программами, такими как Multiple Alignment using Fast Fourier Transform (MAFFT) [39]. Однако в случае большой базы данных MAFFT обеспечивает высокую скорость за счет параллельных вычислений.

Для второго эксперимента использовался набор данных о штаммах, представленных в исследовании [6]. После извлечения их последовательностей из публичных баз данных, поскольку PhyloTraVis требует входной файл без дубликатов, мы очистили извлеченную базу данных и получили 153 последовательности. Следует обратить внимание, что если последовательность имеет дубликат в базе данных, мы удаляем как оригинальный штамм, так и штаммы с такой же последовательностью. Причина такой фильтрации заключается в том, что нам необходимо присвоить последовательности ее координаты в антигенной картографии, представленной в [6]. Поскольку дублирование последовательностей означает, что у нас есть несколько координат для одной последовательности, мы не можем определить, какая из них должна быть назначена для нее. Поэтому мы исключаем их из нашего эксперимента. Последовательности HA в обоих файлах выровнены и состоят из 329 аминокислот. PhyloTraVis требует, чтобы пользователь определил тип генетических последовательностей (нуклеотид/аминокислота), прежде чем загружать файл.

3.2. Настройка параметров

После успешной загрузки входного FASTA-файла система переходит ко второму этапу – настройке параметров. На этом этапе можно установить все параметры как для филогенетического анализа, так и для визуализации. Стоит отметить, что система может обеспечить двухэтапную реконструкцию филогенетического дерева. Другими словами, исходное дерево, сгенерированное FastTree, подается в первую модель. Выходное дерево из первой модели называется средним деревом, и оно служит входным для второй модели с целью более тонкой настройки. Итоговое филогенетическое дерево без корней является выходом второй модели. Применение второй модели является необязательным и может быть активировано пользователем. Параметры, установлен-

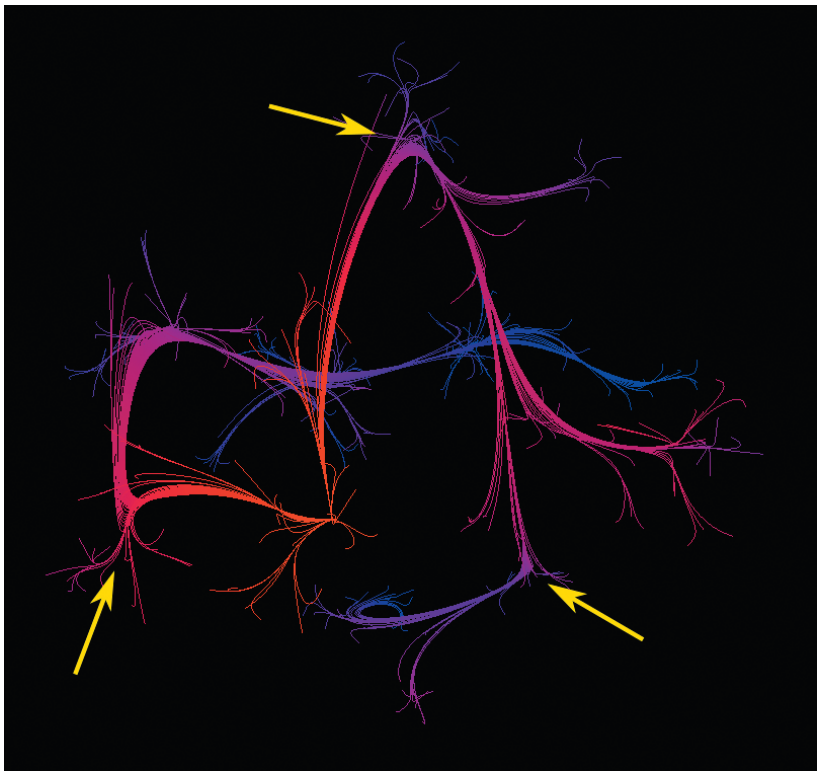

Рис. 6. Визуализация эволюции для выборки из 512 штаммов вируса гриппа подтипа H3N2, собранных в течение 1986–2007 гг. Для простоты визуализации метки штаммов не отображаются. Цветом обозначено евклидово расстояние до корня. Ближайшее расстояние до корня выделено красным цветом, а синий цвет обозначает самое большое расстояние до корня. Желтые стрелки указывают на сильные изменения в генетическом содержании в процессе эволюции.

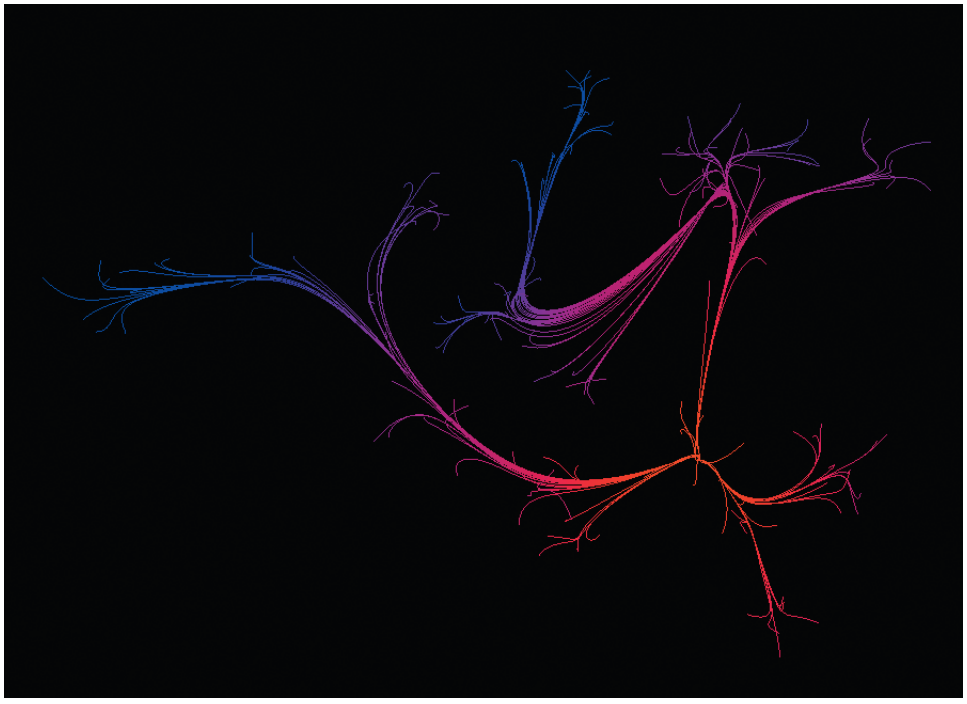

Рис. 7. Визуализация малой выборки из 152 штаммов подтипа H3N2, собранных в течение 1968–2004 годов, представленная в [6]. Евклидово расстояние между координатами штаммов во встроенном пространстве имеет умеренную корреляцию 0.61 (p-value=1e-05) с евклидовым расстоянием между их позициями на антигеновой карте.

Таблица 3. Результат теста Мантеля для вычисления корреляции матрицы евклидовых расстояний штаммов в трехмерном пространстве с матрицей расстояний Хэмминга и матрицей евклидовых расстояний штаммов на антигенной карте. Тест Мантеля был рассчитан методом Пирсона, 100000 перестановок и двухстороннего критерия (two-sided tail)

Матрица расстояний	Матрица расстояний Хэмминга	Матрица евклидового расстояния штаммов в антигенной карте
Матрица евклидовых расстояний между штаммами в 3D-пространстве	0.63 (p-value=1e-05)	0.61 (p-value=1e-05)

ные для обоих экспериментов, представлены в табл. 2.

3.3. Визуализация

После отправки задания, веб-страница периодически проверяет статус отправленного задания и сообщает об этом пользователю. Когда задание выполнено, система генерирует CSV-файл, содержащий координаты сглаженных траекторий. CSV-файл передается в программу Vrungel для показа построенных эволюционных траекторий в трехмерном пространстве пользователю. Типичная визуализация с указанной установкой для обоих наборов данных представлена на рис. 6 и 7. Она также доступна онлайн по ссылке https://github.com/view-zavr/vr-flu-evolution.

3.4. Результаты

В отличие от нашего недавно разработанного подхода [24], в котором каждый путь по отдельности встраивается в низкоразмерное пространство, PhyloTraVis рассматривает все узлы филогенетического дерева и вкладывает их в низкоразмерное пространство вместе. Полученные координаты путей зависимы на трех уровнях. На первом уровне топология дерева определяется через RAxML с помощью модели замены. Таким образом, порядок предковых узлов определяет порядок контрольных точек для кривой Безье после вложения в низкоразмерное пространство. Второй уровень модификации пути зависит от техники вложения, которая отображает объекты из высокоразмерного пространства в низкоразмерное. Третий уровень модификации траектории происходит в процессе сглаживания, когда контрольные точки участвуют в построении кривой Безье. Вообще говоря, координата каждой контрольной точки так или иначе зависит от всех узлов дерева, поэтому любое изменение топологии дерева влияет на конечную визуализацию.

Поскольку визуализация основана на генетической последовательности, предложенный подход старается сохранить соответствующее расстояние между генетически сходными таксонами. Как и ожидалось, генетически близкие штаммы также близки в трехмерном пространстве, в то время как генетически далекие штаммы имеют соответственно более выраженную дистанцию в пространстве. Одной из отличительных особенностей нашего подхода является его кастомизация. Как уже упоминалось, координаты находятся под влиянием процессов кодирования и вложения. Таким образом, пользователь может настроить визуализацию с помощью этих двух процессов. Кодирование определяет, как генетическая информация отражается в числовом пространстве, в то время как метод вложения определяет сходство между объектами в низкоразмерном пространстве. В нашем эксперименте не существует приоритета между аминокислотами, но его можно учитывать при кодировании с помощью AAindex [40]. Кроме того, концепция сходства может быть определена на основе техники вложения. Это предоставляет пользователю широкий выбор для изучения взаимосвязи таксонов. Наши эксперименты показывают, что t-SNE обеспечивает лучшее качество визуализации, чем MDS, за счет генерации более сглаженных путей.

Для того чтобы продемонстрировать эффективность предложенного подхода, мы сравнили расстояние между объектами в трехмерном пространстве с их генетическим расстоянием, а также с расстоянием на антигенной карте в исследовании [6]. В начале мы провели эксперимент для 153 белков HA вируса гриппа H3N2. Использовав PhyloTraVis, мы получили новые координаты в 3D-пространстве для каждого штамма. Затем мы вычислили попарное евклидово расстояние между штаммами и создали матрицу расстояний. Аналогично рассчитывается расстояние Хэмминга для каждой пары штаммов. С целью сравнения двух матриц расстояний мы использовали тест

Мантеля [41]. Тест Мантеля – это непараметрический тест, который проверяет значимость корреляции между двумя матрицами расстояний путем перестановки строк и столбцов. Тест дает коэффициент корреляции, который находится между –1 и 1. Коэффициент, близкий к 1, указывает на сильную положительную корреляцию, в то время как близкий к –1 означает, что существует сильная отрицательная корреляция. Если коэффициент близок к нулю, то это указывает на отсутствие корреляции между матрицами.

Таблица 3 иллюстрирует результаты теста Мантеля. Интересно, что результаты показывают, что матрицы евклидовых расстояний штаммов в трехмерном пространстве и пространстве антигенной карты показывают умеренную корреляцию 0.61 (p-value=1e-05). Это подчеркивает, что данная визуализация имеет потенциал для использования в моделировании антигенной эволюции. Более того, табл. 3 показывает, что существует умеренная корреляция между матрицей расстояний Хэмминга и матрицей евклидовых расстояний во вложенном пространстве. Это указывает на то, что большинство генетических вариаций выражено во вложенном пространстве. По итогам представленной визуализации можно сказать, что данный подход может служить в качестве дополнительного инструмента для филогенетического анализа. Полученные результаты визуализации в свою очередь можно использовать для моделирования фенотипа.

4. ЗАКЛЮЧЕНИЕ

В целом, результаты проведенных экспериментов показывают, что наш подход и платформа PhyloTraVis не столько заменяет классическое представление филогенетического дерева, сколько служит дополнительным исследовательским и аналитическим инструментом для изучения и моделирования эволюции, в том числе вирусов. Благодаря такой визуализации в 3D-пространстве, можно заметить, что в определенные моменты филогенеза происходят резкие изменения направления эволюции, то есть значительные изменения в генетических последовательностях (в белках или нуклеиновых кислотах).

Понимание поворотных событий в эволюции инфекционных агентов имеет важное значение для прогнозирования их мутаций и даже принятия превентивных мер для предотвращения условий, в которых происходит ускорение их изменчивости. Проведя ретроспективный анализ условий среды того момента времени, в котором произошло такое изменение, можно выявить и проанализировать факторы давления окружающей среды, которые могли стать критически значимыми для макроизменчивости. Кроме того, понимание закономерностей в эволюции патогеннов способствует выявлению особо изменчивых и консервативных участков белков и/или генома, что необходимо для разработки лекарственных препаратов и средств вакцинопрофилактики.

С технической точки зрения, единственным существенным недостатком нашего подхода является вычислительная сложность построения филогенетического дерева, а также алгоритмов вложения для большого числа различных таксонов.

Несмотря на то, что для визуализации эволюции вируса гриппа в данном исследовании используется только аминокислотная последовательность белка гемагглютинина (HA), представленная визуализация является частичным отражением эволюции. С целью получения более глубоких представлений об эволюции вируса необходимо использовать анализ полных геномов, имеющихся в публичных базах данных. Помимо этого, данный подход не ограничивается исследованиями в вирусологии и может быть применен для анализа эволюции других видов живых существ.

К тому же, если кодировать аминокислоты с помощью известной базы данных AAindex [40], аминокислоты могут быть рассмотрены с точки зрения различных биологических или физико-химических свойств. Это даст возможность визуализировать и изучать эволюцию с различных точек зрения, таких как гидрофильность или липофильность. Помимо AAindex, также возможно применить к нашей визуализации упрощенные аминокислотные алфавиты. В будущем предстоит работа по предложенным методам кодирования и соответствующим критериям (метрикам, пределам и т.д.), которые позволят автоматически распознавать таксономические варианты и кластеры из представленной визуализации.

БЛАГОДАРНОСТИ

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 19-31-60025.

СПИСОК ЛИТЕРАТУРЫ

1. *Orengo C., Jones D., Thornton J.* (ed.). Bioinformatics: genes, proteins and computers // Taylor & Francis, 2003.
2. *Xu X. et al.* Facilitating Antiviral Drug Discovery Using Genetic and Evolutionary Knowledge // Viruses. 2021. V. 13. № 11. P. 2117.

3. *Moelling K., Broecker F.* Viruses and evolution viruses first? A personal perspective // Frontiers in Microbiology. 2019. V. 10. P. 523.
4. *Novella I.S., Presloid J.B., Taylor R.T.* RNA replication errors and the evolution of virus pathogenicity and virulence // Current Opinion in Virology. 2014. V. 9. P. 143–147.
5. *Harvey W.T. et al.* Identification of low-and high-impact hemagglutinin amino acid substitutions that drive antigenic drift of influenza A (H1N1) viruses // PLoS pathogens. 2016. V. 12. № 4. P. e1005526.
6. *Smith D.J. et al.* Mapping the antigenic and genetic evolution of influenza virus // Science. 2004. V. 305. № 5682. P. 371–376.
7. *Forghani M., Khachay M.* Convolutional Neural Network Based Approach to In Silico Non-Anticipating Prediction of Antigenic Distance for Influenza Virus // Viruses. 2020. V. 12. № 9. P. 1019.
8. *Klingen T.R. et al.* In silico vaccine strain prediction for human influenza viruses // Trends in Microbiology. 2018. № 2. P. 119–131.
9. *Jordan G.E., Piel W.H.* web-based visualizations for the tree of life // Bioinformatics. 2008. V. 24. № 14. P. 1641–1642.
10. *Forghani M., Averbukh V.L. et al.* Three-dimensional visualization for phylogenetic tree // Scientific Visualization. 2017. V. 9. № 4. P. 59–66. URL: http://sv-journal.org/2017-4/06/
11. *Авербух В.Л.* Семиотика и основания теории компьютерной визуализации // Философские проблемы информационных технологий и киберпространства. Электронный научный журнал (ISSN: 2305-3763) 2013. № 1. С. 26–41. URL: http://www.cv.imm.uran.ru/e/3241413
12. *Wang C. et al.* Visualization by example // Proceedings of the ACM on Programming Languages. 2019. V. 4. № POPL. P. 1–28.
13. *Ito K. et al.* Gnarled-trunk evolutionary model of influenza A virus hemagglutinin // PloS One. 2011. V. 6. № 10. P. e25953.
14. *Cox M.A.A., Cox T.F.* Multidimensional scaling // Handbook of data visualization. Berlin, Heidelberg: Springer, 2008. P. 315–347.
15. *Neher R.A. et al.* Prediction, dynamics, and visualization of antigenic phenotypes of seasonal influenza viruses // Proceedings of the National Academy of Sciences. 2016. V. 113. № 12. P. E1701–E1709.
16. *Kimura M.* A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences // Journal of molecular evolution. 1980. V. 16. № 2. P. 111–120.
17. *Page R.D.M.* Tree View: An application to display phylogenetic trees on personal computers // Bioinformatics. 1996. V. 12. № 4. P. 357–358.
18. *Galtier N., Gouy M., Gautier C.* SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny // Bioinformatics. 1996. V. 12. № 6. P. 543–548.
19. FigTree. http://tree.bio.ed.ac.uk/software/figtree/.
20. *Letunic I., Bork P.* Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation // Bioinformatics. 2007. V. 23. № 1. P. 127–128.
21. *Robinson O., Dylus D., Dessimoz C.* Phylo. io: interactive viewing and comparison of large phylogenetic trees on the web // Molecular biology and evolution. 2016. V. 33. № 8. P. 2163–2166.
22. *Ranwez V. et al.* PhyloExplorer: a web server to validate, explore and query phylogenetic trees // BMC evolutionary biology. 2009. V. 9. № 1. P. 1–13.
23. *Wang L.G. et al.* Treeio: an R package for phylogenetic tree input and output with richly annotated and associated data // Molecular biology and evolution. 2020. V. 37. № 2. P. 599–603.
24. *Forghani M. et al.* Visualization of the Evolutionary Path: an Influenza Case Study // CEUR Workshop Proceedings. CEUR-WS, 2021. V. 3027. P. 358–368.
25. *Steinparz C.A. et al.* Visualization of Rubik's Cube Solution Algorithms // EuroVA@ EuroVis. 2019. P. 19–23.
26. *Van der Maaten L., Hinton G.* Visualizing data using t-SNE // Journal of Machine Learning Research. 2008. V. 9. № 11.
27. *Grinberg M.* Flask web development: developing web applications with python // O'Reilly Media, Inc., 2018.
28. *Cock P.J.A. et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics // Bioinformatics. 2009. V. 25. № 11. P. 1422–1423.
29. *Pedregosa F. et al.* Scikit-learn: Machine learning in Python // Journal of Machine Learning Research. 2011. V. 12. P. 2825–2830.
30. *Price M.N., Dehal P.S., Arkin A.P.* FastTree 2 approximately maximum-likelihood trees for large alignments // PloS One. 2010. V. 5. № 3. P. e9490.
31. *Stamatakis A.* RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies // Bioinformatics. 2014. V. 30. № 9. P. 1312–1313.
32. *Forghani M. et al.* TBEV analyzer platform for evolutionary analysis and monitoring tick-borne encephalitis virus: 2020 update // Biostatistics & Epidemiology. 2021. P. 1–17.
33. *Baydas S., Karakas B.* Defining a curve as a Bezier curve // Journal of Taibah University for Science. 2019. V. 13. № 1. P. 522–528.
34. Vasev P., Vrungel, https://github.com/viewzavr/vrungel.
35. *Dirksen J.* Learning Three. js: the JavaScript 3D library for WebGL // Packt Publishing Ltd., 2013.
36. *Авербух В.Л., Байдалин А.Ю., Исмагилов Д.Р., Казанцев А.Ю., Тимошпольский С.П.* Использование трехмерных метафор визуализации // 14-я Международная конференция по компьютерной графике

и зрению ГрафиКон'2004, 6–10 сентября, 2004, Москва, Россия. Труды конференции. МГУ им. М.В. Ломоносова. С. 295–298. URL: http://www.cv.imm.uran.ru/e/3549

37. *Wang P. et al.* Predicting influenza antigenicity by matrix completion with antigen and antiserum similarity // Frontiers in Microbiology. 2018. V. 9. P. 2500.

38. *Edgar R.C.* MUSCLE: multiple sequence alignment with high accuracy and high throughput // Nucleic Acids Res. 2004. V. 32. № 5. P. 1792–1797.

39. *John Rozewicki, Songling Li, Karlou Mar Amada, Daron M Standley, Kazutaka Katoh.* MAFFT-DASH: integrated protein sequence and structural alignment // Nucleic Acids Res. 2019. V. 47. Iss. W1. P. W5–W10. https://doi.org/10.1093/nar/gkz342

40. *Kawashima S. et al.* AAindex: amino acid index database, progress report 2008 // Nucleic acids Res. 2007. V. 36. Suppl. 1. P. D202–D205. URL: https://www.genome.jp/aaindex

41. *Mantel N.* The detection of disease clustering and a generalized regression approach // Cancer Res. 1967. V. 27. № 2. Part 1. P. 209–220. PMID 6018555.