

=====ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ =====

Предсказательное моделирование эволюции респираторных вирусов: современные возможности и ограничения

Тычкова В.И.¹, Леоненко В.Н.^{2,1}, Даниленко Д.М.*^{1,2}

¹ФГБУ «НИИ гриппа им. А.А. Смородинцева» МЗ РФ, Санкт-Петербург, Россия

²Университет ИТМО, Санкт-Петербург, Россия

Аннотация. Моделирование траекторий вирусной эволюции имеет первостепенное практическое значение при выборе штаммов, включаемых в состав вакцин против гриппа и SARS-CoV-2. Регулярное обновление состава вакцин оказывает выраженный эффект на их эффективность, который зависит, прежде всего, от соответствия между штаммами, выбранными для включения в вакцину, и штаммами, которые будут циркулировать в предстоящий эпидемический сезон. Указанный выбор перспективных штаммов является сложной задачей, решение которой основывается на большом количестве экспериментальных данных. Кроме того, выбор вакцинных штаммов должен быть осуществлен заранее для заблаговременного развертывания производства вакцин. Развитие методов предсказательного моделирования эволюции вирусов может существенно улучшить соответствие выбранных вакцинных штаммов циркулирующим в предстоящем эпидемическом сезоне, что послужило мотивацией для настоящей работы. В работе рассмотрены основные подходы к моделированию вирусной эволюции, описываются их сильные стороны и ограничения. Производится выбор оптимального подхода для предсказания эволюции вирусов гриппа, циркулирующих на территории России. Представлены результаты использования этого подхода с качественной оценкой его прогностической силы на ретроспективных геномных данных подтипа A(H1N1)pdm09 из базы данных EpiFlu GISAID. Предложена методика оценки соответствия набора наиболее приспособленных штаммов, полученных в результате работы предсказательной модели, штаммам, выбранным экспертами для включения в состав гриппозных вакцин в соответствующие эпидемические сезоны. Данный подход может быть расширен для предсказания эволюции других респираторных вирусов, включая SARS-CoV-2.

Ключевые слова: филогенетический анализ, грипп, SARS-CoV-2, вирусная эволюция, предсказательное моделирование.

ВВЕДЕНИЕ

Возбудители вирусных инфекций наносят значительный ущерб системам здравоохранения всего мира на постоянной основе. Среди них особое место принадлежит возбудителям респираторных вирусных инфекций, прежде всего, вирусам гриппа и SARS-CoV-2, возбудителю COVID-19, а также другим возбудителям острых респираторных вирусных инфекций. Сезонные эпидемии гриппа и ОРВИ, включая вызванные SARS-CoV-2, составляют до 90% всех инфекционных заболеваний,

*daria.baibus@gmail.com

регистрируемых в России и во многих других странах мира [1], и представляют существенное экономическое бремя. В этой связи важнейшей задачей современной медицинской науки является профилактика респираторных вирусных инфекций с целью снижения заболеваемости, госпитализации и смертности от этих инфекций, а также минимизации рисков, связанных с развитием осложнений после перенесенных инфекций [2].

Важнейшим обстоятельством, влияющим на эффективность вакцинации, является соответствие штаммов, введенных в состав вакцин, тем, которые реально циркулируют в тот или иной эпидемический сезон [3, 4, 5]. Проблема адекватного выбора связана с двумя ключевыми обстоятельствами: 1) выбор штамма для включения в состав вакцин происходит заранее, до момента ее применения, и этот интервал значителен по времени (в случае гриппозных вакцин выбор штаммов происходит за 9 месяцев до начала вакцинации); 2) рассматриваемые вирусы являются высокоизменчивыми, что приводит к изменению доминирующих линий на постоянной основе.

Изменения популяции вирусов и смена доминирующих кладов (отдельных генетических групп) происходят неравномерно во времени и пространстве – в разных географических точках мира могут циркулировать различные варианты вирусов. Вакцинные штаммы для гриппозных вакцин рекомендуются ВОЗ для всех производителей гриппозных вакцин один раз в году [6] для каждого из полушарий. Для SARS-CoV-2 в настоящее время такие рекомендации формируются отдельной экспертной группой ВОЗ [7], однако выбор конкретного штамма определяется на страновом уровне или на уровне производителя и может варьировать.

Селекция штаммов для включения их в состав вакцин – важнейшее решение, от которого будет зависеть здоровьесбережение целых стран, в которых вакцинируют от гриппа и SARS-CoV-2. В настоящее время принятие таких экспертных решений осуществляется на основе ключевых генотипических и фенотипических признаков, включающих в себя оценку антигенных и генетических свойств основных кладов, а также данные популяционного иммунитета. При этом экспертный выбор базируется на экспериментальных данных, получение которых является длительным и крайне трудоемким, поскольку требует работы с инфекционными вирусами, получение специфических антисывороток и т.д. [8], что ограничивает возможности проведения массового анализа циркулирующих вариантов вирусов.

В этой связи в последнее десятилетие принимаются активные попытки внедрения прогностических алгоритмов, позволяющих анализировать «приспособленность» вирусов, основные направления их эволюции, влияние отдельных мутаций в определенных участках генома вирусов, кодирующих поверхностные белки (отвечающих за антигенные свойства вирусов) [9]. Кроме того, создаются комплексные модели, позволяющие учитывать и данные заболеваемости, и популяционный иммунитет среди населения отдельных стран или регионов ВОЗ. Основная цель моделей – дать краткосрочный (6 месяцев – 1 год) прогноз эволюции вирусов, с тем чтобы выбрать ту генетическую линию, которая с наибольшей вероятностью будет доминировать в мире или в рассматриваемом регионе через определенный промежуток времени. Такой прогноз призван поддержать решение экспертов, отбирающих вакцинные штаммы на основе экспериментальных данных, а в идеале и дать возможность делать прогноз с использованием минимального числа экспериментальных данных, особенно тех, которые требуют работы с инфекционным вирусом [10].

Развитие предсказательных алгоритмов и моделей стало возможным благодаря массовому внедрению полногеномного секвенирования патогенов в практику геномного надзора: возможность анализировать геномные данные вирусов гриппа и SARS-CoV-2

по всему миру открывает беспрецедентные возможности как для отслеживания геномной эпидемиологии возбудителей, так и для создания моделей вирусной эволюции.

В данной работе был произведён анализ основных подходов к предсказательному моделированию вирусной эволюции с описанием их сильных сторон и ограничений, на базе которого был выбран оптимальный подход для предсказания эволюции вирусов гриппа, циркулирующих на территории России. Представлены результаты использования этого подхода с оценкой его прогностической силы на ретроспективных геномных данных подтипа A(H1N1)pdm09 из базы данных EpiFlu GISAID. Предложена методика оценки соответствия набора наиболее приспособленных штаммов, полученных в результате работы предсказательной модели, штаммам, выбранным экспертами для включения в состав гриппозных вакцин в соответствующие эпидемические сезоны. Необходимо отметить, что группа ОРВИ – это обширная группа возбудителей вирусных инфекций, среди которых в рамках исследований, проводимых авторами, рассматривается только два возбудителя – вирус гриппа (типа А и В) и SARS-CoV-2, поскольку только в отношении этих двух возбудителей существуют средства иммунопрофилактики – специфические вакцины.

2. ОБЗОР СУЩЕСТВУЮЩИХ РЕШЕНИЙ

Для практического использования результатов анализа и предсказательного моделирования эволюции вирусов гриппа и других высокоизменчивых вирусов необходим метод определения степени приспособленности вируса, а также выбор наиболее подходящего метода предсказания связанных с приспособленностью характеристик.

Генетическая приспособленность (genetic fitness, для вирусов – viral fitness) — это ключевое понятие эволюционной биологии, характеризующее способность организма или его генотипа передавать свои гены в последующие поколения. Понятие приспособленности служит основным индикатором адаптивности организма в условиях естественного отбора: высокая приспособленность предполагает большую вероятность передачи генов потомству, что ведет к их распространению в популяции. Как правило, в количественных расчётах используется абсолютная приспособленность (absolute fitness): общее число потомков с заданным генотипом, оставленных организмом. Этот параметр рассматривается в сравнении с другими представителями популяции и зависит от множества факторов, включая репродуктивный успех, выживаемость и адаптацию к условиям среды.

Фактом, затрудняющим количественный анализ и сравнение приспособленности, является низкая практически достижимая точность эмпирической оценки приспособленности для генома целиком или его участков значительной длины. Данное обстоятельство связано с тем, что адаптивный ландшафт среды обитания является существенным, а в некоторых случаях (например, для вирусов) — ключевым фактором, влияющим на приспособленность. Таким образом, оценка не может производиться за счёт использования лабораторной модели популяции — объектом исследования должна являться выборка популяции в естественной среде. При этом, если исследуется не изолированная мутация, а значительный сегмент генома (например, отдельный сегмент вирусного генома), возникают проблемы получения репрезентативной выборки, минимальная требуемая длина которой растёт экспоненциально с ростом длины исследуемого сегмента. При исследовании приспособленности вирусов в факторы окружающей среды входят и свойства популяции носителей. Репрезентативная выборка носителей вируса не гарантирует репрезентативную выборку вирусов, выделенных у этих носителей. Кроме того, важно учитывать, что вирусы с высокой скоростью

изменчивости (например, вирус гриппа, вирус SARS-CoV-2), существуют в среде с непрерывно меняющимся адаптивным ландшафтом ввиду изменений иммунитета популяции носителей. Каждая прошедшая волна заболеваемости (эпидемия) приводит к появлению значительного числа индивидуумов, обладающих иммунитетом в отношении циркулирующих вариантов вируса, и потому, являющаяся к ним уже невосприимчивой.

Описанные выше сложности не делают экспериментальное определение приспособленности вирусов принципиально невозможным, однако его нельзя считать абсолютно достоверным. Важно подчеркнуть, что оценка способности вируса к репликации в клетках носителя и эффективность иммунного отклика для отдельных индивидуумов является практически достижимой, однако достоверное моделирование коллективного иммунитета популяции или воспроизведение популяции носителей с целью оценки эффективности инфицирования не представляются возможными, — а значит, рассчитывать на сугубо экспериментальные методы получения «точного» значения приспособленности нецелесообразно. В силу этой причины использование модельного подхода для количественных оценок приспособленности становится ещё более актуально для прикладных исследований в области эволюции респираторных вирусов.

Методы оценки приспособленности можно разделить на следующие группы:

1. Экспериментальные методы оценки приспособленности, среди которых можно назвать глубокое мутационное сканирование (Deep Mutational Scanning, DMS) [11, 12], исследования на базе клеточных и животных моделей [13, 14], конкурентные эксперименты (Competition Assays) [15].
2. Биоинформатические методы, среди наиболее распространенных - филогенетический анализ [16, 17, 18], моделирование трехмерной структуры вирусных белков и использование алгоритмов анализа влияния мутаций [19, 20, 21], анализ больших данных с использованием алгоритмов машинного обучения (МО) [22, 23].
3. Методы на основе статистических и математических моделей, в частности, статистический анализ распространения [24, 25], анализ соотношения частот несинонимичных и синонимичных замен [26, 27], фитнес-ландшафты [28, 29].

Как видно из представленного обзора применяемых методов, существует большое количество различных способов напрямую или косвенно оценить приспособленность вирусных штаммов, как с большей опорой на экспериментальные результаты и анализ внутренней структуры патогенов, так и на основе статистических и математических подходов. В настоящей работе авторы остановились на методе, который в большей части использует общие математические принципы анализа филогенетических деревьев, связанные с процессами рождения и гибели, и позволяет анализировать приспособленность с помощью анализа генетических последовательностей.

3. ЧИСЛЕННЫЕ ЭКСПЕРИМЕНТЫ

1. Описание метода

В качестве наиболее подходящего метода для анализа эволюции респираторных вирусов в Российской Федерации было решено выбрать метод прогнозирования эволюции вируса гриппа, предложенный в публикации Неера и др. [30]. Метод основывается на оценке приспособленности вирусов на основе анализа филогенетических деревьев. Рассматривая эпидемические по гриппу сезоны как «поколения», авторы показывают,

что для субъединицы белка HA1 вируса гриппа A/H3N2 (основной антигенной детерминанты вирусов гриппа) такой метод позволяет предсказать «основного» (т.е. наиболее приспособленного) «родителя» штаммов следующего поколения не хуже, чем популярный и потому взятый в качестве референсного метода алгоритм на основе анализа эпитопных мутаций из работы [32].

Авторы метода исходят из предположения о постепенном непрерывном характере мутаций – таком, при котором с одной стороны, происходит непрерывный генетический дрейф, но с другой — отсутствуют резкие «скачкообразные» мутации, что согласуется с известными данными об изменчивости сезонных вирусов гриппа [33]. В рамках описанного предположения о непрерывном характере мутаций авторы строят математическую модель оценки правдоподобности генеалогического дерева для некоторого набора последовательностей, а на основе этой модели — итеративный метод построения наиболее правдоподобного генеалогического дерева. На основе структуры восстановленного генеалогического дерева выполняется оценка приспособленности различных штаммов. Распределение приспособленностей внутренних узлов филогенетического дерева $x = x_0, x_1, \dots$ вычисляется на основе формулы

$$P(x|T) = \frac{p_0(x_0)}{Z(T)} \prod_{i=0}^{n_{int}} g(x_{i_1}, t_{i_1} | x_i, t_i) g(x_{i_2}, t_{i_2} | x_i, t_i),$$

где $P(x|T)$ – совместное распределение вероятностей для вектора приспособленностей внутренних узлов дерева $x = x_0, x_1, \dots, x_{n_{int}}$ и дерева T , являющееся оценкой правдоподобности дерева. В этой формуле $p_0(x)$ – распределение приспособленностей в популяции, i – индекс узла, n_{int} – число внутренних узлов дерева, индексы i_1, i_2 обозначают двух потомков узла i , $Z^{-1}(T)$ – нормализующий множитель. Множители вида $g(\cdot|\cdot)$ называются «распространителями по ветвям» (branch propagators) и описывают вероятность связи между предком с приспособленностью x_i в момент времени t_i с потомком (не обязательно прямым) с приспособленностью x_j в момент времени $t_j > t_i$. При этом принято предположение, что ветвь, соединяющая узлы i и j , подразумевает, что все выбранные потомки i также являются потомками j (нет ветвления между i и j). Величина значения $P(x|T)$ может быть рассчитана итерационно путём нескольких проходов по филогенетическому дереву с помощью алгоритма «передачи сообщений» (message passing) от узла к узлу.

В качестве дальнейшего шага авторы метода Неера [30] предлагают эвристическую метрику LBI (local branching index) $\lambda_i(\tau)$ – «локальный коэффициент ветвления», который коррелирует с приспособленностью штаммов и может быть использован для её оценки. В отличие от непосредственного расчёта апостериорного распределения приспособленности $P(x|T)$, алгоритм расчёта $\lambda_i(\tau)$ не зависит от значений параметров модели динамики распространения величины приспособленности вдоль ветвей, поэтому использование данной величины для ранжирования узлов более удобно в практическом смысле.

Таким образом, итоговое ранжирование вирусных штаммов, циркулирующих в настоящий момент в популяции, по их прогнозируемой приспособленности, т.е. способности дать многочисленное потомство в следующем эпидемическом сезоне, опирается на показатель LBI и получается для заданного генеалогического дерева выполнением следующей цепочки шагов:

1. На вход алгоритма подаётся последовательность секвенированных вирусных штаммов.

2. На основе методов максимального правдоподобия производится восстановление предков секвенированных штаммов (внутренние узлы дерева). В [30] для этого предлагается использовать метод FastTree [31], состоящий из следующих шагов:
 - (a) В качестве начального приближения алгоритма на основе загруженных геномных последовательностей строится дерево с помощью метода эвристического присоединения соседей (Heuristic Neighbor-Joining).
 - (b) Производится техническое сокращение длины дерева путём сокращения лишних добавленных узлов.
 - (c) Итеративно максимизируется правдоподобие дерева.
3. Для всех узлов дерева – внутренних, являющимися предками, и внешних, т. е. листьев дерева, которые соответствуют секвенированным штаммам, – рассчитываются значения приспособленности с помощью индикатора LBI;
4. Производится ранжирование листьев дерева по значениям LBI, что даёт упорядоченный список наиболее приспособленных из секвенированных штаммов, могущих получить максимальное распространение в грядущем эпидемическом сезоне.

В настоящей работе исследуется применимость метода к белку НА (гемагглютинина) в целом (в отличие от рассмотрения отдельного домена НА1) на примере современных вирусов гриппа A(H1N1)pdm09, за весь период их распространения в России с 2009 по 2024 гг. Данные геномных последовательностей для анализа взяты из базы EpiFlu GISAID [34]. Цель исследования состояла в оценке применимости восстановленных генеалогических деревьев для выбора доминирующей группы циркулирующих штаммов вирусов гриппа в определенном сезоне и ее соответствия штамму, предложенному в качестве компонента гриппозных вакцин для подтипа A(H1N1)pdm09 экспертами Всемирной организации здравоохранения.

В отличие от статьи-источника, в настоящей работе метод Neher et al. используется на существенно иной выборке входных данных. Ключевыми отличиями является использование последовательностей белка НА целиком, выбор гриппа А подтипа H1N1(pdm09), отказ от бутстреппинга и географической локализации исследуемых образцов. Первое влияет на «зашумлённость» данных. В то время как авторы демонстрируют, что метод не является чувствительным к внеэпитопным мутациям, выявляя «ключевые» замены при построении генеалогического дерева, важно убедиться, что это свойство сохраняется и для последовательностей, содержащих множество антигенных сайтов с продолжительными некодирующими (в антигенном смысле) участками. Выбор другого подтипа вируса гриппа А обусловлен необходимостью подтвердить широкую применимость метода. Авторы метода Неера использовали бутстреппинг для ресемплинга выборки на основе географической локализации штаммов, — в таком случае для образцов, распространенность в одном регионе (нами были выбраны образцы, собранные на территории РФ), необходимость в таком усреднении отсутствует.

2. Данные

Исходные данные, данные после обработки, код исполняемых скриптов, а также результаты их выполнения опубликованы на портале GitHub*. Алгоритм предварительной обработки данных включал в себя следующие шаги:

* <https://github.com/TychkovaV/FitnessInference>

1. Исходные данные представляли собой только полные нуклеотидные последовательности белка НА (1752 п.о.) вируса гриппа A(H1N1)pdm09 образцов, секвенированных в России за сезоны 2009-2024 гг., и депонированных в базу данных EpiFlu GISAID (файл RF-2009-2024.fasta).
2. Далее было проведено разбиение выборки на подвыборки с различными временными периодами регистрации последовательностей. Все последовательности начинались в 2009 г. и продолжались до каждого следующего фиксированного года, формируя таким образом следующие временные диапазоны: 2009 г., 2009-2010 гг., 2009-2011 гг., ..., 2009-2024 гг. за исключением 2009-2019 гг. и 2009-2022 гг. Условно началом эпидемического сезона года считается 1 сентября, а концом — 31 мая следующего года (скрипт `fasta_filter.py`)
3. Выполнялось множественное выравнивание каждой из подвыборок с помощью средства MAFFT [35], из полученных выравниваний с помощью утилиты AliView [36], исключались выбросы и не кодирующие последовательности по краям выравнивания (скрипт `align.sh`).
4. Наконец, с помощью SRAToolbox [37] после удаления дубликатов имён последовательностей, совпадающие последовательности были объединены под общим именем: например, при полном совпадении последовательностей «А» и «В» они заменялись последовательностью с названием «А=В».
5. Получившиеся выравнивания были использованы как входные данные для метода Неера. Приложенный к исследованию [30] открытый код был отлажен и доработан для совместимости с компилятором языка Python3, что, в отличие от исходного кода, позволяет обеспечить воспроизводимость результатов статьи. Также были дописаны процедуры для визуализации филогенетических деревьев для облегчения интерпретации получаемых результатов, которые в первоначальной версии были недостаточно информативны.

3. Результаты

Для каждой из исследованных подвыборок были построены генеалогические деревья с цветовой палитрой, соответствующей приспособленности группы: от синей (наименьшая приспособленность) к красной (наибольшая). Образцы, относящиеся к последнему сезону (и только к нему, в случае если последовательность соответствует нескольким образцам) были помечены символом «*».

В связи с тем что построенные филогенетические деревья содержат несколько тысяч геномов в каждой подвыборке, приведение рисунков в статье (даже в кластеризованном виде) не представляется возможным; авторы отсылают читателя к изображениям, размещенным на портале GitHub[†] [35].

Среди результатов наиболее интересными для обсуждения являются результаты предсказания доминирующих линий (вариантов) для последнего эпидемического сезона по гриппу 2023-2024 гг. Проведенный анализ показал, что среди множества отдельных кластеров вирусов гриппа A(H1N1)pdm09, циркулировавших в данном эпидемическом сезоне, возможно отобрать три кластера с наибольшим значением LBI, условно отражающим наивысшую приспособленность вирусов, относящихся к данному кластеру (таблица 1).

[†] <https://github.com/TychkovaV/FitnessInference>

Таблица 1. Репрезентативные штаммы вирусов гриппа из кластеров, обладающих наивысшей приспособленностью по данным локального коэффициента ветвления LBI

Наименование вируса	Ранг	LBI	Ранг родителя	LBI родителя
Сезонные вирусы гриппа, циркулировавшие в России в сезон 2023-2024 и ранее				
A/Chita/RII-MH128099S/2022 EPI_ISL_17646635	24	1,158	20	3,025
A/Yaroslavl/301-t14V/2023 EPI_ISL_17762850	90	0,657	55	1,962
A/Gorno-Altai/RII-MH124790S/2022 EPI_ISL_16738418	153	0,542	35	2,420
Вакцинные штаммы для компонента A(H1N1)pdm09, предложенные экспертами ВОЗ				
A/Wisconsin/67/2022 EPI_ISL_15928563	2120	0,168	344	0,659
A/Victoria/4897/2022 EPI_ISL_18879345	2355	0,153	541	0,460

Интересно отметить, что родительский LBI во всех случаях был выше, чем LBI потомка, что совпадает с мнением авторов метода, говорящих о том, что наиболее приспособленные родители дают более приспособленных потомков. Кроме того, необходимо отметить, что группа вирусов, получившая доминирование в сезоне 2023-2024 гг. в России, была сформирована и начала свою циркуляцию до начала указанного эпидсезона, т.к. схожие по свойствам вирусы циркулировали и в более ранний период (что подтверждается третьим по уровню приспособленности штаммом из более раннего 2022 года).

Следующий этап исследования являлся ответ на вопрос, насколько же вирусы, циркулировавшие в России, соответствовали штаммам, рекомендованным экспертами ВОЗ для включения в состав гриппозных вакцин по компоненту A(H1N1)pdm09.

Позиционирование вакцинных штаммов A/Wisconsin/67/2022 и A/Victoria/4897/2022 на дереве, состоящем из штаммов, циркулировавших только на территории России приводит к тому, что указанные вакцинные штаммы имеют очень высокий ранг и низкий LBI. Однако необходимо заметить, что вакцинные штаммы отбираются на основе данных всего мира, а не его отдельных областей, в связи с чем полученный результат возможно считать закономерным, поскольку исследовалась только узкая географическая область, пусть и с большой территорией.

Вместе с тем, проведенный анализ аминокислотных замен в выбранных нами репрезентативных штаммах из доминирующих кластеров показал, что все штаммы имели одни и те же замены по сравнению с вакцинными (таблица 2). Это свидетельствует о том, что все три доминирующие варианта обладали схожими аминокислотными профилями, которые определили их эволюционное преимущество.

Анализ аминокислотных отличий показал, что многие из них связаны с изменением свойств вирусов. Так, мутации S137P и A216T в белке гемагглютинина (HA) вируса H1N1pdm09 связаны с повышенной специфичностью связывания с рецепторами, способствуя улучшенной адаптации к рецепторам клеток человека (-2,6-типа) и влияет на эффективность репликации вируса; мутация K142R вовлечена в изменение антигенных эпитопов, что может повлиять на распознавание вируса антителами хозяина. Мутация A277T связана с модификациями в участках гликозилирования, что также потенциально изменяет вирусную антигенность и повышает его способность уклоняться от иммунного

Таблица 2. Аминокислотные отличия между репрезентативными штаммами вирусов гриппа, циркулировавших в России и кандидатными штаммами гриппозных вакцин

Наименование штамма/ Номер АК позиции в HA	137	142	216	223	260	277	356	451
Репрезентативные штаммы из доминирующих кластеров сезонных вирусов гриппа, циркулировавших в России								
A/Chita/RII-MH128099S/2022	P	K	T	Q	D	T	E	N
A/Gorno-Altai/RII-MH124790S/2022	P	K	T	Q	D	T	E	N
A/Yaroslavl/301-t14V/2023	P	K	T	Q	D	T	E	N
Вакцинные штаммы								
A/Victoria/4897/2022	S	R	A	R	E	A	D	H
A/Wisconsin/67/2022	S	R	T	Q	E	A	D	H

ответа. Замена R223Q определяется только субстратом выращивания вируса – эта замена характерна для штаммов а(H1N1)pdm09, культивируемых в куриных эмбрионах, что необходимо для производства большинства гриппозных вакцин. Вакцинный штамм A/Wisconsin/67/2022 не имеет этой замены, поскольку он предназначен для производства вакцин на основе клеточных культур. Мутация E260D может влиять на стабильность белка HA. Роль других мутаций в настоящее время требует уточнения. Таким образом, большинство выявленных отличий связаны с приспособлением вируса к циркуляции в условиях предсуществующего иммунного ответа у индивидуумов в популяции (оба вакцинных штамма – 2022 года).

4. ВЫВОДЫ

Проведенные исследования с использованием филогенетических деревьев для определения эволюции вирусов гриппа позволили выявить наиболее приспособленные генетические кластеры вирусов, циркулировавшие в Российской Федерации в сезоне 2023–2024 гг. Оценка их наибольшей приспособленности, базирующаяся на локальном коэффициенте ветвления, позволила идентифицировать эволюционно доминирующие варианты исключительно с использованием данных секвенирования, что открывает широкие возможности с точки зрения оптимизации методов оценки приспособленности вирусов гриппа и их соответствия штаммам, введенным в состав вакцин.

Для дальнейшего изучения применимости метода в оценке эволюции сезонных вирусов гриппа и возможности предсказания доминирующих групп следует в первую очередь поставить вопрос о том, насколько согласованы понятия «основной родитель следующего поколения» и «оптимальный вакцинный штамм». Очевидна также необходимость учитывать другие свойства вирусов, такие как патогенность штамма, а также исследовать адаптивные ландшафты: к примеру, скорость генетического дрейфа может быть недостаточной для того, чтобы наиболее приспособленный штамм гриппа стал наиболее распространённым в очередной сезонной эпидемии. Наконец, важнейшим аспектом при выборе вакцинных штаммов гриппа для борьбы с сезонными вспышками является антигенное расстояние, которое не является синонимом генетического расстояния, а потому протокол выбора вакцинных штаммов на основе исследуемого метода потребует внедрения исследования эффективности перекрёстного иммунитета (например, на основе антигенных карт [38]), или других методов предсказания антигенных свойств циркулирующих вирусов как заключительного этапа.

Помимо эмпирической проверки точности прогнозов важно оценить расхождение оценки приспособленности исследуемого метода с существующими аналитическими методами (например, на основе эпигенетических мутаций [32]). Дополнительным важным направлением для дальнейших исследований является введение алгоритма или, по крайней мере, эвристики для подбора оптимальных параметров применения метода на основе минимальных исходных данных, например, исключительно даты получения образца — хотя бы для географически локализованной выборки.

Наконец, текущее представление результатов в виде дерева без аннотации не может считаться адекватным, — как, впрочем, и использованное в настоящей работе представление обширных генеалогических деревьев целиком. Таким образом, актуальным и крайне важным является вопрос о методе компактной, но содержательной визуализации результатов.

В заключение необходимо отметить, что бремя гриппа сохраняется несмотря на прогресс в области вакцинопрофилактики. Совершенствование гриппозных вакцин является важнейшей задачей современности. Однако технологии производства вакцин в настоящее время являются длительными, и требуют выбора штаммов-кандидатов за 9–12 месяцев до реального применения вакцин, что дает вирусу длительный период для эволюции. Предсказание эволюции хотя бы на 6 месяцев вперед даст новые возможности для совершенствования гриппозных вакцин и повышения их эффективности. Это свидетельствует о необходимости развития предсказательного моделирования с целью совершенствования стратегий выбора вакцинных штаммов, а также о потенциальной необходимости выбора вакцинных штаммов для отдельных стран и территорий при наличии возможности производить собственные гриппозные вакцины.

Работа выполнена при поддержке гранта НЦМУ «Передовые цифровые технологии» (в рамках Соглашения о предоставлении гранта от 20.04.2022 № 075-15-2022-313).

СПИСОК ЛИТЕРАТУРЫ

1. Global Burden of Disease Study 2013 Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013. *Lancet*. 2015. V. 386. No. 9995. P. 743–800. doi: 10.1016/S0140-6736(15)60692-4
2. Waterer G. The global burden of respiratory infectious diseases before and beyond COVID. *Respirology*. 2023. V. 28. No. 2. P. 95–96. doi: 10.1111/resp.14423.
3. Adabor E.S. A statistical analysis of antigenic similarity among influenza A (H3N2) viruses. *Heliyon*. 2021. V. 7. No. 11. Article No. e08384. doi: 10.1016/j.heliyon.2021.e08384.
4. Russell C.A, Jones T.C., Barr I.G., Cox N.J., Garten R.J., Gregory V., Gust I. D., Hampson A.W., Hay A.J, Hurt A.C et al. Influenza vaccine strain selection and recent studies on the global migration of seasonal influenza viruses. *Vaccine*. 2008. V. 26. P. D31–D34. doi: 10.1016/j.vaccine.2008.07.078
5. WHO Writing Group; Ampofo W.K., Baylor N., Cobey S., Cox N.J., Cox N.J., Daves S., Edwards S., Ferguson N., Grohmann G., Hay A., Katz J., et al. Improving influenza vaccine virus selection: report of a WHO informal consultation held at WHO headquarters, Geneva, Switzerland, 14-16 June 2010. *Influenza Other Respir. Viruses*. 2012. V. 6. No. 2. P. 142–152. doi: 10.1111/j.1750-2659.2011.00277.x
6. Zhang W., Hirve S., Kieny M.P. Seasonal vaccines - Critical path to pandemic influenza response. *Vaccine*. 2017. V. 35. P. 851–852.
7. *Technical Advisory Group on COVID-19 Vaccine Composition Home Page*. URL: [https://www.who.int/groups/technical-advisory-group-on-covid-19-vaccine-composition-\(tag-co-vac\)](https://www.who.int/groups/technical-advisory-group-on-covid-19-vaccine-composition-(tag-co-vac)). (дата обращения: 20.11.2024).
8. Petrova V., Russell C. The evolution of seasonal influenza viruses. *Nat. Rev. Microbiol.* 2018. V. 16. P. 47–60. doi: 10.1038/nrmicro.2017.118
9. Morris D.H, Gostic K.M., Pompei S., Bedford T., Łuksza M., Neher R.A., Grenfell B.T., Lässig M., McCauley J. W. Predictive Modeling of Influenza Shows the Promise of Applied Evolutionary Biology. *Trends Microbiol.* 2018. V. 26. No. 2. P. 102–118. doi: 10.1016/j.tim.2017.09.004
10. Gouma S., Anderson E.M., Hensley S.E. Challenges of Making Effective Influenza Vaccines. *Annu. Rev. Virol.* 2020. V. 7. No. 1. P. 495–512. doi: 10.1146/annurev-virology-010320-044746
11. Dadonaite B., Ahn J.J., Ort J.T., Yu J., Furey C., Dosey A., Hannon W.W., Vincent Baker A. L., Webby R.J. et al. Deep mutational scanning of H5 hemagglutinin to inform influenza virus surveillance. *PLoS Biology*. 2024. V. 22. No. 11. Article No. e3002916. doi: 10.1371/journal.pbio.3002916
12. Starr T.N., Greaney A.J., Hilton S.K., Ellis D., Crawford K.H.D., Dingens A.S., Navarro M.J., Bowen J.E., Tortorici M.A., Walls A.C., et al. Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. *Cell*. 2020. V. 182. No. 5. P. 1295–1310.e1–e10. doi: 10.1016/j.cell.2020.08.012
13. Timofeeva T.A., Rudneva I.A., Lomakina N.F., Timofeeva E.B., Kupriyanova I.M., Lyashko A.V., Shcherbinin D.N., Shilov A.A., Shmarov M.M., Ryazanova E.L., Mochalova L.V., Timofeev B.I. Mutations in the genome of avian influenza viruses of the H1 and H5 subtypes responsible for adaptation to mammals. *MIR J*. 2021. V. 8. No. 1. P. 50–61. doi: 10.18527/2500-2236-2021-8-1-50-61
14. Lan J., Ge J., Yu J., Shan S, Zhou H., Fan S., Zhang Q., Shi X., Wang Q., Zhang L., Wang X. Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature*. 2020. V. 581. P. 215–220. doi: 10.1038/s41586-020-2180-5

15. Di Franco S., Bianca P., Sardina D.S., Turdo A., Gaggianesi M., Veschi V., Nicotra A., Mangiapane L.R., Iacono M.L., Pillitteri I. et al. Adipose stem cell niche reprograms the colorectal cancer stem cell metastatic machinery. *Nat. Commun.* 2021. V. 12. Article No. 5006. doi: 10.1038/s41467-021-25333-9
16. Rambaut A., Pybus O., Nelson M., Viboud C., Taubenberger J.K., Holmes E.C. The genomic and epidemiological dynamics of human influenza A virus. *Nature*. 2008. V. 453. P. 615–619. doi: 10.1038/nature06945
17. Ning Z., Nan Z., Huafeng F., Jie D., Xingyu X., Xiaoqing D., Xiaoxiao D., Dandan X., Xiaoyu M., Yan Y., Hongjin G., Lingfeng M., Min H. Mutations and Phylogenetic Analyses of SARS-CoV-2 Among Imported COVID-19 From Abroad in Nanjing, China. *Frontiers in Microbiology*. 2022. V. 13. doi: 10.3389/fmicb.2022.851323
18. Lefrancq N., Duret L., Bouchez V., Brisse S., Parkhill J., Salje H. Learning the fitness dynamics of pathogens from phylogenies. *medRxiv*. 2023. No. 2023.12.23.23300456. doi: 10.1101/2023.12.23.23300456
19. Waterhouse A., Bertoni M., Bienert S., Studer G., Tauriello G., Gumienny R., Heer F.T., de Beer T.A.P., Rempfer C., Bordoli L., Lepore R., Schwede T. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Research*. 2018. V. 46. No. W1. P. W296–W303. doi: 10.1093/nar/gky427
20. Waman V.P., Ashford P., Lam S.D., Sen N., Abbasian M., Woodridge L., Goldtzvik Y., Bordin N., Wu J., Sillitoe I., Orengo C.A. Predicting human and viral protein variants affecting COVID-19 susceptibility and repurposing therapeutics. *Scientific Reports*. 2024. V. 14. No. 1. Article No. 14208.
21. Padhi A.K., Kalita P., Maurya S., Poluri K.M., Tripathi T. From De Novo Design to Redesign: Harnessing Computational Protein Design for Understanding SARS-CoV-2 Molecular Mechanisms and Developing Therapeutics. *J. Phys. Chem. B*. 2023. V. 127. No. 41. P. 8717–8735. doi: 10.1021/acs.jpcc.3c04542
22. Li J., Zhang S., Li B., Hu Y., Kang X.P., Wu X.Y., Huang M.T., Li Y.C., Zhao Z.P., Qin C.F., Jiang T. Machine learning methods for predicting human-adaptive influenza A viruses based on viral nucleotide compositions. *Molecular biology and evolution*. 2020. V. 37. No. 4. P. 1224–1236.
23. Hie B., Zhong E.D., Berger B., Bryson B. Learning the language of viral evolution and escape. *Science*. 2021. V. 371. No. 6526. P. 284–288.
24. Holder B. P., Simon P., Liao L.E., Abed Y., Bouhy X., Beauchemin C.A.A., Boivin G. Assessing the in vitro fitness of an oseltamivir-resistant seasonal A/H1N1 influenza strain using a mathematical model. *PloS One*. 2011. V. 6. No. 3. Article No. e14767. doi: 10.1371/journal.pone.0014767
25. Kepler L., Hamins-Puertolas M., Rasmussen D.A. Decomposing the sources of SARS-CoV-2 fitness variation in the United States. *Virus Evolution*. 2021. V. 7. No. 2. Article No. veab073. doi: 10.1093/ve/veab073
26. Gu X. $\Delta\text{-}d_N/d_S$: A New Criteria to Distinguish among Different Selection Modes in Gene Evolution *bioRxiv*. 2020. No. 2020.02.21.960450. doi: 10.1101/2020.02.21.960450
27. Tonkin-Hill G., Martincorena I., Amato R., Lawson A.R.J., Gerstung M., Johnston I., Jackson D.K., Park N., Lensing S.V., Quail M.A. al Patterns of within-host genetic diversity in SARS-CoV-2. *eLife*. 2021. V. 10. Article No. e66857. doi: 10.7554/eLife.66857
28. Lei R., Hernandez Garcia A., Tan T.J.C., Teo Q.W., Wang Y., Zhang X., Luo S., Nair S.K., Peng J., Wu N.C. Mutational fitness landscape of human influenza H3N2 neuraminidase. *Cell reports*. 2023. V. 42. No. 1. Article No. 111951.
29. Flynn J.M., Samant N., Schneider-Nachum G., Barkan D.T., Yilmaz N.K., Schiffer C.A., Moquin S.A., Dovala D., Bolon D.N.A. Comprehensive fitness landscape of SARS-CoV-2

- Mpro reveals insights into viral resistance mechanisms. *Elife*. 2022. V. 11. Article No. e77433.
30. Neher R.A., Russell C.A., Shraiman B.I. Predicting evolution from the shape of genealogical trees. *eLife*. 2014. V. 3. Article No. e03568. doi: 10.7554/eLife.03568
 31. Price M.N., Dehal P.S., Arkin A.P. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Molecular biology and evolution*. 2009. V. 26. No. 7. P. 1641–1650.
 32. Łuksza M., Lässig M. A predictive fitness model for influenza. *Nature*. 2014. V. 507. P. 57–61. doi: 10.1038/nature13087
 33. Bedford T., Riley S., Barr I.G., Broor S., Chadha M. Global circulation patterns of seasonal influenza viruses vary with antigenic drift. *Nature*. 2015. V. 523. No. 7559. P. 217–220. doi: 10.1038/nature14460
 34. Shu Y, McCauley J. GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill*. 2017. V. 22. No. 13. Article No. 30494. doi: 10.2807/1560-7917.ES.2017.22.13.30494
 35. Katoh K., Standley D.M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*. 2013. V. 30. No. 4. Article No. 772–780. doi: 10.1093/molbev/mst010
 36. Larsson A. AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics*. 2014. V. 30. No. 22. P. 3276–3278. doi: 10.1093/bioinformatics/btu531
 37. Aparicio-Puerta E., Lebrón R., Rueda A., Gómez-Martín C., Giannoukakos S., Jaspez D., Medina J.M., Zubkovic A., Jurak I., Fromm B., Marchal J.A., Oliver J., Hackenberg M. sRNAbench and sRNAtoolbox 2019: intuitive fast small RNA profiling and differential expression. *Nucleic Acids Research*. 2019. V. 47. No. W1. P. W530–W535. doi: 10.1093/nar/gkz415
 38. Smith D.J., Lapedes A.S., de Jong J.C., Bestebroer T.M., Rimmelzwaan G.F. Mapping the antigenic and genetic evolution of influenza virus. *Science*. 2004. V. 305. No. 5682. P. 371–376. doi: 10.1126/science.1097211

Рукопись поступила в редакцию 12.12.2024.

Переработанный вариант поступил 27.12.2024.

Дата опубликования 31.12.2024.

Predictive Modeling of Respiratory Virus Evolution: Current Capabilities and Limitations

Tychkova V.I. * ¹, Leonenko V.N. ^{2,1}, Danilenko D.M. ^{1,2}

¹Smorodintsev Research Institute for Influenza, St. Petersburg, Russia

²ITMO University, St. Petersburg, Russia

Abstract. Modeling of viral evolution trajectories is of primary practical importance when selecting strains to be included in influenza and SARS-CoV-2 vaccines. Regular updating of the vaccine composition has a pronounced effect on their efficacy, which depends, first of all, on the correspondence of the strains selected for inclusion in the vaccine to those that will circulate during the epidemic season. Given the complexity of the choice, which is based on a large amount of experimental data, and the need to make it in advance for early deployment of vaccine production, the development of predictive methods can significantly improve the correspondence of the selected vaccine strains to those circulating in the upcoming epidemic season. The article considers the main approaches to modeling viral evolution, describes their strengths and limitations. The optimal approach is selected for predicting the evolution of influenza viruses circulating in Russia. The results of using this approach with an assessment of its predictive power on retrospective genomic data of the A(H1N1)pdm09 subtype from the EpiFlu GISAID database are presented. A method for assessing the correspondence of the set of the most adapted strains obtained as a result of the predictive model to the strains selected by experts for inclusion in influenza vaccines in the corresponding epidemic seasons is proposed. This approach can be expanded to predict the evolution of other respiratory viruses, including SARS-CoV-2.

Key words: *phylogenetic analysis, influenza, SARS-CoV-2, viral evolution, predictive modeling.*