МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ
БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ФАКУЛЬТЕТ ПРИКЛАДНОЙ МАТЕМАТИКИ И ИНФОРМАТИКИ
Кафедра дискретной математики и алгоритмики

ТЫЛЕЦКИЙ Павел Сергеевич

СТАТИСТИЧЕСКИЕ МЕТОДЫ ИССЛЕДОВАНИЯ РАСПРОСТРАНЕНИЯ ШТАММОВ ВИРУСОВ

Магистерская диссертация
1-31 80 09 «Прикладная математика и информатика»

Научный руководитель:
Баханович Сергей Викторович, кандидат физико-математических наук

Допущена к защите
«____» _____ 2021 г.
Заведующий кафедрой дискретной математики и алгоритмики

_____
Котов Владимир Михайлович
доктор физико-математических наук, профессор

Минск, 2021

ОГЛАВЛЕНИЕ

ПЕРЕЧЕНЬ УСЛОВНЫХ ОБОЗНАЧЕНИЙ И СОКРАЩЕНИЙ 3

ВВЕДЕНИЕ 4

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ 5

ГЛАВА 1. ОСНОВЫ ФИЛОГЕНЕТИКИ 8

1.1 Мутационная эволюция 8
1.2 Секвенирование ДНК 9
1.3 Задача построения филогенетических деревьев 10
1.4 Штамм B.1.1.7 11

ГЛАВА 2. ПОДГОТОВКА ДАННЫХ 14

2.1 Декомпозиция 14
2.2 Выбор штаммов 14
2.3 Нормализация данных 15
2.4 Линейная аппроксимация 18

ГЛАВА 3. БАЙЕСОВСКИЙ ВЫВОД ОБОБЩЕННОЙ ЛОГИСТИЧЕСКОЙ МОДЕЛИ 20

3.1 Обоснование 20
3.2 Математическая модель эпидемических волн с перекрывающимися субэпидемиями 22
3.3 Использование ОЛМ с субэпидемиями для оценки трансмиссивности 25
3.4 Байесовский вывод 27
3.5 ОЛМ с байесовским выводом 30
3.6 Результаты 32

ЗАКЛЮЧЕНИЕ 34

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 36

ПРИЛОЖЕНИЕ А 37

ПЕРЕЧЕНЬ УСЛОВНЫХ ОБОЗНАЧЕНИЙ И СОКРАЩЕНИЙ

ДНК – дезоксирибонуклеиновая кислота
ОЛМ – обобщенная логистическая модель
ВО – вариант озабоченности

ВВЕДЕНИЕ

Тема магистерской диссертации посвящена развитию методов моделирования распространения эпидемий и оценки трансмиссивности штаммов вирусов. Данные методы являются основным инструментом прогнозирования эволюции эпидемии, своевременного выявления штаммов, представляющих угрозу, и численной оценки их опасности. Применение данных методом организациями общественного здравоохранения позволяет своевременно и точно реагировать на возникшую угрозу, планировать и корректировать ответные мероприятия по ограничению распространения эпидемии, работу медицинских учреждений и протоколы лечения. Таким образом, учитывая крайне сложную эпидемиологическую обстановку в мире в связи с широким распространением вируса SARS-CoV-2 и его стремительными мутациями, разработка достоверных и эффективных математических моделей распространения штаммов является крайне актуальной задачей.

Рассматриваемая в работе тема анализа филогенетических данных связана в первую очередь с исследованием эволюционных сценариев вирусов с целью выявления высокозаразных штаммов. В связи с широким распространением вирусов вырастает вероятность мутаций, приводящим к значительным изменениям характеристик вируса, в том числе трансмиссивности – вероятности передачи другому носителю.

Популярность анализа генома обусловлена резким снижением стоимости секвенирования в последние годы, а также широким вовлечением исследователей и правительств в борьбу с пандемией COVID-19, что позволило получить существенное количество данных: за 2020г. секвенировано более 200 тысяч последовательностей вируса, полученных анализами у пациентов в разное время года по всему миру. Все эти данные, в том числе и сопутствующая метаинформация, находятся в открытом доступе.

Трудностей в анализе генетических данных несколько. С одной стороны, инструменты проведения анализа не являются высокоточными. С другой, сами по себе генетические данные не несут исчерпывающей информации о глобальном развитии эпидемии. Однако их большое количество позволяет делать эвристические выводы и детектировать потенциально интересные для более пристального рассмотрения штаммы.

Инструментальной базой для работы являются открытые базы данных секвенированных различными методами образцов вирусов, позволяющие проводить информационный анализ (проверять построенные теории и разработанные методы), не вкладывая в это дополнительных (значительных на фоне стоимости самого секвенирования и необходимости проверки большого количества гипотез) затрат.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Магистерская диссертация, 40 страниц, 14 рисунков, 14 источников, 1 приложение.

СТАТИСТИЧЕСКИЕ МЕТОДЫ ИССЛЕДОВАНИЯ РАСПРОСТРАНЕНИЯ ШТАММОВ ВИРУСОВ

Ключевые слова: SARS-CoV-2, БИОИНФОРМАТИКА, ФИЛОГЕНИЯ, ФИЛОГЕНЕТИЧЕСКИЕ ДЕРЕВЬЯ, СТАТИСТИЧЕСКИЙ АНАЛИЗ, МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ, ОБОБЩЕННАЯ ЛОГИСТИЧЕСКАЯ МОДЕЛЬ, ТРАНСМИССИВНОСТЬ, СУБЭПИДЕМИЧЕСКАЯ МОДЕЛЬ, БАЙЕСОВСКИЙ ВЫВОД.

Объект исследования — математические модели прогнозирования эпидемии и их применение для оценки трансмиссивности.

Цели работы — разработка математической модели оценки относительной трансмиссивности штаммов вирусов; оценка трансмиссивности штамма B.1.1.7 вируса SARS-CoV-2.

Результат работы — разработана математическая модель оценки относительной трансмиссивности штаммов вирусов и полученная с ее помощью оценка трансмиссивности штамма B.1.1.7 вируса SARS-CoV-2.

Область применения — оперативное выявление высокотрансмиссивных штаммов вирусов и численная оценка их трансмиссивности в целях общественного здравоохранения и научных исследований в области эпидемиологии.

АГУЛЬНАЯ ХАРАКТЭРЫСТЫКА ПРАЦЫ

Магістарская дысертацыя, 40 старонак, 14 малюнкаў, 14 крыніц, 1 дадатак.

СТАТЫСТЫЧНЫЯ МЕТАДЫ ДАСЛЕДВАННЯ РАСПАЎСЮДЖВАННЯ ШТАМАЎ ВІРУСАЎ

Ключавыя словы: SARS-COV-2, БІЯІНФАРМАТЫКА, ФІЛАГЕНІЯ, ФІЛАГЕНЭТЫЧНЫЯ ДРЭВЫ, СТАТЫСТЫЧНЫ АНАЛІЗ, МАТЭМАТЫЧНАЕ МАДЭЛЯВАННЕ, АБАГУЛЬНЕНАЯ ЛАГІСТЫЧНАЯ МАДЭЛЬ, ТРАНСМІСІЎНАСЦЬ, СУБЭПІДЭМІЧНАЯ МАДЭЛЬ, БАЙЕСАЎСКАЯ ВЫСНОВА.

Аб'ект даследавання — матэматычныя мадэлі прагназавання эпідэміі і іх выкарыстанне з мэтай вызначэння узроўню трансмісіўнасці.

Мэты працы — распрацоўка матэматычнай мадэлі вызначэння узроўню трансмісіўнасці штамаў вірусаў; вызначэнне узроўню трансмісіўнасці штама B.1.1.7 віруса SARS-CoV-2.

Вынікі — распрацована матэматычная мадэль вызначэння узроўню трансмісіўнасці штамаў вірусаў і атрыман з яе дапамогай узровень трансмісіўнасці штама B.1.1.7 віруса SARS-CoV-2.

Галіна выкарыстання — хуткае вызначэнне высокатрансмісіўных штамаў вірусаў і лічбавая характэрыстыка іх трансмісіўнасці дзеля грамадзкай аховы здароўя і навуковых даследваній у галіне эпідэміялогіі.

ABSTRACT

Master thesis, 40 pages, 14 figures, 14 sources, 1 appendix.

STATISTICAL METHODS OF VIRUS STRAINS DISTRIBUTION RESEARCH

Keywords: SARS-CoV-2, BIOINFORMATICS, PHILOGENY, PHILOGENETIC TREES, STATISTICAL ANALYSIS, MATHEMATICAL MODELING, GENERALIZED LOGISTIC MODEL, TRANSMISSION, SUBEPIDEMIC MODEL, BAYESIAN INFERENCE.

Object of study — mathematical models for forecasting the epidemic and their application to assess transmissibility.

Objectives of study — development of a mathematical model for assessing the relative transmissibility of virus strains; estimation of the transmissibility of the SARS-CoV-2 virus strain B.1.1.7.

Results — a mathematical model for assessing the relative transmissibility of viral strains and an estimate of the transmissibility of the SARS-CoV-2 virus strain B.1.1.7 obtained with its help.

Field of application — rapid detection of highly transmissible viral strains and the numerical assessment of their transmissibility for the purposes of public health and scientific research in the field of epidemiology.

ГЛАВА 1.
ОСНОВЫ ФИЛОГЕНЕТИКИ

1.1 Мутационная эволюция

Рассмотрим ситуацию, в которой вирус порождается одной клеткой и распространяется до момента проведения исследования. Распространение вируса связано с постоянным воспроизведением его частиц. Однако механизмы репликации, рекомбинации и репарации ДНК не совершенны и вызывают мутации – стойкие изменения участков генома. Обычно при рассмотрении подобных процессов используют следующие два предположения:

- Предположение о бесконечности генома:
  - Существует бесконечное количество оснований, в которых может произойти мутация;
  - Ни одна мутация не происходит в одном и том же основании за историю развития вируса.
- Предположение о наследовании генома: однажды представленная мутация будет представлена во всех потомках данной клетки.

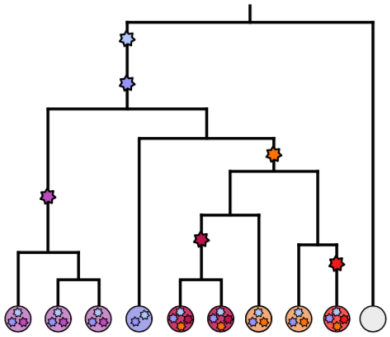

Рис. 1. Пример генеалогического дерева. Шестиконечные звезды обозначают мутации, случившиеся во время деления клетки.

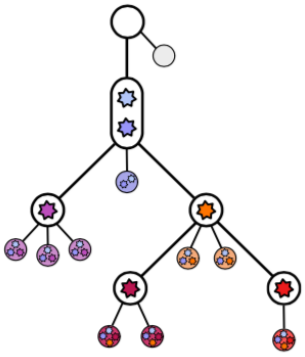

Рис. 2. Дерево мутаций, эквивалентное генеалогическому дереву на Рис. 1

Таким образом, можно выделить штаммы – группы клеток с одинаковым геномом. Если собрать всю историю эволюции вируса, то все ее штаммы можно собрать в иерархическую структуру – бинарное дерево, в котором вершинами являются клетки, а ребра показывают на отношение наследования. Такая структура называется генеалогическим деревом (рис. 1).

Альтернативным представлением может считаться дерево мутаций (рис. 2), в котором вершинами являются наборы мутаций, а ребра показывают на последовательность появление мутаций в ходе эволюции.

Отдельно стоит упомянуть, что данные исследований [1] указывают на невыполнение описанных выше предположений в полученных данных. Это заставляет в разработки методов анализа генетических данных использовать различные минимизационные и вероятностные подходы.

1.2 Секвенирование ДНК

Секвенирование – процесс определения нуклеотидной последовательности, в результате которого получают формальное описание структуры молекулы в текстовом виде. Инструменты для секвенирования требуют большое количество (многократно большее содержащегося в одной клетке) генетического материала для анализа[2], что обуславливает разделение методов секвенирования на две основные группы: массовое и одноклеточное.

Массовое, в первую очередь – секвенирование нового поколения, представляет собой обработку группы клеток с целью определения частоты встречающихся аллелей, что позволяет составить представление о колониях – группах клеток с одинаковым геномом.

Основными свойствами такого подхода являются получение данных о большом сегменте ткани разом, выделение крупнейших колоний и дешевизна подхода. С то же время, все малочисленные популяции теряются в результате такого анализа, а данные о соотношении клеток требуют дополнительной обработки и не гарантируют восстановление информации об изучаемой популяции.

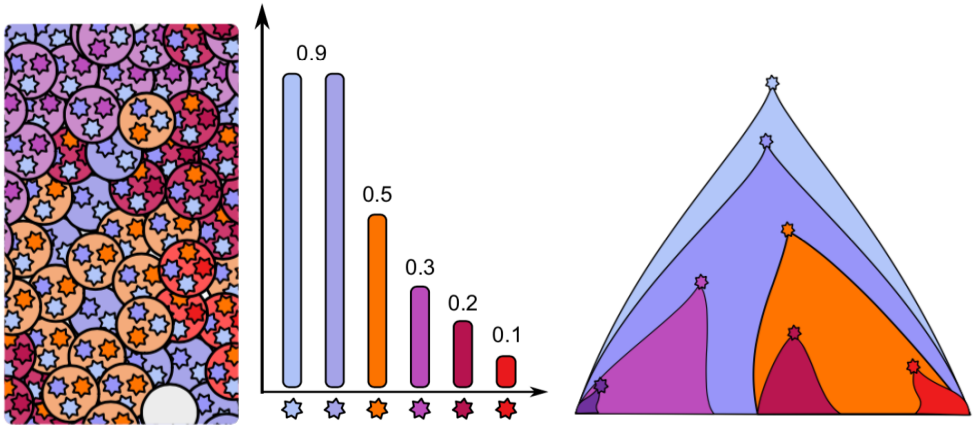

Рис. 3. Слева направо – схематическое изображение исследуемой популяции клеток; частота встречаемости мутации в исследуемой популяции; схематическое изображение экспансии мутаций в исследуемой популяции.

Одноклеточное секвенирование направлено на получение генетической информации одной конкретной клетки. Для этого используются инструменты выделения точечной клетки из ткани и полимеразная цепная реакция для увеличения количества генома до минимального распознаваемого секвенатором. Основное преимущество ОКС заключается в том, что после увеличения количества генетического материала распределение нуклеотидных последовательностей остается таким же, как и в исходной клетке, а значит, появляется возможность довольно точно восстановить нуклеотидную последовательность исходной клетки.

Стоит заметить, что используемые технологии имеют довольно значительные ошибки оценок. Методы ОКС, например, имеют следующие априорные оценки на ошибки секвенирования[3]: ложноположительные (определение некоторой последовательности в случае, когда она не представлена в анализируемой клетке) до 10^{-5} % и ложноотрицательные более 10%, что обуславливает необходимость использовать эти предварительные знания для минимизации влияния ложноотрицательных ошибок. С другой стороны, довольно точные оценки на ошибки в конкретных методах ОКС позволяют использовать полученные апостериорные вероятности ошибок для оценки полученного результата алгоритма более высокого уровня.

1.3 Задача построения филогенетических деревьев

Для оценки эволюционных процессов удобно использовать филогенетические деревья. Однако их построение является нетривиальной задачей, не имеющей решения, гарантированно достоверно отражающего эволюционные процессы. Тема построения деревьев не является задачей данной работы, однако следует учитывать характеристики анализируемых деревьев. Среди таких есть следующие:

- Для заданного набора последовательностей не существует однозначно верного дерева, описывающего мутационную эволюцию. Это исходит как из ошибок самого секвенирования, так и из невыполнения предположений из пункта 1.1;
- Нестабильность – незначительные изменения входных данных могут привести к значительному изменению результата;
- Отсутствие локальности – свойства близких (по количеству отличающихся аллелей) последовательностей находится близко (по расстоянию до ближайшего общего предка) в построенном дереве;
- Отсутствие учета хронологии проведения анализов, которая может помочь упорядочить мутации и тем самым отнести их к одному поддереву.

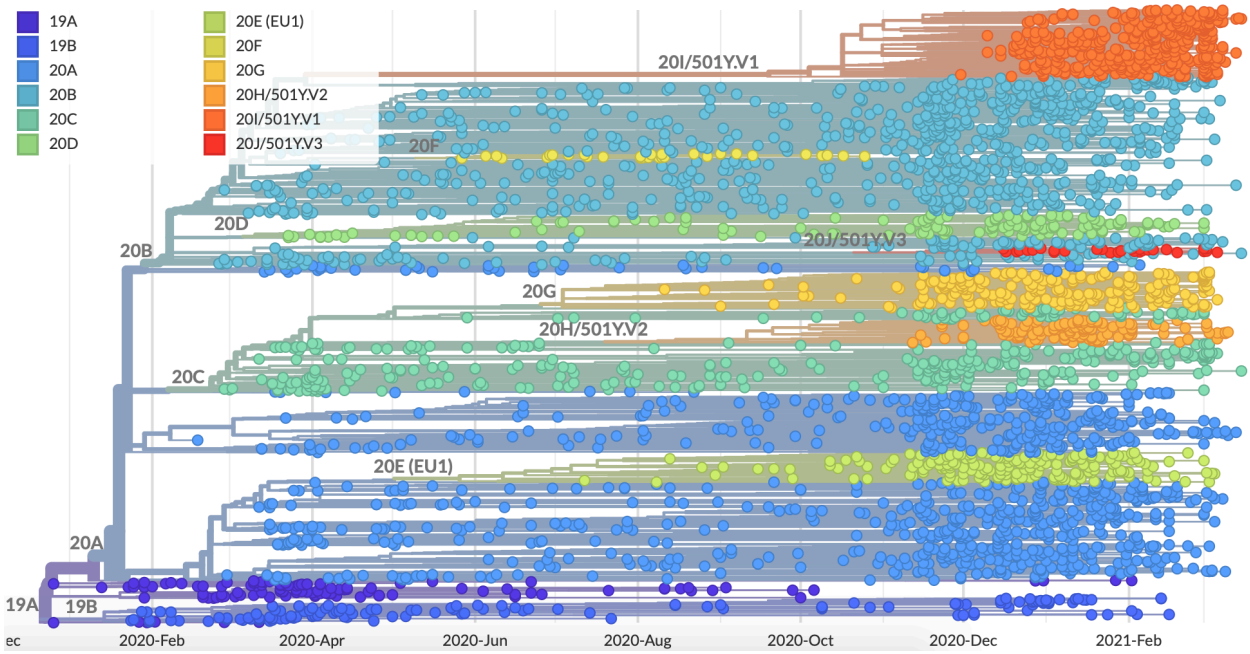

Рис. 4. Филогенетическое дерево на временной шкале для вируса SARS-CoV-2. Оранжевым цветом выделено поддерево штамма B.1.1.7

1.4 Штамм B.1.1.7

Новый штамм SARS-CoV-2, первоначально называвшийся вариантом B.1.1.7 и «вариантом озабоченности» (BO), быстро расширяет свой географический диапазон и частоту встречаемости в Великобритании и других странах мира. Штамм была обнаружен в ноябре 2020 года, и, вероятно, возник в сентябре 2020 года в юго-восточном регионе Англии. Штамм обладает большим количеством несинонимичных замен, имеющих иммунологическое значение.

Предварительный анализ[4] этих наблюдений представлен на рис. 5. Скорость молекулярной эволюции внутри штамма B.1.1.7 аналогична скорости молекулярной эволюции других родственных штаммов. Однако штамм B.1.1.7 больше отличается от филогенетического корня пандемии, что указывает на более высокую скорость молекулярной эволюции на филогенетической ветви, непосредственно предшествующей B.1.1.7.

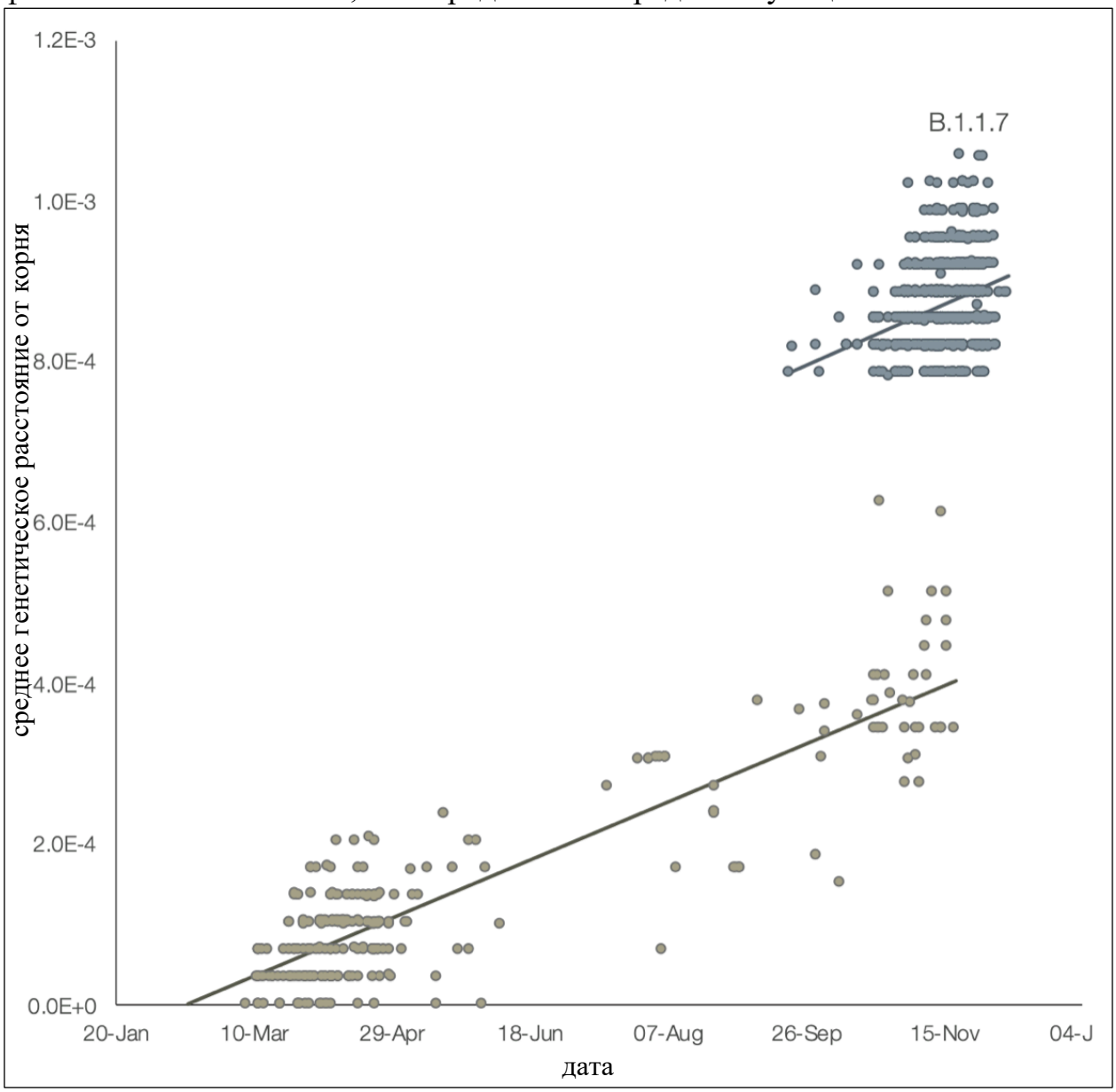

Рис. 5. Регрессия генетических расстояний от корня до листьев в зависимости от дат получения данных для последовательностей, принадлежащих к штамму B.1.1.7, и последовательностей, входящих в его непосредственную внешнюю группу в глобальном филогенетическом дереве. Регрессия проведена для двух наборов независимо друг от друга.

Филогенетические исследования[4, 5], проведенные UK COVID-19 Genomics Consortium, предоставили первое указание на то, что BO имеет необычный набор замен и растет с большей скоростью по сравнению с другими циркулирующими штаммами. Были проанализированы полные геномы BO, собранные в период с начала октября по 5 декабря 2020 года, и

обнаружено, что скорость увеличения частоты встречаемости ВО соответствует большей трансмиссивности по сравнению с другими циркулирующими штаммами в Великобритании. Чтобы подтвердить эти результаты, были исследованы временные тенденции в пропорции ПЦР-тестов, демонстрирующих изменения в спайковом белке, по всей Великобритании (275000 тестов) в качестве биомаркера ВО, и исследована взаимосвязь между ростом местной эпидемии и частотой ВО. Продемонстрировано, что увеличение базового репродуктивного числа связано с увеличением частоты изменений в спайковом белке среди зарегистрированных случаев – биомаркера инфекции ВО, и подтверждает связь с помощью аналитических подходов. Важно отметить, что находятся доказательства того, что нефармацевтических вмешательств было достаточно для контроля не-ВО линий до базового репродуктивного числа ниже 1 во время блокировки в ноябре 2020 года в Англии, но в то же время было недостаточно для контроля ВО.

ГЛАВА 2.
ПОДГОТОВКА ДАННЫХ

2.1 Декомпозиция

Для анализа распространения штаммов был выбран подход локального анализа – разбиения данных на регионы средней гранулярности (до десятков тысяч секвенированных последовательностей). Этот выбор обусловлен следующими причинами:

- Разделение данных на более мелкие группы позволяет замечать опасные тренды, еще не успевшие распространиться за пределы такой группы, намного раньше, так как внутри группы заражения происходят намного интенсивнее, чем за ее пределами;
- Ограничения построенного филогенетического дерева не позволяют однозначно трактовать поддеревья как штаммы, так как они могут быть сгруппированы ошибочно;
- Построение и обработка дерева для имеющегося количества последовательностей (более 200 тысяч последовательностей суммарным объемом более 6Гбайт) требует значительных вычислительных ресурсов.

В исходном наборе данных возможность группировки по территориальному признаку заканчивается на уровне государств и штатов (только для крупных федеративных государств). В связи с этим были отобран ряд стран, данные из которых содержат достаточное для анализа количество данных, но при этом количество этих данных не превышает вычислительные возможности применения алгоритмов по построению деревьев.

Для последовательностей, сгруппированным по выбранным странам, были построены филогенетические деревья, которые были использованы для дальнейшего анализа.

2.2 Выбор штаммов

Перед исследованием свойств штаммов следует определить способы выделения штаммов для анализа и их именования. Исходя из того, что в филогенетическом дереве последовательности сгруппированы по некоторой близости (в отношении мутаций), можно сделать вывод, что листья некоторого поддерева будут являться представителями соответствующего ему штамма. Однако все поддеревья не представляют практический интерес для исследования и требуется осуществить предварительный отбор.

Микроорганизмы, вирусы и плазмиды имеют обозначения, состоящие из букв и порядковых номеров. Обычно рекомендуется включать в обозначение инициалы работника или описательный символ местности, лаборатории и так далее. Каждому новому штамму, мутанту, изоляту или производному присваивается новое (серийное) обозначение. Чтобы избежать использования того же обозначения, что и у известного штамма, обозначение сверяется с базой данных публикации. На практике же это означает, что такой подход (не основанный на уникально идентифицирующих данных штамма) не дает возможности определить положение некоторого штамма в дереве исключительно из его названия. В связи с этим приходится обращаться к метаинформации, которая содержит описательные данные штамма (но даже при таком подходе полной автоматизации процесса выделения штаммов добиться будет крайне сложно). В данной работе выборка целевых штаммов осуществлялась вручную по списку несинонимичных замен генома на визуализированном представлении филогенетического дерева.

На первом этапе для простоты деревья выбираются исходя из их размера: подходящими считаем деревья размером от 30 до 100 вершин, что позволяет как увидеть тенденции в эволюционной динамике штамма через изменение частоты встречаемости этого штамма, так и не пропустить вновь появившиеся мутации.

Вторым этапом из данных стоит отсеять те штаммы, которые оказались нежизнеспособны и пропали из анализов достаточно давно, так как они либо были замещены более приспособленными штаммами и тем самым не представляют угрозу, либо были ошибочно сгруппированы именно таким образом алгоритмом построения деревьев.

2.3 Нормализация данных

Анализируя динамику распространения штаммов, нельзя не заметить, что распределение изначальных данных не является равномерным, что обусловлено количеством заболевших людей в целом и заинтересованностью исследователей в проведении подобных исследований. По этой причине данные требуется нормализовать, то есть привести количество последовательностей штамма к виду, соответствующему равномерному распределению исходной статистики.

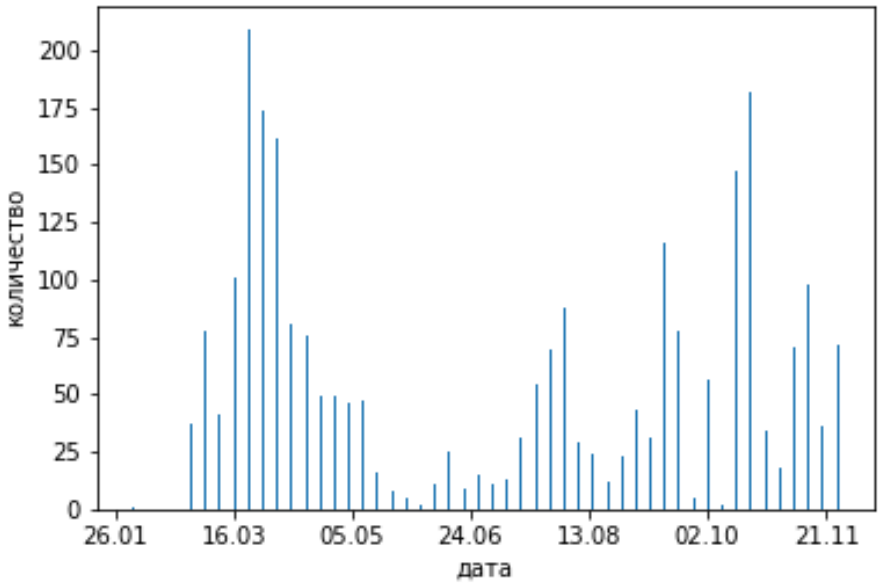

Рис. 6. Гистограмма всех анализов из Бельгии

Для нормализации данных были проверены два способа:

- Нормализация через полиномиальную аппроксимацию и вычисление доли. Такой способ позволяет не привязываться к сетке исходной гистограммы и получать приближенные значения исходного распределения для нормализации по ним, что дает возможность сгладить тренды исходного распределения (интервал порядка года) и использовать их в более мелкой сетке для анализируемого штамма (интервал порядка нескольких месяцев). Минусами такого подхода являются:
    - Возможные отрицательные значения полученного полинома на заданном отрезке, что противоречит смыслу приближаемой функции и заставляет отдельно обрабатывать такие ситуации в дискретной нормализации;
    - Значительные флуктуации для выбросов и малых величин. Таким образом, в даты с максимальным и минимальным количеством анализом отношения истинного и приближенного значения становятся слишком велики и вносят серьезные искажения в нормализованные данные.

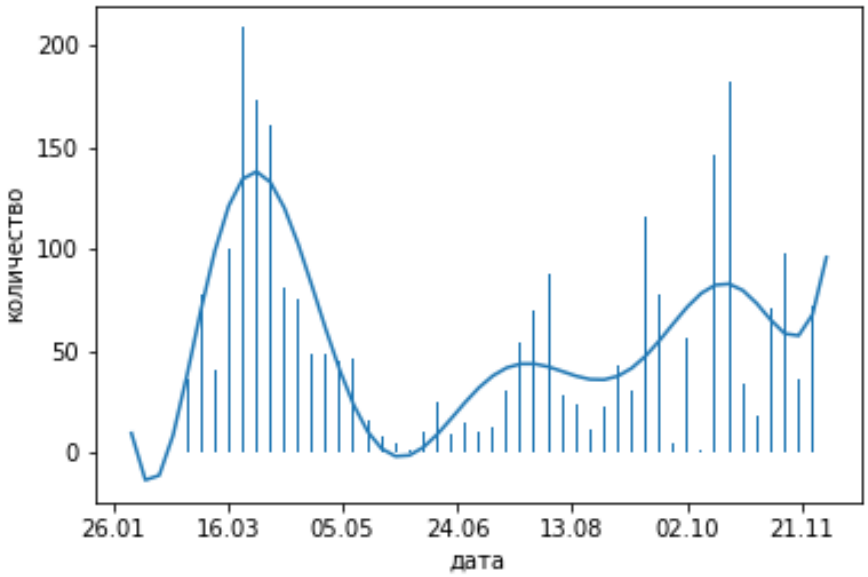

Рис. 7. Гистограмма и многочлен степени 10, аппроксимирующий ее значения

- Прямое вычисления доли по гистограмме с фиксированной шириной ячейки. Является более простым и стабильным способом, позволяющим получить точные значения нормализованных данных, однако ширина ячейки должна быть выбрана заранее. В данной работе для этого заранее вычисляется гистограмма всех данных для страны с шириной ячейки в одну неделю и используется для последующей нормализации. Выбор ширины ячейки обусловлен сглаживанием периодических колебаний в рамках недели (например, сильно меньшее количество анализов в выходные дни) при минимальном размере ячейки. Однако такой способ плохо отвечает потребности анализировать распространение недавно появившегося штамма, так как он может занимать всего несколько ячеек гистограммы.

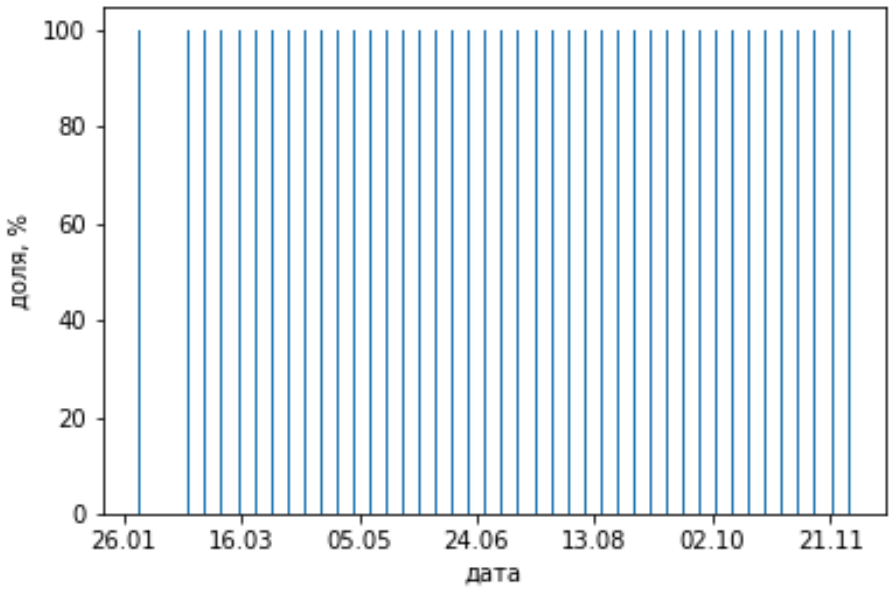

Рис. 8. Нормализованная гистограмма всех анализов Бельгии (оригинал на рис. 6)

Для более точной нормализации данных можно использовать тренд-сезонные подходы, которые позволят сгладить не только глобальные флуктуации данных (волны эпидемии), но и локальные в рамках недели.

2.4 Линейная аппроксимация

Основной задачей данной работы является детектирование штаммов вирусов, которые демонстрируют устойчивый быстрый рост своей популяции. Для этого требуется иметь инструменты численной оценки такого неформального понятия. Для этого опишем поставленную задачу.

Предположим, что на входе алгоритма имеется корректное филогенетическое дерево некоторого вируса и нормализованные метаданные для него – список всех анализов и сопутствующей информации о них, в первую очередь даты проведения исследований. Требуется спроектировать алгоритм, способный численно характеризовать степень скорости роста количества случаев встречания представителей заданного штамма (поддерева филогенетического дерева).

Рассмотрим простейшим способ анализа роста значений последовательности. Таковым является приближение значений гистограммы прямой. Такая прямая описывает собой тренд – направление роста/убывания функции. В качестве численного значения можно использовать коэффициент наклона прямой.

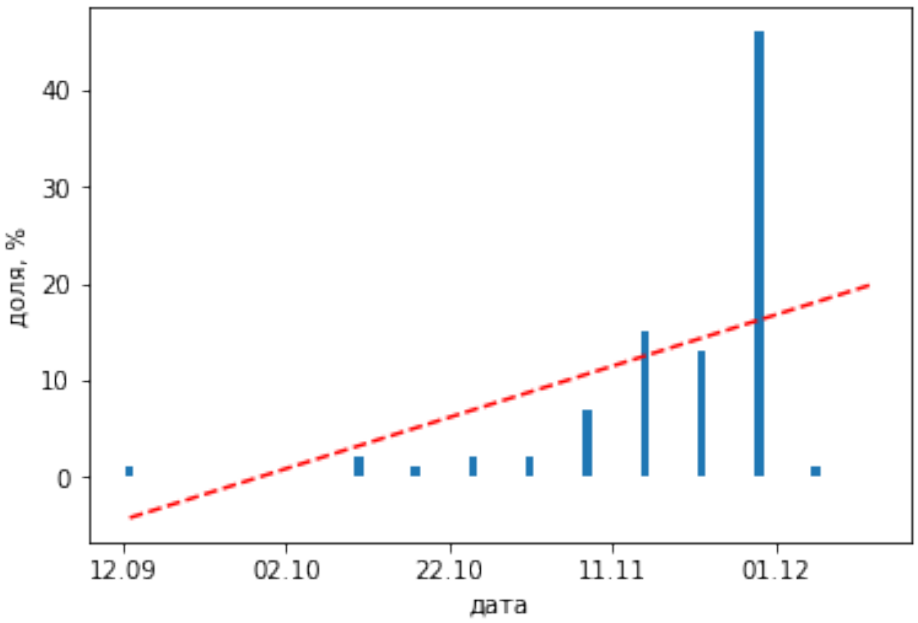

Рис. 9. Пример использования линейной аппроксимации

Несмотря на простоту данного способа, он может быть использован для предварительного отсеивания неинтересующих штаммов, имеющих

затухающую или нейтральную динамику распространения. Метод обладает следующими плюсами:

- Точная численная оценка глобального тренда;
- Вычислительная простота;
- Устойчивость к выбросам;

Среди минусов же можно выделить:

- Потеря информации о локальных трендах;
- Слабый учет затухания в конце периода;
- Отсутствие выравнивания на временные рамки (исходя из предположения, что доминирующий штамм должен продолжать расти вплоть до самых актуальных последовательностей);
- Отсутствие выравнивания последней ячейки гистограммы (может быть исправлена на этапе нормализации путем прогнозирования или отрезания ячейки).

ГЛАВА 3.
БАЙЕСОВСКИЙ ВЫВОД ОБОБЩЕННОЙ ЛОГИСТИЧЕСКОЙ МОДЕЛИ

3.1 Обоснование

Множество взаимосвязанных и часто ненаблюдаемых факторов, влияющих на распространение патогенов в различных пространственных и временных масштабах, создают серьезные проблемы для прогнозирования динамики передачи инфекционных заболеваний. Факторами, влияющими на передачу инфекционного заболевания, являются: способ передачи, сеть взаимодействий, история болезни, поведение отдельных лиц, ответные меры системы здравоохранения, изменение поведения в ответ на эпидемию, а также коллективный иммунитет населения, сформированный генетическими факторами, взаимодействием с вирусом или вакцинацией. Способность составлять точные прогнозы эпидемии усложнена недостаточностью данных на индивидуальном и групповом уровнях, которые влияют на динамику передачи инфекционных заболеваний.

Точность прогнозов эпидемий также снижается из-за отсутствия подробных данных о частоте вспышек и контактных данных. Обычно модели прогнозирования должны основываться на совокупных зарегистрированных случаях заболеваемости, выявленных при появлении симптомов или постановке диагноза. Данные о заболеваемости эпидемиями являются ценным эпидемиологическим инструментом для оценки и прогнозирования тенденций и потенциала передачи в режиме реального времени. Однако агрегированные данные о случаях редко содержат информацию о путях передачи и других характеристиках населения, необходимых для создания реалистичной модели передачи болезни. Например, в течение первых нескольких месяцев эпидемии Эболы в Западной Африке[5] в 2014 – 2016 гг. Всемирная организация здравоохранения публиковала еженедельные кривые эпидемии на национальном уровне для Гвинеи, Либерии и Сьерра-Леоне. Напротив, вирус Эбола сначала поразил деревню Гекеду в Гвинее, и цепи передачи быстро пересекли близлежащие границы Сьерра-Леоне и Либерии. Следовательно, кривые эпидемии с более точным пространственным и временным разрешением, охватывающие соответствующие взаимодействующие сообщества, были бы более подходящими для оценки модели распространения и руководства усилиями по сдерживанию заболевания.

Недостаточные данные об эпидемии ограничивают сложность математических моделей с точки зрения механизмов и количества параметров, оценку которых можно построить на основе данных. В этих моделях часто

используется структура метапопуляции для учета однородности путем разделения населения на социально-демографические группы на основе восприимчивости к вирусу, моделей передвижения или других индивидуальных характеристик, связанных с динамикой распространения заболевания. Предполагается, что особи в одной группе однородны, а неоднородность популяции ограничивается числом групп. Даже когда количество параметров, которые можно оценить на основе ограниченных данных, невелико, модель должна включать достаточно сложную математическую закономерность, чтобы учесть лежащую в основе динамику передачи. Исследования[6,7] показывают, что простые модели роста логистического типа имеют тенденцию недооценивать пиковые сроки и продолжительность эпидемических вспышек. Кроме того, простые феноменологические модели роста логистического типа, как правило, могут поддерживать только одноволновую траекторию эпидемии, характеризующуюся единственным пиком числа новых инфекций, за которым следует период «выгорания», если только нет внешних движущих сил, таких как сезонные изменения контактов.

Для улучшения оценки параметров можно использовать структуру моделирования субэпидемии, которая поддерживает различные траектории волн эпидемии, включая стабильные модели заболеваемости с устойчивыми или затухающими колебаниями. Для этого население делится на группы и используются перекрывающиеся субэпидемии в этих группах в качестве математических строительных блоков для понимания и прогнозирования эпидемии, наблюдаемой в более грубых масштабах. Следовательно, наблюдаемая в крупном масштабе эпидемия создается из совокупности перекрывающихся субэпидемий в группах, которые имеют регулярную структуру. Эти субэпидемии обычно не наблюдаются и формируются неоднородностью населения. Группы определяются восприимчивостью основных популяций, моделями мобильности населения, естественным течением болезни, перераспределением между разными группами риска, различными вмешательствами в области общественного здравоохранения и быстро меняющимися факторами окружающей среды. Такой подход позволяет моделировать прогноз в зависимости от изменений в составе отдельных групп на основе временных изменений здравоохранения или местных изменений в поведении, которые влияют на заболеваемость в данной территориальной области или субпопуляциях.

В гетерогенных популяциях крупномасштабные эпидемии редко можно охарактеризовать простой математической моделью. Подход пересекающихся субэпидемических блоков помогает понять, как разложить крупномасштабные модели эпидемических волн на несколько кривых заболеваемости, которые могут быть сформированы множеством факторов. Крупномасштабную эпидемическую волну можно исследовать как

совокупность регулярных и перекрывающихся субэпидемий, которые связаны друг с другом некоторым систематическим образом. Это уменьшает количество свободных параметров, необходимых для соотнесения субэпидемий друг с другом.

3.2 Математическая модель эпидемических волн с перекрывающимися субэпидемиями

Для модели эпидемических волн с перекрывающимися субэпидемиями каждую групповую субэпидемию моделируем с помощью обобщенной логистической модели роста (ОЛМ), которая показала многообещающие результаты для краткосрочного прогнозирования траектории новых вспышек инфекционных заболеваний[8]. ОЛМ задается следующим дифференциальным уравнением:

$$\frac{dX(t)}{dt} = f\ X(t)^d \left(1 - \frac{X(t)}{Q_0}\right), \tag{1}$$

где $\frac{dX(t)}{dt}$ описывает кривую заболеваемости во времени t. Кумулятивное количество случаев в момент времени t определяется выражением $X(t)$, в то время как f – положительный параметр, обозначающий скорость роста, Q_0 – окончательный размер эпидемии, а $\mathrm{d} \in [0,1]$ – параметр масштабирования роста. Если $d = 0$, это уравнение описывает постоянную заболеваемость во времени, а если $d = 1$, уравнение становится моделью экспоненциального роста. Промежуточные значения d описывают субэкспоненциальные модели роста.

Затем моделируем эпидемическую волну, состоящую из n перекрывающихся субэпидемий, имеющих однородную структуру, с использованием следующей системы дифференциальных уравнений:

$$\frac{dX_i(t)}{dt} = f_i\ A_i(t)\ X_i(t)^d \left(1 - \frac{X_i(t)}{Q_i}\right), \tag{2}$$

где $X_i(t)$ отслеживает кумулятивное количество инфекций для субэпидемии i, а Q_i – размер i-й субэпидемии, где $\mathrm{i} = 1..\mathrm{n}$. Таким образом, модель сводится к простой модели логистического типа, когда $\mathrm{n} = 1$. Чтобы смоделировать время начала $(\mathrm{i}+1)$-й субэпидемии, мы используем индикаторную функцию $A_i(t)$ так, чтобы субэпидемии, включающие эпидемическую волну, имели регулярную структуру, потому что $(\mathrm{i}+1)$-я субэпидемия запускается, когда

кумулятивное количество случаев для i-й субэпидемии превышает общее количество случаев Q_{thr} и перекрывается, поскольку $(i+1)$-я субэпидемия начинается до того, как i-я завершает свое течение.

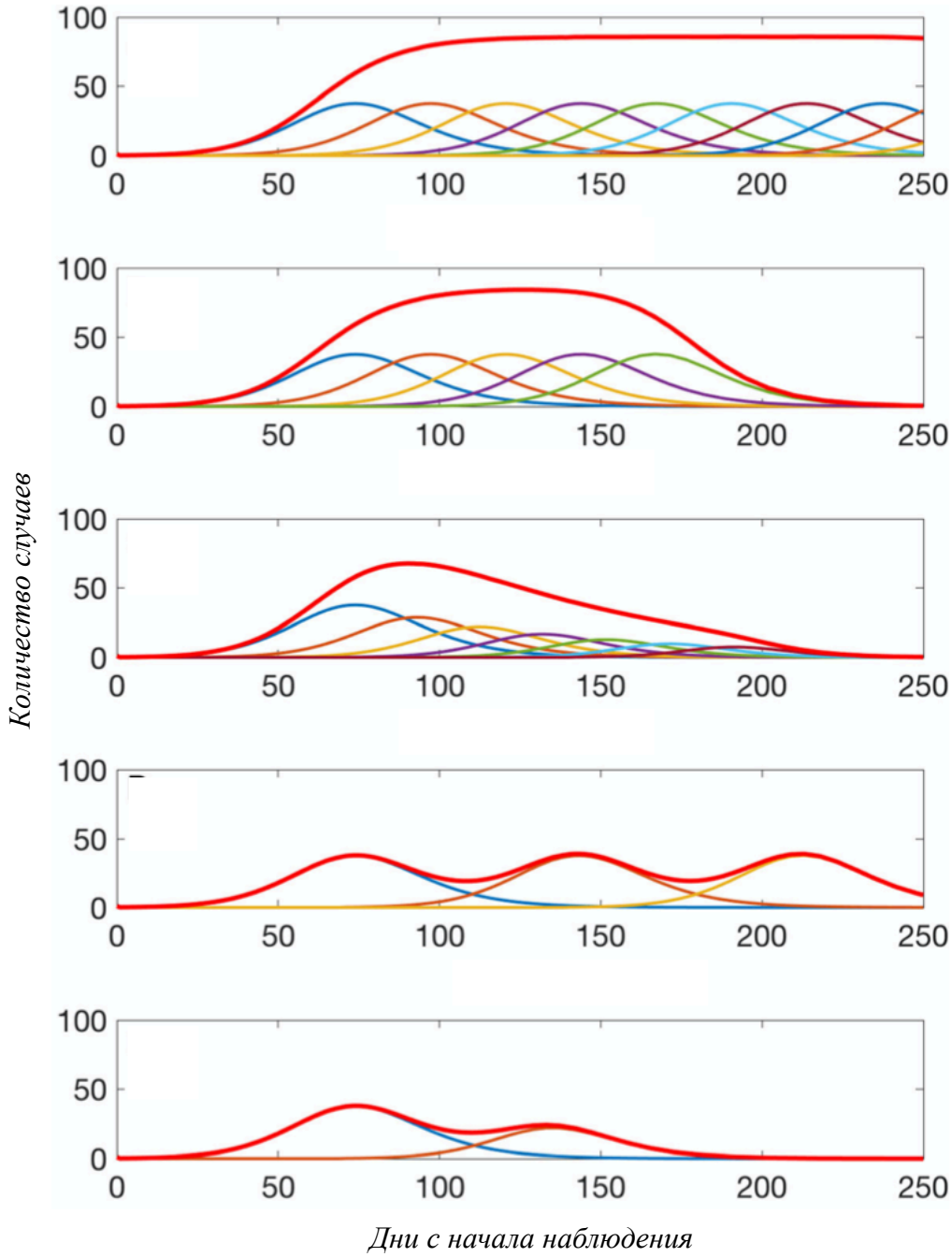

Рис. 10. Примеры разбиения кривых заболеваемости на субэпидемические компоненты

Следовательно,

$$A_i(t) = \begin{cases} 1, Q_i(t) > Q_{thr} \\ 0, \text{Иначе} \end{cases} \quad i = 1..n, \tag{3}$$

где $1 \le Q_{thr} \le Q_0$ и $A_1(t) = 1$. Более того, размер i-й субэпидемии Q_i экспоненциально уменьшается со скоростью q для последующих субэпидемий из-за множества факторов, включая эффекты сезонной передачи, постепенно

увеличивающееся влияние вмешательств здравоохранения или изменения поведения населения, которые уменьшают темпы распространения. Если $q = 0$, модель предсказывает волну эпидемии, включающую субэпидемии одинакового размера. Предполагая, что последующие размеры субэпидемии экспоненциально уменьшаются, мы имеем:

$$\mathrm{Q_i} = \mathrm{Q_0 e^{-q(i-1)}} \tag{4}$$

где Q_0 – размер начальной субэпидемии ($Q_1 = Q_0$). Следовательно, когда $q > 0$, общее количество субэпидемий, поддерживаемых моделью, зависит от Q_{thr}, q и, Q_0, так как $(\mathrm{i}+1)$-я субэпидемия запускается, только если $Q_{thr} \leq Q_i$.

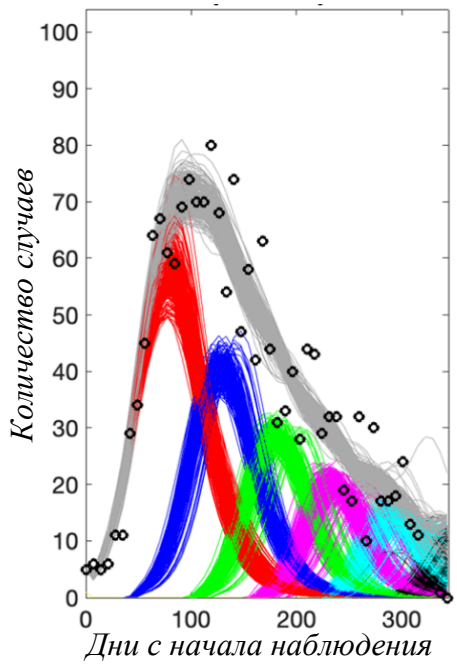

Рис. 11. Пример применения модели на данных эпидемии чумы в Мадагаскаре из оригинальной статьи[9]. Серым цветом показаны суммарные данные заболеваемости, черные точки – наблюдаемые еженедельные случаи, остальными цветами показаны предсказанные волны субэпидемии для различных значений параметров

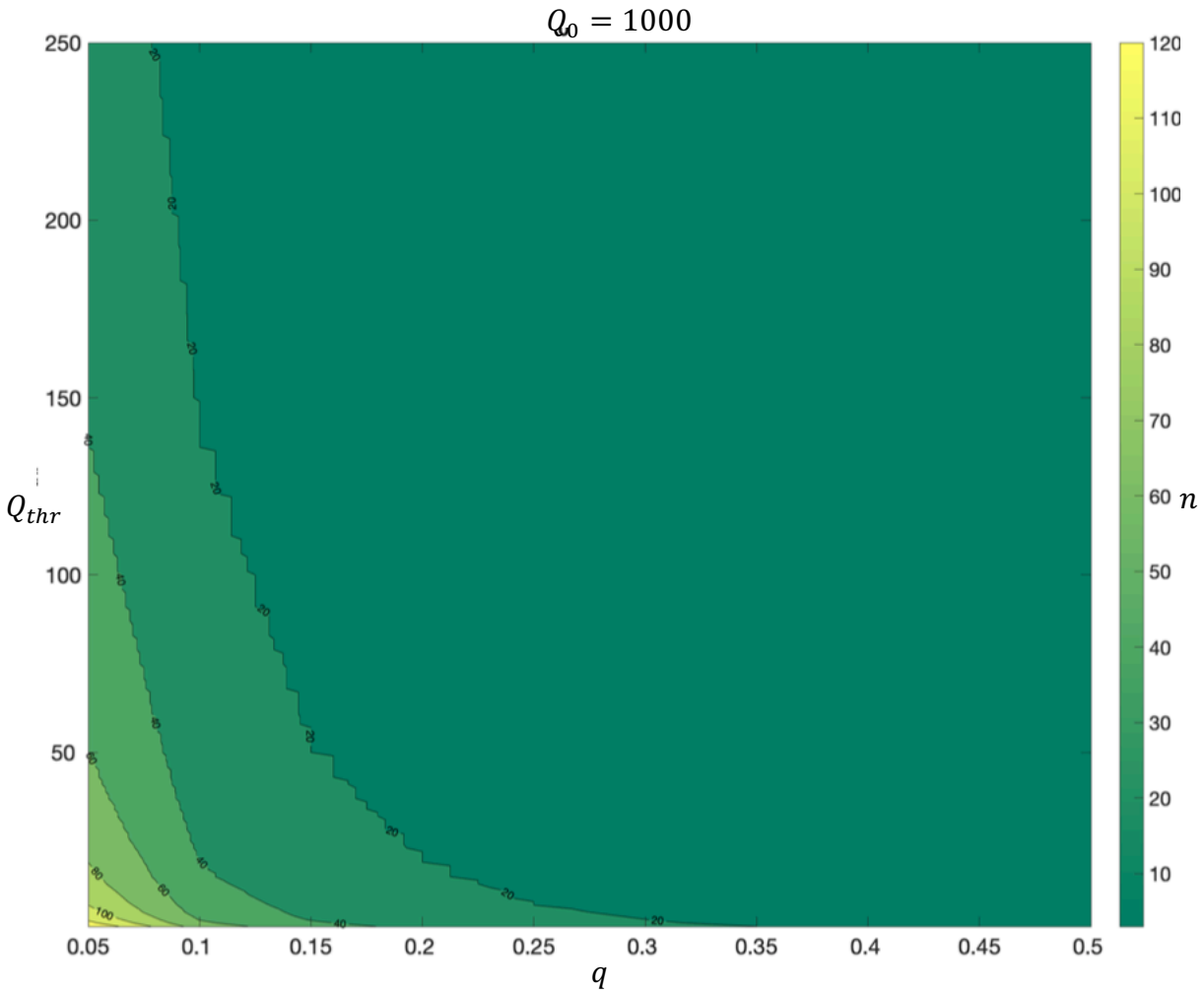

Рис. 12. Контурный график зависимости количества субэпидемий от параметров модели

Параметры модели $\Theta = (Q_{thr}, q, f, d, Q_0)$ предлагается оценивать методом наименьших квадратов по наблюдаемым данным заболеваемости.

3.3 Использование ОЛМ с субэпидемиями для оценки трансмиссивности

Для оценки трансмиссивности некоторого штамма вируса будем использовать модифицированную модель из предыдущего пункта. Основное отличие заключается в том, что описанная модель используется для прогнозирования течения эпидемии, задача же данной работы – оценка трансмиссивности одного штамма вируса относительно другого (в том числе среднего). С точки зрения алгоритма это позволяет нам почерпнуть из данных больше информации: в такой постановке мы сами можем определять количество субэпидемий исходя из построенного филогенетического дерева, предполагая, что некоторое его поддерево и образует некоторую субэпидемию. Такое предположение уточняет определенное в оригинальном алгоритме предположение о возможной природе субэпидемий.

Численным значением трансмиссивности будем считать базовое репродуктивное число[8, с.103–121] R_0 – это ключевая величина, используемая для оценки трансмиссивности инфекционных заболеваний. Теоретически R_0 определяется как среднее количество вторичных случаев, вызванных одним первичным случаем в течение всего периода его трансмиссивности в полностью восприимчивой популяции. Число воспроизводств R_0 напрямую связано с типом и интенсивностью вмешательств, необходимых для борьбы с эпидемией, поскольку цель усилий здравоохранения – как можно скорее достичь $R_0 < 1$. Одна из наиболее известных функций R_0 – определение критического охвата иммунизацией, необходимого для искоренения болезни в случайно смешанной популяции. R_0 обычно использовался для оценки серьезности эпидемии, потому что доля инфицированных в конце эпидемии (то есть окончательный размер) зависит только от R_0.

Хотя R_0 может быть полезным для понимания трансмиссивности болезни и разработки различных стратегий вмешательства, классическая пороговая величина теоретически предполагает, что эпидемия распространяется в полностью восприимчивой популяции. Помимо R_0, практическое значение имеет оценка зависящих от времени изменений потенциала передачи. Объяснение динамики эпидемии во времени может быть частично достигнуто путем оценки эффективного репродуктивного числа, $R(t)$, определяемого как фактическое среднее число вторичных случаев на первичный случай в календарное время t. $R(t)$ показывает изменение во времени из-за снижения числа восприимчивых людей (внутренние факторы) и применения сдерживающих мер (внешние факторы).

Если предположить, что мы наблюдали случаи Y_i с начала эпидемии, то репродуктивное число $R_c(t_i)$ рассчитывается как

$$R_c(t_i) = \frac{Y_i}{\sum_{j=0}^{n} Y_{i-j} s_j}, \tag{5}$$

где s_j – дискретизированный “последовательный интервал”, который определяется как время от начала первичного случая до начала вторичного случая.

Как видно из формулы (5), для оценки трансмиссивности некоторого штамма нам потребуется спрогнозировать количество случаев заражения им в каждый день эпидемии. Для этого воспользуемся моделью из п.3.2 и данными субэпидемий из имеющегося филогенетического дерева. Однако данная модель не предполагает использование таких данных, в связи с чем потребуется ее модификация.

3.4 Байесовский вывод

В статистическом анализе есть две широкие категории интерпретаций вероятности: байесовский и частотный выводы[9, с. 443–458]. Эти подходы часто расходятся друг с другом относительно фундаментальной природы вероятности. Частотный вывод определяет вероятность как предел относительной частоты события в большом количестве испытаний и только в контексте экспериментов, которые являются случайными и четко определенными. С другой стороны, байесовский вывод может назначать вероятности любому утверждению, даже если этому не сопоставлен некоторый случайный процесс. В байесовском выводе вероятность – это способ представить степень уверенности в утверждении или данных.

Основа для байесовского вывода вытекает из теоремы Байеса. Ее уравнение:

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}$$

Замена события B наблюдениями y и события A набором параметров Θ приводит к следующему:

$$p(\Theta|y) = \frac{p(y|\Theta)p(\Theta)}{p(y)},$$

где $p(y)$ будет обсуждаться ниже, $p(\Theta)$ – это набор априорных распределений набора параметров Θ до того, как будет наблюдаться y, $p(y|\Theta)$ – вероятность y в рамках модели, а $p(\Theta|y)$ является совместным апостериорным распределением набора параметров Θ, которое выражает неопределенность относительно набора параметров Θ после учета как априорного знания, так и данных. Поскольку обычно существует несколько параметров, Θ представляет собой набор из j параметров и может рассматриваться в дальнейшем в этой статье как

$$\Theta = (\theta_1, \dots, \theta_j)$$

Знаменатель

$$\mathrm{p}(y) = \int p(y|\Theta)p(\Theta)d\Theta$$

определяет маргинальное распределение y, или априорное предсказывающее распределение y, и может быть заменен на неизвестную константу c. Предыдущее прогнозирующее распределение указывает, как y должен выглядеть с учетом модели до того, как y будет наблюдаться. Только набор априорных вероятностей и функция правдоподобия модели

используются для предельного правдоподобия y. Наличие маргинального распределения y нормализует совместное апостериорное распределение $p(y|\Theta)$, гарантируя, что данное распределение удовлетворяет свойству ограниченности еденицей.

Заменяя $p(y)$ на c – константу пропорциональности, формулировка теоремы Байеса на основании модели становится следующей:

$$p(\Theta|y) = \frac{p(y|\Theta)p(\Theta)}{\mathrm{c}}$$

Или, перейдя от точного равенства к пропорциональности:

$$\mathrm{p}(\Theta|\,\mathrm{y}) \sim p(y|\Theta)p(\Theta)$$

Эта форма может быть определена как ненормированное совместное апостериорное распределение с учетом априорного знания. Однако цель байесовского вывода, основанного на модели, обычно состоит не в том, чтобы суммировать ненормированное совместное апостериорное распределение, а в том, чтобы суммировать маргинальное распределение параметров. Полный набор параметров Θ обычно можно разделить на

$$\Theta = \{\Phi, \Lambda\}$$

где Φ – представляющий интерес подвектор, а Λ – дополнительный подвектор Θ – вектор мешающих параметров. В байесовской модели наличие мешающих параметров не создает формальных теоретических проблем. Мешающий параметр – это параметр, который существует в совместном апостериорном распределении модели, но не интересующий с точки зрения модели. Функцию предельного правдоподобия апостериорного распределение интересующего параметра Φ можно просто записать как

$$\mathrm{p}(\Phi|y) = \int p(\Phi, \Lambda|y)d\Lambda$$

В модели, основанной на байесовском выводе, теорема Байеса используется для оценки ненормализованного совместного апостериорного распределения и дает возможность оценить и сделать выводы из маргинальных апостериорных распределений.

Априорные распределения можно оценить в рамках модели с помощью гиперприорных распределений. Параметры гиперприорных распределений называются гиперпараметрами. Использование гиперприорных

распределений для оценки априорных называется иерархическим байесовским выводом. Теоретически этот процесс может продолжаться и дальше, используя гиперприорные распределения для оценки гиперприорных распределений. Оценка априорных значений через гиперприоры и на основе данных – это метод выявления оптимальных априорных распределений. Одно из многих естественных применений иерархического Байеса – многоуровневое моделирование.

Так как ненормализованное совместное апостериорное распределение пропорционально вероятности, умноженной на априорное распределение

$$p(\Theta|\, y) \sim p(y|\Theta)p(\Theta)$$

Простейшая иерархическая байесовская модель имеет вид

$$p(\Theta, \Phi|\, y) \sim p(y|\Theta)p(\Theta|\Phi)p(\Phi),$$

где Ф – множество гиперприорных распределений. При чтении уравнения справа налево оно начинается с гиперприоров Ф, которые используются условно для оценки априорных значений $p(\Theta|\Phi)$, которые, в свою очередь, используются, как обычно, для оценки правдоподобия $p(y|\Theta)$, и, наконец, апостериор – $p(\Theta, \Phi|\, y)$.

Чтобы завершить определение байесовской модели, как априорные распределения, так и вероятность должны быть аппроксимированы или полностью определены. Вероятность y в рамках модели, функция правдоподобия или $p(y|\Theta)$ содержат информацию, предоставленную выборкой. Вероятность равна

$$\mathrm{p}(y|\Theta) = \prod_{i=1}^{n} \mathrm{p}(y_i|\Theta)$$

Данные y влияют на апостериорное распределение $p(\Theta|\, y)$ только через функцию правдоподобия $p(y|\Theta)$. Таким образом, байесовский вывод подчиняется принципу правдоподобия, который гласит, что для данной выборки данных любые две вероятностные модели $p(y|\Theta)$, которые имеют одинаковую функцию правдоподобия, дают одинаковый вывод для Θ.

3.5 ОЛМ с байесовским выводом

Опишем метод оценки темпов роста и основных показателей воспроизводства для субпопуляций геномов вирусов, распространяющихся в популяции восприимчивых людей. Формально нам дана вирусная популяция $P = P_1 \cup \ldots \cup P_n$, состоящая из n субпопуляций с различными фенотипическими особенностями, и цель состоит в том, чтобы оценить наиболее вероятную функцию $f: P_i \to f_i$. Для оценки точности будет использована модель для $n = 2$, но описанный подход применим и для большего количества субпопуляций. Предположим, что вирусная популяция была отобрана за дискретный интервал времени $\tau = (\tau_1 \ldots \tau_S)$. Для каждой дискретной временной точки τ_j рассматриваем наблюдаемое количество субпопуляций $k^j = \left(k_1(\tau_j), \ldots, k_n(\tau_j)\right)$, общее количество секвенированных последовательностей $l(\tau_j) = \sum_{i=1}^{n} k_i(\tau_j)$ и наблюдаемую частоту встречаемости $c(\tau_j)$. Каждая субпопуляция P_j также имеет время первого обнаружения $t_i \leq min\{\tau : k_i(\tau) > 0\}$.

Для описания волн будем использовать систему однородных дифференциальных уравнений, аналогичную (2) [10]. Индикаторную же функцию заменим с использованием информации о времени начала субэпидемии t_i на следующую:

$$\mathrm{A}_i(\mathrm{t}) = \begin{cases} 1, t \geq \mathrm{t_i} \\ 0, \text{Иначе} \end{cases} \quad i = 1..n, \tag{6}$$

Помимо темпов роста f_i, другими параметрами модели являются параметр «масштабирования роста» d и максимальный размер субэпидемии Q_i (4). Кроме того, далее рассматривается вектор ежедневных случаев заражения $\overline{Y}(\tau_j) = \left(Y_1(\tau_j), \ldots, Y_n(\tau_j)\right)$, где $Y_1(\tau_j) = X_i(\tau_j) - X_i(\tau_{j-1}), i = 1, \ldots, n$.

Мы оцениваем параметры модели f, Q_0, d, q, максимизируя апостериорную вероятность $p(f, d, Q_0, q, t|C, k)$ параметров модели, заданных наблюдаемыми данными зарегистрированных случаев заболевания $C = \left(c(\tau_1), \ldots, c(\tau_s)\right)$ и выборочные данные о субпопуляции $k = (k_1, \ldots, k_s)$ – секвенированные последовательности из построенного дерева, байесовским способом:

$$p(f, d, Q_0, q, t|C, k) \sim p(C|f, d, Q_0, q, t)p(k|f, d, Q_0, q, t) \times \\ \times \left(\prod_{i=1}^{n} p(f_i)p(t_i)\right) p(d)p(Q_0)p(q) \tag{7}$$

Вероятности $p(C|f,d,Q_0,q,t)$ и $p(k|f,d,Q_0,q,t)$ определяются исходя из предположения, что для каждого момента времени τ_i:

1. наблюдаемая частота $c(\tau_i)$ берется из распределения Пуассона со средним значением, равным расчетной общей заболеваемости на основе модели $Y(\tau_j) = \sum_{i=1}^{n} Y_i(\tau_j)$
2. наблюдаемые данные о субпопуляции $k^j = \left(k_1(\tau_j), \dots, k_n(\tau_j)\right)$ соответствует полиномиальному распределению с вероятностями $p^j = \left(p_1^j, \dots, p_n^j\right)$, где $p_i^j = \frac{Y_i(\tau_j)}{Y(\tau_j)}$.

Априорные значения $p(f_i)$, $\mathrm{p}(d)$, $\mathrm{p}(Q_0)$, $\mathrm{p}(q)$, $p(t_i)$ были определены исходя из предположении, что соответствующие параметры распределены равномерно на интервалах $[0, f_{max}]$, $[0, p_{max}]$, $[0, Q_{max}]$, $[0, q_{max}]$, $[\tau_1, min\{\tau: k_i(\tau) > 0\}]$ соответственно. Таким образом, после логарифмирования и отбрасывания постоянных параметры оценивались путем решения следующей оптимизационной задачи:

$$(\mathrm{f}^*, \mathrm{d}^*, \mathrm{Q_0}^*, \mathrm{q}^*, \mathrm{t}^*) = \arg \max_{f,d,Q_0,q,t} (C - l) \cdot ln(\mathrm{Y}) - 1 \cdot Y + \sum_{j=1}^{S} k^j \cdot ln\left(\overline{Y}\right) \quad (8)$$

с ограничениями $0 \le f_i \le f^{max}, 0 \le d \le d^{max}, 0 < Q_0 \le Q_0^{max}, 0 \le q \le q^{max}, \tau_1 \le t_i \le min\{\tau: k_i(\tau) > 0\}$. Здесь $\overline{Y}$ и Y – функции от (f,d,Q_0,q,t), символ $\cdot$ обозначает скалярное умножение векторов, а ln – покоординатный натуральный логарифм.

Для решения (8) используем методы численной безградиентной оптимизации, так как целевая функция дискретна и, соответственно, не дифференцируема. Количество оптимизируемых параметров данной задачи равно $2n + 3$. В качестве тестовой реализации был выбран метод сопряженных направлений Пауэлла, реализованный в пакете SciPy[11]. Однако целевая функция так же не является заданной явно – Y зависят от X, вид которого не известен, а задан через однородное дифференциальное уравнение. Для этого уравнения из его эпидемиологического смысла можно сформулировать условия Коши, а именно $\mathrm{X}_i(0) = 0$. Будем решать такие уравнения для каждого i численно. В тестовой реализации для этого был использован явный метод Рунге-Кутты 4 порядка, также реализованный в пакете SciPy.

Полученные оптимальные параметры модели использовались для оценки базового репродуктивного числа, связанного с каждой субпопуляцией. Предсказанные данные SARS-CoV-2 были смоделированы с помощью параметрического бутстрэпинга с учетом гамма-распределения со средним значением $\mu = 5{,}2$ дня и стандартным отклонением $\sigma = 1{,}72$ дня[12]. Затем,

если $\rho(t)$ - это распределение вероятностей интервала сгенерированных t, а $Y_i^*(\tau_j)$ - основанная на модели частота i-й субэпидемии, рассчитанная с использованием оптимальных параметров, то i-е базовое репродуктивное число было рассчитано с использованием уравнения восстановления (п.3.3) следующим образом:

$$R_{\tau_j}^i = \frac{Y_i^*(\tau_j)}{\sum_{l=1}^{j} Y_i^*(\tau_j - \tau_l)\rho(\tau_l)} \quad (9)$$

Здесь числитель представляет общее количество новых случаев в данный момент времени τ_j, а знаменатель – общее количество случаев, которые вносят вклад (как первичные случаи) в создание новых случаев в момент времени τ_j.

3.6 Результаты

Максимальные апостериорные базовые числа репродукции $R_0(1)$ и $R_0(2)$ были оценены для штамма B.1.1.7 и всех остальных с использованием методов, описанных выше. Чтобы генерировать распределение параметров использовался параметрический бутстрэп с 500 наборами и пуассоновским шумом, добавленным к наблюдаемому общему числу случаев и наблюдаемым случаям субпопуляций. Верхние границы для байесовского вывода были выбраны следующими: $f^{max} = 2, p^{max} = q^{max} = 1, Q_0^{max} = 10^8$.

Расчетное среднее максимальное апостериорное отношение репродуктивных базовых чисел субпопуляций B.1.1.7 и не–B.1.1.7 составило $R_0(2)/R_0(1) = 1{,}641$ (95% доверительный интервал [1,615, 1,754]), рис. 11. Таким образом, оценочная трансмиссивность SARS-CoV-2 штамма B.1.1.7 примерно на 64% выше, чем для штаммов, отличных от B.1.1.7 ($p < 0{,}001$, критерий Краскела–Уоллиса). Эта оценка согласуется с оценками относительной трансмиссивности новых британских вариантов SARS-CoV-2, представленными в других ранних исследованиях [13, 14].

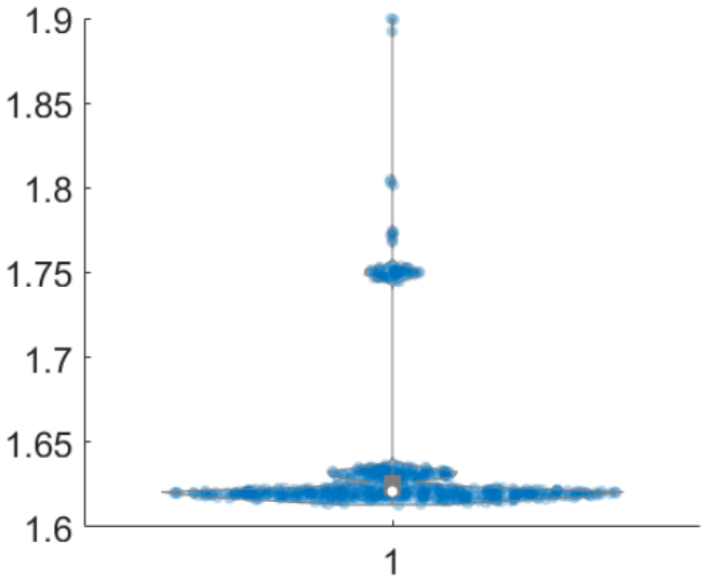

Рис. 13. График отношения базовых репродуктивных чисел для штаммов B.1.1.7 и не–B.1.1.7

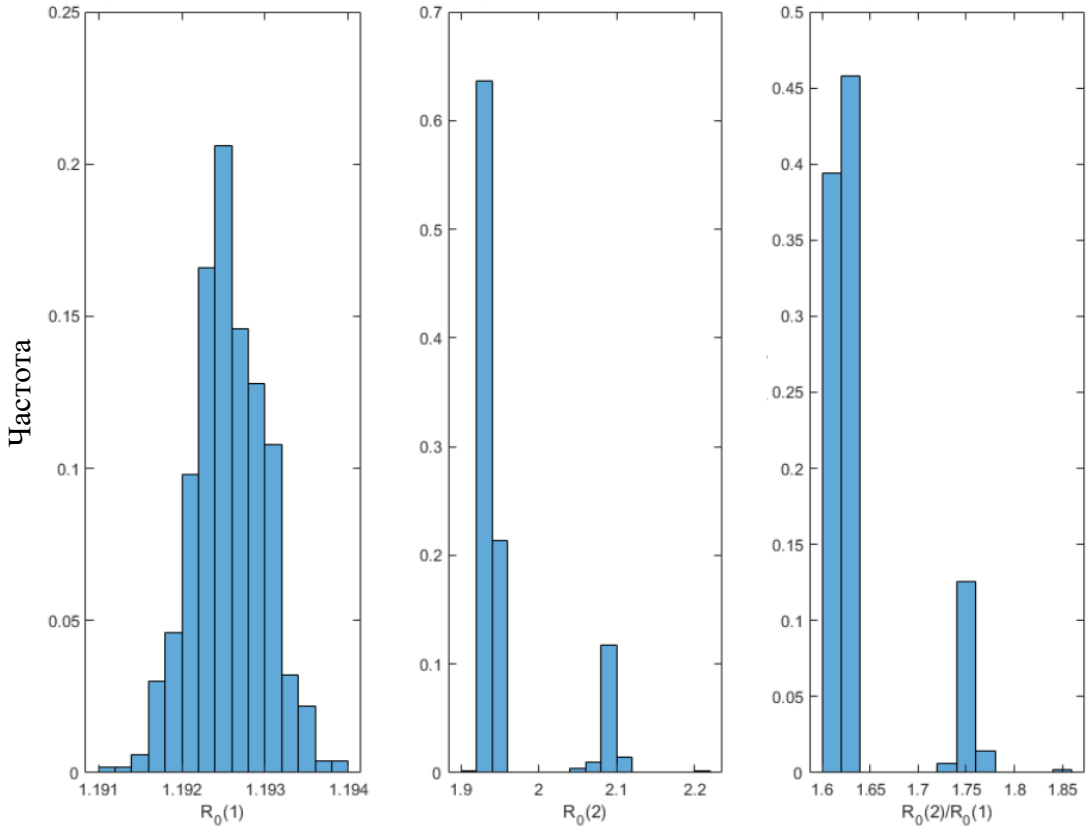

Рис. 14. Графики распределения оцененных значений базовых репродуктивных чисел и их отношения в результате применения бутстрэпинга

ЗАКЛЮЧЕНИЕ

В данной работе рассмотрены основы обработки данных секвенирования, построения на их основе филогенетических деревьев, выделения из них штаммов, предварительной подготовки данных и информационно-статистического анализа темпов их распространения.

В первой главе изложены основные теоретические данные, на основе которых происходит анализ клеточного генома, описана эволюционная теория развития вирусов и способ представления истории эволюции в виде графов. Далее, приведены методы секвенирования и даны априорные оценки их ошибок, описаны сценарии использования. В конце главы сформулирована задача построения филогенетических деревьев и особенностей получаемых результатов.

Вторая глава посвящена подготовке данных для их последующего анализа. Рассмотрены вопросы декомпозиции задачи на подходящие подзадачи, выбора штаммов и нормализации данных. Проверено и проанализировано два метода нормализации данных и предложены их потенциальные улучшения. Описан и формализован непосредственно способ индикаторного анализа эволюции штаммов. Реализован и проверен простейший подходящий способ численно описать темпы роста распространенности мутации, описаны его преимущества и недостатки.

В третьей главе описана обобщенная логистическая модель с перекрывающимися эпидемиями, ее теоретическое обоснование и нижележащая ОЛМ. Описан стандартный способ оценки трансмиссивности вирусов, который предложено использовать для оценки трансмиссивности штаммов. Выведена вероятностная модель, позволяющая оптимизировать параметры исходя из наблюдаемых данных при помощи байесовского вывода. Описана задача оптимизации и предложены методы, позволяющие ее численно решить и получить оптимальные параметры модели с субэпидемиями на основании построенного филогенетического дерева, его штаммов и общих численных данных о течении эпидемии. Полученные оптимальные параметры использованы для оценки количества случаев заражения в каждый момент времени каждым рассматриваемым штаммом вируса, что в последующем использовано для вычисления трансмиссивности штаммов.

Данная модель не является идеальной, как и другие существующие модели, однако она на каждом шаге своей работы максимизирует вероятность быть корректной на основании имеющихся данных. Так же модель не проверена для большего количества рассматриваемых субэпидемий, чем две, не оценена точность и производительность использованных в ходе ее

оптимизации численных методов, которые могут накапливать значительные ошибки из-за большого количества итераций.

Построенная математическая модель проверена на данных штамма B.1.1.7 вируса SARS-CoV-2 и получены результаты относительной трансмиссивности, схожие с полученными в других исследованиях данного штамма, что позволяет говорить о состоятельности данного подхода как такового. Данная модель использует хорошо показавший себя для прогнозирования фреймворк ОЛМ с субэпидемиями и традиционно используемый в биоинформатике байесовский вывод, что позволяет расценивать данную модель как одну из самых точных для оценки трансмиссивности штаммов.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Dissecting the clonal origins of childhood acute lymphoblastic leukemia by single-cell genomics / [Electronic resource]. — 2014. Mode of access: https://www.pnas.org/content/111/50/17947 — Date of access: 11.10.2019.
2. DNA sequencing / [Electronic resource]. Mode of access: https://en.wikipedia.org/wiki/DNA_sequencing — Date of access: 11.10.2019.
3. Advances in understanding tumour evolution through single-cell sequencing / [Electronic resource]. — 2017. Mode of access: https://www.sciencedirect.com/science/article/pii/S0304419X17300392 — Date of access: 03.10.2019.
4. Preliminary genomic characterisation of an emergent sars-cov-2 lineage in the UK defined by a novel set of spike mutations. / Andrew Rambaut [etc.] – Virological, 2020 – 6 p.
5. Mathematical and Statistical Estimation Approaches in Epidemiology / Hiroshi Nishiura [etc.] – Springer, 2009 – 360 p.
6. Bruce Pell. Using phenomenological models for forecasting the 2015 Ebola challenge. / Bruce Pell [etc.] – Epidemics, 2018– 70p.
7. Gerardo Chowell. Fitting dynamic models to epidemic outbreaks with quantified uncertainty / Gerardo Chowell – Infect Dis Model, 2017 – 25p.
8. Gerardo Chowell. A novel sub-epidemic modeling framework for short-term forecasting epidemic waves. / Gerardo Chowell, Amna Tariq, and James M Hyman – BMC medicine, 2019 – 164 p.
9. Tomohiro Ando. Bayesian Predictive Information Criterion for the Evaluation of Hierarchical Bayesian and Empirical Bayes Models / Tomohiro Ando [etc.] – Biometrika, 2007 – 584 p.
10. Pavel Skums. Bayesian assessment of the transmissibility of an emergent SARS-CoV-2 lineage in the UK / Pavel Skums – 2021 – 6p.
11. Powell's Method / [Electronic resource] – Mode of access https://docs.scipy.org/doc/scipy/reference/optimize.minimize-powell.html – Date of access: 11.01.2021.
12. Estimating the generation interval for coronavirus disease (covid- 19) based on symptom onset data, march 2020 / Tapiwa Ganyani [etc.] – Eurosurveillance, 2020 – 29p.
13. Early empirical assessment of the n501y mutant strains of sars-cov-2 in the united kingdom, october to november 2020 / Kathy Leung [etc.] – medRxiv, 2020.
14. Estimated transmissibility and severity of novel sars-cov-2 variant of concern 202012/01 in England / Nicholas G Davies [etc.] – medRxiv, 2020.

ПРИЛОЖЕНИЕ А

```
import io
from newick import load
from collections import defaultdict
from datetime import datetime
import matplotlib.pyplot as plt
import matplotlib.dates as mdates
import numpy as np

def load_tree(path):
    with io.open(path, encoding='utf8') as fp:
        return load(fp)[0]

def get_date_from_meta(meta):
    return meta.split("|")[2]

def get_id_from_meta(meta):
    return meta.split("/")[2]

def get_timestamp(date: str):
    return mdates.epoch2num(datetime.strptime(date, '%Y-%m-%d').timestamp(
))

def generate_bins_from_dates(start_date, end_date, timerange):
    start = pd.Timestamp(start_date)
    end = pd.Timestamp(end_date)
    parts = list(pd.date_range(start, end, freq=timerange))
    parts_with_borders = []
    if len(parts) != 0:
        parts_with_borders =
            [parts[0] - pd.to_timedelta(timerange)] + parts + \
            [parts[-1] + pd.to_timedelta(timerange)]
    else:
        number_of_days = int(start.weekday()) + 1
        parts = [start - pd.to_timedelta(f'{number_of_days}d')]
        parts_with_borders = parts + [parts[-1] + pd.to_timedelta(timerang
e)]

    return list(map(lambda x: x.to_pydatetime().timestamp(), parts_with_bo
rders))

def show_hist(n, bins, polynom=None, approx_polynom=None, width=1):
    fig, ax = plt.subplots(1,1)
    ax.bar(bins[:-1], n, width=width)
    ax.xaxis.set_major_formatter(mdates.DateFormatter('%d.%m'))
    if polynom is not None:
        ax.plot(bins, polynom(bins), "r--")
    plt.xlabel('date')
    plt.ylabel('%')
    if approx_polynom is not None:
        y = [np.polyval(approx_polynom, i - bins[0]) for i in bins]
        ax.plot(bins, y)
    plt.show()

def prepare_hist(leaves):
    dates = []
    for leaf in leaves:
        try:
            dates.append(datetime.strptime(
```

```
                get_date_from_meta(leaf), '%Y-%m-%d'))
        except:
            try:
                dates.append(datetime.strptime(
                    get_date_from_meta(leaf), '%Y-%m-00'))
            except:
                pass
    bins = list(map(mdates.epoch2num, generate_bins_from_dates(min(dates),
max(dates), '1w')))
    return np.histogram(list(map(lambda x: mdates.epoch2num(x.timestamp())
, dates)), bins=bins)

def draw_date_histogramm_by_subtree(
       leaves, totals, force_draw=False, min_k=0.02):
    ns, bins = prepare_hist(leaves)
    n1 = []
    for n, b in zip (ns, bins):
        n1.append(n / totals[b] * 100 if n != 0 else 0)
    ns = n1
    z = np.polyfit(bins[:-1], ns, 1)
    if z[0] > min_k or force_draw:
        print(f'Subtree size: {len(leaves)}')
        print(f'Trendline coefficient: {z[0]}')
        p = np.poly1d(z)
        show_hist(ns, bins, p)

def prepare_approx(n, bins):
    approx = np.polyfit(bins[:-1] - bins[0], n, 10)
    return np.poly1d(approx), n.mean()

def prepare_totals(tree):
    leaves = tree.get_leaf_names()
    ns, bins = prepare_hist(leaves)
    return {b: n for n, b in zip(ns, bins)}

def _dig(node, stop_fn):
    if not stop_fn(node):
        res = []
        for descendant in node.descendants:
            if not descendant.is_leaf:
                res += _dig(descendant, stop_fn)
        return res
    else:
        return [node]

def analyze_subtrees(tree, size_range=(30, 100)):
    n, bins = prepare_hist(tree.get_leaf_names())
    totals = prepare_totals(tree)
    for i in
       _dig(tree, lambda node: len(node.get_leaves()) < size_range[1]):
        leaves = i.get_leaf_names()
        if len(leaves) < size_range[0]:
            continue
        draw_date_histogramm_by_subtree(leaves, totals, min_k=0.1)

top_countries = ['Belgium', 'Netherlands', 'Denmark', 'France', 'Japan', '
Portugal', 'Spain', 'Switzerland'] # 'South Africa'

for country in top_countries:
    print(f'\n\n{country}\n')
    tree = load_tree(os.path.join('COVID data', 'countries', country, 'tre
e.nwk'))
```

```
    analyze_subtrees(tree)

import numpy as np
from scipy.integrate import solve_ivp
from scipy.optimize import *
from scipy.stats import gamma

# x = f, d, Q0, q, t

class model:
    def __init__(self, data, samples):
        self.n = len(samples[0])
        self.c = data
        self.s = len(samples)
        self.k = samples # [t[subpopulation]]
        self.x_max = self.s
        self.x = (*[0.1]*n,0.1,0.1,0.1,*list(map(get_first_non_zero_idx, s
amples.T)))
        self.l = np.array(list(sum(self.k[t]) for t in range(self.x_max)))

    def dX(self, i, t, x):
        x = x[0]
        f = self.x[:self.n]
        d, Q0, q = self.x[self.n:-self.n]
        T = self.x[-self.n:]
        return f[i] * self.a(i, t, T) * x**d * (1 - x / self.Q(i, Q0, q, T
))

    def X(self, i, t):
        sol = solve_ivp(
            lambda t_, _x: self.dX(i, t_, _x),
            [0, self.x_max],
            [1],
            t_eval=[t])
        return sol.y[0][0]

    def Y(self, i, t):
        return self.X(i, t) - self.X(i, t - 1)

    def Y_sum(self, t):
        return sum(self._Y(t))

    def Y_deltas(self):
        return np.array(list(self.Y_sum(t) for t in range(1, self.x_max)))

    def Y_(self, t):
        return np.array(list(self._Y(t)))

    def f(self, x):
        self.x = x
        res = (self.c - self.l)[1:].dot(np.log(self.Y_deltas())) + \
              sum(self.k[j].dot(np.log(self.Y_(j))) for j in range(1, sel
f.s)) - \
              np.ones(self.s - 1).dot(self.Y_deltas())
        return res

    # Helpers
    def _Y(self, t):
        return (self.Y(i, t) for i in range(self.n))

    def a(self, i, t, T):
        return 0 if t < T[i] else 1
```

```
    def Q(self, i, Q0, q, t):
        return Q0 * np.exp(q * t[i])

class P:
    def __init__(self, mu, sigma):
        self.alpha = 1. / sigma
        self.beta = mu / sigma

    def __call__(self, t):
        return gamma.pdf(t, self.beta, self.alpha)

def get_first_non_zero_idx(arr):
    return next((i for i, x in enumerate(arr) if x), None)

def R(i, j, x):
    m.x = x
    p = P(5.2, 1.72)
    s = sum(m.Y(i, j - l) * p(l) for l in range(1, j))
    return m.Y(i, j) / s

def relative_transmissibility(n, data, samples):
    t_j = len(samples) - 1
    res = minimize(
        lambda x: m.f(x),
        (*[1]*n, .5, .5, 10**5, *list(map(lambda x: get_first_non_zero_idx
(x) // 2, samples.T))),
        bounds = zip(
            (*[0]*n, 0, 0, 0,     *[0]*n),
            (*[2]*n, 1, 1, 10**8, *list(map(get_first_non_zero_idx, sample
s.T)))
        ),
        method = 'Powell'
    )
    return R(0, t_j, res.x) / R(1, t_j, res.x)

tree = load_tree(
    os.path.join('COVID data', 'countries', 'United Kingdom', 'tree.nwk'))

def analyze_subtrees(tree, size_range=(30, 100), N=10):
    rel_trans = []
    for i in \
        _dig(tree, lambda node: len(node.get_leaves()) < size_range[1]):
        leaves = i.get_leaf_names()
        if len(leaves) < size_range[0]:
            continue
        rel_trans.append(relative_transmissibility)

    top_n = sorted(rel_trans, key = lambda x: x, reverse = True)[:N]
    for z, ns, bins, p in top_n:
        print(f'Trendline coefficient: {z[0]}')
        show_hist(ns, bins, p)


for country in top_countries:
    print(f'\n\n{country}\n')
    tree = load_tree(
        os.path.join('COVID data', 'countries', country, 'tree.nwk'))
    analyze_subtrees(tree)
```