Институт нефтехимии и катализа - обособленное структурное подразделение
Федерального государственного бюджетного научного учреждения Уфимского
федерального исследовательского центра
Российской академии наук
(ИНК УФИЦ РАН)

На правах рукописи

Ахметзянова Лиана Ульфатовна

КОМПЬЮТЕРНЫЙ ДИЗАЙН ПРАЙМЕРОВ ДЛЯ ПЕТЛЕВОЙ ИЗОТЕРМИЧЕСКОЙ АМПЛИФИКАЦИИ

1.2.2 – Математическое моделирование, численные методы и комплексы программ

Автореферат на соискание учёной степени
кандидата технических наук

Новосибирск-2024

Работа выполнена в лаборатории математической химии Института нефтехимии и катализа – обособленном структурном подразделении Федерального государственного бюджетного научного учреждения Уфимского федерального исследовательского центра Российской академии наук

Научный руководитель: **Губайдуллин Ирек Марсович**
доктор физико-математических наук, профессор, заведующий лабораторией математической химии ИНК УФИЦ РАН

Научный консультант: **Гарафутдинов Равиль Ринатович**
кандидат биологических наук,
заведующий лабораторией структуры и функций биополимеров ИБГ УФИЦ РАН

Официальные оппоненты:
(предварительно планируемые)

Никитина Алла Валерьевна
– доктор технических наук, профессор кафедры Интеллектуальных и многопроцессорных систем Института компьютерных технологий и информационной безопасности Федерального государственного автономного образовательного учреждения высшего образования «Южный федеральный университет».

Ломзов Александр Анатольевич,
кандидат физико-математических наук, заведующий лабораторией структурной биологии Института химической биологии и фундаментальной медицины СО РАН

Ведущая организация:
(предварительно планируемая)

Федеральное государственное бюджетное учреждение науки Институт теоретической и экспериментальной биофизики Российской академии наук

Защита диссертации состоится 00 месяц 2024 года в 00 час. 00 мин. на заседании диссертационного совета...
Автореферат разослан « » _____2024 г

Ученый секретарь диссертационного совета, уч.степень ФИО

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы. В настоящее время все более широкое применение находит анализ биологических образцов путем обнаружения в них специфических последовательностей нуклеиновых кислот (НК) с помощью амплификации – ферментативной реакции, приводящей к наработке фрагментов ДНК в количестве, которое может быть детектировано инструментальными методами. Разработано множество методов амплификации НК, различающихся вариантами проведения реакции и способами детекции ее результатов. Самым популярным методом остается полимеразная цепная реакция (ПЦР), которая включает несколько этапов, протекающих при разных температурах. Из-за необходимости термоциклирования продолжительность ПЦР составляет около 1,5 ч, в связи с чем вызывают интерес более быстрые реакции амплификации, протекающие при постоянной температуре. Среди таковых наиболее популярной стала петлевая изотермическая амплификация (LAMP); LAMP-анализ занимает, как правило, до 40 мин.

Незаменимым компонентом любой амплификационной системы являются праймеры – короткие синтетические олигонуклеотиды длиной 16-30 нуклеотидов, для подбора которых используются специализированные компьютерные программы, базирующиеся на различных алгоритмах. Некоторые программы дизайна праймеров представлены в виде бесплатных web-сервисов, их дистрибутивы могут быть скачаны с соответствующих сайтов или получены от разработчиков. Для части программ их код находится в открытом доступе, ряд программ доступен на платной основе.

Для проведения классической ПЦР необходимы два праймера (прямой и обратный), для подбора которых написаны сотни различных компьютерных программ. В отличие от ПЦР, для проведения LAMP нужно как минимум четыре праймера: два внешних и два внутренних, последовательности которых гомологичны шести участкам нуклеотидной последовательности мишени. Дизайн LAMP-праймеров представляет собой более трудную задачу, при этом количество соответствующих программ не превышает десяти. Из них только две доступны онлайн, одна работает на Linux. Данные программы имеют определенные ограничения, например, позволяют осуществлять подбор праймеров только на основе нуклеотидных последовательностей, не учитывают образование димеров. Это приводит к подбору праймеров недостаточно высокого качества, следствием чего может стать получение недостоверных результатов амплификационного анализа.

Таким образом, создание инструмента для дизайна «качественных» праймеров для петлевой изотермической амплификации является актуальной задачей.

Целью настоящей работы является разработка комплекса программ для компьютерного моделирования праймерных систем, позволяющих подбирать специфичные праймеры для проведения петлевой изотермической амплификации.

Задачи исследования. Для достижения данной цели, были поставлены следующие задачи:

1. Смоделировать новую систему подбора праймеров для LAMP на основе расчета температуры отжига, включающую учет длины праймеров, возможность расположения их на максимально близком расстоянии друг от друга и исключения потенциальных димерных структур.
2. Предложить модификацию алгоритма прямого перебора с использованием трафаретного подхода, учитывающего GC-состав и температуру отжига праймеров.
3. Разработать комплекс программ с дружелюбным интерфейсом и многокритериальными условиями дизайна праймеров для LAMP. Провести верификацию нового программного продукта. Сравнить с существующими аналогами (NEB LAMP, PrimerExplorer).
4. Провести тестирование функционала программ на нуклеотидных последовательностях различной структуры и экспериментальную валидацию праймеров на примере обнаружения генетического материала коронавируса SARS-CoV-2.

Научная новизна заключается: 1) в моделировании новой системы дизайна праймеров для петлевой изотермической амплификации, подразумевающей задание жестких критериев отбора и возможность сближения последовательностей праймеров в пределах одного комплекта; 2) в модификации метода прямого перебора с использованием трафаретного подхода для поиска праймеров; 3) в разработке комплекса программ с расширенным функционалом, который позволяет подбирать специфичные праймеры для получения более точных диагностических результатов.

Практическая значимость работы определяется созданием программного комплекса для моделирования (дизайна) наборов специфичных LAMP–праймеров на основе новой системы их подбора. Разработанный программный комплекс может быть использован в молекулярной биологии, для создания диагностических тест-систем, обеспечивающих высокую чувствительность и достоверность обнаружения специфических ДНК и РНК. Программный комплекс может применяться в научно-исследовательских институтах и лабораториях, занимающихся амплификацией нуклеиновых кислот.

Методология и методы исследования. В основе разработанного алгоритма лежит линейный поиск образца в строке с учетом критериев подбора праймеров для LAMP. Программный комплекс дизайна LAMP–праймеров реализован на языке программирования Python, с использованием библиотек bioPython для работы с нуклеотидными последовательностями и фреймворка Qt для разработки интерфейса.

Положения, выносимые на защиту.

- Смоделирована новая система подбора праймеров для LAMP, включающая формулу расчета температуры отжига праймеров, возможность расположения их на максимально близком расстоянии и исключения потенциальных димерных структур. Полученная модель позволяет предотвратить получение ложных результатов (обеспечивает высокую специфичность амплификации).

- Предложена модификация метода прямого перебора с использованием трафаретного подхода, учитывающего GC-состав и температуру отжига и позволяющего снизить сложность перебора.

- Разработан программный комплекс LAMPrimers iQ для дизайна специфичных праймеров для проведения LAMP-амплификации, позволяющий подбирать наборы праймеров на основе нуклеотидных последовательностей различной структуры и длины. В амплификационных экспериментах было показано, что праймеры, подобранные с помощью программы LAMPrimers iQ, обеспечивают более высокую специфичность анализа.

Степень достоверности результатов. Высокая степень достоверности представленных результатов подтверждается совпадением расчетных и экспериментальных данных, воспроизведением работ других авторов, а также обсуждением результатов работы на международных и всероссийских конференциях и публикацией в серьезных научных журналах.

Апробация работы. Основные результаты диссертационной работы докладывались и обсуждались:

- на научных семинарах следующих институтов: Институт нефтехимии и катализа УФИЦ РАН; Институт биохимии и генетики УФИЦ РАН; Институт вычислительной математики и математической геофизики СО РАН; Институт цитологии и генетики СО РАН; Институт теоретической и экспериментальной биофизики РАН; Институт математических проблем биологии РАН - филиал ИПМ им. М.В. Келдыша РАН.
- на международных и всероссийских конференциях: Международная научная конференция «Параллельные вычислительные технологии (ПаВТ)» (г. Пермь, 2020 г.), (г. Волгоград, 2021); Международная научная конференция «Уфимская осенняя математическая школа» (г. Уфа, 2020 г., 2021 г.); XXVIII Международная конференция студентов, аспирантов и молодых учёных «Ломоносов», секция «Биоинженерия и биоинформатика» (г. Москва, 2020 г.); IX Международная научная молодежная школа-семинар «Математическое моделирование, численные методы и комплексы программ» имени Е.В. Воскресенского (г. Саранск, 2020 г). Международная научно-практическая конференция «Интеллектуальные информационные технологии и математическое моделирование» (пос. Дивноморское, г. Геленджик, 2021 г., 2022 г.); The Thirteenth International Multiconference. Bioinformatics of Genome Regulation and Structure/Systems Biology (BGRS/SB-2022), (г. Новосибирск, 2022 г.).

Работа выполнена при поддержке гранта РФФИ (Аспиранты) № 20-37-90091 (2020-2022) «Разработка программы дизайна праймеров для Loop AMPlification - петлевой изотермической амплификации на основе технологий машинного обучения».

Личный вклад автора. Определение темы диссертационной работы, цели и задач исследования проводились автором совместно с научным руководителем. Объект исследования выбирался совместно с научным консультантом. Личный вклад

автора состоит в анализе математических алгоритмов поиска образцов в строке, обзоре существующих компьютерных программ, позволяющих проводить дизайн праймеров для LAMP-амплификации, разработке программного кода и интерфейса, компьютерного подбора праймеров для обнаружения генетического материала коронавируса SARS-CoV-2, интерпретации и обсуждении полученных результатов, их апробации, подготовки статей и тезисов докладов по теме работы.

Публикации. По материалам диссертационной работы опубликовано 17 научных трудов, из них 5 статей в журналах, индексируемых в WoS/Scopus и/или из перечня ВАК, 3 статьи в журналах, индексируемых в РИНЦ, 7 тезисов докладов на всероссийских и международных конференциях. Получено 2 свидетельства о регистрации программ для ЭВМ.

Структура и объем работы. Диссертационная работа изложена на 107 страницах, состоит из введения, 4 глав, заключения и списка литературы из 111 наименований и содержит 12 таблиц и 34 рисунка.

Работа выполнена в лаборатории математической химии Федерального государственного бюджетного учреждения науки Института нефтехимии и катализа Российской академии наук при финансовой поддержке Российского фонда фундаментальных исследований (проект № 20-37-90091).

Благодарности. Автор выражает искреннюю благодарность научному руководителю, доктору физико-математических наук, профессору Губайдуллину Иреку Марсовичу за постоянную поддержку и полезные советы на всех этапах выполнения диссертационной работы.

Автор также признателен коллегам из Института биохимии и генетики УФИЦ РАН: доктору биологических наук, профессору Чемерису Алексею Викторовичу за участие в выборе объекта исследования и обсуждение полученных результатов, кандидату биологических наук Гарафутдинову Равилю Ринатовичу за помощь в освоении методов молекулярной биологии, необходимых для выполнения диссертационной работы и консультации в ходе её выполнения, кандидату биологических наук Сахабутдиновой Ассоль Рафиковне за экспериментальную апробацию подобранных праймеров и интерпретацию полученных результатов.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Введение

Современные компьютерные технологии всё чаще применяются в различных областях не только науки, но и в повседневной жизни человека. Развиваются различные области, где данные цифровизируются, анализируются и моделируются с помощью современных возможностей компьютера. Эта тенденция коснулась и молекулярную биологию, в том числе анализ биологических образцов путем обнаружения в них специфических последовательностей нуклеиновых кислот (НК) с помощью амплификации. Одним из наиболее популярных методов является петлевая изотермическая амплификация (loop-mediated isothermal amplification (LAMP)) и для успешного проведения данной реакции необходимо правильно подобрать праймеры (последовательности из 16-30 нуклеотидов) к определенному участку, число копий которого необходимо увеличить. Дизайну LAMP-праймеров уделяют особое внимание. Известно, что последовательность ДНК содержит четыре нуклеотида: А – аденин и Т – тимин, Г – гуанин и Ц – цитозин и перестановок этих нуклеотидов огромное количество. Например, если взять последовательность из 20-ти нуклеотидов, их количество может быть более миллиарда (4^16 = 4 294 967 296, где 4 – четыре нуклеотида (A, T, G или C), а 16 – длина последовательности) и проанализировать вручную такое количество данных практически невозможно. К тому же существует ряд определенных критерий, который необходимо учитывать при дизайне праймерных систем именно для LAMP-амплификации, таких как длина праймеров, состав гуанина и цитозина (GC-состав), температура отжига праймеров (Tm) и.т.д. Поэтому необходимы новые современные компьютерные решения для дизайна специфичных праймеров для проведения именно петлевой изотермической амплификации.

1. Обзор современного состояния исследуемого объекта

1.1 Методы амплификации нуклеиновых кислот

Амплификация НК является основой современных методов исследования биологических объектов, включая такие прикладные области как молекулярная диагностика инфекционных и наследственных заболеваний, установление родства, ДНК-криминалистика, анализ объектов окружающей среды, продуктов питания и т.п. После разработки полимеразной цепной реакции (ПЦР) она очень быстро стала и остается диагностическим методом №1. ПЦР обеспечивает наработку продуктов (ампликонов) в геометрической прогрессии, упрощенно описываемой формулой 2^n, где 2 – две цепи денатурированной исходной ДНК, а n – число циклов этой реакции. Протекание ПЦР регулируется сменой температурных режимов, при которых идут следующие процессы: денатурация (обычно при 95°C), отжиг праймеров (обычно при 50-60°C), элонгация (обычно при 72°C). Каждый их этих этапов начинается по достижении соответствующей температуры, однако смена температуры в

реакционном блоке происходит не мгновенно, что приводит к искусственному сдерживанию ПЦР.

В связи с вышесказанным значительный интерес представляют изотермические реакции, протекающие при постоянной температуре (обычно 60-65°C). Достойной альтернативой ПЦР и вторым по масштабам применения в молекулярной диагностике является петлевая изотермическая амплификация (LAMP). Схема расположения LAMP-праймеров представлена на рисунке 1.

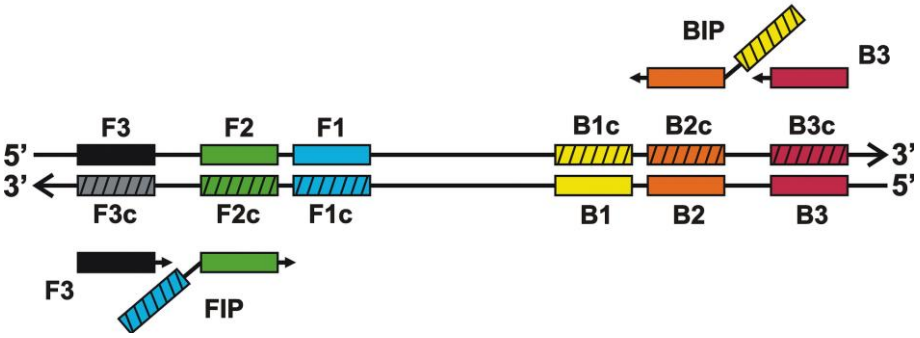

Рисунок 1 – Схема расположения праймеров для петлевой изотермической амплификации (LAMP)

В отличие от ПЦР, в классической LAMP используется четыре праймера: два внешних (F3 (Forward), B3 (Backward)) и два внутренних (FIP (Forward Inner Primer), BIP (Backward Inner Primer)). Предложен также вариант LAMP с 6 праймерами, подразумевающий добавление пары петлевых праймеров для увеличения точности реакции. Поскольку именно праймеры обеспечивают специфичность амплификации, к их подбору предъявляются жесткие требования. Для конструирования LAMP-праймеров написано несколько специальных компьютерных программ, однако их функционал ограничен и не удовлетворяет потребностей конечных пользователей.

1.2. Обзор существующих компьютерных программ для LAMP-амплификации

Разработано достаточно большое количество компьютерных программ (более 120) для подбора ПЦР-праймеров. Однако для подбора праймеров для LAMP написано не более десяти специальных программ. Наиболее широко используемым бесплатным онлайн-инструментом, обеспечивающим дизайн праймеров для LAMP, является программа **PrimerExplorer**, разработанная компанией Eiken Chemical Co LTD, Токио, Япония, версии 4 и 5. Программа достаточно быстрая, простая в использовании. В PrimerExplorer V5 существует три режима подбора праймеров (Automatic Judgment, Normal, User Assignment), а именно, создание автоматического, стандартного и определенного наборов праймеров. Поддерживаются три типа файловых форматов: простой текстовый формат (только последовательность), формат FASTA и формат GenBank. Однако программа ограничена в анализируемой нуклеотидной последовательности (до 2000 нуклеотидов) и не предусмотрено

изменение параметров для дизайна петлевых праймеров. Так же программа не учитывает такой важный критерий как сближенное расстояние праймеров.

Еще одна программа дизайна праймеров – **LAMP Designer** американской фирмы Premier Biosoft, позволяет подбирать набор как из четырех, так и из шести (с учетом петлевых) праймеров. Однако LAMP Designer является коммерческой программой и не обладает должным функционалом в плане необходимости расширять возможности экспериментаторов при дизайне необычных праймерных систем.

Программное обеспечение **FastPCR** финской фирмы Primer Digital Ltd. позволяет разрабатывать праймеры для различных видов ПЦР (стандартной, в реальном времени, мультиплексной и т.д.) и LAMP (в варианте четыре праймера). FastPCR является коммерческой программой и может быть установлено исключительно на платформе Microsoft Windows.

Программа **GLAPD** позволяет создавать наборы праймеров LAMP на основе последовательности полного генома или его участка. Затем отдельные праймеры объединяются в наборы LAMP-праймеров и сопоставляются с целевым участком и полным геномом. После этого выводятся наборы праймеров. GLAPD работает только в операционной системе Linux. Необходимы perl (высокоуровневый интерпретируемый динамический язык программирования общего назначения) и gcc (набор компиляторов для различных языков программирования).

В 2020 году была разработана программа **Lamprim**, функционирующая в двух режимах: дизайн праймеров и анализ праймеров. Программа написана на языке Python 3 с использованием библиотеки biopython и поддерживает платформы операционных систем Windows и Linux.

Компания New England Biolabs в августе 2020 г. выпустила бесплатную онлайн программу **NEB LAMP** PrimerDesignTool. Однако программа не учитывает повторы нуклеотидов в одном праймере, зачастую подбирает только один петлевой праймер, что нецелесообразно, и не учитывает возможность образования димеров праймеров в одном наборе, что может привести к ложным результатам.

Таким образом, разработка новой компьютерной программы дизайна специфичных праймерных систем для проведения LAMP-амплификации является весьма актуальной и востребованной задачей.

1.3. Обзор математических алгоритмов поиска праймеров в нуклеотидной последовательности

Задачу поиска праймеров в нуклеотидной последовательности можно сопоставить с задачей поиска подстроки (некого образца) в более длинной строке (или тексте), где праймер будет являться образцом, а нуклеотидная последовательность текстом. Пусть задана короткая подстрока A, именуемая образцом (праймер), и более длинная строка T, именуемая текстом (нуклеотидная последовательность). Задача заключается в отыскании всех вхождений подстроки A в

строке T. В случае стандартной LAMP: поиск шести образцов (мест присоединения праймеров) в тексте. В расширенной LAMP: поиск восьми образцов в тексте.

Существуют алгоритмы, которые можно применить для решения данной задачи. Рассмотрим некоторые из них. Самым простым алгоритмом, который может быть использован в данном случае, является **прямой поиск**. Сравнение происходит слева направо. Крайний левый конец образца находится на одной позиции с крайним левым концом текста. В случае несовпадения между образцом и строкой сдвиг образца происходит на одну позицию, или при полном совпадении всех элементов подстроки и строки (тогда возникает необходимость обозначить вхождение праймеров нуклеотидной последовательности). Поиск продолжается до достижения конца длинной строки. Сложность алгоритма составляет O(n·m), где n – длина подстроки *A,* m – длина строки *T,* т.е. напрямую зависит от количества найденных образцов и длины строки.

Одним из классических алгоритмов поиска подстроки в строке является **алгоритм Кнута-Морриса-Пратта**. Данный алгоритм является наиболее популярным алгоритмом с линейным временем для поиска точных совпадений подстроки со строкой. Идея алгоритма Кнута-Морриса-Пратта заключается в увеличении расстояния сдвига подстроки по строке, тем самым сократив количество сравнений. Сложность алгоритма также зависит от объёма входных данных и определяется как O(n+m), где n – длина подстроки, m – длина строки T.

Рассмотрен еще один классический алгоритм поиска подстроки в строке. Это **алгоритм Бойера-Мура**. Он является самым быстрым алгоритмом общего назначения для поиска подстроки в строке среди известных классических алгоритмов. Суть данного алгоритма аналогична алгоритму Кнута-Морриса-Пратта. Однако есть существенное различие, которое заключается в том, что сравнение совпадения при использовании алгоритма Бойера-Мура ведется справа налево, иными словами, проверка начинается с последнего символа подстроки, которую нужно найти. Также есть правило сдвига плохого символа, которое позволяет сдвигаться сразу на несколько позиций, значительно сократив время поиска.

При дизайне праймеров для LAMP необходимо проводить поиск сразу нескольких образцов. Поэтому были рассмотрены алгоритмы, которые можно применить для поиска сразу нескольких вхождений.

Был рассмотрен **алгоритм Рабина-Карпа**. Он был разработан в 1987 году Майклом Рабином (Израиль) и Ричардом Карпом (США). Алгоритм работает на основе хеширования. Хеширование, или хеш-функция – это функция, которая преобразует массив данных произвольной длины в битовую строку фиксированной длины (хэш). Сложность данного алгоритма можно оценить как O(n), где n – это длина текста. Но для такого хорошего результата необходимо правильно выбрать хеш-функцию. В ином случае сложность алгоритма будет равна O(m·n), где n – длина текста, m – длина шаблона, что является одной из причин того, почему данный алгоритм не слишком широко используется.

Для сравнения был рассмотрен ещё один алгоритм – это **алгоритм Ахо-Корасик**, разработанный Альфредом Ахо (Канада) и Маргарет Корасик в 1975 году. Он позволяет найти сразу все вхождения образцов в тексте. В нем используется конечный автомат, в результате которого образуется префиксное дерево, бор. Узлы дерева должны соответствовать префиксам исходного образца. А для того, чтобы проводить поиск и переходить по узлам бора, необходимы суффиксные ссылки, которые находятся на узле самого длинного суффикса и позволяют продолжать поиск. Сложность алгоритма линейно зависит от объема входных данных и определяется как O(n+m+z), где n – длина образца, m – длина строки, z – общее количество вхождений образца в текст.

Однако данные алгоритмы целесообразно применять в случаях, когда уже известна последовательность, которую необходимо найти, а для решения поставленной цели и задач, приведенные выше алгоритмы не вполне удовлетворяют, поскольку имеются определенные критерии подбора праймеров для LAMP, которые необходимо учитывать, а именно:

– длина праймера (15-30 нуклеотидов, для внутренних праймеров 30-50 нуклеотидов);

– GC-состав (оптимально 40-60%);

– температура отжига (оптимальная Tm в пределах 55-65°C);

– исключение образования димерности праймеров (с перекрытием более 2 нуклеотидов);

– исключение повторяющихся нуклеотидов в праймере.

2. Дизайн праймеров для LAMP-амплификации на основе существующих алгоритмов

2.1 Компьютерная реализация алгоритма Рабина-Карпа

Учитывая тот факт, что для LAMP-амплификации необходимо большое количество праймеров, целесообразно рассмотреть алгоритм Рабина-Карпа и алгоритм Ахо-Корасик. Для ускорения работы алгоритма Рабина-Карпа был реализован параллельный поиск праймеров, позволяющий производить одновременный поиск нескольких образцов в тексте. Каждому потоку отдается один из шести праймеров. Далее проводился поиск вхождения праймера в нуклеотидной последовательности.

Алгоритм Рабина-Карпа был реализован на языке программирования Python - высокоуровневый язык программирования. Выбор данного языка обусловлен наличием самой большой и популярной библиотеки biopython, которая содержит различные подмодули для решения задач биоинформатики.

Результаты ускорения для двух и четырех процессов представлены на рисунке 2. Расчеты проводились на ПК с характеристиками: процессор: Intel(R) Core(TM) i3-3210 CPU 3,20 ГГц, 2 ядра, ОЗУ 8 ГБ.

С увеличением мощности процессора увеличивается эффективность многозадачных вычислений. Среднее ускорение на двух процессах составило 1,7, на четырех процессах 2,6.

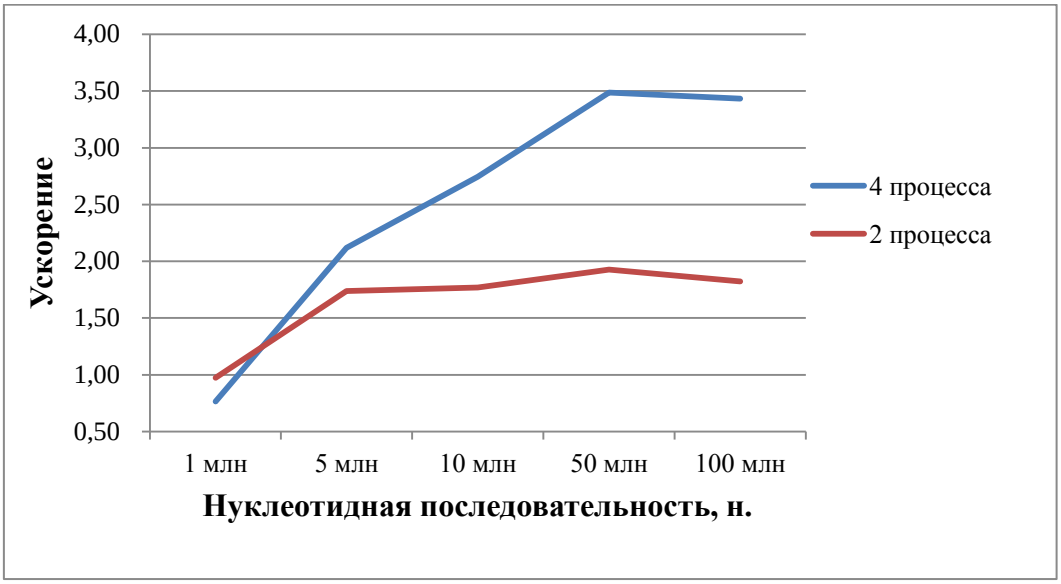

Рисунок 2 - Влияние длины нуклеотидной последовательности на ускорение работы алгоритма Рабина-Карпа в зависимости от количества процессов

2.2 Компьютерная реализация алгоритма Ахо-Корасик

Так же был рассмотрен алгоритм Ахо-Корасик для поиска праймеров в нуклеотидных последовательностях различной длины. На рисунке 3 представлены графики зависимости работы алгоритма для последовательного и многоядерного поиска по времени в зависимости от длины нуклеотидной последовательности. Среднее ускорение составило 1,78.

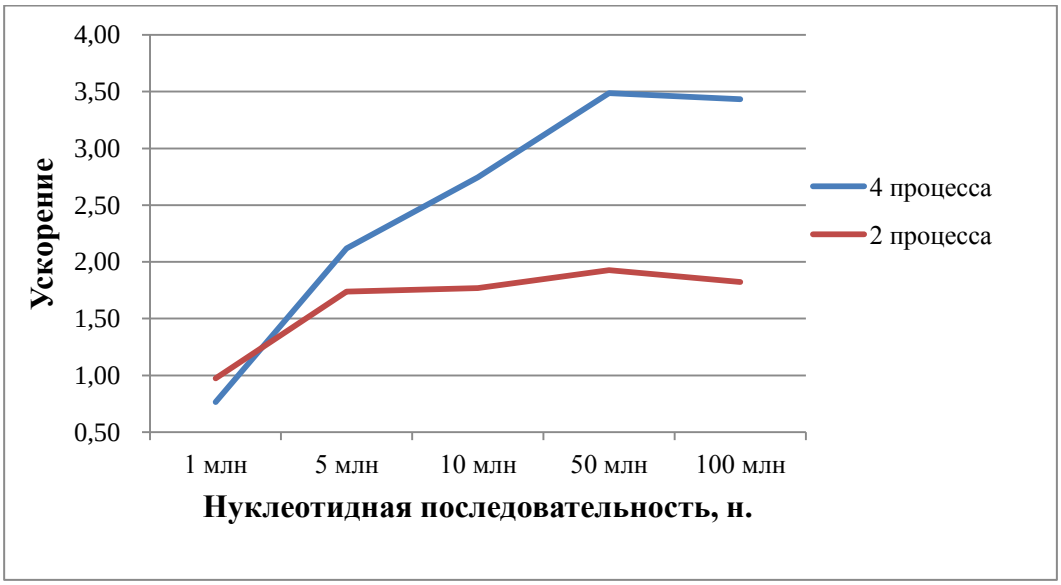

Рисунок 3 – Влияние длины нуклеотидной последовательности на длительность работы алгоритма Ахо-Корасик

Использование реализованных алгоритмов для решения поставленной цели и задач в данной работе приведет к делению дизайна праймеров на несколько этапов, хранению и анализу последовательности по несколько раз и в больших объемах.

Поэтому возникает необходимость разработать новую систему дизайна LAMP-праймеров, позволяющую снизить сложность перебора за счет использования трафаретного подхода, который в процессе поиска будет сразу рассчитывать GC-состав и температуру отжига праймеров. Новая система позволит организовать дизайн праймеров за короткое время и с минимальными затратами компьютерных ресурсов.

3. Дизайн праймерных систем для проведения LAMP-амплификации

3.1. Алгоритм расчета GC-состава и температуры отжига праймеров

Задача поиска праймеров в нуклеотидной последовательности также представляется как поиск образца *A* в строке *T*, однако поиск будет усложнен за счет учета длины праймера, состав гуанина и цитозина (GC-состав), и температуры отжига (Tm) праймеров. Поиск заключается в отыскании всех образцов *A,* удовлетворяющих условиям необходимым для проведения LAMP-амплификации, а именно длина праймера (15-30 нуклеотидов), GC-состав (40-60%) и Tm (55-65°C) в строке *T*. Учитывая что для стандартной LAMP-амплификации необходимо четыре праймера (два внешних, два внутренних), то мест отжига необходимо шесть, за счет удвоенной длины внутренних праймеров. А в случае LAMP-амплификация с петлевыми праймерами (к основным праймерам добавляются еще два петлевых) – всего восемь.

Предположим, что наша задача найти все последовательности, состоящие из 15 нуклеотидов, GC-состав от 45% до 55% и Tm от 62° до 65°.

За основу используется метод прямого перебора (прямой поиск). Сравнение будет происходить слева направо, однако вместо сравнения одного нуклеотида, анализируются первые 15, т.е. рассчитывается GC-состав. Если он удовлетворяет заданным условиям, рассчитывается Tm, в ином случае происходит сдвиг на одну позицию. Если же новый (добавленный / исключённый из последовательности) нуклеотид является C или G, происходит перерасчет GC-состава, и при изменении GC-состава, пересчитывается Tm.

Пример такого поиска представлен на рисунке 4.

																↓	↓	↓	↓	↓		
строка	A	G	T	C	T	A	T	A	C	C	T	A	C	G	G	A	T	A	G	C	T	
образцы	A	G	T	C	T	A	T	A	C	C	T	A	C	G	G	A						-
		G	T	C	T	A	T	A	C	C	T	A	C	G	G	A	T					+
			T	C	T	A	T	A	C	C	T	A	C	G	G	A	T	A					-
				C	T	A	T	A	C	C	T	A	C	G	G	A	T	A	G			+
					T	A	T	A	C	C	T	A	C	G	G	A	T	A	G	C		+
						A	T	A	C	C	T	A	C	G	G	A	T	A	G	C	T	-

Рисунок 4 – Схематичное изображение поиска праймера удовлетворяющего заданным условиям («+»-перерасчет)

В данном примере проведено 10 сравнений символов образца и строки.

На рисунках 5 представлена блок-схема дизайна праймеров по GC-составу: расчет происходит только для первой последовательности, далее алгоритм обращает внимание только на последний (добавленный) нуклеотид, либо же на первый (исключённый из последовательности) нуклеотид, и если тот является C или G, то состав пересчитывается, если иным, то происходит сдвиг на одну позицию. GC-состав рассчитывается пропорционально длине праймера и должен быть в пределах от 40% до 60%.

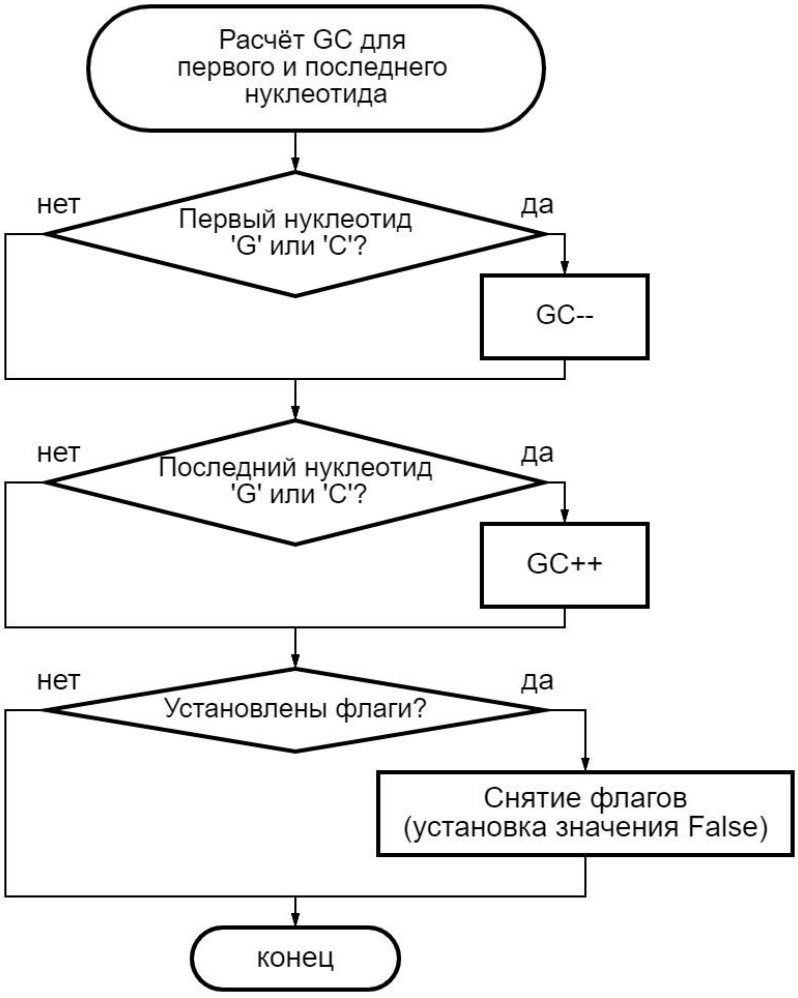

Рисунок 5 – Расчет GC-состава последовательности

Тоже происходит и с температурой отжига праймеров. Расчет Tm происходит только в том случае, если меняется GC-состав, иначе нет. На рисунках 6 представлена схема дизайна праймеров по температуре отжига праймеров.

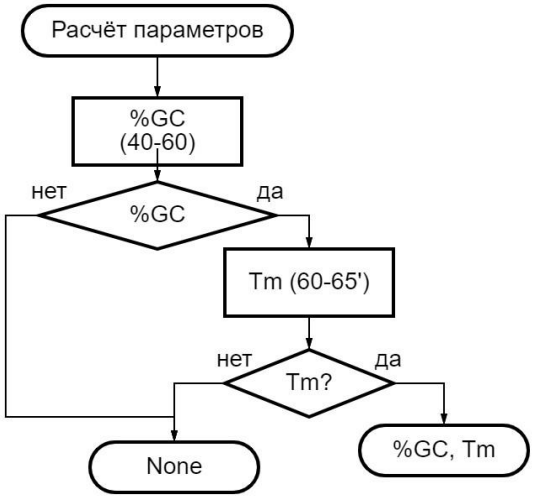

Рисунок 6 – Расчет Tm праймеров в зависимости от GC-состава

Решением обратной задачи была выведена новая формула расчета Tm которая учитывает длину праймера, GC-состав и концентрацию [Na^+]:

$$Tm = 81.5 + 16.6 \cdot (log_{10}[Na^+]) + 0.41 \cdot (\%G + \%C) - \frac{548}{L} \qquad (1)$$

где L обозначает длину праймера; [Na^+] – молярная концентрация ионов Na^+; %G+%C – процентное содержание GC-пар в исследуемом участке нуклеотидной последовательности.

На рисунке 7 представлены кривые, построенные по значениям Tm, полученным с помощью выведенной нами формулы, известной формулы (2) и онлайн-утилиты OligoAnalyzer:

$$Tm = 81.5 + 16.6 \cdot (log_{10}[Na^+]) + 0.41 \cdot (\%G + \%C) - \frac{600}{length} \qquad (2)$$

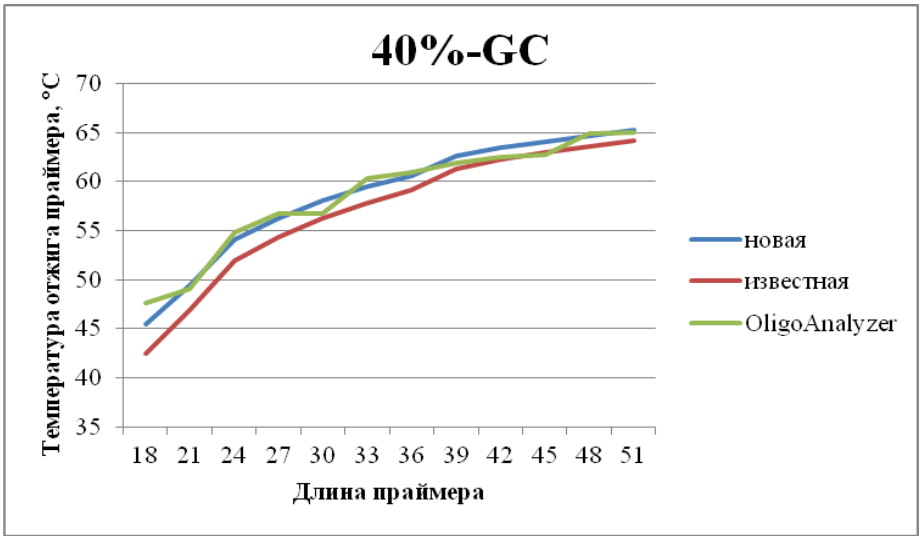

Рисунок 7 – Сравнение значений Tm праймеров, полученных с помощью различных формул расчета (для GC-состава 40%)

Выведенная формула расчета температуры отжига праймеров более близка к данным онлайн-утилиты OligoAnalyzer, по сравнению с известной формулой (2).

Для оценки функционала разработанного алгоритма проведено тестирование поиска всех возможных праймеров в нуклеотидных последовательностях различной длины (результаты представлены в таблице 1).

Таблица 1 – Продолжительность поиск праймеров на примере нуклеотидных последовательностей разной длины (расчет проводился на ноутбуке с параметрами: процессор Intel(R) Core(TM) i7-10750H CPU, 2.60GHz, 6 ядер. 16 ГБ ОЗУ).

Организм	Размер нуклеотидной последовательности, п.о.	Продолжительность поиска праймеров, с
SARS-CoV-2	29 903	0,31
Escherichia virus T4	168 903	1,73
Mycoplasma	580 076	5,43
Helicobacter pylori	1 624 458	18,11
Escherichia coli	4 641 652	71,68 (1,2 мин)
Caenorhabditis	100 286 401	1082,53 (18 мин)

Продолжительность поиска экспоненциально зависит от длины последовательности. Разработанный алгоритм обеспечивает высокую скорость поиска праймеров. Так, даже для протяженных последовательностей (например, длиной до 4 млн. нуклеотидов) продолжительность поиска занимает менее 1 мин.

После завершения поиска потенциальных праймеров их необходимо сформировать в наборы с учетом расстояния между ними. Рекомендуемое расстояние для F3/F2 и B3/B2 составляет 1-10, для F2/F1c – 10-25, для F1c/B1c – 0-30 нт, что обеспечит получение ампликонов длиной 120-220 п.о. Кроме того, необходимо провести анализ праймеров на предмет образования ими гомо- или гетеродимерных структур, а также учесть разницу в заданной величине Tm (ΔTm = 2-3°) в пределах набора.

Общая сложность алгоритма в худшем случае $O(m \cdot n)$, где n - длина праймера, амортизированная временная сложность составляет $O(m)$, где m - длина последовательности. В целом, продолжительность работы алгоритма зависит от того, насколько часто встречаются фрагменты нуклеотидной последовательности, удовлетворяющие критериям поиска.

3.2. Компьютерная программа дизайна праймеров для LAMP - амплификации.

Комплекс программ был реализован на высокоуровневом языке программирования Python, с применением библиотеки biopython.

Функциональные возможности программы:

1) Загрузка последовательности из файла (простой текстовый формат (только последовательность), формат FASTA, GenBank) или через буфер обмена.

2) Выбор последовательности для поиска праймеров.

3) Поиск праймеров:

Праймеры объединяются по критериям:

- проверка длины ампликона;
- учет расстояния между праймерами;
- учет отклонения Tm (<3°C);
- исключение образования димерных структуры;

(если все условия соблюдены, то такой набор праймеров рассматривается как пригодный для LAMP).

4) Вывод подобранных наборов праймеров на экран и/или сохранение в файл в формате *.xls.

На рисунках 8 и 9 представлен интерфейс программы, разработанный с помощью Фреймворка Qt, позволяющего создавать приложения с графическим интерфейсом.

Кнопка «Primer Design Parameters» позволяет изменить дополнительные конфигурации дизайна праймеров. Стандартные значения установлены по умолчанию.

Кнопка «Search» запускает дизайн праймеров. Далее, если удалось подобрать наборы праймеров согласно заданным параметрам, открывается окно нажатием на кнопку «Open Results».

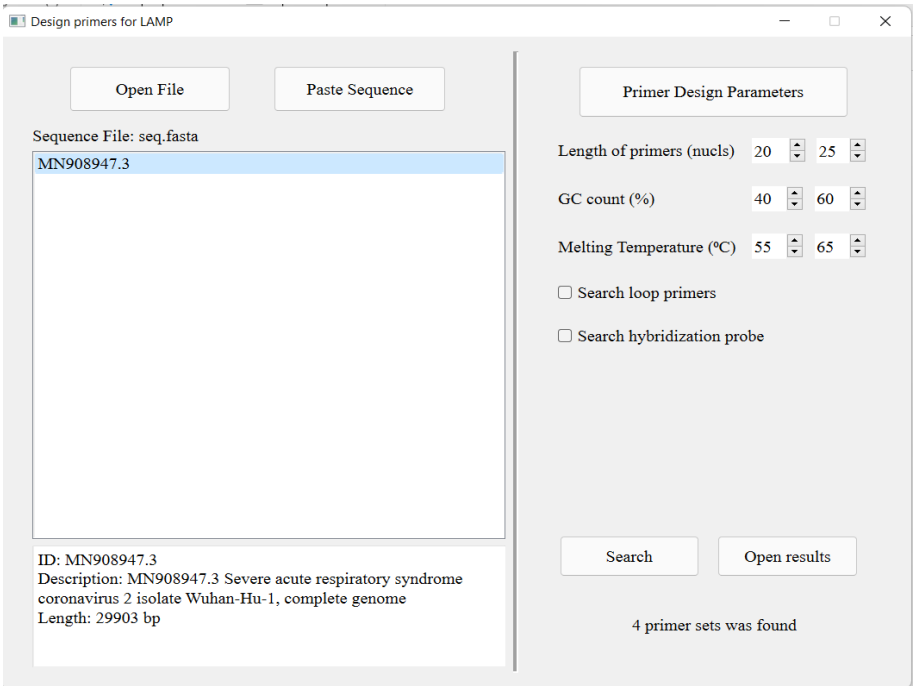

Рисунок 8 – Интерфейс программы LAMPrimers iQ

Текущий набор либо все наборы вместе сохраняются в формате Excel при нажатии на кнопку «Save to Excel».

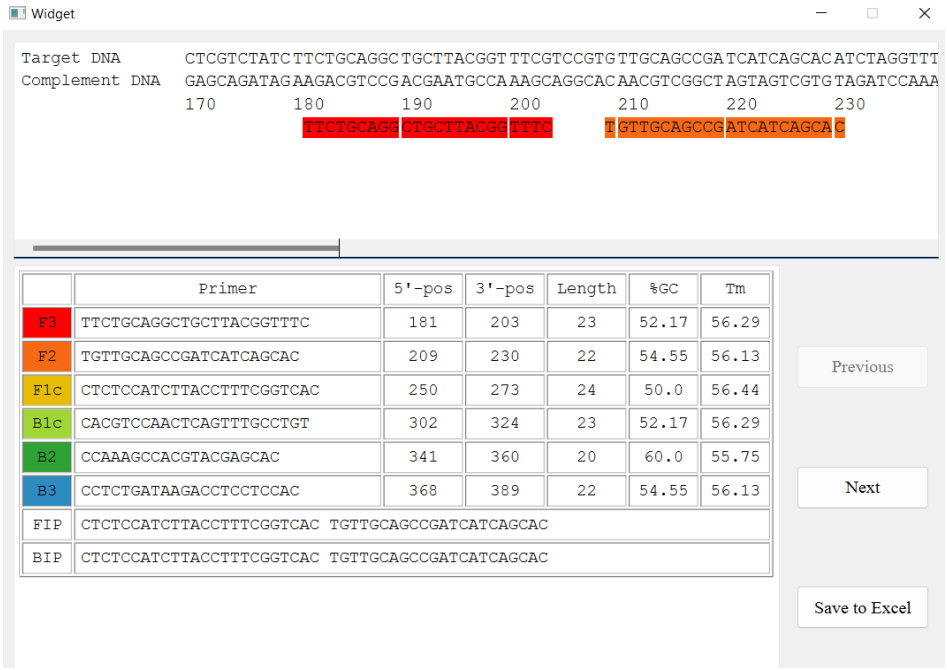

Рисунок 9 – Интерфейс результатов работы программы

Комплекс программ дизайна праймеров для LAMP-амплификации зарегистрирован в Реестре программ для ЭВМ под названиями: LAMPrimers iQ № 2022617417 от 20 апреля 2022 г. и LAMPrimers iQ_loop № 2023662840 от 14 июня 2023 г. Код программы в открытом доступе: https://github.com/Restily/LAMPrimers-iQ/blob/main/lamp/start_lamp.py.

Разработанный алгоритм, модель и код программы были протестированы на нуклеотидных последовательностях различной структуры. Результаты представлены в следующей главе.

4. Тестирование и экспериментальное исследование

4.1 Результаты подбора праймеров в зависимости от заданных параметров

Тестирование функционала программы проводилось на примере бактериофага *Lambda* длиной 48502 нуклеотидов (~48,5 Кб). Выбор данной последовательности обусловлен большим количеством G и C в его составе (GC-богатая нуклеотидная последовательность). В таблице 2 приведено количество полученных наборов праймеров в зависимости от заданных критериев подбора, а именно, варьировался параметр GC-состава (от 40% до 65%), разница Tm в одном наборе (2° или 5°) и размеры амплифицируемой области (от 160 п.о. до 300 п.о.). В случае жестких условий программа может не выдать ни одного набора.

Оказалось, что значительное влияние на количество наборов оказывает размер амплифицируемой области: при длине 160 п.о. формируются единичные наборы праймеров, или они не находятся вовсе. Следствием этого является необходимость анализа более длинных нуклеотидных последовательностей.

Таблица 2 – Количество наборов праймеров в зависимости от заданных параметров поиска

GC%	ΔTm,°	максимальный размер амплифицируемой области, п.о.		
		300	230	160
40-60	5	213	119	3
	2	198	184	3
45-55	5	132	80	0
	2	116	64	0
50-60	5	195	185	4
	2	181	164	4
55-65	5	160	147	4
	2	134	105	0

Для относительно коротких нуклеотидных последовательностей (до 2000 нуклеотидов) подбор праймеров занимает менее одной секунды. На рисунке 10 приведены данные анализа нуклеотидной последовательности бактериофага *Lambda*, полученные на ноутбуке с параметрами: процессор Intel(R) Core(TM) i7-10750H CPU, 2.60GHz, 6 ядер. 16 ГБ ОЗУ; заданы средние критерии подбора праймеров (40-60% GC, ΔTm = 5°, длина анализируемого участка до 300 нуклеотидов).

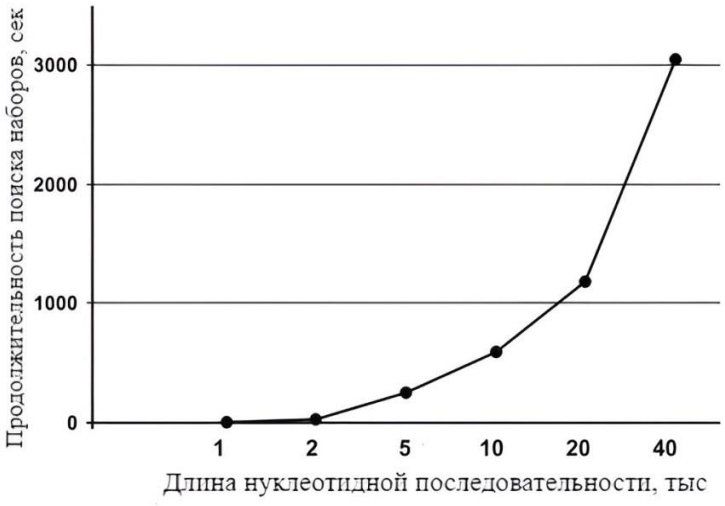

Рисунок 10 - Влияние длины нуклеотидной последовательности на длительность подбора наборов праймеров (найдено для нуклеотидной последовательности бактериофага *Lambda*)

С увеличением длины последовательности длительность поиска праймеров увеличивается экспоненциально.

4.2 Лабораторная экспериментальная оценка эффективности работы LAMPrimers iQ

Для оценки качества праймеров, подбираемых с помощью программы LAMPrimers iQ, коллегами из Института биохимии и генетики УФИЦ РАН была проведена серия экспериментов по обнаружению генетического материала коронавируса SARS-CoV-2. Интерес к данному объекту обусловлен высокой скоростью распространения вирусной инфекции, которая вызывает у человека опасную болезнь, получившую название COVID-19. Данная нуклеотидная последовательность является АТ-богатой и праймеры, удовлетворяющие критериям подбора будут более длинными.

Для проведения натурных экспериментов к одному и тому же участку гена *S* белка этого вируса были подобраны комплекты праймеров с помощью программы LAMPrimers iQ и доступных онлайн-утилит от компаний New England Biolabs (Neb LAMP) и Eiken Chemical (PrimerExplorer).

На рисунке 11 представлены кривые LAMP-амплификации, проведенной с разными комплектами праймеров. Показано, что праймеры, подобранные с помощью программы LAMPrimers iQ, обеспечивают более высокую специфичность анализа, обусловленную снижением скорости протекания неспецифической реакции. Это выражается в отсутствии подъема кривых для образцов отрицательного контроля (L-) в отличие от тестируемых образцов, содержащих генетический материал вируса (L+).

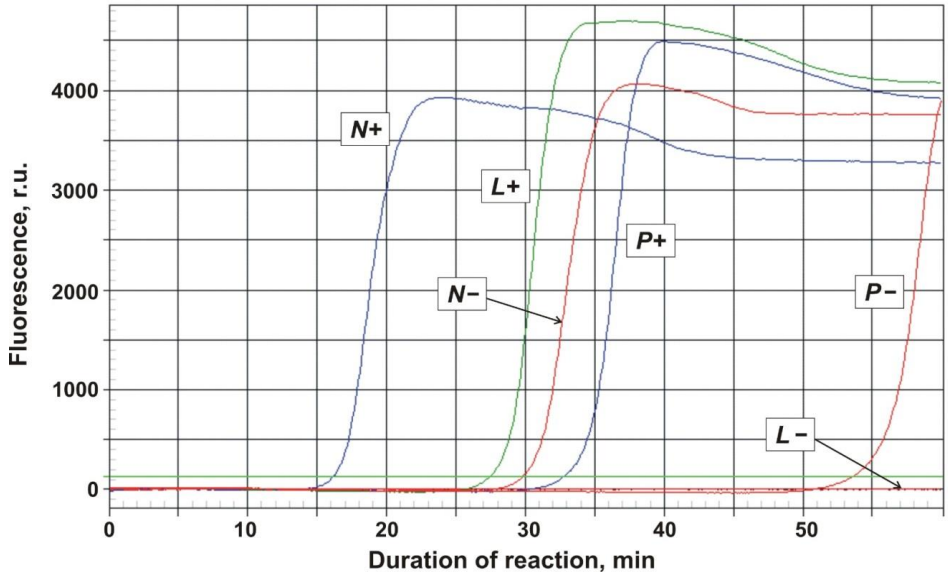

Рисунок 11 – Кривые LAMP-амплификации, проведенной с разными комплектами праймеров: L – праймеры, полученные с помощью LAMPrimers iQ; N –NEB LAMP Primer Design; P – PrimerExplorer; "+" образцы содержали РНК коронавируса SARS-CoV-2; "-" образцы – контрольные образцы (не содержали нуклеиновых кислот).

Применимость праймеров, подобранных с помощью LAMPrimers iQ, для выявления патогенной РНК была также оценена на выборке клинических образцов,

полученных от больных с подтвержденным с помощью ПЦР диагнозом COVID-19 (рисунок 12).

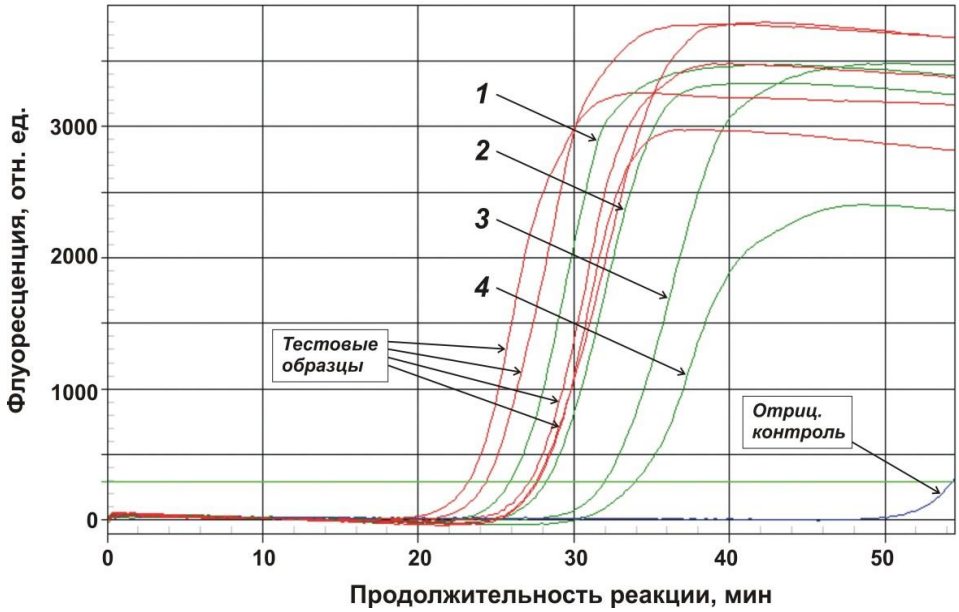

Рисунок 12 - Кривые амплификации, полученные при детекции РНК коронавируса SARS-CoV-2 с использованием праймеров набора L: 1-4 - кривые, соответствующие последовательным разбавлениям смешанного клинического образца Rmix (выделены зеленым цветом: 1 – без разбавления, 2 – разбавление в 10 раз, 3 – в 100 раз, 4 – в 1000 раз). Тестовые образцы – образцы, содержавшие индивидуальные лизаты носоглоточных мазков больных COVID-19 (выделены красным цветом), отрицательный контроль – образец, не содержавший нуклеиновых кислот (выделен синим цветом).

Образцы от больных COVID-19 показали ранний подъем кривых амплификации (в диапазоне 20-30 мин от начала реакции). Разница в величинах порогового времени для данных образцов была обусловлена различием в количестве РНК-мишеней (вирусной нагрузки). Таким образом, программа LAMPrimers iQ обеспечивает подбор «качественных» праймеров, обеспечивающих достоверное обнаружение специфических НК-мишеней.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ РАБОТЫ И ВЫВОДЫ

1. Смоделирована новая система подбора праймеров для LAMP, включающая новую формулу расчета температуры отжига праймеров, возможность расположения их на максимально близком расстоянии и исключения потенциальных димерных структур. Полученная модель позволяет предотвратить получение ложных результатов амплификации.

2. Предложен комплексный алгоритм прямого перебора с использованием трафаретного подхода, учитывающего GC-состав, температуру отжига и позволяющего снизить сложность перебора. Проведено тестирование предложенного алгоритма на нуклеотидных последовательностях различной структуры (геномы

SARS-CoV-2, Escherichia virus T4, Mycoplasma, Helicobacter pylori , Escherichia coli, Caenorhabditis). Тестирование показало быструю скорость и адекватность работы алгоритма (для *SARS-CoV-2 coronavirus* подбираются ~8 000 праймеров за 0,3сек)

3. Разработан комплекс программ с дружелюбным интерфейсом, учитывающий критерии дизайна праймеров. Получены свидетельства о регистрации программ для ЭВМ (№ 2022617417 «LAMPrimers-iQ», № 2023662840 «LAMPrimers iQ_loop»). Программа доступна по адресу: https://github.com/Restily/LAMPrimers-iQ. На примере генетического материала коронавируса SARS-CoV-2 показано, что для успешной амплификации РНК, характеризующейся склонностью к разрушению, необходимо использовать сближенные праймеры.

4. В амплификационных экспериментах было показано, что праймеры, подобранные с помощью программы LAMPrimers iQ, обеспечивают более высокую специфичность анализа, по сравнению с существующими аналогами (NEB LAMP, PrimerExploper), обусловленную снижением вероятности протекания неспецифической реакции.

На основе разработанного комплекса программ планируется продолжить ряд экспериментальных исследований, описать полученные кривые амплификации, повысить функционал программ.

ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

По материалам диссертационной работы опубликовано 17 научных трудов, из них 5 – статьи в журналах, индексируемых в WoS/Scopus и/или из перечня ВАК, 3 – статьи в РИНЦ, 7 тезисов докладов на всероссийских и международных конференциях. Получено 2 свидетельства о регистрации программ для ЭВМ.

1. **Akhmetzianova L.U,** Davletkulov T.M, Sakhabutdinova A.R, Chemeris A.V, Gubaydullin I.M, Garafutdinov R.R. LAMPrimers iQ: New primer design software for loop-mediated isothermal amplification (LAMP) // Anal Biochem. 2023 Nov 2;684:115376. doi: 10.1016/j.ab.2023.115376. WoS, Scopus, ВАК, РИНЦ.
2. **Ахметзянова Л.У.**, Давлеткулов Т.М., Гарафутдинов Р.Р., Губайдуллин И.М. Применение алгоритма Ахо-Корасик для подбора праймеров для петлевой изотермической амплификации // Математическая биология и биоинформатика. 2022. Т. 17. № 2. С.250-265. doi: 10.17537/2022.17. ВАК, Scopus.
3. **Ахметзянова, Л.У.** Компьютерная программа подбора праймеров для LAMP-амплификации // Advanced engineering research. 2024. Т. 24(1). С. 98-108. ВАК.
4. **Akhmetzianova L.U.**, Davletkulov T.M., Gubaidullin I.M., Islamgulov A.R. Parallel implementation of the primer search algorithm for loop-mediated isothermal amplification // Journal of Physics: Conference Series. 2021. V. 2131, № 2. Paper 022004. doi.org/10.1088/1742-6596/2131/2/022004. Scopus.
5. Кирьянова О.Ю., **Ахметзянова Л.У.**, Губайдуллин И.М. Алгоритмы поиска в задачах анализа нуклеотидных последовательностей с целью однозначной идентификации геномов // Вестник Башкирского университета. 2020. Т. 25. № 2. С. 285-290. doi.org/10.33184/bulletin-bsu-2020.2.10. ВАК.

6. Свидетельство о государственной регистрации программы для ЭВМ № 2022617417 от 20.04.2022. «LAMPrimers iQ (Loop-mediated isothermal amplification Primers iQ)». Авторы: **Л.У. Ахметзянова**, Т.М. Давлеткулов, Р.Р. Гарафутдинов, А.В. Чемерис, И.М. Губайдуллин.
7. Свидетельство о государственной регистрации программы для ЭВМ № № 2023662840 от 14.06.2023. «Дизайн праймеров для петлевой изотермической амплификации (LAMPrimers iQ_loop)». Авторы: **Л.У. Ахметзянова**, Т.М. Давлеткулов, В.Д. Хайритдинов, Р.Р. Гарафутдинов, И.М. Губайдуллин.
8. Kiryanova O. Y., **Akhmetzianova L.U.**, Kuluev B.R., Gubaydullin I. M., Chemeris A.V. Computer modelling of primers search in the DNA chain // Computational mathematics and information technologies. 2019. Vol. 1. № 1. P. 29-34. РИНЦ
9. Гарафутдинов Р.Р., Чемерис Д.А., Мавзютов А.Р., **Ахметзянова Л.У.**, Давлеткулов Т.М., Губайдуллин И.М., Чемерис А.В. Петлевая LAMP амплификация нуклеиновых кислот. I. Два десятилетия развития и совершенствования // Биомика. 2021. Т. 13. № 2. С. 176-226. РИНЦ
10. Гарафутдинов Р.Р., **Ахметзянова Л.У.**, Сахабутдинова А.Р., Чемерис Д.А., Губайдуллин И.М., Чемерис А.В. Петлевая LAMP амплификация нуклеиновых кислот. II. Детекция коронавируса SARS-CoV-2 с помощью различных вариантов обратно-транскрипционной петлевой амплификации // Биомика. 2023. Т. 15. № 4. С. 272-290. РИНЦ
11. Kiryanova O.Yu., Kiryanov I.I., **Akhmetzianova L.U.**, Kuluev B.R., Chemeris A.V., Gubaydullin I.M. Parallel implementation of search algorithm for RNA guide design // Сборник: Параллельные вычислительные технологии (ПаВТ'2020). Короткие статьи и описания плакатов. 2020. С. 52-58. *Короткая статья по результатам конференции. РИНЦ*
12. **Ахметзянова Л.У.**, Кирьянова О.Ю., Губайдуллин И.М. Компьютерное моделирование петлевой изотермической амплификации. В книге: Уфимская осенняя математическая школа – 2020 // Сборник тезисов международной научной конференции. Уфа. 2020. С. 169-171.
13. Kiryanova O.Yu., Kiryanov I.I., **Akhmetzianova L.U.**, Kuluev B.R., Chemeris A.V. The method of generation barcode for DNA certification of plants and organisms // Сборник трудов по материалам VI Международной конференции и молодежной школы. В 4-х томах. Под редакцией В.А. Фурсова. 2020. С. 292-296.
14. **Ахметзянова Л.У.**, Кирьянова О.Ю., Губайдуллин И.М. Поиск коротких фрагментов в анализе нуклеотидных последовательностей // IX Международная научная молодежная школа-семинар «Математическое моделирование, численные методы и комплексы программ» имени Е.В. Воскресенского. 2020 г.
15. **Ахметзянова Л.У.**, Давлеткулов Т.М., Гарафутдинов Р.Р., Чемерис А.В., Губайдуллин И.М. Параллельный поиск с использованием алгоритма Рабина−Карпа для петлевой изотермической амплификации ДНК // В сборнике: Параллельные вычислительные технологии (ПаВТ-2021). Короткие статьи и описания плакатов. XV международная конференция. Челябинск, 2021. С. 278.

16. **Ахметзянова Л.У.**, Давлеткулов Т.М., Губайдуллин И.М. Дизайн праймеров для петлевой изотермической амплификации // Уфимская осенняя математическая школа: Материалы международной научной конференции (г. Уфа, 6-9 октября, 2021 г.). Т. 2. 2021. С.144-146.
17. **Akhmetzianova L.U.**, Davletkulov T.M., Garafutdinov R.R., Gubaydullin I.M., Chemeris A.V. Design primers for LAMP-amplification // Bioinformatics of Genome Regulation and Structure/Systems Biology (BGRS/SB-2022): Abstracts the Thirteenth International Multiconference, Novosibirsk, 04–08 July 2022 г. 2022. P. 24-25.