Математическая биология и биоинформатика
2022. Т. 17. № 2. С. 250–265. doi: 10.17537/2022.17.250

========== **БИОИНФОРМАТИКА** ==========

Применение алгоритма Ахо-Корасик для подбора праймеров для петлевой изотермической амплификации

Ахметзянова Л.У.[*] [1,2], Давлеткулов Т.М.[1], Гарафутдинов Р.Р.[3], Губайдуллин И.М.[1,2]

[1]*Уфимский государственный нефтяной технический университет, Уфа, Россия*
[2]*Институт нефтехимии и катализа – обособленное структурное подразделение Федерального государственного бюджетного научного учреждения УФИЦ РАН, Уфа, Россия*
[3]*Институт биохимии и генетики – обособленное структурное подразделение Федерального государственного бюджетного научного учреждения УФИЦ РАН. Уфа, Россия*

Аннотация. В работе представлены результаты разработки компьютерной программы, позволяющей проводить дизайн (подбор) праймеров для выявления целевой нуклеотидной последовательности с помощью петлевой изотермической амплификации (loop-mediated isothermal amplification, LAMP). Приведен обзор наиболее популярных программ дизайна праймеров для LAMP.

В работе описаны условия, которые необходимо учитывать при подборе праймеров для петлевой изотермической амплификации, а именно: длина праймеров, GC-состав, средний размер ампликона, температура отжига праймеров, расстояние между праймерами.

При подборе праймеров необходимо проводить поиск позиций вхождения нескольких образцов (праймеров) в нуклеотидной последовательности. Так как для проведения петлевой изотермической амплификации используется набор как минимум из четырех праймеров, имеющих гомологию с шестью участками нуклеотидной последовательности, для реализации поиска был реализован алгоритм Ахо-Корасик, который позволяет производить одновременный поиск нескольких вхождений в более длинной последовательности.

Разработанная программа позволяет находить праймеры для последовательностей разной длины и группировать их по наборам, которые формируются согласно критериям подбора и начальным условиям, определяемым пользователем. В результате проведенного компьютерного анализа пользователь может выбрать из предложенного перечня наборов праймеров наиболее оптимальный для конкретного эксперимента. Тестовый набор праймеров подбирался для проведения петлевой изотермической амплификации генома с целью выявления РНК коронавируса SARS-CoV-2, вызывающего заболевание COVID-19.

Программа реализована на языке программирования Python с применением библиотек biopython, pyahocorasick и доступна по адресу: https://cloud.mail.ru/public/C7av/QCkSiUomz.

Ключевые слова: *поиск образца в строке, алгоритм Ахо-Корасик, Python, дизайн праймеров, компьютерное моделирование, LAMP.*

[*] www.lianab@mail.ru

ВВЕДЕНИЕ

Амплификация нуклеиновых кислот (НК) – это ферментативный процесс, в ходе которого происходит образование фрагментов нуклеотидной последовательности конкретных ДНК или РНК. Амплификация НК используется в фундаментальных исследованиях и на практике при диагностике наследственной патологии и для обнаружения инфекционных агентов, при оценке качества пищевых продуктов, анализе объектов окружающей среды, в ДНК-криминалистике, при установлении родства. Предложены различные методы амплификации НК: полимеразная цепная реакция (ПЦР) [1–3], петлевая изотермическая амплификация (Loop-mediated isothermal AMPlification, LAMP)) [4], амплификация «катящимся кольцом» [5] и др.

Наиболее широко используется ПЦР. Толчок к развитию ПЦР получила благодаря международной программе «Геном человека», для выполнения которой были созданы технологии секвенирования, основанные на применении ПЦР. В настоящее время существуют различные модификации ПЦР, показана возможность создания тест-систем для обнаружения микроорганизмов, выявления точечных мутаций, описаны десятки различных применений метода. Без преувеличения можно сказать, что открытие метода ПЦР стало одним из наиболее выдающихся событий в области молекулярной биологии за последние десятилетия. Это позволило поднять медицинскую диагностику на качественно новый уровень.

Полимеразная цепная реакция – это ферментативное копирование определенного фрагмента ДНК с помощью ДНК-полимеразы, т.е. селективная амплификация ДНК. В основе методики лежит циклическое повторение трех стадий со своими оптимальными температурными режимами и временем: денатурации ДНК (93–95 °C), отжига праймеров (50–65 °C), элонгации (72 °C) копий целевой последовательности. Оптимальное число повторений цикла зависит от начальных условий реакции и подбирается эмпирически.

Границы копируемого фрагмента задаются праймерами. Праймеры – это последовательности нуклеотидов, которые представляют собой искусственно синтезированные короткие одноцепочечные фрагменты ДНК длиной около 20–25 звеньев, служащие в качестве так называемой «затравки» при построении новых цепей.

В ПЦР используются, как правило, два праймера – прямой (forward) и обратный (backward). К 3’-концу праймера ДНК-полимераза присоединяет нуклеотиды, комплементарные одноцепочечной матрице. Затравки подбираются таким образом, что прямой праймер комплементарен участку прямой цепи молекулы ДНК на 5’-конце целевой последовательности, а обратный – участку на 5’-конце обратной цепи этой же последовательности генома.

Схема расположения праймеров для ПЦР представлена на рисунке 1.

Рис. 1. Схема расположения праймеров для ПЦР.

Принципы дизайна ПЦР-праймеров разработаны достаточно хорошо [6]. Праймеры для ПЦР должны отвечать определенным требованиям для обеспечения специфичности процесса амплификации и его эффективности. Специфичность ПЦР во многом зависит от длины праймеров, лучше использовать праймеры длиной от 18 до 25 нуклеотидов.

На специфичность праймеров также влияют праймерные димеры. Они могут образовываться между разными праймерами (прямым и обратным), формируя гетеродимеры. Также любой из праймеров может сформировать гомодимер из-за самокомплементарности. Образование праймерных димеров способно приводить как к ложно-позитивным, так и к ложно-негативным результатам ПЦР.

Длина амплифицируемых фрагментов генома от 100 до 200 п.н. удовлетворяет многим требованиям для ПЦР. Однако при этом надо иметь в виду, что GC-состав ампликонов желательно, чтобы находился в диапазоне от 40 до 60 %, поскольку GC-богатые участки с большей вероятностью могут формировать вторичные структуры, вследствие чего будет снижаться эффективность ПЦР.

Для обеспечения специфичности и эффективности ПЦР еще одним важным критерием является выбор оптимальной температуры отжига праймеров. В случае применения заниженной температуры возрастает вероятность возникновения ложно-позитивной ПЦР, а при завышенной может иметь место плохая наработка целевых ампликонов, вплоть до их непоявления, что грозит возникновением ложно-негативной ПЦР.

Для обеспечения максимальной специфичности ПЦР также требуется, чтобы праймеры были полностью комплементарны местам отжига и не допускали неполного спаривания своих последовательностей с матрицей ДНК или РНК.

Достойной альтернативой ПЦР и второй по масштабам применения методикой является изотермическая петлевая амплификация (LAMP) [7, 8]. Это достаточно простой и чувствительный метод. Главным преимуществом LAMP является его изотермичность (около 63 °C), за счет постоянства температуры вся реакция проходит максимально быстро (в пределах одного часа, а то и меньше).

Успех амплификации во многом зависит от правильно подобранных праймеров. Так же, как для ПЦР, праймеры для LAMP подбираются по правилу комплементарности, праймеры отжигаются на определенных участках ДНК, тем самым способствуя увеличению концентрации целевого участка. Однако отличительной особенностью является количество праймеров, необходимых для метода LAMP.

Базовый набор праймеров для LAMP содержит четыре праймера, которые распознают шесть различных областей в целевой ДНК (за счет удвоенной длины внутренних праймеров) и называются F3 и B3 – внешние праймеры, F2, F1c и B1c, B2 – удвоенные внутренние праймеры. Для того чтобы ускорить процесс наработки нужных участков в реакционную смесь было добавлено еще два праймера, отжигающихся на петлях образуя в процессе амплификации некую гантелеобразную форму участка мишени и потому получивших название петлевые праймеры (Loop primers) – FL и BL [9]. Процесс протекания LAMP нагляднее всего представлены в анимации фирмы EikenChemicalCo [10] и в видео фирмы NewEnglandBiolabs [11].

Схема LAMP представлена на рисунке 2 [12].

Для достижения успешного проведения LAMP и получения наиболее достоверных результатов амплификации необходимо выполнить ряда требований к подбору (дизайну) праймеров:

1. длина: от 15 до 30 нуклеотидов (для внутренних праймеров – удвоенная длина: от 30 до 50 нуклеотидов);
2. GC-состав: от 40 % до 60 %;
3. средний размер целевого участка генома (расстояние от 5'-конца первого внешнего праймера до 5'-конца другого): 120–250 пар оснований;
4. температура плавления Tm. Она зависит от GC-состава и варьируется в пределах от 55 до 65 °C;
5. расстояния между праймерами. Рекомендуемое расстояние: «F3 / F2»: 1–10 нуклеотидов, «F2 / F1c»: 10–25 нуклеотидов, «F1c / B1c»: 0–30 нуклеотидов;

6. вторичная структура (для предотвращения образования ложных налипаний праймеров важно убедиться, что 3'-концы не соединяются ни с какими другими участками праймеров);

7. стабильность 3’-концов (этот критерий варьируется в зависимости от целевой последовательности и условий конструирования праймера).

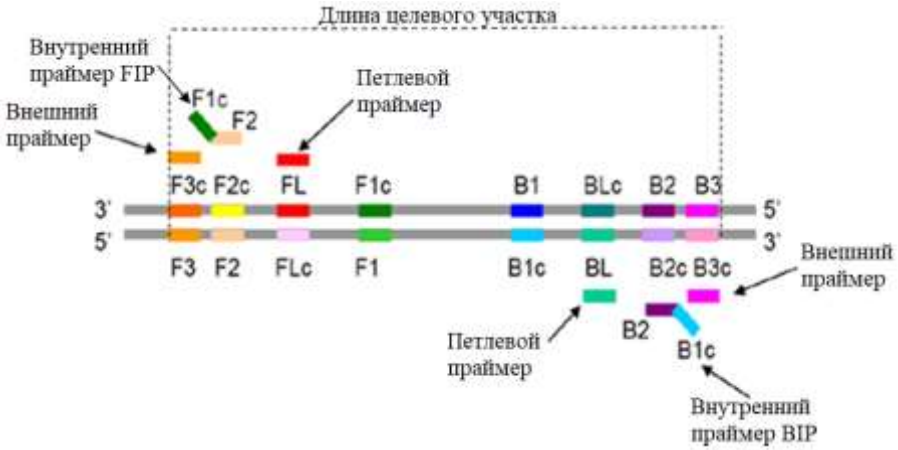

Рис. 2. Схема расположения праймеров для LAMP.

Дизайн праймеров для амплификации целевого участка ДНК является очень важным шагом для эффективного проведения LAMP. Необходимо отметить, что дизайн праймеров для LAMP более сложен, чем для ПЦР. Это связано с тем, что для LAMP необходим набор минимум из четырех праймеров, которые должны отжечься на шести участках целевой ДНК. Или же в расширенном наборе из шести праймеров (два дополнительных петлевых праймера) для достижения более высокой специфичности [13, 14].

Подбор праймеров можно рассматривать как задачу классификации большого набора данных. Например, если длина праймера 20 нуклеотидов, то комбинаций может быть более триллиона (точнее 4^{20}), и проанализировать их вручную практически невозможно.

В одной из немногочисленных отечественных обзорных публикаций по LAMP [15] упоминается, что для этой реакции «праймеры можно подбирать вручную, однако из-за их большого числа это весьма трудоемкая задача».

ОБЗОР СУЩЕСТВУЮЩИХ КОМПЬЮТЕРНЫХ ПРОГРАММ ДЛЯ ПРОВЕДЕНИЯ LAMP

Разработано достаточно большое количество компьютерных программ, позволяющих подобрать праймеры для ПЦР, число которых уже превышает полторы сотни [16]. Многие программы для дизайна существуют в виде бесплатных web-сервисов, ряд программ доступны из коммерческих источников. С накоплением информации о нуклеотидных последовательностях отдельных генов и целых геномов обязательным этапом подбора стала проверка праймеров с помощью программы BLAST [17] на вероятность их неспецифичности, что позволяет в дальнейшем избежать появления ложных результатов реакции. Имеется немало компьютерных программ, оценивающих температуру плавления/отжига праймеров, в том числе работающих в качестве web-сервисов и применяющих для этого различные алгоритмы. На сайтах некоторых фирм имеются специальные страницы, где в режиме on-line можно провести анализ нуклеотидных последовательностей, выбранных в качестве праймеров.

Однако для LAMP существует небольшое число специальных программ для дизайна LAMP–праймеров.

Наиболее широко используемым бесплатным онлайн-инструментом, обеспечивающим дизайн праймеров для LAMP является программа PrimerExplorer разработанная компанией Eiken ChemicalCo. LTD, Токио, Япония [18, 19]. Программа эта достаточно быстрая, проста в использовании. В PrimerExplorer V5 существует три режима подбора праймеров (Automatic Judgment, Normal, User Assignment), а именно создание автоматического, стандартного и определенного наборов праймеров. При создании автоматического набора праймеров в качестве входных данных требуется только анализируемая область генома (до 2000 нуклеотидов), программа автоматически определяет GC-содержание в целевой последовательности, учитывается, что GC-содержание должно быть в пределах от 45 % до 60 %. При стандартном режиме настроены все стандартные параметры по умолчанию. А при определенном (настраиваемом) режиме пользователь сам может установить некоторые параметры необходимые для подбора праймеров. На рисунке 3 представлен скриншот интерфейса программы PrimerExplorer V5.

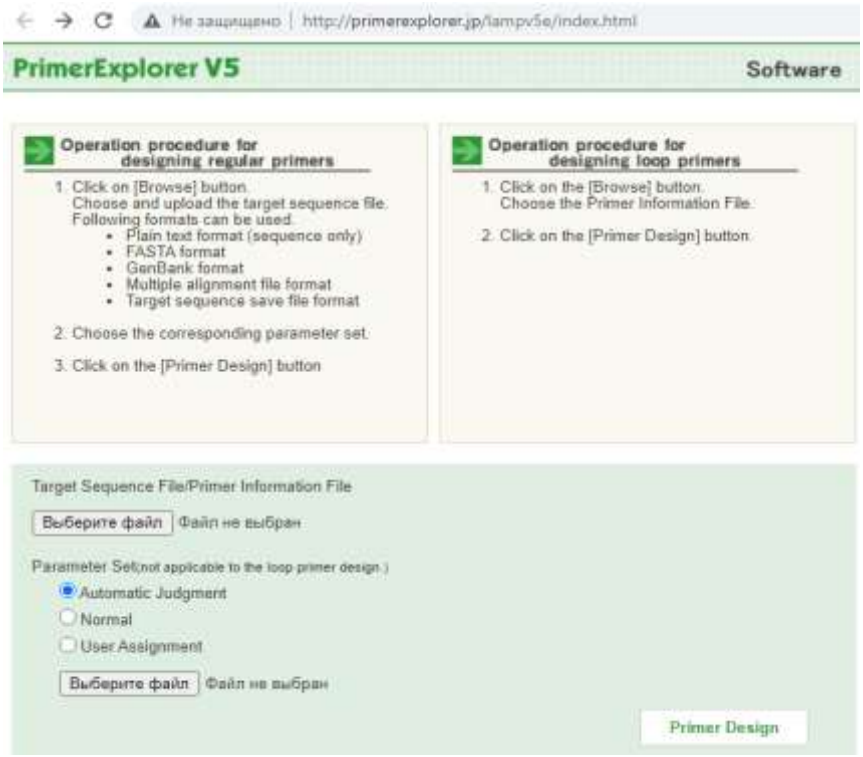

Рис. 3. Начальная страница программы PrimerExplorer V5.

Еще одна программа дизайна праймеров для петлевой амплификации LAMP Designer американской фирмы Premier Biosoft [20], позволяет подбирать наборы праймеров как из четырех праймеров, так и набор из шести (с учетом петлевых) праймеров.

Инструмент, LAMP Designer, автоматически интерпретирует результаты поиска и избегает области, которые могут иметь перекрестную гомологию с базой данных BLAST, для проверки на специфичность. Программа анализирует термодинамические свойства перекрестных взаимодействий между праймерами.

Однако LAMP Designer является коммерческой программой и не обладает должным функционалом в плане необходимости расширять возможности экспериментаторов при дизайне праймерных систем для продвинутой LAMP амплификации.

Существует еще один программный продукт FastPCR [21] финской фирмы Primer Digital Ltd [22]. Это инструментальная среда, которая позволяет разрабатывать праймеры для различных видов ПЦР (стандартной, в реальном времени,

мультиплексной и т.д.) и LAMP (четыре праймера). Программа анализирует последовательности с пересечением GC или AT, процентное содержание GC и GA, генерирует случайные последовательности ДНК и определяет температуру плавления; анализирует ферменты рестрикционного типа I-II-III [23], проводит поиск или создает сайты рестрикции ферментов для анализируемых последовательностей.

Время работы программы пропорционально количеству ампликонов и размеру целевого участка [24].

Интерфейс программы содержит меню, панели инструментов и ленту. Лента предназначена для того, чтобы помочь пользователю быстро найти команды, необходимые для выполнения задачи. На рисунке 4 представлен интерфейс программы. Файлы могут быть сохранены в различных форматах, .rtf , .xls или .txt.

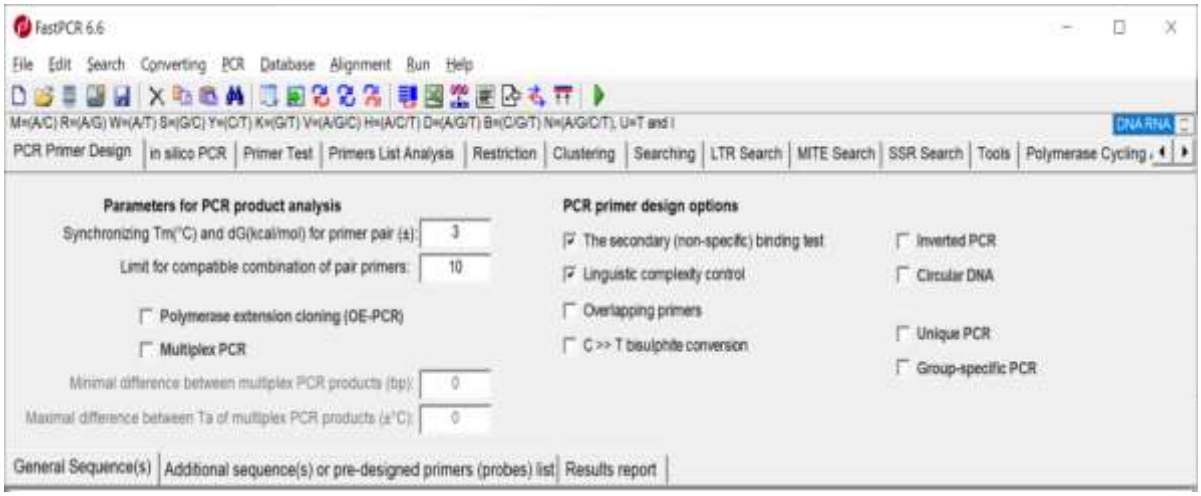

Рис. 4. Интерфейс программы FastPCR.

Программное обеспечение FastPCR может быть установлено исключительно на платформе Microsoft Windows, так же FastPCR является коммерческой программой. На их сайтах можно скачать демо-версию, которая будет активирована в течение семи дней. В течение всего периода предоставляется полный доступ ко всем функциям программы. Однако подобрать праймеры для проведения LAMP в пробной версии FastPCR не удалось.

Программа GLAPD [25] позволяет создавать наборы праймеров LAMP на основе полного генома [26]. Она также может подбирать праймеры LAMP для целевого участка. Программа идентифицирует все возможные участки генома с одним праймером, затем отдельные праймеры объединяются в наборы для LAMP и сопоставляются с целевым участком и полным геномом. После этого выводятся наборы праймеров.

GLAPD работает в три этапа: первый – идентификация возможных участков с одним праймером; второй – объединение отдельных праймеров в набор праймеров LAMP; и третий – проверка набора праймеров LAMP.

GLAPD работает только в операционной системе Linux. Необходимы perl (высокоуровневый интерпретируемый динамический язык программирования общего назначения) и gcc (набор компиляторов для различных языков программирования).

Для запуска программы может потребоваться программное обеспечение Bowtie, который можно загрузить по ссылке: http://bowtie-bio.sourceforge.net/index.shtml и драйвер CUDA, который можно загрузить по другой ссылке: http://www.nvidia.com.

В 2020 году была разработана программа Lamprim [27]. Она позволяет подобрать праймеры LAMP для целевой последовательности. Работает в двух режимах: дизайн праймеров и анализ праймеров. Результатами программы являются наборы праймеров,

информация о расстоянии между праймерами, длине, GC-состав, температуре плавления, стабильности 5’- и 3’-концов. Программа написана на языке Python 3 с использованием библиотеки biopython и поддерживается в операционных системах Windows и Linux.

Компания NewEnglandBiolabs в августе 2020 года выпустила программу NEB LAMP PrimerDesignTool [28, 29]. На рисунке 5 представлен скриншот интерфейса программы.

Рис. 5. Интерфейс программы NEB LAMP PrimerDesignTool.

Она подбирает наборы из четырех основных праймеров, и также есть возможность сгенерировать петлевые праймеры. Длина анализируемой последовательности ограничена: от 100 до 2000 нуклеотидов. Однако есть возможность загрузить более длинные последовательности, для этого необходимо указать начало и конец целевой последовательности. Есть две возможности запустить программу: автоматический (на основе GC-состава) и настраиваемый (с указанием длины праймеров, GC-состава, стабильности праймера, расстояния между праймерами) [30].

Однако программа не учитывает повторы нуклеотидов в одном праймере и зачастую подбирает только один петлевой праймер, что не целесообразно для проведения LAMP. И так же не учитывает гомо- и гетородимерность праймеров в одном наборе, что может привести к ненужным налипаниям и тем самым к ложным результатам.

Очевидно, что подбор праймеров для LAMP представляет собой нетривиальную задачу. Существующие программные решения имеют много ограничений и недоработок. Мы разработали компьютерную программу с удобным интерфейсом, решающую проблему подбора расширенного набора эффективных праймеров с учетом ряда требований, предъявляемых условиями эксперимента.

ПОСТАНОВКА ЗАДАЧИ

Целью данной работы является разработка специальной компьютерной программы, позволяющей подбирать оптимальные наборы праймеров для проведения LAMP. Программа должна учитывать начальные условия подбора праймеров, проводить поиск праймеров используя математические алгоритмы, анализировать соответствие

праймеров рекомендуемым условиям и группировать их по наборам с минимальной разницей температуры плавления праймеров в одном наборе.

Задачами работы являются: поиск праймеров в геноме, сравнение и анализ нуклеотидных последовательностей, сортировка праймеров в наборы по ряду признаков.

АЛГОРИТМЫ ПОИСКА

Задачу поиска праймеров в геноме можно сопоставить с задачей поиска подстроки (образца) в более длинной строке (тексте), где праймер будет являться образцом, а геном текстом. Пусть задан текст A, именуемая образцом (праймер), и более длинная строка T, именуемая текстом (геном). Задача заключается в отыскании всех вхождений образца A в текст T. В случае стандартной LAMP: поиск шести образцов (мест присоединения праймеров) в тексте. В расширенной LAMP: поиск восьми образцов в тексте.

Существует ряд алгоритмов, которые могут быть применены к решению данной задачи. Рассмотрим некоторые из них.

Ранее нами был проведен сравнительный анализ трех классических алгоритмов поиска образца в тексте: прямой поиск, алгоритм Кнута – Морриса – Пратта, алгоритм Бойера – Мура [31]. Все эти алгоритмы применимы для решения задачи поиска образца в тексте, однако было принято решение реализовать алгоритм Бойера-Мура для поиска праймеров в нуклеотидной последовательности для проведения ПЦР, т.к. он является самым быстрым алгоритмом среди известных классических алгоритмов общего назначения для поиска образца в тексте [32].

При дизайне праймеров для LAMP необходимо проводить поиск сразу нескольких образцов. Поэтому мы рассмотрели алгоритмы, которые можно применить для поиска сразу нескольких вхождений. Для этого мы применили алгоритм Рабина – Карпа [33]. Алгоритм работает на основе хеширования. Хеширование, или хеш-функция – это функция, которая преобразует массив данных произвольной длины в битовую строку фиксированной длины (хэш).

Алгоритм выполняет поочередный поиск позиций нахождения праймера в геноме:

1. вычисляется множитель для вычисления скользящего хэша;
2. вычисляется хэш-функция праймера и отрезка генома, который равен длине праймера, с помощью метода Хорнера и модульной арифметики;
3. сравниваются хэши; если они совпадают, то проводится сравнение отрезка генома и праймера из-за возможности ложного срабатывания;
4. с помощью скользящей хэш-функции вычисляется хэш для следующего отрезка генома и снова сравниваются хэши. В случае несовпадения поиск сдвигается на символ правее;
5. если хеши совпали, этот участок проверяется посимвольно.

Так обрабатываются по очереди все позиции генома.

На рисунке 6 представлена наглядная схема поиска образца AGCAGC в строке GTAGCAGCTGAGCAGCAGCAGCTC согласно алгоритму Рабина-Карпа.

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 $O(m \cdot n)$, где n – длина текста, m – длина шаблона, что является одной из причин того, почему данный алгоритм не слишком широко используется [34].

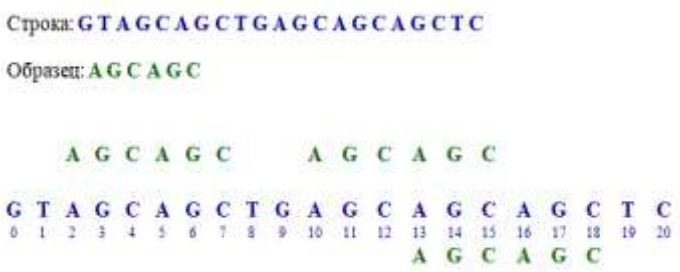

Рис. 6. Схема поиска образца в строке согласно алгоритму Рабина – Карпа.

Мы распараллелили поиск для шести возможных участков в геномах разной длины, запустив два или четырех процесса. Каждому потоку отдается один из шести праймеров.

Результаты ускорения для двух и четырех процессов представлены на рисунке 7.

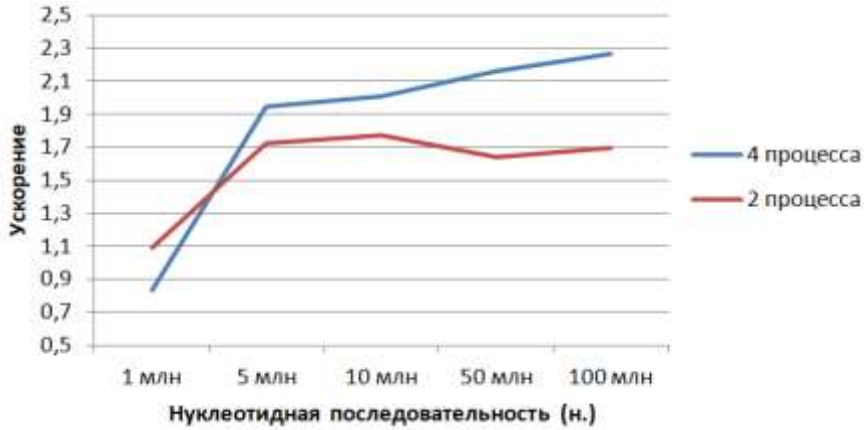

Рис. 7. График зависимости времени поиска от длины нуклеотидной последовательности.

С увеличением мощности процессора увеличивается эффективность многозадачных вычислений. Среднее ускорение на двух процессах составило 1.7, на четырех процессах 2.6. Результаты реализации алгоритма представлены в работах [35, 36].

Для сравнения был рассмотрен ещё один подход, основанный на применении алгоритма, разработанного Альфредом Ахо и Маргарет Корасик в 1975 году. Он позволяет найти сразу все вхождения образцов в тексте [37]. В нем используется конечный автомат, в результате работы 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.

Пример построенного бора для строк: AC, TAC, AGT, ACGT используя алгоритм Ахо-Корасик, представлен на рисунке 8:

1. строится пустая вершина;
2. далее, двигаясь по ребру бора, сравнивается очередная буква с подстрокой, которую нужно найти;
3. если буква на ребре удовлетворяет, то поиск продолжается;
4. если буква не подходит, происходит сдвиг к вершине и повторяется поиск, сдвигаясь на другое ребро бора.

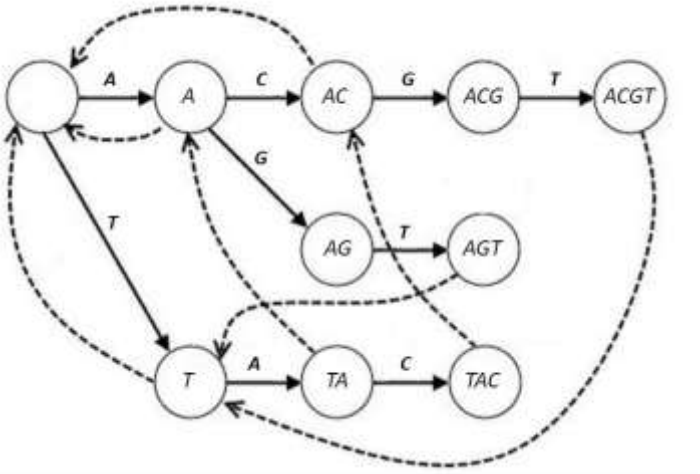

Рис. 8.Схематичное изображение автомата алгоритма Ахо – Корасик для набора строк {AC, TAC, AGT, ACGT}.

Данный алгоритм позволяет искать сразу несколько вхождений, то есть несколько праймеров. 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 [38, 39]. Сравнив два алгоритма, мы отдали предпочтение алгоритму Ахо – Корасик, потому что для хорошего результата алгоритм Рабина – Карпа необходимо правильно выбрать хеш-функцию, а это весьма сложно. В ином случае сложность алгоритма будет значительно выше, чем у алгоритма Ахо – Корасик.

На рисунке 9 представлены графики зависимости времени поиска от длины нуклеотидной последовательности для последовательного поиска и многоядерного поиска. Среднее ускорение составило 1.78.

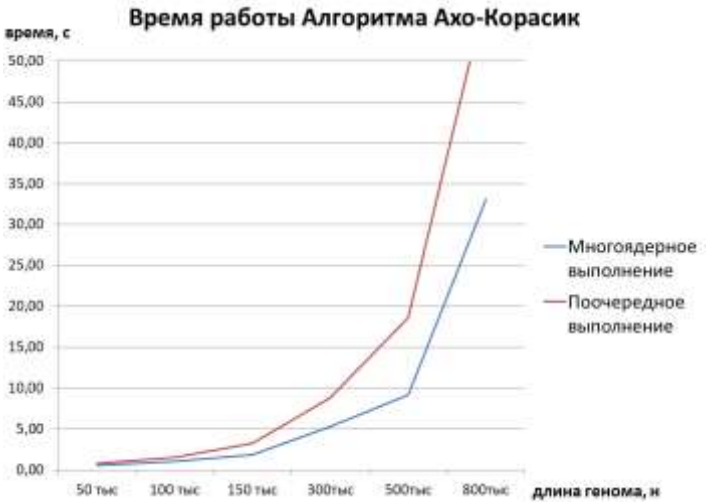

Рис. 9. График зависимости времени от длины нуклеотидной последовательности с применением алгоритма Ахо – Корасик.

РЕАЛИЗАЦИЯ ПРОГРАММЫ

Программа дизайна праймеров для LAMP была реализована на языке программирования Python [40]. Выбор данного языка обусловлен наличием пакета biopython [41], самого большого и популярного пакета утилит для задач биоинформатики.

Общий принцип работы алгоритма представлен на рисунке 10.

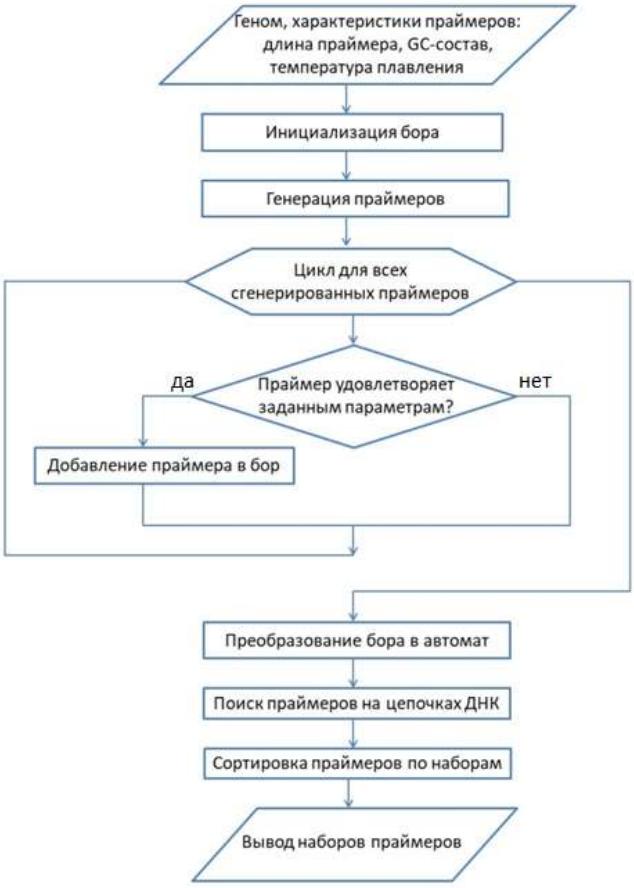

Рис. 10. Блок-схема работы алгоритма для формирования набора праймеров для LAMP.

Для работы алгоритма Ахо – Корасик происходит инициализация дерева (бора) с помощью библиотеки pyahocorasick. Так же метод product из библиотеки itertools генерирует праймеры согласно требованиям, которые указаны на входе, а именно рассчитывается температура плавления с помощью метода calcTm из модуля primer3-py, определяется GC-состав.

Для определения температуры плавления молекул ДНК использовалась формула:

$$Tm = 81.5 + 16.6(\lg[Na^+]) + 0.1 \cdot GCcont - 575/length,$$

где *length* обозначает длину праймера; $[Na^+]$ – молярная концентрация ионов натрия; *GCcont* – процентное содержание нуклеотидов G и C в исследуемом участке. Данная формула была выведена эмпирически, и за ее основу была взята зависимость $Tm = 81.5 + 16.6(\lg[Na^+]) + 0.41 \cdot GCcont - 600/length$, доступная из открытых источников (например, http://biotools.nubic.northwestern.edu/OligoCalc.html) и предложенная ранее для расчета температур отжига олигонуклеотидных праймеров привычной длины (20–25 нт). Указанная формула была модифицирована с учетом необходимости расчета Tm для более протяженных LAMP-праймеров и для обеспечения большего соответствия значений температур отжига праймеров значениям Tm, получаемым с помощью удобной онлайн-утилиты OligoAnalyzer (https://eu.idtdna.com/pages/tools/oligoanalyzer?returnurl=%2Fcalc%2Fanalyzer), обеспечивающей качественный подбор праймеров для различных видов реакций амплификации. Фрагмент кода, в котором рассчитывается температура плавления и процентное содержание нуклеотидов G и C представлен в разделе 1 дополнительных материалов к статье.

Если найденный праймер удовлетворяет заданным параметрам, то он добавляется в бор. Затем бор преобразуется в автомат, происходит поиск полного совпадения праймеров по обеим цепям ДНК.

Далее найденные праймеры, удовлетворяющие заданным параметрам, сортируются в наборы по четыре праймера (два внешних – F3, B3 и два двойных внутренних – FIP (F2, F1c), BIP (B2, B1c)) с минимальной разницей температур плавления и выводятся на экран пользователю.

На рисунке 11 представлен скриншот рабочего окна программы с полученными наборами праймеров, которые сформировались для всего генома коронавируса SARS-CoV-2 [42], длиной 29 844 нуклеотидов. В окнах в верхней части копии экрана можно видеть параметры, которые может задавать пользователь для подбора праймеров на заданном участке ДНК.

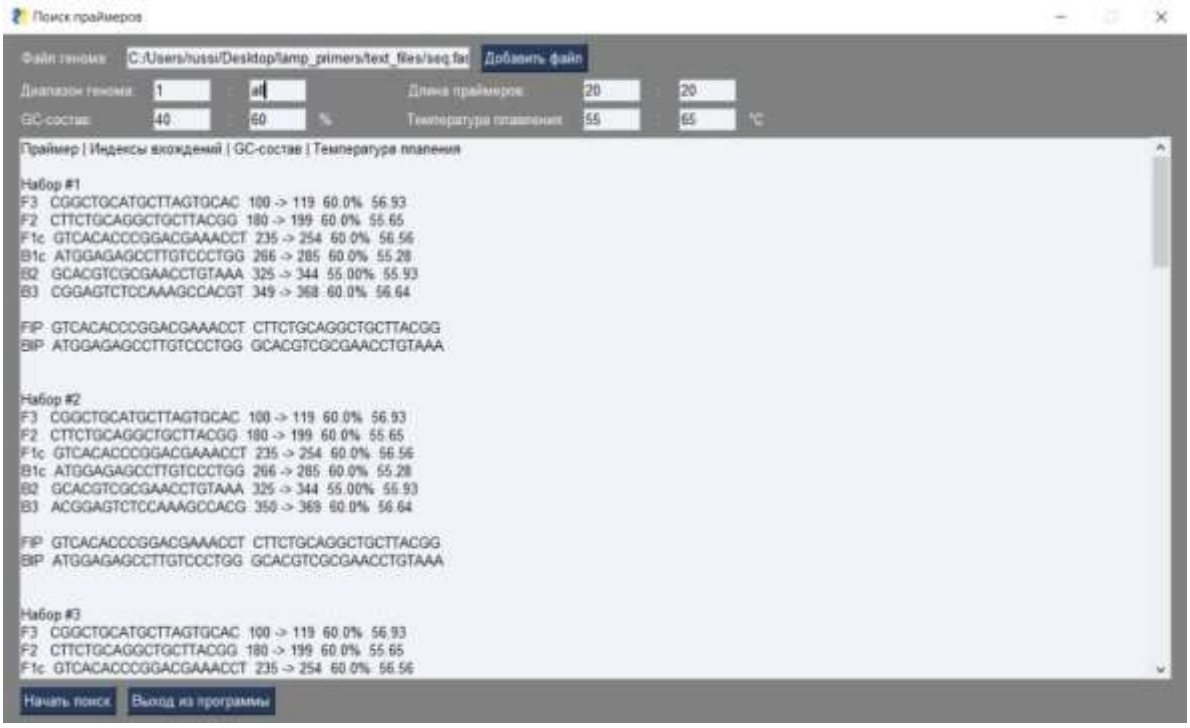

Рис. 11. Пример вывода программы с наборами найденных праймеров.

В окне указаны названия праймеров, например, набор номер 1 состоит из:

F3 CGGCTGCATGCTTAGTGCAC – внешний праймер;

F2 CTTCTGCAGGCTGCTTACGG – часть внутреннего праймера (FIP);

F1c GTCACACCCGGACGAAACCT – часть внутреннего праймера (FIP);

B1c ATGGAGAGCCTTGTCCCTGG – часть внутреннего праймера (BIP);

B2 GCACGTCGCGAACCTGTAAA – часть внутреннего праймера (BIP);

B3 CGGAGTCTCCAAAGCCACGT – внешний праймер;

FIP GTCACACCCGGACGAAACCT CTTCTGCAGGCTGCTTACGG – внутренний праймер;

BIP ATGGAGAGCCTTGTCCCTGG GCACGTCGCGAACCTGTAAA – внутренний праймер.

Для каждого праймера указаны геномные координаты начала и конца, значения GC-состава в процентах и температуры плавления.

ЗАКЛЮЧЕНИЕ

В данной работе приведен обзор наиболее популярных программ дизайна праймеров для LAMP, отмечены их возможности и слабые стороны. Разработан

алгоритм для нахождения нескольких альтернативных наборов праймеров для одной последовательности ДНК. Программная реализация алгоритма на языке программирования Python позволяет осуществлять дизайн эффективных праймеров для LAMP с учетом рекомендуемых и заданных условий подбора праймеров. Пользователь может сам изменять условия подбора праймеров, такие как: GC-состав, длина праймера, температура плавления и диапазон целевого участка генома, в зависимости от объекта исследования. Программа находит праймеры в целевом участке ДНК или полном геноме без ограничений на длину целевой последовательности, в то время как существующие онлайн доступные программы позволяют вводить последовательности любой длины, однако длина анализируемой области должна быть меньше 2000 нуклеотидов.

В результате реализации алгоритма Ахо-Корасик были получены пробные наборы праймеров, сгруппированные в наборы с минимальной разницей температур плавления. Полученные наборы пригодны для проведения натурных экспериментов методом LAMP. Исполняемый файл доступен для скачивания по ссылке: https://cloud.mail.ru/public/C7av/QCkSiUomz, планируется экспериментальная проверка эффективности найденных праймеров для амплификации коронавируса.

Для увеличения точности и достоверности результата планируется дополнить набор параметров алгоритма структурными и термодинамическими характеристиками ДНК, такими, как наличие мононуклеотидных треков, возможность образования праймерами нежелательных вторичных структур (гомо- и гетеродимеров), а также добавить петлевые праймеры, которые будут служить дополнительным идентификатором амплификации искомой последовательности, а также ускорят процесс наработки нужных участков для проведения реакции.

Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований (проект № 20-37-90091).

СПИСОК ЛИТЕРАТУРЫ

1. Mullis K., Faloona F. Specific synthesis of DNA in vitro via a polymerase catalyzed chain reaction. *Meth. Enzymol*. 1987. V.155. P. 335–350.
2. Saiki R.K., Scharf S., Faloona F., Mullis K.B., Horn G.T., Erlich H.A., Arnheim N. Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science*. 1985. V. 230. P. 1350–1354. doi: 10.1126/science.2999980
3. Dietrich D., Uhl B., Sailer V., Holmes E.E., Jung M., Meller S., Kristiansen G. Improved PCR performance using template DNA from formalinfixed and paraffinembedded tissues by overcoming PCR inhibition. *PLoS ONE*. 2013. V. 8. Article No. e77771.
4. Notomi T., Okayam H., Masubuchi H., Yonekawa T., Watanabe K., Amino N., Hase T. Loop-mediated isothermal amplification of DNA. *Nucleic Acids Research*. 2000. V. 28. Article No. e63. doi: 10.1093/nar/28.12.e63
5. Гарафутдинов Р.Р., Сахабутдинова А.Р., Гильванов А.Р., Чемерис А.В. Амлификация нуклеиновых кислот "катящимся кольцом" - универсальный метод анализа широкого круга биологических мишеней. *Биоорг. химия.* 2021. Т. 47. № 6. С. 721–740. doi: 10.31857/S0132342321060075
6. Гарафутдинов Р.Р., Баймиев Ан.Х., Малеев Г.В., Алексеев Я.И., Зубов В.В., Чемерис Д.А., Кирьянова О.Ю., Губайдуллин И.М., Матниязов Р.Т., Сахабутдинова А.Р., Никоноров Ю.М., Кулуев Б.Р., БаймиевАл.Х., Чемерис А.В. Разнообразие праймеров для ПЦР и принципы их подбора. *Биомика*. 2019. Т. 11. № 1. С. 23–70. doi: 10.31301/2221-6197.bmcs.2019-04

7. Mori Y., Notomi T. Loop-mediated isothermal amplification (LAMP): arapid, accurate, and cost-effective diagnostic method for infectious diseases. *Chemother*. 2009. V. 15. P. 62–69. doi: 10.1007/s10156-009-0669-9
8. Chen S.H., Lin C.Y., Cho C.S., Lo C.Z., Hsiung C.A. Primer Design Assistant (PDA): A web-based primer design tool. *Nucleic Acids Res*. 2003. V. 31. P. 3751–3754. doi: 10.1093/nar/gkg560
9. Nagamine K., Hase T., Notomi T. Accelerated reaction by loop-mediated isothermal amplification using loop primers. *Mol Cell Probes*. 2002. V. 16. № 3. P. 223–229. doi: 10.1006/mcpr.2002.0415
10. *Eiken Chemical Co*. URL: http://loopamp.eiken.co.jp/e/lamp/anim.html (дата обращения: 02.06.2020).
11. *Видеофирмы New England Biolabs.* URL: https://international.neb.com/products/m1800-warmstart-colorimetric-lamp-2x-master-mix-dna-rna#Product%20Information (дата обращения: 02.11.2022).
12. Loop-Mediated Isothermal Amplification (LAMP): Primer Design and Assay Optimization. URL: https://youtu.be/GJkvQqDufh0/ (дата обращения: 02.06.2020)
13. Nkere C.K., Oyekanmi J.O., Silva G., Bömer M., Atiri G.I., Onyeka J., Maroya N.G., Seal S.E., Kumar P.L. Chromogenic Detection of Yam Mosaic Virus by Closed-Tube Reverse Transcription Loop-Mediated Isothermal Amplification (CT-RT-LAMP). *Arch. Virol*. 2018. V. 163. № 14. P. 1057–1061. doi: 10.1007/s00705-018-3706-0
14. Wang C., Shen X., Lu J., Zhang L. Development of a Reverse Transcription-Loop-Mediated Isothermal Amplification (RT-LAMP) System for Rapid Detection of HDV Genotype 1. *Lett. Appl. Microbiol*. 2013. V. 56. № 3. P. 229–235. doi: 10.1111/lam.12039
15. Макарова Ю.А., Зотиков А.А., Белякова Г.А., Алексеев Б.Я., Шкурников М.Ю. Изотермическая петлевая амплификация: эффективный метод экспресс-диагностики в онкологии. *Онокоурология*. 2018. Т. 14. № 2. С. 88–99.
16. Чемерис Д.А., Кирьянова О.Ю., Губайдуллин И.М., Чемерис А.В. Дизайн праймеров для полимеразной цепной реакции (краткий обзор компьютерных программ и баз данных). *Биомика*. 2016. Т. 8. № 3. С. 215–238.
17. *Basic Local Alignment Search Tool*. URL: https://blast.ncbi.nlm.nih.gov/Blast.cgi (дата обращения: 02.11.2022).
18. *Primer Explorer*. URL: http://primerexplorer.jp/e (дата обращения: 02.11.2022).
19. *LAMP primer designing software. Primer Explorer*. URL: https://primerexplorer.jp/e/v4_manual/index.html (дата обращения: 02.11.2022).
20. LAMP Designer. Design Primers for Loop Mediated Isothermal Amplification. URL: http://www.premierbiosoft.com/isothermal/lamp.html (дата обращения: 02.11.2022).
21. Kalendar R., Samuilova O., Ivanov K.I. 2017. FastPCR: an in silico tool for fast primer and probe design and advanced sequence analysis. *Genomics*. 2017. V. 109. Article No. 312319. doi: 10.1016/j.ygeno.2017.05.005
22. Kalendar R., Tselykh T.V., Khassenov B., Ramanculov E.M. Introduction on using the FastPCR software and the related Java web tools for PCR and oligonucleotide assembly and analysis in PCR. *Methods in Molecular Biology*. 2017. V. 1620. P. 33–64. doi: 10.1007/978-1-4939-7060-5_2
23. *Фермент рестрикции - Restriction enzyme*. URL: https://dev.abcdef.wiki/wiki/Restriction_enzyme (дата обращения: 02.11.2022).
24. Kalendar R., Lee D., Schulman A. H. Java web tools for PCR, in silico PCR, and oligonucleotide assembly and analysis. *Genomics*. 2011. V. 98. № 2. P. 137–144. doi: 10.1016/j.ygeno.2011.04.009
25. *Программа GLAPD*. URL: https://github.com/jiqingxiaoxi/GLAPD (дата обращения: 02.11.2022).

26. Jia B., Li X., Liu W., Lu C., Lu X., Ma L., Li YY., Wei C. GLAPD: Whole Genome Based LAMP Primer Design for a Set of Target Genomes. *Front. Microbiol*. 2019. V. 10. P. 1–9. doi: 10.3389/fmicb.2019.02860
27. *Maximato/lamprim*. URL: https://github.com/Maximato/lamprim (дата обращения: 02.11.2022).
28. *Isothermal Amplification*. URL: https://lamp.neb.com/#!/ (дата обращения: 02.11.2022).
29. *Loop Mediated Isothermal Amplification (LAMP) Tutorial*. URL: Loop Mediated Isothermal Amplification (LAMP) Tutorial | NEB (дата обращения: 02.11.2022).
30. *NEB LAMP Primer Design Tool*. URL: https://lamp.neb.com/#!/help (дата обращения: 02.11.2022).
31. 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. 285–289.
32. *Боровский А. Алгоритмы поиска в тексте. RSDN Magazine*. URL: https://rsdn.org/article/alg/textsearch.xml (дата обращения: 02.11.2022).
33. Кормен Т., Лейзерсон Ч., Ривест Р., Штайн К. *Алгоритмы: построение и анализ*. М.: Вильямс, 2013. 1328 с.
34. R*abin-Karp Algorithm for Pattern Searching*. URL: https://www.geeksforgeeks.org/rabin-karp-algorithm-for-pattern-searching/ (дата обращения: 02.11.2022).
35. 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. *Короткие статьи и описания плакатов. ПаВТ-2021*. Издательский центр ЮУрГУ. 2021. С. 278. doi: 10.14529/pct2021
36. 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. Article No. 022004. doi: 10.1088/1742-6596/2131/2/022004
37. Aho A.V., Corasick M.J. Efficient string matching: An aid to bibliographic search. *Commun. ACM*. 1975. V. 18. № 6. P. 333–340.
38. *Алгоритм Ахо-Корасик.* URL: https://e-maxx.ru/algo/aho_corasick (дата обращения: 02.11.2022).
39. *Алгоритм Ахо-Корасик.* URL: https://neerc.ifmo.ru/wiki/index.php?title=Алгоритм_Ахо-Корасик (дата обращения: 02.11.2022).
40. Федоров Д.Ю. *Программирование на языке высокого уровня Python*: учебное пособие для прикладного бакалавриата. 2-е издание. М.: Издательство Юрайт, 2019. 161 с.
41. *Biopython*. URL: https://biopython.org/ (дата обращения: 02.11.2022).
42. *Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome*. URL: https://www.ncbi.nlm.nih.gov/nuccore/MN908947 (дата обращения: 02.11.2022).

Рукопись поступила в редакцию 26.06.2022, переработанный вариант поступил 03.10.2022.
Дата опубликования 14.11.2022.

Application of the Aho-Korasik Algorithm for the Selection of Primers for Loop Isothermal Amplification

Akhmetzianova L.U.[1,2], Davletkulov T.M.[1], Garafutdinov R.R.[3], Gubaydullin I.M.[1,2]

[1]Ufa State Petroleum Technical University, Ufa
[2]Institute of Petrochemistry and Catalysis of Russian Academy of Sciences, Ufa
[3]Institute of Biochemistry and Genetics of Russian Academy of Sciences, Ufa

Abstract. This paper presents a program which allows user to do primer design for identifying DNA target site or a whole genome with a goal of performing loop-mediated isothermal amplification. The review of the most popular existing primer design programs for LAMP is carried out.

Recommended conditions are presented in the paper. They are required to be taken in consideration during the process of primer design for loop-mediated isothermal amplification. These are the conditions: primer's length, GC-content, amplicon average size, annealing temperature and distance between primers.

A search for primer positions in genome is needed since loop-mediated isothermal amplification requires primer kits that consist of 6 primers in order for primer design to be done. The Aho–Corasick algorithm was proposed for a search implementation. This algorithm is capable of simultaneous search for a number of sample (primer) entries in a longer sequence (a fragment or a whole genome).

This software allows the search for primers in genomes of various length and it groups primers by kits, which in turn could be applied in laboratory experiments. These kits are formed according both to the recommended conditions of primer selection for performing loop-mediated isothermal amplification and to the initial conditions, which are determined by the user before the process. After that, the user may choose the best option for their case from a list of primer kits that are being created as a result of performed computer analysis. The test run of the program was done during the search for a specific primer kit that is meant to be used for performing loop-mediated isothermal amplification of genome with a goal of detection of novel coronavirus infection SARS-CoV-2, a virus that triggers a dangerous disease, COVID-19.

The software was developed using Python with BioPython and Pyahocorasick libraries and available at the link: https://cloud.mail.ru/public/C7av/QCkSiUomz.

Key words*: search for a pattern in a string, Aho-Korasik algorithm, Python, primers design, computer modeling, LAMP.*