АЛГОРИТМ СЕЛЕКЦИИ СПЕЦИФИЧНЫХ ЗОНДОВ ДЛЯ ВЫЯВЛЕНИЯ АКТУАЛЬНЫХ ВОЗБУДИТЕЛЕЙ ЗАБОЛЕВАНИЙ ЧЕЛОВЕКА С ПОМОЩЬЮ ТЕХНОЛОГИИ ДНК-БИОЧИПОВ

DOI: 10.17691/stm2022.14.1.01
УДК 616.9:577.21
Поступила 7.12.2021 г.

© **Е.Н. Филатова,** к.б.н., ведущий научный сотрудник лаборатории молекулярной биологии и биотехнологии[1];
А.С. Чайкина, студентка[2];
Н.Ф. Бруснигина, к.м.н., доцент, зав. лабораторией метагеномики и молекулярной индикации патогенов[1];
М.А. Махова, к.б.н., старший научный сотрудник лаборатории метагеномики и молекулярной индикации патогенов[1];
О.В. Уткин, к.б.н., зав. лабораторией молекулярной биологии и биотехнологии[1]

[1]Нижегородский научно-исследовательский институт эпидемиологии и микробиологии им. академика И.Н. Блохиной Роспотребнадзора, Н. Новгород, ул. Малая Ямская, 71, 603950;
[2]Приволжский исследовательский медицинский университет, пл. Минина и Пожарского, 10/1, Н. Новгород, 603005

Цель исследования — разработка алгоритма селекции дискриминирующих зондов для определения актуальных возбудителей инфекционных заболеваний человека.

Материалы и методы. Алгоритм подбора зондов, предназначенных для выявления широкого спектра патогенов, был реализован в виде программы для ЭВМ disprose (DIScrimination PRObe SElection), написанной на языке программирования R. Для осуществления некоторых функций программы использовали стороннее программное обеспечение: семейство программ BLAST+ и ViennaRNA Package. Алгоритм тестировали, подбирая специфичные зонды для детекции атипичного возбудителя внебольничной пневмонии (ВП) *Chlamydophila (Chlamydia) pneumoniae*. Нуклеотидные последовательности для анализа загружали из банка данных NCBI.

Результаты. Разработан алгоритм селекции специфичных зондов для детекции актуальных возбудителей инфекционных заболеваний человека. Алгоритм реализован в виде модульной программы disprose, позволяющей осуществлять все этапы подбора зондов: загрузку нуклеотидных последовательностей и их метаданных из доступных банков, создание локальных баз данных, формирование пула зондов, расчет их физико-химических параметров, выравнивание зондов и последовательностей, содержащихся в локальных базах, обработку и оценку результатов выравниваний. Проведен анализ производительности модулей алгоритма и определена оптимальная последовательность их выполнения для оптимизации скорости работы. Алгоритм апробирован на примере подбора зондов для специфичной детекции одного из возбудителей ВП — *Chlamydophila pneumoniae*. Рассчитанные *in silico* параметры специфичности подобранных зондов свидетельствуют о низком риске их неспецифичного связывания и возможности применения в целом ряде молекулярно-генетических инструментов диагностики (ДНК-биочипы, ПЦР).

Заключение. Разработан и реализован в виде модульной программы disprose алгоритм для селекции специфичных зондов, детектирующих широкий спектр актуальных возбудителей заболеваний человека бактериальной и вирусной природы в клиническом биоматериале. Подобранные с использованием программы зонды способны составлять функциональную основу ДНК-ориентированных биочипов, предназначенных для дифференциальной диагностики возбудителей полиэтиологических заболеваний, таких, например, как ВП. Вследствие гибкости и открытости программы область ее применения может быть расширена.

Для контактов: Филатова Елена Николаевна, e-mail: filatova@nniiem.ru

Ключевые слова: алгоритм подбора зондов; ДНК-биочип; дизайн ДНК-биочипа; внебольничная пневмония; *Chlamydophila pneumoniae.*

Как цитировать: Filatova E.N., Chaikina A.S., Brusnigina N.F., Makhova M.A., Utkin O.V. An algorithm for the selection of probes for specific detection of human disease pathogens using the DNA microarray technology. Sovremennye tehnologii v medicine 2022; 14(1): 6, https://doi.org/10.17691/stm2022.14.1.01

English

An Algorithm for the Selection of Probes for Specific Detection of Human Disease Pathogens Using the DNA Microarray Technology

E.N. Filatova, PhD, Leading Researcher, Laboratory of Molecular Biology and Biotechnology[1];
A.S. Chaikina, Student[2];
N.F. Brusnigina, MD, PhD, Associate Professor, Head of the Laboratory for Metagenomics and Molecular Indication of Pathogens[1];
M.A. Makhova, PhD, Senior Researcher, Laboratory for Metagenomics and Molecular Indication of Pathogens[1];
O.V. Utkin, PhD, Head of the Laboratory of Molecular Biology and Biotechnology[1]

[1]Blokhina Scientific Research Institute of Epidemiology and Microbiology of Nizhny Novgorod, Federal Service for Surveillance on Consumer Rights Protection and Human Wellbeing (Rospotrebnadzor), 71 Malaya Yamskaya St., Nizhny Novgorod, 603950, Russia;
[2]Privolzhsky Research Medical University, 10/1 Minin and Pozharsky Square, Nizhny Novgorod, 603005, Russia

The aim of the study was to develop an algorithm for the selection of discriminating probes to identify a wide range of causative agents of human infectious diseases.

Materials and Methods. The algorithm for selecting the probes was implemented in the form of the disprose (DIScrimination PRObe SElection) computer program written in the R language. Additionally, third-party software was used: the BLAST+ and ViennaRNA Package programs. The developed algorithm was tested by selecting specific probes for detecting *Chlamydophila (Chlamydia) pneumoniae* — an atypical bacterial pathogen causing community-acquired pneumonia (CAP). Nucleotide sequences for analysis were downloaded from the NCBI databank.

Results. An algorithm for the selection of specific probes capable of detecting human infectious pathogens has been developed. The algorithm is implemented in the form of the disprose modular program, which allows for performing all stages of the probe selection process: loading the nucleotide sequences and their metadata from available databanks, creating local databases, forming a pool of probes, calculating their physicochemical parameters, aligning the probes and sequences contained in local databases, processing and evaluating the alignment results. The algorithm was successfully tested and its performance was confirmed by selecting a set of probes for the specific detection of *Chlamydophila pneumoniae*. The specificity of the selected probes calculated *in silico* indicated a low risk of their nonspecific binding and a high potential of using them as molecular genetic diagnostic tools (DNA microarrays, PCR).

Conclusion. An algorithm for the selection of specific probes detecting a wide range of human pathogens in clinical biomaterial has been developed and implemented in the form of the disprose modular program. The probes selected using this program can serve as the functional basis of DNA-oriented microarrays able to identify causative agents of polyetiological diseases, such as CAP. Due to the flexibility and openness of the program, the scope of its application can be expanded.

Key words: probe selection algorithm; DNA microarray; DNA microarray design; community-acquired pneumonia; *Chlamydophila pneumoniae.*

Введение

С учетом сложившейся эпидемической ситуации на сегодняшний день резко возросла потребность в диагностических ДНК-биочипах, предназначенных для детекции бактериальных и вирусных патогенов.

Эффективность диагностического ДНК-биочипа обусловлена качеством выбора специфичных олигонуклеотидных зондов. Трудности их подбора обусловлены сложностью химического и видового составов клинических образцов, что является причиной высокого риска кросс-гибридизации зондов с нецелевыми ДНК, и появлением ложноположительных результатов [1].

Существующие алгоритмы подбора зондов используют различные способы оценки соответствия кандидатных последовательностей критериям специфичности и гомогенности [2–5]. Как правило, программное обеспечение, реализующее известные алгоритмы, не позволяет пользователю модифицировать способ расчета этих критериев, порядок их применения и/или параметры. При этом в зависимости от спектра

детектируемых патогенов, их таксономического разнообразия, а также характеристик биологического образца и целей диагностики (выявление факторов патогенности, определение антибиотикорезистентности и др.) вес критериев и целесообразность их применения могут варьировать. Дополнительно усложнение процедуры отбора зондов неизбежно приводит к увеличению затрачиваемого на работу времени, а также повышению требований к вычислительному оборудованию. В связи с этим для селекции специфичных зондов, составляющих основу диагностических ДНК-биочипов, предпочтительными являются алгоритмы, реализованные в виде модульных программ, допускающих модификации.

Целью исследования явилась разработка алгоритма селекции дискриминирующих зондов для определения широкого спектра возбудителей инфекционных заболеваний человека.

Материалы и методы

Для выполнения цели исследования решали следующие задачи: разработать алгоритм подбора специфичных зондов; реализовать его в виде программы для ЭВМ; оптимизировать производительность алгоритма; протестировать его работу путем подбора зондов для детекции *Chlamydophila (Chlamydia) pneumoniae.*

Алгоритм подбора зондов, предназначенных для массовой дифференциальной детекции бактериальных и вирусных возбудителей заболеваний человека, реализован в виде программы для ЭВМ disprose, написанной на языке программирования R и оформленной в виде пакета функций. Пакет распространяется по лицензии GNU GPL-3 (2007 г.) и доступен для загрузки из официального международного репозитория «The Comprehensive R Archive Network» (https://CRAN.R-project.org/package=disprose).

Расчеты проводили на рабочей станции Intel Xeon 2560 (x2), 128 GB RAM. Алгоритм тестировали, выполняя поиск зондов, позволяющих специфично детектировать «атипичный» возбудитель внебольничной пневмонии (ВП) *C. pneumoniae*. Генетические последовательности *C. pneumoniae* были получены из базы данных NCBI Nucleotide [6].

Расчет минимальной энергии фолдинга (МЭФ) олигонуклеотидных последовательностей кандидатных зондов выполняли с применением набора программ ViennaRNA Package (версия 2.4.14) [7, 8]. Расчет температуры плавления ($T_{пл}$) проводили на основании установленного набора термодинамических параметров [9] методом ближайших соседей [10].

Локальное выравнивание нуклеотидных последовательностей осуществляли с применением программы blastn из пакета программ BLAST+ (версия 2.10.0) [11]. Поиск соответствий выполняли в полноразмерной загружаемой базе данных NCBI Nucleotide collection, а также в локальных базах нуклеотидных последовательностей, сформированных с помощью программы blastdbcmd.

В данной работе не использована информация, нарушающая чью-либо конфиденциальность. Исследование выполнено без участия людей и животных.

Результаты

Алгоритм подбора дискриминирующих зондов и его реализация в программе disprose

Отбор целевых нуклеотидных последовательностей. Перед запуском процедуры отбора зондов необходимо определить, какие нуклеотидные последовательности являются целевыми (т.е. такими, с которыми зонды должны эффективно гибридизоваться), а какие — неспецифичными (гибридизация зондов с которыми нежелательна).

Целевые последовательности могут быть получены исследователем самостоятельно или быть загружены из доступных банков данных. На сегодняшний день в программе disprose реализованы функции загрузки как самих последовательностей, так и их метаданных из нескольких крупных банков: NCBI (базы данных Nucleotide, GenBank, RefSeq) и GISAID. На основании получаемых метаданных исследователь может отобрать из всего набора имеющихся в наличии последовательностей только те, которые представляют для него интерес. Отобранные нуклеотидные последовательности составляют локальную целевую базу, предназначенную для тестирования способности кандидатных зондов гибридизоваться с целевыми последовательностями.

Отбор неспецифичных последовательностей. На сегодняшний день загружаемая база NCBI Nucleotide collection содержит более 71 млн. последовательностей объемом свыше 466 гигабаз (Гб). Использование такой базы для тестирования способности зондов к неспецифической гибридизации требует слишком больших временны́х затрат. Для ускорения работы алгоритма по отбору зондов целесообразным является сокращение объема неспецифичной базы за счет включения в нее только таких последовательностей, которые гипотетически могут присутствовать в тестируемых образцах.

Для детектируемых нами возбудителей ВП и используемого с этой целью биологического материала (мазки со слизистой оболочки носоглотки, ротоглотки, мокрота и т.д.) по данным литературы нами установлены более 40 таксонов микроорганизмов, генетический материал представителей которых может встречаться в анализируемых образцах. Полный перечень таксонов представлен в приложении 1 (https://drive.google.com/file/d/1JpRQzi7gJ8lzY1o6TMmULSBJvfiO20h/view). Для каждого таксона с помощью программы disprose из базы данных NCBI Nucleotide collection были выгружены все ассоциированные с ним последовательности, которые сформи-

ровали используемую нами далее неспецифичную базу (8,7 млн. последовательностей объемом 165 Гб). Уменьшение объема базы позволило сократить время, затрачиваемое в дальнейшем на тестирование специфичности зондов, в среднем в 5 раз (рис. 1).

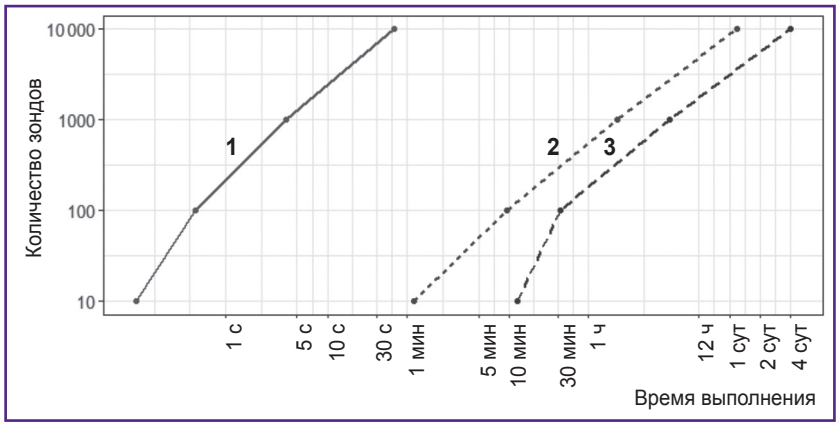

Рис. 1. Зависимость времени, затрачиваемого на процедуру выравнивания последовательностей алгоритмом BLAST, от количества зондов и объема локальной базы

Проводили выравнивание указанного количества зондов, предназначенных для выявления последовательностей *C. pneumoniae*, и содержимого баз разных объемов: *1* — база целевых последовательностей, 0,02 Гб; *2* — база неспецифичных последовательностей, 165 Гб; *3* — база неспецифичных последовательностей, 466 Гб. Указано время, затрачиваемое на процедуру выравнивания без обработки результата. Использован логарифмический масштаб осей

Формирование пула кандидатных зондов. Пул кандидатных зондов формируется путем виртуального нарезания выбранной пользователем «материнской» последовательности на участки установленной длины. Последовательность для нарезки определяется пользователем (частым выбором служит референсный геном патогена) и может быть представлена в виде FASTA-файла, полученного из любого источника, либо загружена непосредственно из банка NCBI.

Тестирование гомогенности — проверка физико-химических свойств зондов. Поскольку все зонды ДНК-биочипа должны взаимодействовать с целевыми последовательностями в одинаковых условиях (например, при одной температуре гибридизации), важным этапом является определение физико-химических свойств кандидатных зондов. В программе disprose реализована проверка четырех физико-химических параметров, позволяющих контролировать условия гибридизации и стабильность вторичной структуры зондов: процентного содержания нуклеотидов гуанина и цитозина (GC), количества гомоповторов, расчетной температуры плавления ($T_{пл}$) и МЭФ (см. таблицу). При этом программа предоставляет возможность изменять параметры расчета.

Тестирование специфичности зондов. Алгоритм предполагает двухшаговое тестирование специфичности зондов. На первом шаге проводится проверка способности зондов к целевой гибридизации путем выравнивания их и отобранных ранее последовательностей патогена алгоритмом BLAST. Далее осуществляют обработку результатов с помощью специализированных функций пакета disprose. В процессе обработки для каждого зонда подсчитывают число целевых последовательностей, с которыми он был выровнен в соответствии с требуемыми характеристиками (минимальные длина

Тестируемые физико-химические параметры зондов

Параметр	Влияние на характеристики зонда	Значения по умолчанию	Литературные источники
Размер зонда	С возрастанием длины зонда снижается дискриминационный потенциал, но повышается эффективность и уровень сигнала гибридизации	24–32 н.о.	[1, 12, 13]
Содержание GC	Влияет на температуру плавления: низкое содержание GC снижает эффективность гибридизации, высокое — повышает риск неспецифичной гибридизации	40–60%	[2, 14–17]
Количество гомоповторов (повторяющихся нуклеотидов подряд)	Более четырех одинаковых нуклеотидов подряд повышает вероятность неспецифичного связывания	<5 н.о.	[18]
Минимальная энергия фолдинга	Чем ниже показатель, тем выше вероятность формирования зондом стабильных вторичных структур, снижения его чувствительности и эффективности гибридизации	≥–3 ккал/моль	[7, 8, 13, 19]
Температура плавления	Основное условие реакции гибридизации определяет характеристики буферных растворов. Температура гибридизации примерно на 5° ниже температуры плавления	55–60°C	[10, 13, 16]

выравнивания и процент покрытия, значения показателей score и E-value и др.).

На втором шаге зонды тестируют на специфичность путем их выравнивания на содержимое неспецифичной базы.

В итоге для каждого зонда получают список целевых и неспецифичных последовательностей, с которыми он потенциально взаимодействует. Анализ как количества, так и состава специфичных и неспецифичных взаимодействий с помощью функций программы disprose позволит отобрать зонды с контролируемыми *in silico* параметрами специфичности.

Заключительный этап анализа. В том случае, если отобранные зонды не будут специфичными в отношении всех целевых последовательностей, цикл анализа запускают заново. При этом неохваченные последовательности образуют новый целевой банк меньшего объема, а пул кандидатных зондов формируется из новой «материнской» последовательности. Циклы анализа повторяют до тех пор, пока для каждой из целевых последовательностей патогена не будут подобраны зонды, способные эффективно с ней гибридизоваться.

Таким образом, предлагаемый нами алгоритм подбора дискриминирующих зондов состоит из трех основных этапов, выполняемых последовательно:

1) определение перечней целевых и неспецифичных последовательностей, формирование локальных банков последовательностей;

2) генерация пула кандидатных зондов, проверка их физико-химических параметров;

3) проверка способности кандидатных зондов гибридизоваться с целевыми и неспецифичными последовательностями.

Функции программы disprose, обеспечивающие реализацию алгоритма, и их краткие характеристики представлены в приложении 2 (https://drive.google.com/file/d/1JpRQzi7gJ8Iz-Y1o6TMmULSBJvfiO20h/view).

Дополнительные возможности алгоритма и его оптимизация

Помимо основного алгоритма селекции специфичных зондов в программе disprose реализованы дополнительные функции: добавление нуклеотидных адаптеров к последовательностям зондов и аннотирование участков генома патогена, взаимодействующих с зондами. Обеспечена возможность использования последовательностей проектов полногеномного секвенирования — WGS (whole genome shotgun). Для успешной реализации также проведена оптимизация быстродействия алгоритма.

Работа с WGS-проектами. Проекты WGS представляют собой неполные сборки геномов или хромосом прокариот и эукариот, состоящие из набора контигов. Включение таких последовательностей в перечень целевых последовательностей затруднено, поскольку каждый контиг при выравнивании алгоритмом BLAST рассматривается как самостоятельная единица. При этом возникает необходимость подбора специфичного зонда к каждому контигу, что приводит к подбору избыточного количества зондов и, следовательно, к увеличению временны́х затрат в десятки раз.

Работа предлагаемого алгоритма с последовательностями проектов WGS возможна благодаря использованию специализированных функций программы disprose, позволяющих рассматривать все контиги одного генома как одну виртуальную последовательность. В этом случае зонды, подобранные к одному из контигов, расцениваются как специфичные в отношении целого генома WGS-проекта.

Оптимизация производительности алгоритма. В связи с тем, что подбор специфичных зондов предполагает оценку нескольких миллионов кандидатных последовательностей, при разработке алгоритма одной из основных задач служило обеспечение его высокой производительности. В программе disprose используются стандартные технические приемы увеличения производительности алгоритма, такие как хранение промежуточных данных в базе SQL и запуск большинства функций в параллельном режиме (применимо для многоядерной конфигурации вычислительного сервера). Однако главным аспектом, определяющим скорость расчетов, является порядок применения функций программы.

Как показано на рис. 2 и 3, операции алгоритма различаются производительностью, при этом максимальные временны́е затраты приходятся на выравнивание зондов и содержимого локальных банков при проверке специфичности. Именно поэтому мы рекомендуем тестировать специфичность зондов на самом последнем этапе, когда бóльшая их часть уже исключена из списка кандидатных.

Сравнительно медленными являются и процедуры определения физико-химических показателей зондов. Расчет каждого показателя выполняется с разной скоростью (см. рис. 2). Этапы расчета процентного содержания GC и гомоповторов представляются наиболее производительными и позволяют сразу исключать зонды с заведомо неудовлетворительными характеристиками. Эти этапы должны выполняться первыми. Так, при подборе зондов для детекции *C. pneumoniae* последовательное исключение кандидатных зондов, не удовлетворяющих условиям содержания GC, наличия гомоповторов и величине МЭФ, позволило к началу этапа расчета $T_{пл}$ (наиболее низкопроизводительная из всех процедура) сократить количество кандидатных зондов с 11,0 до 2,6 млн. Это уменьшило время, затрачиваемое на расчет $T_{пл}$, в 3,7 раза (20,1 вместо 75,2 мин).

Поиск специфичных зондов для детекции Chlamydophila pneumoniae. Предложенный алгоритм и программа disprose использовались нами для поиска специфичных зондов, позволяющих детектировать

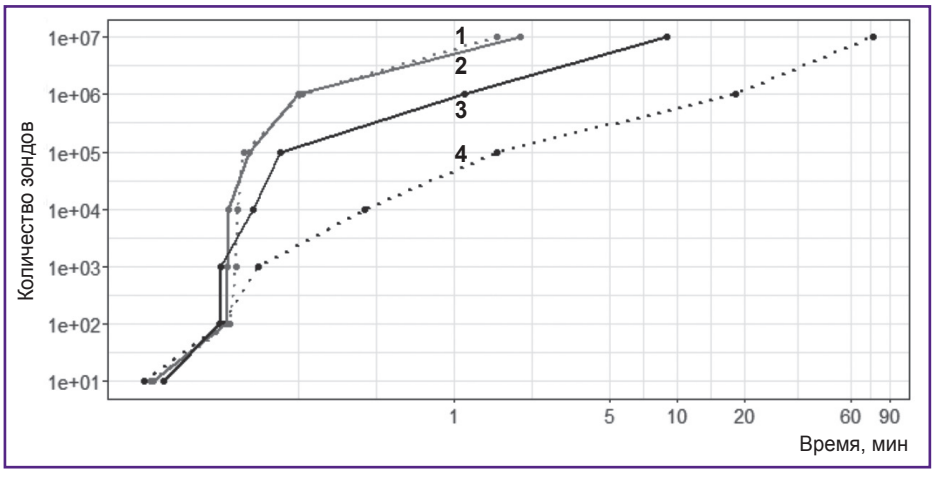

Рис. 2. Зависимость затрачиваемого времени от количества зондов и характера выполняемого расчета
Рассчитывали физико-химические свойства зондов для детекции последовательностей *C. pneumoniae*: *1* — проверка наличия гомоповторов; *2* — расчет процентного содержания GC; *3* — расчет минимальной энергии фолдинга; *4* — расчет температуры плавления. Использован логарифмический масштаб осей

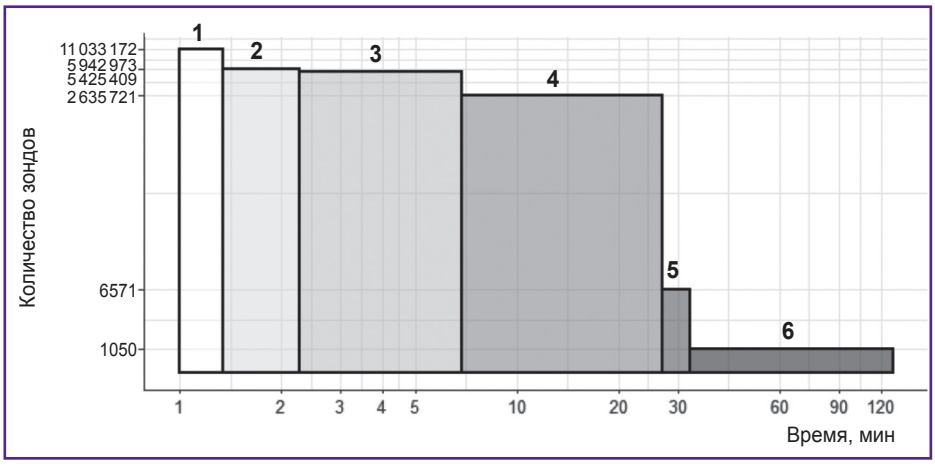

Рис. 3. Время, затрачиваемое на подбор зондов для детекции последовательностей *C. pneumoniae*
Отображены этапы расчета: *1* — расчет процентного содержания GC; *2* — проверка наличия нуклеотидных гомоповторов; *3* — расчет минимальной энергии фолдинга; *4* — расчет температуры плавления; *5* — проверка гибридизации, выравнивание последовательностей зондов на базу целевых последовательностей; *6* — проверка специфичности, выравнивание последовательностей зондов на базу неспецифичных последовательностей. При выполнении процедур количество анализируемых зондов последовательно уменьшалось за счет выбраковки зондов с неудовлетворительными характеристиками. Для процедур выравнивания указано время, затрачиваемое на саму процедуру и обработку результата. Использован логарифмический масштаб осей

C. pneumoniae. Для создания перечней специфичных последовательностей из базы данных NCBI Nucleotide collection были выгружены метаданные последовательностей, доступных по поисковому запросу «Chlamydia pneumoniae * [organism] OR Chlamydophila pneumoniae * [organism]». Всего в набор метаданных вошли 9062 записи. В результате анализа в качестве целевых были отобраны 17 полногеномных последовательностей и 165 последовательностей WGS-проекта. Объем целевой базы составил 0,02 Гб. Перечень целевых последовательностей представлен в приложении 3 (https://drive.google.com/file/d/1JpRQzi7gJ8lz-Y1o6TMmULSB-JvfiO20h/view). Нецелевую базу составили отобранные ранее последовательности, ассоциированные с генетическим материалом человека, представителями его нормофлоры и возбудителей заболеваний.

В качестве «материнской» последовательности для формирования пула кандидатных зондов была выбрана последовательность «Chlamydia pneumoniae TW-183, complete sequence» (идентификационный номер NC_005043 NCBI RefSeq). «Материнская» последовательность была нарезана на все возможные

отрезки длиной от 24 до 32 нуклеотидных оснований (н.о.), которые составили пул кандидатных зондов (11 033 172 зонда).

При проверке физико-химических свойств условиями отбора зондов стали следующие: содержание G и C в диапазоне 40–60%; отсутствие гомоповторов длиной 5 н.о. и более; МЭФ не ниже 0 ккал/моль. В конце рассчитывали $T_{пл}$ зондов. Большинство зондов обладали расчетной $T_{пл}$, приближенной к 57°C, поэтому для ускорения процедуры отбора количество зондов было сокращено с сохранением таких из них, расчетная $T_{пл}$ которых находилась в пределах 56,97–57,03°C. Итоговое количество кандидатных зондов, отобранных для следующего этапа, составило 6571.

Для тестирования способности кандидатных зондов гибридизоваться с целевыми последовательностями проводили их выравнивание алгоритмом BLAST с целевой и неспецифической базами. Для получения максимально специфичных зондов мы установили следующие условия эффективной гибридизации: для гибридизации с целевыми последовательностями идентичность должна составлять 100% при отсутствии точечных несовпадений и пропусков нуклеотидов; для гибридизации с неспецифичными последовательностями идентичность должна составлять 50% и более.

Из кандидатного пула 6380 зондов эффективно взаимодействовали со всеми целевыми последовательностями *in silico.* Для сокращения временны́х затрат на тестирование возможности неспецифической гибридизации количество кандидатных зондов было уменьшено до 1050 путем выбора еще более узкого диапазона допустимых значений расчетной $T_{пл}$ (56,994–57,006°C). Зонды, которые эффективно гибридизовались хотя бы с одной неспецифичной последовательностью, в итоговый пул не вошли.

В результате работы алгоритма отобраны 100 специфичных дискриминирующих зондов, позволяющих осуществлять дифференциальную детекцию *C. pneumoniae* среди других патогенов. Ввиду высокой специфичности отобранные зонды могут использоваться в качестве функциональной основы ДНК-биочипов, предназначенных для выявления актуальных возбудителей ВП. Суммарное время подбора зондов с помощью программы disprose составило 130 мин. Перечень отобранных зондов и их характеристики представлены в приложении 4 (https://drive.google.com/file/d/1JpRQzi7gJ8lzY1o6TMmULSBJvfiO20h/view).

Обсуждение

Ключевым этапом работы алгоритма селекции зондов, кандидатных на включение в состав ДНК-биочипа, является определение целевых последовательностей. От консерватизма целевого участка напрямую зависят специфичность и дискриминационный потенциал зондов, т.е. способность ДНК-биочипа дифференциально детектировать патогены, относящиеся к таксонам различного уровня. Зонды к высококонсервативным участкам генома, таким как гены *16S* и *23S* рРНК, менее специфичны, но позволяют дифференцировать бактерии, принадлежащие к разным видам. Зонды к менее консервативным последовательностям, например бактериальным генам *recA*, *gyrB*, *rpoB*, способны различать штаммы микроорганизмов в пределах одного вида [1, 12].

В случае отсутствия информации о наличии генетических участков нужной степени консерватизма возможен поиск таковых в совокупности целевых последовательностей путем их множественного выравнивания. Однако множественное выравнивание большого количества длинных последовательностей (например, геномов) требует применения значительных вычислительных мощностей и занимает много времени [20].

В программе disprose был реализован иной подход к поиску целевых участков, идея которого заключалась в генерации большого количества коротких зондов и их выравнивании на совокупность последовательностей с применением алгоритма BLAST. Описанный процесс менее требователен к оборудованию, хорошо реализуется в параллельном режиме и в десятки раз менее затратен по времени. Зонды, выровненные на полный перечень целевых последовательностей со 100% покрытием, считаются зондами к консервативным участкам пула указанных последовательностей. Таким образом, варьирование перечня целевых последовательностей с применением программы disprose дает возможность подбора зондов для конструирования ДНК-биочипов с различным дискриминационным потенциалом.

Разработанный алгоритм апробирован нами при поиске специфичных зондов, позволяющих проводить дифференциальную детекцию *C. pneumoniae* в клинических образцах. *C. pneumoniae* является одним из множества микроорганизмов, вызывающих ВП. Совокупная доля этого и других «атипичных» возбудителей ВП составляет от 8 до 30% случаев заболевания [21, 22], и их своевременное выявление может способствовать решению проблемы недостаточно эффективной диагностики ВП [23]. Применение программы disprose с целью поиска зондов для детекции *C. pneumoniae* продемонстрировало возможности использования разработанного алгоритма.

В результате работы алгоритма из пула кандидатных зондов большого объема были отобраны сто зондов, обладающих высокой специфичностью в отношении целевого патогена. Сопоставление участков происхождения зондов с фрагментами аннотированного референсного генома показало, что большинство зондов сформированы из областей, кодирующих ферменты, белки семейства шаперонов, а также регуляторы клеточного цикла (см. приложение 4). Эти гены содержат высоко консервативные в отношении

C. pneumoniae участки и могут представлять интерес не только для разработки молекулярно-генетических диагностических тестов, но и для филогенетических исследований.

Заключение

Разработан алгоритм поиска специфичных зондов, позволяющих детектировать широкий спектр бактериальных и вирусных патогенов человека. Он реализован в виде программы для ЭВМ disprose, написанной на языке программирования R, и был использован при решении задачи поиска зондов для детекции *C. pneumoniae*. Алгоритм и программа подбора зондов обладают рядом достоинств:

универсальность — алгоритм направлен на поиск специфичных участков в наборе последовательностей любого объема и может быть легко адаптирован под решение широкого спектра задач;

модульность — исполнение алгоритма может происходить в несколько этапов, их порядок определяется пользователем, при этом любой этап может быть исключен или выполнен с помощью стороннего программного продукта;

открытость — исходный код пакета disprose находится в открытом доступе и может быть модифицирован в соответствии с задачей;

удобство использования — алгоритм может работать с популярными банками данных генетической информации (NCBI, GISAID) и локальными базами.

Вследствие гибкости и открытости программы область ее применения может быть расширена.

Источники финансирования. Данное исследование профинансировано из средств государственного бюджета в рамках выполнения отраслевой научно-исследовательской программы Роспотребнадзора на период 2021–2025 гг. «Научное обеспечение эпидемиологического надзора и санитарной охраны территории Российской Федерации. Создание новых технологий, средств и методов контроля и профилактики инфекционных и паразитарных болезней».

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Литература/References

1. Kostić T., Sessitsch A. Microbial diagnostic microarrays for the detection and typing of food- and water-borne (bacterial) pathogens. *Microarrays (Basel)* 2011; 1(1): 3–24, https://doi.org/10.3390/microarrays1010003.

2. Rouillard J.M., Zuker M., Gulari E. OligoArray 2.0: design of oligonucleotide probes for DNA microarrays using a thermodynamic approach. *Nucleic Acids Res* 2003; 31(12): 3057–3062, https://doi.org/10.1093/nar/gkg426.

3. Sung W.K., Lee W.H. Fast and accurate probe selection algorithm for large genomes. *Proc IEEE Comput Soc Bioinform Conf* 2003; 2: 65–74, https://doi.org/10.1109/csb.2003.1227305.

4. Urisman A., Fischer K.F., Chiu C.Y., Kistler A.L., Beck S., Wang D., DeRisi J.L. E-Predict: a computational strategy for species identification based on observed DNA microarray hybridization patterns. *Genome Biol* 2005; 6(9): R78, https://doi.org/10.1186/gb-2005-6-9-r78.

5. Watson M., Dukes J., Abu-Median A.B., King D.P., Britton P. DetectiV: visualization, normalization and significance testing for pathogen-detection microarray data. *Genome Biol* 2007; 8(9): R190, https://doi.org/10.1186/gb-2007-8-9-r190.

6. National Center for Biotechnology Information. *Nucleotide*. Bethesda (MD): National Library of Medicine (US); 2021. URL: https://www.ncbi.nlm.nih.gov/nucleotide/.

7. Lorenz R., Bernhart S.H., Höner Zu Siederdissen C., Tafer H., Flamm C., Stadler P.F., Hofacker I.L. ViennaRNA Package 2.0. *Algorithms Mol Biol* 2011; 6(1): 26, https://doi.org/10.1186/1748-7188-6-26.

8. McCaskill J.S. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers* 1990; 29(6–7): 1105–1119, https://doi.org/10.1002/bip.360290621.

9. Junhui L. *TmCalculator: melting temperature of nucleic acid sequences. R package version 1.0.1*. 2020. URL: https://CRAN.R-project.org/package=TmCalculator.

10. SantaLucia J. Jr. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc Natl Acad Sci U S A* 1998; 95(4): 1460–1465, https://doi.org/10.1073/pnas.95.4.1460.

11. Camacho C., Coulouris G., Avagyan V., Ma N., Papadopoulos J., Bealer K., Madden T.L. BLAST+: architecture and applications. *BMC Bioinformatics* 2009; 10: 421, https://doi.org/10.1186/1471-2105-10-421.

12. Bodrossy L., Sessitsch A. Oligonucleotide microarrays in microbial diagnostics. *Curr Opin Microbiol* 2004; 7(3): 245–254, https://doi.org/10.1016/j.mib.2004.04.005.

13. Sanguin H., Herrera A., Oger-Desfeux C., Dechesne A., Simonet P., Navarro E., Vogel T.M., Moënne-Loccoz Y., Nesme X., Grundmann G.L. Development and validation of a prototype 16S rRNA-based taxonomic microarray for Alphaproteobacteria. *Environ Microbiol* 2006; 8(2): 289–307, https://doi.org/10.1111/j.1462-2920.2005.00895.x.

14. Maskos U., Southern E.M. A study of oligonucleotide reassociation using large arrays of oligonucleotides synthesised on a glass support. *Nucleic Acids Res* 1993; 21(20): 4663–4669, https://doi.org/10.1093/nar/21.20.4663.

15. Raddatz G., Dehio M., Meyer T.F., Dehio C. PrimeArray: genome-scale primer design for DNA-microarray construction. *Bioinformatics* 2001; 17(1): 98–99, https://doi.org/10.1093/bioinformatics/17.1.98.

16. Wong C.W., Albert T.J., Vega V.B., Norton J.E., Cutler D.J., Richmond T.A., Stanton L.W., Liu E.T., Miller L.D. Tracking the evolution of the SARS coronavirus using high-throughput, high-density resequencing arrays. *Genome Res* 2004; 14(3): 398–405, https://doi.org/10.1101/gr.2141004.

17. Wong C.W., Heng C.L.W., Wan Yee L., Soh S.W.L., Kartasasmita C.B., Simoes E.A.F., Hibberd M.L., Sung W.K., Miller L.D. Optimization and clinical validation of a pathogen detection microarray. *Genome Bio* 2007; 8(5): R93, https://doi.org/10.1186/gb-2007-8-5-r93.

18. Yoo S.M., Keum K.C., Yoo S.Y., Choi J.Y., Chang K.H., Yoo N.C., Yoo W.M., Kim J.M., Lee D., Lee S.Y. Development of DNA microarray for pathogen detection. *Biotechnol*

Bioprocess Engin 2004; 9(2): 93–99, https://doi.org/10.1007/bf02932990.

19. Zuker M., Mathews D.H., Turner D.H. Algorithms and thermodynamics for RNA secondary structure prediction: a practical guide. In: Barciszewski J., Clark B.F.C. (editors). *RNA biochemistry and biotechnology*. Springer; 1999; p. 11–43, https://doi.org/10.1007/978-94-011-4485-8_2.

20. Pais F.S.M., Ruy P.C., Oliveria G., Coimbra R.S. Assessing the efficiency of multiple sequence alignment programs. *Algorithms Mol Biol* 2014; 9(1): 4, https://doi.org/10.1186/1748-7188-9-4.

21. Рачина С.А., Бобылев А.А. Атипичные возбудители внебольничной пневмонии: от эпидемиологии к особенностям диагностики и лечения. *Практическая пульмонология* 2016; 2: 20–27.

Rachina S.A., Bobylev A.A. Atypical pathogens of community-acquired pneumonia: epidemiology, diagnosis, and treatment. *Prakticeskaa pul'monologia* 2016; 2: 20–27.

22. Nair G.B., Niederman M.S. Updates on community acquired pneumonia management in the ICU. *Pharmacol Ther* 2021; 217: 107663, https://doi.org/10.1016/j.pharmthera.2020.107663.

23. Зайцев А.А. Внебольничная пневмония: возможности диагностики, лечения и вакцинопрофилактики в условиях пандемии COVID-19. *Практическая пульмонология* 2020; 1: 14–20.

Zaitsev A.A. Community-acquired pneumonia: diagnostic, treatment and vaccine prevention opportunities in the context of the COVID-19 pandemic. *Prakticeskaa pul'monologia* 2020; 1: 14–20.