Отчет о проделанной работе с использованием оборудования ИВЦ НГУ

1. Аннотация работы

В результате анализа данных запусков панелей для таргетного NGS установлены характеристики праймеров и ампликонов, оказывающие наибольшее влияние на равномерность покрытия целевых регионов, и определены диапазоны оптимальных значений данных параметров. На основании полученных диапазонов оптимальных параметров и литературных данных были протестированы различные алгоритмы машинного и глубокого обучения. Была разработана не зависящая от модуля primer3 программа дизайна праймеров для мультиплексной ПЦР, содержащая алгоритмы проверки праймеров на образование различных нецелевых продуктов и расположение SNP в сайте посадки праймера.

2. Тема работы

Разработка высокопроизводительной программы дизайна праймеров для таргетного секвенирования нового поколения (NGS).

3. Состав коллектива

- Михеева Регина Евгеньевна, mikheevare@yandex.ru, студент НГУ
- Кечин Андрей Андреевич, н.с., к.б.н., aa_kechin@niboch.nsc.ru

4. Информация о гранте

Работа выполняется без поддержки гранта.

5. Научное содержание работы

5.1. Постановка задачи

Целью данной работы является разработка высокопроизводительной программы дизайна праймеров для таргетного NGS. Для её достижения решались следующие задачи:

1) анализ и определение по ранее полученным данным таргетного NGS влияния характеристик праймеров и получаемых ПЦР-продуктов на равномерность покрытия целевых регионов;
2) разработка нового алгоритма дизайна олигонуклеотидов для таргетного NGS, учитывающего все возможные взаимодействия праймеров между собой и с фрагментами ДНК образца;

5.2. Современное состояние проблемы

Проблема конструирования эффективных праймеров возникла с появлением самой полимеразной цепной реакции (ПЦР). На сегодняшний день существует огромное множество коммерческих и свободно доступных программ, способных сконструировать олигонуклеотиды для различных целей. Наиболее простые программы предлагают провести дизайн одной пары праймеров, которые проведут амплификацию заданного района генома. Другие – нацелены на выявление какой-либо мутации за счёт различия в эффективности ПЦР при наличии замены на 3’-конце (аллель-специфичные праймеры), определение длины повторов (VNTR-анализ) или количественную оценку числа копий локуса ДНК или РНК (количественная ПЦР). Оптимальным по своей экономичности и трудозатратности является использование мультиплексной ПЦР – разновидности ПЦР, при которой в одной реакционной смеси происходит амплификация одновременно многих выбранных локусов. В этом случае требуется меньше не только реагентов (ДНК-полимеразы, дезоксирибонуклеозидтрифосфатов и компонентов буфера), но и самого исследуемого образца ДНК.

Контроль эффективности амплификации каждого локуса в мультиплексной ПЦР необходим для достижения равномерной представленности копий целевых последовательностей в конечной смеси. В случае достижения равномерности становится

возможным анализировать не только точечные мутации во всех исследуемых районах генома, но и оценивать число копий каждого локуса (copy number variations – CNV) [1]. Для амплификации каждого целевого района в мультиплексе с одинаковой эффективностью, то есть чтобы за каждый цикл число целевых молекул увеличивалось в одно и то же число раз, необходимо, чтобы с одинаковой эффективностью происходили три стадии ПЦР: (1) взаимодействие праймеров с ДНК-матрицей; (2) формирование комплекса ДНК-полимераза-праймер-ДНК-матрица и (3) удлинение праймера до длины ампликона, то есть элонгация [2]. На сегодняшний день существует несколько моделей протекания количественной моноплексной ПЦР [3–9]. Большинство представленных моделей построены на предположении, что каждый цикл ПЦР может быть описан уравнением Михаэлиса-Мэнтена, то есть представлен одностадийным процессом. Однако для того, чтобы определить влияние каждого из компонентов, необходимо разделить процесс амплификации на соответствующие стадии, что и сделано в модели Booth и соавторов. В данной модели учтено влияние на эффективность начальной концентрации ДНК, праймеров, продолжительностей элонгации и отжига, длины целевой последовательности, нарушения последовательности матрицы, а также денатурация ДНК-полимеразы с каждым циклом. В представленном проекте планируется усложнение этой модели путём учёта большого числа участвующих в мультиплексной ПЦР праймеров, учёта эффективности гибридизации праймеров с ДНК-матрицей и друг с другом и модификации нуклеотидов в составе ДНК-матрицы вследствие фиксации гистологического материала, чего не сделано ни в одной из представленных в научной литературе моделей.

Полученная модель оценки эффективности амплификации будет использована в разрабатываемой в проекте программе по конструированию олигонуклеотидов для мультиплексного обогащения NGS-библиотек целевыми последовательностями. На данный момент существует множество программ, предлагающих разработку праймеров для тех или иных задач. Подавляющее большинство из них перечислены в статье, описывающей программу NGS-PrimerPlex [10]. При этом сама программа на данный момент остается единственной свободно доступной программой, которая позволяет быстро проводить дизайн праймеров для новых таргетных NGS-панелей. В недавней работе Xie и соавторов был предложен алгоритм подбора праймеров для мультиплексной ПЦР [11], основанный на случайном переборе сочетаний праймеров. Однако для валидации подхода авторами была выбрана достаточно простая задача – анализ 384 непересекающихся SNP генома. В этом случае вероятность комплементарности праймеров, а, следовательно, и их 3'-концов значительно снижается по сравнению с панелями, в которых требуется секвенировать всю кодирующую последовательность какого-либо гена, и пары праймеров располагаются друг за другом. Кроме того, предложенный авторами алгоритм не был оформлен в виде программы, что усложняет его проверку. В работе Yuan и соавторы была описано программа Ultiplex [12], однако на момент написания заявки она уже была недоступна для использования и также основана на пакете primer3, что усложняет дизайн праймеров для мультиплексной ПЦР.

В работе Xie и соавторы был частично изучен и описан вопрос формирования праймер-димеров при проведении мультиплексной ПЦР [11]. Из топ-5 праймер-димеров, предсказанных по вычисленному ими показателю «badness», только один присутствовал в списке топ-5 праймер-димеров, прочитанных при секвенировании. Это может свидетельствовать, с одной стороны, о слишком сильном упрощении в схеме расчёта этого показателя для праймер-димеров, а, с другой стороны – о том, что необходимо учитывать кинетику наработки таких праймер-димеров, в том числе зависимость кинетики от последовательности димера. Так, в работе предполагается использовать общепринятый подход расчет термодинамических характеристик ДНК-дуплексов, основанный на модели ближайшего

соседа [13,14], а также подход TEEM (toehold exchange energy measurement) для оценки термодинамических показателей взаимодействия олигонуклеотидов между собой [13].

Кроме того, научным руководителем дипломной работы было описано влияние удаленности мисматча между матрицей ДНК и праймером от 3’-конца праймера на количество соответствующего ПЦР-продукта.

Таким образом, до сих остается мало исследованной область моделирования процессов формирования целевых и нецелевых продуктов мультиплексной ПЦР, и задачи, предложенные для решения в проекте, являются актуальными.

Основными научными конкурентами в России являются Центр алгоритмических биотехнологий в Санкт-Петербургском государственном университете (директор Алла Лапидус), Научно-исследовательский институт искусственного интеллекта (г. Москва), в мире – множество лабораторий, включая департамент биоинженерии Университета Райса (руководитель – Ганг Бао), компания NuProbe (руководитель – Давид Жанг), Центр компьютерной биологии Национального университета Сингапура (руководитель Энрикко Петретто), Центр секвенирования генома человека Медицинского колледжа Бейлора (руководитель – Ричард Гиббс).

Список литературы:

1. Chen Y. et al. SeqCNV: a novel method for identification of copy number variations in targeted next-generation sequencing data // BMC Bioinformatics. BioMed Central, 2017. Vol. 18, № 1. P. 147.
2. Louw T.M. et al. Experimental Validation of a Fundamental Model for PCR Efficiency. [Electronic resource] // Chemical engineering science. NIH Public Access, 2011. Vol. 66, № 8. P. 1783–1789. URL: http://www.ncbi.nlm.nih.gov/pubmed/21822325 (accessed: 31.01.2019).
3. Gevertz J.L., Dunn S.M., Roth C.M. Mathematical model of real-time PCR kinetics // Biotechnol Bioeng. 2005. Vol. 92, № 3. P. 346–355.
4. Cobbs G. Stepwise kinetic equilibrium models of quantitative polymerase chain reaction. // BMC Bioinformatics. BioMed Central, 2012. Vol. 13. P. 203.
5. Chigansky P., Jagers P., Klebaner F.C. What can be observed in real time PCR and when does it show? // J Math Biol. Springer, 2018. Vol. 76, № 3. P. 679–695.
6. Booth C.S. et al. Efficiency of the polymerase chain reaction [Electronic resource] // Chemical Engineering Science. NIH Public Access, 2010. Vol. 65, № 17. P. 4996–5006. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3142788 (accessed: 31.01.2019).
7. Chervoneva I. et al. Modeling qRT-PCR dynamics with application to cancer biomarker quantification. // Stat Methods Med Res. NIH Public Access, 2018. Vol. 27, № 9. P. 2581–2595.
8. Sánchez A. et al. Modeling Real-Time PCR Kinetics: Richards Reparametrized Equation for Quantitative Estimation of European Hake (Merluccius merluccius) // J Agric Food Chem. 2013. Vol. 61, № 14. P. 3488–3493.
9. Marimuthu K., Jing C., Chakrabarti R. Sequence-Dependent Biophysical Modeling of DNA Amplification // Biophys J. The Biophysical Society, 2014. Vol. 107, № 7. P. 1731.
10. Kechin A. et al. NGS-PrimerPlex: High-throughput primer design for multiplex polymerase chain reactions // PLoS Comput Biol / ed. Pertea M. Public Library of Science, 2020. Vol. 16, № 12. P. e1008468.
11. Xie N.G. et al. Designing highly multiplex PCR primer sets with Simulated Annealing Design using Dimer Likelihood Estimation (SADDLE) // Nature Communications 2022 13:1. Nature Publishing Group, 2022. Vol. 13, № 1. P. 1–10.

12. Yuan J. et al. The web-based multiplex PCR primer design software Ultiplex and the associated experimental workflow: up to 100- plex multiplicity // BMC Genomics. BioMed Central Ltd, 2021. Vol. 22, № 1. P. 1–17.
13. Bae J.H., Zhang D.Y. Predicting stability of DNA bulge at mononucleotide microsatellite // Nucleic Acids Res. Oxford Academic, 2021. Vol. 49, № 14. P. 7901–7908.
14. SantaLucia J. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics // Proc Natl Acad Sci U S A. National Academy of Sciences, 1998. Vol. 95, № 4. P. 1460–1465.

5.3. Подробное описание работы, включая используемые алгоритмы

Были проанализированы девять ранее сконструированных другой программой (NGS-PrimerPlex) панелей для таргетного секвенирования, включающих 564 пары праймеров, и результаты запусков NGS с их использованием. Были протестированы следующие характеристики праймеров: GC-составы праймеров и ампликонов, температуры плавления праймеров, разница температур плавления между прямым и обратным праймерами, длины праймеров и ампликонов.

Полученные результаты анализа параметров праймеров и получаемых ПЦР-продуктов позволили предположить возможность применения технологий машинного обучения с целью классификации праймеров и, как следствие, отбора и дальнейшего анализа только лучших по характеристикам праймеров.

Было опробовано четыре различных модели машинного обучения: модель градиентного бустинга, модель случайного леса, модель логистической регрессии и модель Бернулли. Также была протестирована нейросеть, работающая с табличными данными и имеющая высокую эффективность работы с небольшими выборками (https://github.com/automl/TabPFN).

Обучение моделей производилось на различных выборках из имеющихся данных, состоящих из всевозможных комбинаций характеристик праймеров и ампликонов. Тренировочная выборка, содержала зависимый бинарный параметр, отражающий эффективность праймера: 1 — для праймеров, при использовании которых кратность покрытия находится в диапазоне от 0,24 до 4; 0 — для остальных праймеров. Соотношение тренировочной выборки к тестовой составляло 70:30. Непосредственно перед обучением моделей было произведено масштабирование данных, чтобы исключить возможность неравномерного влияния параметров на итоговый результат, однако для нейросети масштабирования не осуществлялось, так как она имеет встроенную нормализацию данных.

5.4. Полученные результаты

При тестировании алгоритмов машинного и глубокого обучения лучшие результаты показало обучение моделей на данных из девяти параметров: длины и температуры плавления праймера, GC-состава праймера и его 3`-конца, разницы температур плавления прямого и обратного праймеров, длины и GC-состава ампликона, медианное ΔG праймер-димера, медианное ΔG сцепленного 3`-конца праймер-димера и категориальный признак для нейросети, показывающий имеет ли праймер-димер сцепленный 3`-конец. Было опробовано два варианта обучения моделей — на сбалансированной выборке (1:1) и несбалансированной, состоящей из всех имеющихся данных (соотношение эффективных праймеров и неэффективных — 3:1). Результаты обучения моделей представлены в таблице 1.

Таблица 1. Результаты применения моделей машинного и глубокого обучения с целью классификации праймеров.

Название модели	Значение AUC для сбалансированной выборки	Значение F1 для несбалансированной выборки
Модель градиентного бустинга	0,5584	0,8288
Модель логистической регрессии	0,5582	0,8385
Модель случайного леса	0,5888	0,6814
Модель Бернулли	0,5031	0,8362
Табличная нейросеть	0,5405	0,8344

В ходе работы был создан собственный алгоритм дизайна праймеров для моноплексной ПЦР. Алгоритм выполняет функции конструирования праймеров так же, как и модуль primer3. Необходимость данного этапа работы объясняется отсутствием возможности контролировать сортировку и фильтрацию праймеров при использовании модуля primer3.

Был разработан алгоритм проверки ранее сконструированных праймеров на нецелевую гибридизацию, наличие SNP в сайтах посадки праймеров и дальнейшее составления мультиплексных наборов. Входными параметрами алгоритма являются последовательности праймеров с предыдущего этапа и количество вариантов для каждого мультиплексного набора.

После создания алгоритма дизайна праймеров для мультиплексной ПЦР была произведена оценка сложности его частей. Поскольку оценка сложности в зависимости от времени работы программы и объема используемой памяти является неточной в связи с разными последовательностями целевых регионов и, как следствие, разным количеством праймеров и их характеристик, то была выполнена приблизительная оценка для каждого из блоков программы в соответствии с количеством выполняемых операций.

Сложность этапа конструирования праймеров для каждого целевого региона зависит от количества целевых регионов и соотношения перекрывающихся и неперекрывающихся пар праймеров, то есть $O((k - m) \times N)$, где k — общее количество праймеров для региона, m — количество праймеров, чьи последовательности перекрываются с предыдущими, N — число целевых регионов генома.

Сложность проверки праймеров на нецелевую гибридизацию также зависит от количества целевых регионов и числа пар праймеров для проверки для каждого из регионов, то есть $O(k \times N)$, где k — количество праймеров для целевого региона, отобранных для проверки согласно значению SCORE, N — количество целевых регионов генома.

Поскольку проверка сайтов посадки праймеров на наличие SNP была сведена к проверке последовательностей целевых регионов на перекрывание SNP, то сложность данного этапа программы составляет O(N), где N — количество целевых регионов генома.

Сложность последнего этапа программы, а именно составления комбинаций и праймеров и их проверки на димеризацию зависит от количества блоков, числа пар праймеров для каждого блока и среднего количества комплементарных участков, содержащихся в праймерах, то есть приблизительно $O(N \times (k/2)mN + Z \times N)$, где N — число блоков, k — среднее

число пар праймеров для блока, m — число пар праймеров, которое необходимо для покрытия одного блока, Z — число комбинаций внутри блока, которые будут перебираться, чтобы получить нужное количество наборов пар праймеров между блоками (обычно 1–10).

5.5. Эффект от использования кластера в достижении целей работы

Непосредственно на кластере было проведено тестирование моделей машинного и глубокого обучения, производящих бинарную классификацию праймеров. Данный этап был необходим для перехода к следующей стадии дипломной работы – отбора наиболее подходящих вариантов праймеров для мультиплексной ПЦР из множества возможных.

5.6. Перечень публикаций, содержащих результаты работы

На данный момент публикации отсутствуют.