УДК 572.112

ОБ ОБРАБОТКЕ ДАННЫХ ВЫСОКОПРОИЗВОДИТЕЛЬНОГО СЕКВЕНИРОВАНИЯ

Протько М.А.[1], студент гр.050502

Белорусский государственный университет информатики и радиоэлектроники[1]
г. Минск, Республика Беларусь

Борисенко О.Ф. – канд. физ.-мат. наук

Аннотация. В данной работе представлены краткие сведения, необходимые для изменений и оптимизаций существующих алгоритмов, используемых в сфере биоинформатики, а именно, анализа данных секвенирования ДНК человека.

Ключевые слова. Секвенирование Illumina/Solexa, анализ геномных данных, индексирование, аннотация, выравнивание, болезни экспансии.

Мною была поставлена задача по оптимизации поиска экспансии тринуклеотидных повторов. На данный момент, не смотря на огромное количество производительных инструментов, большая часть действий совершается человеком. А именно: берется выборка из всех доступных инструментов на каждом этапе геномного анализа и, на основании опыта исследователя, меняются некоторые параметры до тех пор, пока не получается повторяемый несколько раз результат с возможностью получить одни и те же выводы. Примеры такого процесса можно увидеть в статьях Академии Наук, где выделен отдельный подпункт «Объекты и методы исследования» [1-2]. Данные методы работы полны субъективных оценок и параметров, особенно при аргументировании доказательств верности выводов. Отсутствие должной методологии привело и приводит к катастрофическим последствиям, показательным примером которых является создание нескольких форматов одного и того же файла данных секвенирования fastq без возможности определить, к какой именно версии файл относится (форматы Illumina, Phred, Solexa, Sanger [3]). Данная ошибка привела к путанице и невозможности доказать верность или повторить процесс получения данных многих исследователей. Прецеденты таких ошибок далеко не исключительны, настолько, что был проведен их статистический анализ с выводом критериев, по которым возможно определить, что исследование ошибочно [4].

Из всего вышеописанного был сделан следующий вывод: необходимо тщательно проверять всю информацию, объясняя подробно каждый подпункт, что особенно важно, ведь применение решения изначальной задачи будет относиться к нахождению генетических заболеваний человека. Не хотелось бы по неосмотрительности выставлять неизлечимый диагноз. Но, в то же время, хотелось бы обнаружить всех индивидуумов в зоне риска.

Поиск тринуклеотидных повторов позволяет диагностировать около 22 болезней экспансии [5]. Данные заболевания коварны, не только потому что протекают болезненно и неизлечимы, но и потому, что могут оставаться незамеченными в геноме очень долгое время, пока не проявятся (иногда даже в возрасте 40-55 лет при атаксиях). Это говорит о том, что начальная стадия заболевания проходит бессимптомно, что позволяет обнаружить ее только при геномном анализе и анализе семейных заболеваний. Получение столь ценной информации на раннем этапе течения болезни при должных профилактических мерах позволит значительно улучшить качество жизни.

Задача данной работы: разработать специальный алгоритм для нахождения тринуклеотидных повторов при болезнях экспансии. Поскольку объем материалов и данных для ее решения невозможно осветить одной статьей, было решено разбить ее на подзадачи. Поэтому, цель данной статьи такова: сравнить существующие алгоритмы для анализа данных секвенирования Illumina (с упором на выравнивание), а также поиск тех фактов и оптимизаций, которые позволяют ускорить этот процесс.

Стоит добавить, что в данной статье не раскрываются общепринятые термины биоинформатики.

«Константы»

В процессе работы данные будут всегда содержать следующие закономерности [6]:

– Прочтение человеческого генома содержит две комплементарные друг другу цепи – т.е., прочтение является парным, или же: данные секвенирования будут содержать два файла-копии. Необходимо стремиться к тому, чтобы оба парных чтения были максимально подобны друг другу.

– Частота встречаемости последовательности 5'-CG-3' составляет от 1/2 до 1/5 частот 5'-GC-3', что приблизительно равно 0,23 и 1,15 соответственно для генома человека.

– Кодирующие участки ДНК (по ним «собирается» полипептид) составляют около ~3-5% от всего генома.

– Каждый геном человека имеет около 1,6 – 3,2 миллионов однонуклеотидных различий на всю длину генома (~3,2 млрд нуклеотидов) на протяжении 1000-2000 пар оснований.

– Набор, получаемый из секвенирования, диплоидный (две копии, перемешанные в одном файле, одного и того же гена).

– Болезни экспансии определяются наличием многократных повторов определенной последовательности (к примеру ‘CAG’), локализующихся как в кодируемой, так и не кодируемой области. Для развития некоторых заболеваний необходим один ген с повторами в аллельных генах, для некоторых – необходимо два (доминантные и рецессивные соответственно). То есть, необходимо искать в худшем случае две копии одного и того же гена.

Подробнее о математических свойствах и закономерностях генетических кодов в [7-8].

Возможные решения (постановка задачи)

Исходя из поставленной задачи для нахождения болезней экспансии необходимо совершить следующую последовательность действий:

Фильтрация данных: проводится с помощью статистических инструментов (fastqc) и trimmomatic (или Cutadapt) по стратегии ‘crop and trim’. Все, что посредством анализа не проходит поставленный фильтр не считается качественными данными.

В инструментах фильтрации (fastqc), помимо базовой статистики (распределение длины ридов и качества, поиск адаптеров), используются инструменты поиска взаимосвязи меду разными прочтениями, чтобы на раннем этапе обнаружить исключительные ситуации, возникающие из-за несовершенства оборудования (пропущенные фрагменты большой длины, ПЦР-дубликаты и т.д.).

Аннотация: поскольку в данной работе необходимо найти болезни экспансии, нет необходимости полностью картировать весь геном, только его части, содержащие тринуклеотидные повторы. Учитывая короткую длину ридов, с которыми необходимо работать (100-200 bp), могут возникнуть пять случаев:

1. Нужная последовательность находится посередине (с обоих концов ее можно определить, сравнив с референсным геномом).
2. Имеется только левая часть (не повторы, а кодируемый набор).
3. Имеется только правая часть.
4. Имеются только повторы.
5. Повторы прочитаны не были.

Также могут быть случаи нахождения в кодируемой (экзон) и некодируемой (интрон) области (в последнем случае процесс выравнивания усложняется из-за огромной вариативности последней).

Исходя из пяти случаев для каждого имеется свой алгоритм действий:

Случаи 1-3: поиск всех тринуклеотидных повторов в референсном геноме, выравнивание всех найденных областей. Поиск области с минимальной разницей.

Случаи 4-5: сборка элементов (ридов) по перекрытиям в скаффолд или континг, индексирование референсного генома (аннотирование посредством выравнивания с областью), подсчет наиболее вероятного случая (в случае, если был собран скаффолд с неизвестными областями, находящимися в районе повторов).

Также стоит учитывать необходимость поиска двойного набора и распределения по аллелям генов для определения рецессивных заболеваний (необходимо две копии с мутацией для развития признака).

Этап анализа (в него входит определение экспрессивности найденных генов и определение их функций (создает фермент, полипептид и прочее), поиск взаимосвязи между различными локализациями мутаций и одной болезнью) не имеет необходимости, поскольку болезни экспансии определены только повторами в конкретных местах и не имеют взаимосвязи друг с другом. Параметры экспрессии считать повторно не нужно, ведь они имеют прямую зависимость от количества повторов.

Также, следует не забывать про рекомендации [9], по которым всегда нужно находить ответы на следующие вопросы:

– Размер фрагментов имеет правильный разброс? (качество ДНК)

– Параметры качества на фрагмент/основание одного чтения достаточны для того, чтобы им доверять?

– Пропорция уникальных и повторяющихся ридов (GC) соответствует допустимой?

– Сборка de novo или сборка относительно референсного генома?

– Выбор вида выравнивания: уникальное, не уникальное, локальное или глобальное.

– Насколько оправдано использование эвристического подхода (как он влияет на качество и покрытие данных).

Далее рассмотрим часть вопросов, которые обязательно возникнут при разработке.

Как производить поиск подобных последовательностей (индексирование)

Задача индексирования (подобна разбиению слов на языке на суффиксы с созданием словаря по ним) позволяет значительно ускорить работу (конечно, никто не отменяет метод грубой силы, он также применим, но совершенно не оптимизирован).

Для индексирования существуют следующие алгоритмы [10]:

– Использующие хэш-таблицы: fasta, BLAST, SSAHA, BLAT, GMAP, MAQ.

– Использующие BWT-FM: BWT-SW, BWA, Bowtie, SOAP2, TopHat, BS-Seeker.

– Использующие суффиксы: QPALMA, seqemehl, VMATCH, LAST, BLASR, Masai.

Хэш-таблицы используются значительно чаще из-за их простоты в имплементации и быстроты работы. Но хэш-таблицы весьма громоздки и не позволяют находить «неточные» совпадения.

Пример алгоритма хэширования, используемого в инструменте fasta [11]:

– Провести экстракцию сидов (так называется короткий элемент рида, описывающий хэш-таблицу) на референсном геноме. Хэш-таблица будет содержать: сид и список расположений этого сида в риде. Размер слов в таблице от 4-6 оснований.

– Провести экстракцию сидов из ридов секвенирования.

– Проводить выравнивание по совпадающим запросам (фильтрация перед выравниванием для поиска наиболее подходящих вариантов).

Выравнивание двух ридов

Данный шаг позволяет численно определить различия между двумя прочтениями: рида относительно референсного генома. Алгоритмы выравнивания должны определять минимальное количество различий между двумя геномными последовательностями, природу каждого различия и его местоположение в одной из двух заданных последовательностей. Такая информация представляет собой совокупность оптимальных местоположений и типов каждой правки.

Популярнейшие алгоритмы выравнивания и инструменты, что их используют:

– Алгоритм Смита-Уотермана (локальное выравнивание, динамический): fasta, Gapped BLAST, BLASTZ, BWT_SW, MAQ, Zoom.

– Алгоритм Нидлмана-Вунша (глобальное выравнивание, динамический): fasta, SSAHA, GMAP, GNUMAP, GenomeMapper, PASS.

– Расстояние Хэмминга (не динамический): RMAP, BRAT, BSMAP, Bowtie, MOM, PerM.

– Битовый вектор Майерса (не динамический); DREAM-Yara, Masai, Hobbes2.

Также достойно упоминания: Алгоритм Рабина-Карпа, полуглобальное выравнивание, алгоритм Ландау-Вишкина, алгоритмы без использования динамического программирования с опорой на эвристический подход.

Когда необходим поиск генетических замен, вставок и удалений, предпочтение отдается алгоритмам, основанным на динамическом программировании, а не на алгоритмах, его не использующих. В целом, алгоритм локального выравнивания предпочтительнее глобального, когда ожидается, что только часть считанных данных будет совпадать с некоторыми областями референсного генома из-за, например, больших структурных вариаций. Поскольку алгоритмы динамического программирования для выравнивания имеют квадратическую сложность вычисления, они не используются при больших объемах данных.

Плюсы использования коротких последовательности: они меньше подвержены ошибкам при выравнивании, их покрытие всего генома больше, на нем легче проводить локальное выравнивание, легче находить полиморфизмы. Существенным минусом использования коротких последовательностей является сложность обнаружения структурных вариантов (или же множественных замен, болезни экспансии к ним относятся). Исходя из этого, возможен вариант сборки по перекрытиям большого фрагмента и затем его выравнивание.

Сравнение из работы [10] показывает, что самыми быстрыми и точными реализациями являются те, которые не ограничиваются одним алгоритмом.

Определим корректное выравнивание следующим образом [12]: рид корректно выровнен, если он не нарушает критериев выравнивания (иное не математическое определение: рид корректно выравнен, если он соответствует своему исходному местоположению в геноме). Данные критерии определяются по методу оценки Рабема и оценке точности.

В целом, производительность инструментов оценивается с учетом трех аспектов, а именно пропускной способности (bps/sec) или времени выполнения, объема занимаемой памяти и процента выравниваний.

Работа с диплоидным набором

Согласно высокому уровню абстракции, для определения двух разных гетерозиготных генов (haplotype phasing [13]) необходимо произвести следующие действия:

Во время выравнивания ридов к референсному геному, использовать не уникальные выравнивания (учитывая то, что имеются две копии гена, которые могут быть гетерезогитоными), согласно такой стратегии, возможны четыре возможные сочетания двух копий: две копии одинаковые, обе разные (по сравнению с референсым геномом), одна из копий соответствует референсному геному (гетерозиготные позиции, первые две – гомозиготные).

Для решения задачи определения гетерозиготных позиций без использования специальных библиотек (составляемых при подготовке материала для секвенатора) используются длинные риды. В процессе обработки (во время выравнивания) решается проблема максимального среза графа. В графе каждый узел это одиночный полиморфизм рида, каждое ребро в графе обозначает, что в двух ридах один и тот же полиморфизм (используется в инструменте HapCut), или же решая проблему максимального веса связующих деревьев в графе (в HapCompass).

Также, решение поставленной задачи для диплоидного набора требует альтернативного решения NP-проблем: проблемы восстановления сообщества (необходимо для каждого взаимосвязанного рида найти его принадлежность к одному из двух аллельных генов, причем по ребру связного графа), что также сводится в проблеме поиска максимально среза графа.

Заключение

Проведя сравнение большинства инструментов, было выяснено, что большая часть их них использует определенные проверенные алгоритмы в категории state-of-art (так они обозначаются в большой части материалов из списка литературы). По этой причине, стоит исследовать алгоритмы отдельно от их практической реализации, поскольку последняя содержит список оптимизаций, иногда приводящих к неточным результатам (то есть, невозможно гарантировать, что все необходимые данные будут найдены).

Большая часть методов является эвристическими, потому что призвана решать изначально NP-задачу. Большая часть инструментов имеет один и тот же алгоритм, но разные методы перехода от качества в скорость и выполняемость. Данных очень много – чем-то приходится жертвовать.

Для столь специфической задачи, как поиск болезней экспансии, необходимо найти свою пропорцию скорость/качество, как и свое множество применяемых алгоритмов.

Также было выяснено, что однозначно определится с использованием пенальти при алгоритмах выравнивания и оптимизаций без сравнения реализаций невозможно.

На основе анализа алгоритмов и практических реализаций (описанных в данной статье и ее источниках), для этапа разработки первого варианта моего инструмента были выбраны следующие параметры:

– Алгоритм индексирования с использованием ранее созданной базы сидов на референсном геноме hg38 с использованием хэш-таблиц.

– Для лучших случаев (случаи 1-3) было решено использовать алгоритмы Нидлмана-Вуша и Смита-Уоттермана.

– Использование коротких ридов (в случаях 1-3).

– Сборка контингов по перекрытиям по жадному алгоритму (Overlap-Layout-Consensus) и использование выравнивания по большим фрагментам FANGS [14] (также работает с алгоритмом индексирования на основе хэш-таблиц) в случаях 4-5, а также при исследуемых заболеваниях рецессивного типа.

– В обоих случаях допускается использование не уникальных выравниваний.

Список использованных источников:

1. Малышева, О. М. [и др.]: Роль генетических нарушений в формировании инвалидизирующих последствий у недоношенного новорожденного. / Proceedings of the National Academy of Sciences of Belarus. Biological series, 2020, vol. 65, no. 3, pp. 328–341

2. Пилипчук, Т. А. [и др.]: Особенности молекулярно-генетической организации pseudomonas phage БИМ BV-45 Д. / Proceedings of the National Academy of Sciences of Belarus. Biological series, 2022, vol. 67, no. 2, pp. 190–196

3. Peter J. A. Cock [u др.]: The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants / Nucleic Acids Research, Volume 38, Issue 6, 1 April 2010, Pages 1767–1771, December 2009.

4. John, P. A. Ioannidis: Why Most Published Research Findings Are False. / PLOS Medicine 19(8): e1004085.

5. Albert R. La Spada, J. Paul Taylor: Repeat expansion disease: Progress and puzzles in disease pathogenesis / Nat Rev Genet. 2010 Apr; 11(4): 247–258

6. Житкевич, Т.И. Молекулярная медицина: молекулярные основы генных болезней. / Курс лекций. Минск «ИВЦ Минфина». 2018. 95 стр. Учреждение образования «Международный государственный экологический институт имени А. Д. Сахарова» Белорусского государственного университета. УДК 577:61:575 ББК 28.070+28.04+5я7. ISBN 978-985-7205-40-0

7. Протько, М. А. Алгоритм кодирования процесса трансляции белков в клетке / М. А. Протько, О. Ф. Борисенко // Технологии передачи и обработки информации : материалы Международного научно-технического семинара, Минск, март-апрель 2023 г. / Белорусский государственный университет информатики и радиоэлектроники; редкол.: В. Ю. Цветков [и др.]. – Минск, 2023. – С. 108–112.

8. Козлов, Н.Н.: Математический анализ генетического кода / БИНОМ. Лаборатория знаний, 2010. — 215 с. : ил., [8] с. цв. вкл. — (Математическое моделирование).

9. Applications of Clinical Microbial Next-Generation Sequencing/ American Academy of Microbiology. 2016. 65 p.

10. Mohammed Alser [u др.], Technology dictates algorithms: recent developments in read alignment / Genome Biology volume 22, Article number: 249 (2021), 26 August 2021

11. Dr. Mamta C. Padole, Search Algorithm Used in FASTA / Conference: CONMICRO-2005 - Current Trends in Computer Technology & BioInformaticsAt: Lucknow, India, May 2005.

12. Ayat Hatem [u др.]: Benchmarking short sequence mapping tools / BMC Bioinformatics / Volume 14, article number 184, (2013), 7 June 2013.

13. Data Science for High-Throughput Sequencing: Lecture 10: Haplotype Phasing - Community Recovery [эл.ресурс] / Режим доступа – https://data-science -sequencing.github.io/Win2018/lectures/lecture10. Дата доступа – 8.04.2024.

14. Sanchit Misr [u др.]: FANGS: High Speed Sequence Mapping for Next generation Sequencers / Electrical Engineering and Computer Science Northwestern University Evanston, IL 60208.

575.112

PROCESSING OF HIGH-THROUGH SEQUENCING DATA

Protsko M.A.[1]

Belarusian State University of Informatics and Radioelectronics[1]*, Minsk, Republic of Belarus*

Borisenko O.F. – PhD in Physics and Mathematics

Annotation. This paper provides a summary of the necessary changes and optimizations to existing algorithms used in the field of bioinformatics, namely the analysis of human DNA sequencing data.

Keywords. lumina/Solexa sequencing, genomic data analysis, indexing, annotation, alignment, expansion diseases.