СОЗДАНИЕ МАТЕМАТИЧЕСКОЙ МОДЕЛИ ДАННЫХ КОРОТКОГО СЕКВЕНИРОВАНИЯ

Протько М. А.
Кафедра программного обеспечения информационных технологий,
Белорусский государственный университет информатики и радиоэлектороники
Минск, Республика Беларусь
E-mail: mari.protko@mail.ru

Для создания алгоритма анализа данных секвенирования коротких ридов, полученных по технологии Illumina/Sloexa, для поиска по ним тандемных повторов необходимо убедиться в реализуемости поставленной задачи с достаточной точностью. Для этого нужно определить необходимые параметры точности для каждой ключевой характеристики системы, а затем смоделировать ее работу на худших и лучших случаях. По анализу данной модели возможно оценить реализуемость задачи.

Введение

Автором была поставлена задача поиска болезней экспансии по данным секвенирования Illumina/Solexa. Данные и сама задача были получены из института цитологии и генетики НАН РБ.

Если перефразировать задачу, то необходимо найти на данных с алфавитом, состоящим из четырех букв (A,T,C,G), необходимые последовательности «слов» (триплетов CAG, как в работе [1]), причем сам «текст» (файл fastq секвенируемого генома) состоит из множественного числа пропусков и опечаток, а страницы, на которых он был напечатан, были перепутаны с его копиями.

Возвращаясь к изначальной формулировке задачи, помимо основных повторений (также их называют «тандемными повторами»), характеризующих болезни экспансии, необходимо найти соседствующие с ними последовательности (на обоих концах повтора), которые называют маркирующими. Посредством маркирующих последовательностей найденный повтор возможно выровнять на представленный референсный геном и, соответственно, определить представленное заболевание.

Поставленная задача выполнима на технологии секвенирования Nanopore (ONT) и PacBio, что показано в работе [2], но возможна ли данная процедура на коротких ридах Illumina/Solexa?

I. Зачем нужна математическая модель?

Поскольку идея использовать данные секвенирования для диагностирования (а не для исследования) достаточна нова в научном сообществе, практические разработки с четкими и доказанными примерами применения отсутствуют, а те, что имеются, принадлежат специфической и крайне недоступной технологии длинных ридов, получение которых весьма дорого.

Обычно процедура проверки наличия в геноме болезней экспансии подкрепляется сравнением результатов посредством иного, более точного теста. Данные сравниваются чаще всего с некими химическими тестами на болезни экспансии ([3] – причем здесь только проверка наличия повторов, само же заболевание определяется по симптомам, что весьма проблематично для диагностирования некоторых атаксий на начальных стадиях) и после анализируются (процесс анализа состоит из очистки данных генома от заведомо ошибочных последовательностей и частичной сборки («выравнивания») на референтную последовательность). Проблема такого анализа в том, что каждый его этап состоит из множества допущений, исходящих из эвристического подхода, который постепенно в течение многих лет внедрялся по среднестатистическому успешному результату. Из вышеописанного следует, что специализированных инструментов анализа данных для поиска тандемных повторов нет, особенно на коротких рядах.

Учитывая то, что автор не располагает ни финансовыми, ни вычислительными ресурсами для проверки своих алгоритмов и гипотез, было решено смоделировать необходимые данные, чтобы оценить качество анализа данных fastq для секвенирования Illumina/Solexa.

II. Определенные «проблемы»

При работе с созданием программного обеспечения для анализа данных генома существуют некоторые особенности, которые не позволяют сделать процесс максимально точным. Такими особенностями являются используемые алгоритмы (неправильное построение матрицы цен при выравнивании), неточность физико-химических реакций (ПЦР-амплификации и считывания цвета люминисценции при секвенировании, к примеру) и использование неподходящих инструментов на каком-либо этапе (обрезка праймеров с помощью trimmomatic, который автоматически избавляется от повторяющихся последовательностей CAG, что заведомо гарантирует невозможность обнаружить тандемные повторы с ним).

Автором в процессе работы с множеством этих «особенностей» были выделены ключевые, которые значительно влияют на результат. Здесь

и далее они будут называны «проблемами», отличие их от «особенностей» в том, что «проблема» поддается анализу и математическому описанию (вероятность возникновения «проблемы» можно определить, а также присутсвует градация для худших и лучших случаев).

Для решения поставленной задачи, необходимо проверить худшие и лучшие случаи, исходя из действия следующих «проблем»:

1. Короткие риды – длинная последовательность;
2. вероятность ошибки на чтении;
3. ошибки ПЦР-амплификации;
4. неправильный референсный геном;
5. остатки праймеров в анализируемых данных;
6. пропущенный фрагмент при парном чтении слишком большой.

Также достойны упоминания следующие упрощения поставленной задачи: автором рассматриваются только болезни экспансии в экзонной («кодируемой») области гена, причем для проявления болезни достаточна встреча только в одной паре гена из аллеля.

III. Пример работы с «проблемой»

Для работы с каждой определенной «проблемой» необходимо совершить следующие действия:

- Определить причину ее происхождения и параметры;
- определить формулы расчета ошибок (для каждого пункта 2.1-2.6 свои);
- определить «лучший» и «худший» возможные случаи;
- определить статистические закономерности, не учитывая конкретные свойства системы.

На выходе из всех этих шагов возможно получить математическую модель для поставленной «проблемы».

Рассмотрим подробнее «проблему» из списка 2.2. Задача – построить ее математическую модель таким образом, чтобы получить все возможные ошибки с заранее известной вероятностью.

Для ошибки секвенирования параметры ошибок подробно расписаны в инструкциях к моделям секвенатора [4].

Согласно официальному источнику [5], минимальным порогом качества является расчет процента выравнивания PhiX. Согласно ему, рекомендуемый минимум: 5% выравнивания данных(худший случай); для оптимальной производительности: 32%; для большого количества полиморфизмов: 40% . Самый лучший случай, которого возможно добиться на новом оборудовании – 94,9 %.

Автором был проведен анализ полученных данных секвенирования, согласно которому, с учетом вышеописанных свойств и параметров, математическая модель 2.2. представляет собой создание данных с нужным error rate (для характеристик ошибки секвенирования используются величины p и q, из которых по формуле из [6] получается error rate (e)).

Добиться нужного error rate, дающего подобные анализируемым экспериментальным данным показатели распределения, возможно по формуле экспоненциального закона с применением метода инверсии функции распределения:

$$Y = a - \frac{1}{\lambda}\ln(1 - e^{-\lambda(b-a)}) \qquad (1)$$

где a и b – левая и правая границы интервала соответственно; λ – параметр масштаба интервала; x – число, взятое на промежутке значений из нормального распределения; Y – число, удовлетворяющее экспоненциальному закону (искомый error rate).

Подбор параметров формулы (1) совершается индуктивным методом на основе проведенного анализа закономерностей.

Заключение

Для успешного получения данных для полного покрытия тестами и получения ответа на вопрос, о возможности решения поставленной задачи на данной технологии необходимо построить такую математическую модель, которая позволяет регулировать выраженность каждого пункта (через вероятность возникновения ошибки) 2.1-2.6 как по отдельности, так и вместе. Таким образом, будет получен набор данных, в котором с абсолютной точностью будут определены представленные ошибки, что позволит проверять каждый последующий разрабатываемый алгоритм и его характеристики.

1. Протько, М. А. Об обработке данных высокопроизводительного секвенирования // Компьютерные системы и сети: сборник статей 60–й научной конференции аспирантов, магистрантов и студентов, Минск, 22–26 апреля 2024 г. / Белорусский государственный университет информатики и радиоэлектроники. – Минск, 2024. –С. 18–22.
2. Ebbert Mark T. W. [и др.]: Long-read sequencing across the C9orf72 ‘GGGGCC’ repeat expansion: implications for clinical use and genetic discovery efforts in human disease // Molecular Neurodegeneration volume 13, Article number: 46 (2018)
3. Warner J.P. [и др.]: A general method for the detection of large CAG repeat expansions by fluorescent PCR // Journal of Medical Genetics 33(12):1022– 6 DOI: 10.1136/jmg.33.12.1022, January 1997.
4. Illumina DNAPrep Reference Guide Document#1000000025416v09 // ILLUMINA PROPRIETARY /June 2020
5. SAMUEL KARLIN [и др.]: Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes // Proc. Natl. Acad. Sci. USA Vol. 87, pp. 2264-2268, March 1990 Evolution December 26, 1989
6. Mechanical DNA Fragmentation with the Q800R2 Sonicator / 2017, Illumina Prepare Library