ОТЗЫВ

члена диссертационного совета на диссертацию Пржибельского Андрея Дмитриевича на тему: «Алгоритмы для сборки геномов и транскриптомов», представленную на соискание ученой степени кандидата физико-математических наук по специальности 03.01.09 "Математическая биология, биоинформатика",

Развитие и все более широкое использование новых технологий секвенирования (NGS) позволяет получать гигантские объемы информации. Однако сырые данные не представляют особого интереса, пока они не будут предварительно обработаны. Исходные данные, как правило, представляют наборы сотен миллионов коротких прочтений ДНК, которые для дальнейшего биологического анализа следует собрать в непрерывные последовательности – геномы. Другим источником таких данных являются данные о транскрипции – прочитанные короткие фрагменты РНК, которые были скопированы с ДНК. Здесь тоже стоит задача сборки из коротких фрагментов полноразмерных последовательностей РНК.

Структура диссертации обычна. Ведение содержит основные сведения о работе. Литературный обзор, представленный в первой главе весьма краток (5 страниц), хотя и содержит достаточно большое количество ссылок, посвящен лишь общему описанию проблемы и известным подходам к их решению.

Стандартная процедура сборки геномов основывается на данных, полученных из сотен тысяч одинаковых клеток, что предполагает предварительную культивацию культур. Однако значительная часть бактериальных клеток являются «некультивируемыми», т.е. для них не известны условия, при которых можно из отдельной клетки получить культуру. Это, в частности, бактерии, образующие биопленки. Некоторые виды таких бактерий являются патогенами. Для того, чтобы получить геномы этих бактерий, применяют стратегию «одноклеточного» секвенирования, когда независимо читается геном каждой отдельной клетки. При секвенировании массива клеток получается очень большой массив информации, который позволяет, в частности, исправить возникающие инструментальные ошибки. Секвенирование же одиночных клеток не позволяет исправлять такие ошибки. Эти редкие ошибки не представляют особой угрозы для понимания генома, поэтому их можно и пропустить, но такие ошибки становятся критичными при сборке геномов – поскольку не позволяют по точному перекрытию объединить отдельные фрагменты. Кроме того, неравномерность покорытия также приводит к ряду алгоритмических

09/2-139 от 25.02.2020

трудностей. Решению именно этой проблемы – сборки геномов при одноклеточном секвенировании – и посвящена первая часть диссертации. В работе, в частности, предложен метод итеративной сборки генома, который позволяет частично решить проблему низкого покрытия. Другой проблемой сборки геномов, в частности сборки геномов при одноклеточном секвенировании – это проблема повторов. Если длина повтора превышает длину прочтения, то принципиально невозможно восстановить последовательность генома. Однако для решения этой проблемы применяют дополнительные экспериментальные процедуры – парно-концевое чтение и метод mate-pair. Данные, получаемые в результате таких экспериментов, позволяют во многоих случаях разрешить повторы. В работе предложен эвристический подход, основанный на анализе расстояний между прочтениями в парно-концевых и mate-pair данных. Все эти подходы и алгоритмы легли в основу программы SPAdes. В конце главы представлены результаты тестирования алгоритма SPAdes, которые по большинству параметров показали преимущество по сравнению с другими методами. Использование перспективных технологий секвенирования длинных прочтений (PacBio) в комбинации с современными методами NGS показало, что они позволяют существенно повысить качество сборки. При таких подходах к секвенированию разработанная автором программа сборки генома SPAdes также показало наилучшее качество сборки по большинству критериев.

Вторая часть диссертации посвящена другой, но родственной проблеме – сборка транскриптомов de-novo. Эта задача возникает тогда, кога исследуется новый не секвенированный полностью организм, либо в случаях, когда транскриптом интенсивно редектируется, например, для инфузорий. В этих случаях применяют технологию секвенирования РНК, для того, чтобы узнать, какие гены работают, исследовать их дифференциальную экспрессию и ряда других биологических задач. Для случая, когда геном исследуемого организма известен и не слишком полимерфен применяются методы, основанные на картировании прочтений на геном. Если же геном организма не известен, то возникает задача сборки транскриптома по прочтениям. Эта задача родственна задаче сборки генома, однако есть существенные отличия. Дело в том, что, во-первых, покрытие не равномерно, а во-вторых, в эукариотических огранизмах есть процесс сплайсинга, в том числе альтернативного. Тесты показывают, что непосредственное использование геномных сборщиков дают посредственные результаты. В диссертации представлен ряд алгоритмических приемов, основанных на анализе графов, которые позволили

восстанавливать транскриптомы при наличии альтернативного сплайсинга, в частности при наличии минорных изоформ. Проведено истенсивное тестирование программы rnaSPAdes и показано, что по большинству критериев эта программа дает наилучшие результаты, в том числе и на слабо покрытых регионах.

Замечания.

1. Хотелось бы более подробного введения, предполагающего меньший уровень предварительных знаний читателя.

2. Основные главы начинаются к краткого введения. При этом не дано четкой границы, где кончается введение, а где начинаются результаты автора.

3. Представленные алгоритмы имеют ряд существенне ограничение при применении к эукариотам. Это полиморфизм генома и диплоидность. Если у человека полиморфизм составляет 1 замена на 1000 позиций, то у некоторых других видов полиморфизм может достигат 1 на 10 позиций. В этом случае предстпавленные алгоритмы будут испытывать существенные трудности.

4. Алгоритм сборки транскриптовмов используют только прочтения, полученные в эксперименте. Однако есть большой массив данных о транскриптомах других, в том числе родственных, организмов и последовательностей белков. Использование этих данных может еще повысить качество сборки.

5. Величина Support принимает дискретные значения (0 или 1), основываясь на распределении расстояний с использованием заранее определенного порога. Представляется более разумным использовать вместо этой величины байесову оценку вероятности пути и на ее основе находить путь. При этом в результате будет оценена еще и достоверность сборки, а также достоверность отдельных фрагментов сборки. Впрочем при таком подходе может оказаться затруднительным построение эффективного алгоритма.

6. стр. 24. Использование функции взвешивания дает почти всегда единственный результат, Но не ясно, почему использование функции взвешивания дает правильный результат.

7. Редакционные замечания.

   Стр. 15. Аббревиатура ONT не определена

   Стр. 19. Отмечено наблюдение, что для одноклеточного секвенирования оптимальные значения k равны 21, 33, 55 не зависимо от длины прочтения. Хорошо бы иметь хоть какую-то интерпретацию.

   Стр. 22. “ Density(e 1 , e 2 , d) = 0 при Expected(e 1 , e 2 , d) = 0”, но Expected(e 1 , e 2 , d) стоит в знаменателе!

Сделанные замечания носят редакционный или дискуссионный характер и не умаляют вклад соискателя в решение важной научной и прикладной проблемы.

Диссертация Пржибельского Андрея Дмитриевича на тему: «Алгоритмы для сборки геномов и транскриптомов» соответствует/не соответствует основным требованиям, установленным Приказом от 01.09.2016 № 6821/1 «О порядке присуждения ученых степеней в Санкт-Петербургском государственном университете», соискатель Пржибельский Андрей Дмитриевич заслуживает присуждения ученой степени кандидата физико-математических наук по специальности 03.01.09 “Математическая биология, биоинформатика”. Пункт 11 указанного Порядка диссертантом не нарушен.

Член диссертационного совета

Доктор биологических наук, кандидат физ.-мат. Наук, профессор, профессор Факультета Биоинженерии и биоинформатики МГУ им. М.В.Ломоносова А.А.Миронов

Дата 5.02.20

Заверяю:

Менеджер персонала

МИХАЙЛОВА И.М.

05.02.2020

МГУ
Отдел кадров
ФББ