ОТЗЫВ

Члена диссертационного совета на диссертацию Михеенко Аллы Александровны на тему: «Разработка вычислительных методов для анализа и визуализации эукариотических геномных сборок», представленную на соискание ученой степени кандидата физико-математических наук по специальности 03.01.09 – «Математическая биология, биоинформатика»

Развитие эффективных методов определения генетических последовательностей – секвенирования ставит новые задачи для биоинформатики. Биохимические методы секвенирования практически всегда предполагают фрагментирование цельных молекул ДНК. Следует упомянуть,что молекулу ДНК являются физически очень длинными. Например, суммарная ДНК человека имеет физическую длину около 2 метров, распределенную между 46 хромосомами, при этом это не самая длинная молекула – есть биологические объекты с гораздо более длинной ДНК. Поэтому фрагментация ДНК будет еще долгое время является необходимым этапом. После фрагментации и прочтения фрагментов возникает биоинформатическая задача сборки целого генома из фрагментов. Этой задаче было посвящено множество работ, были предложены разнообразные алгоритмы сборки и еще больше их реализаций. Все алгоритмы сборки так или иначе связаны с задачами на строках и графах. Основной проблемой сборки, кроме объема данных, геномов является проблема повторов. Наличие повторов является принципиально не преодолимой проблемой. При этом надо понимать, что, во-первых, чем длиннее геном, тем больше повторов, а, во-вторых, тандемные повторы характерны для особых уастков хромосом – центромерам и теломерам.

Настоящая работа посвящена разработке методов оценки качества сборки геномов, визуализации графов сборки а также подходам к решению проблемы тандемных повторов. Построение работы несколько отличается от традиционного -- вместо общего обзора литературы каждая глава имеет собственный мини-обзор. Диссертация начинается с краткого введения, где обосновывается постановка задач, а также указаны формальные характеристики работы.

Первая глава посвящена программному комплексу QUAST-LG позволяющему оценивать качество сборки. Комплекс производит стандартные оценки, а также производит выравнивание контигов, полученных в результате сборки, с референсным геномом. Здесь использования оригинальная идея построения цепного выравнивания с использованием специально отобранных k-меров, минимайзеров. Выравнивания сборок на референсный геном могут показывать ошибки, которые таковыми не являются, а являются биологическими вариациями, в том числе, связанными с мобильными генетическими элементами. Один из разделов главы посвящен проблеме мобильных элементов, которые могут быть восприняты, как ошибки сборки. Введено понятие теоретически оптимальной сборки, которое ставит теоретические верхние пределы показателей качества. Глава завершается разделом «результаты», где комплекс QUAST-LG применен к различным сборщикам на данных генома человека. Показано, что не существует оптимального сборщика, а по одним параметрам лучше работает один сборщик, а по другим – другой.

Во второй главе рассмотрена проблема тандемных повторов – сателлитной, мини- и микро-сателлитной ДНК. Построение выравниваний к последовательностям, насыщенным тандемными повторами является достаточно сложной задачей, особенно при использовании длинных прочтений, содержащих ошибки. Впрочем, ошибки типа замен или инделов могут иметь и биологическую природу и объясняться геномной вариабельностью. В работе по-прежнему используется техника k-меров, Однако для выбора k-меров применяются дополнительные эвристики, основанные на анализе частот встречаемости. После получения чернового выравнивания была применена процедура полировки выравнивании, устраняющие мелкие недочеты. Предложены метрики оценки качества сборки, специально настроенные на анализ последовательностей, насыщенных повторами, например, центромерных областей. Поскольку для таких последовательностей трудно предложить надежные референсные сборки, для оценки качества сборщиков была предложена модель, строящая высоко случайные повторные последовательности соласно заданному набору правил. Мне представляется такой подход наиболее адекватным в отсутствие надежных референсных последовательностей. Сравнение методов картирования показало значительное превосходство TandemMapper над другими алгоритмами.

Наконец, последняя глава посвящена визуализации сборок. В частности, представлен визуализатор Icarus. Идеология этой программы аналогична геномным браузерам, что удобно для пользователя. Однако в отличие от существующих программ, эта программа ориентирована именно на задачи сборки геномов и позволяет показывать особенности именно сборок. Другая программа, AGB, ориентирована на представлениях графов сборки. Эти графы имеют тысячи и десятки тысяч вершин, поэтому их графическое представление представляет собой достаточно сложную задачу. Был предложен метод разбиения больших графов на подграфы, основанный на поиске минимальных разрезов. Преложенное представление графов позволяет показывать достаточно большой спектр дополнительной информации, например, наличие повторностей, сбаллансированность вершин и пр. Эти данные позволяют указывать разнообразные возможные ошибки сборки.

Вопросы и замечания:

1. Одним из способов оценки качества сборки является сравнение сборки с референсным геномом. Однако, референсный геном был также получен в результате сборки фрагментов с помощью какого-то сборщика. Поэтому наивный читатель может предположить, что оценка качества сборки сравнением с референсной сборкой это всего лишь сравнение одного (тестируемого) сборщика с некоторым «стандартным» сборщиком.

2. Геномы, как правило, диплоидны, при этом иногда материнский набор хромосом может отличаться от отцовского заметным количеством перестроек. Более того раковые геномы зачастую также имеют значительное количество перестроек. Референсный геном, как правило, гаплоидный. Последние сборки не являются непрерывными и допускают вариации (графовое представление геномов). Как предложенные методы могут учитывать эту вариабельность.

3. Не во всех литературных ссылках указан год издания (напр. [44, 45])

4. В списке литературы разные публикации представлены по-разному, например ссылка [12] начинается с названия статьи, а [13] со списка авторов. Не говоря уже о таких мелочах, как не одинаковое использование курсива.

5. Нестандартные подписи к рисункам. Обычный формат: *Рисунок 3.14. Заголовок всего рисунка. А – то, что показано на панели А. Б – вверху: ..., внизу: ....*

6. п.2.3: Хотелось бы увидеть график распределения покрытия на симулированных данных

7. Таблица 6 – в заголовке указано, что лучшие значения выделены жирным шрифтом, однако в таблице нет такого выделения.

Высказанные замечания носят скорее редакционный характер и не умаляют достоинств работы. Все представленные программные продукты доступны по указанным в работе адресам и сопровождаются подробной документацией, примерами использования, протоколами изменений. Основные результаты данной работы описаны в четырех статьях, опубликованных в журналах, индексируемых в базах данных Web of Science Core Collection и Scopus (все журналы в настоящее время ранжируются как Q1). Отмечу, что во всех представленных публикациях соискатель отмечен как первый автор. Следует указать, что соискатель имеет еще некоторое количество научных работ, в том числе весьма значимых, не вошедших в диссертацию.

Диссертация Михеенко Аллы Александровны на тему: «Разработка вычислительных методов для анализа и визуализации эукариотических геномных сборок» соответствует основным требованиям, установленным Приказом от 01.09.2016 № 6821/1 «О порядке присуждения ученых степеней в Санкт-Петербургском государственном университете», соискатель Михеенко Алла Александровна 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 физико-математических наук, профессор, профессор Факультета биоинженерии и биоинформатики МГУ им. М.В.Ломоносова — А.А.Миронов

3 декабря 2020