ОТЗЫВ

члена диссертационного совета на диссертацию Банкевича Антона Викторовича на тему: «Модели, численные методы и комплекс программ для сборки геномов из нестандартных данных секвенирования», представленную на соискание ученой степени кандидата кандидата физико-математических наук по специальности Специальность 05.13.18 —«Математическое моделирование, численные методы и комплексы программ»

.Биоинформатика стала формироваться как отдельное направление биологической науки в самом начале 80-х годов 20-го века. Мощным толчком для этого развития послужило создание эффективных и достаточно рутинных методов определения нуклеотидных последовательностей. Уже самые первые работы по секвенированию небольших плазмид и фаговых геномов были основаны на методе дробовика, когда геном разбивается на достаточно короткие фрагменты, которые потом собираются в полную последовательность по перекрытию. Самые первые ппрочитанные и собранные последовательности составляли несколько тысяч нуклеотидов и собирались из десятков фрагментов. Задача сборки в то время не представляло особой проблемы даже для весьма маломощных компьютеров и решалась достаточно примитивными методами, часто вручную. Развитие технологий и переход к более мощным секвенаторам и к прочтению больших геномов поставил новые вполне серьезные алгоритмические задачи, когда геном надо собирать из миллионов фрагментов. Этот технологический скачок потребовал разработки принципиально новых подходов. Одним из таких подходов был переход от Гамильтоновой задачи сборке с помощью графов де Брюйна. В результате появилось семейство геномных сборщиков, основанных на таком подходе. При традиционном подходе секвенируется одновременно большое количество клеток, что позволяет эффективно решать проблему ошибок секвенирования. Однако дальнейшее развитие технологий позволило перейти к секвенированию геномов индивидуальных клеток и проблема ошибок секвенирования вновь возникла. Другим развитием методов секвенирования является баркодирование, которое должно облегчить решение проблемы повторов. Диссертационная работа Антона Викторовича Банкевича посвящена развитию методов сборки геномов для одноклеточного секвенирования и применения баркодирования.

Работа состоит из введения и четырех глав. Во введении дается краткое описание проблемы, дана постановка задачи и даны основные формальные сведения о работе. Первая глава посвящена подробному анализу современного состояния проблемы геномной сборки. При этом рассмотрены как алгоритмические аспекты проблемы, так и вероятностные модели процесса секвенирования.

Вторая глава посвящена проблеме сборки геномов при одноклеточном секвенировании. Рассмотрены подходы к упрощению графа де Брюйна. При одноклеточном секвенировании весьма актуальой является проблема ошибок секвенирования, в том числе проблема химерных прочтений. Для решения этой проблемы используются различные алгоритмические подходы с использованием, в том числе, теории потоков в сетях.

Третья глава посвящена разработке алгоритмов для новой технологии баркодного секвенирования. Для решения проблемы повторов раньше применялся экспериментальный подход,

09/2-93 05 04.06 18

основанный на субклонировании в BAC и в коспиды, которые затем секвенировались, собиральсь отдельно, а затем собирался полный геном. Новая технология баркодирования по сути является иммитацией метода субклонирования и позволяет в одном эксперименте получить прочтения, адресованные к разным участкам генома. В диссертации разработаны алгоритмы и методы обработки таких данных. Здесь также есть несколько проблем, связанных с химеризмом и ошибками, решением которых занимается автор. Одной из проблем секвенирования геномных сообществ (метагеномов) является оценка полной длины всех геномов в сообществе. Это необходимо для понимания того, насколько полученные данные хорошо покрывают все разнообразие организмов. Если проблема определения длины генома для отдельного организма решается достаточно просто, то в соообществе присутствуют клетки с существенно разной представленностью. Для решения этой проблемы разработан математический аппарат, позволяющий оценить полноту покрытия. Отмечу, что распределение представленности геномов в сообществе имеет тяжелый хвост. Поэтому решить эту проблему точно достаточно трудно, если вообще возможно. Автор предлагает методику приближенной оценки, пренебрегая совсем редкими геномами.

Наконец, четвертая глава посвящена программной реализации предложенных алгоритмов и ее тестированию.

По работе можно высказать ряд мелких замечаний:

1. Во введении представляется лишним упоминание о ранних математических динамических моделях – они не имеют отношения. С другой стороны, целая эпоха, связанная с секвенированием методами Максама-Гильберта и Сенгера упомянута весьма поверхностно, а ведь именно Сенгером были секвенированы первые бактериофаги бактерии и там как раз появились первые, достаточно наивные по современным представлениям, сборщики.
2. Стр13. «Мы будем считать, что замены в прочтениях Illumina производятся независимо и случайно в каждом нуклеотиде с некоторой вероятностью p», что не совсем верно, поскольку качество прочтения падает к концу прочтения.
3. Стр. 15 «геномные фрагменты, которым соответствуют различные контиги, не пересекаются;» на самом деле допускается пересечение на некоторое заданное количество нуклеотидов – предельный случай есть два контига, проекции которых на геном пересекаются на 1 нуклеотид. Разумеется, при сборке у нас нет никаких оснований их соединять.
4. Стр. 35. Rep занимает только незначительную часть генома. Вычисление длины Rep ... что для 95% протестированных геномов длина Rep не превышает 90% длины генома.
5. Стр. 68. это похожие бактериальные подвиды (стрейны) → штаммы