

ОТЗЫВ

члена диссертационного совета на диссертационную работу Банкевича Антона Викторовича «МОДЕЛИ, ЧИСЛЕННЫЕ МЕТОДЫ И КОМПЛЕКС ПРОГРАММ ДЛЯ СБОРКИ ГЕНОМОВ ИЗ НЕСТАНДАРТНЫХ ДАННЫХ СЕКВЕНИРОВАНИЯ», представленной на соискание ученой степени кандидата физико-математических наук по специальности 05.13.18 — «Математическое моделирование, численные методы и комплексы программ»

Рецензент получил удовольствие от прочтения диссертационной работы Антона Банкевича, посвященной разработке методов и программного обеспечения для сборки геномов из различных исходных типов данных высокопроизводительного секвенирования, в том числе секвенирования геномов единичных бактериальных клеток, а также баркодированных данных, позволяющих осуществлять *in silico* сборку длинных прочтений по 1-10Кб.

Коллектив Павла Аркадьевича Певзнера является одним из признанных мировых лидеров в области разработки алгоритмов и программного обеспечения для сборки геномов, а также для решения ряда других задач в области анализа данных высокопроизводительного секвенирования.

Вклад Антона Банкевича в развитие этого направления в целом, и в развитие специализированных алгоритмов и программного обеспечения для решения обозначенных выше задач бесспорен. Эти методы и программы востребованы в современных исследованиях, все плотнее использующих анализ геномов единичных клеток и технологии молекулярного баркодирования, и, очевидно, продолжат непрерывную эволюцию, параллельно с развитием молекулярных технологий, позволяющих получать такие данные.

Диссертация состоит из введения, четырёх глав и заключения. Во введении автор дает представление о цели данной работы и поставленных задачах. В первой главе описываются подходы к моделированию данных секвенирования, формулируются задачи сборки геномов и упрощения графов де Бройна. Вторая и третья главы посвящены основному содержанию работы - сборке данных одноклеточного секвенирования, баркодной сборке синтетических прочтений, численному методу оценки длины генома. В четвертой главе описывается разработанное программное обеспечение, а также результаты сравнительного анализа эффективности алгоритмов сборки. В целом результаты работы изложены достаточно внятно, и позволяют оценить качество проделанной работы и результирующих алгоритмов.

Пожалуй, основным замечанием рецензента к работе является нехватка иллюстративного материала, затрудняющего восприятие для читателя, не вовлеченного непосредственно в задачи сборки геномов. Так, исходно к описанию принципа графов Де Брюйна не прилагается понятного развернутого рисунка, из которого несведущий наблюдатель мог бы быстро представить себе картину происходящего - граф, переход к сжатому графу, типичные артефакты.

Для лучшего понимания повествования также часто не хватает конкретики. Лишь на 30ой странице впервые приводится длина k-мера, используемого для построения графа де Брюйна, и при этом не приводится обоснования выбора этой длины.

Стр. 13: «Мы будем считать, что замены в прочтениях Illumina производятся независимо и случайно в каждом нуклеотиде с некоторой вероятностью p ($p = 0,01$).» - алгоритмы, разрабатываемые автором, по всей видимости не учитывают информации о качестве каждого прочтенного нуклеотида, заложенного в выходной результат секвенирования Illumina (?). Вероятно, это оправдано с точки зрения рационального расходования вычислительных мощностей при геномной сборке, однако, по меньшей мере, требует комментария.

Из рисунка 1.2 можно заключить, что среднее покрытие обычно превышает 100, что не всегда верно. Такое покрытие удорожает геномное секвенирование, ограничивая практическое применение. Обсуждение компромиссов при выставлении трешхолда отсечения покрытия было бы полезным.

Из повествования остается не вполне ясным, почему фильтрацию низко-покрытых k-меров принято проводить после построения графа Де-Брюйна, а не до.

Стр. 28: "Химерические прочтения и рёбра иногда встречаются и в обычных библиотеках прочтений (без применения амплификации), но не являются проблемой, поскольку имеют очень низкое покрытие и могут быть отфильтрованы. Однако, химерические прочтения, возникающие при одноклеточном секвенировании, вызваны ошибками в процессе амплификации: если при копировании ДНК появилось химерическое соединение, оно копируется на последующих итерациях амплификации". Насколько известно рецензенту, на сегодняшний день при геномном секвенировании всё ещё часто используют амплификацию на этапе приготовления библиотек. Соответственно, можно ожидать наличия химерических продуктов, амплифицированных с редких исходных событий. В то же время, amplification-free технологии также производят химеры - при лигировании, и не защищены от неравномерного

покрытия. Соответственно, фильтрация низко-покрытых, потенциально химерических к-меров может приводить к потерям информации. Возможная дупликация химер на твердой фазе Illumina усложняет игру.

Рис. 2.1 вызывает много вопросов. Представлены данные для одной конкретной бактерии или многих? Если для многих, то представлены воспроизводимые химеры? Если для одной, каким образом исключена возможность геномных перестроек, приводящих к реальному формированию ребер показанных как химерные?

Стр. 30: "Количество химерических рёбер не превосходит 150 на каждом наборе данных" - есть ли воспроизводимые химерические ребра?

Главы 2.4, 2.5. Не хватает иллюстративного материала.

Рис 3.2 справа. Не вполне ясно как может происходить такое событие, с учетом того что исходно лигируются адаптеры, закрывающие концы. Адаптер оказался бы внутри если смена матрицы происходила бы по прочтении всей молекулы. Самозатравка с недостроенной молекулой - наиболее вероятное объяснение для длинных фрагментов, но требует соответствующей иллюстрации и описания. Возможно также, что показанная химерическая молекула является результатом ошибки транспозазы в ходе фрагментирования и вставления адаптеров Nextera. Понимание фундаментальных причин ошибок, возникающих в ходе пробоподготовки, является, с точки зрения рецензента, важной основой для разработки эффективных алгоритмов коррекции ошибок в ходе анализа данных.

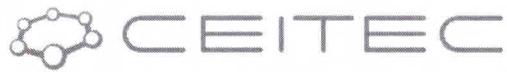
Работа не лишена небольших неточностей и просто опечаток, а возможно и не просто опечаток: Так, выражение "новыЯ" модели геномной сборки можно интерпретировать как сентиментальную ссылку к дореформенной русской орфографии.

"Недавно открытый метод одноклеточного секвенирования" - за вычетом эзотерических объяснений (на чем рецензент не настаивает), обычно принято писать, что методы разрабатывают, а не открывают.

к сожалению, на рисунке нет

Рис 3.2: «конец синего фрагмента» - к сожалению, на рисунке нет ничего синего.

Перечисленные выше недочеты не снижают ценности работы.



Central European Institute of Technology
BRNO | CZECH REPUBLIC

Опубликованные статьи отражают содержание диссертации. Диссертация Банкевича Антона Викторовича на тему: «МОДЕЛИ, ЧИСЛЕННЫЕ МЕТОДЫ И КОМПЛЕКС ПРОГРАММ ДЛЯ СБОРКИ ГЕНОМОВ ИЗ НЕСТАНДАРТНЫХ ДАННЫХ СЕКВЕНИРОВАНИЯ» является законченной научно-квалификационной работой, которая по своему содержанию, уровню проведенных исследований, актуальности выбранной темы, достоверности полученных результатов и выводов, их научной и практической ценности в полной мере отвечает всем требованиям, установленным Приказом от 01.09.2016 № 6821/1 «О порядке присуждения ученых степеней в Санкт-Петербургском государственном университете», соискатель Банкевич Антон Викторович заслуживает присуждения ученой степени кандидата физико-математических наук по специальности 05.13.18 — «Математическое моделирование, численные методы и комплексы программ».

Член диссертационного совета,
заведующий лабораторией «Adaptive Immunity Group»,
Central European Institute of Technology,
Университет им. Масарика, Брно, Чехия
д.б.н., профессор Дмитрий Михайлович Чудаков
1 июня 2018 г.

Д.М. Чудаков