ОТЗЫВ

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

С появлением технологий секвенирования второго поколения в начале 2000-х, которые позволили значительно снизить цену чтения генома, задача сборки нуклеотидных последовательностей приобрела широкую популярность. За прошедшие 20 лет было создано несколько десятков различных программ для сборки геномов и транскриптомов с нуля. Однако, постоянно обновляющиеся протоколы секвенирования и меняющиеся характеристики данных требуют поддержки существующего, а в некоторых случаях и разработки полностью нового программного обеспечения. В частности, разработанный в 2007 году метод полногеномной амплификации, позволивший секвенировать бактериальные геномы всего лишь по одной клетке, оказался не по зубам существующим сборщикам, и требовал разработки новых алгоритмов. Что же касается задачи сборки транскриптома, то несмотря на то, что в последние годы можно выделить ассемблер Trinity, занимающий лидирующие позиции в этой области, даже его результаты зачастую оставляют желать лучшего.

В диссертации Пржибельского А.Д. предложены новые методы для сборки геномов по данным, полученным по одной клетке, а также по стандартным транскриптомным данным. В части посвященной геномной сборке, упор сделан на задаче разрешения повторов по парно-концевым прочтениям. Текст диссертации содержит связанное повествование о создании нескольких методов сборки на единой алгоритмической базе, что создает целостное впечатление о проделанной работе. Среди минусов можно выделить достаточно краткое введение в область, а также некоторая размытость границ между описаниями алгоритмов, созданных автором, и существовавших ранее методик, в том числе разработанных его коллегами.

Качество проделанной работы отражено в двух секциях с результатами, которые демонстрируют существенное преимущество разработанных методов над имеющимися аналогами. О важности и новизне диссертации можно судить по информации, представленной в введении и заключении. В них приведены примеры использования созданных сборщиков в реальных биоинформатических проектах, а также их высокий уровень цитирования, указывающий на широкое распространение ассемблеров SPAdes и rnaSPAdes.

В порядке дискуссии хотелось бы задать автору несколько вопросов и высказать некоторые пожелания.

09/2-141 от 27.02.2020

- В тексте диссертации подчеркивается важность правильного выбора длины к-мера при сборке. Методика определения оптимального значения k подробно описана для транскриптомных сборок (стр. 56), однако практически отсутствует аналогичное описание для сборки геномов (стр. 23). Каким образом выбирались значения длин к-меров для сборки геномных данных (21,33,55)?
- Один из ключевых алгоритмов, разработанных автором диссертации — алгоритм для разрешения повторов exSPAnder. Несмотря на подробное описание алгоритма текстом, отсутствует псевдокод или блок-схема, которые могли бы заметно упростить его понимание.
- В диссертации кратко описано каким образом SPAdes разрешаст тандемные повторы при помощи парных ридов (стр. 28-29). Утверждается также, что при предполагаемом количестве копий больше 1 алгоритм вставляет последовательность из символов N, то есть не разрешает повтор точно. Можно ли усовершенствовать эту методику, в том числе используя другие типы данных секвенирования?
- В таблицах с результатами как геномных, так и транскриптомных сборок, как правило, отсутствует описание организма (представлено только научное наименование). Если *E.coli, H.sapiens* у всех на слуху, остальные виды могут не быть известны широкому кругу.
- Химерические ребра при сборке РНК именуются шпильками (стр. 55), что является плохим выбором термина, так как шпилька — частый элемент вторичной структуры РНК (и не имеет отношения к химическим ребрам в графе де Брюйна).
- Стр. 69: Ген считается экспрессирующимся если он имеет нуклеотидное покрытие не менее 5. Каким образом была выбрана эта отсечка?
- Стр. 41: Сочетаются русские и английские аббревиатуры (kbp и кбп).
- Стр. 17: опечатка, наличиЯ.

Возникшие вопросы ни в коей мере не умаляют высокого качества диссертации.

Диссертация Пржибельского Андрея Дмитриевича на тему: «Разработка алгоритмов для сборки геномов и транскриптомов» 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.

Председатель диссертационного совета
доктор биологических наук, профессор кафедры цитологии и гистологии, Биологического факультета, СПбГУ
Дата

12.02.2020

Подгорная О.И.