Skoltech

Skolkovo Institute of Science and Technology

ОТЗЫВ

члена диссертационного совета
на диссертацию Аллы Александровны Михеенко
«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.

В диссертации А.А. Михеенко описаны разработанные ею модули оценки качества сборок, вошедшие в состав программного комплекса QUAST: TandemMapper для картирования ридов в областях с повторами; QUAST-LG, улучшенный вариант программы для оценки качества сборок больших геномов со сложной структурой; TandemQUAST для оценки сборок тандемных повторов; Icarus и AGB для визуализации сборок и графов сборок, соответственно.

Все эти программы, несомненно, полезны; в соответствующих разделах автор убедительно показывает, что разработанные ею программы превосходят по качеству и достоверности результатов имеющиеся аналоги. Они актуальны: прогресс в технике секвенирования, выразившийся, в частности, в увеличении длины читаемых фрагментов дает новые возможности, в частности, определять последовательность теломерных областей, но реализовать эти возможности невозможно без соответствующего программного обеспечения. Актуальность и практическая важность работы А.А. Михеенко доказаны востребованностью разработанных ею программ, в частности, в проекте Telomere-to-Telomere.

Следует подчеркнуть, что вклад А.А. Михеенко в разработку программ является решающим, что доказывается тем, что она является первым автором всех основных публикаций: по теме диссертации опубликовано пять статей в ведущем профильном журнале по алгоритмической биоинформатике *Bioinformatics*. Я также считаю нужным отметить, хотя это впрямую не относится к теме отзыва, что работа А.А. Михеенко этими публикациями не исчерпывается: помимо статей, составивших основу диссертации, она участвовала в целом ряде публикаций по анализу масс-спектрометрических данных.

№ 09/2-02-744 от 29.12.2020

Тем самым, замечания к диссертации носят по большей части стилистический и редакторских характер. Пожалуй, единственное содержательное, что мне показалось странным, — это принципиальный отказ использовать идентификацию повторов для фильтрации и анализа мисассемблов (с мотивировкой, что это «выходит за рамки данной работы», стр. 24). Автор лишает себя мощного средства, при том, что даже относительно примитивный фильтр на повторы (с учетом как известной структуры референсного генома, так и локального покрытия) могло бы оказаться вполне полезным.

С общей структурой работы имеется та же проблема, что с любой диссертацией, написанной по опубликованным статьям: практически отсутствует общее введение и обсуждение, которое бы придавало цельность исследованию. Вместо этого каждая глава начинается с собственного обзорного раздела. В первой главе этот обзор местами производит странное впечатление, например, в §1.1.3 зачем-то приводятся сведения из учебника молекулярной биологии (про опероны, экзоны и интроны), причем не вполне точные (например, «У прокариот несколько генов могут быть организованы на одном опероне: они расположены рядом и управляют ферментами, которые осуществляют последовательные или близкие реакции синтеза» — опероны, кодирующие рибосомные белки, ABC-транспортеры или, скажем, компоненты жгутика явно не покрываются этим описанием). В §1.2.4 сказано, что наиболее активные и распространенные повторы в геноме человека — это LINE-1: а как же ALU? С другой стороны, в первой главе отсутствует обсуждение

Часто «мотивировочные» пассажи, которым место в обзоре, проникают в разделы с результатами, например, второй абзац в §1.2.3. Аналогично, местами плохо разграничены результаты и их обсуждение, например, последний параграф на стр. 32.

Странное впечатление производит глоссарий — непонятно, по какому принципу выбирались термины, которые следует объяснить: скажем, *мутация* есть, а *юнит* — нет. *Семя* как русский аналог *seed* — это плохой выбор, адекватнее *затравка*. Вместо *прикладывание* (*mapping*) обычно говорят *картирование*.

Еще о терминах — на стр. 18 читаем: «Выбирается самый маленький (в соответствии с предварительно определенным порядком)», — тут сходу подобать термин не удается, (*ранний? старший?*), но всегда можно переформулировать весь пассаж…

Некоторые формулировки формально неточны, например: «*Nx* — это длина контига, для которой все контиги этой или большей длины составляют не менее *x*% длины сборки» (стр. 16): пропущено слово «*максимальная* длина контига». Местами изложение недостаточно формально. Скажем, ни из текста §1.2.4, ни из подписи к рис. 2 не следует, все ли логически возможные случаи перечислены.

Пара еще более частных замечаний.

Стр. 8: «в статье QUAST-LG [14]» — тут должна быть ссылка [12]; [14] — это статья про MetaQUAST, которая в текст не вошла.

Рис. 1: конечно, «сходство» нетранзитивно, но все-таки следовало бы ожидать, что левая часть фрагмента E будет выравниваться с правой частью фрагмента, находящегося под D.

Замечание формальное, но, тем не менее, необходимое. Список литературы оформлен крайне небрежно: в нескольких разных форматах, причем во многих пунктах отсутствует часть библиографической информации (например, 3 —библиографическая информация отсутствует; 4 — есть только URL; 5 — не указан год...). Огорчительно, что этим дефектом обладают даже описания собственных работ автора, по которым защищается диссертация: 10 — нет года, 11 — зачем-то указан месяц, 12 — нет тома и страниц, 13 — нет года, 14 — указан месяц. У меня возникло подозрение, что это могло быть вызвано интеграцией списков из нескольких публикаций, но нет: все публикации сделаны в одном журнале, стало быть, исходные списки, коль скоро таковые существовали, были выдержаны в одном формате. Более того, ссылка 1 сохраняет следы форматирования в TEX-е: тем более удивительно, как вообще можно было добиться такого эффекта.

Еще раз укажу, что сделанные замечания не влияют на мою общую высокую оценку представленной к защите работы.

Диссертация А.А. Михеенко «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». Пункт 11 указанного Порядка диссертантом не нарушен.

Алла Александровна Михеенко заслуживает присуждения ученой степени кандидата физико-математических наук по специальности 03.01.09 — Математическая биология, биоинформатика.

12.12.2020 М.С.Гельфанд

Skoltech