Skoltech

Skolkovo Institute of Science and Technology

ОТЗЫВ

члена диссертационного совета на диссертацию Дмитрия Юрьевича Антипова на тему: «Разработка алгоритмов для специальных задач сборки геномов», представленную на соискание ученой степени кандидата физико-математических наук по специальности «1.5.8. Математическая биология, биоинформатика»

Предварительное замечание. *На рецензию представлены сразу два текста, русский и английский, — видимо, это какая-то местная традиция (я видел подобное и в других диссертациях, защищенных в СПбГУ). Я буду рассматривать русский текст.*

Актуальность диссертации не вызывает сомнений — сборка геномов является первым (точнее, первым нетривиальным) и ключевым этапом обработки геномных и метагеномных данных, от успешного прохождения которого критически зависит дальнейший биологический анализ. Сразу следует отметить, что написанные Д.Ю.Антиповым программы широко используются в практике, в т.ч. другими группами (на момент написания рецензии hybridSPAdes использована в >270, plasmidSPAdes — в >180 работ). При независимом тестировании (Arredondo-Alonso et al., Microb. Genom., 2017) программа plasmidSPAdes показала лучший результат (чувствительность 82%, специфичность 75%) из четырех сравниваемых программ, хотя и уступила ряду более поздних программ, во всяком случае, по утверждениям их авторов (MOB-recon, Roberrson & Nash, Microb. Genom. 2018; SAUTE, Souvorov & Agarwala, BMC Bioinformatics 2021). plasmidSPAdes показала наилучшие результаты при сборке коротких плазмид, хотя уступила MOB-suite при реконструкции длинных (Paganini et al., Microorganisms 2021). Аналогично, ряд методических работ рекомендуют hybridSPAdes для работы со смешанными данными секвенаторов второго и третьего поколений (George et al., Microb. Genom. 2017; Brown et al., Sci. Rep. 2021). Собственно говоря, на этом отзыв можно было бы и закончить.

Работа построена по традиционной схеме: первая глава содержит обзор, вторая и третья — описание созданных автором программ: соответственно, hybridSPAdes для гибридной сборки с использованием длинных и коротких чтений и plasmidSPAdes и metaplasmidSPAdes для сборки плазмид по данным секвенирования прокариотических геномов и метагеномов. В заключении обсуждается дальнейшее развитие разработанных подходов и алгоритмов.

Обзор носит достаточно формальный характер. В нем приведены самые общие сведения о решаемых задачах, даны основные алгоритмические понятия и описан комплекс SPAdes, в который как самостоятельные модули входят написанные автором программы. Обсуждение других программ, решающих аналогичные задачи, содержится во вступительных разделах соответствующих глав.

Во второй главе описан алгоритм сборки с использованием длинных чтений. В принципе, изложение достаточно подробное и логичное, однако автору не вполне удалось решить проблемы разделения общей конструкции и отдельных эвристических и алгоритмических улучшений: из-за этого не всегда понятно, какие точки в изложении являются критическими. Приведены результаты сравнения качества сборки hybridSPAdes и рядом современных алгоритмов. Тут следует отметить сложность, с которой сталкиваются все разработчики программ: область быстро развивается, и программы, бывшие лучшими в момент публикации, часто перестают быть таковыми ко времени защиты. Хорошо, что hybridSPAdes все еще вполне конкурентоспособен. (Вторая стандартная проблема, что меняется объем и качество данных, автором упомянута, но подробно не обсуждена.)

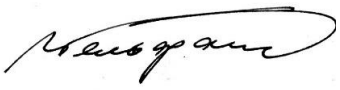
Аналогично построена и третья глава, разве что в ней изложены сразу три программы: базовая plasmidSPAdes, metaplasmidSPAdes, решающая задачу сборки плазмид в метагеномах, и plasmidVerify для оценки происхождения контигов. Последняя программа стоит несколько особняком, поскольку она использует не формальные комбинаторно-алгоритмические приемы, а анализирует функции генов, содержащихся в контиге. Представляется, что схему тестирования этой программы (и/или всего комплекса) стоило бы описать чуть подробнее; скорее, даже, существенны не столько схемы тестирования, сколько описание типичных ситуаций, в которых могли бы применяться указанные программы (при том качестве предсказаний, которые они дают), и примеры биологических выводов, которые были бы сделаны с использованием этих программ (включая обсуждение робастности этих выводов относительно не вполне стопроцентного качества сборок и предсказаний). Следует отметить, что в этой главе автор пошел по минималистскому пути, отослав читателя за более подробной информацией к собственной опубликованной статье — я не уверен, что это правильно: всё-таки, обычно предполагается, что диссертация самодостаточна, и, если местные правила не разрешают защиту по набору публикаций, существенные факты из публикаций должны быть в диссертацию перенесены. С другой стороны, можно понять и нежелание автора переводить и переписывать уже опубликованный текст.

Можно высказать ряд мелких редакционных, скорее даже корректорских замечаний. Кажется, что в тексте диссертаций исчез типографский знак «тире» — всюду, где читатель ожидал бы его увидеть, имеется очень длинный пробел, но самого тире нет. Видимо, это какой-то сбой верстающей программы. Второе мелкое редакционное замечание — текст несвободен от англицизмов. И если «индел», видимо, уже является термином, который заменить нечем, то без «мисматча», «мизассемблов» и «полишинга» явно можно было обойтись. «Химерические» чтения — лучше сказать «химерные». В тексте также не хватает некоторого количества запятых, в частности, при деепричастных оборотах.

Автором опубликовано десять статей в ведущих профессиональных журналах, в пяти из которых представлен материал рассматриваемой диссертации; из этих пяти статей в трех ключевых он является первым автором; эти статьи опубликованы в журналах первого квартиля. Результаты доложены на международных конференциях высокого уровня, в т.ч RECOMB и MCCMB.

Диссертация Дмитрия Юрьевича Антипова на тему: «Разработка алгоритмов для специальных задач сборки геномов» соответствует основным требованиям, установленным Приказом от 19.11.2021 № 11181/1 «О порядке присуждения ученых степеней в Санкт-Петербургском государственном университете», соискатель заслуживает присуждения ученой степени кандидата физико-математических наук по специальности «1.5.8. Математическая биология, биоинформатика». Пункты 9 и 11 указанного Порядка диссертантом не нарушены.

Член диссертационного совета, доктор биологических наук кандидат физико-математических наук, профессор, вице-президент Сколковского института науки и технологий по биомедицинским исследованиям



Михаил Сергеевич Гельфанд

1.1.2022