Skoltech

Skolkovo Institute of Science and Technology

ОТЗЫВ

члена диссертационного совета на диссертацию Татьяны Евгеньевны Дворкиной на тему: «Разработка алгоритмов для анализа графов геномной сборки и геномных сборок», представленную на соискание ученой степени кандидата физико-математических наук по научной специальности 1.5.8 «Математическая биология, биоинформатика»

Предварительное замечание. По неведомой для меня причине текст диссертации существует сразу на двух языках, русском и английском. Насколько я могу судить, просмотрев названия разделов и рисунки, содержание этих вариантов идентично. Нижеследующий отзыв следует русскоязычному варианту. Для удобства замечания пронумерованы в квадратных скобках.

В диссертации Т.Е. Дворкиной описаны важные и полезные инструменты, позволяющие превратить неполностью собранные геномы в полезный ресурс для содержательного биологического анализа. Для создания этих программ потребовалась разработка принципиально новых алгоритмов, поэтому естественно, что диссертация представлена на соискание степени кандидата именно физико-математических наук.

Структура работы в основном следует структуре опубликованных статей. Эти статьи были опубликованы в 2020–2022 гг. в хороших журналах (BMC Bioinformatics, Microbiome, Bioinformatics, Genome Research); в четырех из них Татьяна Евгеньевна является первым автором (в некоторых университетах, в частности, в Сколтехе и на факультете компьютерных наук ВШЭ этого было бы достаточно для защиты по совокупности работ).

Следствием такого решения является то, что в тексте имеются три независимых введения, три методические части и т.п. Этот отход от традиционной структуры является оправданным, поскольку в каждой из вводных подглав излагается именно тот научный контекст, который необходим для понимания и оценки результатов. Взаимосвязь отдельных проектов раскрыта во Введении и в Заключении. Вторым следствием является то, что [1] название кажется до некоторой степени очень общим — оно скорее подошло бы для докторской, а не кандидатской диссертации.

В главе 1 описаны алгоритмы и программы выравнивания на граф сборке в нуклеотидном (длинные прочтения) и аминокислотном (белковые последовательности или модели скрытых марковских цепей) вариантах. Следует отметить, что [2] для выравнивания аминокислотных последовательностей реализованы только линейные (не аффинные) штрафа за вставки/удаления (§ 1.3.2); это представляется не очень реалистичным. [3] Можно заметить также некоторое лукавство автора при описании

результатов сравнения собственной программы SPAligner и альтернативной GraphAligner (§ 1.4.1): способность SPAligner выравнивать больше прочтений проявляется только в основном варианте, требующем заметного времени (на один-два порядка больше, чем альтернативы), а при ускорении время счета сокращается (хотя по-прежнему заметно превосходит таковое для обоих вариантов GraphAligner), но и чувствительность падает и во многих случаях становится меньше, чем у GraphAligner.

В главе 2 этот подход применяется для поиска новых генов инсектицидных белков в метагеномах. Полный конвейер, включающий идентификацию открытых рамок считывания, выравнивание и кластеризацию результатов, сначала тестируется на модельных примерах, а затем применяется к реальным данным.

Очень интересна и важна глава 3, в которой фактически предложен новый язык для описания повторяющихся последовательностей с иерархией, таких как центромеры и теломеры.

Следует особо подчеркнуть, что все разработаные программы широко используются не только в группе, где они были созданы, но и в других лабораториях, а алгоритмы получили дальнейшее развитие в ряде только что опубликованных статей.

Хотя текст диссертации в целом написан логично и понятно, в некоторых биологических описаниях встречаются неточности. Так, [4] не стоит называть хроматиды «половинками хромосомы» (стр. 6) – на самом деле, это две копии молекулы ДНК до разделения центромер; [5] аналогично, нехорошо писать, что в процессе клеточного деления происходит «расщепление ДНК с помощью кинетохора» (стр. 50). Писать, что [6] растительные пестициды — это обычно «пестицидные белки из генов, встроенных в ДНК растений» (стр. 33), во-первых, неточно (что такое «белки из генов»? кем они были «встроены»), а во-вторых, и не очень правильно: это верно только для генно-модифицированных организмов, и не учитывает природные вторичные метаболиты, синтезируемые растениями, например, классический пиретрум. [7] В правой панели рис. 2.1 стоило бы включить не все белки UniProt, а только бактериальные. Кроме того, [8] не очень хорошо описана процедура тестирования выравнивания аминокислотных последовательностей на граф сборки § 1.4.2, в частности, не вполне ясно, каковы были критерии качества. [9] Стоило бы указать, какие кодоны рассматривались в качестве стартовых в § 2.2 – только ATG (AUG) или также GTG и TTG: ясно, что можно любым способом, но интересно, как это было реализовано и есть ли у пользователя возможность на это влиять. Кстати, и для стоп-кодонов можно было предусмотреть некоторую вариативность, учитывая неуниверсальность генетического кода.

Остальные замечания к работе носят почти исключительно редакционный и стилистический характер.

[10] «vg заточен на выравнивание коротких прочтений» (стр. 18) выглядит как-то очень уж жаргонно.

[11] «Тестирования» (во множественном числе) в заголовке § 1.4.2 тоже странно звучит.

[12] «Компл<u>и</u>ментарное направление в графе» (стр. 37) — это, конечно, «компл<u>е</u>ментарное направление».

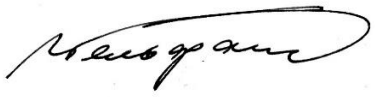
[13] «Металлофосфуэстеразы» (стр. 48-49) — это, видимо, опечатка, а вот «дисульфидные редуктазы» в том же предложении — уже не очень уверенный перевод с английского.

[14] «Юниты повтора» (стр. 53) — даже и не перевод, а прямая транслитерация. Известная уроженка Санкт-Петербурга В.И. Матвиенко, занимающая важный государственный пост, говорила: «Есть же аналоги у русского языка. Для чего мы засоряем наш красивый, замечательный, уникальный язык такими разными словами? Это надо и на законодательном уровне посмотреть» (news.rambler.ru/politics/48871257/).

[15] «При репликации … приклеивалась вторая половина повтора» (стр. 69) — вряд ли «приклеивалась», видимо, механистически происходило что-то другое.

Диссертация Татьяны Евгеньевны Дворкиной на тему: «Разработка алгоритмов для анализа графов геномной сборки и геномных сборок» соответствует основным требованиям, установленным Приказом от 19.11.2021 № 11181/1 «О порядке присуждения ученых степеней в Санкт-Петербургском государственном университете», соискатель Татьяна Евгеньевна Дворкина заслуживает присуждения ученой степени кандидата физико-математических наук по научной специальности 1.5.8 «Математическая биология, биоинформатика». Нарушения пунктов 9 и 11 указанного Порядка в диссертации не установлены.

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



Михаил Сергеевич Гельфанд

07.07.2023