ОТЗЫВ

Члена диссертационного совета на диссертацию Дворкиной Татьяны Евгеньевны 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

Развитие новых молекулярно-биологических методов, продуцирующих данные, ставит новые задачи перед биоинформатикой. В частности, появление методов секвенирования второго поколения потребовало разработки новых методов анализа данных, в том числе методов сборки геномов и транскриптомов, которые порождают гигантские графы сборки. Появление методов секвенирования третьего поколения привело к появлению новых, ранее не рассматривавшихся задач биоинформатики, в частности, к задачам гибридной сборки – сборки геномов на основе данных секвенирования второго и третьего поколений. Одной из подзадач такой гибридной сборки является выравнивание длинных прочтений третьего поколения на граф сборки, полученный на данных секвенирования второго поколения. Однако не только интеграция данных второго и третьего поколений приводит к подобного рода задачам. Здесь видны также задачи поиска различных генов в разного типа графах сборки. Диссертация Дворкиной Татьяны Евгеньевны как раз посвящена как раз проблеме выравнивания последовательностей на графы выравниваний. Кстати, мне представляется более адекватным названием работы «Задачи выравнивания последовательностей на графы сборки».

Диссертация начинается введением, где обосновывается актуальность работы, сформулированы цели и задачи исследования, поясняется практическая значимость работы. Далее идут три главы, посвященные разным аспектам рассматриваемой проблемы и трем сформулированным задачам. В диссертации отсутствует единый литературный обзор, но каждая глава предваряется своим литературным обзором. Всего диссертация ссылается на 137 источника.

Первая глава посвящена выравниванию последовательностей на графы сборки. При этом рассмотрена задача выравнивания длинных прочтений третьего поколения, а также задача выравнивания аминокислотных последовательностей на графы сборки. Описаны разработанные автором алгоритмы и программные инструменты (программа SPAligner) для решения этих задач.

Во второй главе развивается задача выравнивания аминокислотных последовательностей на графы сборки. Рассмотрена проблема поиска известных генов в графах сборки на примере задачи поиска генов токсинов насекомых в геномах бактерий и в метагеномах, причем геномы (метагеномы) представлены графами сборки. В последовательности для поиска представлены либо набором известных генов из других бактерий, либо CMM профилем. Был разработан вычислительный инструментарий в виде пакета ORFograph и он был применен для поиска генов инсектицидов в графах сборки. Применение этого инструментария позволило обнаружить ряд новых генов.

Третья глава посвящена исследованию архитектуры центромер. Эта задача также является специальным случаем задачи выравнивания. Разработаны инструменты

StringDecomposer, CentromereArchitect и HORmon, которые позволяют анализировать длинные прочтения третьего поколения и предсказывать структуру центромерных повторов.

В целом диссертация представляет из себя законченную научную работу. В диссертации удачно сбалансированы разработка инструментов и их тестирование с одной стороны и их применение к биологическим задачам.

Тем не менее по работе есть ряд замечаний:

1. Хорошо бы во введении перечислить явно публикации автора.
2. Стр.23 «для шести возможных сдвигов» – имеется в виду 6 рамок считывания? Надо понимать, что, есть 3 рамки по прямой цепи и 3 рамки по комплементарной цепи и два якоря, принадлежащих прямой и комплементарной цепи никогда не могут появиться в одном выравнивании.
3. Стр.23 «аминокислотные последовательности...могут быть восстановлены из одного якоря» – не совсем ясно. Это зависит от эволюционного расстояния. В стандартном BLAST на одно выравнивание приходится десятки первичных якорей, которые потом расширяются и большинство отбрасывается.
4. Хотелось бы иметь численные оценки, например, характерное количество символов в метке ребра как в геномных, так и в метагеномных данных. Может быть в большинстве случаев длина метки ребра сравнима с длиной белка и в большинстве случаев реально приходится строить выравнивание, проходящее через одну-две вершины. Тогда задача становится тривиальной. Далее говорится, что может возникнуть более одного пути – это крайне маловероятно, за исключением случая, когда 1-2 аминокислоты с концов белка имеют спорные картирования на продолжение ребер. Неясно, как обрабатываются случаи, когда кодон разорван вершиной графа.
5. Стр.26 неясно, за что отвечает параметр AG?
6. Стр.35 Неясно, на вход ORFograph поступает граф сборки индивидуального генома или граф сборки метагенома.
7. По-видимому в подписи к рис. 2.3 перепутаны синий и желтый цвета.
8. Стр.46 «не присутствовали в базе данных белков BLAST» – BLAST – не база данных, а программа поиска сходства.
9. Хотелось бы увидеть множественное выравнивание последовательностей Cry1, в котором отмечены новые находки. Также на дереве рис.2.4 хотелось бы также увидеть новые находки, например выделив их цветом подписи.
10. В тексте хотелось, чтобы собственные работы автора как-то выделялись, например «в **нашей** работе [xyz] было показано».

Сделанные замечания говорят скорее о неаккуратности формулировок, чем о научном содержании работы.

Диссертация Дворкиной Татьяны Евгеньевны 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 Дворкина Татьяна Евгеньевна 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

Профессор Факультета биоинженерии и биоинформатики,
доктор биологических наук,
кандидат физико-математических наук,
профессор А.А.Миронов

ПОДПИСЬ УДОСТОВЕРЯЮ
ЗАВ. КАНЦЕЛЯРИЕЙ
НН САДО...

Федеральное государственное образовательное учреждение высшего образования «Московский государственный университет имени М.В. Ломоносова»
ОГРН 1037700258694
Факультет биоинженерии и биоинформатики