Отзыв официального оппонента на диссертационную работу Гершгорина Романа Александровича «Кратчайшее преобразование и реконструкция хромосомных структур», представленную на соискание учёной степени кандидата физико-математических наук по специальности 03.01.09 – математическая биология, биоинформатика

Актуальность выбранной темы для науки и практики

В настоящий момент в общедоступных базах данных хранится очень большое количество секвенированных геномов различных организмов и органелл. Актуальной проблемой является разработка программного обеспечения для сравнительного анализа этих биологических последовательностей. Основным подходом при таком анализе является выравнивание последовательностей. Ограничением этого подхода является то, что с его помощью удобно исследовать лишь такие различия между геномами, которые вызваны локальными мутациями, то есть точечными заменами нуклеотидов, а также небольшими (до нескольких десятков нуклеотидов) вставками и делециями. В то же время хорошо известно, что в процессе эволюции происходят также крупные перестройки геномов, такие как инверсии, транслокации, крупные вставки и делеции. Программное обеспечение, облегчающее сравнительный анализ геномов различных организмов с учётом крупных перестроек, также существует, но несомненно нуждается в совершенствовании.

Диссертация Р.А. Гершгорина посвящена разработке алгоритмов и компьютерных программ для сравнительного анализа геномов с учётом их крупных перестроек. Часть этих алгоритмов предназначена для оценки минимального числа перестроек, необходимых для преобразования одного заданного генома в другой, что позволяет получать альтернативную, по сравнению с методами, основанными на выравниваниях отдельных генов, оценку времени расхождения организмов. Диссертантом разработаны и реализованы также алгоритмы для реконструкции предковых структур геномов на основе структур современных геномов. Программы, основанные на таких алгоритмах, могут быть востребованы специалистами, чья работа включает необходимость реконструкции эволюционных деревьев и оценки эволюционных расстояний между организмами или их самостоятельно эволюционирующими частями (клетками, органеллами, плазмидами, мобильными элементами и др.), в том числе занимающимися проблемами эволюционной биологии, эпидемиологии, биотехнологии. Кроме того, полученные теоретические

результаты могут послужить основой дальнейшего развития данного направления.

Содержание диссертации

Диссертация изложена на 127 страницах и состоит из введения, четырёх глав и списка использованных источников. Работа включает 14 таблиц и 75 рисунков. В списке использованных источников 77 наименований.

Введение состоит из четырёх подразделов; первые три фактически представляют собой немного изменённый вариант автореферата, в то время как четвёртый — это очень краткий (всего три страницы) обзор литературы.

После введения приведён список публикаций автора по теме диссертации, включающий пять статей в рецензируемых журналах и три кратких сообщения в сборниках тезисов конференций.

Глава 1 называется «Кратчайшее преобразование хромосомных структур без паралогов: неравный генный состав и неравные цены операций». В ней описывается и решается следующая задача. Пусть имеется два генома, каждый состоит из своего набора линейных и кольцевых хромосом. В данной работе каждая хромосома описывается как состоящая из генов, при этом существенным является взаимное расположение и ориентация этих генов. Все гены имеют названия, при этом названия генов одного генома могут совпадать с названиями генов другого генома (т.е. отношения ортологичности считаются уже установленными), в то же время названия генов внутри каждого генома не повторяются («отсутствие паралогов»). В реальных приложениях роль генов могут играть синтеничные участки. Тем самым каждый геном описывается своей т.н. «хромосомной структурой», что соответствует игнорированию локальных различий (точечных мутаций и небольших инделей) и позволяет сосредоточиться на изучении крупных перестроек, на уровне взаимного расположения участков, описанных как гены. Задача состоит в том, чтобы описать кратчайшее преобразование одной хромосомной структуры в другую. Для определения кратчайшего преобразования используются шесть элементарных операций над хромосомными структурами. Если присвоить каждой элементарной операции цену (положительное число), то кратчайшим преобразованием, по определению, будет последовательность элементарных операций, переводящая первую структуру во вторую и такая, что сумма цен входящих в неё операций будет минимально возможной. Решения подобной задачи известны из литературы, однако автором впервые рассмотрен случай, когда одновременно допускаются неравные цены операций и неравный генный состав (то есть гены, не имеющие ортологов в другом геноме). Предложен алгоритм, который может быть применён для любых значений цен операций, но гарантированно находит именно кратчайшее преобразование только при

специальном условии на цены, а именно, все цены, кроме одной (цены вставки связной цепочки генов) равны между собой, а цена вставки не более чем вдвое превышает цену остальных операций. Для этого алгоритма понадобилось новое определение т.н. «общего графа» двух структур, обобщающее определение, известное из литературы. Значительную часть первой главы (стр. 41–53) занимает разбор примеров работы алгоритма на конкретных парах хромосомных структур.

Глава 2 называется «Реконструкция хромосомных структур вдоль дерева: специальное расстояние и неравный генный состав». Здесь рассматривается другая задача: не определение минимальной последовательности перестроек, переводящих одну хромосомную структуру в другую, а реконструкция предковых состояний на основе рассмотрения трёх и более структур, при наличии предполагаемого филогенетического дерева этих структур. Предковые состояния в узлах дерева реконструируются на основе требования минимальной суммарной длины дерева, то есть минимальной суммы расстояния между структурами в соседних узлах и листьях. Расстояние при этом может определяться по-разному: и как минимальная суммарная цена операций, переводящих одну структуру в другую, и более простым образом (так называемое «брейкпойнтовое расстояние»). В главе 2 описываются алгоритмы, минимизирующие именно сумму брейкпойнтовых расстояний. Для случая отсутствия паралогов выходом разработанного алгоритма является сама разметка дерева, то есть предковые хромосомные структуры в узлах. Для общего случая, когда возможны паралоги, то есть одинаково названные гены в пределах одной структуры, выходом алгоритма является сведение задачи к задаче булева линейного программирования. Сами разработанные алгоритмы имеют квадратичную по объёму входных данных сложность, но в случае наличия паралогов полученная задача булева линейного программирования имеет факториальную сложность. Глава заканчивается кратким разбором небольшого примера.

Глава 3 носит название «Преобразование и реконструкция хромосомных структур, согласование контигов сведением к целочисленному линейному программированию: с паралогами и равными ценами». В этой главе цены всех элементарных операций предполагаются равными. Прежде всего рассматривается задача нахождения кратчайшего преобразования одной хромосомной структуры в другую в случае наличия паралогов. Описаны алгоритмы, сводящие эту задачу к задаче целочисленного линейного программирования (ЦЛП): линейной сложности для случая отсутствия незамкнутых (линейных) хромосом и квадратичный для общего случая. Затем описывается алгоритм кубической сложности, сводящий к задаче ЦЛП задачу о реконструкции предковых состояний, минимизирующих теперь уже сумму

редакционных расстояний (количество элементарных операций) по дереву. Алгоритм является точным (то есть действительно находящим требуемый минимум) только при специальном условии (фактически отсутствии событий исчезновения целой кольцевой хромосомы или появления полностью новой кольцевой хромосомы в минимальном сценарии эволюции). Наконец, рассматривается так называемая «задача согласования контигов», сводящаяся к задаче соединения двух наборов линейных хромосомных структур каждый в свою циклическую структуру, с минимальным расстоянием между полученными циклическим структурами. Предложен алгоритм, сводящий эту задачу к задаче ЦЛП. Глава завершается разбором работы алгоритмов, решающих задачи реконструкции, на двух примерах: с циклическими и с линейными хромосомами.

Последняя, четвёртая глава носит название «Филогенетические деревья и реконструкция хромосомных структур митохондрий инфузорий и споровиков, пластид родофитной ветви и бактерий рода Rhizobium» и содержит примеры применения разработанных алгоритмов и написанных на их основе программ на реальных биологических примерах.

Выводы в конце диссертации отсутствуют, но имеются во введении.

Обоснованность и достоверность изложенных научных результатов

Научные результаты, изложенные в диссертации, представляются обоснованными и достоверными. Разработанные алгоритмы изложены с исчерпывающей полнотой. Утверждения снабжены подробными доказательствами. Работа созданных компьютерных программ протестирована как на искусственных, так и на биологических примерах. Результаты опубликованы, в том числе, в ведущих рецензируемых журналах в области компьютерной биологии.

Новизна и значимость изложенных научных результатов

В диссертации описано несколько новых алгоритмов и компьютерных программ, превосходящих по ряду параметров имеющиеся аналоги. Полученные результаты важны для развития компьютерных методов молекулярной биологии. Прежде всего, созданные компьютерные программы могут непосредственно применяться исследователями, занимающимися сравнительным анализом геномов. Такой анализ имеет как фундаментальное, так и практическое значение, например, при эпидемиологических исследованиях, для повышения эффективности биотехнологий, в сортоведении и других областях. Кроме того, разработанные подходы и алгоритмы могут послужить основой для дальнейшего прогресса в области компьютерной биологии и биоинформатики.

Замечания

Из недостатков текста диссертации прежде всего хотелось бы отметить слишком краткий литературный обзор. Хотя упомянуто 14 работ, но каждой из них уделено всего по одному небольшому абзацу. По моему мнению, в диссертации на подобную тему следует излагать работы предшественников подробнее. Например, стоило кратко изложить алгоритм из статьи Compeau 2013 (номер 13 в списке использованных источников), а не ограничиваться лишь его упоминанием. Из текста диссертации непросто понять, чем каждый из описанных в ней пяти новых алгоритмов превосходит существовавшие ранее аналоги.

Две из восьми публикаций автора, приведённых как публикации по теме диссертации, из них одна — в реферируемом журнале, по темам слабо связаны (если вообще связаны) с темой диссертации. Впрочем, остальные четыре публикации в реферируемых журналах несомненно связаны с темой работы и достаточно полно описывают её основные результаты.

Для статьи, имеющей номер 15 в списке использованных источников, указан неверный DOI (от другой статьи).

К сожалению, автор не провёл тестирования различных наборов весов элементарных операций, это можно было бы сделать на имеющихся примерах сравнением межгеномных расстояний, посчитанных по хромосомным перестройкам, с одной стороны, и по локальным заменам в отдельных генах, с другой. При этом в диссертации (стр. 39) приводятся конкретные наборы весов как «удобные» для линейных или же циклических хромосом, но не указывается, каким образом подбирались эти веса. Нелишним было бы и сравнение «брейкпойнтового» расстояния с расстоянием, определяемым как суммарная цена операций, например, по тому, какое из них и насколько лучше коррелирует с числом локальных мутаций.

Наконец, общее замечание ко всему тексту — известная сумбурность изложения. Например, во введении сказано, что описанный в главе 1 алгоритм имеет линейную по времени и памяти сложность, но в самой главе 1 про сложность алгоритма не сказано вообще ничего. Определения прописаны не всегда достаточно чётко, например в (ключевом) определении общего графа двух структур на стр. 20 написано: «Ребра общего графа соединяют следующие пары вершин: <…> 4) Если линейный блок не имеет соседних общих генов (т.е. является цепью), то в общем графе ему соответствует изолированная особая вершина». В математическом тексте предпочтительна большая структурированность и чёткость. Ссылка на программную реализацию алгоритмов, которой следовало бы посвятить отдельный раздел введения, приводится «между делом» в четвёртой главе.

Указанные замечания не умаляют важности проделанной работы и не влияют на её общую положительную оценку.

Заключение

Диссертация Р.А. Гершгорина посвящена актуальной теме, содержит новые важные научные результаты и свидетельствует о личном вкладе автора в науку. Основные результаты работы опубликованы в рецензируемых научных изданиях. Автореферат соответствует содержанию диссертации.

Можно заключить, что диссертация Р.А. Гершгорина на тему «Кратчайшее преобразование и реконструкция хромосомных структур» является законченной научно-квалификационной работой и соответствует всем критериям, установленным «Положением о порядке присуждения учёных степеней», утверждённым Постановлением правительства РФ № 842 от 4 сентября 2013 г., а её автор, Роман Александрович Гершгорин, заслуживает присуждения ему учёной степени кандидата физико-математических наук по специальности 03.01.09 – «Математическая биология, биоинформатика».

Кандидат физико-математических наук,
ведущий научный сотрудник
НИИ физико-химической биологии имени А.Н. Белозерского
Московского государственного университета имени М.В. Ломоносова
Спирин Сергей Александрович
28.01.2019 /С.А. Спирин/