ОТЗЫВ

официального оппонента на диссертационную работу

Гершгорина Романа Александровича

«Кратчайшее преобразование и реконструкция хромосомных структур», представленной на соискание ученой степени кандидата физико-математических наук по специальности 03.01.09 – математическая биология, биоинформатика

Актуальность выбранной темы для науки и практики

Диссертация Р.А. Гершгорина посвящена разработке алгоритма преобразования одного ориентированного графа, состоящего из цепей и циклов (каждому ребру приписано имя – натуральное число) в другой такой граф операциями из фиксированного списка. Такие графы традиционно называются хромосомными структурами (или структурами), компоненты графа – хромосомами, а их рёбра – генами. Также она посвящена разработке алгоритма расстановки структур, заданных в листьях данного дерева, в его внутренних вершинах. Точнее, каждой операции сопоставлена её цена – рациональное число. Ищется преобразование с минимальной суммарной ценой и соответственно расстановка с минимальной суммарной ценой по всем рёбрам. Для двух структур минимальная суммарная цена называется кратчайшим расстоянием. Обе задачи важны, в частности, в биоинформатике, где структуры понимаются как отражение биологических хромосомных структур, а реконструкция понимается как отражение эволюции хромосомных структур видов вдоль дерева видов. Полученные диссертантом алгоритмы применяются им для построения эволюционных деревьев хромосомных структур митохондрий, инфузорий и споровиков, пластид водорослей и споровиков, бактерий. Такие исследования широко ведутся с начала 90-х годов прошлого века, по ним опубликованы сотни работ. Ведущий специалист по биоинформатике П.А. Певзнер и его последователи рассматривали различные частные случаи задач преобразования и реконструкции структур, как правило, ограничиваясь случаем равного генного состава структур и тем, что имена генов не повторяются, а цены операций не рассматриваются. Диссертант допускает все эти возможности, тем самым, рассматривая задачи в наиболее общей постановке, которая реально востребована в биоинформатике. У него структура может состоять из любого

числа линейных и кольцевых хромосом. Операции над структурами полно отражают биологические изменения в геноме. Это – двойная, полуторная и одинарные переклейки (известные как DCJ-операции), а также удаление и вставка связного участка генов. Реальные геномы содержат десятки тысяч генов, поэтому важный вопрос – оценка сложности получаемых алгоритмов, а также – доказательство точности алгоритмов (естественно, при некоторых условиях и при правильном ответе оракула). Насколько могу судить, в биоинформатике для современных достаточно сложных алгоритмов не принято рассматривать обе эти важные характеристики. В диссертации они тщательно рассматриваются. Если имена рёбер не повторяются, диссертант предложил алгоритм для задачи преобразования, решающий задачу за время, линейное от размера исходных графов. Этот алгоритм (глава 1) весьма нетривиален. Если имена повторяются, задача становится NP-трудной и в предположении P≠NP она не может быть решена точным полиномиальным алгоритмом. Диссертант описал сведение исходной задачи к задаче целочисленного линейного программирования (ЦЛП), что позволяет применять известные программы для ЦЛП, хорошо зарекомендовавшие себя на практике. Время их работы критически зависит от размера задачи ЦЛП, что делает важным приведённые диссертантом оценки: для задачи преобразования размер соответствующей задачи ЦЛП не более чем квадратичный, а для задачи реконструкции – не более чем кубичный. Этот доказанный диссертантом результат неожиданный и с теоретической точки зрения.

Содержание диссертации. Диссертационная работа Гершгорина Романа Александровича имеет традиционную структуру. Она состоит из обзора литературы, четырех глав, содержащих методы и результаты, список использованной литературы из 77 работ. Она изложена на 127 страницах и включает 75 рисунков и 14 таблиц.

Во Введении приведены постановки задач, формулировки результатов, обзор литературы.

В Главе 1 рассматривается задача о преобразовании одной структуры в другую DCJ-операциями, дополненными операциями удаления и вставки. В ней предполагается отсутствие паралогов.

В разделе 1.1 описывается ключевая идея алгоритма: представить структуры a и b в виде неориентированного графа $a+b$ и свести задачу преобразования a в b к задаче

приведения графа $a+b$ к финальному виду $c+c$. Операциями над $a+b$ служат аналоги операций над структурами кроме операции вставки, использования которой удаётся избежать.

В разделе 1.2 описывается оригинальный линейный по сложности алгоритм решения задачи приведения $a+b$ к финальному виду. Его точность доказана (не диссертантом).

В разделе 1.3 приведено тестирование этого алгоритма на искусственных примерах.

В Главе 2 описаны алгоритмы решения задачи реконструкции структур вдоль данного дерева для специального расстояния между структурами. Это расстояние – упрощённое расстояние из главы 1: допускаются лишь операции разреза и склейки хромосом, удаления и вставки одиночных генов, не соседствующих с другими генами.

В разделе 2.1 приводится постановка задачи: дано дерево и структуры в его листьях, нужно расставить структуры по внутренним вершинам так, чтобы суммарное по всем рёбрам расстояние между структурами для всех рёбер минимизировалось.

В разделе 2.2 в предположении отсутствия паралогов для любых цен операций приводится квадратичный по времени работы и используемой памяти алгоритм решения этой задачи. Доказывается его точность (теорема 2 для равных цен и теорема 3 для неравных).

В разделе 2.3 при допущении паралогов для любых цен операций приводится описание сведения рассматриваемой задачи к задаче булева линейного программирования квадратичного размера.

В разделе 2.4 приведено тестирование описанных алгоритмов на искусственных примерах.

В Главе 3 рассматриваются те же задачи преобразования и реконструкции структур, но в предположении наличия паралогов и равных цен операций.

В разделе 3.1 описано сведение задачи преобразования к задаче ЦЛП для случая, если все хромосомы в структурах кольцевые. В этом случае алгоритм сведения приобретает существенно более простой, хотя и далеко нетривиальный, вид по сравнению с общим случаем. Размер задачи ЦЛП также невелик: число переменных и ограничений квадратично зависит от суммарного числа паралогов в данных структурах, а при фиксированном числе паралогов – линейно от суммарного размера исходных структур.

Поскольку число паралогов обычно невелико (заведомо не больше квадратного корня от размера структур), размер задачи ЦЛП можно считать линейным.

В разделе 3.2 описано сведение задачи преобразования к задаче ЦЛП для общего случая. Размер задачи ЦЛП квадратичный. Ключевая идея сведения – устранение паралогов введением переменных, указывающих на соответствие гена в одной структуре гену в другой структуре (или на отсутствие соответствия). Минимизация целевого функционала даёт соответствие, при котором минимально расстояние между структурами, уже не имеющими паралогов, после чего алгоритм из главы 1 строит соответствующую последовательность операций.

В разделе 3.3 описано сведение задачи реконструкции к задаче ЦЛП кубического размера. Ключевая идея та же, что и в предыдущем разделе с тем усложнением, что вводятся дополнительные переменные, задающие неизвестные структуры во внутренних вершинах дерева.

В разделе 3.4 описано сведение к задаче ЦЛП линейного размера задачи согласования двух произвольных множеств цепей (линейных структур). Схема решения соответствует методике, использовавшейся в разделе 3.1.

В разделе 3.5 приведено тестирование описанных алгоритмов на искусственных примерах.

В Главе 4 диссертант применяет ранее описанные алгоритмы для построения филогенетических деревьев хромосомных структур митохондрий инфузорий и споровиков, а также пластид и бактерий. Эти алгоритмы применяются для вычисления матрицы расстояний между данными структурами, а затем известным методом по матрице строится дерево.

Новизна и значимость основных научных результатов, полученных диссертантом

Диссертационная работа Р.А. Гершгорина носит теоретический характер. Основные результаты, полученные соискателем.

- Получен линейной сложности точный алгоритм решения задачи преобразования структур без паралогов.
- Для специального расстояния, отсутствия паралогов и любых цен операций, получен квадратичной сложности точный алгоритм решения задачи реконструкции.

В случае того же расстояния, присутствия паралогов и любых цен операций, получен квадратичной сложности точный алгоритм для решения задачи реконструкции сведением к задаче квадратичного булева линейного программирования. Точность алгоритмов доказана.

- Получены алгоритмы для решения задачи преобразования, с паралогами и равными ценами, сведением к целочисленному линейному программированию: для циклических хромосомных структур – к ЦЛП линейного размера и для произвольных структур – к ЦЛП квадратичного размера. Точность алгоритмов доказана. Сложность алгоритмов сведения соответственно линейная и квадратичная.
- Получен кубической сложности точный алгоритм для решения задачи реконструкции, с паралогами и равными ценами, произвольных структур сведением к ЦЛП кубического размера.
- Получен линейной сложности точный алгоритм для решения задачи согласования, с паралогами и равными ценами, множеств контигов сведением к ЦЛП линейного размера.
- На основе компьютерных реализаций алгоритмов, которые предложены в пунктах 1–5, и стандартного пакета решения задачи ЦЛП построены филогенетические деревья хромосомных структур митохондрий инфузорий и споровиков из класса *Aconoidasida*, пластид родофитной ветви у водорослей и споровиков, бактерий рода *Rhizobium*.

Имеются все основания утверждать, что автор реализовал поставленные цели и задачи и обосновал заявленную новизну исследования и научную значимость полученных результатов. Все результаты диссертации являются новыми и подробно доказанными. Актуальность диссертационного исследования не вызывает сомнений в свете необходимости обрабатывать всё возрастающие объёмы геномных данных. Построение сценариев эволюции и, в частности, эволюционных деревьев является одной из сложнейших и не решённых до конца проблем биоинформатики. Сравнение деревьев, построенных разными способами, позволяет выявить перспективные направления в разработке новых алгоритмов их построения или усовершенствовании имеющихся.

Отсюда вытекает практическая ценность предложенного диссертантом способа построения эволюционных деревьев на основе сравнения хромосомных структур изучаемых таксономических групп.

Замечания.

1. Автор не всегда чётко разграничивает свои результаты от упоминаемых им результатов других авторов. Например, во Введении текста диссертации автор приводит условие точности основного из его алгоритмов: «Получено эффективное решение задачи преобразования структур общего вида *без паралогов*; кроме того, оно является точным, если ….» (стр. 9). Хотя из автореферата, дальнейшего текста диссертации и из соответствующих ссылок ясно, что этот результат получен не автором, было бы лучше подчеркнуть это и во Введении.

2. Автор не приводит никаких данных об эффективности применяемого им пакета решения задач ЦЛП. Очевидно, он не разбирался в деталях реализованного там алгоритма.

3. Во Введении автор называет свои алгоритмы эффективными, относя эту характеристику к сложности алгоритмов сведения к ЦЛП и финального алгоритма из главы 1, который по подсказке ЦЛП строит решение исходной графовой задачи. Конечно, о самой задаче ЦЛП и, тем более, о конкретном пакете, который диссертант использовал для её решения, нет аналогичных результатов. Действительно, в теории алгоритмов с оракулами характеристики сложности и точности всегда не касаются самого оракула, для которого они обычно не известны или тривиально экспоненциальные. Однако в контексте прикладной работы это можно было бы объяснить явно.

4. Описывая применения своих алгоритмов, автор не указывает время решения задач. Не приводит автор и сравнений результатов вычислений одной и той же задачи в разных пакетах ЦЛП, что позволило бы увидеть, насколько полученные решения близки.

Заключение

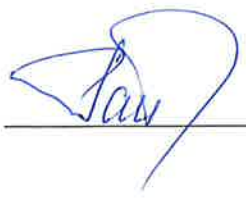
Замечания не снижают положительного мнения о диссертационном исследовании, представляющем собой обстоятельный и завершенный труд, в котором содержится подробное описание алгоритмических решений важных научных проблем, имеющих практические приложения. Диссертационная работа выполнена автором самостоятельно

на высоком научном уровне. Полученные результаты достоверны, выводы и заключения обоснованы. По каждой главе в работе сделаны убедительные выводы, соответствующие поставленным задачам и полученным результатам. Автореферат соответствует содержанию диссертации.

Основные результаты диссертации представлены в ряде статей и неоднократно обсуждались на отечественных и международных конференциях.

Учитывая все вышесказанное, диссертационная работа Гершгорина Романа Александровича на тему «Кратчайшее преобразование и реконструкция хромосомных структур» без сомнения является законченным научно-квалификационным исследованием и соответствует всем требованиям «Положения о порядке присуждения ученых степеней», утвержденного Постановлением правительства РФ от 4 сентября 2013 г. №842. Гершгорин Роман Александрович заслуживает присуждения ученой степени кандидата физико-математических наук по специальности 03.01.09 – «математическая биология, биоинформатика».

доктор физико-математических наук,
главный научный сотрудник – заведующий
лабораторией продвинутой комбинаторики и
сетевых приложений Федерального государственного
автономного образовательного учреждения
высшего образования Московский физико-технический
институт (государственный университет)
Райгородский Андрей Михайлович



/Райгородский А.М./

21 января 2019 года

Подпись д.ф.-м.н. А.М. Райгородского заверяю

Ученый секретарь МФТИ
к.ф.-м.н., доцент
Скалько Юрий Иванович