МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное бюджетное учреждение науки
Институт общей генетики им. Н.И. Вавилова
Российской академии наук
(ИОГен РАН)

ул. Губкина, д. 3, г. Москва, ГСП-1, 119991
Тел.: (499) 135-62-13, (499) 135-20-41
Факс: (499) 132-89-62

E-mail: iogen@vigg.ru
http: www.vigg.ru

УТВЕРЖДАЮ
Директор Института общей
генетики им. Н.И. Вавилова РАН,
д.б.н., проф. А. М. Кудрявцев

24 января 2019

ОТЗЫВ

ведущей организации 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

Развитие технологий высокопроизводительного секвенирования привело к появлению большого числа нуклеотидных последовательностей хромосом для большого числа биологических видов. Это даёт большой простор для изучения хромосомной эволюции, что в свою очередь приводит к необходимости разработки эффективных алгоритмов, способных работать с полногеномными последовательностями. Диссертация Р.А. Гершгорина посвящена исследованию эволюции хромосомных структур. А именно, разработке алгоритмов для вычисления преобразования одной хромосомной структуры в другую, реконструкции структур вдоль дерева видов и применению найденных алгоритмов к исследованию хромосомных структур митохондрий инфузорий и споровиков из класса

Aconoidasida, пластид родофитной ветви у водорослей и споровиков, бактерий рода Rhizobium. Диссертация относится к большой теме исследований, которые особенно интенсивно начались около 1992 года; с этого времени по теме опубликованы более сотни работ. Диссертант рассматривает общий случай неравного генного состава, наперёд заданных цен операций и присутствия паралогичных генов. Поставленная в диссертации задача NP-трудная и, следовательно, не может быть решена точным полиномиальным алгоритмом. Поэтому перед диссертантом стояли следующие задачи: найти алгоритмы, которые имеют квадратичное (или даже линейное, но не выше кубического) время работы и используют такой же размер памяти вычислительного устройства (относительно размера исходных данных) – для задачи преобразования без паралогов. Для задачи реконструкции (даже без паралогов) найти мотивированный способ формально «устранить паралоги». Такой способ найден диссертантом и состоит в сведении исходных данных к задаче целочисленного линейного программирования (ЦЛП), решение которой содержит оптимальное соответствие паралогов, после чего исходная задача решается как задача преобразования без паралогов. Важно, что задача ЦЛП, к которой выполняется сведение, не более чем кубического размера от размера исходной графовой задачи.

Диссертационная работа изложена на 127 страницах и включает следующие разделы: Введение, которое также содержит выводы и обзор литературы, 4 главы и список использованных источников. Работа также включает 75 рисунков, 14 таблиц. Список использованных источников содержит 77 работ.

Во Введении приводятся постановки задач, полученные результаты, содержание работы, выводы и обзор литературы.

В главе 1 автор рассматривает задачу о преобразовании двух общего вида хромосомных структур a и b, первую во вторую, с помощью фиксированного хорошо известного, стандартного набора операций. В разделе 1.1 приводятся ключевые определения. В разделе 1.2 Р.А. Гершгорин описывает оригинальный линейный по сложности алгоритм решения задачи преобразования без паралогов. В разделе 1.3 приведены результаты тестирования этого алгоритма на искусственных примерах.

В главе 2 автор описывает решение задача реконструкции хромосомных структур вдоль заданного дерева для специального (брейкпоинтового) расстояния и любых цен операций. В разделе 2.1 приводится постановка задачи. В разделе 2.2, автор приводит квадратичный по времени работы и памяти алгоритм решения задачи реконструкции в

отсутствии паралогов. В разделе 2.3 приводится точный квадратичный по времени работы и используемой памяти алгоритм сведения задачи реконструкции к квадратичному булевому линейному программированию в присутствии паралогов. Специальное расстояние – частный, более простой случай кратчайшего расстояния, для которого задачи рассматриваются в главе 3, уже в предположении равных цен операций.

В главе 3 получены алгоритмы сведения к ЦЛП, которые вместе с алгоритмом из главы 1, дают решения исходных задач. В разделе 3.1 описывается точный алгоритм решения задачи преобразования для циклических структур. В разделе 3.2 приводится точный алгоритм решения задачи преобразования для произвольных структур. В разделе 3.3 описывается точный алгоритм решения задачи реконструкции для произвольных структур. В разделе 3.4 предложен точный линейный алгоритм решения задачи согласования двух произвольных множеств цепей, путём сведения её к линейному ЦЛП. В разделе 3.5 приводится тестирование алгоритмов из этой главы на искусственных примерах. По всей диссертации точность понимается естественным образом: относительно точности решения задачи ЦЛП и при условии на алгоритм из главы 1. Оценки сложности везде относятся к алгоритмам сведения и алгоритму из главы 1, который превращает подсказку ЦЛП в решение исходной задачи.

В Главе 4 автор приводит результаты применения алгоритмов, которые описаны в главах 1 и 3, и соответствующих компьютерных программ для построения филогенетических деревьев хромосомных структур митохондрий инфузорий (*Ciliophora*) и споровиков видов класса *Aconoidasida*, пластид родофитной ветви, а также бактерий рода *Rhizobium*.

Новизна и значимость основных научных результатов, полученных диссертантом

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

1. Получен линейной сложности точный алгоритм решения задачи преобразования структур без паралогов.
2. 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.

3. 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.
4. Получен кубической сложности точный алгоритм для решения задачи реконструкции, с паралогами и равными ценами, произвольных структур сведением к ЦЛП кубического размера.
5. Получен линейной сложности точный алгоритм для решения задачи согласования, с паралогами и равными ценами, множеств контигов сведением к ЦЛП линейного размера.
6. 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*.

Новизна результатов состоит в подходе к практическому решению NP-трудной задачи, который основан на целочисленном линейном программировании и алгоритмах сведения. Финальный алгоритм, решающий задачу преобразования хромосомных структур, использует около двадцати взаимосвязанных новых понятий, связанных с графами.

Результаты диссертации получены автором самостоятельно и впервые, и в качестве пяти статей опубликованы в международных рецензируемых журналах BMC Bioinformatics, Молекулярная биология и Journal of Communications Technology and Electronics,.

Рекомендации по использованию результатов и выводов диссертации

Разработанные Р.А. Гершгориным алгоритмы реконструкции могут быть использованы для изучения эволюции хромосомных структур клеточных органелл, геномной синтении и, как следствие, эволюции видов. Результаты работы могут быть использованы в научно-исследовательских организациях, занимающихся исследованиями в областях молекулярной биологии и биофизики, таких как Институт молекулярной биологии им. В.А. Энгельгардта РАН, Институт теоретической и экспериментальной биофизики РАН,

Институт биофизики клетки РАН, Институт цитологии и генетики СО РАН, Институт химической биологии и фундаментальной медицины СО РАН, Институт биофизики СО РАН, Московский государственный университет имени М.В. Ломоносова.

Общие замечания

К сожалению, диссертация не лишена ряда недостатков:

1. Во Введении текста диссертации автор описывает свои алгоритмы словом «эффективные». Данную характеристику для приводимых алгоритмов использовать не совсем правомерно. Статистически значимое исследование их эффективности требует исследования эффективности используемого диссертантом пакета ЦЛП; такое исследование не проведено и вряд ли возможно. Обращение к «черному ящику» (в данном случае – ЦЛП) не позволяет сделать теоретический вывод о сложности рассматриваемой цепочки алгоритмов: сведение+ЦЛП+алгоритм из главы 1. Для биологических данных, на которых цепочка алгоритмов тестировалась, время счёта оказалось приемлемым, что, конечно, не позволяет сделать вывод о его теоретической эффективности на произвольных данных.

2. Во Введении (стр. 9) в тексте диссертации используется выражение «Получено эффективное решение задачи реконструкции» (и аналогично для других задач). Эта фраза неточна: что понимать под «решением задачи реконструкции»? Можно понимать цепочку алгоритмов, которая включает не только оригинальные алгоритмы сведения и финальный, но и обращение к «черному ящику», что не может быть эффективным для произвольных данных.

3. Задачи ЦЛП, которые нужно решить ради решения поставленных задач, имеют специальный вид по сравнению с произвольной задачей целочисленного линейного программирования. Для задач ЦЛП такого вида, по крайней мере, не противоречит известным теоретическим результатам предположение о решении их за хорошее полиномиальное время. Хотелось бы, чтобы диссертант исследовал возникающий таким образом новый подкласс задач ЦЛП. Это не сделано.

4. В тексте диссертации встречаются опечатки и ошибки пунктуации (например, на страницах 105, 83). Что хуже, встречаются повторы частей предложения, например, в первом абзаце пункта Содержание работы во Введении.

Заключение

Диссертационная работа Р.А. Гершгорина представляет собой законченное исследование на актуальную тему, проведенное на высоком научном уровне. Работа

аккуратно оформлена, написана грамотным языком, сделанные выводы хорошо обоснованы. Автореферат соответствует содержанию диссертации. Полученные результаты имеют важное теоретическое и практическое значение. Работа Р.А. Гершгорина отвечает требованиям пункта 9 «Положения о порядке присуждения ученых степеней», утвержденного постановлением Правительства Российской Федерации от 24 сентября 2013 г. №842, предъявляемым к кандидатским диссертациям, а ее автор заслуживает присуждения ученой степени кандидата физико-математических наук по специальности 03.01.09 – «математическая биология, биоинформатика».

Отзыв рассмотрен на расширенном семинаре лаборатории системной биологии и вычислительной генетики ИОГен РАН 21 января 2019 года, протокол №4.

кандидат физико-математических наук,
научный сотрудник
лаборатории системной биологии и вычислительной генетики
Федерального государственного бюджетного учреждения науки
Института общей генетики им. Н. И. Вавилова
Российской академии наук (ИОГен РАН)

_____/Касьянов А.С./
21 января 2019 года

Подпись научного сотрудника лаборатории системной биологии и вычислительной генетики ИОГен РАН, кандидата физико-математических наук Касьянова Артема Сергеевича удостоверяю.

Ученый секретарь ИОГен РАН
д.б.н., проф. Абилев С.К.