На правах рукописи

Гершгорин Роман Александрович

КРАТЧАЙШЕЕ ПРЕОБРАЗОВАНИЕ И РЕКОНСТРУКЦИЯ ХРОМОСОМНЫХ СТРУКТУР

03.01.09 – Математическая биология, биоинформатика

Автореферат диссертации на соискание ученой степени
кандидата физико-математических наук

Москва – 2018

Работа выполнена в Федеральном государственном бюджетном учреждении науки Институте проблем передачи информации им. А.А. Харкевича Российской академии наук (ИППИ РАН)

Научный руководитель:

Любецкий Василий Александрович, доктор физико-математических наук, профессор, Федеральное государственное бюджетное учреждение науки Институт проблем передачи информации им. А.А. Харкевича Российской академии наук (ИППИ РАН), заведующий лабораторией математических методов и моделей в биоинформатике.

Официальные оппоненты:

Андрей Михайлович Райгородский, доктор физико-математических наук, профессор, Федеральное государственное автономное образовательное учреждение высшего образования Московский физико-технический институт (государственный университет), факультет инноваций и высоких технологий, кафедра дискретной математики, главный научный сотрудник – заведующий лабораторией продвинутой комбинаторики и сетевых приложений.

Сергей Александрович Спирин, кандидат физико-математических наук, Научно-исследовательский институт физико-химической биологии им. А.Н. Белозерского Московского государственного университета им. М.В. Ломоносова, ведущий научный сотрудник.

Ведущая организация: Федеральное государственное бюджетное учреждение науки Институт общей генетики им. Н.И. Вавилова Российской академии наук.

Защита состоится …. 2019 года в .... на заседании диссертационного совета Д 002.077.04 на базе ИППИ РАН Большой Каретный пер., д. 19, стр. 1, Москва, ГСП-4, 127994.

С диссертацией можно ознакомиться в библиотеке ИППИ РАН и на сайте www.iitp.ru

Автореферат разослан … 201.. года

Учёный секретарь диссертационного совета,
доктор биологических наук, профессор Г.И. Рожкова

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Рассматриваемые задачи, результаты в целом и актуальность темы. Большое число полно секвенированных хромосом, включая геномы митохондрий и пластид, открыло ещё один путь для построения эволюционных сценариев, кроме традиционного пути, основанного на расхождении гомологичных генов. А именно, появилась возможность строить модель эволюции на основе взаиморасположения генов, иными словами: макроструктуры генома, на расхождении макроструктур. Исследование макроструктуры или, как ещё говорят, геномной или *хромосомной структуры* началось около 1992 года, и к настоящему времени известно большое число результатов, как по эволюции хромосомных структур отдельных видов, так и по алгоритмам и компьютерным программам, которые работают с такими структурами. Однако эти алгоритмы не позволяют работать с хромосомными структурами общего вида и назначать любые цены операций. Модели эволюции рассматривают хромосомные структуры ядерных геномов, геномов органелл, а также – виды в целом. Сравнение эволюционных деревьев, получаемых на основе разных моделей эволюции, приводит к важным биологическим выводам. Это – модели, основанные на гомологичных белках (суперматрице выравниваний), рРНК, ультраконсервативных и высококонсервативных элементах, хромосомных структурах, расстояниях между генами и т.д.

Алгоритм называется *точным* (в противоположность – *эвристическому*), если он сопровождается доказательством того, что для любых данных на его входе выдаётся глобальный минимум функционала из соответствующей задачи. Точные алгоритмы имеют очевидное преимущество перед эвристическими; доказательство точности всегда требует условия на данные, поэтому «хороший эвристический» алгоритм – точный алгоритм, успешно применяемый и вне области выполнения такого условия; конечно, условие должно быть достаточно широким. Алгоритм может точно или эвристически решать исходную задачу и тогда называется *прямым*. Однако алгоритм может состоять в сведении одной задачи к другой и тогда называется *алгоритмом сведения*, в этом случае утверждение о точности относится к алгоритму сведения, а не к решению задачи, которая получается в результате сведения. Представляют интерес *алгоритмы сведения* к каноническим задачам,

которые прошли широкую апробацию и для которых разработаны широко применяемые пакеты компьютерных решений, в том числе, приближённых, а иногда и специализированные вычислительные устройства. Среди таких канонических задач находится линейное программирование и его усложнения, включая целочисленное и булево. Важно, чтобы сама задача минимизации, к которой сводится исходная, имела квадратичный или близкий к нему размер (от размера исходной задачи), а алгоритм сведения имел столь же низкую сложность вычисления. Например, задача о графах с n вершинами и рёбрами сводилась бы к линейной задаче минимизации с n^2 переменными, равенствами и неравенствами.

Хромосомная структура (часто говорят: *структура*) по сравнению со многими предшествующими исследованиями понимается в нашей работе наиболее общим образом. А именно, как граф, состоящий из любого множества непересекающихся цепей и циклов, в которых каждому ребру приписано направление и имя. Такой граф соответствует множеству линейных и кольцевых хромосом, если пренебречь межгенными участками, длиной и составом генов. Тогда ген с его направлением транскрипции и именем изображается ребром указанного графа. В этом смысле ребро вместе с его именем называется далее *геном*, а цепь или цикл – *хромосомой* (или на математическом языке – *компонентой*). *Край* гена понимается как позиция хромосомы, в которой начинается или заканчивается его транскрипция; численное значение этой позиции также не учитывается в хромосомной структуре. Термины *конец* и *начало* (направленного ребра) используются в контексте работы с графами. В хромосомной структуре края соседних генов отождествляются (или, как говорят, *склеиваются*) в вершине графа.

Долгое время основной задачей в этой области было вычисление расстояния между двумя структурами a и b, точнее, нахождение *кратчайшей* последовательности операций из их фиксированного списка, которые преобразуют a в b, и цены кратчайшей последовательности, которую будем называть *кратчайшим расстоянием* между a и b. «Кратчайшая» означает здесь: «с минимальной суммарной ценой операций в последовательности с точностью до некоторой фиксированной аддитивной константы». Точнее, пусть $c(a,b)$ – минимум функционала кратчайшего расстояния, $T(a,b)$ – суммарная цена последовательности, которую выдаёт наш алгоритм; тогда должно выполняться: $|c(a,b)-T(a,b)| \le k$, где

константа k не зависит от a и b, а зависит только от цен операций. Заметим, что гораздо чаще предлагаются алгоритмы «с точностью до некоторой фиксированной мультипликативной константы k», что, конечно, является гораздо более слабым утверждением; оно означает: $\frac{T(a,b)}{c(a,b)} \leq k$. В случае равных цен всех операций кратчайшее расстояние иногда называют (обычным) *расстоянием*.

Искомая минимальная цена (=кратчайшее расстояние) не является обычной метрикой. Нахождение кратчайшего расстояния и кратчайшей последовательности называется **задачей о *кратчайшем преобразовании* a в b**. Эта задача рассматривается в Главе 1 (публикация [1]), где предполагается, что имена в структуре не повторяются (биологически это значит, что *отсутствуют паралоги*). В качестве операций фиксированы *стандартные* – двойная, полуторная и одинарная переклейки, и *дополнительные* – удаление и вставка связного участка рёбер (генов).

В предшествующих результатах, относящихся к задаче преобразования, обычно использовались два очень существенных ограничения: множества имён в a и b совпадают («равный генный состав») и цены всех операций равны, т.е. минимизируется всего лишь число операций в кратчайшей последовательности или, иными словами, использовалось обычное расстояние. Часто использовались и более сильные ограничения. В диссертации (раздел 4 Введения) приведён обзор публикаций, наиболее тесно связанных с нашими результатами. В диссертации **нигде не предполагается равный генный состав, а цены при отсутствии паралогов могут быть разными или, напротив, цены могут быть равными, но допускается любое число паралогов.** Это принципиально усложняет задачу.

При произвольных ценах поставленная задача NP-трудная, т.е. заведомо не поддаётся обоснованному решению; поэтому её можно решать только при тех или иных условиях на цены операций; впервые в работах диссертанта найдены точные алгоритмы решения общей задачи преобразования [1,2].

Решение задачи преобразования ищется в виде *линейного* по сложности алгоритма, т.е. по времени работы и по используемой памяти. Для характерных данных современной биоинформатики, такое или близкое к нему требование (квадратичности или, в крайнем случае, кубичности алгоритма) кажется обязательным. В диссертации предлагаются как *прямые* алгоритмы, так и их

альтернатива – алгоритмы *сведения*. Алгоритм сведения имеет у нас сложность, которая не превосходит размера соответствующей линейной задачи, т.е. числа переменных и ограничений в ней, и, таким образом, сведение не сложнее, чем сложность выписывания этих переменных и ограничений. *Квадратичного* размера называется целочисленное линейное программирование (ЦЛП) с квадратичным от размера исходных данных числом переменных и ограничений, например, от суммарного числа рёбер в исходных графах. В аналогичном смысле употребляется термин *линейное* и *кубическое* ЦЛП; прилагательное везде указывает на размер ЦЛП относительно размера исходной задачи. Аналогичные термины используются для булева линейного программирования (БЛП).

Вторая решаемая в диссертации ***задача*** – ***реконструкция*** структур вдоль дерева, которая, насколько диссертанту известно, ранее не рассматривалась в присутствии паралогов. Она состоит в следующем. Дано корневое (не обязательно бинарное) дерево, каждому его листу приписана структура, в которой имена могут повторяться (тем самым, *допускаются паралоги*). Найти расстановку структур, включая *соответствие паралогов*, по всем внутренним вершинам дерева, для которой расстояние между структурами на концах любого ребра (суммированное по всем рёбрам и называемое *суммарным расстоянием*) минимально. Расстояние между структурами a и b, приписанными концам ребра, можно определить как кратчайшее расстояние (в смысле задачи преобразования) или иначе: как число пар различных краев генов, которые в одной структуре склеены, а в другой – не склеены или отсутствуют, сложенное с числом генов, присутствующих в одной структуре и отсутствующих в другой. Это расстояние может вычисляться также с учётом цен: за каждую пару краёв склеенных в одной структуре и расклеенных в другой начисляется фиксированная цена; аналогично за каждый ген, присутствующий в одной структуре и отсутствующий в другой, начисляется своя фиксированная цена. И тогда минимизируется сумма таких цен по всем событиям на ребре (a,b) и по всем рёбрам. Для ребра (a,b) начало (ближе к корню) a и приписанная ему структура a, как и конец ребра (дальше от корня) b и приписанная ему структура b, обозначаются одинаково. Такое расстояние называется *специальным* (или: *брейкпоинтовым*). Различаются случаи: без цен (т.е. равные цены) и с ценами, которые должны быть заданы. Для случая равных генных составов и без цен такое расстояние предложено

в работах[1,2]. Задача реконструкции со специальным расстоянием рассматривается и для случая, когда цены от a к b и от b к a могут различаться. В главе 2 задача реконструкции рассматривается только вместе со специальным расстоянием, а в Главе 3 – вместе с паралогами, кратчайшим расстоянием и равными ценами.

Допущение паралогов существенно усложняет задачу реконструкции: проблема в том, что специальное и кратчайшее расстояния требуют решения трудного и самого по себе биологически важного вопроса, какие паралоги соответствуют друг другу, т.е. что значит «тот же самый ген» в a и b. Например, в a имеются два гена с именем n, а в b – три гена с тем же именем; заранее неясно, какой из паралогов в a соответствует, какому из паралогов в b. Поэтому необходимо следующее уточнение в постановке задачи. Множество рёбер с именем n в a и b обозначим соответственно $X(n,a)$ и $X(n,b)$. Нужно для каждого имени n найти частично определённое инъективное отображение меньшего по числу элементов из множеств $X(n,a)$ и $X(n,b)$ в большее («*соответствие* паралогов»), для которого графы a' и b' уже с уникальными именами рёбер имеют минимальное кратчайшее расстояние. Точнее, в a' рёбра, одноимённые в a, получают уникальные имена (и аналогично для b' и b), так чтобы уникальные имена сохранялись при этом отображении, которое меняется в зависимости от ребра. Уникальные имена можно получить, например, добавляя вторую позицию к исходным именам, т.е. уникальное имя будет иметь вид $n.k$ одновременно в a и b, если эти $n.k$ *соответствуют* друг другу при отображении, соответствующем ребру (a,b). Иными словами, отображения сохраняют уникальные имена. Гены, которые не имеют паралогов, могут получить на второй позиции значение 0, которое не используется при нумерации паралогов.

На той же основе, что и две указанные задачи, в диссертации рассматривается задача согласования двух произвольных множеств цепей (двух «линейных» структур). При секвенировании возникает ситуация: для генома найдены контиги, составленные каждый из нескольких генов, которые имеют направления транскрипции. В нашей терминологии *контиг* – цепь, в которой рёбра имеют

[1] Watterson G.A., Ewens W.J., Hall T.E. The Chromosome Inversion Problem // *Journal of Theoretical Biology*. 1982, Vol. 99, P. 1–7.
[2] Blanchette M., Bourque G., Sankoff D. Breakpoint phylogenies. In: S. Miyano, T. Takagi Genome Informatics // *Univ. Academy Press*. 1997, P. 25–34.

направления и имена, может быть, повторяющиеся («паралоги»). Контиги из данного множества соединяют в цепь или цикл; эти варианты, по сути, эквивалентны, и мы рассмотрим второй из них, как это сделано в работе[3]. Итак, пусть даны два множества a и b цепей (множества «контигов»). Множество a (как и b) и соответствующий a (как и b) цикл – хромосомные структуры. Нужно соединить контиги из a в цикл (и аналогично – контиги из b в цикл), так чтобы расстояние между этими циклами было минимальным с учётом выбора соответствия паралогов в циклах. Не ограничивая общность, считаем: каждый контиг оканчивается концом своего гена. Эту задачу назовём ***согласованием контигов***. Её биологическое содержание подробно обсуждается в работе[4].

Развитые нами алгоритмы реализованы соответствующими компьютерными программами, тестированы на искусственных данных и затем применены для построения филогенетических деревьев хромосомных структур *митохондрий* инфузорий и споровиков из класса Aconoidasida, а также – *пластид* родофитной ветви и бактерий рода *Rhizobium*. Напомним: пластиды – полуавтономные органеллы, происходящие от цианобактерий; в родофитной ветви они представлены у красных водорослей (Rhodophyta) и у видов с пластидами вторичного и третичного происхождения от пластид Rhodophyta. Среди них находятся фотосинтезирующие и нефотосинтезирующие виды.

Итак, нами рассмотрена *тема* – разработка алгоритмов для работы с хромосомными структурами и применение алгоритмов и соответствующих компьютерных программ для построения филогенетических деревьев хромосомных структур. Тема представляется актуальной, за последние 30 лет по ней опубликовано сотни статей и появляются всё новые.

Цели работы. Найти линейные или близкие к ним по сложности (не выше кубических) точные алгоритмы решения задач преобразования и реконструкции хромосомных структур или сведения их к задачам целочисленного (или булева) линейного программирования. Аналогично – согласование множеств контигов. Реализовать алгоритмы компьютерными программами. Применить полученные программы для построения филогенетических деревьев митохондрий инфузорий и

[3] Chin Lung Lu. An Efficient Algorithm for the Contig Ordering Problem under Algebraic Rearrangement Distance // *Journal of Computational Biology*. 2015, Vol. 22(11), P. 975–987.

споровиков из класса Aconoidasida, пластид родофитной ветви и бактерий рода *Rhizobium*.

Методы исследования. В работе использованы методы теории алгоритмов и организации вычислений с использованием известных и оригинальных программ, в том числе для параллельных вычислений на суперкомпьютерах, методы математической биологии и биоинформатики. Оригинальный подход, предложенный автором, состоял в определении соответствия паралогов с помощью целочисленного или булева линейного программирования с линейным, квадратичным или (самое большое) кубическим числом переменных и ограничений.

Научная новизна. Полученные алгоритмы и компьютерные программы, как и их применения для построения филогенетических деревьев хромосомных структур митохондрий, пластид и бактерий, являются новыми.

Более подробно. Получен линейной сложности точный алгоритм решения задачи преобразования структур общего вида *без паралогов*.

Получен квадратичной сложности точный алгоритм решения задачи реконструкции структур общего вида *без паралогов* для специального расстояния, равных и неравных цен. Его точность доказана.

Получено решение задачи преобразования произвольных структур *с паралогами* и равными ценами сведением её квадратичным точным алгоритмом к задаче целочисленного линейного программирования (ЦЛП) квадратичного размера. Для циклических хромосомных структур *с паралогами* и равными ценами задача преобразования решена сведением её линейным точным алгоритмом к задаче ЦЛП линейного размера.

Получено решение задачи реконструкции *с паралогами* и любыми ценами для специального расстояния сведением её квадратичным точным алгоритмом к задаче булевого линейного программирования квадратичного размера. Получено решение задачи реконструкции *с паралогами* и равными ценами для произвольных структур общего вида и кратчайшего расстояния сведением её кубическим точным алгоритмом к задаче ЦЛП кубического размера. Везде выше допускается неравный генный состав.

Получено решение задачи согласования множеств контигов с неравным генным составом, паралогами и равными ценами сведением линейным точным алгоритмом к ЦЛП линейного размера.

Разработаны программы для решения задач преобразования и реконструкции, которые позволяют эффективно решать задачи для хромосомных структур, содержащих тысячи генов. Это достигается за счёт использования современных методов распараллеливания программ и технологий хранения и обработки больших данных.

На основе оригинальных алгоритмических и программных решений построены разумные филогенетические деревья хромосомных структур митохондрий инфузорий и споровиков из класса Aconoidasida, пластид родофитной ветви и бактерий рода *Rhizobium*. На их основе выявлены особенности эволюции органелл.

Практическая значимость работы. Работа носит теоретический характер. В то же время, исследование может иметь прикладное значение. Реконструкция хромосомных структур может применяться для анализа хромосомных перестроек, что, в частности, важно при многих заболеваниях: хромосомные перестройки меняют уровни экспрессии генов, что служит одной из причин заболевания. Рассмотренные методы могут применяться при сборке секвенируемых геномов.

Апробация работы. Компьютерные программы тестировались на искусственных примерах с известными ответами; рассмотрено около 100 таких примеров. Программы снабжены удобным для пользователя интерфейсом. Результаты работы опубликованы в 5 статьях и 3 тезисах и докладывались на следующих конференциях:

- 39-я конференция «Информационные технологии и системы»: ИТиС'15 (Сочи, 7–11 сентября 2015);
- Международная конференция “Moscow Conference on Computational Molecular Biology”: MCCMB'17 (Москва, 27–30 июля 2017);
- 57-ая научная конференция МФТИ (Москва, 23-28 ноября 2015).

Работа также докладывалась на научных семинарах механико-математического факультета Московского государственного университета им. М.В.

Ломоносова и на семинаре по Математической биологии и биоинформатике Института проблем передачи информации им. А.А. Харкевича РАН.

Публикации. По теме диссертации опубликовано 5 статей и 3 тезисов докладов на конференциях. Все результаты, включённые в диссертацию, получены лично автором.

Структура и объём работы. Работа состоит из введения, четырёх глав и списка литературы. Список литературы содержит 77 наименований. Объём работы составляет 127 страниц, включая 14 таблиц и 75 рисунков.

Основные положения, выносимые на защиту:

1) Предложен точный алгоритм решения задачи преобразования структур. Глава 1, [1].

2) Для специального расстояния, отсутствия паралогов, равных и неравных цен получен точный квадратичный алгоритм решения задачи реконструкции. В случае того же расстояния, присутствия паралогов и любых цен получен точный квадратичный алгоритм сведения задачи реконструкции к задаче квадратичного булева линейного программирования. Точность обоих алгоритмов доказана. Глава 2, [2].

3) Получен алгоритм решения задачи преобразования с равными ценами: циклических хромосомных структур (сведением к ЦЛП линейного размера) и произвольных структур (сведением к ЦЛП квадратичного размера). Глава 3, [4].

4) Получен алгоритм решения задачи реконструкции с равными ценами произвольных структур сведением к ЦЛП кубического размера. Глава 3, [4].

5) Получен алгоритм решения задачи согласования с равными ценами двух множеств контигов сведением к ЦЛП линейного размера. Глава 3, [4].

6) На основе алгоритмических и программных решений, предложенных в пунктах 1–5, построены разумные филогенетические деревья хромосомных структур митохондрий инфузорий и споровиков из класса Aconoidasida, а также – пластид родофитной ветви у водорослей и споровиков и бактерий рода *Rhizobium*. На их основе обсуждаются особенности эволюции этих органелл и видов. Глава 4, [1-4].

Все результаты диссертации опубликованы.

СОДЕРЖАНИЕ РАБОТЫ

Во **Введении** приведены постановки задач и формулировки в целом полученных автором результатов, приведён обзор наиболее близких к ним публикаций.

В **Главе 1**, [1] рассматривается задача о преобразовании двух данных общего вида структур *a* и *b*, первой ко второй, операциями из фиксированного списка. В главе предполагается, что имена рёбер в отдельно взятой структуре не повторяются, т.е. *паралоги отсутствуют*. Упомянутые операции над хромосомной структурой следующие, первые четыре из них называются *стандартными*, последние две – *дополнительными*.

1) *Двойная переклейка*: расклейка двух пар склеенных краёв генов и переклейка полученных четырёх краёв по-новому;

2) *Полуторная переклейка*: расклейка пары склеенных краёв генов и склейка одного из полученных краёв с каким-нибудь несклеенным («свободным») краем;

3) *Разрез*: расклейка пары склеенных краёв генов (образуются два свободных края);

4) *Склейка*: склейка пары свободных краёв генов.

5) *Удаление* связного участка *a*-генов (т.е. присутствующих в *a* и отсутствующих в *b*).

6) *Вставка* связного участка *b*-генов (т.е. присутствующих в *b* и отсутствующих в *a*).

В разделе 1.1 приводится ключевое определение *общего графа a+b* (которое обобщает определение из работы[4]), а задача преобразования переформулируется (теорема 1) как задача приведения неориентированного графа *a+b* к виду *c+c* для некоторой структуры *c*, что обобщает подход, предложенный в указанной работе. Граф вида *c+c* называется *финальным*.

В разделе 1.2 описывается оригинальный линейный по сложности точный алгоритм решения задачи преобразования. Его точность доказана[5], в частности, для

[4] Alekseyev M.A., Pevzner P.A. Multi-Break Rearrangements and Chromosomal Evolution // *Theoretical Computer Science*. 2008, Vol. 395, № 2-3, P. 193–202.

[5] Gorbunov K.Yu., Lyubetsky V.A., A linear algorithm for the shortest transformation of graphs with different operation costs // *Journal of Communications Technology and Electronics*. 2017, Vol. 62, No. 6, P. 653–662.

случая: цена d операции вставки находится в интервале $(c, 2c)$, где c – равные цены оставшихся операций. Аддитивная константа k из формулировки задачи равна $k=d-c$, что при $c=1$ не превышает 1. Из алгоритма сразу следует новый точный линейный алгоритм решения задачи преобразования для случая равных цен операций.

В разделе 1.3 приведено тестирование этого алгоритма на искусственных примерах.

В **Главе 2**, [2-3, 6] решается задача реконструкции структур вдоль данного дерева. В разделе 2.1, [6] приводится постановка задачи с *паралогами* и, возможно, с ценами для *специального* и *кратчайшего* расстояний между структурами на концах рёбер дерева.

В разделе 2.2, [2] в случае специального расстояния и отсутствия паралогов, равных и неравных цен операций, приводится квадратичный точный по времени работы и используемой памяти алгоритм решения задачи реконструкции. Доказывается его точность (теорема 2 для равных цен и теорема 3 для неравных).

В разделе 2.3, [6] в случае специального расстояния и присутствия паралогов, любых цен, приводится квадратичный точный по времени работы и используемой памяти *алгоритм сведения* этой задачи к квадратичному булевому линейному программированию. Точнее, число переменных и ограничений зависит биквадратично (т.е. в 4й степени) от суммарного числа паралогов в листьях, но при фиксированном числе паралогов – квадратично от суммарного размера исходных графов в листьях, например, от суммарного числа рёбер в них.

В [5, 8] содержится оригинальный подход к решению задачи ЦЛП, однако для реального счёта этой задачи использовались стандартные пакеты ЦЛП IBM CPLEX (https://www.ibm.com/ru-ru/marketplace/decision-optimization-cloud), и Ipot (https://projects.coin-or.org/Ipopt).

В разделе 2.4 приведено тестирование алгоритмов из этой главы на искусственных примерах.

В **Главе 3**, [1, 4] рассматриваются четыре задачи, тесно связанные между собой и с задачами из Глав 1–2 как по методу, так и по постановке. Здесь всюду допускаются неравный генный состав и присутствие паралогов, но предполагаются

Gorbunov K.Yu., Lyubetsky V.A., Linear algorithm for minimal rearrangement of structures // *Problems of information Transmission*. 2017, Vol. 53, No. 1, P. 55–72.

равные цены. Структура называется *циклической*, если она состоит из одних циклов (кольцевых хромосом).

В разделе 3.1, [1] содержится линейный точный алгоритм сведения задачи преобразования для циклических структур к линейной ЦЛП. Точнее, число переменных и ограничений квадратично зависит от суммарного числа паралогов в листьях, а при фиксированном числе паралогов – линейно от суммарного размера исходных графов в листьях.

В разделе 3.2, [4] содержится квадратичный точный алгоритм сведения задачи преобразования для произвольных структур к квадратичному ЦЛП.

В разделе 3.3, [4] содержится кубический точный алгоритм сведения задачи реконструкции для произвольных структур к кубическому ЦЛП.

Наконец, в разделе 3.4, [4] содержится линейный точный алгоритм сведения задачи согласования двух произвольных множеств цепей (двух «линейных» структур) к линейному ЦЛП. Точнее, число переменных и ограничений квадратично зависит от суммарного числа паралогов в данных структурах и от суммарного числа цепей; если эти значения фиксированы – линейно от суммарного размера структур. В работе[3] предложен почти линейный (т.е. со сложностью $n \cdot f(n)$, где $f(n)$ – обратная функция Аккермана) алгоритм, который является точным решением задачи при условии: равный генный состав двух множеств контигов и отсутствие паралогов в них. Здесь предлагается решение задачи без этого условия, за счёт допущения паралогов задача становится NP-трудной. В работе[3] решение основано на алгебраической теории перестановок и использует расстояние, которое отличается от расстояния, определённого выше.

В разделе 3.5 приведено тестирование алгоритмов из этой главы на искусственных примерах.

В **Главе 4**, [1, 2, 3] алгоритмы, развитые в Главах 1 и 3, и соответствующие компьютерные программы применяются для построения филогенетических деревьев хромосомных структур митохондрий инфузорий (Ciliophora) и споровиков из класса Aconoidasida, пластид родофитной ветви, а также бактерий рода *Rhizobium*. Используются результаты 1й главы без паралогов с разными ценами и результаты 3й главы с паралогами и равными ценами. Для полученных деревьев обсуждаются особенности эволюции соответствующих митохондрий и пластид; в целом

полученная реконструкция не расходится с принятыми представлениями об их эволюции. Отметим следующие особенности полученных деревьев [7].

1) Отличия между деревьями, построенными по белкам, кодируемым в митохондриях Ciliophora, и по хромосомным структурам незначительны и состоят в различном взаимном расположении видов рода *Tetrahymena*; в каждом из этих двух деревьев род *Tetrahymena* образует кладу, [3].

2) Отличие дерева хромосомных структур митохондрий споровиков от общепринятого дерева видов состоит в перемешивании видов близких родов *Leucocytozoon* и *Plasmodium* между собой; вместе виды этих родов образуют кладу, [1]. В этом поддереве виды с линейными и кольцевыми хромосомами собрались в две соответствующие клады.

3) Дерево пластид родофитной ветви, построенное на основе хромосомных структур ([3], рисунок 6), в целом согласуется с деревом, ожидаемым на основе других данных (по белкам, рРНК, высококонсервативным элементам и др.). В некоторых пластидах наблюдается значительная перестройка хромосомной структуры. В пластидах хромерид *Chromera velia, Vitrella brassicaformis* (синоним *Chromerida RM11*) и багрянки *Porphyridium purpureum* взаимное расположение генов на хромосоме существенно отличается от такового в любых других пластидах. На том же дереве багрянки *Porphyridium purpureum*, как и *Choreocolax polysiphoniae*, отделились от других видов багрянок и друг от друга. Другие багрянки вместе с криптофитовыми водорослями образовали единую кладу. Виды *Chromera velia* и *Vitrella brassicaformis* расположены близко друг к другу, но далеко от других представителей надтипа Alveolata. Споровик *Babesia bovis* отделился от других споровиков и оказался в кладе, образованной видами родов *Nannochloropsis* и *Trachydiscus*.

Эти особенности можно связать как с особенностями эволюции митохондрий и пластид, так и с относительно малым объёмом данных о них.

ВЫВОДЫ

Найдены линейные или близкие к ним по сложности (не выше кубических) точные алгоритмы решения задач преобразования и реконструкции хромосомных структур, согласования множеств контигов, которые реализованы компьютерными программами. Это – прямые алгоритмы или алгоритмы сведения к вариантам ЦЛП

небольшого размера (не выше кубического). Полученные алгоритмы и программы применены для построения филогенетических деревьев митохондрий инфузорий и споровиков из класса Aconoidasida, пластид родофитной ветви, а также бактерий рода *Rhizobium*. Подробнее результаты описаны выше, в разделе «Основные положения, выносимые на защиту».

ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ:

1. Lyubetsky V.A., **Gershgorin** R.A., Seliverstov A.V., Gorbunov K.Yu. Algorithms for Reconstruction of Chromosomal Structures // *BMC Bioinformatics*. 2016, Vol. 17, № 40, 23 pages. DOI: 10.1186/s12859-016-0878-z.
2. Горбунов К.Ю., **Гершгорин** Р.А., Любецкий В.А. Перестройка и реконструкция хромосомных структур // *Молекулярная биология*. 2015, Т. 49, № 3, С. 372–383. DOI: 10.7868/S0026898415030076. Перевод: Gorbunov K.Yu., **Gershgorin** R.A., Lyubetsky V.A. Rearrangement and Inference of Chromosome Structures // *Molecular Biology*. 2015, Vol. 49, № 3, P. 327–338. DOI: 10.1134/S0026893315030073.
3. **Gershgorin** R.A., Gorbunov K.Yu., Zverkov O.A., Rubanov L.I., Seliverstov A.V., Lyubetsky V.A. Highly Conserved Elements and Chromosome Structure Evolution in Mitochondrial Genomes in Ciliates // *Life*. 2017, Vol. 7(1). DOI: 10.3390/life7010009.
4. Lyubetsky V.A., **Gershgorin** R.A., Gorbunov K.Yu. Chromosome structures: reduction of certain problems with unequal gene content and gene paralogs to Integer linear programming // *BMC Bioinformatics*. 2017, Vol. 18, № 537, 18 pages.
5. **Gershgorin** R.A., Rubanov L.I., Seliverstov A.V. Easily Computable Invariants for Hypersurface Recognition // *Journal of Communications Technology and Electronics*. 2015, Vol. 60, № 12, P. 1429–1431. DOI: 10.1134/S1064226915120074.

ТЕЗИСЫ ДОКЛАДОВ:

6. **Gershgorin** R.A., Gorbunov K.Yu., Seliverstov A.V., Lyubetsky V.A. Evolution of Chromosome Structures //*Proceedings of the 39th IITP RAS Interdisciplinary Conference & School "Information Technology and Systems 2015" (ITaS'15)*, Sochi, Russia, Sep 7–11 2015.

7. Lyubetsky V.A., **Gershgorin** R.A., Rubanov L.I., Seliverstov A.V., Zverkov O.A. Evolution and Systematics of Plastids of Rhodophytic Branch // *Proceedings of the International Moscow Conference on Computational Molecular Biology* (MCCMB'17), Moscow, Russia, July 27–30, 2017, 4 стр.
8. **Гершгорин** Р.А., Латкин И.В., Селиверстов А.В. Следы форм высших степеней //*Труды 57-й научной конференции МФТИ*, Москва–Долгопрудный–Жуковский, 24–29 ноября 2014, Управление и прикладная математика. Т. 1, М.: МФТИ, 2014, С. 11–12.