Санкт-Петербургский национальный исследовательский университет
информационных технологий, механики и оптики

Факультет информационных технологий и программирования
Кафедра компьютерных технологий

Авдеев Павел Вадимович

Улучшение алгоритма реконструкции родительских геномов и разработка алгоритма построения филогенетических деревьев

Научный руководитель: Алексеев Максим Александрович

Санкт-Петербург
2013

Содержание

Введение

Благодаря развитию и удешевлению технологий секвенирования, появлению разнообразных геномных ассемблеров, а так же разнообразных инициатив [1], призывающих секвенировать как можно больше организмов, количество информации о геномах различных организмов значительно увеличилось. В результате возникла возможность более точно выявлять закономерности организации и эволюции геномов. Так, в области филогенетики, идентифицирующей и проясняющей эволюционные взаимоотношений среди разных видов жизни на Земле, как современных, так и вымерших, появились новые методы наблюдения эволюционных событий, основанные на сравнении геномов, как последовательностей синтени-блоков. Эволюционные изменения генома, представленными последовательностями синтени-блоков, можно охарактеризовать геномными перестройками.
Таким образом на основе набора родственных геномов организмов можно пытаться ответить на следующие вопросы:

- В какое наиболее вероятное дерево организованы организмы, отражающее эволюционные взаимосвязи между ними и их предками.
- Как выглядели геномы их предков.
- Какие события произошли на этапе эволюции, которые привели геномы предков к текущим геномам организмов.

К сожалению, все существующие алгоритмы имеют ограничения и не отвечают наиболее полно на все эти вопросы. Например, GRAPPA[ссылка] алгоритм и MGR [2] алгоритм не различают надежные и ненадежные перестановки и акцентируются на кратчайшем расстояние между геномами. Алгоритм inferCARS [3] предполагает заданное филогенетическое дерево, которое на основе этих данных еще как-то надо получить. EMRAE[ссылка] алгоритм имеет существенные ограничения, так

как не пытается реконструировать филогенетические деревья, а так же ограничен однохромосомными геномами. Но для всех перечисленных алгоритмов наложено ограничение, исключающее наличие дуплицироанных генов, которое приемлимо для большинства вирусов и митохондрий, или для очень близкородственных геномов, где сохраняются целые участки хромосом, а не отдельные гены. Поэтому часто ограничение на отсутстиве дуплицируемых блоков не выполняется. Так как в ходе эволюции какие-то учатски генома предка могут дуплицироваться, геномы-потомки могут содержать несколько экземпляров некоторых синтени-блоков: изначальные экземпляр, сохранившийся от общего предка, и его копии.

В данной работе описывается обобщение MGRA[ссылка] алгоритма для данных, содержащих дуплицируемые синтени-блоки, улучшение результатов работы алгоритмов на данных, не содержащих дуплицируемые данные, а так же приводится алгоритм реконструкции филогенетических деревьев. Так же дается описание сервера, который был разработан для облегчения работы биологов с данными алгоритмами.

Глава 1. Обзор предметной области

1.1. Биоинформатика

Во многих задачах современной биологии и медицины необходимо работать с большими объемами данных, поэтому для их решения используется вычислительная техника. Таким образом биоинформатика представляет из себя междисциплинарную область науки, которая разрабатывает и усовершенствует методы хранения, поиска, организации и анализа биологических данных.

1.2. Общие сведения о хромосомной эволюции

Перестановки - это геномные "землетрясения изменяющие хромосомную архитектуру. Фундаментальным вопросом в молекулярной эволюции существуют ли "горячие"регионы в хромосомах, где перестановки случаются снова и снова, благодаря высокой вероятности разрыва на этом участке.

1.2.1. Random Breakage Model против Fragile Breakage Model в хромосомной эволюции

В 1970 году Сусуму Оно привел две фундаментальные гипотезы хромосомной эволюции, которые были предметом споров в последние 40 лет. Одна из них Random Breakage Model предложенная Оно [58 в phd] и формализованная Nadeau и Taylor[57 в phd].

Гипотеза 1.1. *Random Breakage Model(RBM) постулирует возможность перестановок в случайных геномных позициях, таким образом подразумевая низкое переиспользование перестановок в конкретных регионах генома.*

Из-за своей пророческой силы предсказания, RBM стало де-факто теории хромосомной эволюции. Только в 2003 году Певзнер и Теслер[в phd 65] опровергли RBM и предложили альтернативную модель хромосомной эволюции Fragile Breakage Model.

Гипотеза 1.2. *Fragile Breakage Model(FBM) постулирует, существование "хрупких"геномных регионов, которые с большей вероятностью подвержены перестановкам, чем остальные части генома, подразумевая высокий уровень повторного использования перестановок в этих "горячих"точках.*

Различные дополнительные исследования подтвердили существование хрупких регионов у огранизмов. Например, Кикута[44 в phd] проанализировал связь между хрупкими участками генома необходимость сохранения нетронутыми регуляторных элементов генома и пришел к выводу, что модель RBM ошибочна.

Так как модель FBM выполнена, то существуют консервативные сегменты генома, которые не подвержены перестановкам. Такие регионы называются блоками синтении. Так как геном состоит из одной или более хромосом, которая содержит наследственную информацию в молекуле ДНК. ДНК имеет двухцепочечную структуру, и транскрипция с основной и комплементарной цепей осуществляется в противоположных направлениях, поэтому хромосому можно представить, как знаковую перестановку синтени-блоков. Знак “+” будет соответствовать прямой ориентации, а знак “-” обратной. Существует достаточное количество алгоритмов[GRIMM алгоритм, Sibelia, Drimm ссылки], которые позволяют на основе геномов, состоящих из последовательности нуклеотидов, получать геномы представленные набором блоков синтении.

1.2.2. Операции перестановок

Если геном представляет собой набор из блоков синтени, описанные в разделе 1.2.1, тогда можно определить возможные геномные перестройки

как операции над знаковой перестановкой:

1. Инверсия или reversal

$$\pi = (p_1 \ \dots \ p_{i-1}p_i \ \dots \ p_jp_{j+1} \ \dots \ p_n)$$

$$\Downarrow$$

$$\pi' = (p_1 \ \dots \ p_{i-1}p_j \ \dots \ p_ip_{j+1} \ \dots \ p_n)$$

2. Транслокация или translocation

$$\pi = (p_1 \ \dots \ p_{i-1}p_i \ \dots \ p_n) \ \sigma = (s_1 \ \dots \ s_{j-1}s_j \ \dots \ s_m)$$

$$\Downarrow$$

$$\pi' = (p_1 \ \dots \ p_{i-1}s_j \ \dots \ s_n) \ \sigma' = (s_1 \ \dots \ s_{j-1}p_i \ \dots \ p_n)$$

3. Слияние или fission

$$\pi = (p_1 \ \dots \ p_n) \ \sigma = (s_1 \ \dots \ s_m)$$

$$\Downarrow$$

$$\pi' = (p_1 \ \dots \ p_ns_1 \ \dots \ s_m)$$

4. Расщепление или fusion

$$\pi = (p_1 \ \dots \ p_{i-1}p_i \ \dots \ p_n)$$

$$\Downarrow$$

$$\pi' = (p_1 \ \dots \ p_{i-1}) \ \sigma' = (p_i \ \dots \ p_n)$$

Стандартные перестановочные операции приведенные выше, могут быть обобщены введением так называемых 2 - break операций в геноме.

Определение 1.3. 2 - break - это операция которая производит разрыв в двух местах генома и склеивает получившиеся фрагменты в новом порядке.

Можно предположить о существование k - break операции в ходе эволюции.

Определение 1.4. k - break - это операция которая делает k разрывов в геноме и склеивает получившиеся фрагменты в новом порядоке.

Множество биологов верят, что k - break перестановки маловероятны для $k > 3$ и относительно редко для $k = 3$. Действительно, биофизические и селективные ограничения серьезны уже для $k = 2$, не говоря уже о $k > 2$. Однако, 3 - break операция, называмая транспозицией, несомненно случается в эволюции, хотя до сих пор не ясно, как часто она происходила в эволюции разнообразных организмов.

- Транспозиция или transposition

$$\pi = (p_1 \ \dots \ p_{i-1}p_i \ \dots \ p_jp_{j+1} \ \dots \ p_kp_{k+1} \ \dots \ p_n)$$

$$\Downarrow$$

$$\pi' = (p_1 \ \dots \ p_{i-1}p_{j+1} \ \dots \ p_kp_i \ \dots \ p_jp_{k+1} \ \dots \ p_n)$$

Задача о нахождении сценария геномных перестроек эквивалентна поиску последовательности приведенных выше операций, преобразующий один набор знаковых перестановок к другому. Так как крупные геномные перестройки на уровне популяций происходят редко, биологи заинтересованы в нахождении кратчайшей последовательности перестроек между геномами разных видов. Тем не менее, в случае сложных комбинаций перестроек, даже кратчайших сценариев может быть несколько. Поэтому будем считать решением этой задачи нахождение любого возможного кратчайшего сценария.

1.2.3. Whole Genome Duplication модель

Whole Genome Duplication модель это вторая гипотеза предложенная Сусуму Оно, которая постулирует новый тип эволюционных событий, которые дублируют некоторый регион генома. Эта гипотеза была предметом споров в течение долгих лет и только в 2004 году было доказана корректность этого утверждения. Kellis[43 в phd] рассмотрел геном дрожжей K.waltii и сравнил с геномом дрожжей S.cerevisiae, и продемонстрировал, что почти каждый регион в R.waltii соотносится с двумя регионами в S.cerevisiae, тем самым доказав, что было целое множество событий дубли-

рования геномов в ходе эволюции дрожжей. За этим открытием быстро последовали открытия дупликации генома у позвоночных и растений. Dehal и Boore[24 в phd] нашли доказательство существования двух этапов дупликации генома на эволюционном пути от ранних позвоночных к человеку. Вскоре после этого Meyer and Van de Peer[54 в phd] нашли этап геномных дупликаций у лучеперых рыб.

Эти недавние исследования обеспечивают неопровержимые доказательства, что whole genome duplication представляет новый тип событий, который может объяснить феномены, которые классическое эволюционное учение с трудом объясняет и поэтому очень важно научить существующие алгоритмы работать с такими геномами.

1.3. Парный breakpoint граф

Впервые, определение breakpoint графа было введено Ханхали и Певзнером[36 и 35 в phd], для разработки полиномиального алгоритма вычисления расстояния реверсиями или reversal distance между двумя знакомыми перестановками.

Определение 1.5. reversal distance — минимальное количество реверсий, необходимых для преобразования одной перестановки в другую.

Так как геном, согласно разделу 1.2.1, является знаковой перестановкой, то примение представления генома ввиде граф возможно. Для простоты понимания, начнем изучения графа в предположение, что геном состоит из одной циклической хромосомы, позже распостранив представление графа для линейных хромосом и мультихромосомных геномов.

1.3.1. В случае циклической хромосомы в унихромосомном геноме

Будем представлять геном P состоящий из одной циклической хромосомы, сформированный блоками синтении $x_1 \dots x_n$, как цикл с n направленными ребрами(соответствующие блокам) и с n ненаправленными

непомеченными ребрами(соответсвующие соединением блоков). Пример такого представления генома проиллюстрирован на рис. 1.1. Направление ребер соответсвует знакам блоков. Мы маркируем начало и конец каждого ребра x_i, как x_i^h и x_i^t соответственно. Вершины в хромосоме соединенные ненаправленными ребрами называются прилегающими.

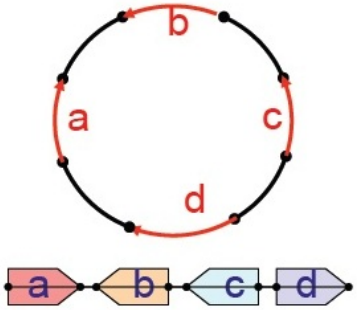

Рис. 1.1: Представление генома ввиде цикла

Пусть P и Q это циклические знаковые перестановки(унихромосомные геномы) над одним и тем же набором блоков синтении Δ.

Определение 1.6. Breakpoint граф $G(P, Q)$ - это граф на наборе вершин $V = \{x^t, x^h | x \in \Delta\}$ с ребрами трех цветов: пунктиные(соединяющие x_i^h и x_i^t), черные(соединяющие прилегающие блоки в геноме P) и серые или зеленые(соединяющие прилегающие блоки в геноме Q).

Пример графа показан на рис. 1.2

Можно заметить, что каждая пара ребер задают переменные циклы(цвет ребер чередуется) в графе G. Черные и серые(зеленые) ребра формируют переменный черно - серый(черно - зеленый) цикл, который играет важную роль в анализе перестановок.

1.3.2. В случае линейной хромосомы в унихромосомном геноме

Теперь, пусть геном P состоит из одной линейной хромосомы, сформированной знаковыми блоками синтении $x_1 \dots x_n$. Будем представлять ге-

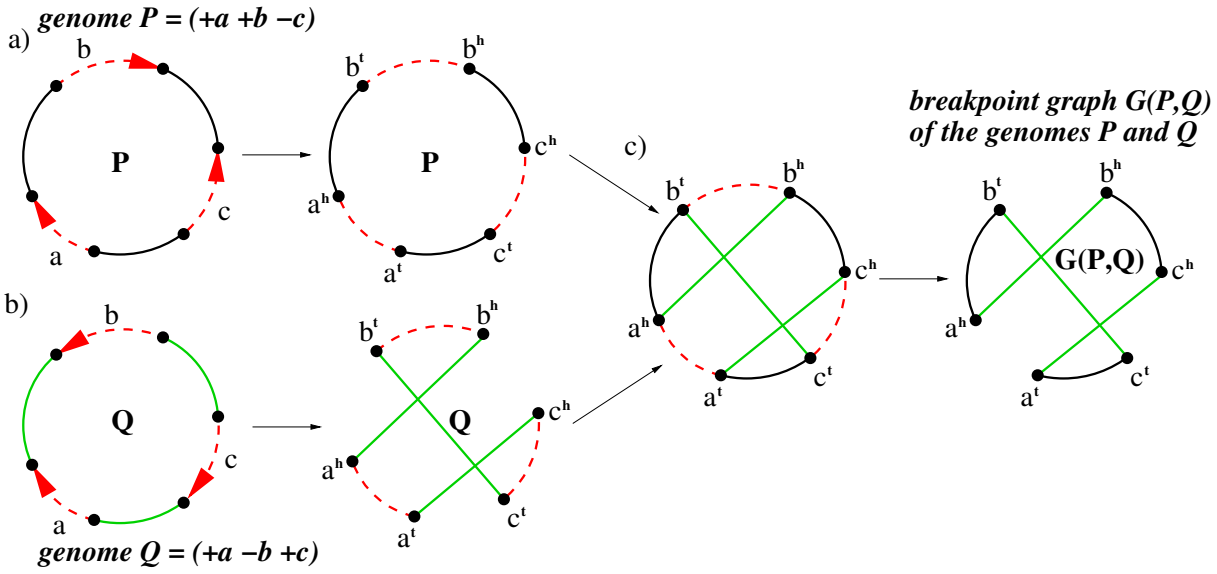

Рис. 1.2: Пример breakpoint графа для двух геномов

ном, как путь из n – пунктирных ребер(кодирующие блок и его направление), а так же с $n-1$ ненаправленными черными ребрами(соединяющие прилегающих блоки). Так же введем новую вершину ∞ и соединим ненаправленными (нерегулярными) черными ребрами с каждыми вершинами представляющими концы хромосомы.

Определение 1.7. Точка в breakpoint графе называется регулярной, если она не является вершиной ∞.

Определение 1.8. Ребро называется регулярным, если обе вершины инцидентные этому ребру регулярные (Ребро называется нерегулярным в остальных случаях).

В таком представление геном состоящий из одной линейной хромосомы - это путь из черных и пунктирных ребер, начинающихся и кончающихся в вершине ∞.

1.3.3. В случае мультихромосомного генома

Расширим наше определение breakpoint графа для генома, состоящего из любого количества хромосов. Breakpoint граф для мультихромосомного генома, отличается от графа унихромосомного генома только

тем, что теперь содержит коллекцию непересекающихся циклов (хромосом) с двумя чередующимися цветами: черный, для ненаправленных ребер и пунктирные цвет для направленных ребер. Мы не будем явно показывать направление ребер, так как они определяются индексом t или h. Отдельно стоит отметить, что в случае линейных хромосом, степень вершины ∞ в breakpoint графе в два раза больше числа хромосом.

1.3.4. Операции над парным breakpoint графом

Для того чтобы проводить дальнейшие филогенетические исследования, нам необхожимо определить геномные перестройки введеные в разделе 1.2.2 для breakpoint графа. Дадим определение, введеной 2-break операции в терминах breakpoint графа:

Определение 1.9. Для любых двух черных ребер графа (u, v) и (x, y) в графе (геноме) P мы определим 2 - break операцию, как замену этих ребер, либо на пару ребер (u, x) и (v, y), либо пару (u, y) и (v, x).

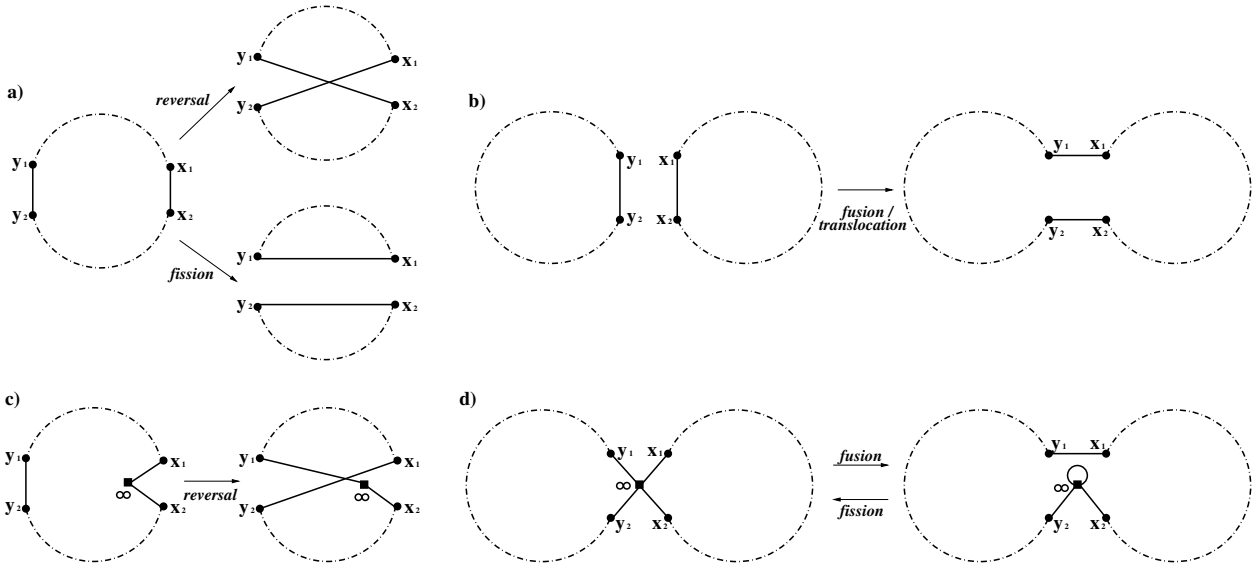

Рис. 1.3: Примеры 2 - break операций на breakpoint графе

Примеры всех таких 2 - break, соответствующие стандартным перестановочным операциям: инверсия, транслокация, деление, слияние, приведены на рис. 1.3. Ключевым наблюдением в исследование парных геномных перестановок является то, что каждая 2 - break трансформация из “черно-

го” генома P в “зеленый“("серый") геном Q, соответствует трансформации парного breakpoint графа G(P, Q) в идентичный breakpoint граф G(Q, Q) 2 - breakами(см рис. 1.4).

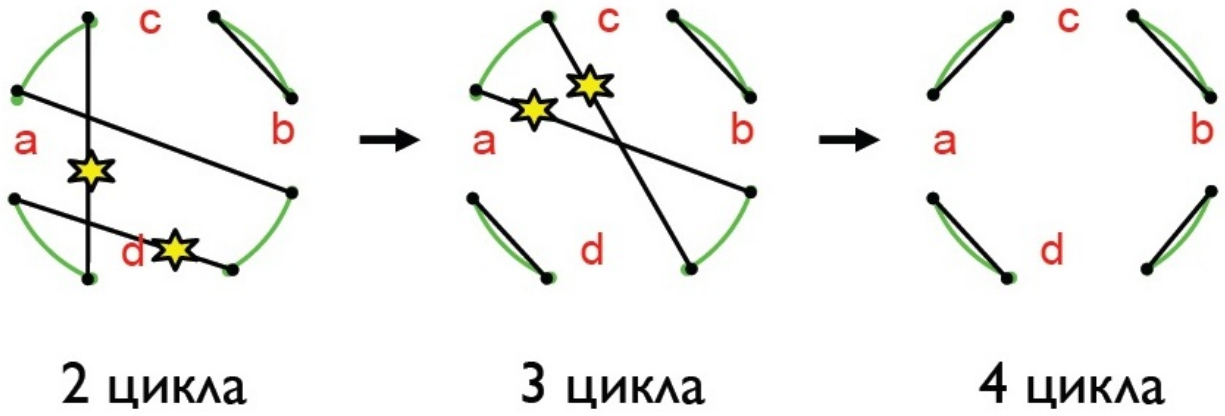

Рис. 1.4: Пример трансформации парного breakpoint графа G(P, Q) в индентичный breakpoint граф G(Q, Q)(самый правый граф).

Логично пытаться посчитать хотя бы минимальное количество 2 - break операций, требующиеся для трансформации одного генома в другой. Введем определения расстояние между двумя геномами:

Определение 1.10. 2 - break расстояние $d_2(P, Q)$ между геномами P и Q определяется, как минимальное число 2 - break, требующихся для трансформации одного генома в другой.

Но как оказалось такое расстояние в терминах парного breakpoint графа можно находить за полиномиальное время[ссылка на все что связано с этим][], так как существует достаточно простая формула дающая это значение:

$$d_2(P, Q) = b(P, Q) - c(P, Q)$$

где $b(P, Q) = |\Delta|$ - это количество блоков синтении в P и Q, и $c(P, Q)$ – это количество черно-зеленых(черно-серых) циклов в G(P, Q).

Стоит отдельно отметить, что в случае линейных хромосом 2 - break операции, включает нерегулярные ребра (см. рис. 1.3), затрагивающие концы хромосом. В таком случае анализ стандартных операций, создает дополнительные алгоритмические проблемы по сравнению с анализом 2 - break в

циклических хромосомах. Однако, сценарий перестановок в линейных хромосомах очень хорошо аппроксимируется сценарием в циклических хромосомах. Таким образом, при использование 2 - break операций в циклических геномах нужно учитывать, что это может привести к нарушению линейности хромосом и создание циклических хромосом.

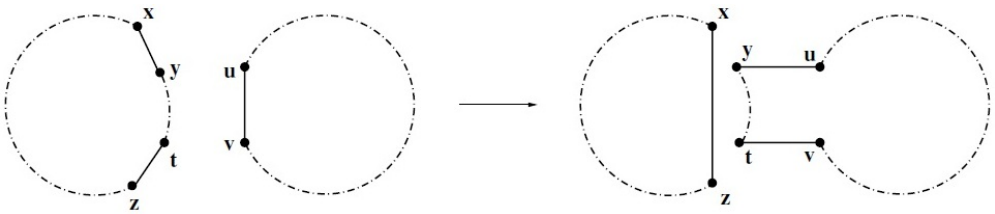

Рис. 1.5: Пример 3 - break операций на breakpoint графе

В разделе 1.2.2 вводилась еще одна операция, называемая транспозицией, которая соответствовала 3 - break операции. Дадим ее определение в терминах breakpoint графа:

Определение 1.11. Для любых трех черных ребер графа (u, v) и (x, y) и (z, t) в графе (геноме) P мы определим операцию транспозиции, как замену этих ребер, либо на тройку ребер (z, x) и (u, y) и (t, v), либо на тройку ребер (z, x) и (y, v) и (u, t).

Пример, такой операции приведен на рис. 1.5. Стоит отметить, что общее определение для 3 - break операций, включает в себя и 2 - break операции. Разные оссобенности 3 - break операции и нахождение расстояния 3 - break расстояния $d_3(P, Q)$ между геномами P и Q здесь рассммотрено не будет, так как в дальнейшем это не используется. Подробную информацию об этом можно прочесть в этих статьях [] [ссылки, много ссылок].

1.4. Breakpoint граф для произвольного количества геномов

В разделе 1.3 breakpoint граф был ограничен двумя геномами. MGRA алгоритм активно использует breakpoint граф построенный на любом количетсве геномов, поэтому ниже приводится обобщение breakpoint графа для n геномов, где $n \geq 2$.

1.4.1. Множественный breakpoint граф

Пусть нам даны произвольные геномы $P_1 \dots P_n$ над одним и тем же множеством блоков синтении Δ. Точно так же, как для парного брейкпоинт графа, множественный breakpoint граф $G(P_1, \dots P_n)$ – является суперпозицией геномов (графов) $P_1, \dots P_n$ над одним и тем же множеством вершин $V = \{b^t, b^h | b \in \Delta\} \cup \{\infty\}$.

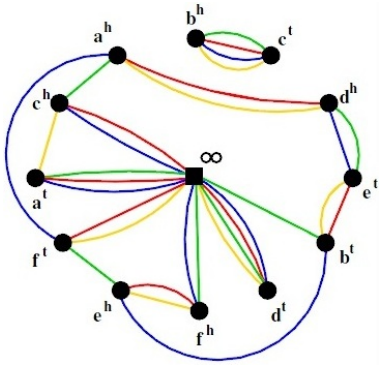

Рис. 1.6: Множественный breakpoint граф $G(P_1, P_2, P_3, P_4)$

Рисунок 1.6 демонстрирует нам множественный breakpoint граф. Введем ряд определений для удобства дальнейшего использования множественного breakpoint графа.

Ребра из $G(P_1, \dots P_n)$ представлены ненаправленными ребрами от геномов $P_1, \dots P_n$ с n различными цветами (следовательно, степень каждой регулярной вершины – k). Для упрощения обозначений, мы используем $P_1, \dots P_n$ для обозначения цветов ребер в множественном breakpoint графе и обозначим множество всех цветов $C = \{P_1 \dots P_n\}$.

Определение 1.12. Любое не пустое подмножество множества C называется мультицветом или multicolor.

Все ребра соединяющие вершины x и y в множественном breakpoint графе формируют мультиребро или multi-edge (x, y) с мультицветом, состоящим из цветов этих ребер. Мультиребра, соответствуют соседним блокам синтении, которые сохраняются на несколько видов существ и представляют ценные филогенетические характеристики. На рис. **??** мультиребро

(e^h, f^h) имеет мультицвет (P_3, P_4), представленные красными и желтыми ребрами.

Определение 1.13. Число мультиребер инцидентных вершине (так же равны числу соседних вершин) называется мультистепенью.

Заметим, что мультистепень может быть меньше обычной степени вершины. Например, на рис. 1.6 вершина e^h имеет степень 4, а мультистепень равна 3.

Так же как и разделе 1.3.4 дадим определение множественного индентичного breakpoint графа, к которому мы хотим привести наш изначальный множественный breakpoint граф.

Определение 1.14. Множественный breakpoint граф называется идентичным breakpoint графом $G(X, \ldots X)$ некоторого генома X если он состоит из полных мультиребер, то есть мультиребра из мультицвета (мультистепень каждой веришны равна единице).

1.4.2. Операции над множественным breakpoint графом

В случае $n \geq 2$ геномов $P_1, \ldots P_n$ на множественном breakpoint графе $G(P_1, \ldots P_n)$ существует $(2^n - 2)$ видов 2 - breakов, столько же, сколько различных мультицветов, сформированных собственными подмножествами из множества C. Каждый такой 2 - break может быть применен к мультиребрам. Однако, не каждая серия таких 2 - breakов имеет смысл с точки зрения реконструкции предковых геномов. Базовым свойством для реконструкции предка геномов, является то, что 2-break на мультиребро с мультицветом $Q \in C$, может быть применена только тогда, когда все геномы, соответствующие цветам из Q могут быть объединены в единый предковый геном. Дадим другое определение такой серии 2 - break:

Определение 1.15. Трансформация (серия 2 - breakов) S на множественном breakpoint графе $G(P_1, P_2, \ldots P_n)$ является надежной, если для любой пары 2-break операций (ρ_1 и ρ_2) на мультиреберах из мультицветов Q_1 и Q_2 выполнено $Q_1 \in Q_2$ тогда ρ_1 предшествуют ρ_2 в S.

1.5. Вывод по главе 1

Глава 2. Постановка задачи

2.1. Теоретическая постановка задачи

Как отмечалось в разделе 1.3.4, ... Предположим, что в отчилие от стандартного анализа breakpoint графа, базирующегося только на 2 - break операциях на черных ребрах, у нас возможны 2 - break операции либо на черных, либо на зеленых(серых) ребрах. Из этого предположения получаем, что происходит приведение breakpoint графа $G(P, Q)$ не к индентичному breakpoint графу $G(Q, Q)$, а к некому индентичному breakpoint графу $G(X, X)$. Но наша серия 2 - break операций над черными и зелеными(серыми) ребрами, соответсвует трансформации $P \to X \to Q$ с помощью m 2 - break операций на черных ребрах. Переход от 2 - break операций на черных ребрах к смеси 2 - break операций на черных и зеленых(серых) ребрах является простой, но мощной парадигмой, которая оказалась полезной в предыдущих исследованиях [] [ССЫЛКА, ССЫЛКА. В статье.]. Поэтому, вместо поиска кратчайшей трансформации $G(P, Q) \to G(Q, Q)$ будем искать кратчайшую трансформацию $G(P, Q)$ в любой идентичный breakpoint граф $G(X, X)$ без знаний о графе(геноме) X заранее.

Аналогично для множественного breakpoint графа $G(P_1, P_2, \ \dots \ P_n)$ сосредоточимся на поиске кратчайшей трансформации в любой идентичный множественый breakpoint граф $G(X, X \ \dots \ X)$ для некоторого неизвестного генома(графа) X. Формализуем множественную геномно перестановочную проблему или Multiple Genome Rearrangement problem(MGRP) в терминах breakpoint графов следующим образом:

Теорема 2.1. *Multiple Genome Rearrangement problem*
Для данных геномов $P_1, P_2 \ \dots \ P_n$ найти кратчайшую надежную серию из 2-break операций, которая трансформирует множественный breakpoint граф $G(P_1, P_2 \ \dots \ P_n)$ в идентичный множественный breakpoint граф.

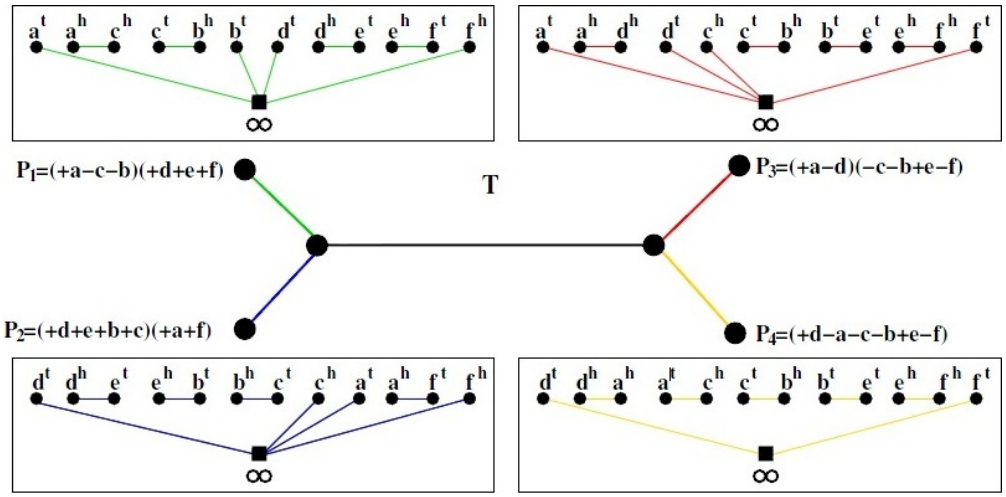

Рис. 2.1: Пример некорневого филогенетического дерева в листьях, которого содержатся геномы

Рассмотрим ситуацию когда есть информация о некорневом филогенетическом дереве T из геномов $P_1, \ldots P_n$ (см. рис. 2.1). Дерево T состоит из n – листьев, $(n-2)$ – внутренних узлов и $(2n-3)$ – ветвей соединяющих пары из узлов. Степень каждого листа равна одному, а степень каждого внутреннего узла равна трем.

Удаление ветви из дерева T разрушает его на два поддерева, каждое из которых индуцировано множеством своих листьев.

Определение 2.2. Мультицвет(multicolor) называется T - согласованным, если он состоит из всех цветов (листьев) любого такого индуцированного поддерева.

Заметим, что если мультицвет Q является T - согласованным, то его дополнение $\overline{Q} = C \setminus Q$ так же T - согласованно. Поэтому, существует взаимно - однозначное соответствие между парами комплементарных T - соглсованных мультицветов и ветвями из T(см. рис. 2.2). Данное наблюдение используется при алгоритме реконструкции деревьев описанного в разделе **??**. Получаем что когда филогенетическое дерево дано, MGRA алгоритм решает урезанную версию MGRP, где 2 - break операции применяются только к мультицветам, соответствующие T - согласованным цветам в филогенетическом дереве. Сформулируем дерево согласованную множесвенную геномно перестановачную проблему или The Tree-Consistent Multiple Genome Rearrangement problem (TCMGRP):

Теорема 2.3. *The Tree-Consistent Multiple Genome Rearrangement problem Для данных геномов $P_1, P_2 \ \dots \ P_n$, как листьев филогенетического дерева T, найти кратчающую надежную серию T-согласованных 2-break операций, трансформирующих множетсвенный breakpoint граф $G(P_1, \ \dots \ P_n)$ в идентичный множественный breakpoint граф.*

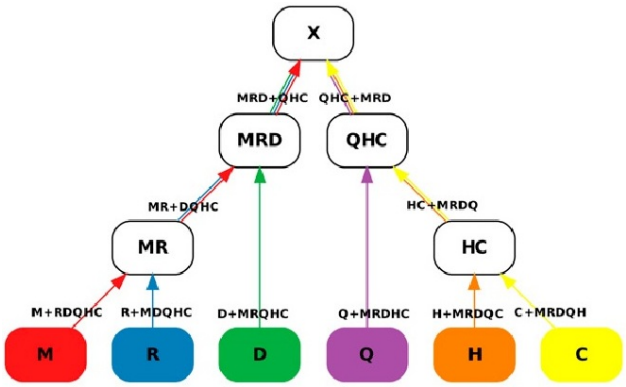

Рис. 2.2: Пример направленного дерева индуцированного цветами геномов

Отметим что MGRP и TCMGRP проблемы, в случае трех унихромосомных геномов, соответствуют проблеме медианы, которая является NP-полной задачей[ССЫЛКИ, ССЫЛКИ]. Поскольку вряд ли существуют точные полиномиальные алгоритмы для решения MGRP и TCMGRP проблем, MGRA использует эвристический подход "уничтожающий"ребра в $G(P_1, P_2 \ \dots \ P_n)$, используя надежные T - согласованные перестановки.

Для дальнейшего анализа будет удобно находить фиксированные ветви χ в филогенетическом дереве T и предполагать, что это ветвь содержит корень X (рассматривается как еще один узел), точное местоположение которого будет определенно позже. Выбор корня X определяет направление “к” X всех ветвей на филогенетическом дереве T. Для наглядности, пометим каждый лист P_i, содержащий геном P_i, в направленном дереве T единичным мультицветом $\{P_i\}$ и затем рекурсивно помечаем каждый внутренний узел, объединением мультицветов начиная с узлов всех входящих в него ветвей (см. рис. 2.2 общий конец ветвей, выходящих из листьев M и R, помечаем цветом M, R).

Определение 2.4. Мультицвет, соответсвующий пометке внутреннего уз-

ла в дереве T называется $\vec{T}$ - согласованным.

Можно дать альтернативное определение $\vec{T}$ - согласованным мультицветам, которые могут быть определены, как T - согласованные мультицвета, чьи индуцированные поддеревья не содержат ветви χ. Заметим, что именно один из мультицветов в каждой паре комплементарных T - согласованных мультицветов является $\vec{T}$ - согласованным и его пометки начальных узлов, соответствуют направлениям в дереве T (за исключением нескольких цветов соответствующих ветвям χ, что оба $\vec{T}$ - согласованные).

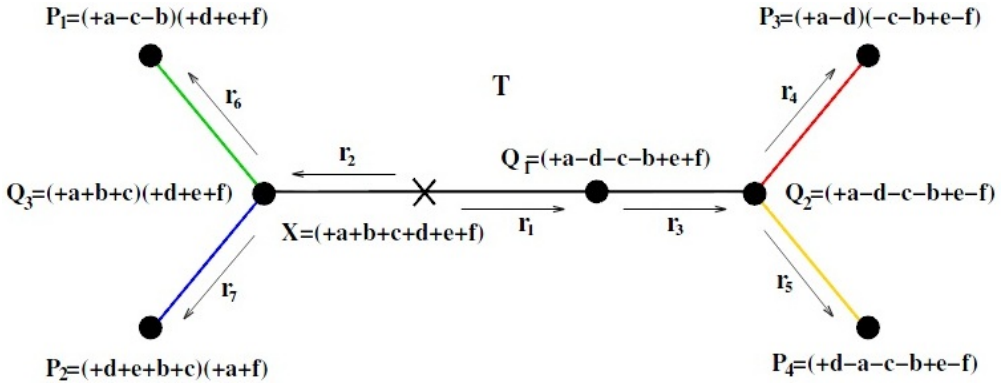

Рис. 2.3: Пример некорневого филогенетического дерева в листьях, которого содержатся геномы

MGRA алгоритм трансформирует геномы $P_1, \ldots P_n$ в X по направленным ветвям из дерева T, используя 2 - break операции на $\vec{T}$ - согласованные мультицвета для мультиребер ($\vec{T}$ - согласованные 2 - break операции). Данное утверждение в терминах breakpoint графов звучит так: MGRA алгоритм устраняет мультиребра в множетсвенном breakpoint графе $G(P_1, \ldots P_n)$ с помощью $\vec{T}$ - согласованных 2 - break операций и трансформирует его в идентичный множественный breakpoint граф $G(X, \ldots X)$. Использование $\vec{T}$ - согласованных 2 - break операций мотивированно важным свойством, что каждое $\vec{T}$ - согласованная трансформация может быть заменена на надежную $\vec{T}$ - согласованную трансформацию генома с изменением порядка в 2 - break операциях. Такие трансформации определим, как реверс трансформации из генома X в геномы $P_1, \ldots P_n$ с $\vec{T}$ - согласованными 2 - break операциями(результат приведен на рис. 2.3). MGRA алгоритм отслеживает перестановки применяющиеся к множественному breakpoint

графу $G(P_1, \ldots P_n)$ во время трансформации в идентичный множественный breakpoint граф $G(X, \ldots X)$. Записанные перестановки (в обратном порядке) определяют обратную трансформацию, которая проходит через каждый внутренний узел в дереве T и таким образом может быть использована, для реконструкции предков генома во внутренних узлах из дерева T.

2.2. Требование предъявляемые к улучшениям MGRA алгоритма

На основе биологических данных, состоящих из 59 штаммов бактерии ecoli, был получен граф содержащий 2876 вершин (см. раздел 3.1.1). Из них всего 208 вершин соответствует событиям дупликаций, непосредственно затрагивая еще 526 вершин. Для оставшегося графа из 2142 вершин разумно предположить, что они соответствуют 2 - break операциям в геноме, так как это наиболее распространненые перестановочные операции в геноме, а так же благодаря тому, что частота транспозиций в эволюции пока не установлена (см. раздел 1.2.2).

2.3. Требования предъявляемые к разрабатываемому серверу

2.4. Вывод по главе 2

Глава 3. Реализация модификаций

3.1. MGRA алгоритм

MGRA алгоритм был реализован на языке C++, поэтому улучшения реализовывались с помощью этого языка. Для графического отображения графов в реализованном алгоритме, используется библиотека graphiz[ссылка]. Текущая версия алгоритма доступна по ссылке [**algo_src**]

3.1.1. Представление графа

Как говорилось в разделе 2.2, теперь степень каждой вершины может быть больше чем n, где n - количетсво геномов, возможно существование мультицветов, в которые цвета входят по нескольку раз, а так же исходящие из вершины мультиребра с мультицветами могут иметь не пустое пересечение друг с другом. Все эти нововведения делали не валидными старые структуры данных, которые сильно использовали свойства наблюдающиеся в геномах без событий дупликаций.

На данный момент множественный breakpoint граф представлен, как набор breakpoint графов для каждого генома, где в каждый "локальный"breakpoint граф поддерживает возможность хранения кратных ребер. Соответствующий код представлен в файле mpbgraph.h. Мультицвета это теперь не просто множество, а multimap отображающий цвет и его количество и весь код реализован в файле mcolor.h. Так же из-за необходимости обрабатывать исходящие ребра из вершины для них созданна отдельная структура, называемая mularcs, поддерживающая разнообразные операции над ними.

После того как все структуры данных были модернизированны, на основе 59 штаммов бактерий ecoli был получен множественный breakpoint граф, приведенный на рис. **??**, содержащий 2876 вершин.

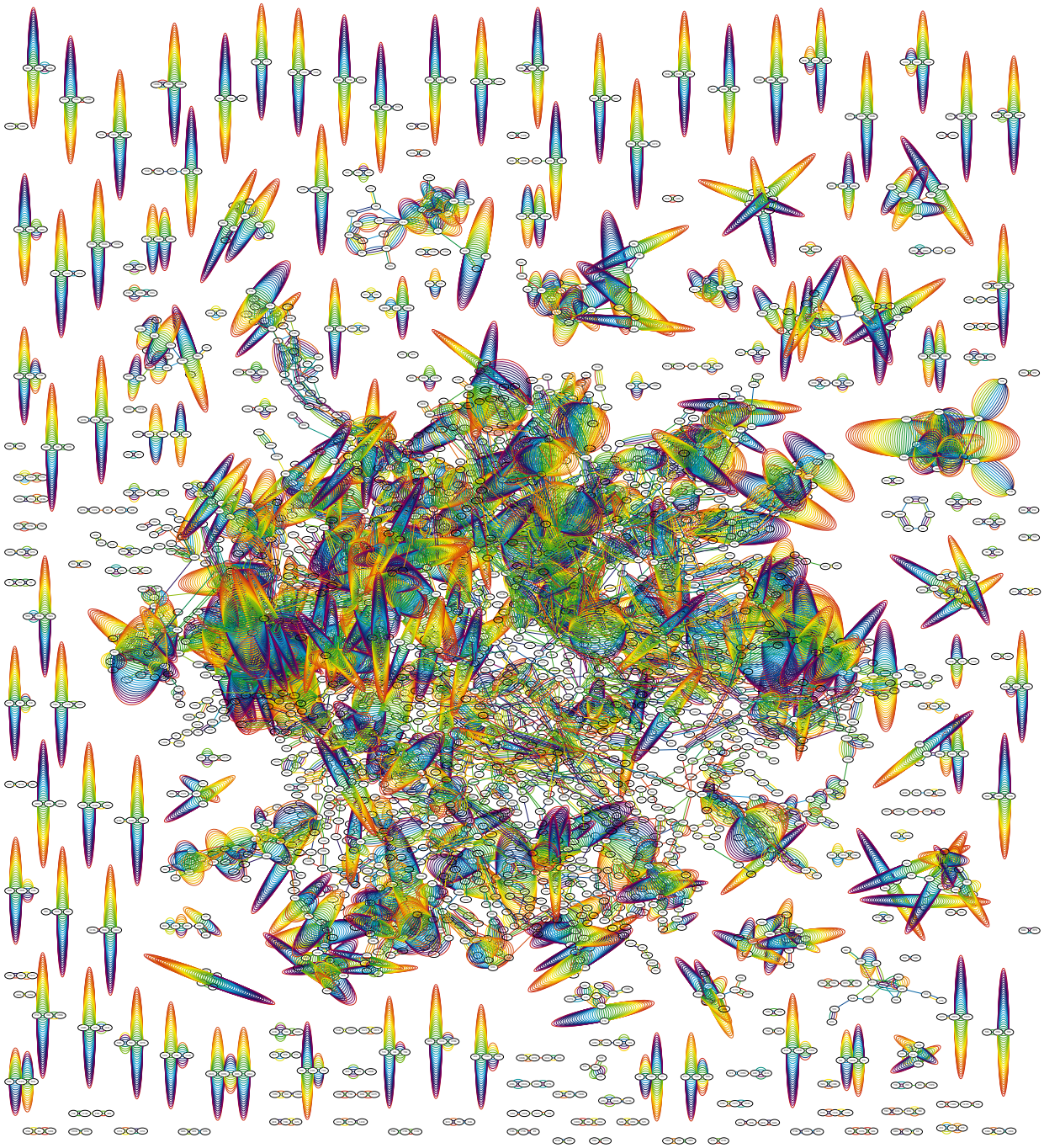

Рис. 3.1: Множественный breakpoint граф построенный на основе 59 штаммов бактерий ecoli

3.1.2. Адаптация первого шага по обработке хороших циклов и путей

Мультиребро называется **нормальным**, если в его мультицвете каждый цвет встречается один раз. Вершина называется **обычной**, если эта регулярная вершина инцидентна двум нормальным мультиребрам,

мультицвета которых имеют пустое пересечение. Мультиребро называется **обычным**, если оно соединяет две обычные вершины. Путь/цикл называется **обычным**, если в него входят все обычные мультиребра, которые альтернируют между Q и $\overline{Q}$ Путь/цикл называется **хорошим**, если все мультицвета ребер T-консистенты

3.1.3. Адаптация второго шага по обработке честных ребер

3.2. Структура MGRA сервера

Данный сервер доступен по адресу [].

3.2.1. Алгоритм реконструкции филогенетических деревьев

3.3. Вывод по главе 3

Глава 4. Экспериментальные результаты модификаций

4.1. Вывод к главе 4

Заключение

Список литературы

1. Genome 10K Project. https://genome10k.soe.ucsc.edu.
2. *Bourque G and Pevzner P A.* Genome-scale evolution: reconstructing gene orders in the ancestral species // Genome Res. 2002. Pp. 26–36.
3. *Ma J and Zhang L and Suh B B and Raney B J and Burhans R C and Kent J W and Blanchette M and Haussler D and and Miller W.* Reconstructing contiguous regions of an ancestral genome // Genome Res. 2006. Pp. 1557–1565.