Декомпозиция графа геномов раковых клеток

Збань И.К., магистрант кафедры компьютерных технологий ИТМО,
izban@mail.ru

Аннотация

В данной статье приведена формулировка задачи декомпозиции графа геномов раковых клеток и предложен алгоритм решения поставленной задачи за линейное время.

Введение

Биоинформатика стала важной частью многих областей биологии. В экспериментальной молекулярной биологии методы биоинформатики, такие как создание изображений и обработка сигналов, позволяют получать полезные результаты из большого количества исходных данных. В области генетики и геномики биоинформатика помогает в упорядочивании и аннотировании геномов и наблюдаемых мутаций. Она играет роль в анализе данных из биологической литературы и развитии биологических и генетических онтологий по организации и запросу биологических данных. Также она используется в анализе гена, экспрессии белка и регуляции.

Инструменты биоинформатики помогают в сравнении генетических и геномных данных и в целом в понимании эволюционных аспектов молекулярной биологии. В общем виде она помогает анализировать и каталогизировать биологические пути и сети, которые являются важной частью системной биологии. В структурной биологии она помогает в симуляции и моделировании ДНК, РНК и белковых структур, а также молекулярных взаимодействий.

В данной работе рассмотрена одна из задач биоинформатики — восстановление последовательной структуры хромосом в геномах раковых клеток в упрощенной модели.

Постановка задачи

Развитие рака может быть рассмотрено как набор соматических мутаций, серьёзно изменяющий здоровый исходный геном. Для упрощения будем считать, что единственные изменения, происходящие за время развития рака, это структурные изменения большого масштаба, и опустим, что опухоли могут быть гетерогенными, а геномы не гаплоидные. В такой упрощенной модели можно решать задачу восстановления последовательной структуры хромосом в геномах раковых клеток.

Мы не можем явно измерить хромосомы напрямую, но с учетом того, что раковые геномы происходят от известных геномов, любая производная хромосома определяет путь или цикл отрезков исходной хромосомы. Есть несколько методов, позволяющих оценить, насколько часто данный отрезок наблюдается в раковом геноме и какие бывают связи между частями в итоговой хромосоме.

Формально, можно рассмотреть неориентированный граф на $2m$ вершинах с выделенным полным паросочетанием. Назовем эти ребра сегментными ребрами. Кроме того, в графе могут быть другие ребра, назовем их ребрами связи. Каждое ребро имеет какую-то положительную кратность, в графе могут быть кратные ребра и петли.

Задача — разбить этот граф на минимальное множество путей и циклов, причем в каждом пути и цикле должны чередоваться сегментные ребра и ребра связи, и кроме этого каждый путь должен начинаться и заканчиваться с сегментного ребра. Такие пути будут соответствовать последовательной структуре ракового генома, полученного из известного здорового генома.

Предлагаемое решение

Предлагаемое решение этой задачи заключается в сведении к модификации алгоритма поиска Эйлерова пути в неориентированном графе.

Во-первых, рассмотрим критерий существования решения. Скажем, что баланс вершины — разность между количеством смежных с ней сегментных ребер и ребер связи, причем петли учитываются дважды. Легко понять, что если есть вершина с отрицательным балансом, решения не существует.

Это верно, потому что, если взять ребра любого цикла в ответе, они не изменят баланс ни одной вершины (потому что два соседних ребра аннулируют вклад друг друга в баланс, поскольку они разного типа), а любой путь лишь увеличивает на единицу баланс двух вершин концов этого пути (крайние ребра пути сегментные, а остальные вершины имеют такой же нулевой вклад в баланс).

Таким образом, если есть вершина с отрицательным балансом, ответ найти не получится.

Утверждается, что иначе решение всегда есть. Сначала добавим произвольным образом ребра связи так, чтобы баланс каждой вершины стал равен нулю — это можно сделать жадным алгоритмом. Сумма балансов положительная четная, а одно ребро связи уменьшает баланс двух вершин на единицу, так что жадный алгоритм решит проблему. Теперь, поскольку баланс каждой вершины нулевой, каждой вершине инцидентно одинаковое количе-

ство сегментных ребер и ребер связи (петлю снова учитываем дважды).

Разобьем в каждой вершине ребра на пары, чтобы в каждой паре было по ребру противоположных типов (это легко, поскольку количества ребер равны).

Скажем, что если мы вошли в вершину по какому-то ребру, то мы выйдем из нее по парному ребру. Поскольку каждое ребро продублировано дважды — в двух вершинах-концах этого ребра, мы сопоставили каждому ребру по два ребра-соседа в цикле. Это значит, что мы только что разбили все ребра на циклы.

Вспомним, что у нас есть добавленные фиктивные ребра связи, которые чинили баланс вершин. Удалим их из циклов, после этого некоторые циклы распадутся на пути. Можно заметить, что такое разбиение на пути и циклы удовлетворяет условию задачи — в каждом пути и цикле типы ребер чередуются и каждый путь начинается и заканчивается сегментным ребром (поскольку мы удаляли лишь ребра связи из циклов).

Вспомним, что мы хотели минимизировать число элементов в разбиении. Для этого можно понять, что если есть два элемента разбиения (пути или циклы, причем один из них цикл) с общей вершиной, то их можно склеить. Как? У этой вершины есть два смежных ребра в каждом из элементов разбиения, и можно вставить ребра цикла после этой вершины в пути в одном из двух порядков (прямом или развернутом), потому что хотя бы один из порядков сохранит, что соседние ребра имеют различный тип.

Таким образом можно избавиться от лишних циклов, и останутся лишь пути, которых ровно столько, сколько нужно — сумма всех балансов пополам, и циклы, которые не содержат в своих компонентах связности вершин с ненулевым балансом.

Докажем, что меньшим числом путей и циклов обойтись нельзя: по построению разбиения выполнено, что каждая компонента связности исходного графа, в которой балансы всех вершины были нулевые, будет покрыта одним циклом. Каждая компонента с ненулевым суммарным балансом b покрыта ровно $\frac{b}{2}$ путями, что является минимально возможным числом путей. Если бы были другие пути, не сошелся бы суммарный баланс, а если бы были еще циклы в разбиении, хотя бы один точно имел общую вершину с хотя бы одним из путей.

Заключение

Разработан эффективный линейный алгоритм для решения поставленной задачи.

По данной теме была подготовлена задача для интернет-соревнования по биоинформатике. За время соревнования верно решить предложенные тестовые данные смогли четыре участника из нескольких сотен.

Литература

[1] Bioinformatics contest 2019 // http://mon.stepik.org/

[2] Постановка задачи // https://stepik.org/lesson/207045/step/2?unit=184369/