Разработка метода восстановления фрагментов нуклеотидной последовательности по парным чтениям.

Сергушичев Алексей Александрович

Научный руководитель: Царев Ф.Н.

Санкт-Петербург, 2011

ДНК

- Молекула ДНК представляет из себя двойную спираль, состоящую из пар нуклеотидов (A, T), (T, A), (G, C) или (C, G).
- Цепочки в ДНК противонаправлены.
- Можно считать, что ДНК – две обратно-комплементарные строки над алфавитом {A, T, G, C}.

Секвенирование генома

- Геном организма состоит из нескольких длинных молекул ДНК (хромосом).
- Секвенирование генома – процесс определения последовательности нуклеотидов в хромосомах.
- Можно считывать только небольшие фрагменты.
- Очень много считываний, чтоб сделать достаточное покрытие.

Парные чтения

- На данный момент достаточное покрытие можно эффективно получать с помощью небольших парных чтений.

Постановка задачи

- Дано большое число парных чтений (сотни гигабайт данных)
- Известно распределение длин фрагментов.
- Для каждой позиции в каждом чтении известна вероятность ошибочного чтения этой позиции.
- Необходимо восстановить геном.

Цель работы

- Разработать алгоритм восстановления фрагментов, который был бы эффективен для данных такого объема.

Граф де Брюина

- Граф де Бриюина размерности *k* над алфавитом Σ – ориентированные граф <V, E>, где

$$V = \Sigma^k$$

$$E = \{ <(v_1, v_2, \ldots, v_k), (v_2, v_3, \ldots, v_{k+1})> : v_i \in \Sigma \}$$

- Ребро однозначно идентифицируется строкой $v_1v_2\ldots v_kv_{k+1}$.

ДНК

- Набор нуклеотидных последовательностей хромосом – подграф графа де Брюина для алфавита Σ = {A, T, G, C}.

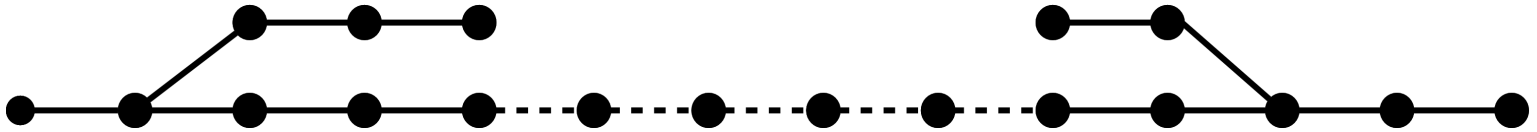
- Задается множеством ребер.

Подграф графа де Брюина

- Ребра — подстроки длины k+1 из прямых и обратно-комплементированных исходных данных, которые встречаются больше некоторого порога раз.
- Всего ребер порядка удвоенной длины генома.
- Можно хранить только половину ребер в хеш-сете, с расходом памяти от восьми (лучше около 12) байт на ребро для $k \leq 31$.

Парные чтения

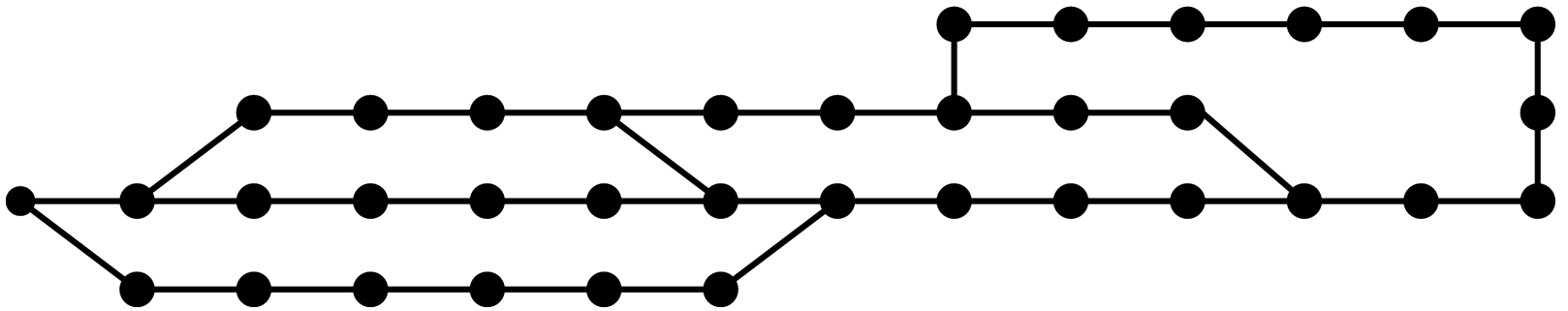
- Парные чтения – концы пути от одной вершины до другой.



- Надо найти путь, похожий на то, что есть

Решение

- Найдем все пути из первой вершины во вторую с длиной между некоторой минимальной и максимальной границей, которые определяются из распределения длин фрагментов.



Решение

- Оцениваем «правдоподобность» всех путей на основе известных нуклеотидов и качества их чтения.
- Не учитываем длину.
- Оставляем только достаточно «правдоподобные» пути.

Полученные пути

- Один путь – хорошо.
- Несколько путей одинаковой длины – можно посмотреть на степень различия, если они похожи, хорошо.
- Несколько путей разной длины – что делать непонятно, фрагмент не восстанавливается.

Проблемы

- Есть плохие места подграфа, где средняя степень вершин велика – число различных путей увеличивается, растет время работы и используемая память.
- Процесс поиска приходится прерывать, если уже рассмотрено достаточно много вершин.

Алгоритм

- Берем вершины, соответствующие началам парных чтений.
- Идем от них соответственно вперед и назад пока можно.
- Запускаем два параллельных обхода в ширину.
- На каждой итерации выкидываем «неправдоподобные» вершины – вершины с маленькой вероятностью их появления на этом расстоянии.
- Если в слое большое число вершин – прерываем работу.

Экспериментальные данные

- На синтезированных данных генома размером 1.8 миллиардов нуклеотидов с длиной фрагментов около 500 при k = 30:
- Восстанавливается 67% процентов фрагментов:
  - Для 7% не нашлось ни одного пути.
  - Восстановление 23% было прервано.
  - У 3% возникла неоднозначность.
- Обработка 350 миллионов парных чтений на 24-х ядерном компьютере с 64 ГБ памяти занимает сутки.
- Распределение длин полученных фрагментов и данное распределение длин всех фрагментов практически полностью совпадает.

Результаты

- Был разработан алгоритм восстановления фрагментов нуклеотидной последовательности по парным чтениям.
- Разработана реализация предложенного алгоритма на языке *Java*.
- Произведено экспериментальное исследование, подтвердившее возможность достаточно эффективного восстановления фрагментов больших геномов.

Спасибо за внимание!