Выравнивание последовательностей на графы метагеномных сборок

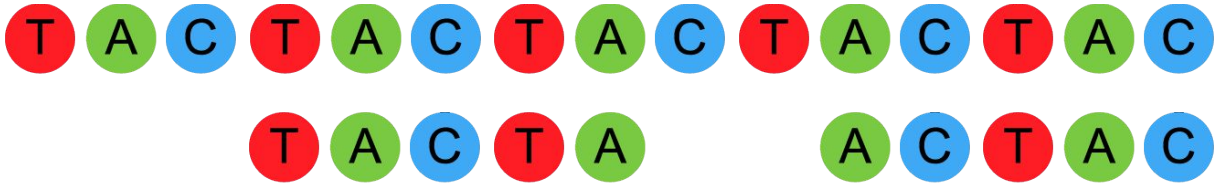
Автор: Богомолов Егор
Научный руководитель: Мелешко Дмитрий

Мотивация: 16S рРНК

- Присутствует у почти всех бактерий
- Функции со временем почти не изменились
- Позволяет с большой точностью определять род и вид бактерий
- Есть уже собранные базы: **SILVA rRNA database project**
- Регион 16S рРНК плохо собирается ассемблером, просто найти его в базе не получается
- Хочется уметь искать в графе метагеномной сборки пути, похожие на последовательности из базы

Мотивация: улучшение качества сборки

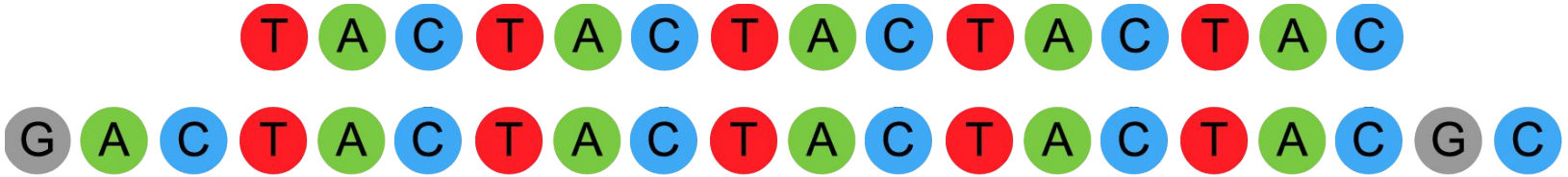
- При сборке генома из коротких ридов есть проблемы, например, тандемные повторы:



- При помощи только коротких ридов хорошего решения нет

Мотивация: улучшение качества сборки

- При сборке генома из коротких ридов есть проблемы, например, тандемные повторы:



- Добавляя длинные риды проблему можно решить
- Хочется искать в графе сборки пути, похожие на длинные риды

Существующие аналоги

1. Разнообразные инструменты для выравнивания последовательностей на строки
2. GABA — проект, находящийся в общем доступе
    - Умеет работать только с деревьями
3. Уже имеющаяся в SPAdes утилита
    - При работе использует сиды фиксированной длины

Существующие аналоги

1. Разнообразные инструменты для выравнивания последовательностей на строки
2. GABA — проект, находящийся в общем доступе
   - Умеет работать только с деревьями
3. Уже имеющаяся в SPAdes утилита
   - При работе использует сиды фиксированной длины
   - Стала основой для данной работы

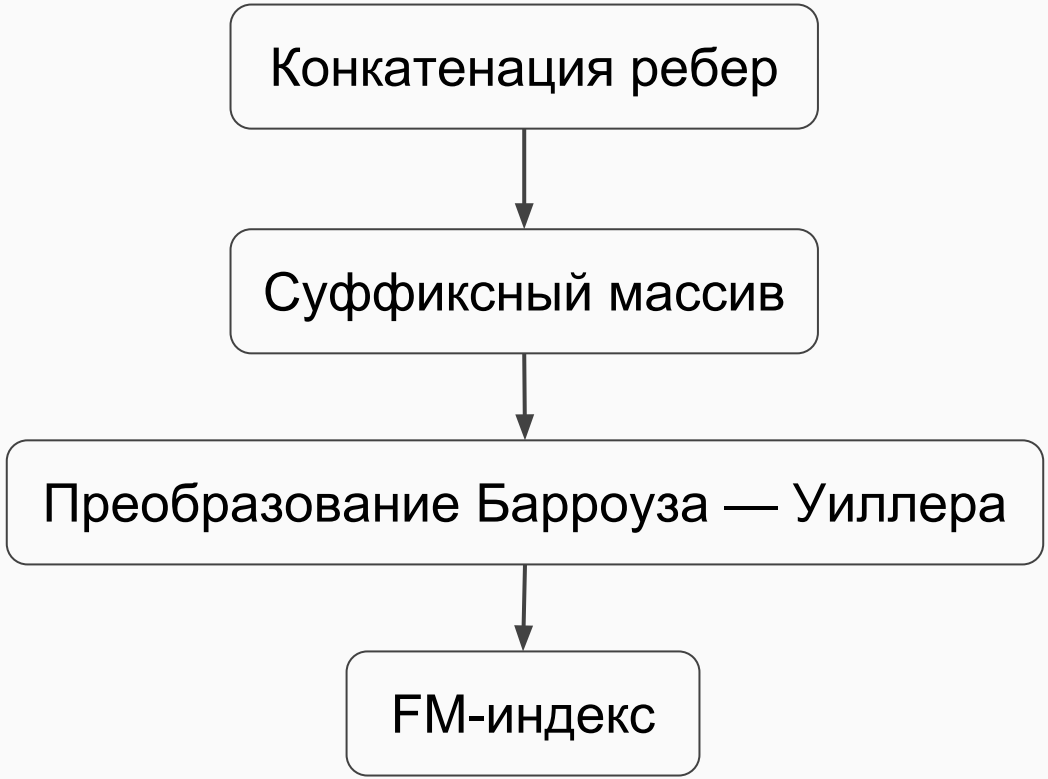
Постановка задачи

1. Написание утилиты, позволяющей выравнивать последовательности нуклеотидов на графы метагеномной сборки
2. Первоначально предполагалось делать это внутри SPAdes, но затем от этой идеи отказались
3. Язык реализации: C++

Принцип работы

- В последовательности могут встречаться ошибки
- Несмотря на это какая-то часть не очень длинных подстрок должна в точности найтись на ребрах графа
- Найдем их вхождения
- Фрагменты между попробуем заменить кратчайшими путями
- Выберем вариант с минимальным редакционным расстоянием

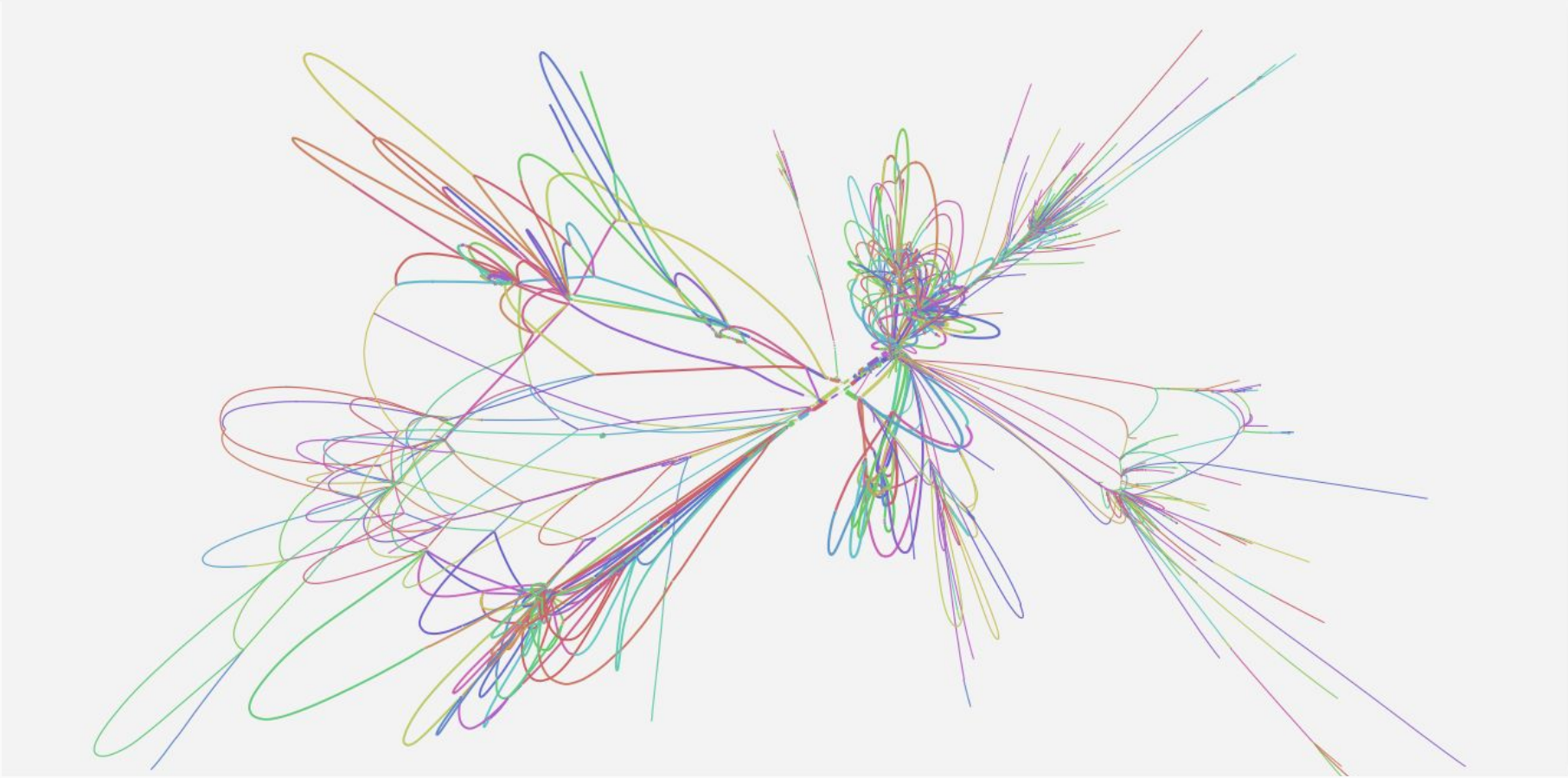
Принцип работы: поиск вхождений



Принцип работы: построение путей

- Для каждой позиции в строке найдем самую длинную подстроку, которая начинается в ней и встречается на ребрах графа
- Итерируемся по ним слева направо
- Поддерживаем набор путей-кандидатов в графе и пытаемся их достраивать
- Отсекаем по расстоянию между префиксом строки и путем

Пример реального графа



Результаты

1. Реализован препроцессинг графа для быстрого поиска вхождений
2. Опробованы две стратегии дальнейшего построения путей
3. Программа протестирована при помощи генерации небольших графов и путей в них
4. Запуски на реальных данных, в результате чего выявлены некоторые проблемы, к данному моменту не решенные

Список литературы

1. Janda JM, Abbott SL. 16S rRNA Gene Sequencing for Bacterial Identification in the Diagnostic Laboratory: Pluses, Perils, and Pitfalls . *Journal of Clinical Microbiology*. 2007;45(9):2761-2764. doi:10.1128/JCM.01228-07.
2. Clarridge JE. Impact of 16S rRNA Gene Sequence Analysis for Identification of Bacteria on Clinical Microbiology and Infectious Diseases. *Clinical Microbiology Reviews*. 2004;17(4):840-862. doi:10.1128/CMR.17.4.840-862.2004.
3. Cheng Yuan, Jikai Lei, James Cole, Yanni Sun; Reconstructing 16S rRNA genes in metagenomic data. *Bioinformatics* 2015; 31 (12): i35-i43. doi: 10.1093/bioinformatics/btv231
4. Ben Langmead*;* Introduction to the Burrows-Wheeler Transform and FM Index. *Department of Computer Science, JHU.* 2013;

Спасибо за внимание!