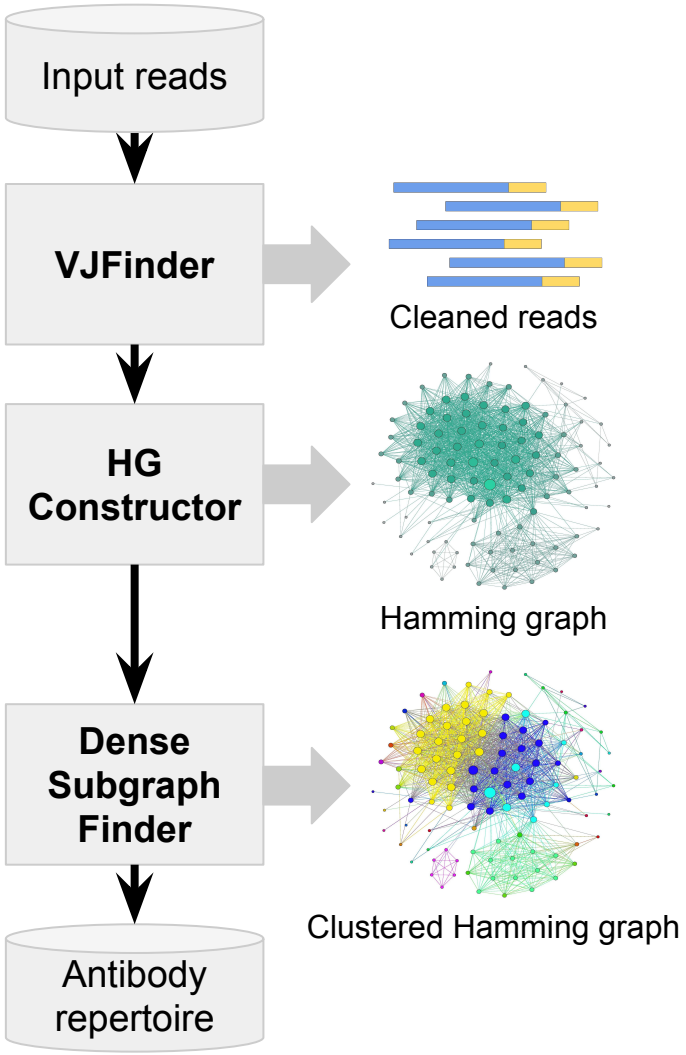
Speed-up VJ Finder

авторы: Алфёров Василий, Винниченко Максим
научный руководитель: Александр Шлемов

IgRepertoireConstructor

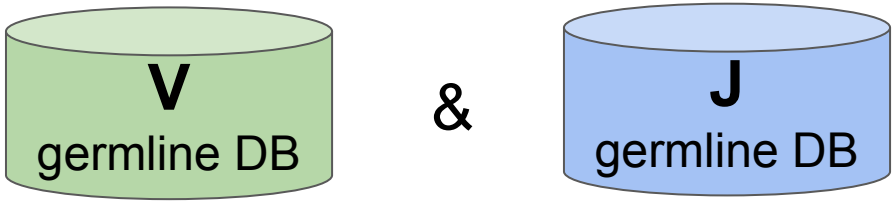
- Иммуноглобулины -- белки, активно участвующие в работе иммунитета
- IgReC по набору иммуноглобулинов строит их репертуар и анализирует полученные данные.
- VJ Finder строит соответствие между иммуноглобулинами и базой.



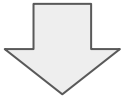
Illumina MiSeq (2014)

- В 2014 году в Illumina Labs построена MiSeq, читающая до 10^8 ридов.
- В мире используется IgBLAST, работающий медленно.

Что делает VJ Finder



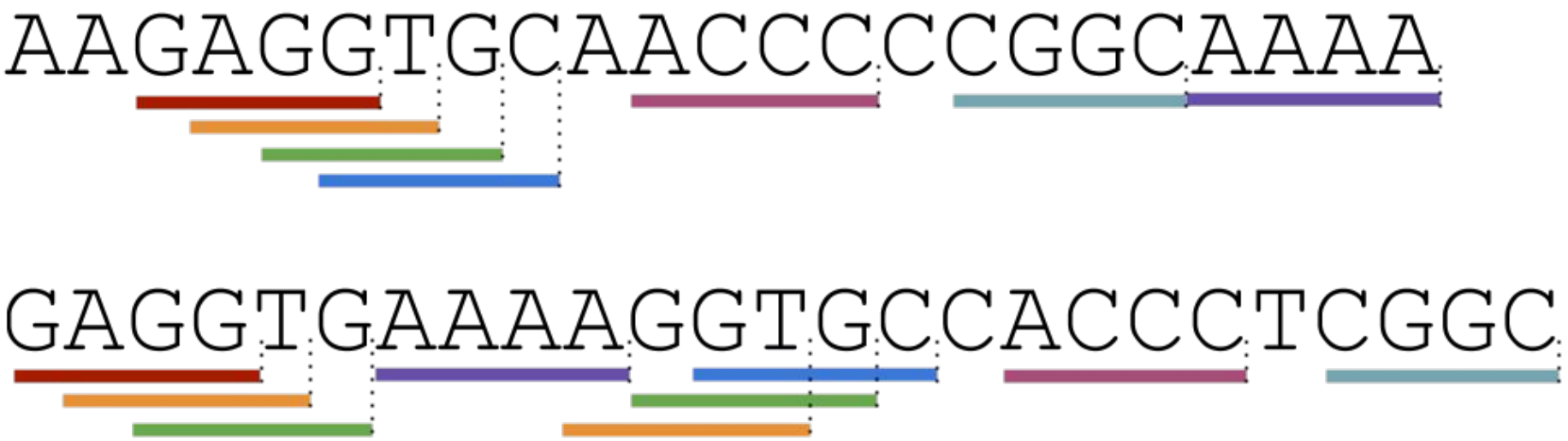
нашли выравнивание



OK

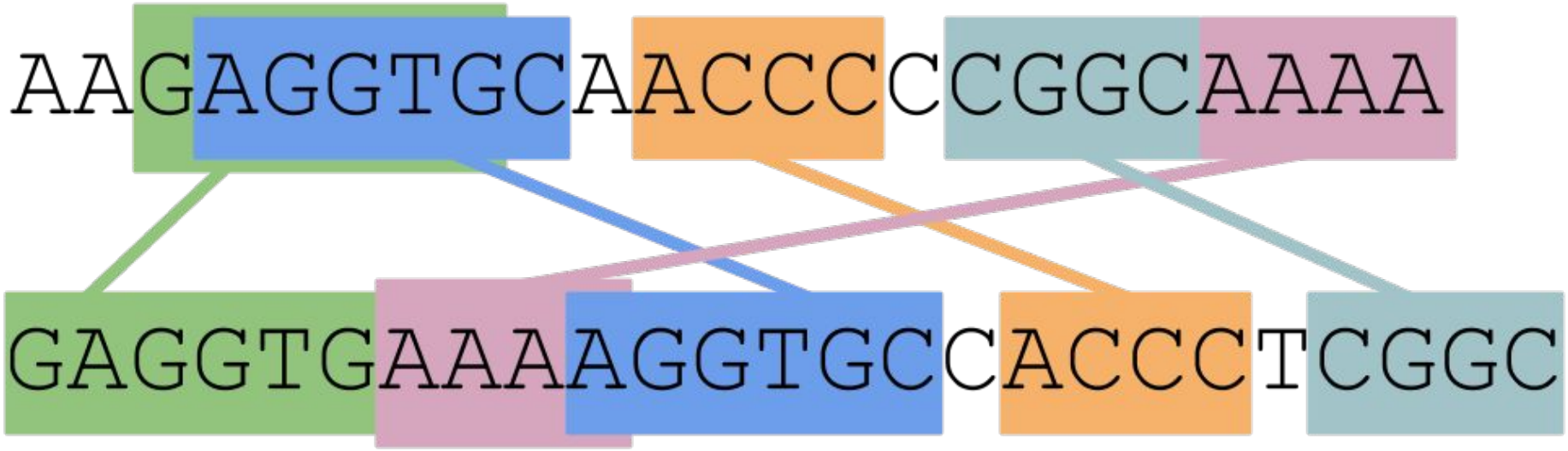
Как?

Ищем общие Kmer match



Как?

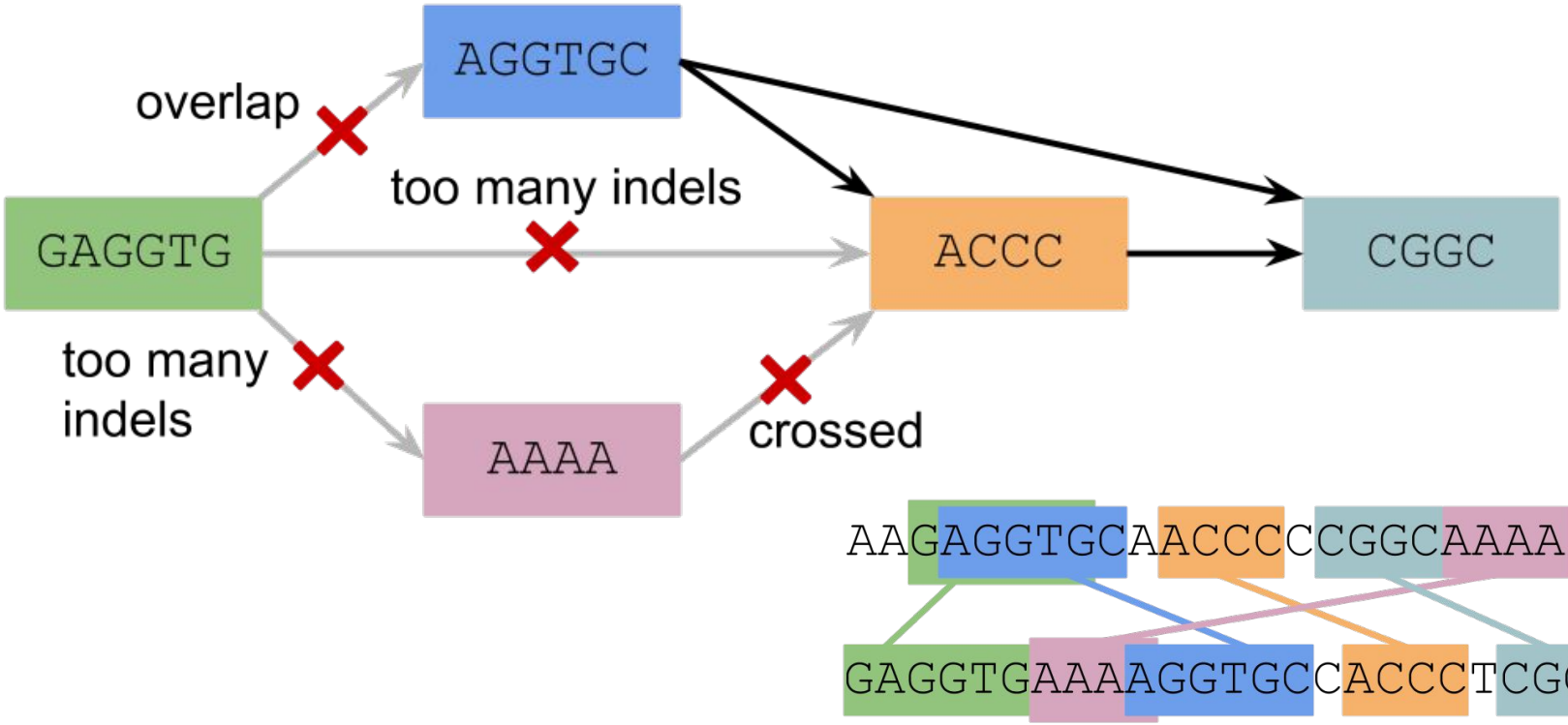
Объединяем последовательные Kmer match в Match



Как?

Строим граф. (вершины -- пары Match, ребро -- другой Match больше по каждой координате)

Ищем самый длинный путь с помощью динамики.

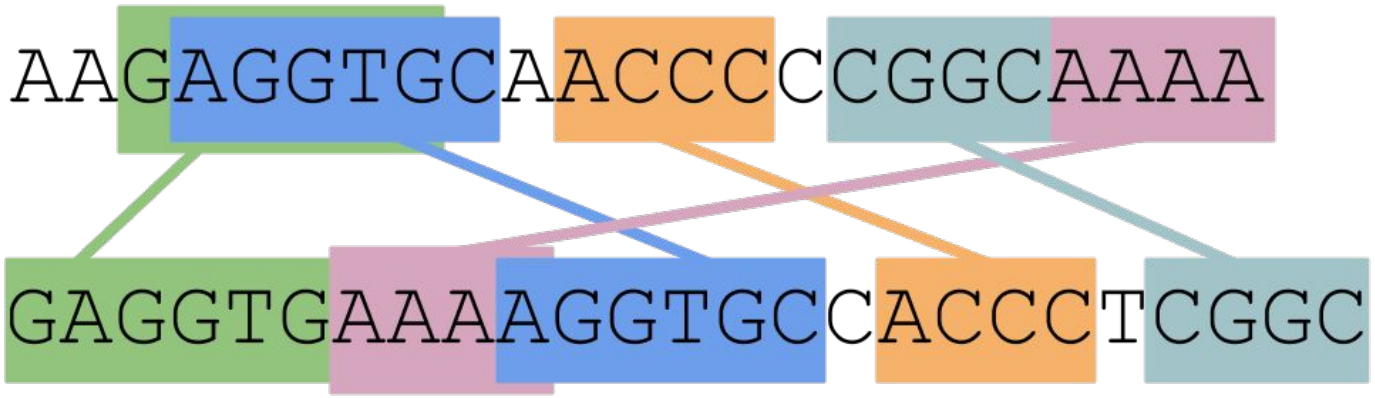


Переформулировка задачи

- Вершину графа можно представить как пару чисел -- позиция совпадения в первой и второй строках.
- Любую пару чисел можно представить как точку на плоскости.
- Путь-ответ должен быть монотонен по обеим координатам.
- Вес пути вычисляется по сложным формулам, которые не получается идейно встроить в алгоритм.

Пробуем улучшить

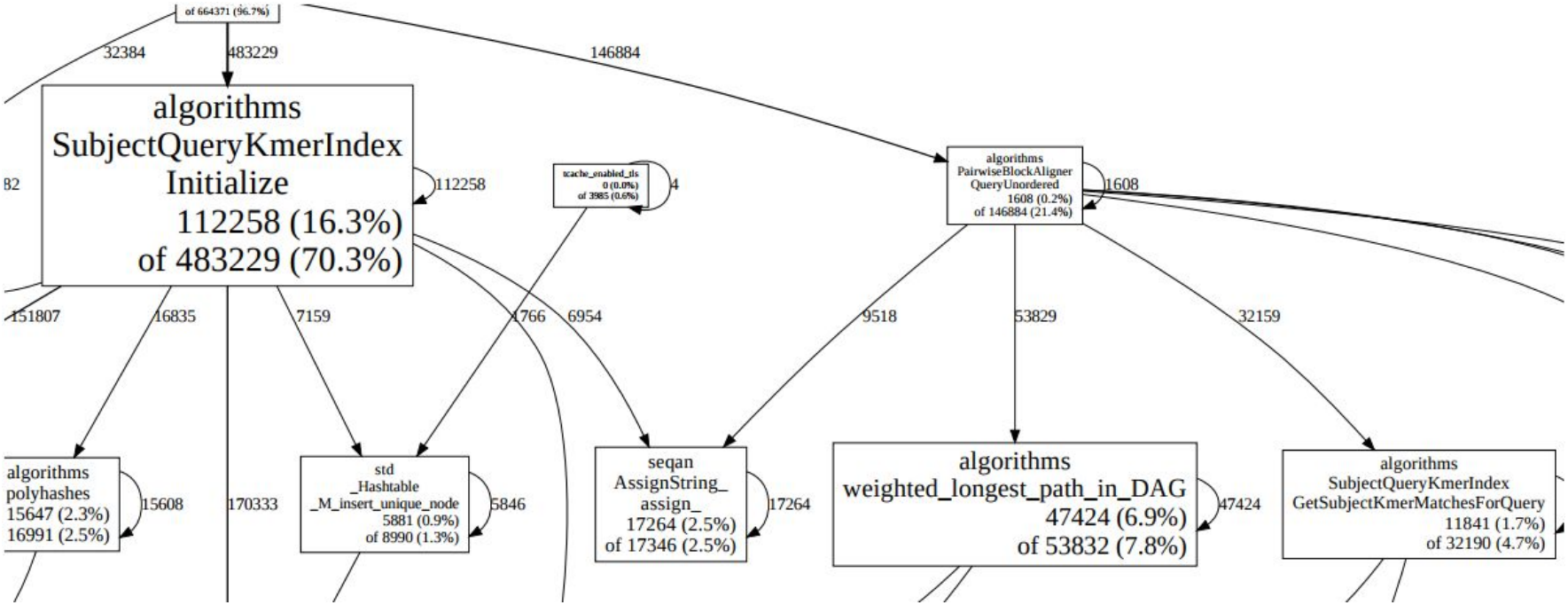
- Некоторые рёбра запрещены. (много вставок/удалений)
- Сортируем по gap ([начало в одном] - [начало в другом])
- Идём вправо/влево с отсечением на большую разницу gap.



Другие способы скоринга

- Раз последовательности монотонны по обеим координатам, то можно искать НВП.
- Угадывает 4% совпадений. ☹

Профилирование



Продолжаем улучшать.

- Замечаем много вызовов обращений unordered_map.
- Т.к. ключи у неё до 2^14 (фактически). Заменяем на vector такого размера.

- с его помощью ищутся общие kmer в match сразу для всей БД V/J сегментов.

Сортировка вставками

- Поскольку итоговый путь должен быть отсортирован по обеим координатам, мы можем вставить новый элемент только после определённых.
- Получится, что каждое возможное ребро в графе является инверсией в массиве элементов, отсортированных по второй координате.
- Все инверсии можно перебрать с помощью InsertionSort.
- Не сильно улучшает время работы, так как данные всё же не рандомные.

Итого

- Найдены проблемы производительности
- Ускорили в 1.5 раза
- https://github.com/vasalf/ig_repertoire_constructor/tree/testing