Улучшение качества геномных сборок: поиск структурных ошибок и заполнение разрывов в скаффолдах.

Лиознова Анна Валерьевна
руководитель: Son Kim Pham, PhD

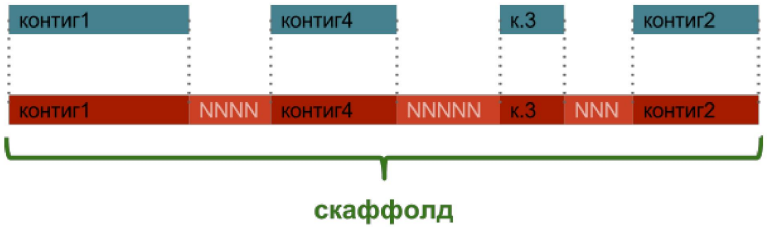
Академический Университет
2015

Определения

Риды — короткие последовательности нуклеотидов, получающиеся при секвенировании

Контиги — результат работы ассемблера, непрерывные последовательности нуклеотидов

Скаффолды — упорядоченная последовательность контигов с заданными расстояниями между ними (и направлением)



Актуальность работы

Одна из целей сборки геномов de novo — изучение эволюции живых организмов.

Ассемблеры допускают ошибки.

Неправильные данные = неправильные выводы.

Алгоритмы не могут выдать готовые хромосомы.

Много разрывов = высокая стоимость ПЦР.

Нужны правильные и полные геномные последовательности!

Цели и задачи

Цель: улучшение качества геномных сборок de novo путем поиска структурных ошибок и заполнения разрывов в скаффолдах

Задачи: разработать и реализовать алгоритмы для

- поиска структурных ошибок в сборке с использованием парных ридов
- закрытия разрывов в скаффолдах с помощью неиспользованных контигов, опираясь на геномы родственных видов

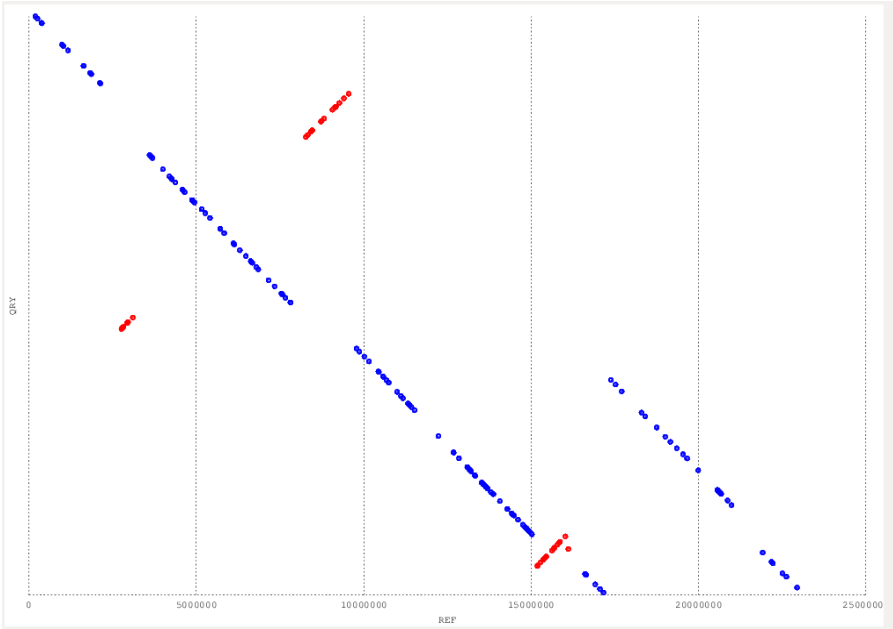
Поиск структурных ошибок

Структурные ошибки

= глобальные

<u>Существующие решения:</u>

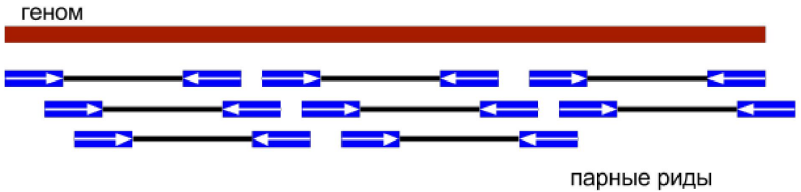
Hagfish (2012), Amosvalidate (2008), REAPR (2013), misSEQuel (2015), Pilon (2014) и др.



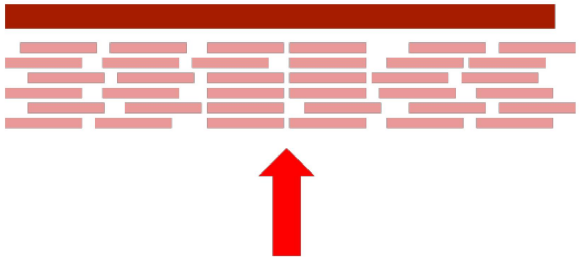
Выравнивание сборки на референс, должна быть одна диагональная линия

Предлагаемое решение

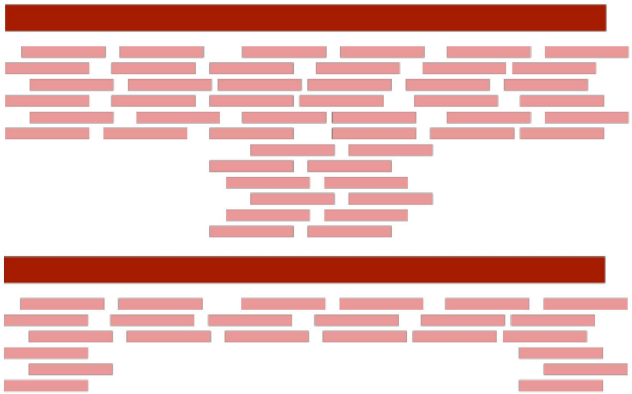
Парные риды — пары ридов на приблизительно известном расстоянии.



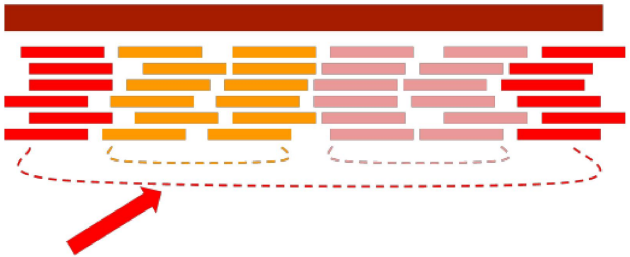
1. Начала и концы



2. Покрытие



3. Расстояние вставки



Результаты

Организм: E.coli

Длина референса: 4.64 Mbp

Количество парных ридов: 28 428 648

	референс	EULER	SPAdes
начала\концы	0	72	33
длина вставки	0	4	1
покрытие	5	24	19
всего	5	149	53
REAPR	77	535	195
QUAST	0 + 0 (лок.)	9 + 31 (лок.)	0 + 5 (лок.)

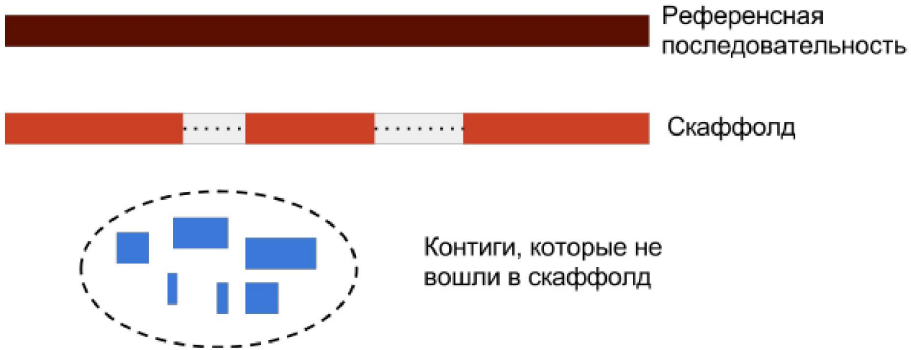
Технологии: python, biopython, numpy, bwa, bowtie2

Заполнение разрывов

<u>Существующие решения:</u>

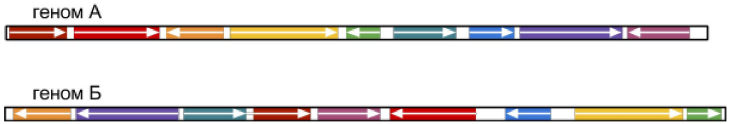
скаффолдинг: **Ragout** (2014), r2cat (2009), Projector2 (2005), OSLay (2007), CONTIGuator (2011), ABACAS (2012), Mauve (2009) и др.

закрытие разрывов с использованием ридов: IMAGE (2012), assisted assembly (2009) и др.

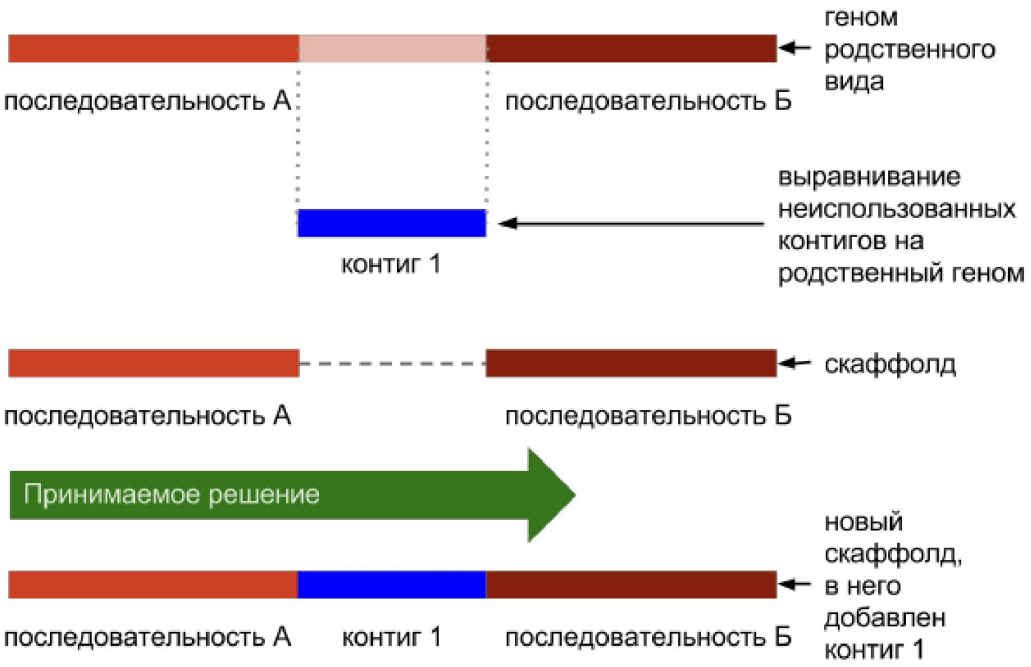


Предлагаемое решение

Синтенные блоки — участки нескольких геномов без структурных различий



Выравнивание контигов на родственные геномы



Результаты

Реализация на примере скаффолдов Ragout.

Организм	Доступные референсы	Свободные контиги	Добавлено контигов (из них уникальных)	Ошибки в итоговых скаффолдах
H.pylori	4	83	59 (59)	0
E.coli	1	37	22 (21)	2

Технологии: python, biopython, bwa, MUMmer, Ragout, Sibelia

Выводы

1. Разработаны новые алгоритмы для выявления структурных ошибок и заполнения разрывов в скаффолдах

2. Реализованы соответствующие утилиты:

   https://github.com/ALioznova/improve_assembly

3. Проведенное тестирование показывает адекватность алгоритмов и их способность решить поставленную задачу

4. Разнообразие существующих методов и полученные результаты демонстрируют наличие широкого спектра возможностей для дальнейших исследований

Благодарности:

- Шон Фам
- Алексей Гуревич
- Ксения Крашенинникова
- Александр Шлемов
- Михаил Колмогоров

Спасибо за внимание!