Исправление ошибок в зашумленных последовательностях при помощи графов сборки

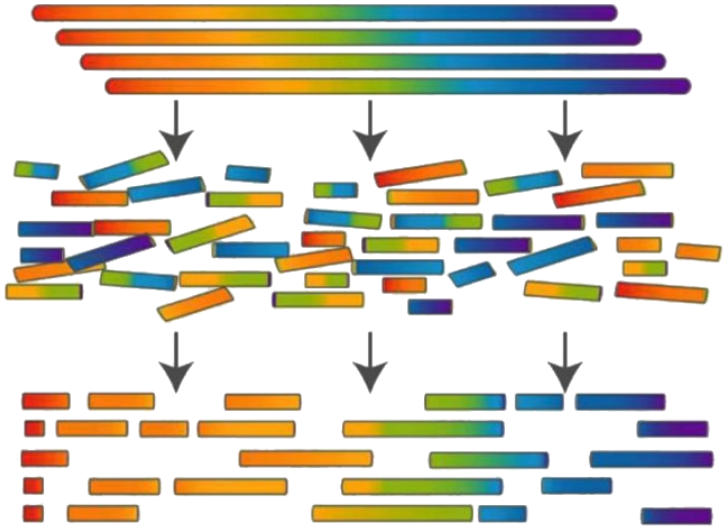
Клещин Антон Сергеевич, 20.M07-мм

Научный руководитель: доц. каф. СП, к.т.н. Ю. В. Литвинов
Консультанты: доц. каф. стат. мод., к.ф.-м.н. А. И. Коробейников
старший н.с., к.ф.-м.н. А. Д. Пржибельский
Рецензент: приглашённый н.с., к.ф.-м.н. С. Ю. Нурк

СПбГУ

5 мая 2022 г.

Терминология



Мотивация

- Исправление сборок из длинных ридов с помощью графа из коротких
- Исправление метагеномных сборок
- Промежуточное исправление в метагеномных сборках геномов

Постановка задачи

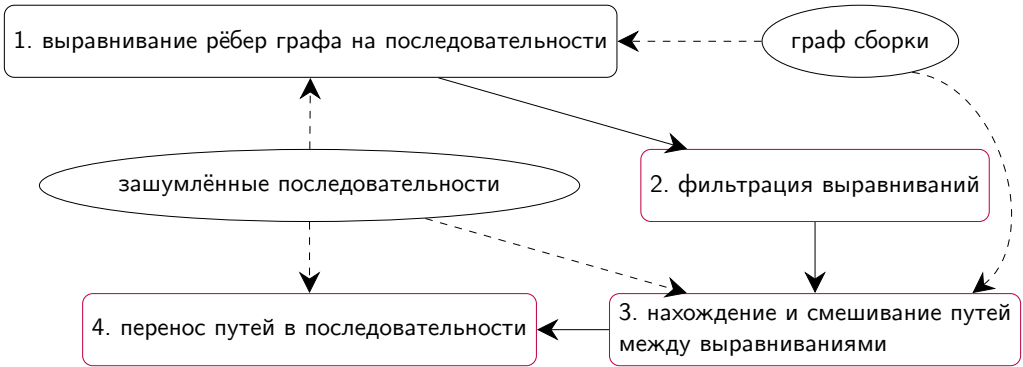
Цель — создание инструмента, позволяющего исправлять ошибки в контигах при помощи графов сборки.

- Формирование критериев фильтрации выравниваний рёбер графа на последовательности.
- Разработка алгоритма исправления ошибок за пределами выравненных рёбер.
- Разработка алгоритма переноса полученных путей в графе обратно в последовательности.
- Реализация итогового алгоритма в виде отдельного инструмента.
- Апробация алгоритма на симулированных и реальных данных.

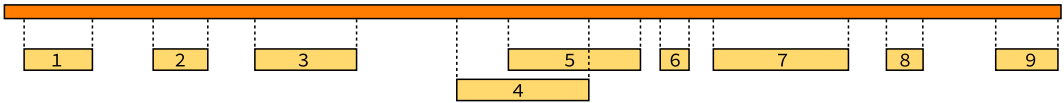
Существующие решения

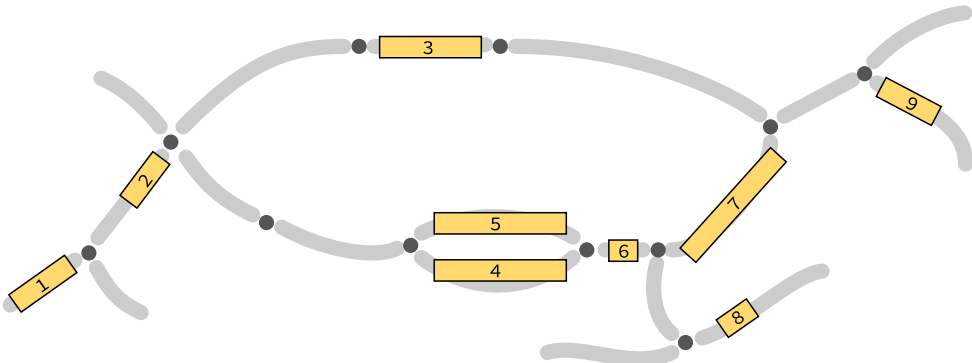
- LoRDEC
- Jabba
- HG-CoLoR
- FMLRC
- CoLoRMap
- Ratatosk

Этапы алгоритма коррекции

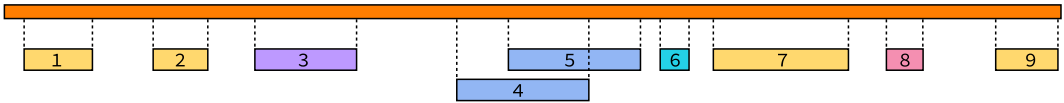


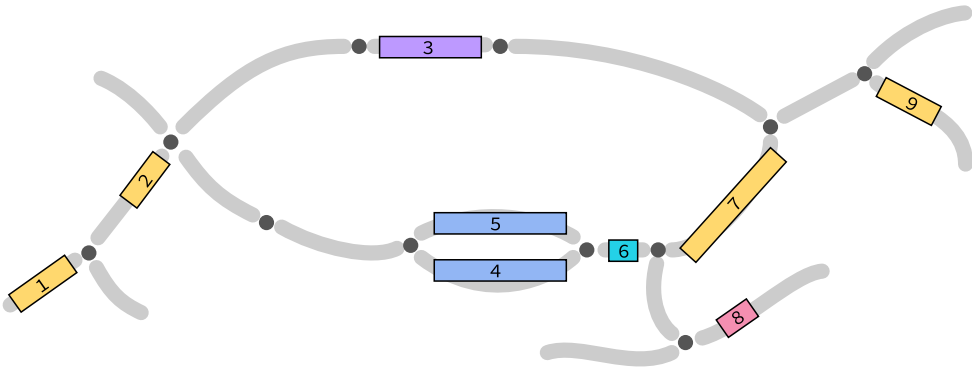
Алгоритм: выравнивание рёбер



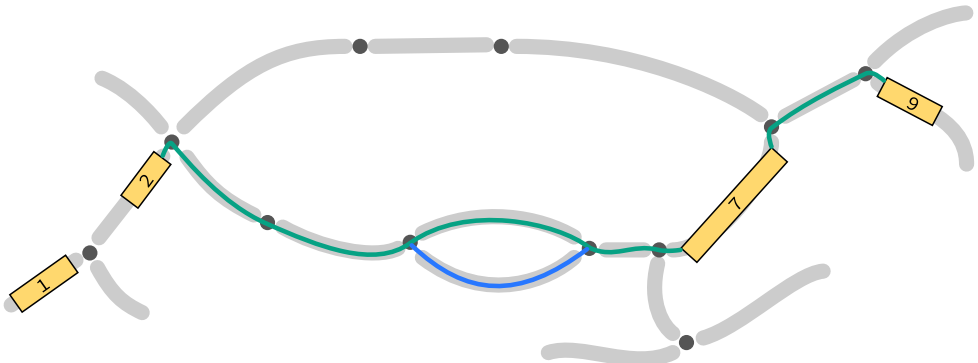


Алгоритм: фильтрация выравниваний





Алгоритм: реконструкция заполняющих путей



Может можно проще?

- Взять какой-нибудь выравниватель на граф
  - Например, GraphAligner
- По-умному смешать найденные пути

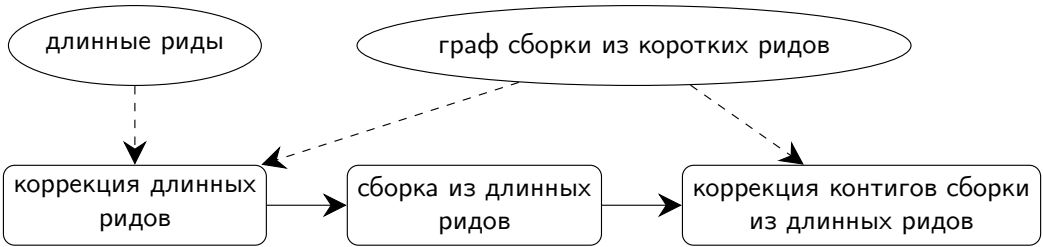
Вставки и удаления, альтернативный алгоритм

	flye	we	mixed
Bacillus subtilis	10111	181	91
Enterococcus faecalis	8393	562	82
Escherichia coli	9213	484	812
Lactobacillus fermentum	3985	33	52
Listeria monocytogenes	7181	78	82
Pseudomonas aeruginosa	4928	42	70
Saccharomyces cerevisiae	32845	7739	6985
Salmonella enterica	10119	166	10433
Staphylococcus aureus	4643	155	106

Замены, альтернативный алгоритм

	flye	we	mixed
Bacillus subtilis	2124	207	269
Enterococcus faecalis	593	547	667
Escherichia coli	13012	1401	10740
Lactobacillus fermentum	412	104	489
Listeria monocytogenes	149	64	840
Pseudomonas aeruginosa	1077	606	937
Saccharomyces cerevisiae	15441	10943	10670
Salmonella enterica	12816	352	199995
Staphylococcus aureus	891	157	1002

Апробация: схема сравнения



Апробация: Симулированные данные

	raw flye	ratatosk contigs	ratatosk reads	our contigs	ratatosk and we	ratatosk ratatosk
Покрытие генома	83.21	83.15	82.52	83.22	82.52	82.47
Структурные ошибки	34	34	50	34	40	55
Замены на 100kbp	369.97	105.08	112.19	121.67	67.22	101.24
Вставки и удаления на 100kbp	688.01	194.1	199.81	175.13	87.29	176.19

Апробация: Bmok12

	raw flye	ratatosk contigs	ratatosk reads	our contigs	ratatosk and we	ratatosk ratatosk
Покрытие генома	62.40	62.36	64.43	62.41	64.45	64.40
Структурные ошибки	72	77	129	72	99	145
Замены на 100kbp	342.69	127.27	203.28	144.39	166.3	180.82
Вставки и удаления на 100kbp	654.71	61.02	88.76	41.56	28.6	46.28

Апробация: Zymo

	raw flye	ratatosk contigs	ratatosk reads	our contigs	ratatosk and we	ratatosk ratatosk
Покрытие генома	97.12	97.09	96.93	97.11	96.91	96.98
Структурные ошибки	69	68	72	67	71	66
Замены на 100kbp	109.39	60.51	42.98	34.02	34.58	36.09
Вставки и удаления на 100kbp	214.49	47.68	21.3	22.22	18.07	21.64

Заключение

- Сформированы критерии фильтрации выравниваний рёбер графа на последовательности.
- Разработан алгоритм исправления ошибок за пределами выравненных рёбер.
- Разработан алгоритм переноса полученных путей в графе обратно в последовательности.
- Итоговый алгоритм реализован в виде отдельного инструмента.
  - Реализация выполнена на языке C++ и является подпроектом для ассемблера SPAdes.
  - Исходный код SPAdes доступен по ссылке: `https://github.com/ablab/spades/`. Реализованный инструмент будет доступен начиная с версии 3.17.
- Проведена апробация алгоритма на симулированных и реальных данных.
  - Сравнимый и лучший результат по сравнению с коррекцией ридов и сборок существующими решениями.
  - Кооперация даёт существенно лучший результат.