Алгоритмы для поиска полиморфизмов в графах геномных сборок

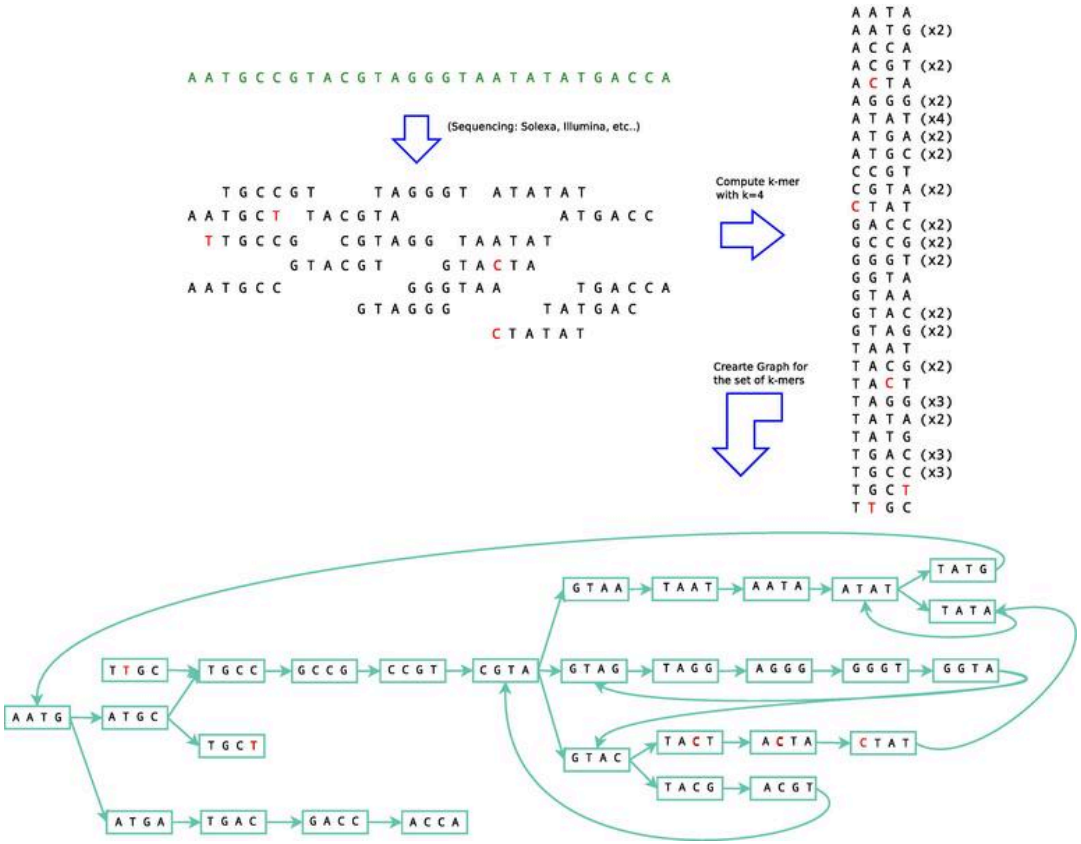
Щербин Егор

Руководитель: Мелешко Дмитрий
СПб АУ РАН

Весна 2017

Графы геномных сборок

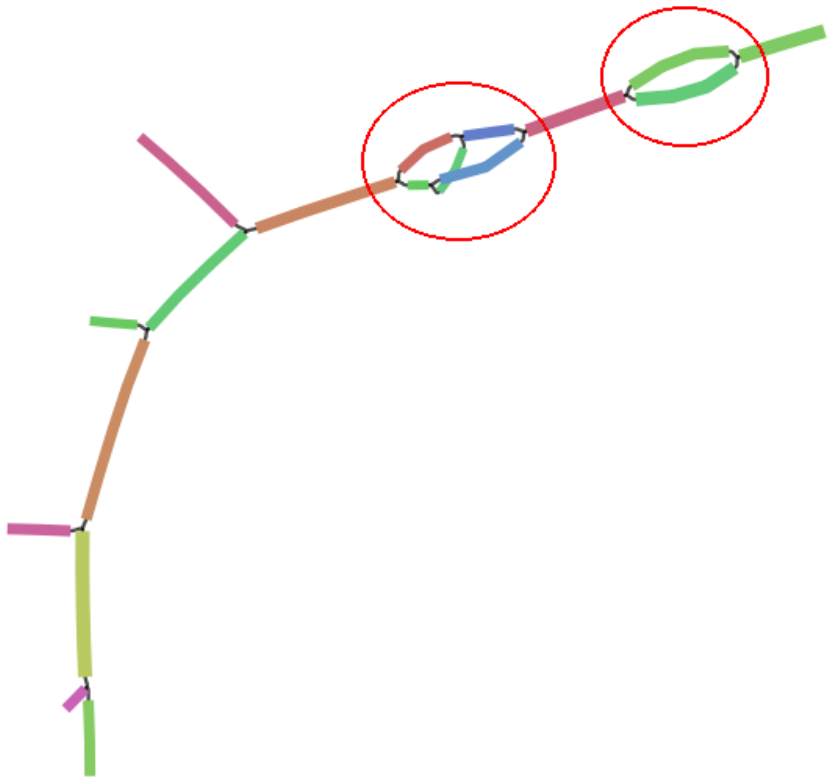
Общая схема работы ассемблера:



Источник: en.wikipedia.org/wiki/Velvet_assembler

Полиморфизмы

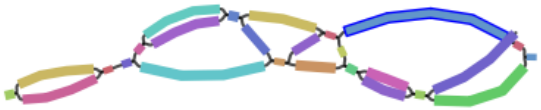
- В графах сборок выделяют некоторые виды подструктур
- Например, полиморфизмы (пузыри):

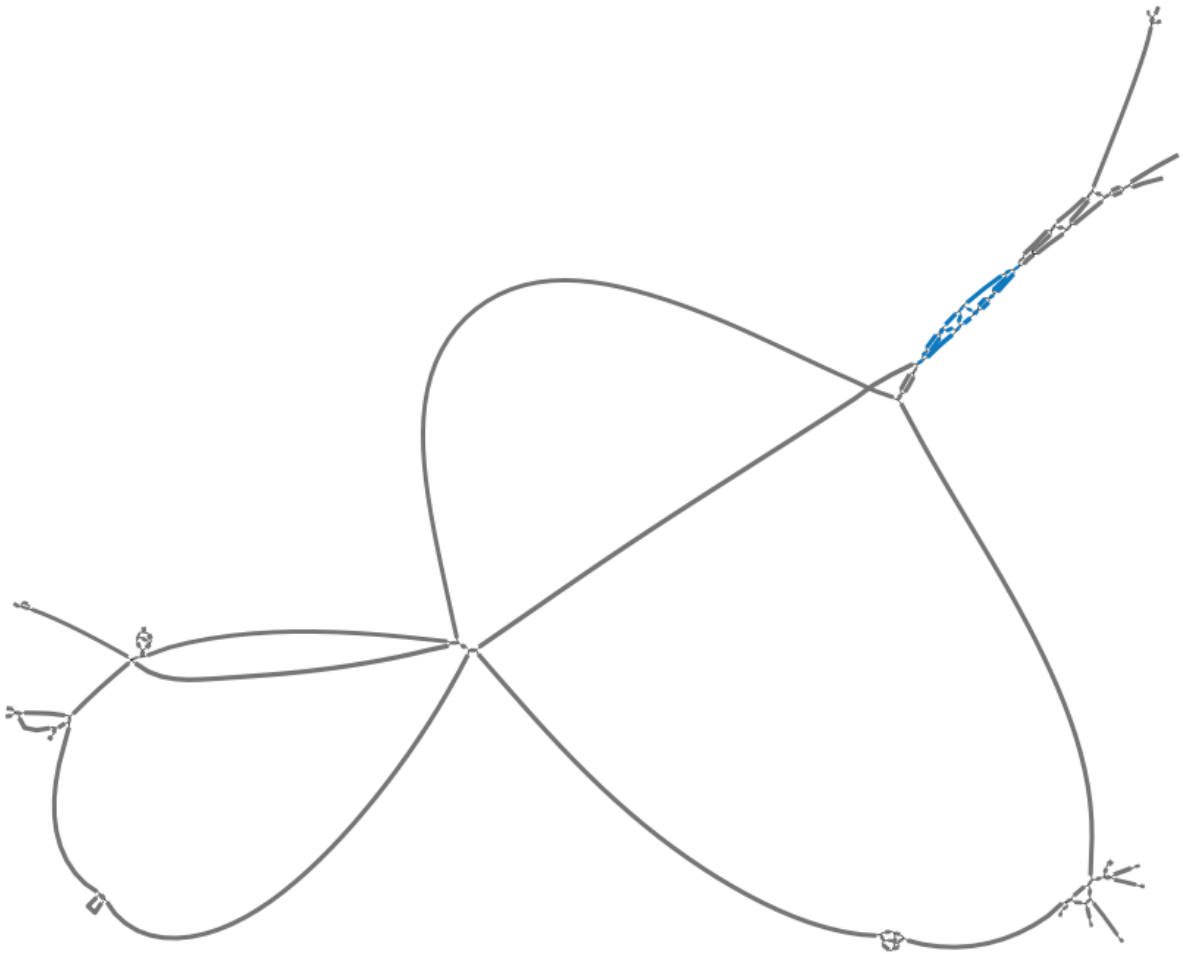


Задачи

- Изучить существующие алгоритмы поиска полиморфизмов или разработать новый.
- Реализовать выбранные алгоритмы как часть ассемблера SPAdes.
- Проанализировать полученные результаты и разработать методы обработки найденных полиморфизмов.

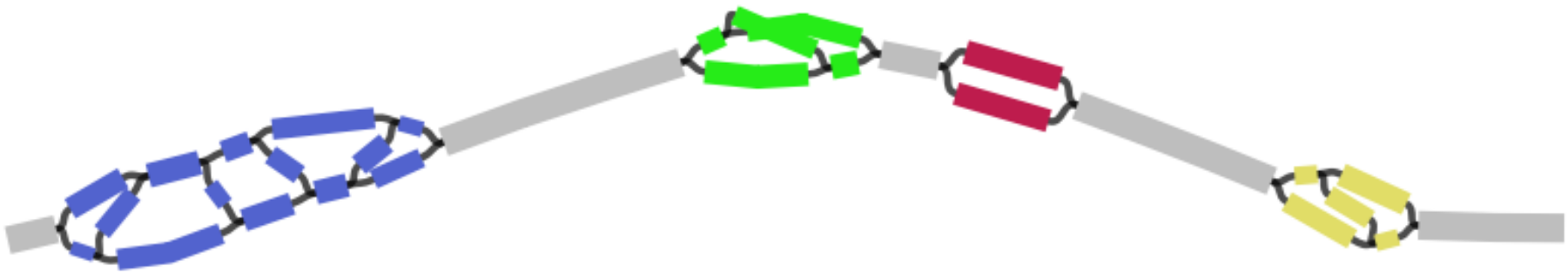
Реальный случай





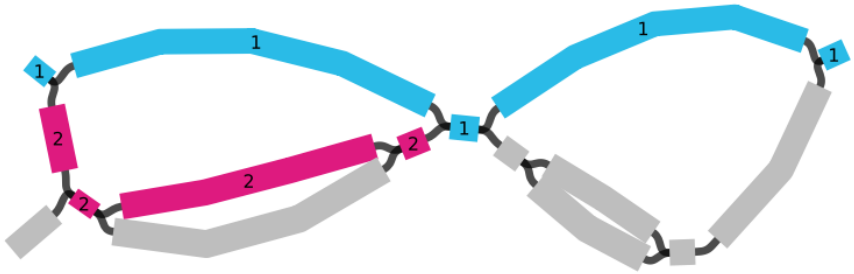
Поиск суперпузырей

- Модификация алгоритма топологической сортировки графа.
- Для каждого потенциального входа в пузырь ищет возможный выход.
- Найденные структуры имеют строгий математический вид.

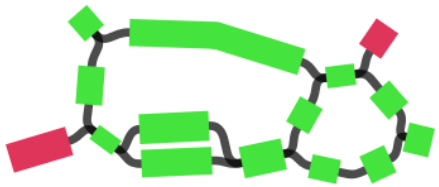


Tour Bus

- Модификация алгоритма Дейкстры.
- Когда заходим в посещенную вершину, откатываемся до общего предка и сливаем ребра пути в одно множество.
- Каждое множество в итоге образует полиморфизм.



Приливаем путь 2 к пути 1.



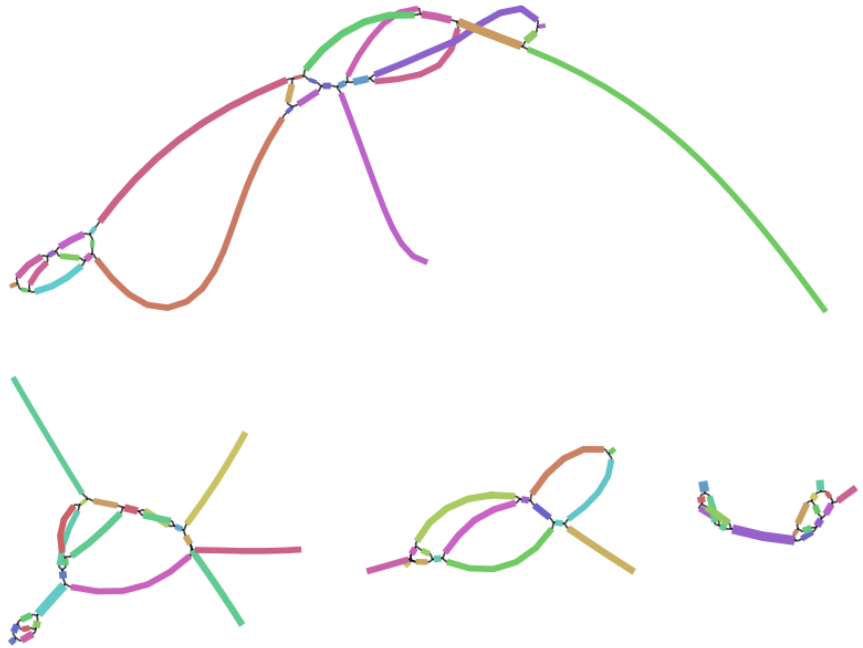
Найденный полиморфизм с отростками.

Что получилось

- Реализованы два метода поиска полиморфизмов в графах геномных сборок.
- Найденные полиморфизмы выводятся в популярном формате представления графов GFA.
- Оба алгоритма написаны на C++ внутри инфраструктуры ассемблера SPAdes.
- Бонус: в SPAdes реализован вывод графов в новом формате GFA2.

Примеры найденных полиморфизмов

Superbubbles



Tour Bus bulges

Пример запуска

```
[end]
GAF$> help find_bubbles
Command `find_bubbles`
Usage:
> find_bubbles
 This command finds all superbubbles in the graph and outputs them to a GFA file.
[end]
GAF$> help find_tourbus_bulges
Command `find_tourbus_bulges`
Usage:
> find_tourbus_bulges
 This command runs a modification of the Tour Bus algorithm and outputs all found bulges to a GFA file.
[end]
GAF$> find_bubbles
2091 superbubbles found
Found superbubbles were written to pictures_presentation/superbubbles.gfa
[end]
GAF$> find_tourbus_bulges
1192 bulges found with 7693 edges in total
Filtered out bulges with several sinks, 1097 bulges left
Found bulges were written to pictures_presentation/tourbus_bulges.gfa
[end]
GAF$>
```

Дальнейшее развитие

- Проанализировать полученные результаты и подобрать подходящие параметры запуска алгоритмов.
- Продумать и реализовать механизм аннотирования найденных полиморфизмов.
- Протестировать написанные алгоритмы и внедрить их в конвейер сборки SPAdes.

Список литературы

1. Nijkamp, J. F., Pop, M., Reinders, M. J. T., & de Ridder, D. (2013). Exploring variation-aware contig graphs for (comparative) metagenomics using MaryGold. *Bioinformatics*, 29(22), 2826–2834. doi:10.1093/bioinformatics/btt502
2. Onodera, T., Sadakane, K., Shibuya, T.: Detecting superbubbles in assembly graphs. In: Darling, A., Stoye, J. (eds.) WABI 2013. LNCS, vol. 8126, pp. 338–348. Springer, Heidelberg (2013). doi:10.1007/978-3-642-40453-5_26
3. Zerbino, D.R., Birney, E.: Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* 18(5), 821–829 (2008). doi:10.1101/gr.074492.107