

ИБ 2020: De Bruijn Graph assembler

Артем Иванов

24 апреля 2021 г.

Программа строит сжатый граф де Брейна и выводит полученные ребра в формате FASTA, а также полученный граф в формате GFA1, который можно визуализировать, например в Bandage.

Доступны две стратегии очистки графа:

1. Отбрасывание хвостиков (**strategy="tails"**) – отбрасываются все тупиковые ребра, у которых покрытие в 2 раза меньше, чем среднее покрытие сборки, или длина ребра меньше $2 \cdot k$.
2. Отбрасывание низкопокрытых ребер (**strategy="lowcov"**) – отбрасываются все ребра, у которых покрытие в 2 раза меньше, чем среднее покрытие сборки, или длина ребра меньше $2 \cdot k$.

Результаты запуска доступны в папке **out**. Ниже приведены результаты сравнения различных стратегий. Если не использовать очистку, то в графе много коротких ребер и он очень шумный. При отбрасывании всех низкопокрытых ребер лишний мусор убирается, однако граф разваливается на части. Отбрасывание хвостиков работает лучше всего, получаются очень длинные контиги с редкими неотброшенными хвостиками и пузырями.

стратегия	статистика	граф
без очистки	nodes=14 edges=8 total len=2648	
отбрасывание хвостиков	nodes=2 edges=0 total len=1786	
отбрасывание всех низкопокрытых	nodes=2 edges=0 total len=1786	

Таблица 1: Сравнение различных стратегий очистки графа для $s_6.first1000$

стратегия	статистика	граф
без очистки	nodes=82 edges=80 total len=24584	
отбрасывание хвостиков	nodes=2 edges=0 total len=20000	
отбрасывание всех низкопокрытых	nodes=4 edges=0 total len=19846	

Таблица 2: Сравнение различных стратегий очистки графа для $s_6.first10000$


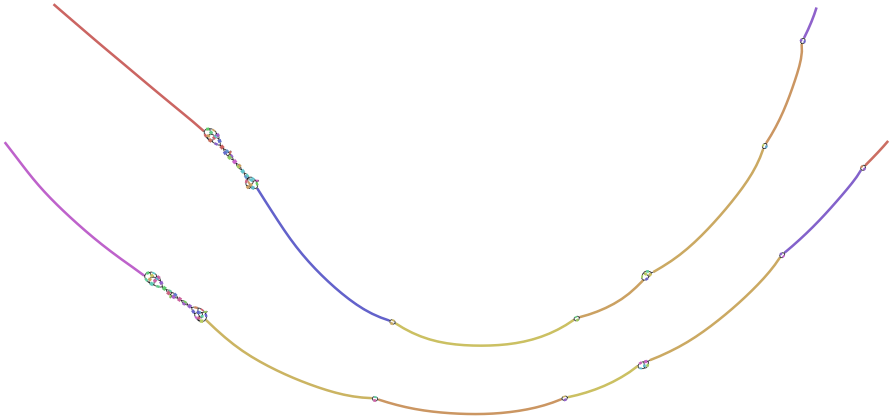
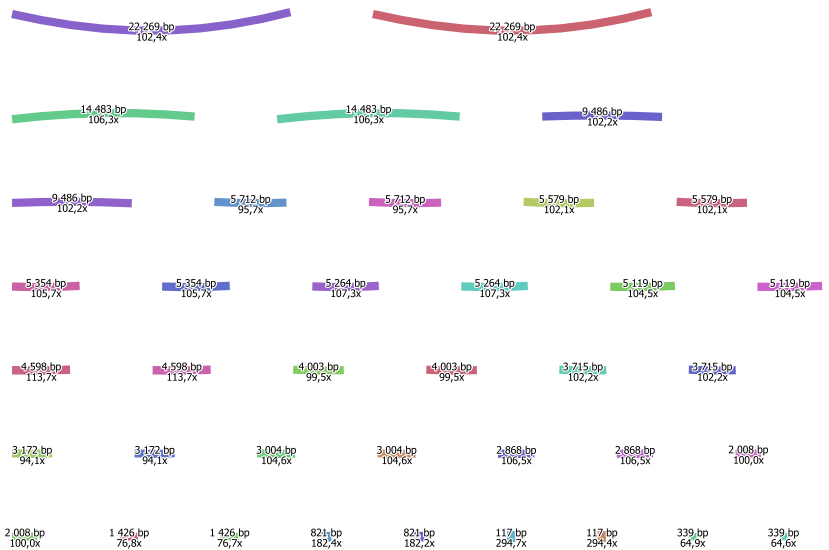
стратегия	статистика	граф
без очистки	nodes=1014 edges=1042 total len=260024	
отбрасывание хвостиков	nodes=104 edges=136 total len=206022	
отбрасывание всех низкопокрытых	nodes=38 edges=0 total len=198674	 <p>22,269.bp 102,7x 22,269.bp 102,7x</p> <p>14,483.bp 106,3x 14,483.bp 106,3x 9,486.bp 102,2x</p> <p>9,486.bp 102,2x 5,712.bp 95,7x 5,712.bp 95,7x 5,579.bp 102,1x 5,579.bp 102,1x</p> <p>5,354.bp 105,7x 5,354.bp 105,7x 5,264.bp 107,3x 5,264.bp 107,3x 5,119.bp 104,5x 5,119.bp 104,5x</p> <p>4,598.bp 113,7x 4,598.bp 113,7x 4,003.bp 99,5x 4,003.bp 99,5x 3,715.bp 102,2x 3,715.bp 102,2x</p> <p>3,172.bp 94,1x 3,172.bp 94,1x 3,004.bp 104,6x 3,004.bp 104,6x 2,868.bp 106,5x 2,868.bp 106,5x 2,008.bp 100,0x</p> <p>2,008.bp 100,0x 1,426.bp 76,8x 1,426.bp 76,7x 821.bp 182,4x 821.bp 182,2x 117.bp 294,7x 117.bp 294,4x 339.bp 64,9x 339.bp 64,6x</p>

Таблица 3: Сравнение различных стратегий очистки графа для $s_6.first100000$