

SVEUČILIŠTE U ZAGREBU  
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

**DOKUMENTACIJA**

**Poboljšanje djelomično  
sastavljenog genoma dugim  
očitanjima**

*Bruno Kovač, Tonko Sabolčec, Fabijan Čorak*

*Voditelj: doc. dr. sc. Krešimir Križanović*

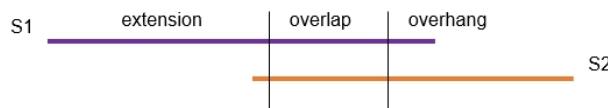
Zagreb, siječanj 2020.

# SADRŽAJ

<b>1. Uvod</b>	<b>1</b>
<b>2. Postupak</b>	<b>2</b>
2.1. Izgradnja grafa . . . . .	2
2.2. Obilazak grafa . . . . .	2
2.3. Obrada putova . . . . .	3
<b>3. Rezultati</b>	<b>5</b>
3.1. Generirani podaci . . . . .	5
3.2. E. Coli . . . . .	6
3.3. C. Jejuni . . . . .	9
3.4. B. Grahamii . . . . .	10
3.5. Vremensko i memorijsko opterećenje . . . . .	11
<b>4. Zaključak</b>	<b>12</b>
<b>5. Literatura</b>	<b>13</b>

# 1. Uvod

Sekvenciranje genoma svodi se na kombiniranje očitanja u jednu cjelinu. Ovaj rad pretpostavlja da su očitanja već sastavljena, ali djelomično - u fragmente. Jedan takav fragment naziva se *contig*. Dakle, zadatok se svodi na što bolje povezivanje *contiga*, što smo učinili postupkom opisanim u [1]. Taj rad definira nekoliko mjera preklopljenosti očitanja koje kombiniraju duljinu područja *overlap* (*OL*), *overhang* (*OH*) i *extension* (*EL*). Mjere su ovdje definirane za dva očitanja  $S_1$  i  $S_2$ ; pripadnost područja određena je indeksom.



**Slika 1.1:** Preklop dvaju očitanja s naznačenim područjima

- *sequence identity (SI)* - omjer ukupnog broja podudarajućih znakova u *overlap* područjima i duljine duljeg od tih dvaju područja

$$SI = \frac{\text{broj\_podudaranja}}{\max(OL_1, OL_2)}$$

- *overlap score (OS)*

$$OS = (OL_1 + OL_2) \frac{SI}{2}$$

- *extension score (ES)* - uz  $S_2$  kao produžetak od  $S_1$

$$ES_2 = OS + \frac{EL_2 - OH_1 - OH_2}{2}$$

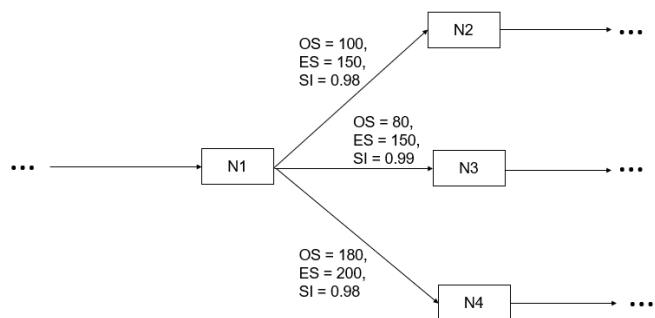
## 2. Postupak

Sastavljanje očitanja u niz modelirano je izgradnjom i obilaskom grafa.

### 2.1. Izgradnja grafa

Svaki *contig* i svako očitanje čine jedan čvor grafa. Dodatno, u skup čvorova dodaje se i reverzna inačica svakog očitanja i *contiga* jer nije unaprijed poznato koja je pogodna orijentacija. Čvor koji predstavlja *contig* zovemo *anchor*. Za svako očitanje Brid postoji između svaka dva čvora čiji je *SI* veći od nekog minimuma. Pritom svaki brid nosi informacije o preklopjenosti čvorova koje povezuje (*SI*, *OS*, *ES*). Te mjerne računaju se na temelju informacija o preklopjenosti dobivenih korištenjem alata *minimap2* opisanog u [3].

### 2.2. Obilazak grafa



**Slika 2.1:** Odabir sljedećeg čvora u obilasku

Slika razmatra grananje u hipotetskom čvoru N1. Prvim postupkom prioritetna lista je: [N4, N2, N3], a drugim postupkom: [N4, N3, N2]. Treći postupak daje vjerojatnosti odabira čvorova {N2: 150/500, N3: 150/500, N4: 200/500}.

Kroz graf se traže putovi čije su krajnje točke *anchor* čvorovi. Za to se koriste tri

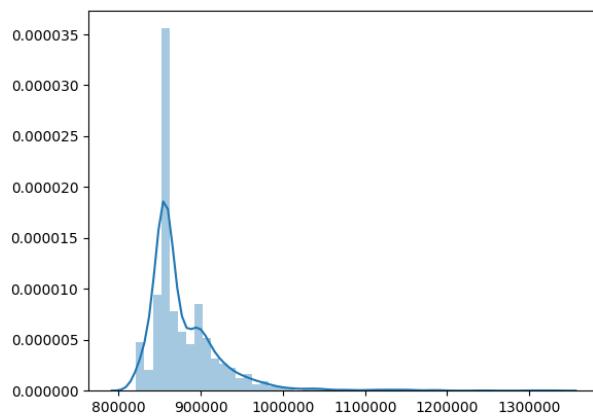
načina obilaska, prilikom kojih je postavljena maksimalna dubina pretraživanja i pamte se obični čvorovi:

1. Iz *anchor* čvora pretraga se nastavlja u sve susjedne čvorove. Iz svakog sljedećeg čvora, pretraga se nastavlja u onaj susjed s kojim je najveći *overlap score*, a kojim se u konačnici dolazi do *anchor* čvora. Ako je *overlap score* jednak, gleda se *sequence identity*. Ako je pak i ta mjeru jednak, gleda se duljina očitanja.
2. Kao i prethodni način, ali umjesto mjeru *overlap score* gleda se *extension score*.
3. U svakom čvoru susjed se odabire probabilistički - s vjerojatnošću odabira proporcionalnom mjeri *extension score*, sve dok se ne dosegne *anchor*. Postupak se pokreće iz svakog *anchor* čvora proizvoljan broj puta. Ovo je tzv. Monte Carlo metoda.

### 2.3. Obrada putova

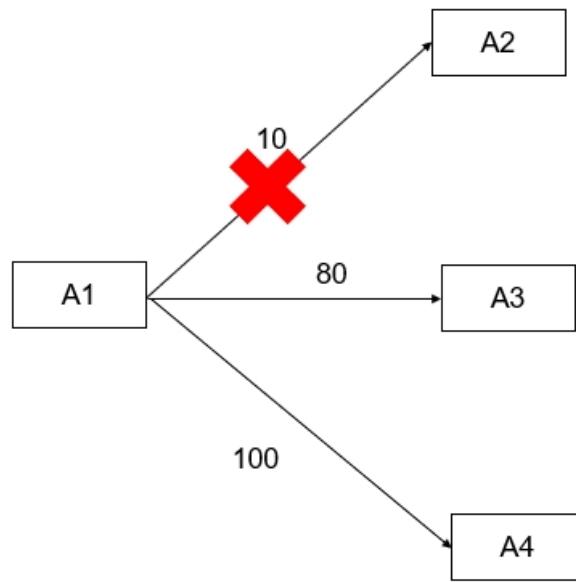
Nakon agregacije putova dobivenih opisanim postupcima, odbacuju se duplikati i obrađuju se putovi između svaka dva *anchor* čvora. Putovi se sortiraju uzlazno prema duljinama i razmatraju se prozori fiksne širine  $W$ . Pritom  $i$ -ti prozor obuhvaća sve putove čija je duljina  $L$ :  $(i-1)W < L \leq iW$  za dobro definirane  $i$ . Koliko je koji prozor frekventan vidljivo je na primjeru povezivanja dvaju *contiga* na slici 2.2. Pretpostavka je da dominacija nekog prozora ukazuje na to da su u njemu putovi koji su izgledni kandidati za povezivanje dvaju *anchor* čvorova između kojih se nalaze. Problem je što to ne mora biti istina, tj. moguće je dominiranje nekog prozora, a da ta dva *anchor* čvora uopće nisu uzastopna.

Za svaki par *anchor* čvorova računa se konsenzus - struktura podataka koja sadrži reprezentativni put te broj valjanih putova između tih dvaju čvorova. Kao reprezentant odabire se proizvoljni čvor maksimalne duljine. Na slici 2.2 to je jedan od putova duljine cca 850000 (najviši stupac). Dodatno, iz susjedstva svakog *anchor* čvora uklanjuju se oni *anchor* čvorovi do kojih je broj putova ispod neke određene granice. Drugim riječima, takvi susjedi vjerojatno ne trebaju biti povezani. U konačnici se konstruiraju poboljšani sljedovi *contiga* i za svako poboljšanje generira se izlazna datoteka. (OVDI DODAT KAKO SE KONSTRUIRAJU POBOLJŠANI SLJEDOVI)



**Slika 2.2:** Distrubucija putova po prozorima fiksne širine

Distrubucija putova između dva *anchor* čvora. Na x-osi označene su širine prozora, a na y-osi frekventnost (udio brojnosti putova u ukupnom broju putova). Najviši stupac dat će i reprezentant povezanosti dvaju čvorova.



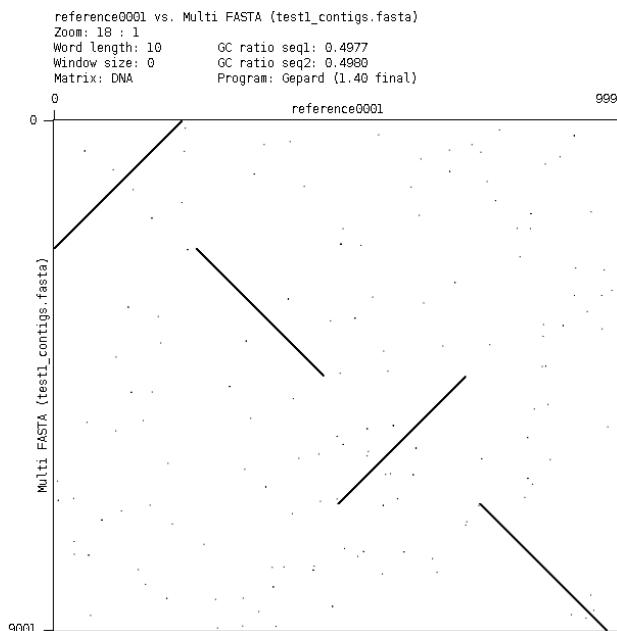
**Slika 2.3:** Otpadanje jedne grane u slučaju kada je granica za odbacivanje 12.5% od najvećeg broja valjanih putova.

# 3. Rezultati

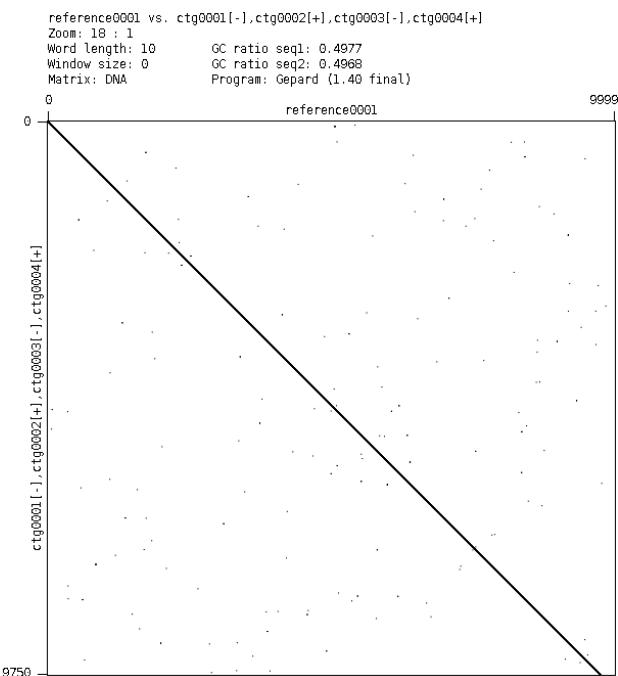
Implementacija je ispitana na skupovima E. Coli, C. Jejuni i B. Grahamii na jednoj dretvi uz procesorsku moć od ... GHz. Priložene su matrice preklapanja između referentnih skupova te prvo neobrađenih, a odmah zatim i obrađenih *contiga*. Za svaki skup podataka prikazana su moguća poboljšanja sastavljanja *contiga*. Sve matrice do bivene su alatom *Gepard* opisanim u [2]. Ishodište je u gornjem lijevom kutu, x-os predstavlja referencu, a y-os predstavlja *contige*. Crne točke označavaju podudaranja. Predznak u opisima matrica označava korištenu orijentaciju *contiga*: + je izvorna, a - reverzna. Npr. +5, -2 znači da je pronađeno povezivanje za izvorno orijentirani *contig* 5 i reverzni 2.

## 3.1. Generirani podaci

Implementacija je ispitana i na umjetno generiranom skupu podataka.



**Slika 3.1:** Matrica preklapanja neobrađenih umjetno generiranih *contiga* i točne reference.

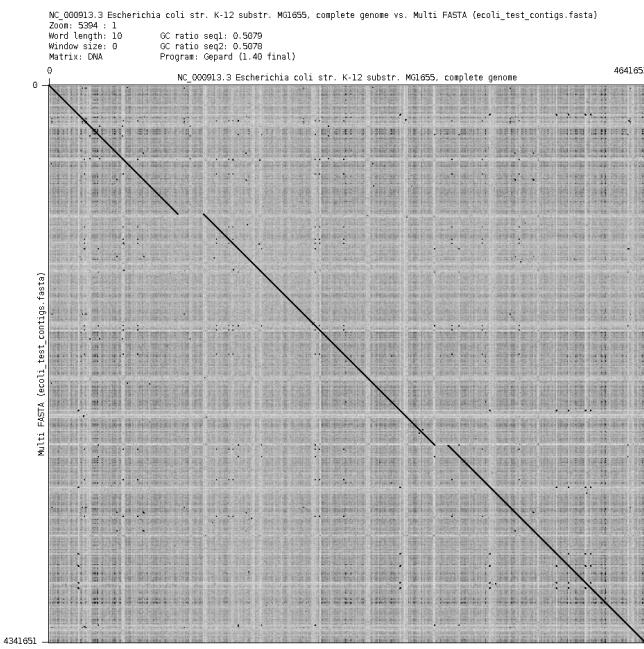


**Slika 3.2:** Poboljšanje generiranog testnog slučaja.

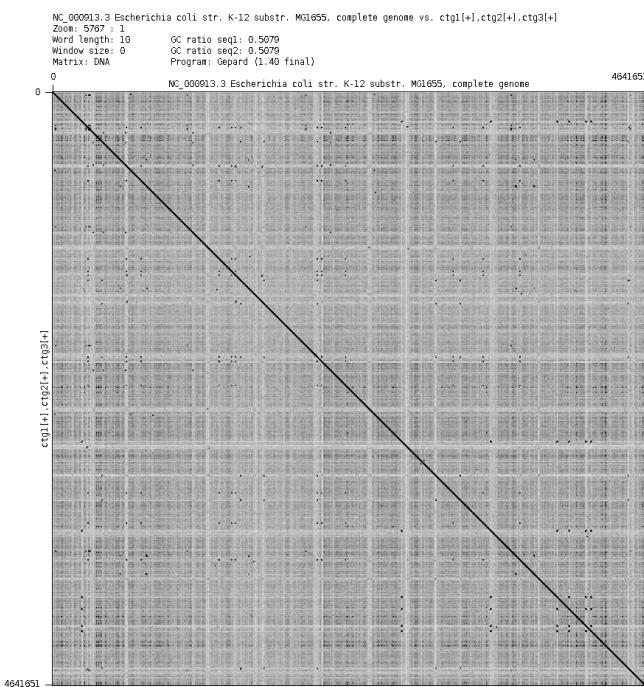
Pronađeni redoslijed *contiga* je -1, +2, -3, +4.

## 3.2. E. Coli

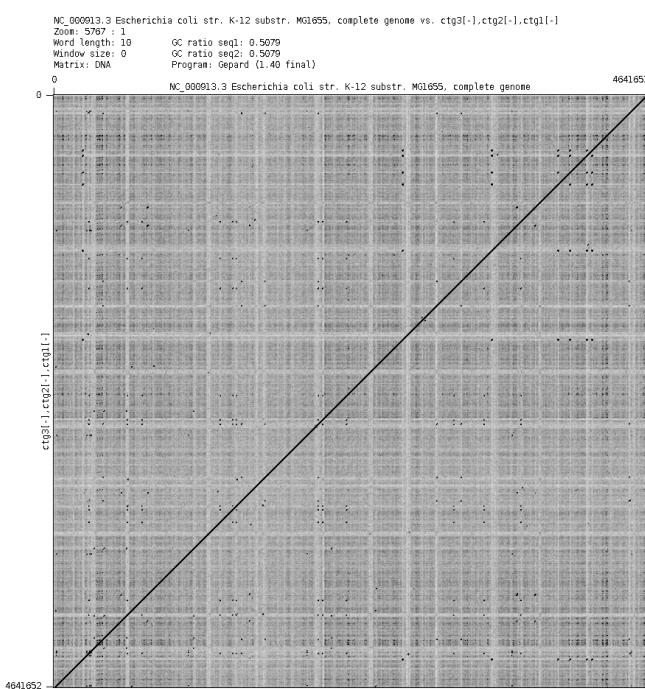
Ovo je sintetski skup podataka, shodno čemu su i ostvareni rezultati bolji u usporedbi s narednim dvama, realnim skupovima. Rezultirajući izlazi programa zapravo su inverzi jedan drugog.



**Slika 3.3:** Matrica preklapanja neobrađenih *contiga* E. Coli i referentnog skupa.



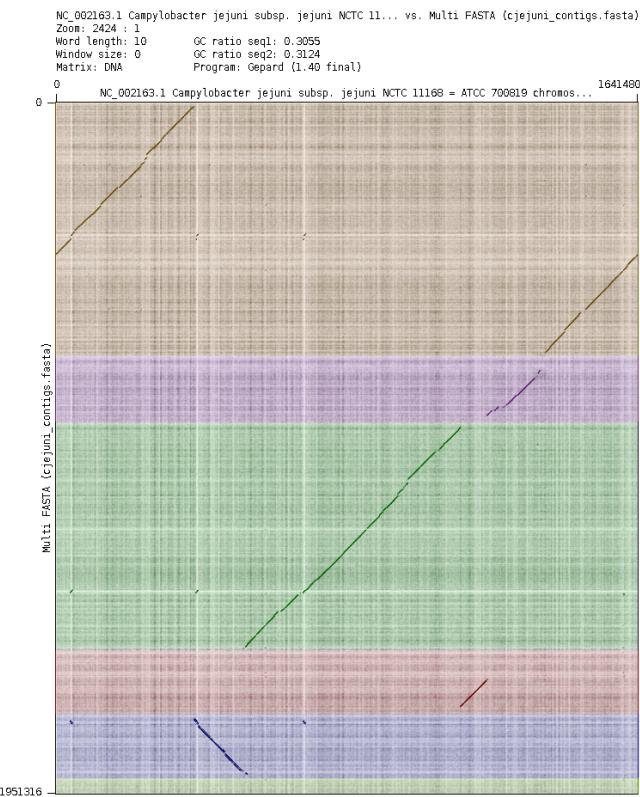
**Slika 3.4:** Prvo poboljšanje: niz *contiga* +1, +2, +3.



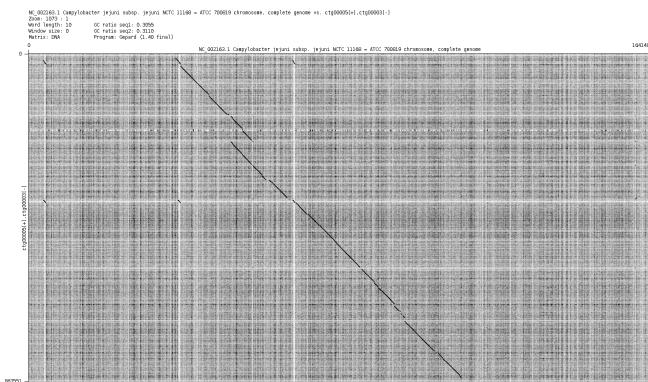
**Slika 3.5:** Drugo poboljšanje: niz *contiga* -3, -2, -1.

### 3.3. C. Jejuni

Prvi realni skup podataka predstavlja očitanja nad genomom bakterije C. Jejuni.



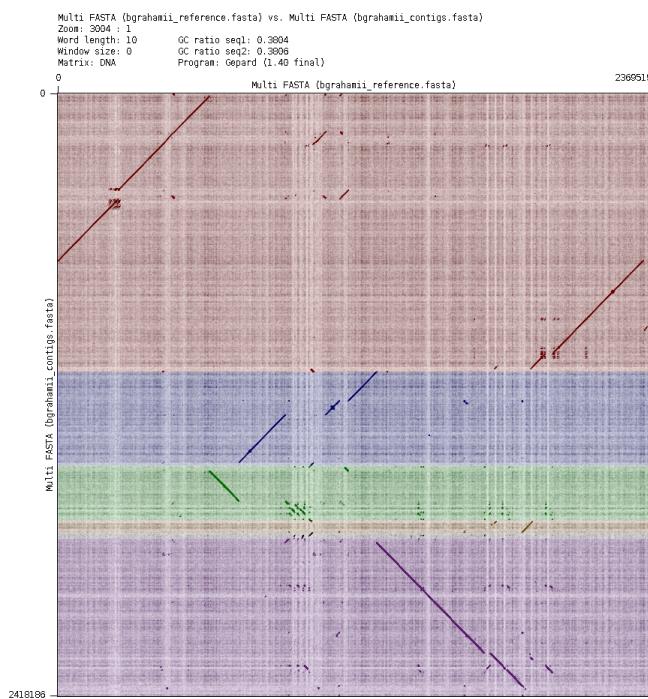
**Slika 3.6:** Matrica preklapanja neobrađenih *contiga* C. Jejuni i referentnog skupa. Svaki contig predstavljen je jednom bojom. *Contig 1* (njegornji) već je dakle u ulaznoj datoteci definiran *izlomljeno*, tj. obuhvaća očitanja sa stvarnog početka i kraja.



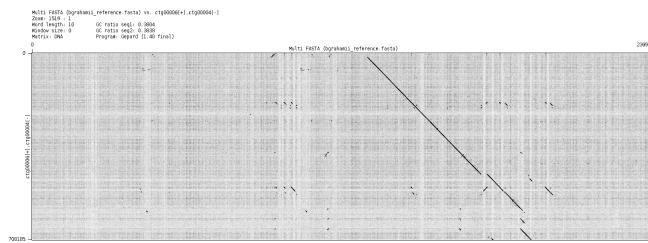
**Slika 3.7:** Poboljšanje: niz *contiga* +5, -3.

### 3.4. B. Grahamii

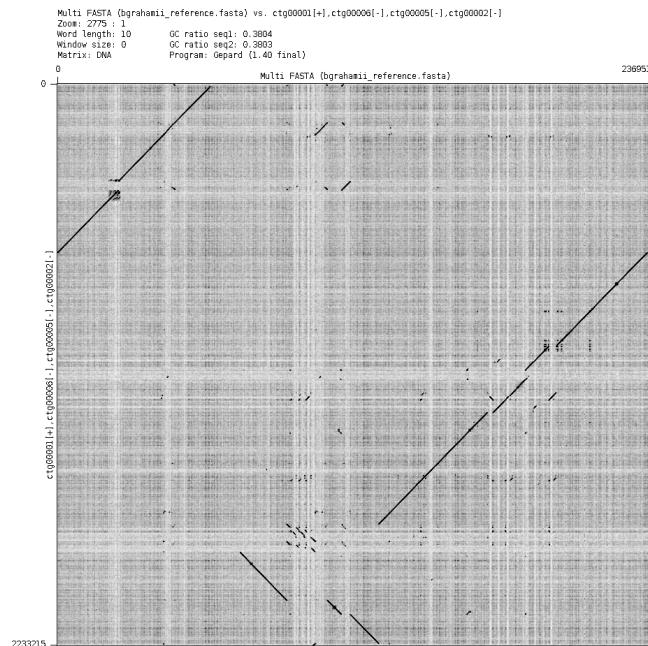
Drugi realni skup podataka čine očitanja genoma bakterije B. Grahamii.



**Slika 3.8:** Matrica preklapanja neobrađenih *contiga* B. Grahamii i referentnog skupa. Svaki contig predstavljen je jednom bojom. Ponovno su vidljive izlomljenosti u ulaznim podacima.



**Slika 3.9:** Poboljšanje: niz *contiga* +6, -4.



**Slika 3.10:** Poboljšanje: niz *contiga* +1, -6, -5, -2.

### 3.5. Vremensko i memorijsko opterećenje

skup podataka	vrijeme (s)	memorija (GiB)
E. Coli	X	Y
C. Jejuni	B	C
B. Grahamii	W	Z

## **4. Zaključak**

Zaključak.

## 5. Literatura

- [1] Huilong Du i Chengzhi Liang. Assembly of chromosome-scale contigs by efficiently resolving repetitive sequences with long reads. *bioRxiv*, 2018. doi: 10.1101/345983. URL <https://www.biorxiv.org/content/early/2018/06/13/345983>.
- [2] Jan Krumsiek, Roland Arnold, i Thomas Rattei. Gepard: a rapid and sensitive tool for creating dotplots on genome scale. *Bioinformatics*, 23(8):1026–1028, 2007.
- [3] Heng Li. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18):3094–3100, 05 2018. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty191. URL <https://doi.org/10.1093/bioinformatics/bty191>.