PROJEKTNI ZADATAK IZ PREDMETA BIOINFORMATIKA 1

Poboljšanje djelomično sastavljenog genoma dugim očitanjima

OPIS PROBLEMA

Korištenjem danas dostupnih alata za čitanje DNA nije moguće napraviti očitanje dovoljno dugačko da pokrije cijeli genom. Zbog ograničenja alata na kraća očitanja za konstrukciju genoma potrebno je kratka očitanja sastaviti u jedinstvenu sekvencu. Najčešći pristup rješavanju ovog problema je tzv. shotgun sekvenciranje. Cijeli genom se replicira te se dostupnim alatima generira mnogo kratkih očitanja koja je potrebno ponovno sastaviti. U procesu sastavljanja nemamo nikakvu informaciju o poretku očitanja, odnosi među očitanjima pokušavaju se rekonstruirati pomoću međusobnih preklapanja. Alati za sastavljanje genoma ne mogu uvijek generirati jedan slijed bez dijelova nepoznatog sadržaja. Ponekad je moguće generirati samo osnovna očitanja-contig-e, koja se sastoje od uspješno spojenog niza očitanja koja se preklapaju, i scaffold-e, skup contig-a s definiranim poretkom. Naknadnim postupcima skup contig-a moguće je povezati u cijeli genom. Postupak scaffolding-a koristi duga očitanja kako bi contig-e djelomično sastavljenog genoma međusobno povezao u dulje sekvence ili čitavi genom. Jedan od algoritama koji se koristi za poboljšanje djelomično sastavljenog genoma, HERA algoritam, implementiran je u sklopu ovog projektnog zadatka.

OPIS ALGORITMA

Ulazni podaci su skup dugih očitanja i već sastavljenih contig-a, preklapanja između contig-a i očitanja te međusobna preklapanja između očitanja. Prvi korak u algoritmu je sastavljanje grafa preklapanja. Graf se sastoji od dva tipa čvorova koji su međusobno povezani. Manji skup čvorova su sidreni čvorovi, oni nastaju iz contig-a, a ostali čvorovi čine skup čvorova očitanja, dobivenih iz dugih očitanja. Svaki čvor očitanja spaja se s maksimalno jednim sidrenim čvorom. Bridovi koji spajaju očitanja koja se preklapaju su usmjereni. Brid se dodaje u graf samo ako očitanje ciljnog čvora proširuje očitanje početnog čvora.

Spojevi između sidrenih čvorova dobivaju se pretraživanjem grafa u dubinu na tri različita načina. Koriste se dvije mjere kvalitete preklapanja, rezultat preklapanja (overlap score) i rezultat produljenja (extension score). Uz oznake: OL za duljinu preklapanja, SI za broj baza koje se preklapaju, overhang OH i duljinu produljenja EL rezultat preklapanja definiran je kao: (OL1+OL2)*SI/2, a rezultat produljenja kao ES2 = OS + EL/2 – (OH1 – OH2)/2.

U prvom načinu pretraživanja se za svaki od čvorova koji se preklapaju sa sidrenim čvorom pokušava generirati put. Sljedeći čvorovi biraju na temelju rezultata preklapanja. U slučaju da nije moguće pronaći put do drugog sidrenog čvora, algoritam se vraća na prijašnji čvor i nastavlja potragu novim čvorom sa drugim najboljim rezultatom. Drugi način sličan je prvom, jedina razlika je u tome što se koristi rezultat produljenja umjesto rezultata preklapanja. Treći način oslanja se na nedeterminističko generiranje putova. Koristi se Monte Carlo pristup kojim se slučajno odabire sljedeći čvor, vjerojatnost da neki čvor bude odabran proporcionalna je rezultatu produljenja.

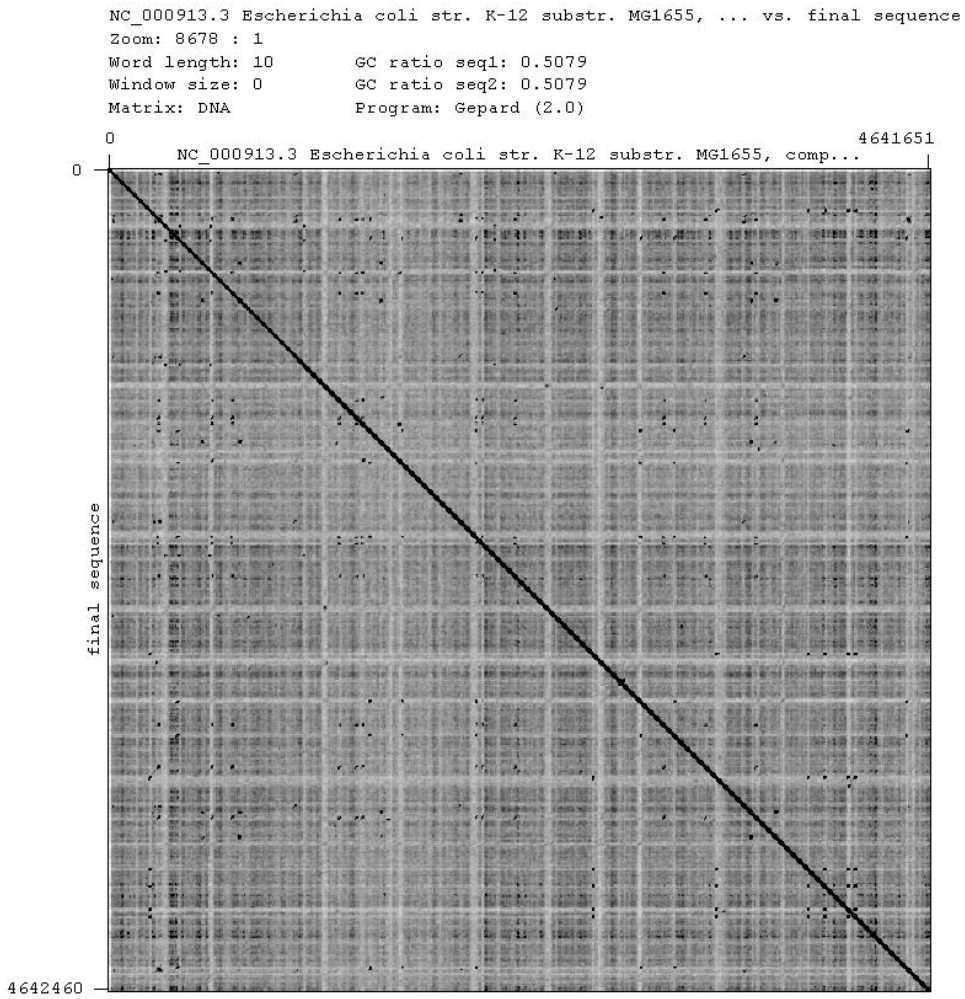
Prethodno objašnjenim metodama generira se skup mogućih puteva iz jednog sidrenog čvora (contiga). Za odabir puta koristi se algoritam koji je detaljnije opisan u HERA radu. Ukratko, skup puteva se

particionira tako da se u jednom podskupu nalaze svi putevi slične dužine. Zatim se za svaki podskup odreduje reprezentativni put i broj ispravnih puteva (valid path count). Reprezentativni put je onaj koji po dužini sličan najvećem broju drugih puteva u particiji, a broj ispravnih puteva je broj puteva kojima je reprezentativni put sličan. Time nastaje skup reprezentativnih puteva (jedan za svaku particiju). Konačan put iz sidrenog čvora određuje se iz skupa reprezentativnih puteva uzimajući u obzir dužinu puta i broj ispravnih puteva. Ovaj postupak se ponavlja za svaki sidreni čvor. Na kraju algoritma svi odabrani putevi se spajaju u jedan put kroz sve sidrene čvorove temeljem kojeg se generira konačna sekvenca.

REZULTATI TESTIRANJA I ZAKLUČAK

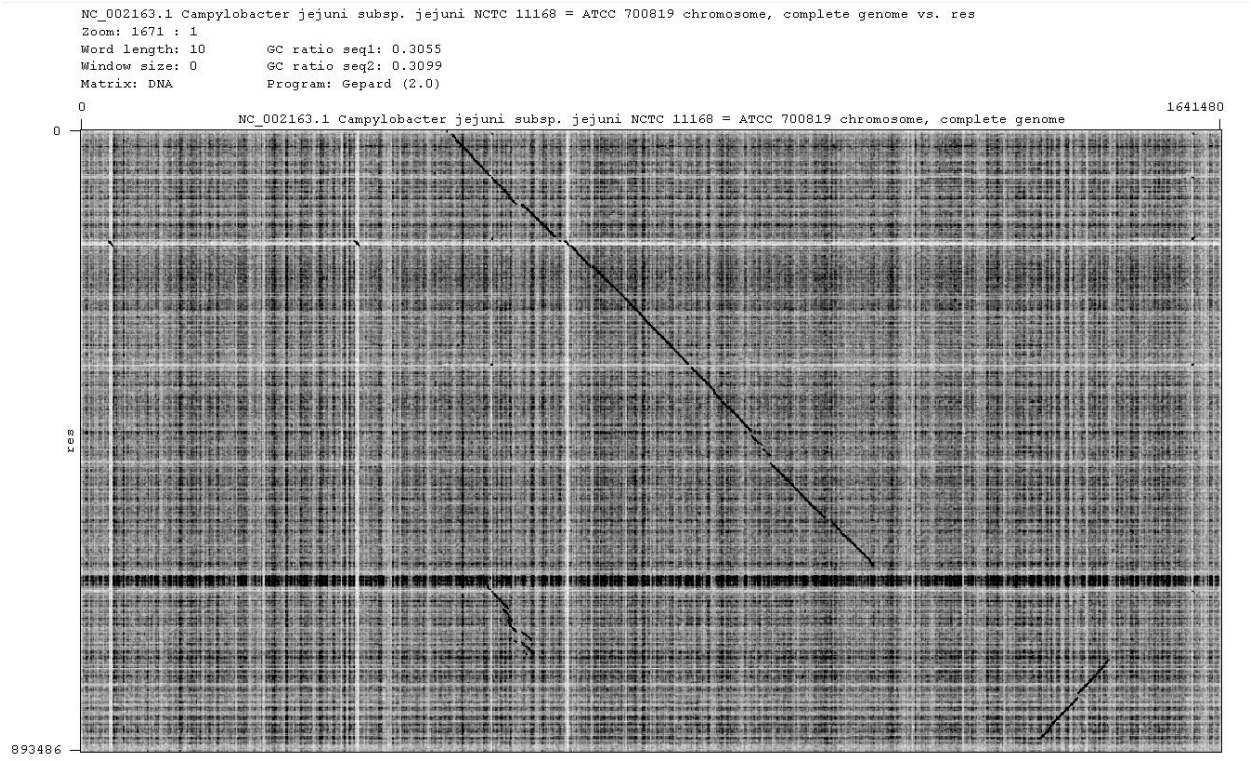
Implementacija algoritma testirana je na jednom primjeru sintetičkih podataka i dva primjera stvarnih.

Na testnim esherihia coli podacima implementacija funkcionira bez problema. Rezultat je prikazan na slici 1. Vrijeme izvođenja je 65 sekundi, a memorijsko zauzeće ne prelazi 1.5 GB



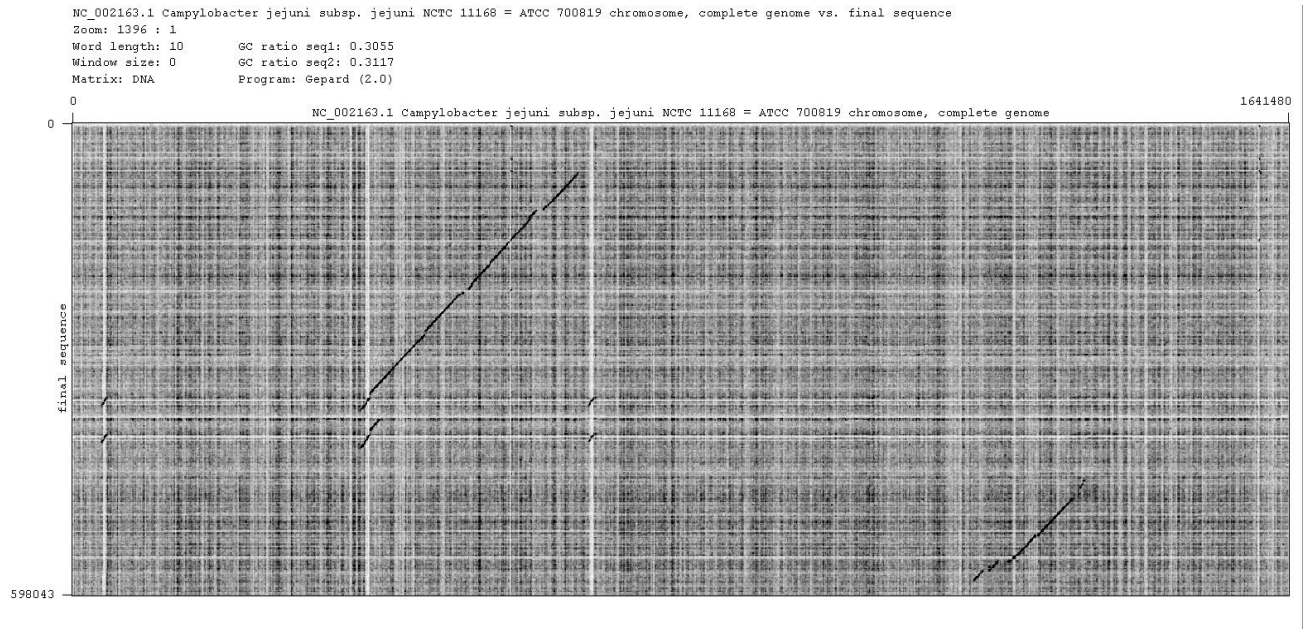
Slika 1. Usporedba generirane sekvence s referentom esherihia coli sekvencom

Na stvarnim primjerima dobiveni su lošiji rezultati. Najbolji rezultat za campylobacter jejuni postignut je za 9 minuta i maksimalno zauzeće memorije od 8 GB. Prikazan je na slici 2.



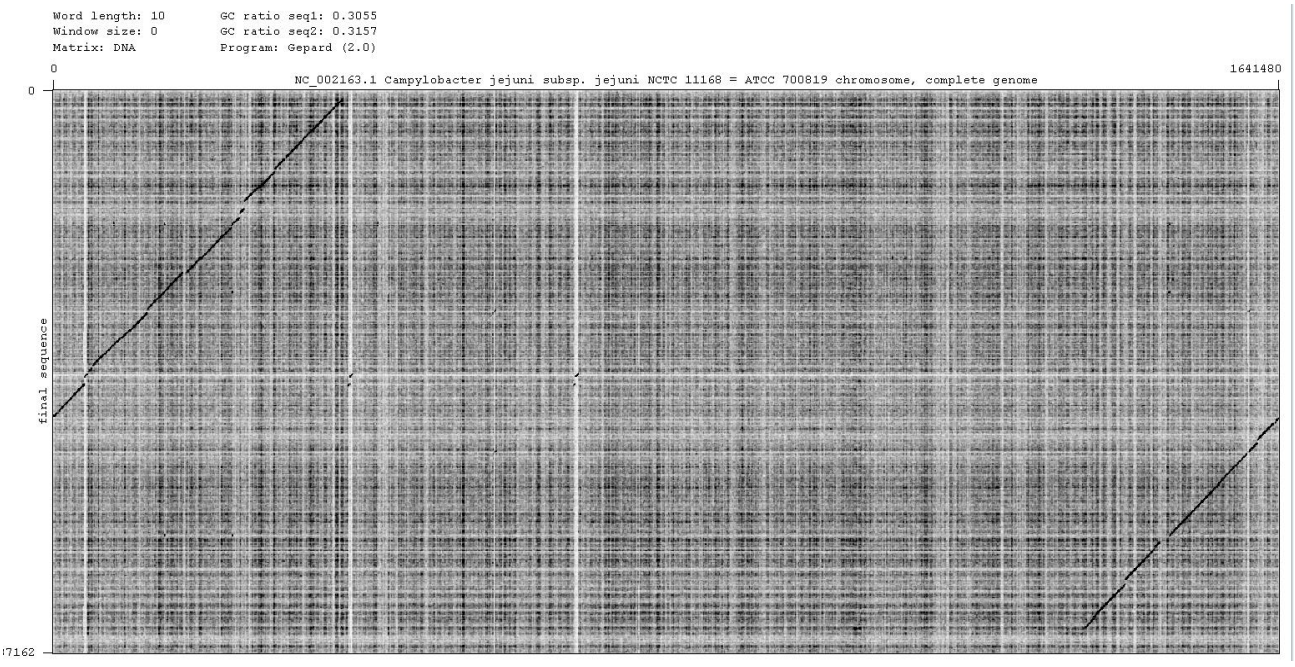
Slika 2. Usporedba generirane sekvence sa campylobacter jejuni

Nakon pronalazaka više grešaka u implementaciji dobivaju se lošiji rezultati. Za najbolji rezultat bilo je potrebno 56 minuta te maksimalno 10 GB memorije. Rezultat je prikazan na slici 3.



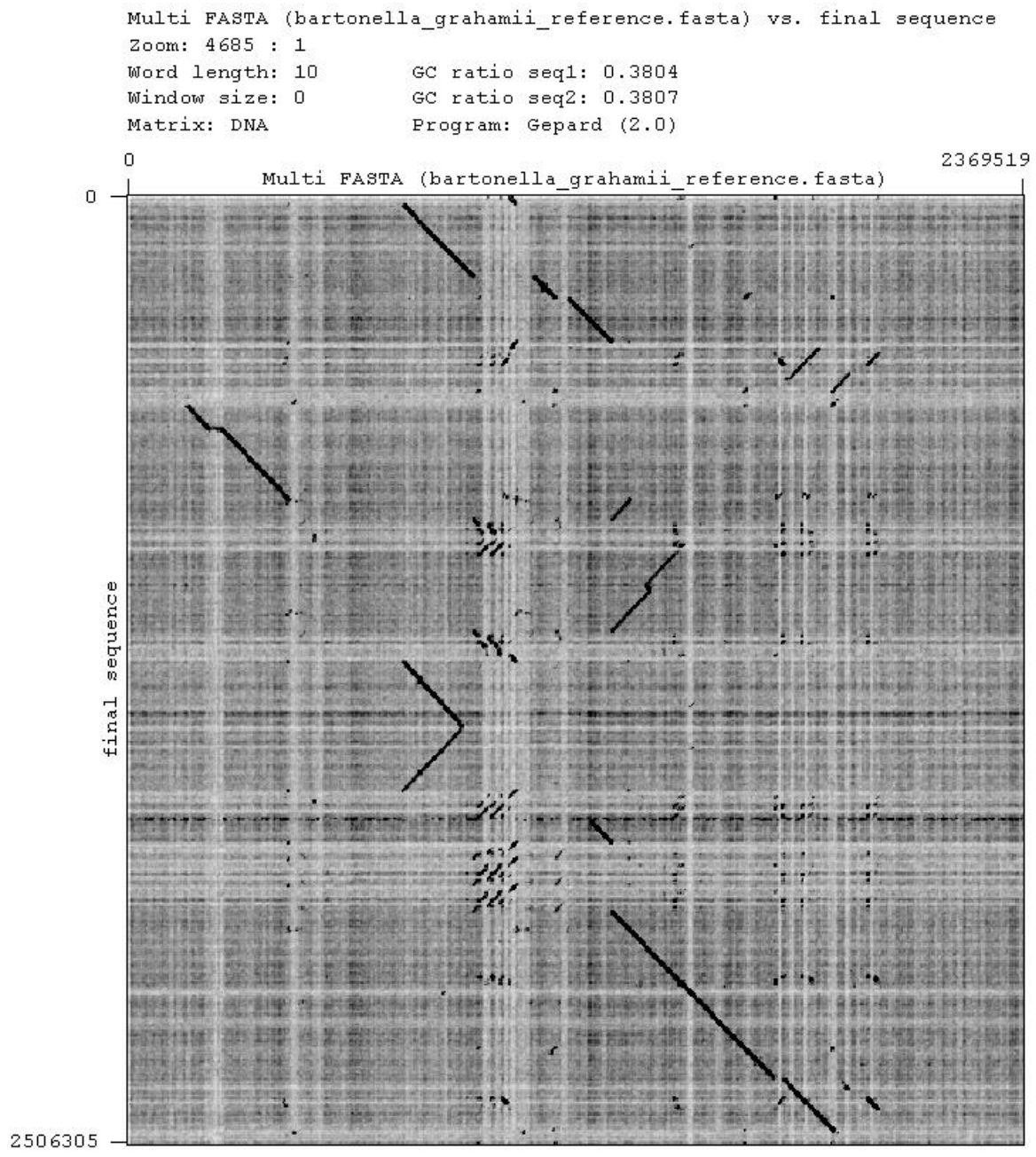
Slika 3. Usporedba generirane sekvence sa campylobacter jejuni

Dobiveni su različiti rezultati uz promjene broja generiranih puteva određenom metodom. Jedan od uspješnijih prikazan je na slici 4.



Slika 4. Usporedba generirane sekvence sa campylobacter jejuni refererencom

Na bartonela grahamii podacima proveden je samo jedan test, rezultat je prikazan na slici 5. Program se izvršava 109 minuta uz maksimalno zauzeće memorije od 12.5 GB iako su generirana maksimalno dva puta za svaki način spajanja.



Slika 5. Usporedba generirane sekvence sa bartonella grahamii referencom

Testiranje na stvarnim podacima pokazuje da postoji još prostora za implementaciju algoritma. Podešavanjem parametara algoritma, tj. promjenom broja staza koje su generirane svakim od tri načina te smanjivanjem ili povećavanjem minimalne tražene kvalitete očitanje mogu se postići različiti rezultati. Daljnjim testiranjem i optimizacijom parametara trenutna implementacija mogla bi dati bolje rezultate.

LITERATURA

Šikić, Domazet-Lošo, Skripta iz Bioinformatike

Huilong Du, Chengzhi Liang; Assembly of chromosome-scale contigs by efficiently resolving repetitive sequences with long reads, bioRxiv 345983; doi: https://doi.org/10.1101/345983