SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

Bioinformatika 1

Projekt

Luka Miličević, Antonio Mišić

Zagreb, lipanj 2024.

Sadržaj

Uvod

Bioinformatika 1 kolegij je na Fakultetu elektrotehnike i računarstva Sveučilišta u Zagrebu koji za polaganje studentima nudi opciju izrade projekta za punih 100 bodova (maksimalan broj bodova na kolegiju). U sklopu navedenog projekta izradili smo program za mapiranje fragmenata (manjih dijelova) genoma na referentni genom, te smo se kroz proces izrade projekta upoznali s osnovnim pojmovima i algoritmima korištenim u bioinformatici, python bibliotekama za bioinformatiku i višedretveno izvođenje te baratanje s podacima, te izradom potpunog programa koji funkcionira na stvarnim podatkovnim skupovima.

Mapiranje genoma je proces kojim se pronalazi položaj fragmenta na referentnom genomu. U našoj implementaciji koristili smo algoritme za pronalazak minimizatora fragmenta i reference s kojima pronalazimo najdužu rastući podniz (LIS – Longest increasing Subsequence) za svaki par fragment-referenca. Nakon pronalaska LIS-a (ako se pronađe prikladan kandidat), koriste se bioinformatički algoritmi poravnanja kako bi se pronašlo najbolje poravnanje između fragmenta i reference. Poravnanje između dvije sekvence opisuje kako iz jedne dobiti drugu i koristi se za pronalazak sličnih regija između sekvenci gena. Koristeći ovaj proces možemo znatno efikasnije mapirati fragmente referentnog genoma na njega.

Projekt su izradili Luka Miličević i Antonio Mišić pod mentorstvom Krešimira Križanovića.

1. Opis dijelova projekta

Projekt se sastoji od 3 glavna dijela: pronalazak minimizatora, pronalazak najdužeg rastućeg podniza minimizatora te poravnanja na području najdužeg rastućeg podniza. Ta 3 dijela se koriste redom kako bi se postiglo mapiranje fragmenta na referentni genom.

1.1. Minimizatori

Minimizatori za bilo koju DNA/RNA sekvencu su specifični podnizovi definirane duljine k (k-mers). Pošto je pronalazak svih mogućih minimizatora vrlo skup koristimo podskup minimizatora objašnjen u sljedećem radu:

https://academic.oup.com/bioinformatics/article/20/18/3363/202143?login=false

Pronalaze se minimizatori za referencu, koji čine indeks minimizatora reference. Taj indeks se koristi u ostatku projekta. Nakon toga se za svaki ulazni fragment također računaju svi minimizatori. Bitno je napomenuti da se pronalaze minimizatori i za sam fragment i komplement odvojeno, s obzirom na to da ne znamo s kojeg lanca je fragment sekvenciran. Bitni parametri za pronalazak minimizatora su k, w i f, koji određuju (redom) duljinu minimizatora, širinu prozora nad kojima se traže i udio najčešćih minimizatora koji se ne koriste u indeksu. U našem slučaju kao unaprijed zadane vrijednosti korišteni su k=15, w=5 i f=0.001.

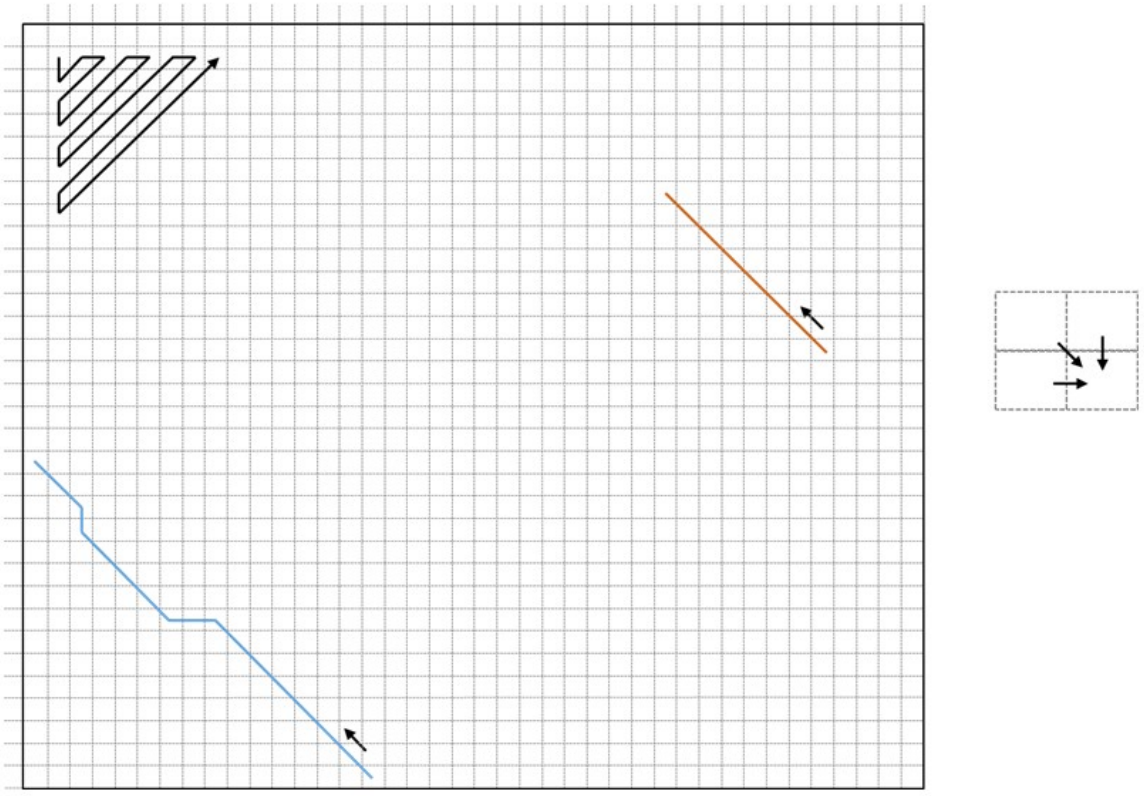
1.2. Najduži rastući podniz

Nakon pronalaska minimizatora na fragmentu i referenci, pronalaze se svi parovi međusobno istih minimizatora na fragmentu i referenci. Ti parovi se onda koriste u algoritmu za pronalazak najdužeg rastućeg podniza (LIS) vremenske složenosti O(nlogn). LIS koji se pronađe predstavlja najbolji kandidat za područje koje bi trebali promatrati pri poravnanju fragmenta na sekvencu. Ne pronalazi se samo jedan LIS, već jedan koristeći minimizatore reverznog komplementa fragmenta, a drugi koristeći minimizatore samog fragmenta. Na kraju za poravnanje koristimo onaj LIS koji je obuhvatio veći dio fragmenta.

Mogu se dogoditi slučajevi u kojima je razlika između dva susjedna para u LIS-u veća od neke granice (u našem slučaju 500 nukleobaza). U tim slučajevima prekidamo traženje LIS-a jer zaključujemo da neće biti moguće lako mapirati taj fragment na referencu. Također se može dogoditi da se jedan minimizator na specifičnoj poziciji koristi u dva para u istom LIS-u. U tim slučajevima ne možemo uzeti taj minimizator već prijeći na sljedeći.

1.3. Algoritmi poravnanja

Pronađeni LIS, uz dodatak s početne i stražnje strane, određuje područje koje će se poravnavati bioinformatičkim algoritmima poravnanja. Poravnanje sekvenci opisuje kako dobiti jednu sekvencu iz druge i koristi se za pronalaženje sličnih regija između DNA, RNA ili proteina. Algoritmi za poravnanje koriste matricu (n + 1) * (m + 1), gdje n i m predstavljaju duljine sekvenci koje se poravnavaju. Postoje algoritmi za globalno, lokalno i poluglobalno poravnanje. U našem slučaju koristimo algoritam Needleman-Wunsch za globalno poravnanje, s obzirom na to da očekujemo da su LIS-ovi fragmenta i reference slični od početka do kraja.



1.4. Mapiranje

Kroz opisani postupak, koji se odvija paralelno za fragmente ulaza, dobili smo izlaz u PAF formatu. PAF je tekstualni format koji se koristi za prikazivanje rezultata poravnanja sekvenci, posebno za prikazivanje parnih poravnanja između dugih sekvenci DNA ili RNA. Dizajniran je da bude jednostavan i kompaktan, što ga čini pogodnim za brzo zapisivanje i čitanje rezultata poravnanja. Zadnji red PAF ispisa za ovaj projekt je CIGAR string. CIGAR je niz koji se koristi za kodiranje operacija poravnanja između fragmenta i referentne sekvence. Svaka operacija označava specifičan način kako je fragment poravnan na ciljanu sekvencu.

2. Implementacija i testiranje

Projekt je implementiran u programskom jeziku python. Korištene su biblioteke Biopython (bioinformatički alati), NumPy (efikasnije baratanje s podacima, osobito nizovima) i Matplotlib (prikaz pomoćnih rezultata u obliku grafa). Za verzioniranje projekta koristio se Git i GitHub. Za komunikaciju između članova tima koristio se Microsoft Teams.

Za testiranje korišten je CLI alat za poravnanje sekvenci Minimap2, koji se često koristi u industriji. Korištene su postavke -map-ont (korištenje indeksa minimizatora) i -c (CIGAR), te Oxford preset. Podaci koje smo koristili su Oxford Nanopore Technologies podaci dobijenim sekvenciranjem genoma Escherichia coli K-12 substr. MG1655. Podaci su dostupni na Loman Labs (MAP-006-1 i/ili MAP-006-2 FASTA datoteke - https://lab.loman.net/2015/09/24/first-sqk-map-006-experiment/), dok je referentni genom dostupan na NCBI (https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/005/845/GCF_000005845.2_ASM584v2/GCF_000005845.2_ASM584v2_genomic.fna.gz).

3. Pristup projektu

Projekt je dostupan na GitHub-u: https://github.com/lLukal/BIO1

Upute za korištenje projekta opisane su README.md datoteci, vidljivoj kao opis projekta na GitHub-u.