Projekti za 100 bodova na predmetu Bioinformatika, 2019./2020.

- broj članova tima: 2-3
- implementacija: C/C++
- opis algoritma, implementacije i testiranje
- dozvoljeno je korištenje pomoćnih knjižnica u zadatcima gdje je tako navedeno, a za ostale situacije možete se dogovoriti s nastavnikom koji je zadao temu
- za svaki dan zakašnjenja umanjuje se konačan broj bodova za 3 boda

Bodovanje zadataka (1) – (7)

	Broj bodova
Program - testiranje • ako program ne radi ispravno na testnim podatcima umanjuje se konačan broj bodova za 10 bodova • prepravke napraviti u roku 2 dana Performanse programa (vrijeme izvođenja i utrošak memorije) • ako se program uspoređuje sa studentskim rješenjem od prošle godine, implementacija mora biti unutar 10% vremena izvođenja i utroška memorije u odnosu na navedenu referencu za isti skup podataka (npr. ako referentni program koristi 1 GB memorije za neki skup podataka, onda Vaša implementacija treba koristiti najviše 1,1 GB memorije) o oduzima se 10 bodova, ako je odstupanje do 20% o oduzima se 15 bodova, ako je odstupanje veće od 20% • ako se program uspoređuje s objavljenim rješenjem, implementacija mora biti unutar 70% vremena izvođenja i utroška memorije u odnosu na navedenu referencu (npr. ako referentni program koristi 1 GB memorije za neki skup podataka, onda Vaša implementacija treba koristiti najviše 1,7 GB memorije) o oduzima se 10 bodova, ako je odstupanje do 100% o oduzima se 15 bodova, ako je odstupanje veće od 100%	60
Testiranje na sintetskim podatcima 10^2-10^6 znakova • svi rezultati moraju biti u dokumentaciji – prikazani u tablici i/ili grafu	10
Testiranje na stvarnim podatcima (*Escherichia coli* ili po dogovoru ovisno o zadatku) • svi rezultati moraju biti u dokumentaciji – prikazani u tablici i/ili grafu	10
Dokumentacija • opis algoritma i vizualizacija na jednostavnom primjeru (5 bodova) • obvezno navesti popis literature i navesti izvore unutar teksta (5 bodova)	15

• za svaki algoritam napraviti analizu točnosti, vremena izvođenja i utroška memorije za različite testne slučaje (5 bodova)	
Prezentacija • oduzimaju se bodovi, ako je prezentacija dulja od predviđenoga vremena (1 bod za svaku minutu prekoračenja)	5

(1) Space-efficient and exact de Bruijn graph representation based on a Bloom filter (Chikhi and Rizk. 2013) (MDL)

- https://almob.biomedcentral.com/articles/10.1186/1748-7188-8-22

U izradi programa:

- dozvoljeno koristiti program/dijelove programa Jellyfish za brojanje k-mera
- dozvoljeno koristiti neku gotovu implementaciju Bloomovog filtera
- testirati za E. coli skup očitanja
- usporediti s originalnom implementacijom (http://minia.genouest.org/)

(2) Improving Bloom Filter Performance on Sequence Data Using k-mer Bloom Filters (Pellow et al 2016) (MDL)

- https://link.springer.com/chapter/10.1007/978-3-319-31957-5_10
- dozvoljeno koristiti neku gotovu implementaciju Bloomovog filtera
- usporediti s originalnom implementacijom: https://github.com/Kingsford-Group/kbf

(3) Dynamic Cuckoo Filter (Chen et al. 2017) (MDL)

- Chen et al. 2017. The dynamic cuckoo filter; https://ieeexplore.ieee.org/abstract/document/8117563
- Fan et al. 2013. Cuckoo Filter: Better Than Bloom; https://www.cs.cmu.edu/~binfan/papers/login_cuckoofilter.pdf
- Fan et al. 2014. Cuckoo Filter: Practically Better Than Bloom; http://www.cs.cmu.edu/%7Ebinfan/papers/conext14_cuckoofilter.pdf
- tražiti slučajne podnizove (k-mere uz različite k, npr. k = 10, 20, 50, 100, 200) u E. coli genomu
- napraviti vlastiti CF i DCF i usporediti ih međusobno te s originalnom implementacijom: https://github.com/efficient/cuckoofilter

(4) FM-stablo (Cheng et al. 2017) (MDL)

- Cheng et al. 2017. FMtree: a fast locating algorithm of FM-indexes for genomic data; https://academic.oup.com/bioinformatics/article/34/3/416/4160683
- usporediti s originalnom implementacijom: https://github.com/chhylp123/FMtree

(5) HiRGC (Liu et al. 2017) (MDL)

- Liu et al. 2017. High-speed and high-ratio referential genome compression; https://academic.oup.com/bioinformatics/article/33/21/3364/3885699
- usporediti s originalnom implementacijom: https://github.com/yuansliu/HiRGC

(6) Određivanje LCP polja korištenjem modificiranog algoritma SA-IS (Fischer, 2011) (MDL)

- Inducing the LCParray (Fischer, 2011) (http://arxiv.org/pdf/1101.3448.pdf)
- originalna implementacija: http://algo2.iti.kit.edu/english/1828.php
- novija implementacija: https://github.com/kurpicz/sais-lite-lcp
- implementirati SA-IS algoritam (Ge Nong et al. 2011) umjesto sais-lite
- usporediti s originalnom i novijom implementacijom

(7) Određivanje poravnanja parova sljedova korištenjem HMM (MDL)

- Yoon, 2009. Hidden Markov Models and their Applications in Biological Sequence Analysis; https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2766791/
- Pairwise alignment using HMMs: http://www.stat.purdue.edu/~junxie/topic4.pdf
- matricu prijelaza i emisije definirati pomoću npr. Needleman-Wunschovog algoritma

(8) Poboljšanje djelomično sastavljenog genoma dugim očitanjima (KK – kresimir.krizanovic@fer.hr)

Cilj: Zadani genom već je djelomično sastavljen nekim od postojećih alata. Međutim, postupak sastavljanja nije bio sasvim uspješan te je rezultat fragmentiran - skup sastavljenih sekvenci (contig-a) za koje ne znamo kako se međusobno povezuju u cijeli genom. Potrebno je implementirati postupak *scaffolding*-a, koji će iskoristiti duga očitanja da bih povezao pojedine contige u dulje sekvence. Pri tome je potrebno implementirati algoritam opisan u radu:

- 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.

Ulazni podaci:

- Skup već sastavljenih contig-a
- Skup očitanja
- Preklapanja između contig-a i očitanja u PAF formatu
- Međusobna preklapanja očitanja u PAF formatu

Izlazni podaci:

- Poboljšani skup sastavljenih contiga u FASTA formatu

Skupovi očitanja i već sastavljenih contiga bit će pripremljeni kao testni podaci. Dok će se preklapanja dobiti pomoći alata Minimap2 (https://github.com/lh3/minimap2), koristeći opciju:

```
./minimap2 -x ava-pb contigs.fa reads.fa > overlaps.paf
```

Za preuzimanje sintetskih i stvarnih testnih podataka potrebno se javiti na kresimir.krizanovic@fer.hr.

Evaluacija:

- Testiranje na sintetskim podacima i usporedba s referencom pomoću alata Gepard, dostupan na http://cube.univie.ac.at/gepard.
- Testiranje na stvarnim podacima, usporedba s referencom pomoću alata Gepard, te usporedba s referentnim rezultatima gledajući mjere:
    - Broj contig-a
    - Duljina najduljeg contig-a

Bodovanje:

	Broj bodova

Program
• ako program ne radi ispravno na testnim podatcima prilikom demonstracije umanjuje se konačan broj bodova za 10 bodova (prepravke napraviti u roku od 2 dana)
• vremensko ograničenje od 60min na 1 dretvi, u protivnom se oduzima 5 bodova
• memorijsko ograničenje od 16 GB RAM-a, u protivnom se oduzima 5 bodova
• točnost rezultata:
  ○ ako program ne radi ispravno na sintetskim podacima oduzima se 40 bodova
  ○ za odstupanje veće od 25% od referentnih rezultata oduzima se 10 bodova
  ○ za za odstupanje veće od 50% od referentnih rezultata oduzima se 25 bodova
	80

Dokumentacija • opis algoritma i vizualizacija na jednostavnom primjeru • obavezno navesti popis literature te navesti izvore unutar teksta • napraviti usporedbu točnosti, vremena izvođenja i utroška memorije vaše implementacije i izvorne	15
Prezentacija • oduzimaju se bodovi, ako je prezentacija dulja od predviđenoga vremena	5

Preporučena literatura:

1. Skripta iz bioinformatike
2. PAF format: https://github.com/lh3/miniasm/blob/master/PAF.md
3. Scaffolding algoritam HERA:
   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.
4. Alat za DOT plot Gepard:
   Jan Krumsiek, Roland Arnold, Thomas Rattei; Gepard: a rapid and sensitive tool for creating dotplots on genome scale, Bioinformatics, Volume 23, Issue 8, 15 April 2007, Pages 1026–1028, https://doi.org/10.1093/bioinformatics/btm039.
5. Alat za računanje preklapanja Minimap2 https://github.com/lh3/minimap2

(9) Pronalaženje varijanti gena iz podataka dobivenih sekvenciranjem (KK – kresimir.krizanovic@fer.hr)

Cilj: Sekvenciran je uzorak koji sadrži nekoliko varijanti istog gena. Potrebno je primijeniti tehnike grupiranja (engl. *clustering*) na očitanja da bi se otkrile sve varijante danog gena koje su prisutne u uzorku. Očitanja je potrebno grupirati na temelju međusobne udaljenosti. Pri tome je za izračun udaljenosti potrebno implementirati algoritme globalnog, poluglobalnog i lokalnog poravnanja. Za računanje centroida pojedine grupe (engl. *cluster*) dopušteno je koristiti postojeću biblioteku SPOA (https://github.com/rvaser/spoa)

Ulazni podaci:

- Skup očitanja

Izlazni podaci:

- Skup otkrivenih varijanti gena u FASTA formatu
- Popis očitanja koja pripadaju kojoj varijanti/grupi/clusteru

Skupovi očitanja bit će pripremljeni kao ulazni podaci, kao i nekoliko uzoraka sa poznatim varijantama.

Za preuzimanje testnih podataka te za detaljnije upute o projektu potrebno se javiti na kresimir.krizanovic@fer.hr.

Evaluacija:

- Testiranje na osnovnim podacima za koje su rezultati poznati.
- Testiranje na podacima za koje stvarni podaci nisu poznati te usporedba s drugim rješenjima.

Bodovanje:

	Broj bodova

Program
• ako program ne radi ispravno na osnovnim podacima prilikom demonstracije umanjuje se konačan broj bodova za 10 bodova (prepravke napraviti u roku od 2 dana)
• vremensko ograničenje od 60min na 1 dretvi, u protivnom se oduzima 5 bodova
• memorijsko ograničenje od 16 GB RAM-a, u protivnom se oduzima 5 bodova
• točnost rezultata:
  ○ ako program ne radi ispravno na osnovnim podacima oduzima se 40 bodova
  ○ za odstupanje veće od 25% od referentnih rezultata oduzima se 10 bodova
  ○ za odstupanje veće od 50% od referentnih rezultata oduzima se 25 bodova
	80

Dokumentacija • opis algoritma i vizualizacija na jednostavnom primjeru • obavezno navesti popis literature te navesti izvore unutar teksta • napraviti ocjenu točnosti, vremena izvođenja i utroška memorije	15
Prezentacija • oduzimaju se bodovi, ako je prezentacija dulja od predviđenoga vremena	5

Preporučena literatura:

1. Skripta iz bioinformatike
2. Biblioteka SPOA (https://github.com/rvaser/spoa)
3. Završni rad Sanje Kosier (mailom nakon prvih konzultacija)

(10) Navarrov algoritam za približno uspoređivanje teksta (KK – kresimir.krizanovic@fer.hr)

Zadatak: Implementirati Navarrov algoritam opisan u radu (Improved approximate pattern matching on hypertext) https://www.sciencedirect.com/science/article/pii/S0304397599003333.

Evaluacija:

Usporediti s bit parallel sequence-to-graph alignment algoritmom (opisanom u radu https://academic.oup.com/bioinformatics/article/35/19/3599/5372677 . Algoritam usporediti na 4 vrste graf topologija koje su opisane u poglavlju 6.2 Graph topology experiment. Skripte za generiranje testnih podataka dostupne su na https://github.com/maickrau/GraphAligner/tree/PaperExperiments/WabiExperimentSnake.

Realizirani algoritam treba biti maksimalno do 50x sporiji od bit parallel sequence-to-graph alignment algoritma.

Bodovanje:

	Broj bodova

Program
• ako program ne radi ispravno na linearnoj topologiji prilikom demonstracije umanjuje se konačan broj bodova za 10 bodova (prepravke napraviti u roku od 2 dana)
• vremensko ograničenje od 60min na 1 dretvi, u protivnom se oduzima 5 bodova
• memorijsko ograničenje od 16 GB RAM-a, u protivnom se oduzima 5 bodova
• točnost rezultata:
  ○ za odstupanje veće od 10% od referentne implementacije oduzima se 10 bodova
  ○ za odstupanje veće od 25% od referentne implementacije oduzima se 25 bodova
	80

Dokumentacija • opis algoritma i vizualizacija na jednostavnom primjeru • obavezno navesti popis literature te navesti izvore unutar teksta • napraviti usporedbu točnosti, vremena izvođenja i utroška memorije vaše implementacije i izvorne	15
Prezentacija • oduzimaju se bodovi, ako je prezentacija dulja od predviđenoga vremena	5

Preporučena literatura:

4. Rad Improved approximate pattern matching on hypertext (https://www.sciencedirect.com/science/article/pii/S0304397599003333)
5. Rad Bit-parallel sequence-to-graph alignment (https://academic.oup.com/bioinformatics/article/35/19/3599/5372677)

(11) Pronalazak mutacija pomoću treće generacije sekvenciranja (RV – robert.vaser@fer.hr)

Ulaz: referentni genom i skup očitanja dobiven sekvenciranjem mutiranog genoma. Obje datoteke su u FASTA formatu.

Cilj: Za dani ulaz, pronaći razlike između referentnog genoma i sekvenciranog mutiranog genoma. Mutacije uključuju jednostruke substitucije, umetanja i brisanja. Očitanja je potrebno mapirati na danu referencu pomoću k-mer indeksa, poravnati ih te iz gomile poravnanja razlučiti mutacije. Zabranjeno je koristiti gotove implementacije.

Izlaz: Lista mutacija u odnosu na referencu (gdje je prvi nukleotid na poziciji 0), u CSV formatu kao što je prikazano u tablici ispod.

Mutacija	**Linija u CSV datoteci**		
Substitucija	X	Pozicija u referenci na kojoj se dogodila substitucija	Zamjenska nukleotidna baza
Umetanje	I	Pozicija u referenci prije koje se dogodilo umetanje	Umetnuta nukleotidna baza
Brisanje	D	Pozicija u referenci na kojoj se dogodilo brisanje	-

Evaluacija: usporediti rezultate s referentnom implementacijom pomoću Jaccardovog indeksa. Za testne skupove, rezultate referentne implementacije i skriptu za evaluaciju potrebno se javiti nastavniku.

Bodovanje:

	Broj bodova

Program
• ako program ne radi ispravno na testnim podatcima prilikom demonstracije umanjuje se konačan broj bodova za 10 bodova (prepravke napraviti u roku od 2 dana)
• vremensko ograničenje od 30min na 1 dretvi, u protivnom se oduzima 5 bodova
• memorijsko ograničenje od 16 GB RAM-a, u protivnom se oduzima 5 bodova
• točnost rezultata:
  ○ za odstupanje veće od 50% od referentne implementacije oduzima se 10 bodova
  ○ za za odstupanje veće od 75% od referentne implementacije oduzima se 25 bodova
	80

Dokumentacija • opis algoritma i vizualizacija na jednostavnom primjeru • obavezno navesti popis literature te navesti izvore unutar teksta • napraviti usporedbu točnosti, vremena izvođenja i utroška memorije vaše implementacije i izvorne	15
Prezentacija • oduzimaju se bodovi, ako je prezentacija dulja od predviđenoga vremena	5

Preporučena literatura:

6. Algoritmi preklapanja - skripta iz bioinformatike
7. Minimizers - https://academic.oup.com/bioinformatics/article/20/18/3363/202143