SVEUČILIŠTE U ZAGREBU

FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

SEMINAR

Metode usporedbe bioloških sljedova bez poravnanja

Ivan Furač

Mentor: *izv. prof. dr. sc. Mirjana Domazet-Lošo*

Zagreb, siječanj 2022.

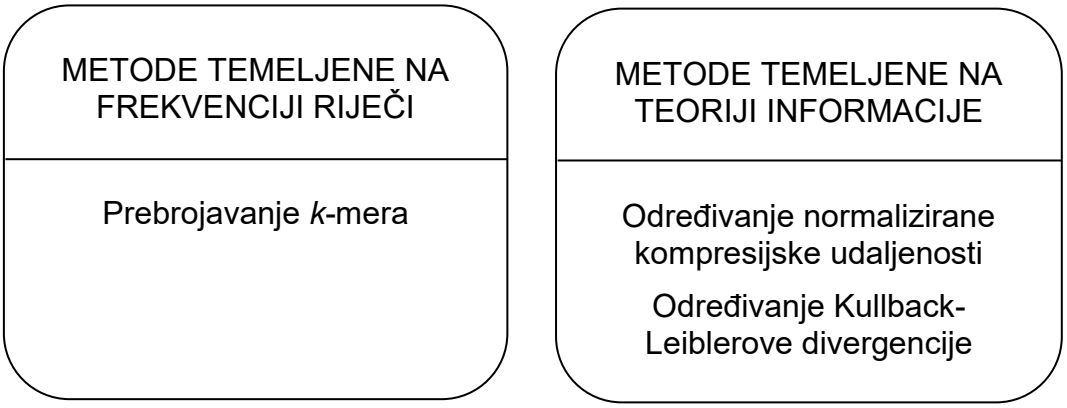
Sadržaj

1. Uvod

Usporedba bioloških sekvenci, odnosno sljedova nukleotida ili aminokiselina u svrhu otkrivanja sličnosti pripada temeljnim problemima koje rješava područje bioinformatike. Otkrivanje sličnosti između dva slijeda upućuje na njihovu homologiju, odnosno zajedničko podrijetlo. Homologni sljedovi imaju sličnu strukturu, a često obavljaju i sličnu funkciju [1]. Razvijene su brojne metode uspoređivanja sljedova koje se temelje na poravnanju, odnosno traženju dijelova koji se poklapaju uzimajući u obzir brisanje, umetanje i zamjenu elemenata slijeda. Primjeri alata koji koriste poravnanje su BLAST, FASTA, ClustalW i mnogi drugi [1]. Međutim, metode usporedbe bioloških sljedova koje koriste isključivo poravnanje imaju brojne nedostatke [2]. Takve metode nužno pretpostavljaju kolinearnost homologni sljedova, odnosno očuvanje linearnog redoslijeda pojedinih gena ili aminokiselina što u stvarnosti ne mora vrijediti zbog raznih evolucijskih promjena. Metode koje koriste poravnanje u pravilu ne prepoznaju homologne sljedove koji imaju manju sličnost (tipično sličnost manju od 20% u slučaju proteinskih sljedova i sličnost manju od 60% u slučaju sljedova nukleotida). Jedan od najvećih problema su velika vremenska i memorijska složenost algoritama poravnanja, kao i činjenica da je izračun poravnanja višestrukih sljedova NP-težak problem, zbog čega mnogi algoritmi koriste heuristike koje ne garantiraju pronalazak optimalnog rješenja. Problem metoda koje se temelje na poravnanju sljedova je i to što rezultati ovise o određenim početnim parametrima koji se najčešće odabiru proizvoljno. Pokazalo se da male promjene u vrijednostima početnih parametara mogu imati velik utjecaj na rezultate poravnanja. Zbog navedenih nedostataka u posljednje se vrijeme sve više istražuju metode usporedbe sljedova koje ne koriste poravnanje (engl. *alignment-free methods*). Postoji više prednosti takvih metoda u odnosu na prethodno spomenute metode koje koriste poravnanje. Jedna od glavnih prednosti je upravo manja složenost koja je u većini metoda linearna, odnosno ovisi isključivo o duljinama sljedova koji se uspoređuju [2]. U nastavku rada će biti opisan način na koji *alignment-free* metode uspoređuju sljedove te će detaljno biti opisane tri metode koje su programski implementirane. Na kraju će biti prikazani rezultati koji su dobiveni na različitim skupovima testnih podataka koji uključuju sljedove nukleotida i aminokiselina.

2. Pregled metoda

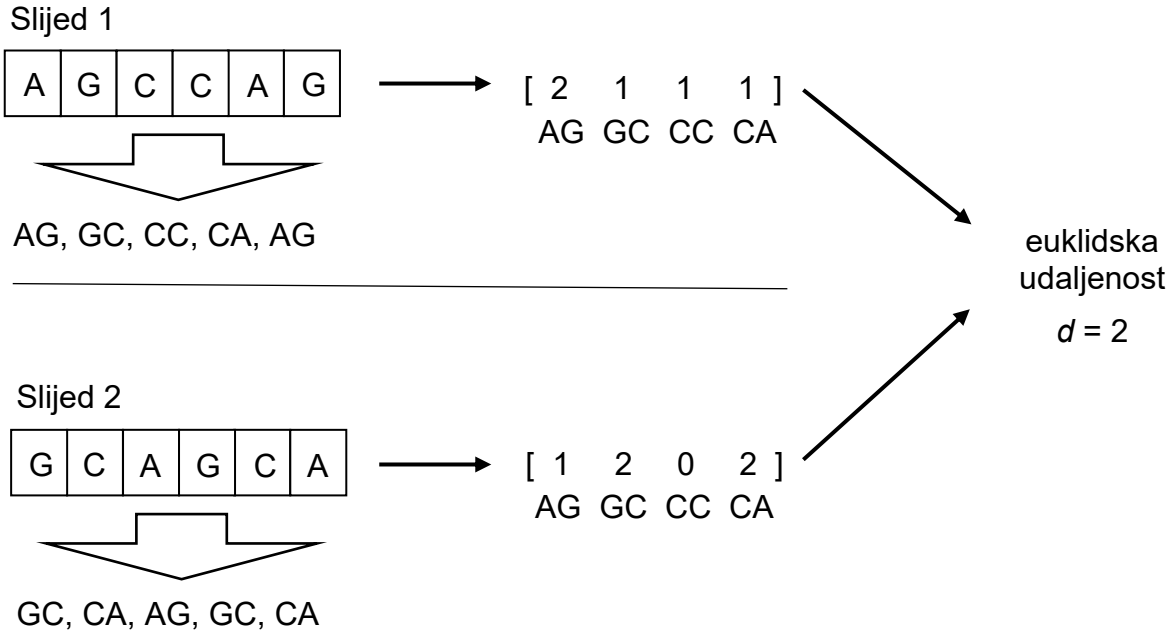
Metode koje ne koriste poravnanje, odnosno skraćeno *alignment-free* metode, najprije kvantificiraju sljedove koje uspoređuju, pretvaraju ih u brojčane vrijednosti nad kojima se primjenjuju postupci vjerojatnosti, statistike i linearne algebre kako bi se dobili konačni rezultati sličnosti sljedova [3]. Kao što ime kaže, ni u jednom trenutku se ne računa poravnanje sljedova. S obzirom na to da uglavnom imaju linearnu složenost, *alignment-free* metode pogodne su za usporedbu dugačkih sljedova, na primjer cjelokupnih genoma organizama [3]. Trenutno je razvijeno preko 100 *alignment-free* metoda i postoji više načina na koje se metode mogu podijeliti u kategorije. Najšira podjela je u dvije kategorije: metode temeljene na frekvenciji pojavljivanja riječi određene duljine unutar sljedova i metode temeljene na teoriji informacije koje uspoređuju količinu informacije sadržanu unutar sljedova [2]. U sklopu ovog rada odabrane su tri metode koje su implementirane u programskom jeziku Python: prebrojavanje k-mera koje spada u prvu kategoriju metoda temeljenih na frekvenciji pojavljivanja riječi te dvije metode temeljene na teoriji informacije, jedna koja uspoređuje Lempel-Ziv složenost sljedova te jedna koja uspoređuje količinu entropije sadržanu unutar sljedova. Zajedničko ovim metodama je to da zapravo računaju udaljenost sljedova, mjeru koja govori koliko su sljedovi različiti. Na slici 2.1 prikazane su korištene *alignment-free* metode.



Slika 2.1 Prikaz korištenih metoda

2.1 Prebrojavanje *k*-mera

Pojam *k*-mer označava podniz duljine *k* koji se pojavljuje unutar nekog većeg niza, u ovom slučaju slijeda nukleotida ili aminokiselina. Ova metoda temelji se na pretpostavci da se u sličnim sljedovima očekivano pojavljuje sličan broj istih podnizova [2]. U sljedovima koji se uspoređuju potrebno je prebrojati sve podnizove zadane duljine k. Na taj način sljedovi se mogu prikazati kao vektori čija je dimenzija broj pronađenih različitih *k*-mera, a pojedine vrijednosti predstavljaju frekvenciju pojavljivanja tih *k*-mera. Udaljenost između vektora moguće je izračunati koristeći različite matematičke funkcije, na primjer euklidsku udaljenost, Manhattan udaljenost ili Mahalanobisovu udaljenost pri čemu se dobiva konačna mjera udaljenosti dva slijeda. U ovom slučaju za implementaciju metode korištena je euklidska udaljenost. Na slici 2.2 prikazani su koraci prilikom izračuna udaljenosti dva kratka slijeda nukleotida u slučaju *k* = 2, odnosno prebrojavanja *k*-mera duljine 2.

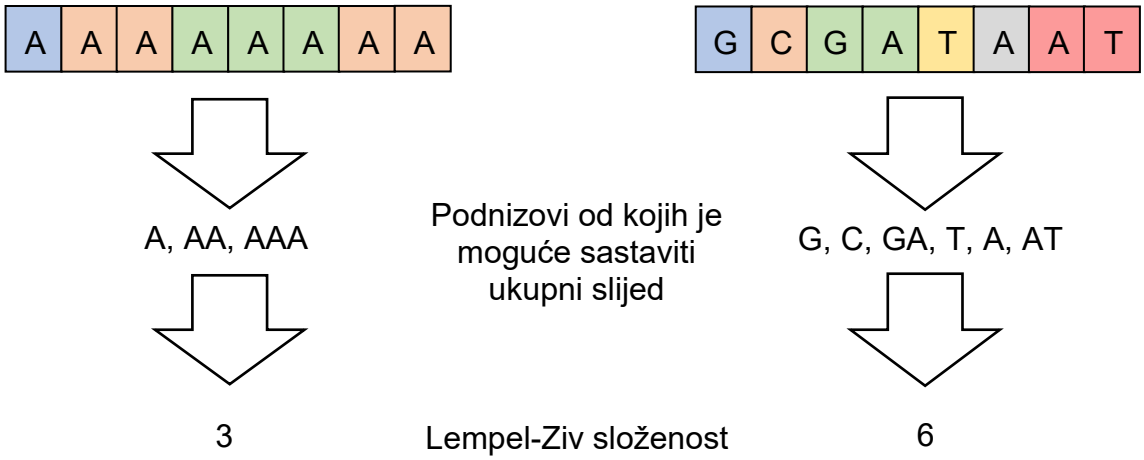


Slika 2.2 Prikaz koraka metode prebrojavanja k-mera

Jedini parametar ove metode je upravo duljina *k*-mera koji se prebrojavaju. Pokazano je da se u slučaju proteinskih sljedova najstabilnija i optimalna rješenja dobivaju odabirom duljine *k*-mera između 2 i 4 dok se u slučaju nukleotidnih sljedova može odabrati i veća duljina [2]. Prilikom analize uspješnosti metoda na testnim skupovima sljedova odabrane su različite vrijednosti parametra *k*, međutim najbolji rezultati dobiveni su korištenjem $k = 4$. Korištenjem većih vrijednosti udaljenosti između sljedova postaju sve veće i međusobno sve sličnije što se u teoriji zove prokletstvo dimenzionalnosti.

2.2 Određivanje normalizirane kompresijske udaljenosti

Ova se metoda temelji na usporedbi Lempel-Ziv složenosti dvaju sljedova. Složenost se općenito koristi u kompresiji podataka kao mjera koja pokazuje koliko je moguće komprimirati niz znakova; složenije nizove teže je komprimirati. Lempel-Ziv složenost računa se kao mjera koja ovisi o broju različitih podnizova unutar nekog većeg niza koji se pojavljuju gledajući od početka do kraja, kao i o frekvenciji pojavljivanja tih podnizova [2, 4]. Ova se složenost često koristi u analizi vremenski diskretnih signala, međutim ima i brojne druge primjene [4]. Na slici 2.3 prikazan je primjer računanja Lempel-Ziv složenosti za dva kratka slijeda nukleotida, jedan koji očekivano ima malu složenost i jedan koji očekivano ima veću složenost.



Slika 2.3 Izračun Lempel-Ziv složenosti

Niz znakova analizira se od početka prema kraju, čita se znak po znak niza i promatra se pročitani podniz. U slučaju da takav podniz do tada nije pronađen, on se dodaje u rječnik i čita se novi podniz počevši od prvog sljedećeg znaka. U slučaju da pročitani podniz već postoji u rječniku, čita se sljedeći znak niza, dodaje se na kraj pročitanog podniza te se postupak ponavlja. Nakon što su pročitani svi znakovi niza, Lempel-Ziv složenost određuje se kao broj podnizova sadržanih u rječniku. Ti su podnizovi dovoljni za opis cjelokupnog niza znakova.

Nakon što su izračunate Lempel-Ziv složenosti pojedinih sljedova koji se uspoređuju, potrebno je na neki način izračunati udaljenost. U tu svrhu, jedan slijed (slijed *x*) konkatenira se na kraj drugog slijeda (slijed *y)* te se računa i Lempel-Ziv složenost novonastalog slijeda (slijed *xy*). Pokazalo se da, ako su sljedovi *x* i *y* koji se žele usporediti međusobno vrlo slični, onda će složenost od *xy* biti vrlo bliska složenosti od *x* ili složenosti od *y*, a ako su sljedovi različiti složenost od *xy* će težiti zbroju pojedinih složenosti [2]. Kao mjera udaljenosti koristi se normalizirana kompresijska udaljenost koja se općenito može izračunati koristeći bilo koju vrstu kompresije, odnosno složenosti [5]. Formula za izračun normalizirane kompresijske udaljenosti prikazana je u nastavku (2.1). Izraz *C(a)* predstavlja rezultat komprimiranja niza *a* kompresijskim algoritmom *C*, a izraz |*C(a)*| predstavlja duljinu rezultata. U ovom slučaju |*C(a)*| zapravo predstavlja Lempel-Ziv složenost.

$$NCD(x,y) = \frac{|C(xy)| - \min\left(|C(x)|,\ |C(y)|\right)}{\max\left(|C(x)|,\ |C(y)|\right)} \tag{2.1}$$

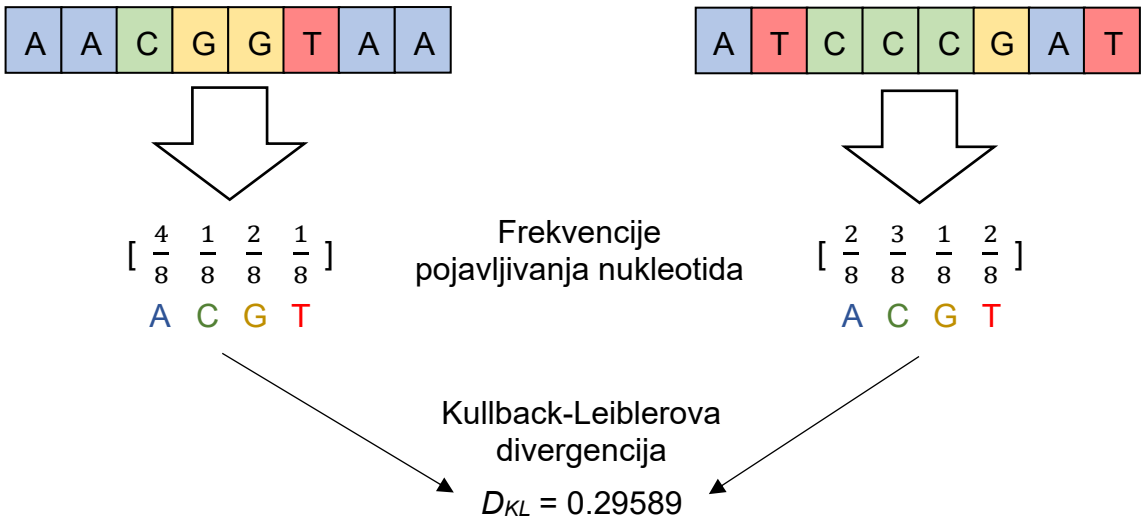
2.3 Određivanje Kullback-Leiblerove divergencije

U teoriji informacija postoji više mjera koje se mogu koristiti za usporedbu dvaju bioloških sljedova. Uz prethodno već spomenutu složenost, moguće je koristiti i entropiju, točnije Shannonovu entropiju. Entropija se može definirati kao neizvjesnost pronalaženja određenog elementa (na primjer nukleotida ili aminokiseline) unutar većeg slijeda [2]. Veća entropija nekog slijeda govori da je unutar tog slijeda sadržano više informacije, odnosno više elemenata koji imaju manju vjerojatnost pojavljivanja.

Entropija je usko povezana sa složenosti, sljedovi koji imaju manju složenost imaju i manju entropiju, kao što sljedovi s većom složenosti imaju i veću entropiju [2]. Kako bi se usporedile entropije dvaju sljedova, koristi se mjera koja se naziva Kullback-Leiblerova divergencija ili relativna entropija. Kullback-Leiblerova divergencija mjera je koja pokazuje koliko su slične dvije vjerojatnosne distribucije, točnije koliko su one udaljene [6]. U nastavku je prikazana formula za izračun spomenute divergencije (2.2). Izrazi p_i i q_i označavaju vjerojatnosti pojavljivanja elementa *i* u sljedovima koji se uspoređuju.

$$D_{KL}(p \parallel q) = \sum_i p_i \log \frac{p_i}{q_i} \tag{2.2}$$

Kako bi se izračunala Kullback-Leiblerova divergencija, potrebno je prvo izračunati distribucije *p* i *q*, odnosno vjerojatnosti pojavljivanja pojedinih elemenata u sljedovima koji se uspoređuju. To se može napraviti jednostavnim prebrojavanjem elemenata i računanjem frekvencije njihovog pojavljivanja. Na slici 2.4 prikazano je računanje Kullback-Leiblerove divergencije dva proizvoljna slijeda. Baza logaritma u (2.2) može biti proizvoljno odabrana, u primjeru na slici odabrana je baza *e*. Treba uzeti u obzir kako Kullback-Leiblerova divergencija nije simetrična mjera, odnosno vrijedi da je $D_{KL}(p \parallel q) \neq D_{KL}(q \parallel p)$, te u slučaju kada je neka od vjerojatnosti p_i ili q_i jednaka nuli može poprimiti beskonačnu vrijednost [6].



Slika 2.4 Izračun Kullback-Leiblerove divergencije

3. Rezultati

Prethodno opisane metode vrlo su jednostavne i u praksi se rjeđe koriste u takvom obliku. Međutim, njihovo vrednovanje na testnim skupom podataka može nam svejedno dati približnu informaciju koja se metoda bolje snalazi s kakvim podatcima. Metode su implementirane u programskom jeziku Python. Kôd se pokreće iz direktorija unutar kojeg se nalaze sljedovi koji se žele usporediti u FASTA formatu podataka. Za svaku od tri metode stvara se zasebna TSV datoteka unutar koje je za svaki par sljedova zapisana udaljenost izračunata tom metodom. Očekivano je da udaljenosti za sljedove za koje se pouzdano zna da su slični ili homologni budu manje, a za parove sljedova koji nisu homologni budu veće. Uspješnost metode može se mjeriti upravo razlikom u udaljenostima između sljedova za koje se zna da su homologni i onih koji to nisu što pokazuje je li metoda sposobna prepoznati homologne parove sljedova. Za testiranje implementiranih metoda korištena je web aplikacija AFproject dostupna na adresi http://afproject.org [7]. Aplikacija nudi više različitih skupova proteinskih i nukleotidnih sljedova koje je moguće preuzeti, izračunati udaljenosti parova sljedova koristeći vlastite metode te izračunate udaljenosti učitati u aplikaciju. Vrednovanje metoda obavlja se automatski nakon čega aplikacija prikazuje uspješnost u odnosu na ostale dostupne *alignment-free* metode. Za testiranje gore opisanih metoda odabrana su tri skupa koja su detaljnije opisana u tablici 3.1.

Tablica 3.1 Opis korištenih skupova sljedova

Sljedovi nukleotida	Skup sadrži ukupno 307 nekodirajućih, odnosno regulacijskih sljedova nukleotida dobivenih iz različitih tkiva dva organizma: vinske mušice i čovjeka. Skup se sastoji od ukupno 7 podskupova, svaki podskup sadrži *n* sljedova iz određenog tkiva jednog organizma te još *n* slučajno odabranih sljedova iste duljine. Očekuje se da su udaljenosti između sljedova koji pripadaju istom tkivu manje nego udaljenosti između nepovezanih sljedova.

Proteinski sljedovi male sličnosti	Skup sadrži ukupno 1066 sljedova proteina čija sličnost nije veća od 40%. Ti se proteinski sljedovi mogu grupirati prema četiri kriterija, odnosno na četiri strukturne razine: obitelj (engl. *family*), superobitelj (engl. *superfamily*), fold i razred (engl. *class*). Prilikom testiranja metode provjerava se jesu li udaljenosti između sljedova koji pripadaju istoj grupi (npr. istoj obitelji ili istom razredu) manje u odnosu na udaljenosti nepovezanih sljedova. Sličnost između proteinskih sljedova koji pripadaju istoj grupi može se smanjiti uslijed evolucijskih promjena.
Proteinski sljedovi velike sličnosti	Skup sadrži ukupno 2128 sljedova proteina čija sličnost može biti do 95%. Kao i u slučaju prethodnog skupa proteinskih sljedova, sljedovi se mogu grupirati na četiri strukturne razine. Provjerava se jesu li udaljenosti između sljedova koji pripadaju istoj grupi manje u odnosu na udaljenosti nepovezanih sljedova.

U nastavku su u tablicama prikazani rezultati na svakom od 3 skupa sljedova. Uz točnosti koje je svaka metoda postigla na pojedinim podskupovima, navedeni su i prosječna točnost te poredak metode u odnosu na sve ostale dostupne *alignment-free* metode koje su autori unijeli u aplikaciju.

3.1 Sljedovi nukleotida

U tablici 3.2 prikazani su rezultati postignuti na skupu regulacijskih sljedova nukleotida. Najbolji rezultati postižu se određivanjem Kullback-Leiblerove divergencije. Ta metoda ima značajno višu točnost na svim podskupovima u odnosu na druge dvije metode, kao i prosječnu točnost. U ukupnom poretku svih dostupnih metoda nalazi se na 3. mjestu. Ako se pogledaju pojedini podskupovi, određivanjem Kullback-Leiblerove divergencije najbolji rezultati postižu se na sljedovima koji pripadaju dišnom sustavu

vinske mušice, s točnošću od 86.1%. Ostale dvije metode svoje najbolje rezultate postižu na sljedovima koji pripadaju mišićnom tkivu čovjeka, međutim treća metoda i na tom podskupu postiže bolje rezultate.

Tablica 3.2 Rezultati na skupu sljedova nukleotida

		Prebrojavanje *k*-mera (k=4)	**Određivanje NCD**	**Određivanje KL divergencije**
Točnosti na podskupovima	fly_blastoderm	41.7%	50.7%	60.7%
	fly_eye	48.5%	49.3%	66.2%
	fly_pns	49.4%	48.6%	60.1%
	fly_tracheal_system	52.8%	52.8%	86.1%
	human_HBB	58.8%	53.7%	69.1%
	human_muslce	60.7%	54.7%	63.3%
	human_liver	38.9%	52.8%	55.6%
Prosječna točnost		51.05%	51.57%	63.42%
Poredak metode		43 / 55	38 / 55	3 / 55

3.2 Proteinski sljedovi male sličnosti

Točnost rezultata na skupu proteinskih sljedova aplikacija mjeri pomoću AUC vrijednosti. AUC vrijednost (engl. *area under the curve*) mjera je koja pokazuje koliko je metoda uspješna u razlikovanju sljedova koji pripadaju istoj grupi od nepovezanih sljedova, odnosno je li sposobna pouzdano prepoznati takve sljedove. Mjeri se točnost prepoznavanja sličnih sljedova na četiri strukturne razine: slični sljedovi unutar iste obitelji, unutar iste superobitelji, unutar istog fold-a i unutar istog razreda. Očekuje se da će udaljenosti biti najmanje između sljedova koji pripadaju istoj grupi unutar neke od strukturnih razina. U tablici 3.3 prikazani su rezultati na ovom skupu sljedova. Za razliku od nukleotidnih sljedova, u slučaju proteinskih sljedova male sličnosti najbolji rezultati dobivaju se uspoređivanjem Lempel-Ziv složenosti sljedova, odnosno

računanjem normalizirane kompresijske udaljenosti (NCD). Ta metoda ostvaruje najbolje rezultate u prepoznavanju sličnih sljedova na svim strukturnim razinama te se u ukupnom poretku metoda nalazi na 14. mjestu. Jako dobru rezultati postižu se i računanjem Kullback-Leiblerove divergencije. Može se primijetiti kako sve tri metode najbolje rezultate ostvaruju u prepoznavanju sljedova koji pripadaju istoj obitelji (engl. *family*), a najteže im je prepoznati sljedove koji pripadaju istom razredu (engl. *class*).

Tablica 3.3 Rezultati na skupu proteinskih sljedova male sličnosti

		Prebrojavanje *k*-mera (k=4)	**Određivanje NCD**	**Određivanje KL divergencije**
Točnosti prema strukturnim razinama	Class	0.48	0.57	0.56
	Fold	0.52	0.70	0.60
	Superfamily	0.49	0.72	0.65
	Family	0.52	0.79	0.69
Prosječna točnost		0.502	0.695	0.625
Poredak metode		47 / 48	14 / 48	35 / 48

3.3 Proteinski sljedovi velike sličnosti

Testiranje metoda na skupu proteinskih sljedova velike sličnosti provodi se na isti način kao i na prethodnom skupu proteinskih sljedova male sličnosti. Kao mjera točnosti koristi se AUC vrijednost, a uspoređuju se i rezultati postignuti na različitim strukturnim razinama. Rezultati su prikazani u tablici 3.4. Može se primijetiti kako su rezultati prepoznavanja sljedova koji pripadaju istoj obitelji (engl. *family*) i superobitelji (engl. *superfamily*) puno bolji u odnosu na proteinske sljedove male sličnosti. Poredak metoda po točnosti i dalje je isti: najbolji rezultati postižu se računanjem normalizirane kompresijske udaljenosti. Sve tri metode svoje najbolje rezultate postižu prilikom otkrivanja sljedova koji pripadaju istoj obitelji (engl. *family*). Metode temeljene na teoriji informacije ovdje postižu jako visoku točnost, AUC vrijednosti su vrlo blizu 1. Rezultati

prepoznavanja sljedova na strukturnim razinama razreda (engl. *class*) i fold-a nisu puno bolji u odnosu na prethodni skup proteinskih sljedova manje sličnosti, ali najbolje rezultate i dalje postiže metoda temeljena na složenosti i izračunu normalizirane kompresijske udaljenosti.

Tablica 3.4 Rezultati na skupu proteinskih sljedova velike sličnosti

		Prebrojavanje *k*-mera (k=4)	**Određivanje NCD**	**Određivanje KL divergencije**
Točnosti prema strukturnim razinama	Class	0.49	0.58	0.56
	Fold	0.52	0.71	0.61
	Superfamily	0.52	0.77	0.71
	Family	0.69	0.97	0.95
Prosječna točnost		0.555	0.758	0.708
Poredak metode		44 / 47	12 / 47	34 / 47

U sva tri skupa sljedova prebrojavanjem *k*-mera postižu se najlošiji rezultati u odnosu na ostale dvije metode. Za duljinu *k*-mera odabrana je vrijednost $k = 4$, s obzirom na to da je u literaturi navedeno kako se za optimalne rezultate u slučaju proteinskih sljedova *k* može postaviti na vrijednost između 2 i 6, a u slučaju sljedova nukleotida i na veće vrijednosti [2]. Odabrano je i nekoliko većih vrijednosti, ali rezultati su bili isti ili još lošiji. Moguće je da bi se detaljnijim pretraživanjem vrijednosti za *k* pronašli i nešto bolji rezultati, međutim to u ovom slučaju nije bilo praktično s obzirom na to da skupovi sadrže velik broj sljedova i izvršavanje kôda traje određeno vrijeme, a svaka datoteka s izračunatim udaljenostima mora se ručno učitati u aplikaciju kako bi se dobile informacije o uspješnosti. Unatoč tome, metode temeljene na teoriji informacije su se sigurno pokazale uspješnijima.

4. Zaključak

Metode usporedbe sljedova bez poravnanja, odnosno *alignment-free* metode imaju mnoge prednosti u odnosu na metode koje koriste poravnanje. Jedna od najvećih prednosti je manja vremenska i memorijska složenost što ih čini puno bržim i efikasnijim u odnosu na metode koje koriste poravnanje zbog čega se mogu koristiti i za usporedbu puno dužih sljedova, na primjer cjelokupnih genoma organizama. Još jedna prednost je što su puno jednostavnije za korištenje i ne ovise o vrsti sljedova koji se žele usporediti. Metode koje su implementirane u ovom radu mogu se koristiti za usporedbu bilo kakvih vrsta sljedova bez ograničenja. Nedostatak *alignment-free* metoda je taj što ne daju toliko informacija kao metode koje koriste poravnanje. Korištenjem metoda s poravnanjem mogu se odrediti točne regije dva slijeda koje se poklapaju ili imaju veliku sličnost (lokalno poravnanje). Moguće je odrediti točne operacije umetanja, brisanja ili zamjene znakova koje je potrebno provesti kako bi se jedan slijed dobio iz drugog. Ovakve informacije mogu biti od velike važnosti. *Alignment-free* metode ne mogu dati takve informacije, već samo računaju brojčanu vrijednost koja predstavlja udaljenost dva slijeda. Takve vrijednosti opet nije moguće samostalno interpretirati, već ih je potrebno usporediti s udaljenostima nekih drugih sljedova kako bi se donijeli zaključci o sličnosti. Dvije vrste metoda usporedbe sljedova mogu se koristiti zajedno, na primjer u slučaju potrebe za uspoređivanjem velikog broja sljedova. Korištenjem a*lignment-free* metoda može se pronaći podskup sljedova koji su potencijalno vrlo slični, a zatim se korištenjem metoda s poravnanjem mogu pronaći točne regije koje se podudaraju. Ovakav pristup povezuje brzinu i efikasnost *alignment-free* metoda s preciznošću metoda koje koriste poravnanje. Metode koje su implementirane u ovom radu vrlo su jednostavne i postoji puno prostora za poboljšanja. Moguće je i kombinirati pristupe više različitih metoda čime bi se mogla dodatno povećati točnost. Iz rezultata se može zaključiti kako su metode koje se temelje na teoriji informacije učinkovitije u odnosu na metode koje prebrojavaju pojavljivanje podnizova u sljedovima. Prednost dviju metoda temeljenih na teoriji informacije je i ta što ne ovise o početnim parametrima koji se moraju ručno odabrati. *Alignment-free* metode mogu imati vrlo široku primjenu i izvan područja bioinformatike, na primjer u analizi teksta te se općenito mogu koristiti za usporedbu bilo kakvih znakovnih nizova.

5. Sažetak

Analizirane su metode usporedbe bioloških sljedova koje ne koriste poravnanje, već se oslanjaju na vjerojatnost, statistiku i teoriju informacije kako bi odredile mjeru sličnosti ili udaljenosti sljedova. Glavna prednost u odnosu na metode koje koriste poravnanje je manja vremenska i memorijska složenost. U programskom jeziku Python implementirane su tri metode: prebrojavanje podnizova proizvoljno odabrane duljine koji se pojavljuju u oba slijeda, uspoređivanje Lempel-Ziv složenosti sljedova pomoću normalizirane kompresijske udaljenosti te uspoređivanje entropije sljedova pomoću Kullback-Leiblerove divergencije. Za testiranje implementiranih metoda korištena je web aplikacija AFproject. Preuzeta su tri skupa sljedova: jedan koji sadrži sljedove nukleotida te dva koja sadrže proteinske sljedove. Najbolje rezultate na skupu sljedova nukleotida postigla je metoda koja uspoređuje entropije sljedova dok je na skupovima proteinskih sljedova najbolje rezultate postigla metoda koja uspoređuje Lempel-Ziv složenosti sljedova.

6. Literatura

[1] W. R. Pearson, “An introduction to sequence similarity (‘homology’) searching,” *Current Protocols in Bioinformatics*, no. SUPPL.42, 2013, doi: 10.1002/0471250953.bi0301s42.

[2] A. Zielezinski, S. Vinga, J. Almeida, and W. M. Karlowski, “Alignment-free sequence comparison: Benefits, applications, and tools,” *Genome Biology*, vol. 18, no. 1. BioMed Central Ltd., Oct. 03, 2017. doi: 10.1186/s13059-017-1319-7.

[3] S. Vinga, “Editorial: Alignment-free methods in computational biology,” *Briefings in Bioinformatics*, vol. 15, no. 3. Oxford University Press, pp. 341–342, 2014. doi: 10.1093/bib/bbu005.

[4] M. Aboy, R. Hornero, D. Abásolo, and D. Álvarez, “Interpretation of the Lempel-Ziv complexity measure in the context of biomedical signal analysis,” *IEEE Transactions on Biomedical Engineering*, vol. 53, no. 11, pp. 2282–2288, Nov. 2006, doi: 10.1109/TBME.2006.883696.

[5] R. S. Borbely, “On normalized compression distance and large malware: Towards a useful definition of normalized compression distance for the classification of large files,” *Journal of Computer Virology and Hacking Techniques*, vol. 12, no. 4, pp. 235–242, Nov. 2016, doi: 10.1007/s11416-015-0260-0.

[6] J. Shlens, “Notes on Kullback-Leibler Divergence and Likelihood,” Apr. 2014, [Online]. Available: http://arxiv.org/abs/1404.2000

[7] A. Zielezinski *et al.*, “Benchmarking of alignment-free sequence comparison methods,” *Genome Biology*, vol. 20, no. 1, Jul. 2019, doi: 10.1186/s13059-019-1755-7.