SVEUČILIŠTE U ZAGREBU

FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

SEMINAR

Određivanje sličnosti sljedova nukleotida korištenjem neizrazitog integrala

Ivan Furač

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

Zagreb, svibanj, 2022.

Sadržaj

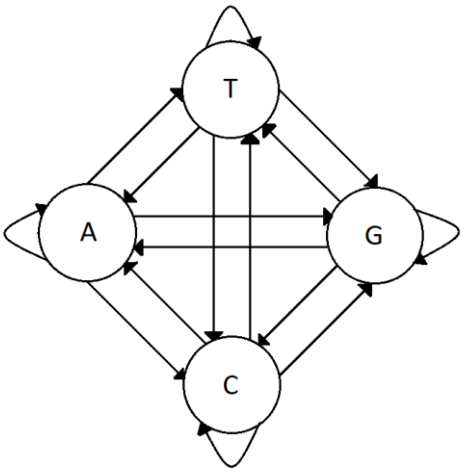
1. Uvod 1
2. Model Markovljevog lanca 2
3. Teorija neizrazitih skupova 4
4. Implementacija i rezultati 7
5. Zaključak 10
6. Literatura 11

1. Uvod

Komparativna analiza bioloških sljedova, odnosno sljedova nukleotida ili aminokiselina pripada temeljnim problemima koje proučava područje bioinformatike. Otkrivanje sličnosti između dva slijeda upućuje na njihovo zajedničko podrijetlo, odnosno na njihovu homologiju. Homologni sljedovi imaju sličnu strukturu, a često obavljaju i sličnu funkciju [1]. Brojne su primjene komparativne analize sljedova, poput predviđanja funkcija gena, sastavljanja sljedova, predviđanja molekularne strukture, analize povezanosti strukture proteina i njegove funkcije te izgradnje filogenetskih stabala [2]. Metode usporedbe sljedova se općenito mogu podijeliti u dvije kategorije. U prvoj kategoriji nalaze se metode koje se temelje na izravnom poravnanju sljedova (engl. *allignment-based methods*), odnosno traženju dijelova koji se podudaraju uzimajući u obzir brisanje, umetanje ili zamjenu pojedinih elemenata sljedova (nukleotida ili pojedinih aminokiselina). Primjeri računalnih alata koji koriste ove metode su BLAST, FASTA, ClustalW i HMMER [1]. Ove se metode danas dominantno koriste u analizi sljedova, međutim imaju izražene nedostatke poput velike računalne složenosti i nemogućnosti otkrivanja homolognih sljedova u slučaju smanjene sličnosti kao posljedice različitih mutacija [2]. U posljednje se vrijeme stoga pojačano istražuje druga kategorija metoda analize sljedova. U toj se kategoriji nalaze metode koje ne pokušavaju izravno poravnati dva slijeda (engl. *allignment-free methods*), već računaju sličnost sljedova korištenjem različitih tehnika iz teorije informacije, vjerojatnosti, statistike i linearne algebre. Glavna prednost metoda iz ove kategorije je njihova računalna složenost koja je puno manja u odnosu na složenost metoda koje se temelje na izravnom poravnanju sljedova, što omogućuje i usporedbu cjelokupnih genoma organizama. U ovom će radu biti prikazana jedna takva metoda detaljno opisana u literaturi [3]. Metoda se temelji na konceptima Markovljevog lanca i teorije neizrazitih skupova. U nastavku će ovi koncepti biti detaljnije objašnjeni kao i način na koji se povezuju u metodu za izračun sličnosti sljedova. Uspješnost metode odredit će se vrednovanjem nad dva različita skupa nukleotidnih sljedova, a rezultati će biti uspoređeni sa rezultatima tri jednostavnije *allignment-free* metode koje su implementirane u prethodnom seminaru: prebrojavanje *k*-mera, određivanje normalizirane kompresijske udaljenosti određivanjem Lempel-Ziv složenosti sljedova te određvanje *Kullback-Leiblerove* divergencije računanjem entropije sljedova.

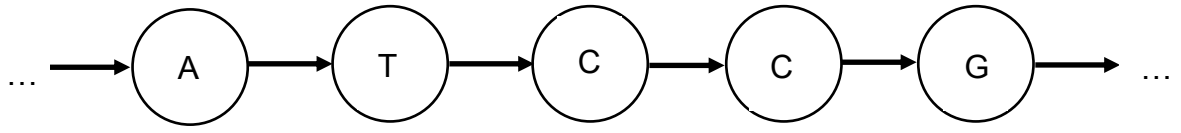
2. Model Markovljevog lanca

Slijed nukleotida može se promatrati kao Markovljev lanac s diskretnim vremenom pri čemu skup stanja sadrži pojedine nukleotide, *S = {A, T, C, G}.* Parametri lanca su vjerojatnosti prijelaza iz jednog stanja u drugo, pri čemu postoji šesnaest mogućih kombinacija. Drugim riječima, slijed nukleotida može se promatrati kao niz generiranih simbola pri čemu vjerojatnost pojavljivanja svakog simbola ovisi o određenom broju prethodnih simbola. Taj broj naziva se red Markovljevog lanca i može se označiti s *k*. Stanja i prijelazi Markovljevog lanca prikazani su na slici (Slika 1 Stanja i prijelazi modela Markovljevog lanca).



Slika 1 Stanja i prijelazi modela Markovljevog lanca

Slijed nukleotida može se promatrati kao lanac opaženih simbola, pri čemu pojavljivanje svakog simbola ima određenu vjerojatnost koja ovisi o tome koji se simbol nalazi na prvom prethodnom mjestu u lancu. U tom slučaju radi se o Markovljevom lancu 1. reda (Slika 2).



Slika 2 Primjer Markovljevog lanca prvog reda

Parametri lanca, odnosno vjerojatnosti pojavljivanja određenog simbola na mjestu *n* ako se na mjestu *n-1* nalazi drugi simbol, mogu se procijeniti iz promatranog slijeda nukleotida. Potrebno je izračunati ukupan broj pojavljivanja parova gdje se na susjednim mjestima nalaze nukleotidi *a* i *b*, za svih šesnaest mogućih kombinacija. Ako s N_{ab} označimo takav broj pojavljivanja, onda je vjerojatnost prijelaza iz nukleotida *a* u nukleotid *b* dana sa sljedećom formulom:

$$p_{ab} = \frac{N_{ab}}{N_{aA} + N_{aT} + N_{aC} + N_{aG}}$$

Nakon što su izračunate vjerojatnosti za sve moguće kombinacije nukleotida, one se mogu sažeto prikazati u matrici prijelaznih vjerojatnosti za Markovljev lanac 1. reda:

$$\boldsymbol{P}^1 = \begin{matrix} p_{AA} & p_{AT} & p_{AC} & p_{AG} \\ p_{TA} & p_{TT} & p_{TC} & p_{TG} \\ p_{CA} & p_{CT} & p_{CC} & p_{CG} \\ p_{GA} & p_{GT} & p_{GC} & p_{GG} \end{matrix}$$

U slučaju da se kao model koristi Markovljev lanac *k*-tog reda, vjerojatnost pojavljivanja nukleotida na mjestu *n* ovisi o vjerojatnostima pojavljivanja *k* nukleotida koji su se prethodno pojavili u lancu. U slučaju *k* = 2, vjerojatnost se može računati ovako:

$$p^2_{AT} = p^1_{AA} * p^1_{AT} + p^1_{AC} * p^1_{CT} + p^1_{AT} * p^1_{TT} + p^1_{AG} * p^1_{GT}$$

Iz formule je vidljivo da se u općenitom slučaju matrica vjerojatnosti prijelaza za model Markovljevog lanca *k*-tog reda može dobiti jednostavnim matričnim množenjem:

$$\boldsymbol{P}^k = \boldsymbol{P}^{k-1} * \boldsymbol{P}^1$$

Na ovaj se način svaki slijed nukleotida može prikazati preko jedinstvene matrice prijelaznih vjerojatnosti, ovisno o odabiru reda Markovljevog lanca koji se koristi kao model.

3. Teorija neizrazitih skupova

Koncept neizrazitih mjera temelji se na činjenici da su podatci koje dobivamo iz vanjskog svijeta često nepotpuni ili neprecizni, a naše je zaključivanje približno [4]. Dok klasične mjere imaju svojstvo aditivnosti, neizrazite mjere imaju svojstvo monotonosti, odnosno nužno im se povećava vrijednost kako se povećava broj elemenata skupa čija se mjera računa. Pretpostavimo da skup *S* sadrži četiri elementa, odnosno sve moguće parove jednog određenog nukleotida i svih ostalih nukleotida, npr. *S = {(AA), (AT), (AC), (AG)}.* Skup *R(S) je* skup svih mogućih podskupova skupa S, uključujući prazan skup i sam skup S. Tada je neizrazita mjera *μ* nad skupom R(S) definirana na sljedeći način:

$$\mu: R(S) \rightarrow [0, 1]$$

Vrijede sljedeća svojstva:

1. *μ(∅)* = 0
2. *μ(S)* = 1
3. *ako je A* ⊆ B, onda vrijedi *μ(A) ≤ μ(B),* odnosno vrijedi svojstvo monotonosti

Neizrazite mjere pojedinih elemenata skupa nazivaju se gustoćama i mogu se interpretirati kao važnost informacije koju element donosi cijelom skupu. U ovom slučaju gustoće pojedinih elemenata su zapravo vjerojatnosti prijelaza dobivene iz matrice u prethodnom koraku algoritma. Pokazalo se da se neizrazita mjera podskupa koji sadrži dva elementa iz originalnog skupa ne može dobiti jednostavnim zbrajanjem mjera pojedinih elemenata [5]. U tu se svrhu koristi takozvana Sugenova λ-mjera za koju vrijedi sljedeće, pri čemu λ > -1 i (*A* ∩ B) = ∅:

$$\mu(A \cup B) = \mu(A) + \mu(B) + \lambda * \mu(\mathrm{A}) * \mu(\mathrm{B})$$

Parametar λ može se izračunati iz gustoća svih *n* elemenata u skupu prema sljedećoj formuli:

$$\lambda + 1 = \prod_{j=1}^{n} (1 + \lambda * \mu^{j})$$

Osim što svaki element skupa nosi određenu informaciju u obliku neizrazite mjere, elementi također imaju i određenu funkciju važnosti koja govori koliko je informacija koju element pridonosi bitna u odnosu na ostale informacije. Kako bi se više različitih mjera agregiralo u jedinstvenu mjeru, koristi se koncept neizrazitog integrala [6]. U problemu određivanja sličnosti sljedova nukleotida, mjere koje su na raspolaganju su vjerojatnosti prijelaza i njih je potrebno povezati u jedinstvenu mjeru koja će reći koliko su sljedovi slični, a za to se koristi neizraziti integral. Dvije glavne klase neizrazitih integrala su Choquetovi integrali i Sugenovi integrali [6]. U ovom se radu koristi Sugenov integral.

Prilikom usporedbe dva slijeda nukleotida, uspoređuju se njihove matrice prijelaznih vjerojatnosti reda *k.* Svaki od šesnaest kombinacija parova nukleotida predstavlja jedan informacijski izvor. Neizrazita mjera izvora definira se na sljedeći način, s obzirom da je poželjnija veća vjerojatnost pojavljivanja para, odnosno bolja očuvanost parova nukleotida:

$$\mu^{ij} = \max\left(\left(\boldsymbol{P}_1^k\right)^{ij}, \left(\boldsymbol{P}_2^k\right)^{ij}\right)$$

Važnost *h* pojedinog elementa može se računati na sljedeći način:

$$h^{ij} = 1 - \left|(\boldsymbol{P}_1^k)^{ij} - (\boldsymbol{P}_2^k)^{ij}\right|$$

Gornji izraz zapravo predstavlja sličnost vjerojatnosti prijelaza iz stanje *i* u stanje *j* u dva promatrana slijeda koji se uspoređuju. Što je sličnost veća, to znači da taj par nukleotida ima veću važnost u određivanju ukupne sličnosti dva siljeda. Prije izračuna Sugenovog neizrazitog integrala, za svaki *i* potrebno je informacijske izvore (*ij), j = {1,2,3,4},* poredati u padajućem poretku ovisno o njihovoj važnosti *h*, odnosno mora vrijediti:

$$h^{i1} \geq h^{i2} \geq h^{i3} \geq h^{i4}$$

Sugenov neizraziti integral *I* sada se može računati prema formuli:

$$I = max\left(max_{i=1}^{4}\left(min_{j=1}^{4}\left(h^{ij}, \mu(A^{ij})\right)\right)\right)$$

pri čemu A^{ij} označava sljedeći skup parova nukleotida za zadani *i*:

$$A^{ij} = \{(i1), (i2), \ldots, (ij)\}$$

Za izračun neizrazite mjere ovakvog skupa potrebno je izračunati parametar λ prema prethodno spomenutoj formuli. Dobivena vrijednost *I* predstavlja sličnost dva slijeda nukleotida. Oznake *i* i *j* predstavljaju pojedine nukleotide iz skupa { A, T, C, G }. Udaljenost između dva slijeda nukleotida može se jednostavno dobiti iz sličnosti sljedećom formulom:

$$D(\boldsymbol{P}_1^k, \boldsymbol{P}_2^k) = 1 - I(\boldsymbol{P}_1^k, \boldsymbol{P}_2^k)$$

Cijeli algoritam može se sažeto prikazati u sljedećim koracima:

- izračun matrica prijelaznih vjerojatnosti za Markovljev lanac reda *k*
- izračun neizrazite mjere i važnosti informacije za svaki par nukleotida na temelju dviju matrica vjerojatnosti
- izračun sličnosti pomoću Sugenovog neizrazitog integrala

S obzirom da se u ovom radu uspoređuju sljedovi nukleotida, a broj mogućih nukleotida je četiri, matrice prijelaznih vjerojatnosti su dimenzija 4x4 i prilikom izračuna parametra λ potrebno je rješavati polinom četvrtog stupnja. Ovaj je algoritam u teoriji primjenjiv i na proteinske sljedove, međutim s obzirom da je broj različitih mogućih aminokiselina puno veći od broja različitih mogućih nukleotida, složenost algoritma postaje vrlo velika i u praksi je teže riješiti takav problem.

4. Implementacija i rezultati

Prethodno opisani algoritam implementiran je u programskom jeziku Python. Za učitavanje sljedova spremljenih u FASTA formatu koristi se sučelje SeqIO iz biblioteke Bio. Za rješavanje polinoma četvrtog stupnja prilikom izračuna parametra λ koristi se klasa poly1d iz biblioteke numpy. Za izračun matrica prijelaznih vjerojatnosti u prvom koraku algoritma potrebno je odrediti parametar *k* koji predstavlja red modela Markovljevog lanca. Isprobano je više vrijednosti za taj parametar u intervalu [*1, 8*], međutim najbolji rezultati dobiveni su odabirom vrijednosti *k=2*. Uspješnost algoritma provjerena je na dva različita skupa sljedova nukleotida te je uspoređena s tri jednostavnije *allignment-free* metode koje su opisane u tablici ispod (Tablica 1).

Tablica 1 Prikaz metoda koje su korištene za usporedbu uspješnosti

Prebrojavanje k-mera	U sljedovima se prebrojavaju podnizovi zadane duljine *k*. Udaljenost se računa kao euklidska udaljenost dva vektora koji sadrže frekvencije pojavljivanja pojedinih podnizova.
Određivanje normalizirane kompresijske udaljenosti	Računaju se Lempel-Ziv složenosti pojedinih sljedova, a udaljenost se računa pomoću normalizirane kompresijske udaljenosti između tih složenosti.
Određivanje Kullback-Leiblerove divergencije	Računaju se entropije pojedinih sljedova, a mjera udaljenosti je Kullback-Leiblerova divergencija.

Za testiranje algoritma na prvom skupu sljedova korištena je stranica AFproject dostupna na adresi http://afproject.org [2]. Ta stranica nudi mogućnost preuzimanja skupova sljedova za testiranje te učitavanje dobivenih vrijednosti u obliku TSV datoteke nakon čega se provodi automatska provjera točnosti. Skup koji je korišten

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 pri čemu 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. Osim što se prikazuje točnost metode na pojedinim podskupovima, također se prikazuje i poredak metode u odnosu na sve ostale dostupne metode. U tablici ispod prikazane su dobivene točnosti (Tablica 2).

Tablica 2 Točnosti na prvom skupu sljedova

		Neizraziti integral	***k*-mer**	**NCD**	**KL-div.**
Točnosti na podskupovima	fly_blastoderm	53.0%	41.7%	50.7%	60.7%
	fly_eye	55.2%	48.5%	49.3%	66.2%
	fly_pns	48.2%	49.4%	48.6%	60.1%
	fly_tracheal_system	66.7%	52.8%	52.8%	86.1%
	human_HBB	63.2%	58.8%	53.7%	69.1%
	human_muscle	59.3%	60.7%	54.7%	63.3%
	human_liver	58.3%	38.9%	52.8%	55.6%
Prosječna točnost		55.54%	51.05%	51.57%	63.42%
Poredak metode		12 / 55	43 / 55	38 / 55	3 / 55

Iz tablice je vidljivo da metoda neizrazitog integrala postiže uglavnom bolje točnosti od prve dvije jednostavnije metode, međutim postiže nešto lošije rezultate od metode koja računa Kullback-Leiblerovu divergenciju i koja se u ukupnom poretku metoda po točnosti nalazi na trećem mjestu. Zanimljivo je primijetiti kako metoda neizrazitog integrala postiže najbolju točnost od svih metoda u tablici na posljednjem podskupu koji sadrži sljedove iz tkiva ljudske jetre.

Drugi skup sljedova nad kojim je metoda testirana sadrži 42 slijeda virusa HIV. Skup je preuzet s online baze podataka sljedova virusa HIV-a, s adrese

https://www.hiv.lanl.gov/content/sequence/NEWALIGN/align.html. Preuzeti sljedovi su unaprijed poravnati i sadrže znak "-" na mjestima na kojima je došlo do brisanja ili umetanja nukleotida, što znači da je moguće izračunati referentnu „stvarnu“ sličnost, a uspješnost pojedinih metoda se mjeri odstupanjem njihove izračunate sličnosti od stvarne sličnosti. Referentna sličnost parova sljedova HIV-a se može izračunati uspoređivanjem znakova na istim mjestima unutar sljedova. Sličnost sljedova je ukupan broj podudaranja podijeljen s ukupnom duljinom sljedova. Udaljenost se pak može izračunati kao broj nepodudaranja podijeljen s ukupnom duljinom sljedova. S obzirom da su sljedovi optimalno poravnati, ovako izračunata mjera je dobar pokazatelj koji su parovi sljedova sličniji u odnosu na druge i može se koristiti za ocjenu uspješnosti metoda. S obzirom na to da udaljenosti koje računaju pojedine metode imaju različite raspone vrijednosti, njihove udaljenosti nije moguće izravno uspoređivati s referentnim udaljenostima, već se uspoređuje poredak. Izračunate udaljenosti između svih parova sljedova se sortiraju te se pogreška metode može mjeriti kao zbroj odstupanja u poretku metode u odnosu na poredak referentnih udaljenosti. Na primjer, ako se udaljenost nekog para sljedova u referentnom poretku nalazi na 50. mjestu, a u poretku udaljenosti koje računa metoda na 100. mjestu, pogreška za taj par sljedova za tu metodu iznosi 50. Zbrojem pogrešaka za sve parove može se dobiti ukupna pogreška. Ukupne pogreške pojedinih metoda prikazane su u tablici (Tablica 3).

Tablica 3 Pogreške metoda u računanju udaljenosti na drugom skupu sljedova

	Neizraziti integral	***k*-mer**	**NCD**	**KL-div.**
Pogreška	243838	199193	251539	230158

U tablici je vidljivo kako na drugom skupu sljedova najbolje rezultate postiže metoda prebrojavanja *k*-mera. Metoda neizrazitog integrala se pokazala otprilike jednako uspješna kao metode računanja normalizirane kompresijske udaljenosti i računanja Kullback-Leiblerove divergencije, iako je za nijansu bolja od prve i lošija od druge, ali to su male razlike u uspješnosti.

5. Zaključak

U ovome radu prikazana je *allignment-free* metoda uspoređivanja sljedova koja se temelji na konceptima Markovljevog lanca i teorije neizrazitih skupova. Općenita prednost *allignment-free* metoda u odnosu na metode koje računaju izravno poravnanje sljedova je njihova brzina i efikasnost što posebno dolazi do izražaja u slučajevima u kojima je skup sljedova koji se uspoređuju iznimno velik ili su sami sljedovi dugački, primjerice sadrže cjelokupne genome organizama. Računanje sličnosti sljedova iz takvih skupova *allignment-free* metode mogu obaviti u vrlo kratkom vremenu koje se može mjeriti u minutama, dok je metodama koje računaju izravno poravnanje za to potrebno puno više vremena. Međutim, nedostatak *allignment-free* metoda je taj što ne daju toliko informacija o sličnosti sljedova koje mogu dati *allignment-based* metode. Također, rezultati pojedinih *allignment-free* metoda se međusobno uvelike razlikuju što je vidljivo i iz rezultata dobivenih u ovom radu. Ne može se pronaći univerzalno najbolja metoda koja će dati najbolje rezultate nad svim skupovima sljedova, već svaka metoda ima određene prednosti i nedostatke koji dolaze do izražaja na različitim skupovima podataka. Dvije kategorije metoda za usporedbu sljedova se mogu kombinirati tako da se prvo korištenjem *allignment-free* metoda pronađe određeni podskup sljedova koji su sličniji, a zatim se korištenjem *allignment-based* metoda može odrediti točno poravnanje sljedova iz tog podskupa. Metoda koja je detaljno prikazana u ovom radu pokazuje dobre rezultate u usporedbi s ostalim *allignment-free* metodama. U prosjeku pokazuje ili djelomično bolje ili jednake rezultate kao i ostale metode ako se u obzir ne uzmu metode koje jako odskaču usvojim rezultatima. Koncept Markovljevog lanca na kojem se temelji ova metoda često se koristi u analizi sljedova i može dati mnoge zanimljive informacije. Koncept neizrazitih mjera i neizrazitog integrala na prvu možda nema puno dodirnih točaka s područjem bioinformatike, međutim može se dobro iskoristiti za analizu pojedinih vrijednosti i njihovu agregaciju u određenu zajedničku mjeru. Rezultati bi se mogli dodatno poboljšati preciznijim odabirom reda Markovljevog lanca *k* ili odabirom drukčije funkcije važnosti prilikom računanja neizrazitog integrala. Ovdje opisani koncepti mogu se sigurno iskoristiti i za brojne druge primjene u području bioinformatike, ali i šire.

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

[3] A. K. Saw, G. Raj, M. Das, N. C. Talukdar, B. C. Tripathy, and S. Nandi, “Alignment-free method for DNA sequence clustering using Fuzzy integral similarity,” *Scientific Reports*, vol. 9, no. 1, Dec. 2019, doi: 10.1038/s41598-019-40452-6.

[4] L. Garmendia, “The Evolution of the Concept of Fuzzy Measure.” [Online]. Available: www.fdi.ucm.es/profesor/lgarmend

[5] H. T. Nguyen, V. Kreinovich, J. Lorkowski, and S. Abu, “Why Sugeno lambda-Measures Why Sugeno lambda-Measures Part of the Computer Sciences Commons,” 2015. [Online]. Available: https://scholarworks.utep.edu/cs_techrephttps://scholarworks.utep.edu/cs_techrep/906

[6] S. Abbaszadeh, M. Eshaghi, and M. de la Sen, “The Sugeno fuzzy integral of log-convex functions,” *Journal of Inequalities and Applications*, vol. 2015, no. 1, pp. 1–12, Dec. 2015, doi: 10.1186/s13660-015-0862-6.