



Univerzitet u Beogradu
Matematički fakultet

Seminarski rad

Analiza, klasifikacija i klasterovanje proteinskih blokova

Mentor:

Prof. dr Nenad Mitić
Katedra za računarstvo i informatiku

Studenti:

Anja Milutinović 235/2021
Đurđa Milošević 84/2021
Smer: Informatika

Datum: 2024/25

Contents

1	Uvod	2
2	O Proteinskim blokovima	3
3	Analiza	6
3.1	Opis podataka	6
3.1.1	Izdvajanje retkih prelaza između proteinskih blokova u podacima	8
3.2	Izdvajanje parova aminokiselina prisutnih u neočekivanim prelazima	8
3.3	Izdvajanje parova sekundarnih struktura koji se javljaju u neočekivanim prelazima	10
3.3.1	Izdvajanje kombinacija aminokiselina i sekundarnih struktura u neočekivanim prelazima	11
3.4	Analiza vrednosti pLDDT parametra	12
3.4.1	Vrednosti pLDDT u opštem slučaju	14
3.4.2	Vrednosti pLDDT prema proteinskim blokovima	15
3.4.3	Vrednosti pLDDT prema aminokiselinama	15
3.4.4	Vrednosti pLDDT prema sekundarnim strukturama	15
3.5	Analiza vrednosti RSA parametra	15
3.6	Ispitivanje zastupljenosti aminokiselina u podacima	15
4	Klasifikacija	18
5	Klasterovanje	19
6	Zaključak	20

1 Uvod

Proteini su osnovni gradivni blokovi svih ćelija u organizmu i kao takvi igraju ključnu ulogu u održavanju života, reprodukciji, odbrani i replikaciji. Funkcije proteina zavise od njihove strukture, zbog čega je analiza strukture proteina od izuzetnog značaja u bioinformatičkim i biohemijskim istraživanjima.

U nastojanju da se postigne precizniji, detaljniji i informativniji prikaz trodimenzionalne strukture proteina, razvijeni su proteinski blokovi. Kao jedni od najistaknutijih predstavnika strukturnih alfabeta, pokazano je da omogućavaju detekciju strukturne sličnosti između proteina sa izuzetnom efikasnošću.

Ovaj seminarski rad se fokusira na proteinske blokove, sa posebnim akcentom na retke i neočekivane prelaze između njih. Niz proteinskih blokova dobijen je analizom humanog proteoma, koji je generisan pomoću AlphaFold2 programa. Seminarski rad se sastoji od tri ključna segmenta: analize, klasifikacije i klasterovanja proteinskih blokova.

Analiza obuhvata izdvajanje aminokiselina i sekundarnih struktura u retkim prelazima, određivanje prirode vrednosti pLDDT (eng. *the predicted local distance difference test*) i RSA (eng. *relative solvent accessibility*) parametara, kao i poređenje zastupljenosti pojedinačnih aminokiselina u podacima u odnosu na očekivane procenete.

Klasifikacija je usredsređena na predviđanje aminokiselina i sekundarnih struktura u prelazima, kao i parova proteinskih blokova koji čine posmatrane prelaze. Ovaj pristup omogućava donošenje interesantnih zaključaka o relevantnosti, ubedljivosti i stabilnosti proteinskih blokova kao strukturnih indikatora.

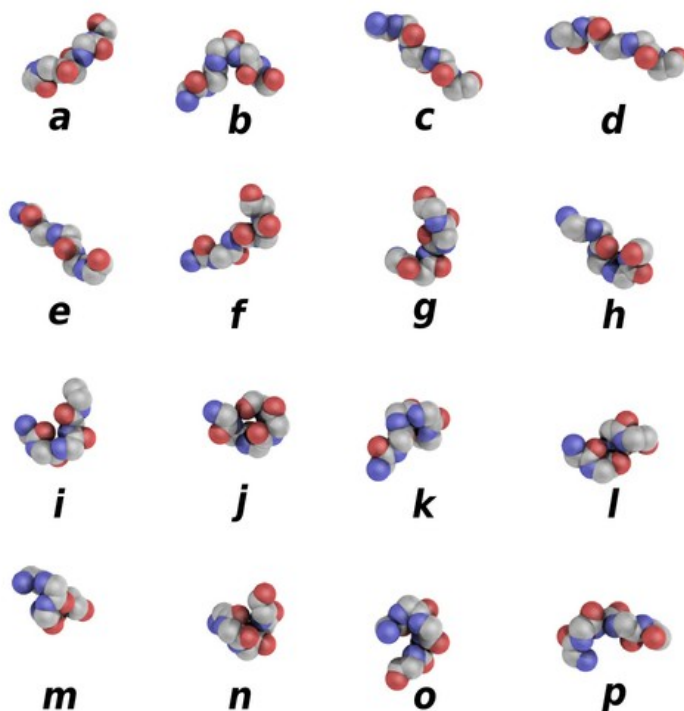
Klasterovanje je izvedeno nad skupom podataka koji sadrži informacije o strukturi proteina, uključujući prelaze između proteinskih blokova, kao i aminokiseline i sekundarne strukture prisutne u tim prelazima.

Seminarski rad je realizovan u okviru kursa „Istraživanje podataka 2” na Matematičkom fakultetu Univerziteta u Beogradu.

2 O Proteinskim blokovima

Pronalaženje sličnosti u prostornoj strukturi proteina je važno jer može da ukaže na sličnosti u funkcionalnosti proteina, koja nije vidljiva ispitivanjem sekvencijalnih informacija o proteinu. Eksperimentalno određivanje trodimenzionalne strukture proteina je skup i vremenski zahtevan proces. Zato, potrebno je da se pronađe efikasan i pouzdan način opisivanja trodimenzionalne strukture proteina, kao i način upoređivanja više struktura proteina međusobno.

Struktura proteina se obično opisuje kao alfa-heliks ili beta-ravan zasnovano na vodoničnim vezama između peptidnih veza unutar glavnog lanca proteina, ali ovaj pristup se pokazao previše uprošćen jer preko 50% strukture proteina ostaje neopisano. Trodimenzionalna struktura može se opisati i korišćenjem aproksimativnih prototipova lokalne strukture proteina. Skup definisanih prototipova lokalne strukture se naziva i strukturni alfabet. Direktno određivanje trodimenzionalne strukture proteina je težak problem, zato se koriste strukturni alfabeti koji opisuju trodimenzionalnu strukturu proteina jednodimenzionalnim nizom strukturnih prototipova.



Slika 1: Šematski prikaz 16 proteinskih blokova označenih slovima od a do p

Jedan od najpoznatijih strukturnih alfabeti jeste Proteinski blokovi (PB),

koje je razvio De Brevern (2000). Proteinski blokovi se sastoji od 16 prototipova, koji su izdvojeni korišćenjem algoritama klasterovanja, Kohenonove samoorganizovajuće mape, nad fragmentima od pet uzastopnih aminokiselina kod proteina sa već poznatom strukturom. U prvom istraživanju je korišćeno 228 poznatih proteina, a kasnije je istraživanje ponovljeno nad 400 poznatih proteina. Procedura klasterovanja se odvijala u tri koraka. U prvom koraku se koristi mera sličnosti fragmenata RMSDA (eng. Root Mean Square Deviation on Angle), u drugom koraku se dodatno koristi i verovatnoća prelaska jednog fragmenta u drugi u sekvenci, dok se u trećem koraku izbacuje ograničenje verovatnoće prelaska. Na kraju je odabrano 16 proteinskih blokova, definisanih sa osam diedarskih uglova, ψ_{i-2} , ϕ_{i-1} , ψ_{i-1} , ϕ_i , ψ_i , ϕ_{i+1} , ψ_{i+1} , ϕ_{i+2} u odnosu na centralnu aminokiselinu u fragmentu dužine pet. (Tabela 1)

PB	ψ_{i-2}	ϕ_{i-1}	ψ_{i-1}	ϕ_i	ψ_i	ϕ_{i+1}	ψ_{i+1}	ϕ_{i+2}
a	41.14	75.53	13.92	-99.80	131.88	-96.27	122.08	-99.68
b	108.24	-90.12	119.54	-92.21	-18.06	-128.93	147.04	-99.90
c	-11.61	-105.66	94.81	-106.09	133.56	-106.93	135.97	-100.63
d	141.98	-112.79	132.20	-114.79	140.11	-111.05	139.54	-103.16
e	133.25	-112.37	137.64	-108.13	133.00	-87.30	120.54	77.40
f	116.40	-105.53	129.32	-96.68	140.72	-74.19	-26.65	-94.51
g	0.40	-81.83	4.91	-100.59	85.50	-71.65	130.78	84.98
h	119.14	-102.58	130.83	-67.91	121.55	76.25	-2.95	-99.88
i	130.68	-56.92	119.26	77.85	10.42	-99.43	141.40	-98.01
j	114.32	-121.47	118.14	82.88	-150.05	-83.81	23.35	-85.82
k	117.16	-95.41	140.40	-59.35	-29.23	-72.39	-25.08	-76.16
l	139.20	-55.96	-32.70	-68.51	-26.09	-74.44	-22.60	-71.74
m	-39.62	-64.73	-39.52	-65.54	-38.88	-66.89	-37.76	-70.19
n	-35.34	-65.03	-38.12	-66.34	-29.51	-89.10	-2.91	77.90
o	-45.29	-67.44	-27.72	-87.27	5.13	77.49	30.71	-93.23
p	-27.09	-86.14	0.30	59.85	21.51	-96.30	132.67	-92.91

Tabela 1: Referentni uglovi Proteinskih blokova

Proteinski blokovi su označeni slovima od a do p (Slika 1). Najčešći proteinski blokovi, m i d, odgovaraju redom alfa-heliksi i beta-ravni. Proteinski blokovi od k do j odgovaraju nespecifičnim strukturama (eng. coil). Prevođenje u sekvencu proteinskih blokova kod proteina sa poznatom 3D strukturom odvija se tako što se svakom fragmentu uzastopnih aminokiselina dužine pet dodeli jedan proteinski blok sa najmanjom vrednošću RMSDA-a. Može se desiti i da se diedralni uglovi ne mogu izračunati u tom se slučaju dodeljuje slovo Z. Na ovaj način svaka aminokiselina učestvuje u pet proteinskih blokova, osim prve i poslednje. Strukturni alfabet Proteinski blokovi se koristi i u drugim podoblastima bioinformatike kao što su nadređivanje 3D strukture proteina, istraživanje strukture proteina, definisanje mesta vezivanja, i analize lokalnih konformacija poremećenih proteina. Postoje alati koji prevode PDB fajlove u sekvence proteinskih blokova, kao što je Plxplore.

3 Analiza

U ovom poglavlju analizirani su prelazi između proteinskih blokova. Fokus analize bio je na identifikaciji neočekivanih prelaza između proteinskih blokova i proučavanju aminokiselina i sekundarnim struktura prisutnih u njima. Dodatno, ispitane su vrednosti relevantnih parametara, kao i zastupljenost pojedinačnih aminokiselina u podacima.

3.1 Opis podataka

Podaci korišćeni u analizi obuhvataju informacije o proteinskim blokovima (PBs), aminokiselinama (AA), sekundarnim strukturama (S2), predviđenoj učestalosti prelaza, kao i dodatnim parametrima poput pLDDT i RSA. Izvor podataka su rezultati generisani pomoću AlphaFold2 programa. Konkretno, analiza je sprovedena nad dve datoteke: prvobitnog, obimnijeg skupa podataka, koji detaljnije opisuje prelaze između proteinskih blokova i manjeg podskupa dobijenog filtriranjem rezultata u skladu sa humanim proteomom.

Podaci su organizovani u tabelarnom formatu, pri čemu svaki red sadrži parove proteinskih blokova koji čine prelaz, parove aminokiselina i sekundarnih struktura koji ga opisuju, predviđenu učestalost prelaza i vrednosti parametara pLDDT i RSA.

U nastavku je dat detaljan opis atributa:

- **Protein_number** – diskretan, kategorijski i redni atribut koji označava pojedinačne proteine u skupu podataka.
- **res_number** – diskretan, kategorijski i redni atribut koji označava poziciju aminokiselinskog ostatka unutar sekvence.
- **PB1, PB2** – diskretni, kategorijski i nominalni atributi koji predstavljaju oznake proteinskih blokova između kojih dolazi do prelaza.
- **AA1, AA2** – diskretni, kategorijski i nominalni atributi koji opisuju aminokiseline prisutne u prelazu. Postoji ukupno 20 različitih vrednosti za ove attribute, u skladu sa standardnim skupom aminokiselina koje grade proteine. U tabeli 2 prikazane su odgovarajuće jednoslovne oznake za svaku aminokiselinu.

Aminokiselina	Jednoslovna oznaka
Alanin	A
Arginin	R
Asparagin	N
Asparaginska kiselina	D
Cistein	C
Glutaminska kiselina	E
Glutamin	Q
Glicin	G
Histidin	H
Izoleucin	I
Leucin	L
Lizin	K
Metionin	M
Fenilalanin	F
Prolin	P
Serin	S
Treonin	T
Triptofan	W
Tirozin	Y
Valin	V

Tabela 2: Aminokiseline i njihove jednoslovne oznake korišćene kao vrednosti atributa AA1 i AA2.

- **S2_1, S2_2** – diskretni, kategorijski i nominalni atributi koji predstavljaju sekundarne strukture prisutne u prelazu. Vrednosti ovih atributa određene su pomoću DSSP (*eng. Dictionary of Secondary Structure of Proteins*) programa. DSSP program se koristi za dodeljivanje jednog od osam stanja sekundarne strukture aminokiselinama tako što se identifikuju vodonične veze između amino i karboksilnih grupa glavnog lanca proteina. U tabeli 3 prikazane su oznake i odgovarajući tipovi sekundarne strukture [1].

Oznaka	Tip sekundarne strukture
H	Alfa-heliks (<i>eng. α-helix</i>)
B	Beta-most (<i>eng. beta bridge</i>)
E	Prošireni beta list (<i>eng. extended beta sheet</i>)
G	3(10)-heliks
I	Pi-heliks (<i>eng. π-helix</i>)
T	Heliks okret (<i>eng. helix-turn</i>)
S	Zavojnica (<i>eng. bend</i>)
C	Nespecifična ili neorganizovana struktura (<i>eng. coil</i>)

Tabela 3: Sekundarne strukture proteina prema DSSP standardu, korišćene kao vrednosti atributa S2_1 i S2_2.

- **expected_frequency** - neprekidan, kvantitativan i razmerni atribut čija vrednost pripada intervalu $[0,1]$ i označava očekivanu učestalost prelaza između proteinskih blokova.
- **pLDDT, RSA1, RSA2** – neprekidni, kvantitativni i razmerni atributi čije su vrednosti u intervalu $[0,100]$. Parametar pLDDT procenjuje pouzdanost predikcije lokalne strukture proteina, dok RSA1 i RSA2 predstavljaju relativnu dostupnost aminokiselinskih ostataka rastvaraču.

3.1.1 Izdvajanje retkih prelaza između proteinskih blokova u podacima

Glavni zadatak analize bio je proučavanje prelaza između proteinskih blokova pri čemu je akcenat stavljen na neočekivane, retke prelaze. Takvi prelazi su, zbog svoje prirode, od posebnog interesa za istraživanje jer mogu doprineti novim saznanjima o nizu proteinskih blokova koji opisuje trodimenzionalnu strukturu proteina.

U ovom radu, retkim prelazom smatra se onaj koji se javlja u manje od 1% slučajeva.

3.2 Izdvajanje parova aminokiselina prisutnih u neočekivanim prelazima

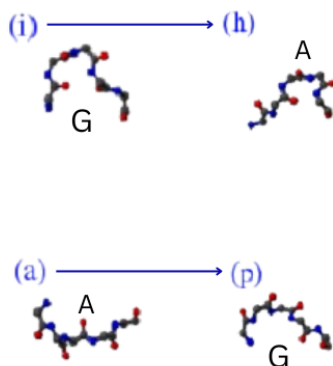
U cilju pronalaženja potencijalnih korelacija između atributa, odnosno između parova aminokiselina vezanih za retki prelaz i proteinskih blokova koji formiraju prelaz, izvršeno je izdvajanje tih parova. Pored toga, izračunate su frekvencije svih parova, a zatim su izdvojeni oni najfrekventiniji.

S obzirom na to da se u podacima korišćenim za ovu analizu prelazi ne nadovezuju, posmatran je pojedinačno svaki prelaz i unutar njega određen

svaki par aminokiselina. Važno je napomenuti da redosled aminokiselina u paru nije irelevantan, zbog čega se parovi aminokiselina (A1, A2) ne mogu smatrati identičnim parovima (A2, A1). Kako bi ova tvrdnja bila intuitivnija i jasnija, potrebno je prvo objasniti zašto je prelaz određen sa dve aminokiseline iako su proteinski blokovi sačinjeni od 5 uzastopnih aminokiselina.

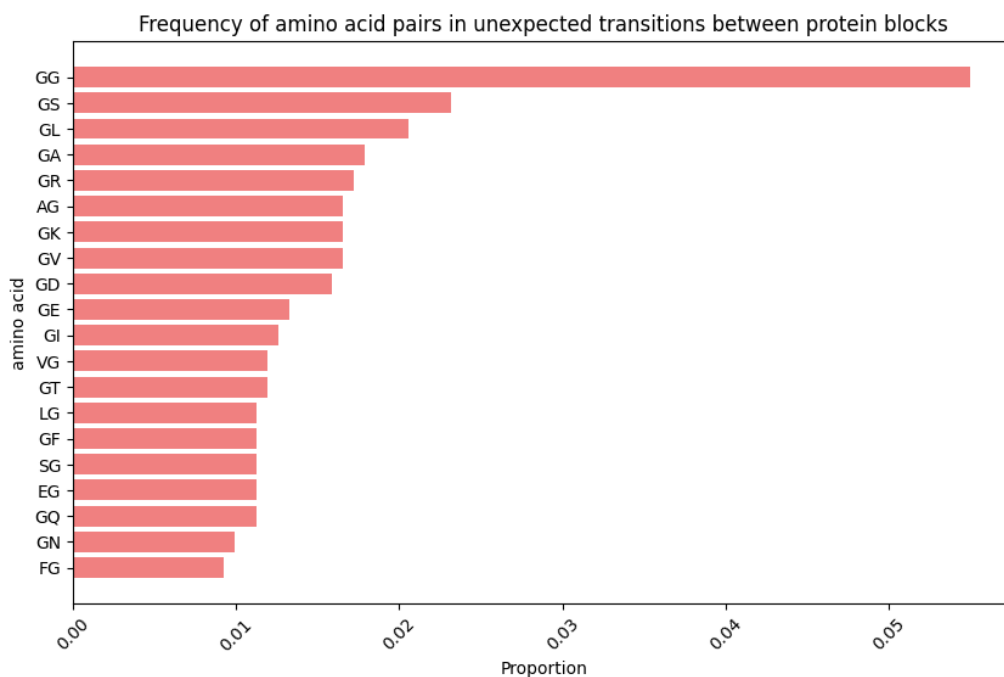
Naime, prilikom dodeljivanja proteinskog bloka fragmentu od 5 aminokiselina, centralna aminokiselina igra ključnu ulogu jer je ona ta koja efektivno definiše fragment i omogućava izračunavanje 8 diedarskih uglova ψ_{i-2} , ϕ_{i-1} , ψ_{i-1} , ϕ_i , ψ_i , ϕ_{i+1} , ψ_{i+1} , ϕ_{i+2} . Ovi uglovi se zatim upoređuju sa već definisanim diedarskim uglovima prototipova proteinskih blokova nakon čega se dodeljuje odgovarajući proteinski blok. Dakle, za svaki prelaz centralna aminokiselina je najvažnija zbog njenog uticaja na ceo fragment i zato je baš ona ta koja je izabrana da bude deo podataka, odnosno da bude predstavnik u prelazu.

Imajući u vidu prethodno navedeno, obrnut redosled aminokiselina u paru može da predstavlja potpuno različit prelaz između proteinskih blokova i zato je važno analizirati ih kao zasebne parove. Slika 2 prikazuje primer iz skupa podataka koji ilustruje iznesenu tvrdnju.



Slika 2: Redosled aminokiselina u paru je važan.

Izračunate su frekvencije svakog para aminokiselina u prelazima. Najčešće se javlja par (G, G). Pregled ostalih 19 najučestalijih parova dat je na slici 3.



Slika 3: Frekvencija parova aminokiselina u retkim prelazima.

Dobijeni najfrekventniji par je od posebnog interesa s obzirom na specifične karakteristike glicina. Glicin, zbog svog malog bočnog lanca, ima ključnu ulogu u omogućavanju fleksibilnosti u proteinu. Takođe, glicin doprinosi stabilnosti tercijalne strukture proteina [2]. S obzirom na to, njegova velika učestalost može ukazivati na stabilnost u prelazima.

3.3 Izdvajanje parova sekundarnih struktura koji se javljaju u neočekivanim prelazima

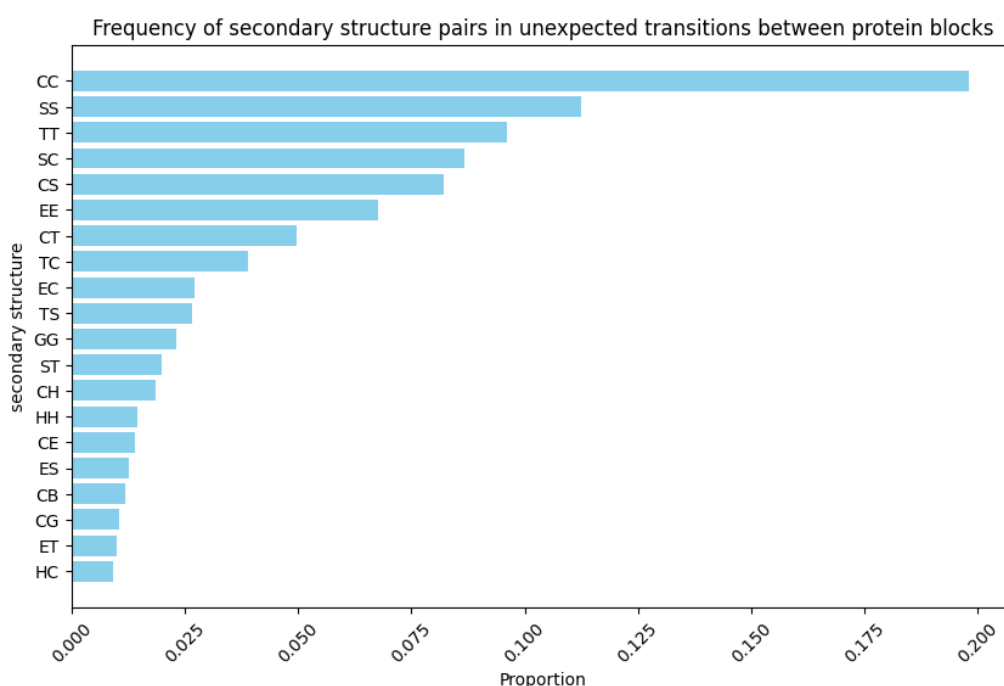
Nakon izdvajanja aminokiselina, sledeći korak u analizi bio je ispitivanje sekundarnih struktura koje se javljaju u prelazima i njihove učestalosti. Kao i u slučaju aminokiselina, izvojeni su parovi sekundarnih struktura tako da je redosled unutar svakog para važan.

Sekundarna struktura proteina definiše lokalne konformacije polipeptidnog lanca koje nastaju usled interakcija između atoma kičme lanca. Najčešća klasifikacija sekundarne strukture je na α -zavojnice i β -ploče, dok se ostale konformacije označavaju kao neregularno stanje (*eng. coil*). Međutim, utvrđeno je da u proseku čak 50% ostataka [3] kolektivno tretira kao neregularno stanje što značajno pojednostavljuje strukturu proteina i dovodi do brojnih ograničenja u njenoj analizi. To je jedan od glavnih razloga formiranja

proteinskih blokova kao koncepta.

U skladu sa tim, očekuje se da će analiza parova sekundarnih struktura u retkim prelazima između proteinskih blokova pokazati pretežnu zastupljenost neregularnih stanja, odnosno *coil*, pošto sami proteinski blokovi treba što vernije da opišu upravo one konformacije koje se ne mogu jednostavno klasifikovati kao α -zavojnice i β -ploče, niti njihove varijacije (E - *extended beta sheet*, G - *(10)-helix*)).

Izračunate su frekvencije parova sekundarnih struktura u prelazima i na slici 4 su prikazani oni najučestaliji, odnosno dvadeset najučestalijih parova.



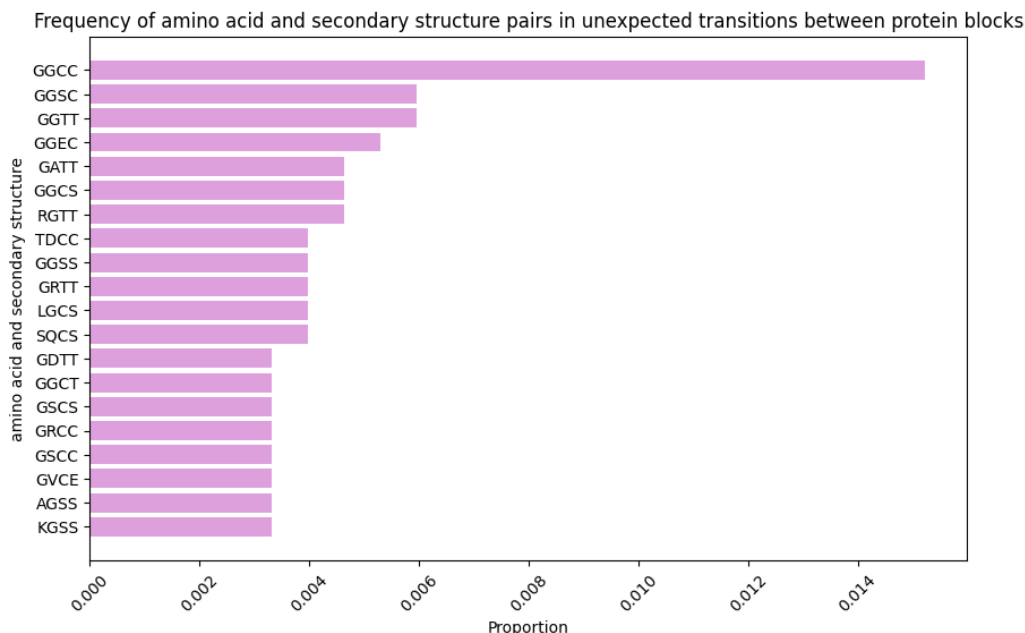
Slika 4: Frekvencija parova sekundarnih struktura u retkim prelazima.

Rezultati analize pokazuju da su najčešće zastupljeni parovi CC (*coil-coil*), SS (*bend-bend*) i TT (*helix-turn-helix-turn*), koji ne pripadaju klasičnim tipovima sekundarnih struktura. Ovo je u skladu sa pretpostavkom da će u retkim prelazima dominirati neregularne sekundarne strukture.

3.3.1 Izdvajanje kombinacija aminokiselina i sekundarnih struktura u neočekivanim prelazima

Pored analize parova aminokiselina i sekundarnih struktura zasebno, izvršeno je i ispitivanje njihovih kombinacija radi pronalaska međusobnih korelacija u prelazima.

Na slici 5 dat je prikaz najfrekventnijih kombinacija.



Slika 5: Frekvencija kombinacija parova aminokiselina i sekundarnih struktura u retkim prelazima.

Rezultati pokazuju da je daleko najzastupljenija kombinacija GG CC. Može se zaključiti da je razlog toga visok stepen fleksibilnosti glicina, koji ga može učiniti sklonijim ka formiranju neregularnih konformacija. Prethodna analiza je pokazala da su pojedinačno najzastupljeniji parovi aminokiselina i sekundarnih struktura baš glicin i neregularno stanje, što dodatno doprinosi tvrdnji da postoji visoka korelisanost kako između njih tako i između njih i proteinskih blokova koji čine prelaz.

3.4 Analiza vrednosti pLDDT parametra

pLDDT (*eng. the predicted local distance difference test*) predstavlja meru pouzdanost lokalne strukture proteina, predviđene AlphaFold2 programom. Vrednosti su na skali od 0 do 100, pri čemu veće vrednosti ukazuju na visoku pouzdanost i precizniju predikciju.

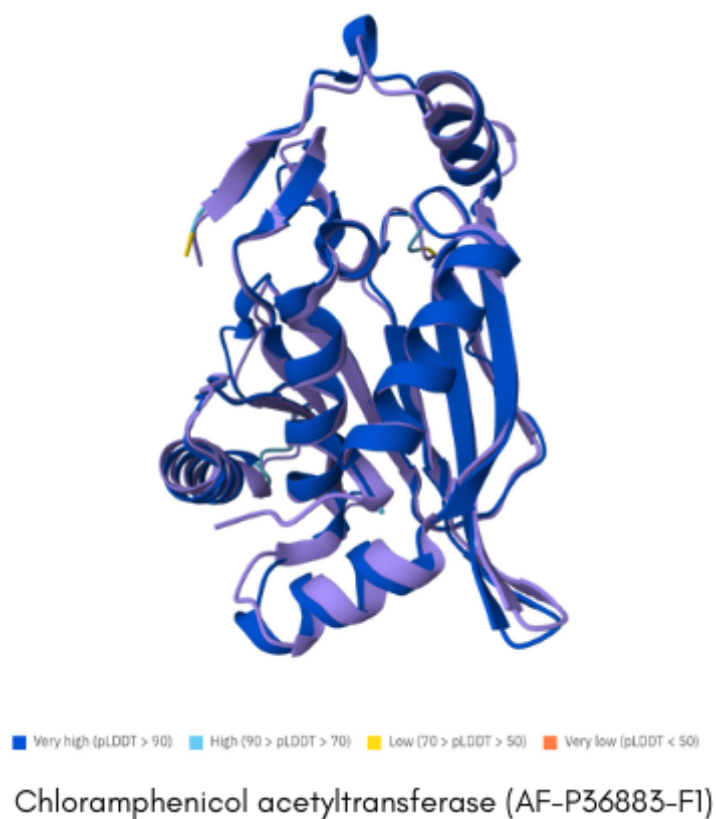
Vrednosti pLDDT veće od 90 se kategorišu kao najveća preciznost i tada se smatra da su i kičma i bočni lanci proteina predviđeni sa visokom preciznošću. Nasuprot tome, vrednosti ispod 70 tipično ukazuju na ispravno predviđenu kičmu, ali sa mogućim greškama u predikciji bočnih lanaca.

Važno je napomenuti da pLDDT vrednost može značajno varirati duž proteinskog

lanca. S obzirom da se ova mera odnosi se na pojedinačne regione strukture, određeni segmenti mogu biti predviđeni s visokom pouzdanošću, dok drugi mogu imati nisku prediktivnu tačnost.

Niska pLDDT vrednost u određenim oblastima proteina može biti posledica više faktora. Najčešći uzroci su visoka fleksibilnost regiona ili odsustvo dobro definisane strukture. Takođe, moguće je da oblast ima stabilnu strukturu, ali da nedostaju relevantni podaci neophodni za pouzdanu predikciju pomoću AlphaFold2 programa. U oba slučaja, ovakvim oblastima se obično pridružuje pLDDT vrednost ispod 50.

Na slici 6 prikazan je primer strukture proteina sa određenim pLDDT vrednostima.



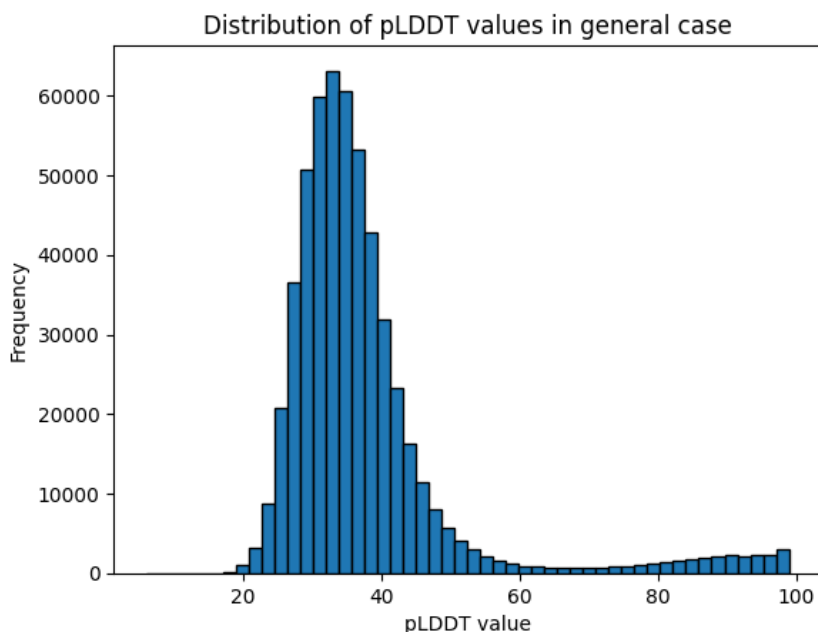
Slika 6: Prosečna pLDDT vrednost za protein Chloramphenicol acetyltransferase iznosi 96.28.

Glavni cilj ovog dela analize je upravo bio ispitivanje prirode vrednosti pLDDT mere u podacima.

3.4.1 Vrednosti pLDDT u opštem slučaju

Ispitivanje vrednosti pLDDT u opštem slučaju podrazumeva analiziranje vrednosti ove mere na celom skupu podataka. Ovakva analiza omogućava donošenje zaključaka o preciznosti predikcija lokalnih struktura proteina čiji su proteinski blokovi bili predmet intraživanja u ovom radu.

Radi utvrđivanja prirode posmatrane mere, najpre je konstruisan histogram vrednosti pLDDT. Prikaz histograma je dat na slici 7.



Slika 7: Histogram raspodele pLDDT vrednosti u celom skupu podataka

Raspodela pokazuje izraženu asimetričnost, sa dominantnim vrhom u intervalu 30–40 i dugim desnim repom. Ovakav oblik raspodele sugerše da može biti aproksimirana gama raspodelom ili nekom drugom asimetričnom raspodelom. Dodatno, primećuje se manja frekvencija pLDDT vrednosti u višem opsegu (iznad 70), što može ukazivati na retkost visoko pouzdanih predikcija u opštem slučaju.

Za preciznije ispitivanje raspodele, potrebno je izračunati osnovne deskriptivne statistike. Srednja vrednost iznosi približno 37.6, 0.75 kvantil 39.6, dok maksimalna pLDDT vrednost koja se javlja u skupu iznosi 98.95. Vrednost ostalih statistika dat je u tabeli 4.

Deskriptivna statistika	pLDDT
mean	37.586782
std	13.145442
min	5.920000
25%	30.540001
50%	34.570000
75%	39.570000
max	98.949997

Tabela 4: Osnovne deskriptivne statistike za pLDDT parametar.

Zaključak je da se u opštem slučaju najčešće javljaju veoma niske pLDDT vrednosti što ukazuje na nizak sveukupni kvalitet predviđenih lokalnih struktura proteina. Međutim, maksimalna vrednost pLDDT-a sugerise na to da za određene proteine postoje delovi strukture koji su predviđeni sa visokim stepenom preciznosti.

3.4.2 Vrednosti pLDDT prema proteinskim blokovima

3.4.3 Vrednosti pLDDT prema aminokiselinama

3.4.4 Vrednosti pLDDT prema sekundarnim strukturama

3.5 Analiza vrednosti RSA parametra

3.6 Ispitivanje zastupljenosti aminokiselina u podacima

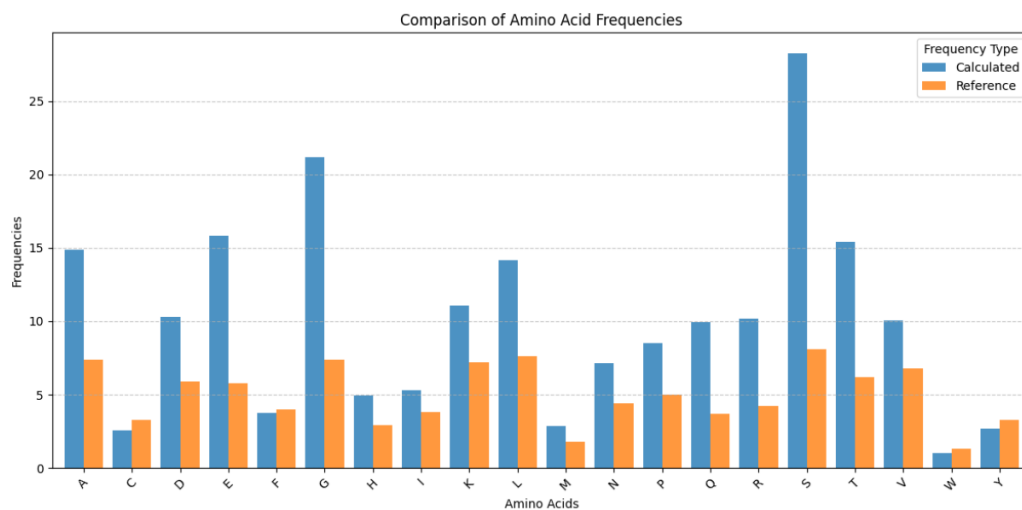
Kodon je triplet nukleotida, odnosno tri uzastopna nukleotida mRNK. Kodoni predstavljaju šifre za aminokiseline. Jedna aminokiselina može biti kodirana jednim kodonom ili sa više kodona. Aminokiseline se dalje kombinuju i kreiraju proteine. Postoje regioni u kodu proteina gde sastav aminokiselina nije od značaja, ali se neke aminokiseline pojavljuju više od drugih. Postavlja se pitanje da li je to slučajno ili postoji neki razlog koji objašnjava prezastupljenost određenih aminokiselina. Određivanje prezastupljenosti i podzastupljenosti aminokiselina vršeno je poređenjem frekvencije aminokiselina u podacima sa očekivanim vrednostima frekvencija. Očekivane vrednosti frekvencije kodona računaju se množenjem prirodnih frekvencija svake DNK baze koja čini kodon, gde se frekvencije u prirodi za uracil 22%, citozin 21.7%, guanin 26.1% i adenin 30.3%. Zatim se očekivana vrednost frekvencije aminokiselina dobija kao suma očekivanih vrednosti svih kodona koji kodiraju tu aminokiselinu. Kako postoje tri kodona koji ne kodiraju aminokiseline već su stop ili nonsense kodoni, prethodnu sumu treba pomnožiti faktorom korekcije 1.057.

Aminokiseline	Kodoni	Posmatrana zastupljenost kod kičmenjaka
Alanine	GCU, GCA, GCC, GCG	7.4%
Arginine	CGU, CGA, CGC, CGG, AGA, AGG	4.2%
Asparagine	AAU, AAC	4.4%
Aspartic Acid	GAU, GAC	5.9%
Cysteine	UGU, UGC	3.3%
Glutamic Acid	GAA, GAG	5.8%
Glutamine	CAA, CAG	3.7%
Glycine	GGU, GGA, GGC, GGG	7.4%
Histidine	CAU, CAC	2.9%
Isoleucine	AUU, AUA, AUC	3.8%
Leucine	CUU, CUA, CUC, CUG, UUA, UUG	7.6%
Lysine	AAA, AAG	7.2%
Methionine	AUG	1.8%
Phenylalanine	UUU, UUC	4.0%
Proline	CCU, CCA, CCC, CCG	5.0%
Serine	UCU, UCA, UCC, UCG, AGU, AGC	8.1%
Threonine	ACU, ACA, ACC, ACG	6.2%
Tryptophan	UGG	1.3%
Tyrosine	UAU, UAC	3.3%
Valine	GUU, GUA, GUC, GUG	6.8%
Stop Codons	UAA, UAG, UGA	—

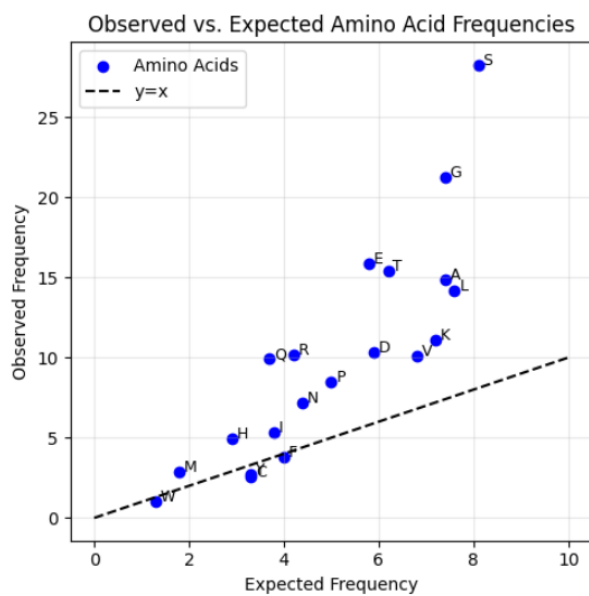
Tabela 5:

Posmatrane frekvencije aminokiselina izračunate su prebrojavanjem svih aminokiselina koje se javljaju u prelazima, računajući i atribut AA1 i atribut AA2, i njihovim deljenjem sa ukupnim brojem prelaza. U našoj analizi retkih prelaza između proteinskih blokova dobijeno je da je većina aminokiseline prezastupljena, jedine koje su blago podzastupljene su C, F, W, Y. Zanimljivo je da je prezastupljenost vrlo izražena kod nekih aminokiseline kao što su G i S je zastupljenost više nego duplo veća. (Slika 8)

Očekivano je da korelacije između očekivanih i posmatranih frekvencija bude visoka, ali u našem radu to nije slučaj. (Slika 9)



Slika 8: Poređenje izračunatih i referentnih zastupljenosti aminokiselina



Slika 9: Poređenje izračunatih i referentnih zastupljenosti aminokiselina

4 Klasifikacija

5 Klasterovanje

6 Zaključak

References

- [1] Phil Carter, Claus A. F. Andersen, Burkhard Rost. *DSSPcont: Continuous Secondary Structure Assignments for Proteins*. Bioinformatics, 19(2):230-231, 2003.
- [2] Hao Dong, Mukesh Sharma, Huan-Xiang Zhou, Timothy A. Cross. *Glycines: Role in α -Helical Membrane Protein Structures and a Potential Indicator for Native Conformation*. Biochemistry, 51(26): 5299-5307, 2012. PMCID: PMC3426646, NIHMSID: NIHMS383776, PMID: 22650985.
- [3] Agnel Praveen Joseph, Garima Agarwal, Swapnil Mahajan, Jean-Christophe Gelly, Lakshmipuram S. Swapna, Bernard Offmann, Frédéric Cadet, Aurélie Bornot, Manoj Tyagi, Hélène Valadié, Bohdan Schneider, Catherine Etchebest, Narayanaswamy Srinivasan, Alexandre G. de Brevern. *A short survey on protein blocks*. Biophysical Reviews, 2(3):137–147, 2010.