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

Sadržaj

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.

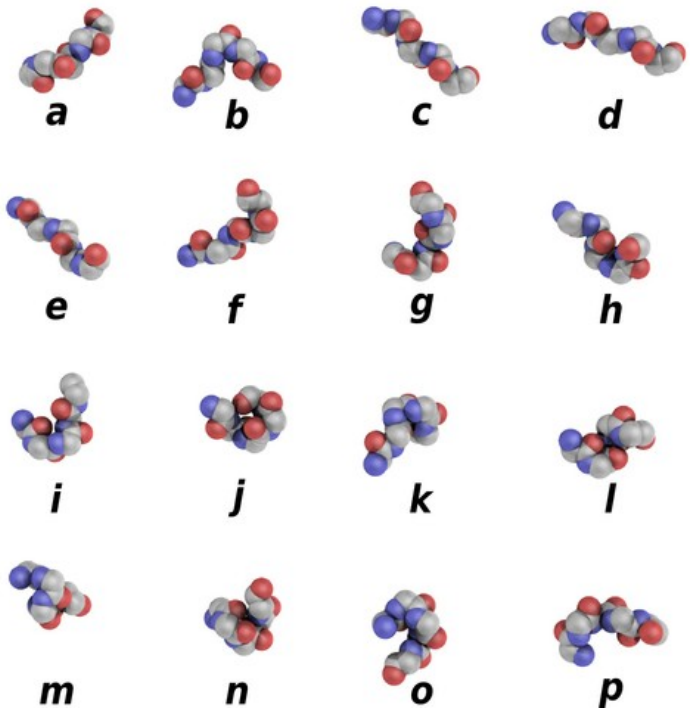
U nastojanju da se postigne precizniji, detaljniji i informativniji prikaz trodimenzionalne strukture proteina, razvijeni su proteinski blokovi. Kao jedni od najistaknutijih predstavnika strukturnih alfabeta, pokazano je da omogućavaju detekciju strukturne sličnosti između proteina sa izuzetnom efikasnošću.

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 procente.
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 trodimenionalne 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 upoređivanje 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.[3] Trodimenzionalna struktura može se opisati i korišćennjem aproksimativnih prototipova lokalne strukture proteina. Skup definisanih prototipova lokalne strukture se naziva i strukturni alfabet. Direktno određivanje trodimenizionalne 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 označenh slovima od a do p

Jedan od najpoznatijih strukturnih alfabeta 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. [7] (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-heliksu i beta-ravni. Proteinski blokovi od g do j odgovaraju nespecificnim strukturama (eng. coil).
Prevođenje u sekvencu proteinskih blokova kod proteina sa poznatom 3D strukturom odvija se tako što se svakom fragmentu uzastopnih aminokselina 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 atribute, 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 (*eng. α-helix*)
B	Beta-most (*eng. beta bridge*)
E	Prošireni beta list (*eng. extended beta sheet*)
G	3(10)-heliks
I	Pi-heliks (*eng. π-helix*)
T	Heliks okret (*eng. helix-turn*)
S	Zavojnica (*eng. bend*)
C	Nespecifična ili neorganizovana struktura (*eng. coil*)

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 izloženost aminokiselinskih ostataka rastvaraču.

3.1.1 Izdvajanje retkih prelaza između proteinskih blokova u podacima

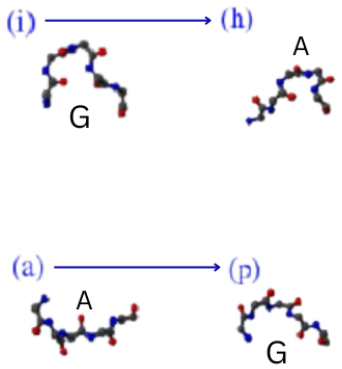
Glavni zadatak analize bio je proučavanje prelaza između proteinskih blokova pri čemu je akcenat stavljen na neočekivane, retke prelaze. Takvi prelazi su, zbog svoje prirode, od posebnog interesa za istraživanje jer mogu doprineti novim saznanjima o nizu proteinskih blokova koji opisuje trodimenzionalnu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

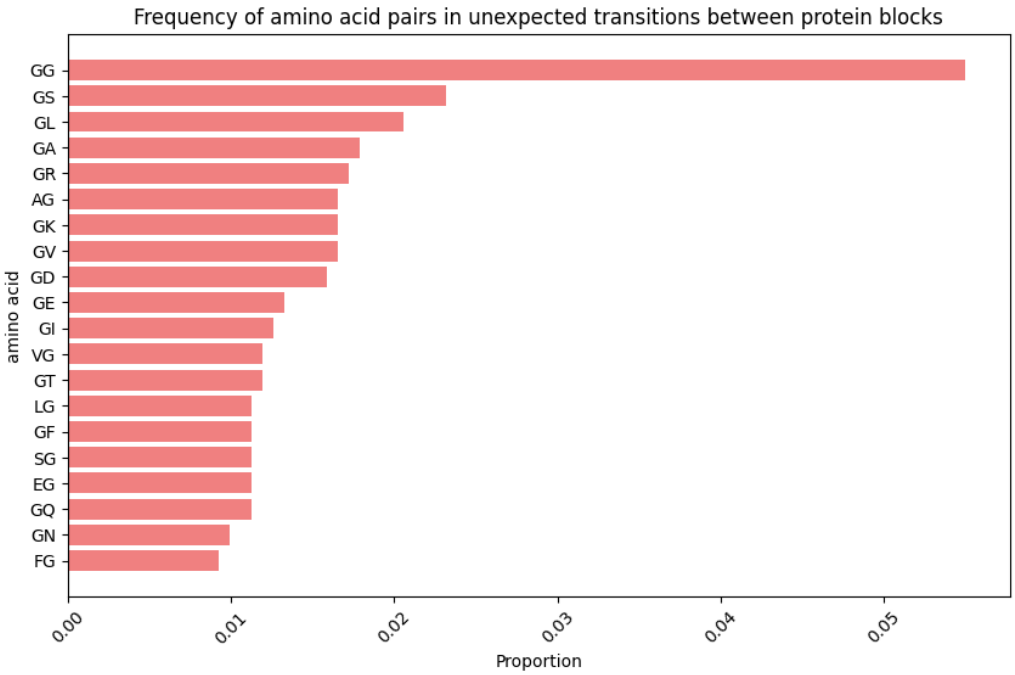
svaki par aminokiselina. Važno je napomenuti da redosled aminokiselina u paru nije irelevantan, zbog čega se parovi aminokiselina (A1, A2) ne mogu smatrati identičnim parovima (A2, A1). Kako bi ova tvrdnja bila intuitivnija i jasnija, potrebno je prvo objasniti zašto je prelaz određen sa dve aminokiseline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



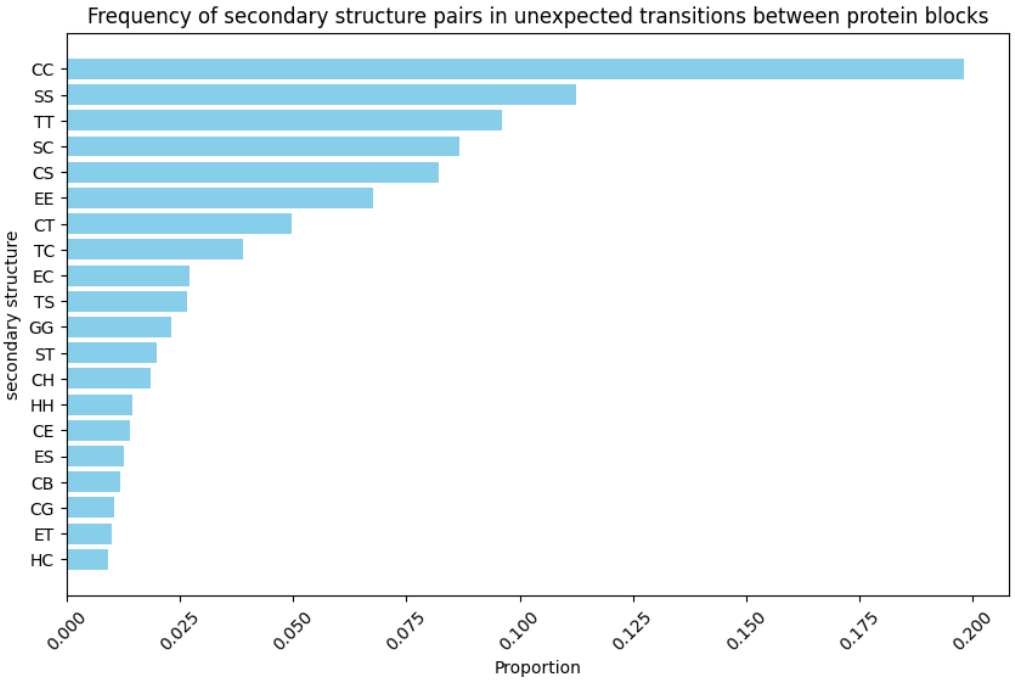
Slika 3: Frekvencija parova aminokiselina u retkim prelazima.

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.



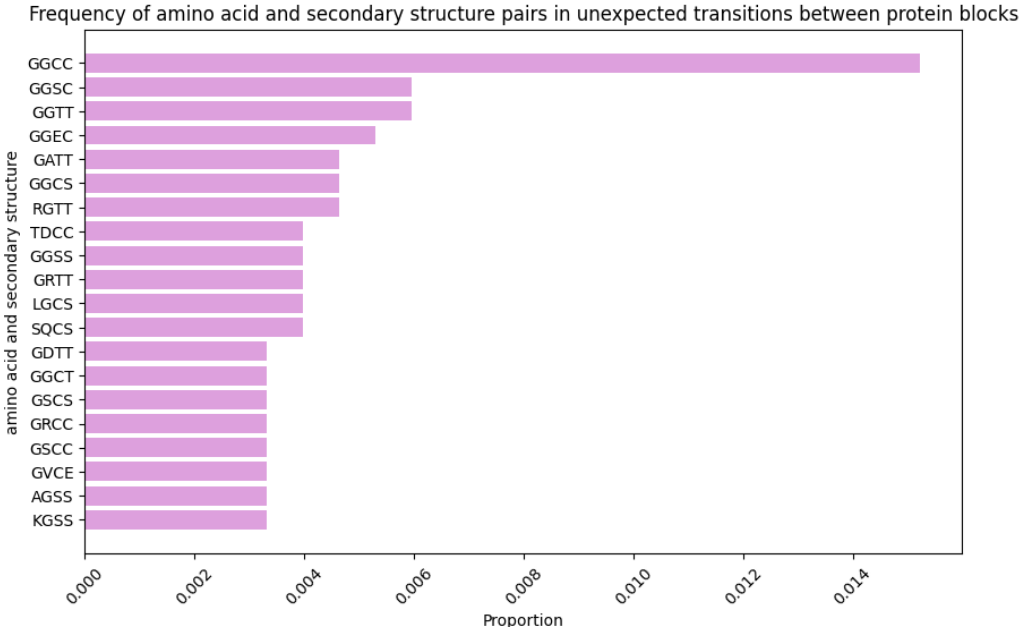
Slika 4: Frekvencija parova sekundarnih struktura u retkim prelazima.

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.



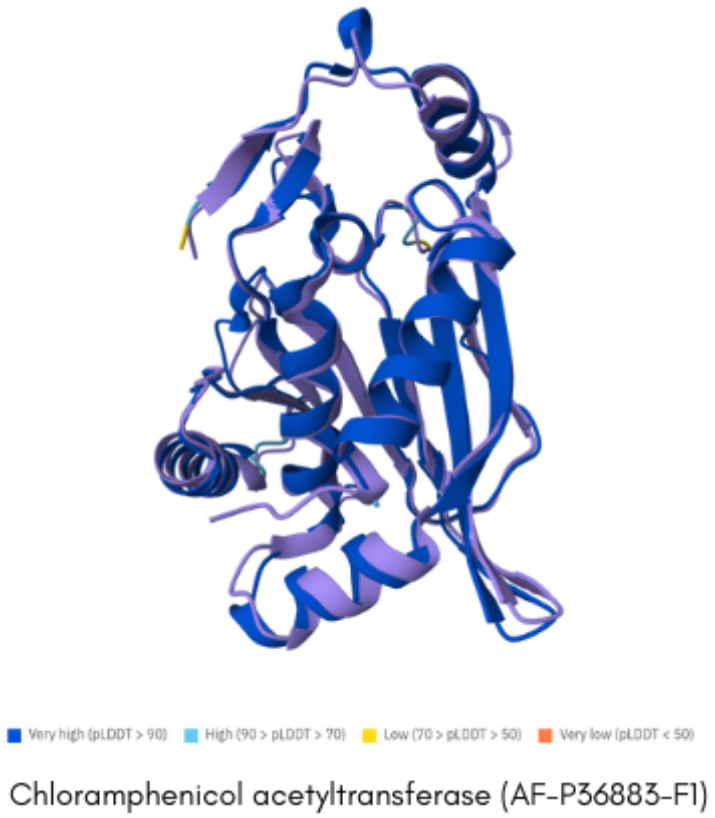
Slika 5: Frekvencija kombinacija parova aminokiselina i sekundarnih struktura u retkim prelazima.

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.

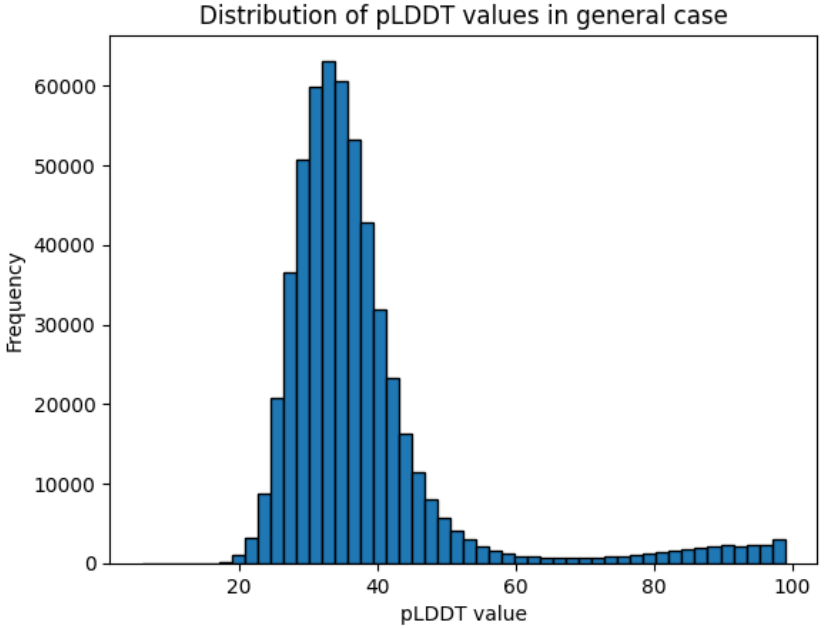


Slika 6: Prosečna pLDDT vrednost za protein Chloramphenicol acetyltransferase iznosi 96.28.

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 sugeriše 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. Ovaj rezultat može da sugeriše na to da analizu nije prikladno vršiti na osnovu celih redova u skupu podataka zbog potencijalne pristrasnosti vrednostima koje se javljaju samo u određenim prelazima, za određene proteine. Iz tog razloga, potrebno je grupisati podatke na osnovu nekog atributa i dalje vršiti analizu.

3.4.2 Vrednosti pLDDT prema proteinskim blokovima

Sledeći korak analize pLDDT-a jeste ispitivanje vrednosti te mere u odnosu na proteinske blokove. Ispitivanje je vršeno tako što su se podaci grupisali na osnovu parova proteinskih blokova. Nakon toga, izračunata je prosečna vrednost pLDDT mere. Na taj način dodeljena je reprezentativna pLDDT vrednost za svaki par proteinskih blokova.
Ideja je da se za određeni prelaz utvrdi da li se pretežno javljaju visoke ili niske vrednosti pLDDT-a i na taj način donese zaključak da li taj prelaz sadrži relativno precizno predviđene lokalne strukture proteina.
Da bi se dobila relevantna informacija o prirodi vrednosti pLDDT u odnosu na sve prelaze, izračunate su desktriptivne statistike za tako grupisane podatke. Pregled vrednosti statistika je dostupan u tabeli 5.

Prosečna vrednost od približno 63.45 ukazuje na to da, posmatrano u odnosu na parove proteinskih blokova, pLDDT vrednost teži visoko pouzdanoj. Takođe, izračunat je broj parova kod kojih se javlja visoka vrednost. Dobijeno je da se u 49 od 106, odnosno u 46.23%, retkih prelaza između proteinskih blokova dostiže precizna predikcija lokalne strukture proteina. U tabeli 6 dat je prikaz parova proteinskih blokova za koje se vezuju veoma visoke vrednosti pLDDT-a. U tim prelazima nisu prisutni proteinski blokovi g, h, ni kao par, ni pojedinačno, dok se i i j javljaju ali veoma retko i ne zajedno. S obzirom na to da oni odgovaraju nespecifičnim strukturama (eng. *coil–coil*), opravdano je da se za njih vezuju niže vrednosti pLDDT-a.

Deskriptivna statistika	pLDDT
mean	63.454415
std	18.023050
min	33.597163
25%	45.888131
50%	65.646074
75%	79.848490
max	90.566951

Tabela 5: Osnovne deskriptivne statistike za pLDDT parametar prema proteinskim blokovima.

PB1	PB2	pLDDT
n	k	90.57
f	o	90.30
p	n	88.73
k	p	88.16
i	m	87.34
m	j	87.13
d	n	85.99
e	n	85.81
k	i	85.79
e	l	85.53

Tabela 6: Parovi proteinskih blokova (PB1-PB2) sa najvišom prosečnom vrednošću pLDDT-a.

3.4.3 Vrednosti pLDDT prema aminokiselinama

U ovom delu analize podaci su grupisani na osnovu parova aminokiselina, izračunata je prosečna vrednost pLDDT mere za svaki par i na kraju određene deskriptivne statistike, date u tabeli 7.

Dobijeni rezultati ukazuju na veoma niske vrednosti posmatrano prema parovima aminokiselina. Maksimalna vrednost u skupu pokazuje da se ni za jedan par ne javljaju visoko pouzdane vrednosti, odnosno vrednosti iznad 70. Naime, razlog ovako niskih vrednosti leži u tome što ista kombinacija aminokiselina može da se pojavi u potpuno različitim delovima različitih proteina, i to u okruženjima sa različitom stabilnošću, fleksibilnošću i strukturnom definisanošću što dovodi do velike varijacije pLDDT vrednosti.

Deskriptivna statistika	pLDDT
mean	38.670799
std	3.802211
min	33.752673
25%	36.140094
50%	37.688579
75%	39.974417
max	61.994733

Tabela 7: Osnovne deskriptivne statistike za pLDDT parametar prema aminokiselinama.

3.4.4 Vrednosti pLDDT prema sekundarnim strukturama

U analizi pLDDT mere ostalo je još ispitati njene vrednosti u odnosu na parove sekudarnih struktura. Ispitivanje je vršeno identično kao u slučaju parova proteinskih blokova i aminokiselina - podaci su najpre grupisani, zatim je izračunata prošečna pLDDT vrednost i na kraju su izračunate vrednosti statistika za novoformirani skup podataka koje su date u tabeli 8.

Deskriptivna statistika	pLDDT
mean	73.448625
std	15.828469
min	35.300866
25%	65.967828
50%	78.462409
75%	85.288241
max	95.586667

Tabela 8: Osnovne deskriptivne statistike za pLDDT parametar prema sekundarnim strukturama.

Vrednosti desktriptivnih statistika, pre svega prosečna vrednost od približno 73.45 i medijana od približno 78.46, ukazuju na visoku pouzdanost u ovom slučaju. Takođe, čak 68.97%, odnosno 40 od 58, parova sekundarnih struktura ima pLDDT vrednost iznad 70. Može se zaključiti da većinu parova, u prelazima čiji su deo, odlikuje precizna predikcija strukture. U tabeli 9 izvojeni su oni parovi kod kojih se javljaju najviše pLDDT vrednosti.

S2_1	S2_2	pLDDT
B	E	95.59
B	H	95.20
G	E	92.51
H	E	92.51
T	B	88.82
S	E	88.16
E	C	87.04
S	B	86.99
G	H	86.99
E	E	86.85

Tabela 9: Parovi sekundarnih struktura (S2_1 - S2_2) sa najvišom prosečnom vrednošću pLDDT-a.

Najpouzdaniji parovi sekundarnih struktura, odnosno parovi čija pLDDT vrednost prelazi 90, uglavnom uključuju elemente B, E i H, što je očekivano jer upravo ovi oblici strukture predstavljaju najstabilnije i najčešće strukturne oblike u proteinima. Odnosno, sasvim je prirodno pretpostaviti da će AlphaFold2 program moći baš te najosnovnije sekundarne strukture da predvidi precizno.

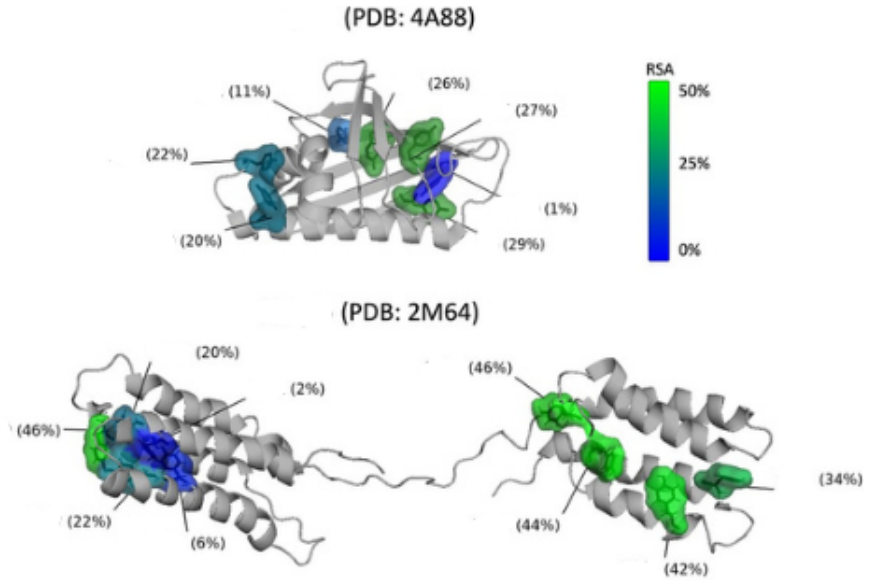
3.5 Analiza vrednosti RSA parametra

Pored pLDDT parametra, od posebnog interesa za analizu je bio i parametar RSA (eng. *relative solvent accessibility*). RSA predstavlja meru izloženosti aminokiselinskog ostatka rastvaraču (vodi) u strukturi proteina i koristi se za opisivanje biofizičkih karakteristika proteina. Vrednost RSA za aminokiselinski ostatak računa se na osnovu formule:

$$\mathrm{RSA} = \frac{\mathrm{ASA}}{\mathrm{maxASA}} \tag{1}$$

gde je *ASA* (eng. *accessibility surface area*) pristupačna površina rastvaraču, a *maxASA* maksimalna moguća pristupačna površina rastvaraču za datu aminokiselinu.
Na osnovu vrednosti RSA, odnosno izračunatoj izloženosti rastvaraču, aminokiselinski ostaci mogu biti „zakopani" (eng. *buried*) ili „izloženi" (eng. *exposed*). Definicije praga za klasifikaciju aminokiselina u jednu od dve kategorije variraju, međutim, u većini slučajeva prag se definiše kao 25% vrednosti RSA, odnosno ukoliko aminokiselinski ostatak ima vrednost RSA manju od 25% onda se može zaključiti da on nije izložen rastvaraču (zakopan), u suprotnom je izložen [4]. S obzirom na to da je prag relativno mali, potrebno

je razumeti da vrednosti iznad praga označavaju različite stepene izloženosti rastvaraču. Naime, vrednosti iznad 90% ukazuju na to da je ostatak gotovo čitavom površinom dostupan rastvaraču i istraživanja pokazuju da se takve konformacije najčešće javljaju u petljama (eng. *loops*) i okretima (*turns*) [5], dok vrednosti bliže 25% označavaju blagu izloženost tj. izloženost manjih delova površine ostatka. Na slici 8 prikazane su 3D strukture proteina proteina glutamin sintaze i K-Ras-a sa izračunatim RSA vrednostima za neke aminokiselinske ostatke.



Slika 8: RSA vrednosti za aminokiselinske ostatke u proteinima.

3.5.1 Vrednosti RSA u opštem slučaju

Radi utvrđivanja RSA vrednosti u skupu podataka izračunate su deskriptivne statistike i određen je broj izloženih i zakopanih aminokiselinskih ostataka. U podacima postoje dve kolone koje se odnose na RSA jer se svaki prelaz između proteinskih blokova sastoji od para aminokiselina za koji je određena RSA vrednost. U skladu sa tim, prilikom računanja brojnosti svake od RSA kategorija sabrane su vrednosti dobijene za prvu i drugu kolonu.
Za obe kolone dobijena je srednja vrednost od približno 92%, dok medijana i treći kvartil iznose 100%. Takođe, dobijeno je da je 97.5% aminokiselinskih ostataka, posmatrano u opštem slučaju, izloženo. Zaključak je da su u podacima daleko zastupljenije izložene aminokiseline i to visoko izložene, odnosno tako da su gotovo čitavom površinom dostupne rastvaraču.

3.5.2 Vrednosti RSA prema proteinskim blokovima

U ovom delu analize, podaci najpre grupisani na osnovu prelaza između proteinskih blokova, analogno pristupu koji je korišćen u analizi parametra pLDDT. Budući da za svaki prelaz postoje dve RSA vrednosti (po jedna za svaki od dva susedna ostatka), izračunata je njihova prosečna vrednost, koja je potom dodeljena kao reprezentativna vrednost za dati prelaz. Nad tako grupisanim podacima izračunate su deskriptivne statistike, čije vrednosti su date u tabeli 10.

Deskriptivna statistika	**RSA**
mean	64.219617
std	20.352375
min	22.983397
25%	48.298345
50%	59.209232
75%	83.560475
max	96.876662

Tabela 10: Osnovne deskriptivne statistike za RSA parametar prema proteinskim blokovima.

Rezultati ukazuju da su aminokiselinski ostaci, posmatrano za svaki pojedinačni prelaz, srednje izloženi rastvaraču odnosno da se nalaze u umereno izloženim regijama strukture. Posebno je značajan podatak da se kod čak 99.06% svih prelaza beleži prosečna RSA vrednost iznad 25%, što sugeriše da se potpuno zakopane aminokiseline praktično ne javljaju kao karakteristika celih parova

proteinskih blokova, već eventualno samo lokalno, na nivou pojedinačnih ostataka.

3.5.3 Vrednosti RSA prema aminokiselinama

Za potrebe detaljnije analize, izvršeno je i ispitivanje prirode vrednosti RSA prema parovima aminokiselina u prelazima. Primenjena metodologija je identična kao u prethodnoj analizi – podaci su grupisani prema svim mogućim parovima aminokiselina koji se javljaju u prelazima, a zatim je za svaki par izračunata prosečna vrednost RSA. Na tako grupisanim podacima izračunate su osnovne deskriptivne statistike prikazane u Tabeli 11.

Deskriptivna statistika	**RSA**
mean	89.822263
std	5.541973
min	65.555131
25%	65.555131
50%	90.880001
75%	93.854497
max	98.135047

Tabela 11: Osnovne deskriptivne statistike za RSA parametar prema aminokiselinama.

Rezultati pokazuju izuzetno visoku prosečnu izloženost aminokiselinskih parova, sa srednjom vrednošću od približno 90%, što znači da se velika većina analiziranih parova nalazi u izloženim regijama proteinskih struktura. Čak i prvi kvartil iznosi 65.56%, što implicira da ni u jednom od posmatranih parova aminokiselina ne dolazi do značajnije zaklonjenosti površine.

3.5.4 Vrednosti RSA prema sekundarnim strukturama

Na kraju ove analize, ispitana je pristupačnost aminokiselinskih ostataka rastvaraču u kontekstu parova sekundarnih struktura koje definišu prelaz. Kao i u prethodnim slučajevima, podaci su grupisani na osnovu para, izračunata je prosečna RSA vrednost i određene su potrebne statistike za analiziranje prirode RSA vrednosti. Rezultati su dati u tabeli 12.

Dobijeni rezultati ukazuju na širi raspon izloženosti aminokiselinskih ostataka u zavisnosti od kombinacije sekundarnih struktura u prelazu. Srednja vrednost i medijana impliciraju umerenu izloženost u većini slučajeva.
Zanimljivo je da je 58.62% svih analiziranih parova sekundarnih struktura imalo prosečnu RSA vrednost veću od 25%, što ukazuje da je više od polovine

Deskriptivna statistika	RSA
mean	41.925664
std	30.019328
min	0.000000
25%	16.518614
50%	43.529417
75%	62.653578
max	97.473198

Tabela 12: Osnovne deskriptivne statistike za RSA parametar prema sekundarnim strukturama.

parova locirano u regijama koje su bar delimično dostupne rastvaraču. S druge strane, prisustvo prelaza sa RSA vrednostima bliskim nuli ukazuje na to da neki tipovi sekundarnih struktura uključuju strukturno zaklonjene pozicije.

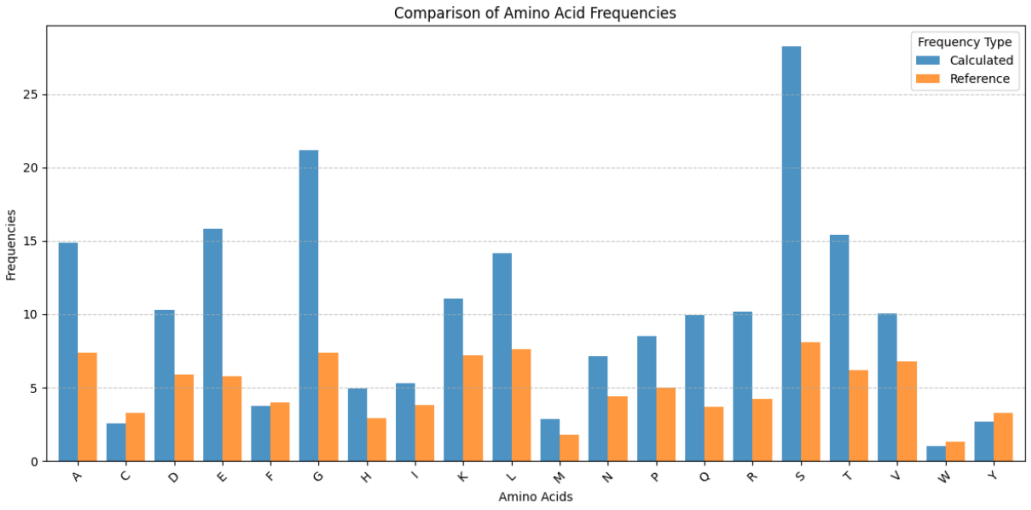
3.6 Ispitivanje zastupljenosti aminokiselina u podacima

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, i gde 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 aminokiseline 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 adeninn 30.3%. Zatim se očekivana vrednost frekvencije aminokiseline 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. [6]

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 proteinski 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 aminokisline kao što su G i S je zastupljenost više nego duplo veća. (Slika 9)

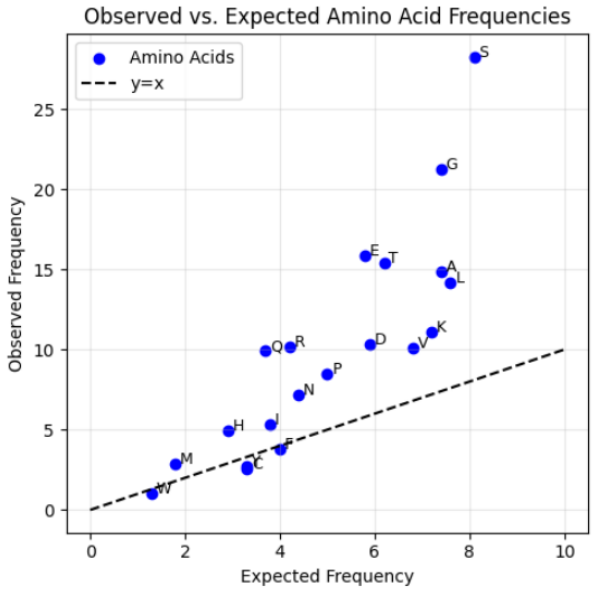
Aminokiseline	Kodoni	Posmatrana zastupljenost kod kičmenjaka
Alanine	GCU, GCA, GCC, GCG	7.4%
Arginine	CGU, CGA, CGC,CGG, AGA, AGG	4.2%
Asparagine	AAU, AAC	4.4%
Aspartic Acid	GAU, GAC	5.9%
Cysteine	UGU, UGC	3.3%
Glutamic Acid	GAA, GAG	5.8%
Glutamine	CAA, CAG	3.7%
Glycine	GGU, GGA, GGC, GGG	7.4%
Histidine	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 13:



Slika 9: Poređenje izračunatih i referentnih zastupljenosti aminokiselina

Očekivano je da korelacije između očekivanih i posmatranih frekvencija bude visoka, ali u našem radu to nije slučaj.(Slika 10)



Slika 10: Poređenje izračunatih i referentnih zastupljenosti aminokiselina

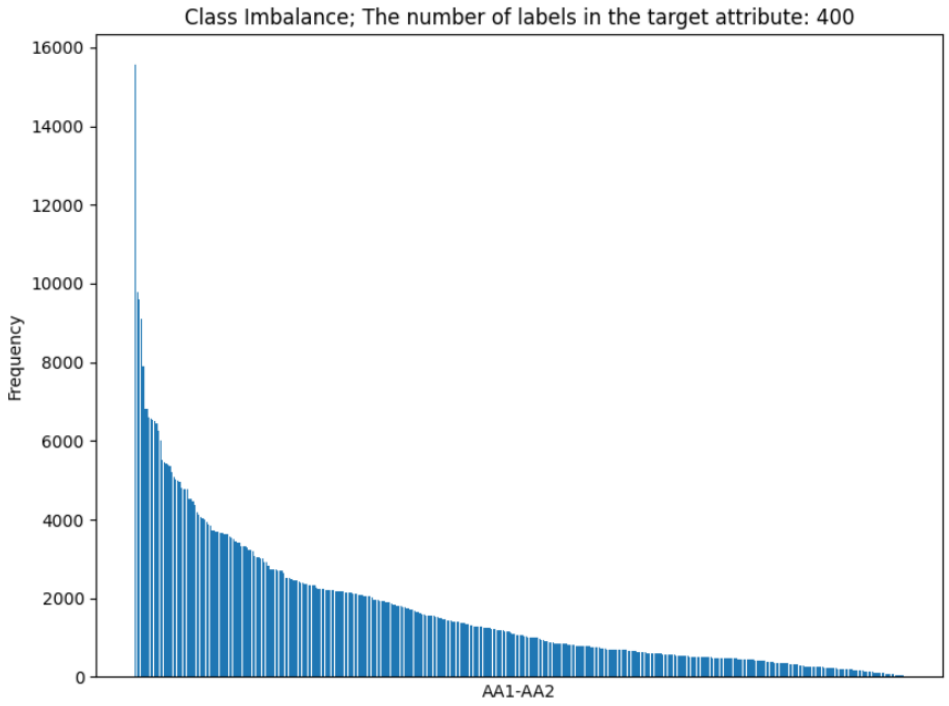
4 Klasifikacija

Algoritmi klasifikacije se u mašinskom učenju koriste za predviđanje ciljnog atributa, klase, na osnovu ostalih atributa. U našem primeru vršimo klasifikaciju za tri različite ciljne promenljive, parove aminokiselina, parove sekundarnih struktura i parove proteinskih blokova na prelazima.

Kako se bavimo prelazima i parovima na njima najpre je neophodno da izvršimo konkatenaciju određenih kolona i napravimo novu kolonu koja će služiti kao ciljna promenljiva. Vrednosti iz prve i druge kolone su konkatenirane uz očuvanje redosleda (npr. 'jb' nije isto što i 'bj').

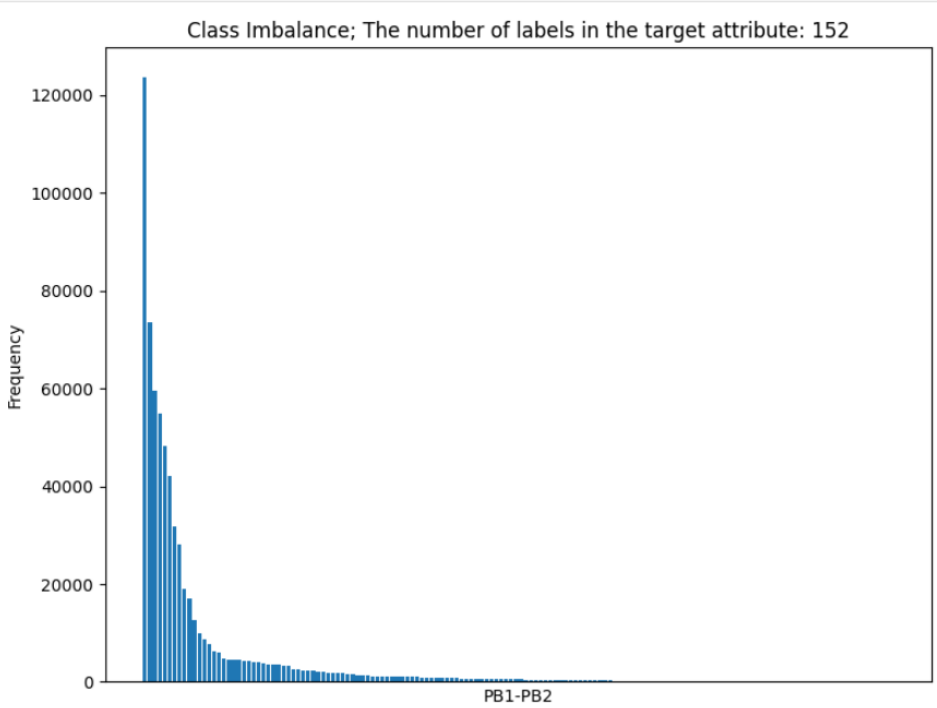
4.1 Balansiranost klasa

Kod problema klasifikacije podataka važno je proveriti balansiranost klasa, odnosno da li se sve klase ravnomerno javljaju ili postoje neke koje su zastupljenije. U slučaju da postoje klase koje su značajno brojnije od drugih može doći do pristrasnosti tim klasama i zanemarivanju ređih klasa, što dovodi do lošije preciznosti modela.



Slika 11: Balansiranost klasa za parove aminokiselina.

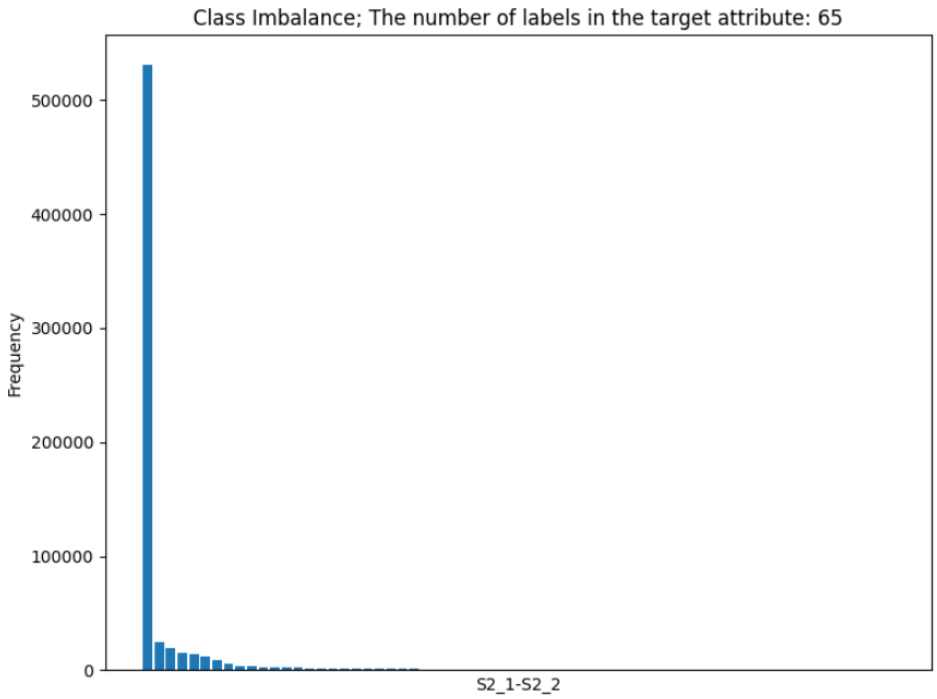
Kako je postoji 20 aminokiselina, a mi predviđamo parove aminokiselina, ukupan broj klasa je 20*20 = 400. Na slici 11 predstavljena je raspodela klasa aminokiselina. Kao što se može videti na slici, distribucija učestalosti ovih klasa pokazuje izrazitu nebalansiranost. Većina primera pripada manjem broju najčešćih klasa, dok veliki broj kombinacija ima relativno mali broj instanci, pa čak i ispod 100 ponavljanja. Posebno je izazovno u ovom slučaju to što se radi o velikom broju klasa (400), što dodatno otežava balansiranje i evaluaciju.



Slika 12: Balansiranost klasa za parove proteinskih blokova.

U slučaju kada je ciljna promenljiva kombinacije parova proteinskih blokova, kojih u skupu podataka ima ukupno 152. Histogram prikazuje izrazitu nebalansiranost u raspodeli ovih klasa (Slika 12). Nekoliko najfrekventnijih kombinacija pojavljuje se i preko 100.000 puta, dok se velika većina ostalih klasa javlja sa znatno manjom učestalošću, od nekoliko stotina do čak ispod 100.

U slučaju kombinacija sekundarnih struktura, pri čemu je ukupno identifikovano 65 različitih kombinacija, vizuelna analiza raspodele klasa (prikazana na slici 13) ukazuje na ekstremnu nebalansiranost. Jedna jedina klasa dominira skupom sa preko 500.000 instanci, dok su ostale klase višestruko manje



Slika 13: Balansiranost klasa za parove sekundarnih struktura.

zastupljene — mnoge sa frekvencijom manjom od 10.000, a neke i blizu nule.

Ovakva distribucija je tipična za biološke podatke, gde određene strukturne konfiguracije imaju prirodno dominantniju ulogu. Međutim, iz perspektive učenja modela, ovolika razlika u broju primera po klasi može izazvati značajnu pristrasnost prilikom treniranja klasifikatora, jer dominantna klasa može preuzeti "težinu" u učenju i prouzrokovati značajnu pristrasnost modela, čime se tačnost predikcija za ređe klase drastično smanjuje. U takvom okruženju, model koji samo predviđa najčešću klasu može imati visoku ukupnu tačnost, ali loše performanse na nivou klasa, što se posebno vidi u metrikama kao što su *recall*.

4.2 CatBoost

CatBoost je savremeni algoritam za mašinsko učenje baziran na gradijentnom buđenju stabala (eng. gradient boosting on decision trees), razvijen od strane kompanije Yandex. Osmišljen je sa ciljem da poboljša efikasnost i tačnost u radu sa strukturisanim podacima, posebno u scenarijima gde su prisutne

i numeričke i kategorijske promenljive. Za razliku od drugih popularnih implementacija kao što su XGBoost i LightGBM, CatBoost ima ugrađenu podršku za direktno korišćenje kategorijskih promenljivih bez potrebe za njihovom prethodnom transformacijom u numerički format (npr. putem one-hot enkodiranja), čime se pojednostavljuje priprema podataka i smanjuje mogućnost preprilagođavanja.

Prilikom treniranja modela korišćen je parametar `auto_class_weights='Balanced'`, koji omogućava CatBoost algoritmu da automatski izračuna težine za svaku klasu na osnovu njihove učestalosti u skupu podataka. Ova tehnika se koristi za ublažavanje uticaja nebalansiranosti klasa, pri čemu ređe klase dobijaju veću težinu tokom treniranja, čime se smanjuje pristrasnost modela prema češće zastupljenim klasama.

Težina w_i za svaku klasu i računa se po sledećoj formuli:

$$w_i = \frac{N}{k \cdot N_i}$$

gde je:

- N — ukupan broj primera u skupu podataka,
- k — broj različitih klasa,
- N_i — broj primera koji pripadaju klasi i.

Aminokiseline

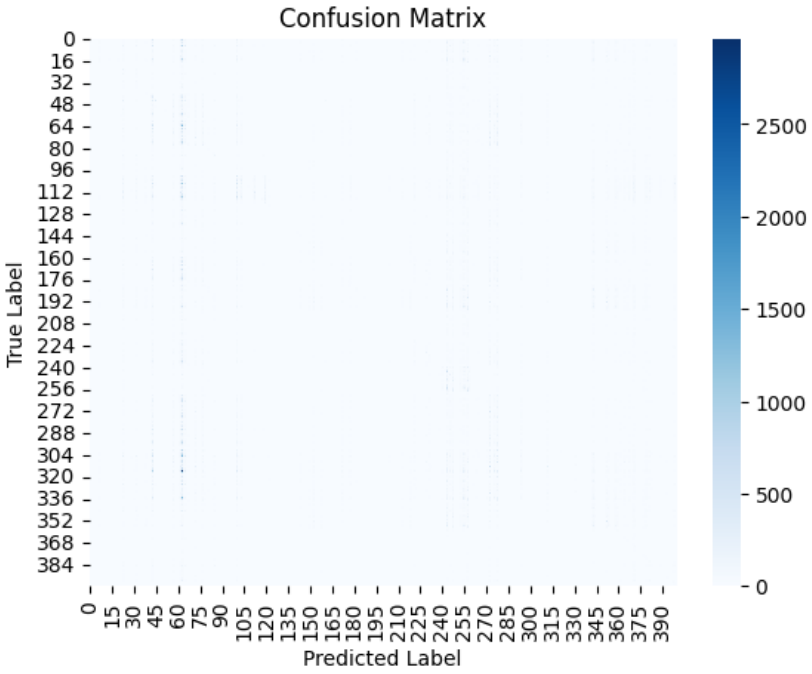
Rezultati pokazuju izuzetno nisku tačnost, odziv (recall) i F1-meru na test i trening skupu, dok je preciznost (precision) relativno visoka ($\sim$0.63). Ovo može ukazivati na neravnotežu klasa u skupu podataka, gde model nauči da klasifikuje samo dominantnu klasu.

- **Tačnost (Accuracy)** je niska na oba skupa ($\sim$2–2.5%), što znači da je model uspešan samo u malom broju slučajeva.
- **Odziv (Recall)** je takođe nizak, što znači da model propušta većinu pozitivnih primera.
- **Preciznost (Precision)** je relativno visoka — kada god model klasifikuje neku instancu kao pozitivnu, uglavnom je u pravu, ali to radi vrlo retko.

S obzirom da su metrike na trening i test skupu veoma slične, možemo zaključiti da nema značajnog preprilagođavanja. Model se ponaša jednako loše na oba skupa.

Tabela 14: Evaluacija modela na test i trening skupu za parove aminokiselina

Metrička vrednost	**Test skup**	**Trening skup**
Tačnost (Accuracy)	0.0227	0.0255
Preciznost (Precision)	0.6285	0.6292
Odziv (Recall)	0.0227	0.0255
F1 mera (F1 Score)	0.0103	0.0122



Slika 14: Matrica konfuzije za parove aminokiselina.

Proteinski blokovi

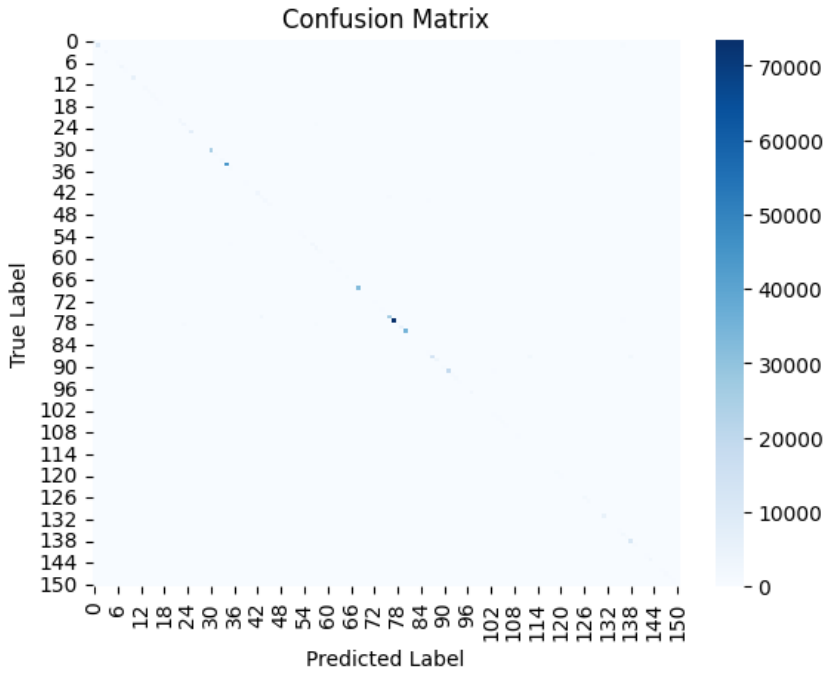
Rezultati pokazuju da model ostvaruje visoku tačnost, odziv, F1-meru i preciznost na oba skupa, što ukazuje na efikasnu klasifikaciju proteinskih blokova. Vrednosti metrika su gotovo identične na test i trening skupu, što znači da nema izraženog preprilagođavanja.

- **Tačnost (Accuracy)** je visoka ($\sim$94.2%) na oba skupa, što ukazuje na generalno uspešnu klasifikaciju većine uzoraka.
- **Preciznost (Precision)** i **odziv (Recall)** su uravnoteženi, što znači da model uspešno prepoznaje većinu pozitivnih primera, uz minimalan broj lažnih pozitivnih.

Tabela 15: Evaluacija modela na test i trening skupu za parove proteinskih blokova

Metrička vrednost	**Test skup**	**Trening skup**
Tačnost (Accuracy)	0.9418	0.9427
Preciznost (Precision)	0.9606	0.9613
Odziv (Recall)	0.9418	0.9427
F1 mera (F1 Score)	0.9472	0.9481

- **F1 mera** potvrđuje ravnotežu između odziva i preciznosti.



Slika 15: Matrica konfuzije za parove proteinskih blokova.

Sekundarne strukture

Prikazane metrike modela na test i trening skupu ukazuju na umereno dobar učinak, sa značajnim neskladom između preciznosti i odziva (recall).

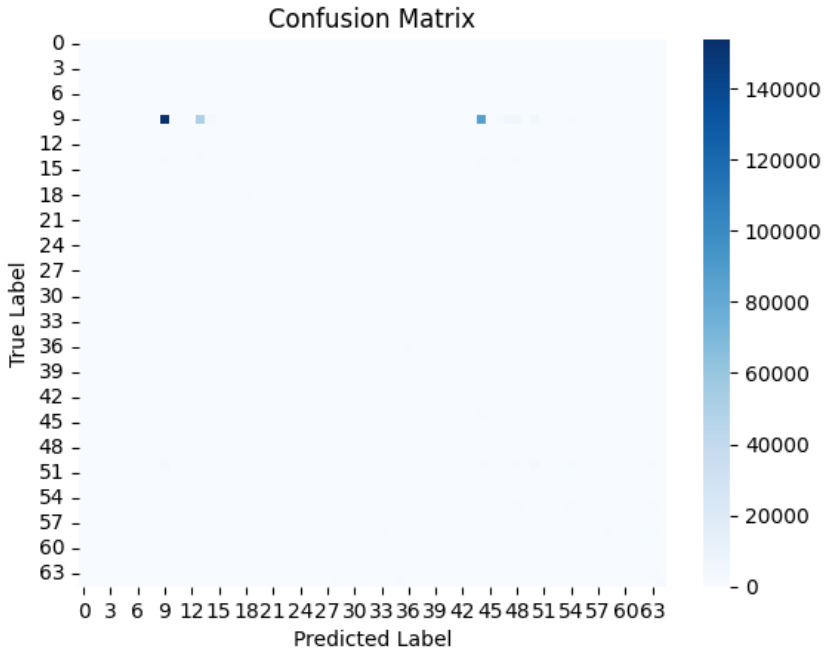
Rezultati pokazuju da model ima visoku preciznost (0.85), što znači da je većina klasifikacija pozitivnih primera tačna, ali je odziv (recall) nizak (0.43), što ukazuje da model propušta veliki broj stvarnih pozitivnih primera.

Tabela 16: Evaluacija modela na test i trening skupu za parove

Metrička vrednost	**Test skup**	**Trening skup**
Tačnost (Accuracy)	0.4251	0.4252
Preciznost (Precision)	0.8489	0.8488
Odziv (Recall)	0.4251	0.4252
F1 mera (F1 Score)	0.5540	0.5536

- **Tačnost (Accuracy)** od oko 42.5% ukazuje na umereno dobru ukupnu preciznost klasifikacije.
- **Nesklad između visoke preciznosti i niskog odziva** može ukazivati na neuravnoteženost klasa ili na to da model često izostavlja pozitivne slučajeve.
- **F1 mera** od oko 0.55 pokazuje da postoji prostor za poboljšanje u balansu između preciznosti i odziva.

S obzirom da su metrike na test i trening skupu veoma slične, nema indikacija o značajnom preprilagođavanju.



Slika 16: Matrica konfuzije za parove sekundarnih struktura.

4.3 Neuronske mreže

Neuronske mreže predstavljaju klasu modela koji se inspirišu strukturom i funkcionisanjem bioloških neuronskih sistema. Njihova moć ogleda se u sposobnosti da modeluju kompleksne i nelinearne odnose u podacima, što ih čini pogodnim za širok spektar zadataka u klasifikaciji, regresiji i obradi nestrukturiranih podataka. Posebno su uspešne u višeklasnoj klasifikaciji, pod uslovom da je ulazni prostor pravilno pripremljen.

Priprema podataka

Ciljna promenljiva, koja je inicijalno sadržila kategoričke oznake klasa, transformisana je u numeričke vrednosti korišćenjem klase `LabelEncoder` iz biblioteke `scikit-learn`. Ova transformacija omogućava jednoznačno mapiranje svake klase na pripadajuću celobrojnu oznaku.

Ulazne promenljive podeljene su na numeričke i kategoričke kolone. Kategoričke promenljive su kodirane tehnikom *One-Hot* kodiranja, gde se svaka diskretna vrednost reprezentuje binarnim vektorom. Numeričke kolone su standardizovane tako da imaju nultu srednju vrednost i jediničnu standardnu devijaciju, čime se obezbeđuje ujednačen doprinos svih numeričkih osobina.

Zbog izražene nebalansiranosti među klasama ciljne promenljive, bilo je neophodno primeniti metod za balansiranje. Korišćena je funkcija `compute_class_weight` sa parametrom `'balanced'`, koja automatski izračunava težine svake klase inverzno proporcionalno njenoj učestalosti u skupu za trening. Dobijene težine se kasnije prosleđuju modelu kao dodatni parametar kako bi se kompenzovala prekomerna zastupljenost dominantnih klasa i omogućilo pravičnije učenje.

Formiranje modela

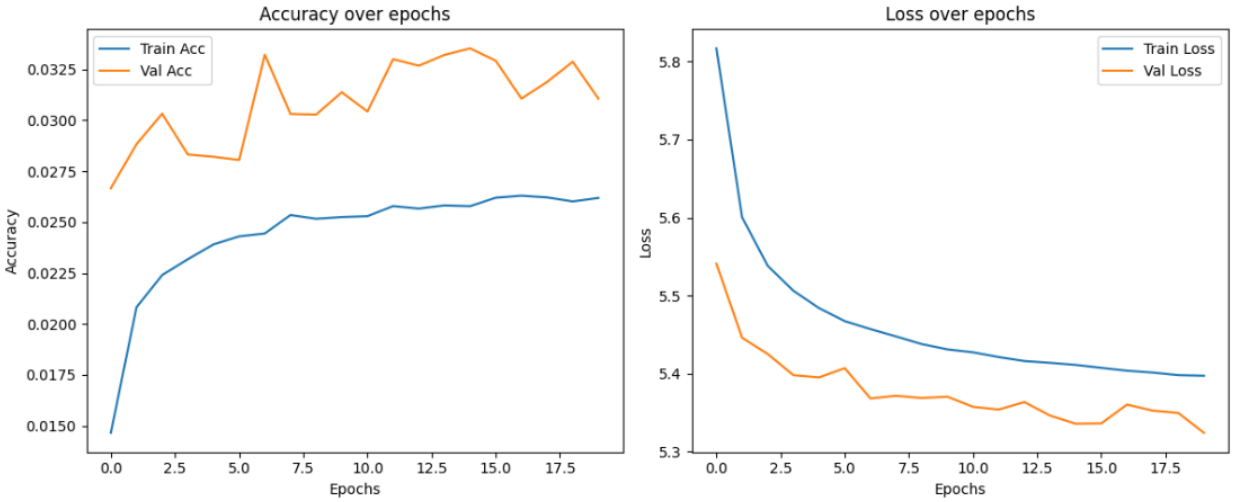
Model veštačke neuronske mreže konstruisan je korišćenjem biblioteke `Keras`, sa ciljem klasifikacije podataka u više klasa. Arhitektura modela sastoji se od dva skrivena sloja, pri čemu prvi sloj sadrži 128, a drugi 64 neurona, oba sa ReLU aktivacionom funkcijom. Radi smanjenja mogućnosti preprilagođavanja (*overfitting*), između slojeva su uključeni Dropout slojevi sa stopom izostavljanja od 30%. Izlazni sloj koristi `softmax` aktivacionu funkciju, što omogućava višeklasnu klasifikaciju.

Model je kompajliran korišćenjem Adam optimizatora sa unapred definisanom brzinom učenja, dok je funkcija greške bila `categorical_crossentropy`, što je standardni izbor za probleme višeklasne klasifikacije. Kao mera uspešnosti modela korišćena je tačnost (*accuracy*).

Radi poboljšanja performansi modela, sprovedena je ručna pretraga hiperparametara (engl. *manual grid search*). Istraživane su različite vrednosti broja neurona

u skrivenim slojevima, stope izostavljanja (dropout), kao i vrednosti stope učenja. Evaluacija svakog modela u okviru pretrage vršena je na osnovu tačnosti na validacionom skupu, čime je odabran skup hiperparametara koji daje najbolje rezultate.

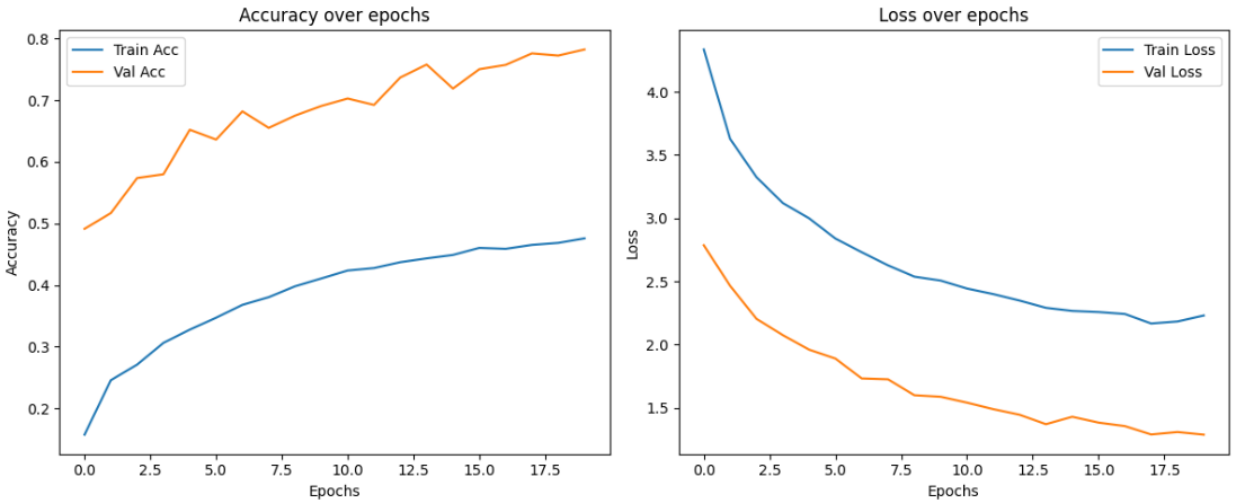
Aminokiseline



Slika 17: Prikaz preciznosti (levo) i gubitka (desno) tokom epoha za najbolji model pronađen Grid pretragom

Na slici 17 prikazana je promena tačnosti i funkcije greške modela tokom 20 epoha treniranja. Može se uočiti da se tačnost modela na trening skupu postepeno povećava, dok se vrednost funkcije greške smanjuje, što ukazuje na uspešno učenje modela. Takođe, validacioni rezultati pokazuju sličan trend: preciznost na validacionom skupu ostaje viša od one na trening skupu, što sugeriše da model generalizuje dobro i ne pokazuje jasne znake preprilagođavanja (engl. *overfitting*) u ovom broju epoha.

Ipak, apsolutne vrednosti preciznosti su niske (ispod 4%), što ukazuje na izazovnost klasifikacionog problema i moguću potrebu za dodatnim unapređenjem modela, kao što su optimizacija arhitekture, izbor relevantnijih osobina ili balansiranje skupa podataka na drugi način. S obzirom na veliku neravnotežu među klasama i veliki broj klasa, ovakvi rezultati su očekivani i zahtevaju dublju analizu distribucije podataka.



Slika 18: Preciznost (levo) i funkcija greške (desno) tokom epoha za klasifikaciju proteinskih blokova

Proteinski blokovi

Na slici 18 prikazani su rezultati treniranja neuronske mreže za klasifikaciju proteinskih blokova. Model pokazuje jasan trend poboljšanja kako u pogledu tačnosti, tako i u pogledu smanjenja funkcije greške. Tačnost na trening skupu konstantno raste, dok se vrednosti funkcije greške smanjuju, što je očekivano ponašanje tokom treniranja.

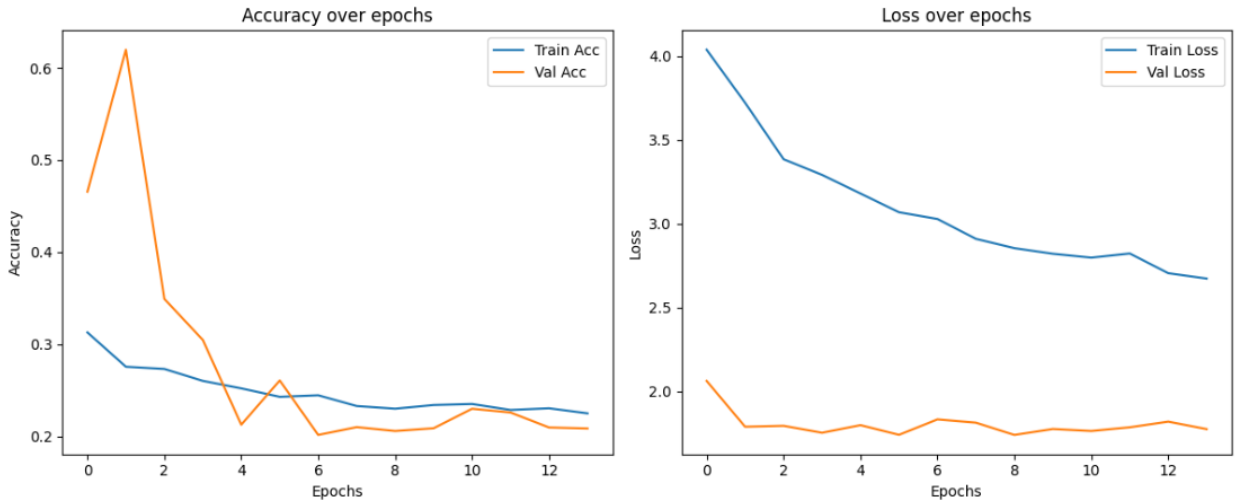
Zanimljivo je da model postiže znatno bolju tačnost na validacionom skupu (preko 75% u poslednjim epohama), nego na trening skupu (oko 47%).

Trendovi funkcije greške pokazuju da validacioni gubitak brzo opada i stabilizuje se, dok trening gubitak opada sporije, što ukazuje na stabilno učenje bez znakova preprilagođavanja (engl. *overfitting*) u posmatranom intervalu epoha.

Ovi rezultati sugerišu da je zadatak klasifikacije proteinskih blokova uspešnije rešiv u odnosu na klasifikaciju aminokiselinskih parova, što može biti posledica boljih ulaznih reprezentacija, većih razlika među klasama ili manje izražene nebalansiranosti skupa podataka.

Sekundarne strukture

Na slici 19 prikazano je praćenje tačnosti i vrednosti funkcije greške tokom epoha treniranja modela za klasifikaciju sekundarnih struktura. Za razliku od prethodnih zadataka, ovde je uočen nestabilan obrazac u metrikama validacije, preciznost validacionog skupa varira značajno kroz epohe, bez



Slika 19: Preciznost (levo) i funkcija greške (desno) tokom epoha za klasifikaciju sekundarnih struktura

jasnog trenda poboljšanja, dok vrednosti gubitka ostaju gotovo konstantne posle početnog pada.

S druge strane, trening metrika pokazuje blagi pad greške i stabilnu, ali relativno nisku tačnost. Ovakvi rezultati mogu ukazivati na nekoliko potencijalnih problema: visoka složenost samog zadatka, izražena klasa nebalansiranost ili manjak reprezentativnosti validacionog skupa.

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.

[4] Wei Wu, Zhiheng Wang, Peisheng Cong, Tonghua Li. *Accurate Prediction of Protein Relative Solvent Accessibility Using a Balanced Model*. Scientific Reports, 10:17560, 2020.

[5] Matthew Z. Tien, Austin G. Meyer, Dariya K. Sydykova, Stephanie J. Spielman, Claus O. Wilke. *Maximum Allowed Solvent Accessibilities of Residues in Proteins*. PLoS ONE, 8(11):e80635, 2013.

[6] National Institute for Mathematical and Biological Synthesis (NIMBioS). *Amino Acid Properties and Functions Web Module*. Available at: `https://legacy.nimbios.org/~gross/bioed/webmodules/aminoacid.htm`, accessed May 27, 2025.

[7] Joseph J. and Agarwal G. *Protein Secondary Structure: An Evolutionary Perspective*. BioPhysical Reviews, Preprint, 2010. Available at: `https://inserm.hal.science/inserm-00512823/file/Joseph_Agarwal_BiohysRev_2010_preprint.pdf`, accessed May 27, 2025.