Univerzitet u Beogradu
Matematički fakultet

Tijana Kostić

Karakterizacija neuređenih regiona proteina pomoću ponavljajućih niski

Master rad

Beograd,
2017.

Podaci o mentoru i članovima komisije

Mentor

prof. dr Nenad Mitić, vanredni profesor, Matematički fakultet, Univerzitet u Beogradu

Članovi komisije

prof. dr Nenad Mitić, vanredni profesor, Matematički fakultet, Univerzitet u Beogradu

prof. dr Saša Malkov, vanredni profesor, Matematički fakultet, Univerzitet u Beogradu

dr Miloš Beljanski, naučni savetnik, Institut za opštu i fizičku hemiju

Datum odbrane:

Sadržaj

Spisak slika

Spisak tabela

1 Uvod

Proteini su osnovni gradivni elementi svih živih bića kao nosioci širokog spektra bioloških funkcija unutar organizma. Funkcionalnost proteina u mnogome zavisi od njegove jedinstvene strukture. Međutim, postoje proteini koji nemaju definisanu tercijarnu strukturu (neuređeni proteini) ili nemaju uređenu tercijarnu strukturu u nekim svojim delovima. Zahvaljujući ovakvim proteinima ili delovima proteina, proteini vrše (ili ne vrše) neke od bioloških funkcija. Proteini koji nemaju definisanu tercijarnu strukturu, imaju velikog značaja u razvoju raznih vrsta bolesti. Dosadašnja istraživanja su pokazala da su neuređeni proteini povezani sa bolestima kao što su kancer, autoimune bolesti, alergije, neuredegenerativne bolesti i druge. Iako se pretpostavlja da neuređeni proteini, odnosno neuređeni delovi proteina, imaju ulogu u nastajanju navedenih bolesti, još uvek se ne zna na koji način. S toga ćemo se pozabaviti analizom neuređenih delova proteina, a sa ciljem uspostavljanja korelacije između ponavljajućih niski unutar proteina sa uređenim, odnosno neuređenim delovima proteina.

Za istraživanje uticaja neuređenih delova proteina na razvoj bolesti, od značaja može biti ispitivanje zavisnosti pojavljivanja ponavljajućih sekvenci unutar neuređenih delova proteina. Drugim rečima, ustanoviti da li su neke ponavljajuće sekvence karakteristične samo za neuređene, odnosno samo za uređene delove proteina, a samim tim i da li je moguće na osnovu ponavljajućih sekvenci razlikovati uređene od neuređenih delova proteina. Upotrebom postojećih alata napravljen je model za predviđanje koji sa određenom tačnošću prepoznaje da li je nešto uređeni ili neuređeni deo proteina.

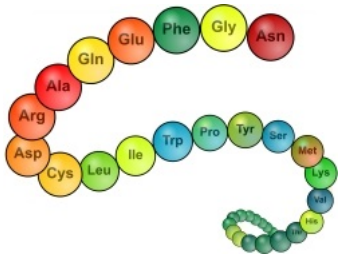
Ponavljajuće sekvence, u koje spadaju palindromi i ponovci(*eng. repeats*), se u proteinima nalaze u znatnom broju i pretpostavlja se da imaju znatnog uticaja na strukturu proteina.

2 Proteini i njihova struktura

Reč protein potiče od grčke reči *proteus* što znači prvi element. Proteini su osnovni elementi rasta i oporavka, dobre funkcionalnosti i strukture svih živih ćelija, samim tim i svakog živog organizma. Hormoni, kao što su insulin, kontrolišu nivo šećera u krvi; enzimi, kao što su amilaze, lipaze, proteaze, su neizostavni u varenju hrane; antitela pomažu u borbi protiv infekcija, itd.

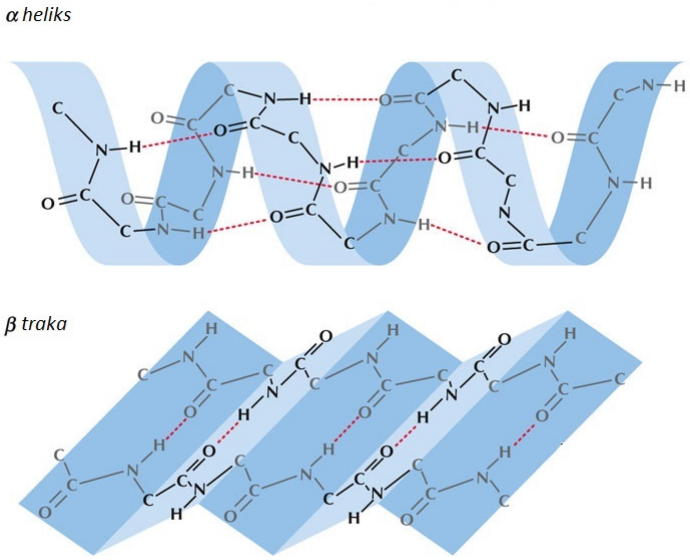
Proteini su polimeri[1] izgrađeni od niza amino kiselina, gradivnih blokova, koji su međusobno povezani. U prirodi je poznato 20 različitih amino kiseline kao i dve koje se, ne tako često, javljaju u proteinima. Kao što se uz pomoć azbuke gradi beskonačno mnogo različitih reči, tako uz pomoć amino kiselina može da se izgradi konačno mnogo različitih proteina. Protein se predstavlja kao niska slova gde svako slovo predstavlja jednu amino kiselinu. U zavisnosti od samog redosleda amino kiselina u proteinu, protein nosi odgovarajuću 3D strukturu u organizmu.

Protein možemo da posmatramo na različitim strukturnim nivoima i tako posmatran, protein ima četiri nivoa strukture. Protein kao lanac povezanih amino kiselina predstavlja primarnu strukturu proteina (slika 1), i takvu reprezentaciju proteina, u računaru, čuvamo kao nisku karaktera. Delovi proteina, odnosno nizovi amino kiselina, u prostoru mogu da imaju različite prostorne oblike među kojima su tri najzastupljenije strukture $\alpha - heliks$, $\beta - traka$ i $\beta(U) - okret$, koje nastaju kao posledica interakcije vodonika i kiseonika unutar različitih amino kiselina u nizu. Međusobnom interakcijom amino kiselina unutar jednog proteina nastaje sekundarna struktura proteina (slika 2).



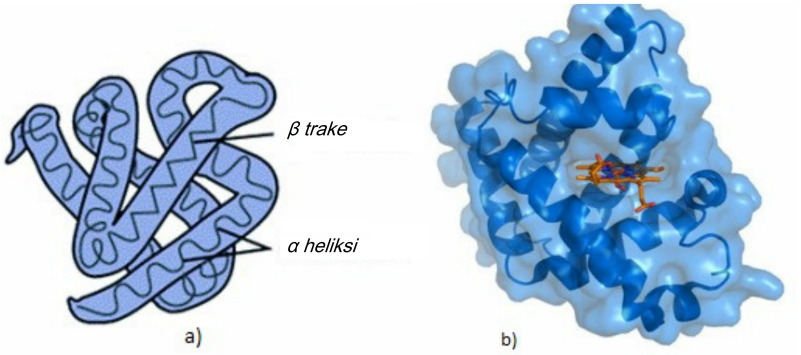
Slika 1: Primarna struktura proteina

[1] Hemijska jedinjenja ili mešavina jedinjenja formiranih polimerizacijom koja se sastoje od ponavljajućih strukturalnih jedinica - monomera. Kod proteina su to amino kiseline [13].



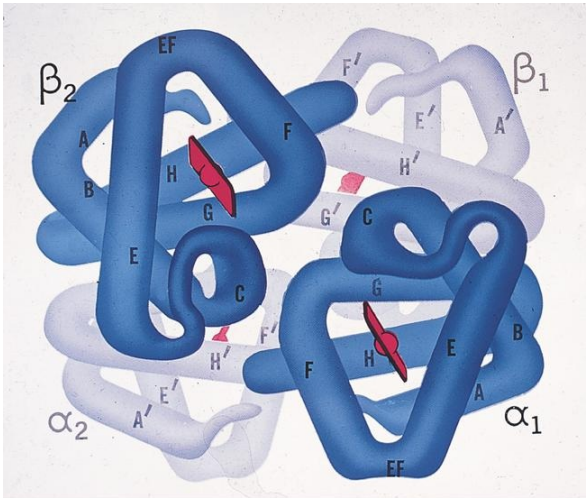
Slika 2: Sekundarna struktura proteina

Sledeći nivo strukture proteina jeste tercijarna struktura koja se dobija interakcijom i spajanjem sekundarnih struktura i predstavlja celokupni oblik polipeptida u prostoru (slika 3).



Slika 3: Tercijarna struktura proteina - a) Grafički prikaz i b) 3D prikaz

Kvaternarna struktura je nivo organizacije proteina koju čine više uvezanih polipeptida od kojih svaki ima uređenu tercijalnu strukturu. Jedan polipeptid u kvaternarnoj strukturi predstavlja jednu podjedinicu, a kvaternarna struktura prostorni raspored više podjedinica (slika 4).



Slika 4: Kvaternarna struktura proteina

3 Uređenost i neuređenost proteina

Veliki broj proteina ima uređenu strukturu u prostoru i takvi proteini se nazivaju uređenim proteinima. Međutim, postoje proteini koji u nekim delovima ili celom svojom dužinom nisu struktuirani i takvi proteini su nazvani suštinski neuređeni proteini (*eng. intrinsically disordered proteins*).

Definicija 3.1 *Tri kriterijuma za definiciju neređenosti proteina [15]:*

1. *Petlje ili uvojci (eng. loops/coils) su neuređeni delovi. Delovi proteina koji imaju strukturu $\alpha - heliksa$, $\beta - heliksa$ i $\beta - traka$ su uređeni regioni, dok se ostali oblici uvojaka smatraju neuređenim regionima. Petlje ili uvojci nisu nužno neuređeni delovi, međutim, neuređeni delovi su jedino nađeni unutar uvojaka. Ukoliko bi se model za predviđanje zasnivao pratećи samo ovu definiciju, gde je prisustvo uvojka potreban ali ne i dovoljan uslov za neuređenost, takav model bi bio suviše slobodan.*

2. *"Vruće petlje" (eng. hot loops) - predstavljaju podskup gore navedenih petlji koje imenuju petlje sa visokim stepenom mobilnosti određen C_α temperaturnim faktorom. Ovakve, visoko dinamične, petlje ukazuju na neuređenost proteina.*

3. *Nedostatak koordinata u strukturi dobijenu $X - zracima$. Neoznačena elektronska gustina često je ukazatelj neuređenosti [11].*

3.1 Prepoznavanje uređenih proteina

U DisProt bazim, verziji 6.03 [7], postoji 694 neuređenih ili delom neuređenih proteina. Uređeni delovi proteina imaju veću gustinu i manji radijus okretanja u odnosu na neuređene (pa s tim i manji nivo interakcije sa drugim molekulima), dok, suprotno tome, neuređene proteine karakteriše manja gustina, pa se uređeni proteini mogu otkriti metodama koje su osetljive na molekulske gustine, molekulske veličine ili hidrodinamički otpor. Najčešće korišćene eksperimentalne metode za otkrivanje strukture proteina su:

- Nuklearno magnetna rezonantna spektroskopija (NMR)
- Difrakciona kristalografija X – zracima
- Cirkularni dihroizam (CD).

3.2 Prepoznavanje neuređenih proteina

Jedna od najranije upotrebljavanih metoda za otkrivanje neuređenih regiona unutar proteina jeste pronalaženje delova sa manjom strukturalnom kompleksnošću. Naravno, ovakva korelacija je daleko od perfektne, iako neuređeni delovi proteina imaju nižu strukturalnu kompleknost. Neke od razvijenih metoda za izučavanje kompleksnosti su: SEG ([19]) and CAST ([12])[15]. Druga vrsta metode se zasniva na hidrofobnosti delova proteina (neuređeni delovi proteina su često vrlo slabo hidrofobni[2]).
Neki od prvih alata razvijenih za predviđanje neuređenih regiona su:

1. PONDR (Predictor of Naturally Disordered Regions) - baziran na veštačkim neuronskim mrežama

2. GlobPlot - zasnovan na algoritmu za predviđanje neuređenih delova baziran na sklonosti ka neuređenosti

3. NORSp - za predviđanje regiona bez sekundarne strukture; prema rečima autora, ovi regioni, ne moraju nužno da budu neuređeni. Neki proteini, iako nemaju sekundarnu strukuturu, mogu da formiraju tercijarnu strukturu koja se sastoji iz blokova - gde su gradivni blokovi uvojci (*eng. coils*).

Predviđanje tercijarne strukture proteina, takođe, može da se koristi kao alternativna metoda za predviđanje neuređenih proteina i delova proteina. Međutim, takve metode su računarski zahtevne i podložne greškama. Takođe, ove metode su konstruisane za predviđanje u globularnim domenima, pa njihova primena na drugim vidovima sekvence može da bude nepredvidiva. Navedeni prediktori se zasnivaju na različitim metodama i koriste različite pristupe i osobine neuređenosti, te tako i ne mogu međusobno biti upoređivani. Do sada je razvijeno više od 50 prediktora i alata koji koriste ove prediktore i svi se mogu svrstati u četiri kategorije po pristupu koji koriste [10]:

1. Pristup "s početka" (*lat. Ab-initio*) u kom se predviđanje zasniva na samoj sekvenci obično korišćenjem tehnika mašinskog učenja (kao što su SVM, neuronske mreže, Bajesovsko klasterovanje). U ovu grupu spadaju RONN, DISOPRED, DisEMBL, VSL2.

[2]Hidrofobnost je stepen odbojnosti prema vodi, odnosno stepen nerastvorivosti u vodi.

2. Pristup zasnovan na obrascima (*eng. template*) u kom se ispituju slične sekvence sa sličnom strukturom. U ovu grupu spadaju prediktori koji su zasnovani na fizičko-hemijskim osobinama amino kiselina u proteinu kao što su PONDR, FoldUnFold, PreLINK, IUPred, FoldIndex.

3. Meta pristup u kom se predviđanje zasniva na kombinaciji nekoliko pristupa (algoritama). Neki od njih su MD, GeneSilico Metadisorder, PONDR-FIT, metaPrDOS.

3.3 Funkcija neuređenih proteina

Tokom proteklih godina pokušavalo se sa razvojem prediktora za predviađanje funkcija neuređenih proteina i neuređenih delova proteina analizom samih sekvenci, ali sa ne tako velikim uspehom, ponajviše zbog širokog varijariteta u dužini neuređenih delova i velikog stepena različitosti samih sekvenci. Sa ciljem dobijanje boljih prediktora za predviđanje funkcije neuređenih proteina, a vođeni lošim rezultatima dobijenih samom analizom sekvenci, prešlo se na alternativni pristup. Neuređeni regioni su nasumično podeljeni u podskupove, za svaki od podskupova razvijen je prediktor analizom sekvenci unutar svakog podskupa, podskupovi su reparticionisani prediktorom koji je davao najbolje rezultate i zatim je postupak ponavljan iterativno dokle god je postajala promena u odnosu na prethodnu iteraciju. Ovakvim pristupom se došlo do tri različita tipa neuređenih regiona [1]:

1. Grupa V - grupa koja je bila bogata ribosamalnim proteinima

2. Grupa S - grupa koja je uključivala veliki broj regiona za vezivanje proteina (eng. Protein-binding region)

3. Grupa C - grupa sa velikim brojem delova proteina koji su podložni promenama.

Neophodni uslovi za obavljanje nekih funkcija jeste veliki broj interakcija koje neuređeni proteini vrše sa različitim molekulima, a to im omogućava visok stepen pokretljivosti što proizilazi iz nedostatka 3D strukture. Samim tim neuređeni proteini mogu da ostvare mnogo širi skup različitih funkcija nego uređeni proteini.

Neuređeni proteini nose ključne biološke funkcije uz pomoć interakcija između dva proteina ili proteina i nukleinske kiseline. Postojanje ovakvih proteina narušava paradigmu struktura - funkcija koja zagovara da je preduslov za postojanje funkcije proteina neophodna dobro definisana 3D struktura proteina i na osnovu koje se

definiše i sama funkcija proteina. Funkcije neuređenih proteina mogu biti podeljene u četiri široke grupe [1]:

- Molekularno prepoznavanje
- Molekularno sklapanje
- Modifikacija proteina
- Entropijske aktivnosti

3.4 Neuređenost proteina i ljudske bolesti, D2 koncept

Neuređeni proteini imaju ulogu u biološkim aktivnostima kao što su regulacija, signalizacija i kontrola gde njihovo vezivanje za druge molekule ima vitalni značaj i time upotpunjuju funkcije uređenih proteina. Nedostatak ili nemogućnost obavljanja funkcija koju nose ovi proteini može rezultovati različitim patološkim stanjima.
Veliki broj bolesti kao što su kancer, kardiovaskularna oboljenja, neurodegenerativne bolesti, diabetes, itd., povezane su na neki način sa neuređenim proteinima. Navedene bolesti spadaju u grupu bolesti koje nastaju kao posledica nemogućnosti da proteini dostignu konformaciju potrebnu za obavljanje odgovarajuće funkcije. Osim nemogućnosti da dostignu svoju funkcionalnu konformaciju, može doći i do interakcije sa nekim manjim molekulima ili drugim proteinima što može dovesti do promene konformacije proteina i time smanjiti sposobnost proteina da dostigne svoju 3D strukturu.

Vladimir N. Uversky, Christopher J. Oldfield i A. Keith Dunker [18] su nedavno sprovodili analizu, sa osvrtom na funkcionalnost proteina, nad celokupnom Swiss-Prot bazom, posmatrajući iz ugla uređenih i iz ugla neuređenih proteina. Ključne funkcionalne reči (reči koji opisuju funkciju proteina) koje se povezuju sa 20 i više proteina su izdvojene zajedno sa skupom koji uključuje proteine koji imaju srodstva sa tom funkcijom. Za svaki par (ključna reč, skup proteina) dodeljen je slučajno izabran skup proteina dužine 1000 amino kiselina. Predviđanje da li je skup uređen ili neuređen izvršena je i za parove (ključna reč, skup proteina) i za slučajno izabran skup. Ako je funkcija koja je opisana ključnom rečju nošena od strane dugačkog regiona neuređenog proteina, pretpostavlja se da bi predikcija unutar skupa povezanog sa tom ključnom rečju imao mnogo veći proj predviđenih neuređenih regiona nego bilo koji od

slučajnih skupova. Broj predviđenih neuređenih regiona bi bio mnogo manji unutar dodeljenog skupa u odnosu na slučajne skupove ukoliko bi funkcija čiji je to skup bila nošena od strane uređenog proteina. Kada su ključne reči particionisane u 11 kategorija (biološki procesi, ćelijske komponente, razvojni procesi, ...), ključne reči povezane sa uređenim proteinima su pronađene u svega 7 kategorija, dok su ključne reči povezane sa neuređenim regionima pronađene u svih 11 kategorija. Ovaj vid istraživanja pokazuje korelaciju bolesti sa predviđanjem neuređenih regiona. Suprotno ovome, nije pronađeno da su proteini koji su povezani sa nekim bolestima u korelaciji sa odustvom neuređenosti.

D2 koncept ili neuređenost u neuređenom (eng. disorder in disorders) predstavlja sve češće upotrebljavan koncept koji označava veliki stepen povezanosti između neuređenog u neuređenim proteinima i različitih oboljenja.

4 Ponavljajuće niske

Simetrija oblika je nešto što se može naći svuda u prirodi. Ljudima je oduvek bila privlačna simetrija u različitim oblicima u prirodi. Prve simetrične rečenice datiraju iz vremena od pre više od 20 vekova. Sa napretkom molekularne biologije i računarskih nauka, otkriveni su novi vidovi simetričnih oblika u prirodi kao što su palindromi i ripiti (eng. *repeat*) u proteinima.

Proizvoljne niske, odnosno par (pod)niski, koje zadovoljavaju određene uslove, se nazivaju ponovci. U zavisnosti od uslova koji ispunjavaju, ponovci mogu biti direktni ili obrnuti, koji, dalje, mogu biti podeljeni u komplementarne i nekompementarne podskupove. ponovci mogu biti precizno definisani definicijom 4.1 [4].

Definicija 4.1 *Neka je $A = \{a,b,c\ ,...\}$ skup proizvoljnih simbola i $L = \{l_1, l_2, ...\}$ jezik definisan nad azbukom A koji obuhvata niske proizvoljne dužine uključujući prazne niske, i neka je $|s|$ dužina niske $s \in L$. Uređena trojka (x, s, p_x) označava podnisku $x \in L$ niske $s \in L$ na poziciji $p_x \geq 1$ ako*

$$\exists x, z \in L : s = yxz \wedge |s| = |x| + |y| + |z| \wedge |z| \geq 1.$$

Ako su funkcije f i g definisane na sledeći način:

- $f : A \longrightarrow A$

$$f(x) = \begin{cases} z, & ako \quad |x| = 1, \quad za\ neko \quad z \in A \\ f(x_1)(x_2), & ako \quad x = x_1x_2 \in L \wedge |x| > 1 \end{cases}$$

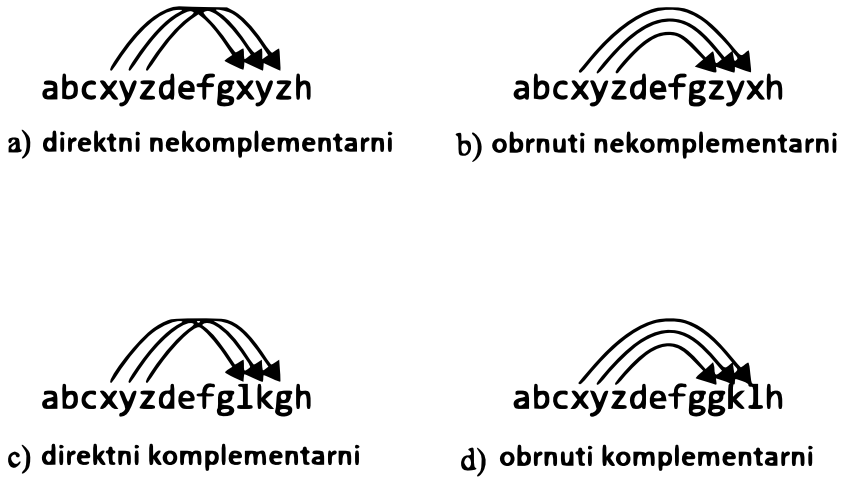
- $g : L \longrightarrow L$

$$f(x) = \begin{cases} yx, & ako \quad |x| = 1 \wedge |y| = 1 \\ yg(x), & ako \quad |y| = 1 \wedge |x| > 1 \\ g(y)x, & ako \quad |y| = 1 \wedge |x| > 1 \\ g(y)g(x), & inače \end{cases}$$

tada, za sve niske $s \in L$, četiri tipa ripita mogu biti definisana (slika 5):

1. *Par niski (a, s, p_a) i (b, s, p_b) je direktni nekomplementarni ripit ako i samo ako $a = b \wedge p_a < p_b$.*

2. *Par niski (a, s, p_a) i (b, s, p_b) je obrnuti nekomplementarni ripit ako i samo ako $a = g(b) \wedge p_a \leq p_b$.*



Slika 5: Grafički prikaz tipova ripita. U ovom primeru su korišćene vrednosti $f(x) = l, f(y) = ki f(z) = g$.

3. *Par niski* (a, s, p_a) *i* (b, s, p_b) *je direktni komplementarni ripit ako i samo ako* $a = f(b) \wedge p_a < p_b$.

4. *Par niski* (a, s, p_a) *i* (b, s, p_b) *je obrnuti komplementarni ripit ako i samo ako* $a = f(g(b)) = g(f(b)) \wedge p_a \leq p_b$.

Obrnuto nekomplementarni ponovci su poznati i kao palindromi.

Sa početkom razvoja molekularne biologije, ranih sedamdesetih, otkriveni su palindromi unutar niske DNK. Palindromska sekvenca (*eng. palindromic sequence*) je sekvenca nukleinskih kiselina dvostruke spirale DNK i/ili RNK koja je ista čitajući je od 5' ka 3' kraju jednog lanca i od 5' ka 3' kraju drugog komplementarnog lanca. Takođe je poznata i kao obrnuta sekvenca (*eng. inverted-reverse sequence*) i javlja se u genomu svakog organizma. Kako su komplementarne azotne baze (nukleoditi) A-T, C-G kod DNK, odnosno A-U, C-G kod RNK, tako bi nukleinska sekvenca jednog lanca bila ista kao obrnuta sekvenca drugog komplementarnog lanca. Primer palindroma unutar nukleinske sekvence je dat u primeru 4.1.

Primer 4.1
```
5' - G A A T T C - 3'
3' - C T T A A G - 5'
```

Poznato je da su obrnute palindromske sekvence unutar DNK značajne za DNK replikaciju i RNK transkripciju [5].

4.1 Palindromski peptidi i njihove funkcije

Da bi se analizirala funkcija palindroma u proteinu, krenulo se od izučavanja uticaja palindroma na strukturu proteina. Smatra se da je pojava palindoma unutar proteina značajna i za njihovu strukturu (a time i za njihovu konformaciju i vezivanje sa druge molekule) i za njihovu funkciju. U istraživanjima koja su sprovedena nad palindromima iz Swiss-Prot baze [17][5], posmatrani su palindromi manje dužine sa savršenom sličnošću sa obrnutom sekvencom (drugim rečima, palindromi odgovaraju definiciji datoj u poglavlju 4). Ovaj uslov obezbeđuje da razlike u strukturi proteina nisu posledica razlike u nepotpunoj simetričnosti palindroma. Zahvaljujući posmatranju palindroma malih dužina, obezbeđen je uslov u kome se obe peptidne strane proteina nalaze blizu jedna drugoj, samim tim u istom(sličnom) okruženju, te se očekuje da formiraju sličnu sekundarnu strukturu. Ukoliko obe strane palindroma formiraju istu strukturu, pretpostavljaće se da imaju sličnu sekundarnu strukturu. Ovakvim posmatranje palindroma i njihove strukture došlo se sledeće podele palindroma:

1. Palindromi kod kojih obe peptidne strane formiraju $\alpha - heliks$ strukturu
2. Palindromi kod kojih obe peptidne strane formiraju $\beta - traka$ strukturu
3. Ostali

U bazi, nad kojom je rađeno istraživanje, ima ukupno 980 palindromskih sekvenci i zastupljenost tih palindroma u gore navedenim grupama je:

- 12.2% palindroma spada u prvu grupu
- 0.7% palindroma spada u drugu grupu
- 87.1% palindroma spada u treću grupu

Procenat zastupljenosti 489720 slučajnih sekvenci u gore navedenim grupama je:

- 3.4% spada u prvu grupu
- 0.6% spada u drugu grupu
- 96% spada u treću grupu

Rezultati ukazuju da palindromi unutar proteina imaju mnogo veću tendeciju formiranja $\alpha - heliks$ struktura od slučajnih ili nasumičnih sekvenci (sekvenci koje ne odgovaraju definiciji palindroma datoj u poglavlju 4) koje se javljaju u proteinu.

Takođe je pokazano i da proteini sa slučajnim sekvencama retko imaju sekundarnu strukturu [16]. Smanjivanjem kompleksnosti kompozicije amino kiselina (broja različitih amino kiselina unutar sekvence) povećava se verovatnoća nastanka palindroma, a time i formiranja sekundarne strukture. Iz prethodnih rezultata dolazi se do zaključka da proteini sa palindromskim sekvencama mogu da formiraju sekundarnu strukture sa naklonosti na $\alpha - heliks$ strukture, što može biti uzrok njihovih čestih pojava unutar proteina [5] [9].

5 Cilj rada

Značajni deo celokupne slike o proteinima odnosi se na neuređenost proteina pa se s tim u vezi postavlja i pitanje ponavljajućih sekvenci u neuređenim delovima. Da bi doprineli u izučavanju ovog dela, pozabavićemo se analizom podataka sa statističke strane karakterišući neuređene delove proteina palindromima i njihovom zastupljenošću u neuređenim regionima koristeći različite metode predviđanja i klasifikacije. Analiza palindroma uključuje njihovu zastupljenost, zastupljenost u odnosu na dužinu palindroma i odnos složenosti i dužine. Cilj rada je:

- Ispitati koliko se palindroma i ripita nalazi u uređenim, odnosno neuređenim delovima proteina.
- Ispitati kolika je zastupljenost kratkih(do 5 amino kiselina)/dugih (više od 5 amino kiselina) palindroma, odnosno ripita u uređenim/neuređenim proteinima, odnosno delovima proteina.
- Tehnikama istraživanja podataka ispitati da li postoji međuzavisnost određenih palindroma i ripita, njihovih dužina i neuređenosti proteina.
- Dobijene klasifikacione modele testirati nad proteinima koji predstavljaju razliku verzija 6.02 i 7.03 DisProt baze.

6 Materijali i metode

Za analizu podataka korišćena je DisProt baza[7] neuređenih proteina, verzija 6.02 i dodatno za proveru dobijenih rezultata verzija 7.03. Verzija 6.03 sadrži 694 proteina, dok verzija 7.03 sadrži 803 neuređenih ili delom neuređenih proteina. DisProt baza je baza ručno sakupljanih ekspirementalnih dokaza neuređenosti iz dostupne literature u kojoj postoji početna i završna pozicija neuređenog regiona kao i metode korišćenje kojim se došlo do rezultata. Svi podaci u daljem tektsu su podaci iz verzije 6.03, dok je verzija 7.03 korišćena za testiranje dobijenih modela klasifikacije u poglavlju 8.2.
U tabeli 1 prikazana je raspodela proteina na osnovu broja amino kiselina od kojih se sastoje.

Tabela 1: Podela proteina prema broju amino kiselina

Dužina proteina	Broj proteina
0-100	69
101-200	149
201-300	111
301-400	86
401-500	66
501-600	53
601-700	41
701-800	26
801-900	18
901-1000	11
1001-2000	48
2001-5000	14
preko 5000	2

Podaci iz DisProt baze[7] su prebačeni u relacionu bazu podataka sa sledećim tabelama:

- Tabela DISPROTFASTA koja sadrži sve proteine sa svojim kodovima.
- Tabela DISPROTPALINDROMI koja sadrži sve palindrome iz svih proteina sa početnom i krajnjom pozicijom palindroma unutar proteina.
- Tabela DISPROTRIPIT koja sadrži sve ponovke iz svih proteina sa početnom i krajnjom pozicijom ripita unutar proteina.

- Tabela DISPROTPOZICIJE koja sadrži početnu i krajnju poziciju uređenog i neuređenog dela proteina za svaki protein.

U DisProt bazi, verziji 6.02, postoje podaci o uređenosti, odnosno intervali dati početnom i krajnom pozicijom, kao i tipom uređenosti (neuređen ili uređen region) za svaki interval unutar sekvence za svaki od proteina. Oznake koje su korišćene za označavanje uređenosti jeste karakter 'D' za neuređeni region i karakter 'O' za uređeni region.

6.1 Metode

Istraživanje podataka (eng.*Data mining*) je disciplina koja se bavi analizom i pronalaženjem skrivenih informacija u velikim skupovima podataka, a sa ciljem pronalaženja značajnih uzoraka podataka, trendova u životnom ciklusu podataka, međuzavisnosti podataka i uspostavljanje pravila među podacima. Kao rezultat izučavanja dobija se model kojim se karakteriše skup podataka nad kojim je istraživanje sprovedeno. Takav model omogućava sagledavanje podataka iz drugog ugla i karakterizaciju novih podataka istog tipa. Tehnike istraživanja podataka koje su ovde korišćene su:

1. **Pravila pridruživanja**(eng. *association rules*) - označava pronalaženje obrazaca koji opisuju međusobno čvrsto povezane osobine podataka. Obrasci se najčešće predstavljeni u obliku pravila $X \to Y$ koje govori da ukoliko podatak zadovoljava skup osobina X, onda zadovoljava i skup osobina Y. Metrike za određivanje kvaliteta pravila su:

   - Podrška - koja govori koliko je pravilo korisno (u koliko se transakcija X i Y javljaju zajedno).

     $$s(X \to Y) = \frac{\sigma(XY)}{N,}$$

     gde je $\sigma(XY)$ broj transakcija u kojima se javlja i X i Y, a N ukupan broj transakcija.

   - Pouzdanost - koja govori koliko je pravilo precizno, odnosno odnos broja transakcija u kojima se javljaju i X i Y i broja transakcija u kojima se javlja X.

     $$c(X \to Y) = \frac{\sigma(XY)}{\sigma(X)},$$

     gde je $\sigma(XY)$ broj tansakcija u kojima se javlja i X i Y, a $\sigma(X)$ broj transakcija u kojima se javlja X.

- Značajna mera koja se takođe koristi je $Lift$ mera. Lift mera u obzir uzima statističke podatke.

$$Lift = \frac{P(Y|X)}{P(Y)},$$

gde je $P(Y|X)$ verovatnoća pojavljivanja Y u transakcijama u kojima se javlja X, a $P(Y)$ verovatnoća pojavljivanja Y u transakcijama.

2. **Klasifikacija** - predstavlja formiranje modela klasifikacije (funkcije) na osnovu kog se ulazni podaci grupišu u već postojeće klase, odnosno definisanje funkcije (podatak posmatramo kao uređeni par (x, y) gde je x skup vrednosti atributa koji opisuju podatak, a y klasa kojoj podatak pripada) kojom se skup atributa preslikava u jednu od predefinisanih klasa. Prilikom izgradnje modela, podaci se razdvajaju na trening i test skupove koji su disjunktni. Trening skup služi za dobijanje modela, a test skup za testiranje tako dobijenog modela. Jedna od osnovnih mera za kvalifikovanje kvaliteta modela jeste *tačnost* modela koja predstavlja odnos ispravno klasifikovanih instanci u odnosu na ukupan broj instanci. Klasifikacija predstavlja *nadgledano* učenje, što znači da su klase klasifikacije unapred poznate. Dobar model klasifikacije mora da dobro klasifikuje i trening i test podatke.

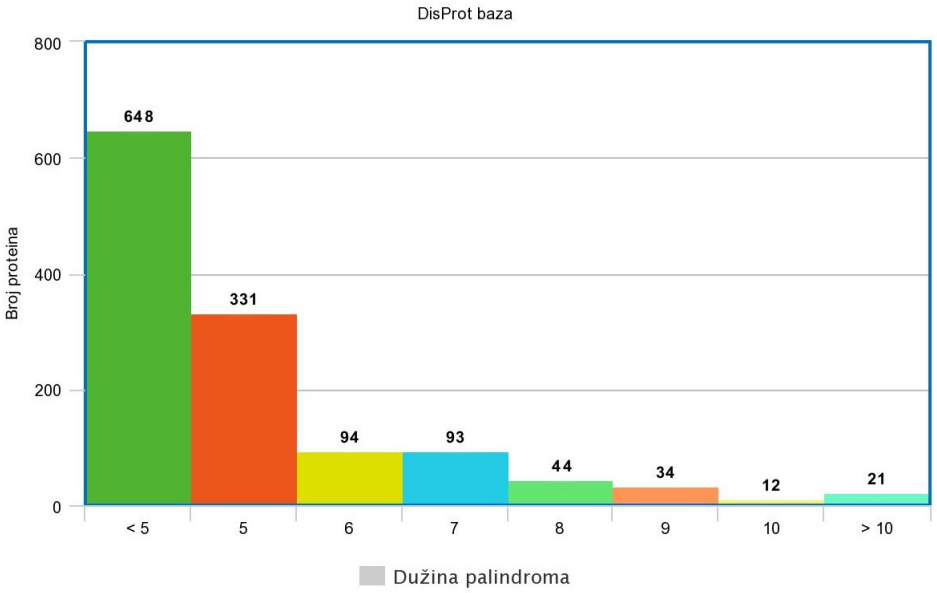
Kao alat za dobijanje klasifikacionog modela, pravila pridruživanja i vizualizaciju dobijenih rezultata korišćen je Intelligent Miner InfoSphere Warehouse kompanije IBM. Svi modeli klasifikacije dobijeni ovim alatom priloženi su u radu, a rezultati opisani u daljem tekstu.

7 Statistička obrada podataka

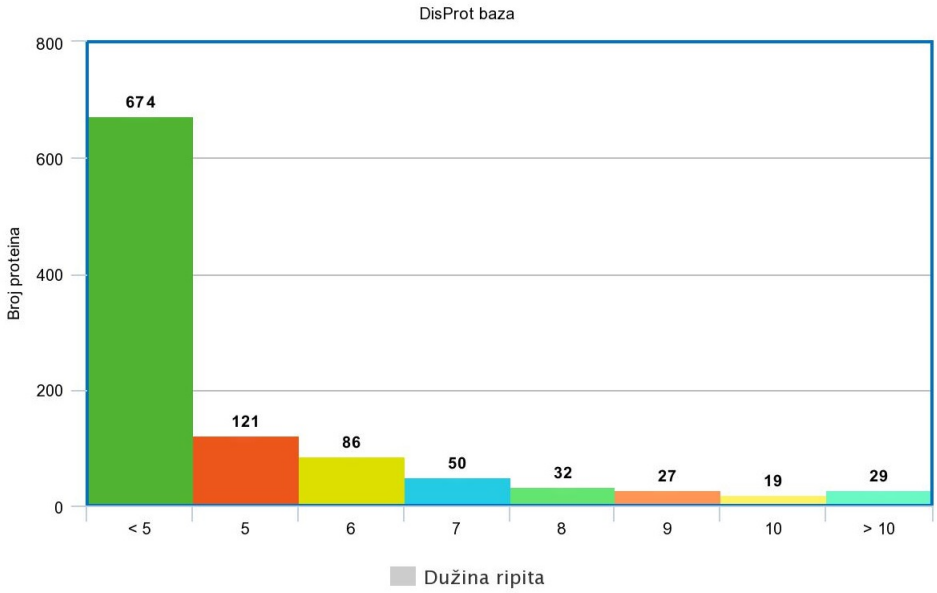
Nad DisProt bazom podataka najpre je izvršena analaza o podacima koji će predstavljati ulaz za klasifikacioni model. Od ukupno 694 proteina koji se nalaze u bazi u verziji navedenoj u poglavlju 6, ne postoji protein koji u sebi sadrži palindrome koji se nalaze isključivo u uređenim regionima, a 643 proteina koji imaju palindrome u neuređenom, nemaju palindrome u uređenom delu. Gledajući ripite, postoji samo jedan protein koji ripite sadrži isključivo u uređenim regionima, a ni jedan ripit u neuređenim regionima. Broj proteina koji sadrže ripite u neuređenim, a nemaju ripite u uređenim regionima je 633. Ovakav skup podataka je pogodan za dobijanje modela za karakteriziju neuređenih delova proteina, odnosno modela koji će sa "dovoljnom" tačnošću reći šta su karakteristični palindromi i ponovci za neuređene delove proteina. Naravno, izraz "dovoljnom tačnošću" zavisi od same primene modela i svrhe u kojoj će se koristiti. U daljem radu će se uz pomoć modela okarakterisati neuređeni delovi proteina, a kroz karakteristične ponavljajuće niske.

Radi boljeg upoznavanja sa podacima, izvršeni su SQL upiti i dobijeni podaci o rasprostranjenost palindroma i ripita po proteinima u zavisnosti od dužine. Rezultati su predstavljeni histogramima na slici 6 i 7. I kod palindroma i kod ripita, zastupljenost značajno opada već sa dužinom pet u odnosu na broj proteina koji imaju dužinu manju od pet, što je delom i zbog posmatranja šireg opsega u slučaju kada se posmatraju svi palindromi i ponovci čija je dužina manja od pet. Ono što je karakteristično jeste da je mnogo veći broj proteina koji sadrži palindrome dužine pet nego proteina koji sadrže ripite dužine pet, što ukazuje da su, možda, palindromi karakterističniji za neuređene regione od ripita. Ostali rezultati prikazani na histogramima su očekivani.

SQL upitima dobijeni su detalji o broju različitih palindroma unutar uređenih i neuređenih delova proteina. Za ispitivanje broja različitih palindroma unutar uređenih i neuređenih delova proteina uzeti su maksimalni palindromi unutar proteina. Rezultati dobijeni upitima predstavljeni su u tabeli 2 i 3. Isto je rađeno i za broj ripita u uređenim i neuređenim regionima i rezultati su predstavljeni u tabelama 4 i 5.



Slika 6: Rasprostranjenost palindroma unutar proteina zavisno od dužine palindroma.



Slika 7: Rasprostranjenost ripita unutar proteina zavisno od dužine ripita.

Tabela 2: Broj palindroma u uređenim i neuređenim regionima proteina u DisProt bazi.

Uređeni region	Neuređeni region
4298	687855

Tabela 3: Broj različitih palindroma u uređenim i neuređenim regionima proteina u DisProt bazi.

Uređeni region	Neuređeni region
736	5418

Tabela 4: Broj ripita u uređenim i neuređenim regionima proteina u DisProt bazi.

Uređeni region	Neuređeni region
3803	662740

Tabela 5: Broj različitih ripita u uređenim i neuređenim regionima proteina u DisProt bazi.

Uređeni region	Neuređeni region
514	5932

Daljim radom, upitima su dobijeni rezultati rasprostranjenosti dugih i kratkih palindroma i ripita i oni su prikazani tabelama 6 i 7.

Iz tabela 6 i 7 se vidi da postoji značajna razlika u broju različitih dužih ripita u neuređenim regionima i dužih palindroma u neuređenim regionima gde je veća zastupljenost ripita, ali je ukupan broj palindroma u neuređenim delovima veći od ukupnog broja ripita u neuređenim regionima, što znači da postoji mnogo veći broj palindroma koji se učestalije javljaju nego što je to slučaj kod ripita. Kod ripita je veća raznovrsnost u strukturi samih ripita.

Tabela 6: Broj palindroma u uređenim i neuređenim delovima proteina za duže i kraće palindrome u DisProt bazi.

Dužina palindroma	Uređeni region	Neuređeni region
≤ 5	4291	686232
> 5	7	1623

Tabela 7: Broja ripita u uređenim i neuređenim delovima proteina za duže i kraće ripite u DisProt bazi.

Dužina ripita	Uređeni region	Neuređeni region
≤ 5	3802	651843
> 5	1	10897

8 Rezultati

8.1 Rezultati dobijeni pravilima pridruživanja

Kao prvi metod uz pomoć kog se mogu karakterisati neuređeni delovi proteina jesu pravila pridruživanja. Pravila pridruživanja su dobijena kroz više iteracija, uzimajući redom sve palindrome dužine veće od šest u prvoj iteraciji, dužine veće od sedam u drugoj iteraciji, i tako dalje, sve do iteracije koja je uzimala u obzir sve palindrome dužine veće od deset. U svakoj od iteracija, jedini atributi koji su učestvovali, bili su sam palindrom proteina i polje tip koji može imati vrednosti neuređeni i uređeni region. Presekom dobijenih pravila dobijena su pravila predstavljena u tabelama 8, 9, 10, 11, 12 i 13 koje bliže opisuju neuređene proteine (u svakom sledećem skupu dobijenih pravila, uzeta su samo pravila koja se prethodno nisu pojavljivala):

Tabela 8: Pravila pridruživanja dobijena nad palindromima dužine veće od 5.

Pravilo	Podrška	Pouzdanost
QGQQGQ → D	10.86	100%
SPSYSP → D	8.23	100%
SDSDSDS → D	2.74	100%
EEEEEE → D	2.26	95%
EEEEEEE → D	2.02	97%
Broj transakcija: 1676		

Tabela 9: Pravila pridruživanja dobijena nad palindromima dužine veće od 6.

Pravilo	Podrška	Pouzdanost
SDSDSDS → D	4.94	100%
QQQQQQQQ → D	2.68	100%
EEEEEEEE → D	2.14	100%
Broj transakcija: 931		

Analizom celokupnog skupa palindroma, bez ograničavanja na dužinu palindroma, kao što je rađeno u prethodnim primerima, dobijaju se pravila ista kao u tabeli 13. Takvi razultati se dobijaju jer je najveći broj palindroma u proteinima

Tabela 10: Pravila pridruživanja dobijena nad palindromima dužine veće od 7.

Pravilo	Podrška	Pouzdanost
APAPAPAPAP → D	3.03	100%
Broj transakcija: 528		

Tabela 11: Pravila pridruživanja dobijena nad palindromima dužine veće od 8.

Pravilo	Podrška	Pouzdanost
PAPAPAPAP → D	3.16	100%
EEEEEEEEE → D	2.63	100%
QPQQPFPQQPQ → D	2.63	100%
GGGGGGGGGGGG → D	2.11	100%
Broj transakcija: 379		

Tabela 12: Pravila pridruživanja dobijena nad palindromima dužine veće od 9.

Pravilo	Podrška	Pouzdanost
GGGGGGGGGGGG → D	2.94	100%
Broj transakcija: 272		

Tabela 13: Pravila pridruživanja dobijena nad palindromima dužine manje od 6.

Pravilo	Podrška	Pouzdanost
KE → D	3.87	99.76%
KK → D	3.38	99.79%
KL → D	3.38	99.66%
EE → D	2.89	99.53%
VE → D	2.87	99.82%
EV → D	2.64	99.78%
EK → D	2.63	99.67%
LK → D	2.63	99.40%
Broj transakcija: 941.231		

dužine manje od šest. Ukupan broj transakcija u tom slučaju je 942.907 transakcija. Pravila koja se se izdvajaju sa većom podrškom od svih ostalih pravila su pravila koja za levu stranu imaju palindrome QGQQGQ i SPSYSP sa podrškama 10.86 i 8.23 respektivno. Ove palindrome, pored onih dužine manje od šest predstavljenih u tabeli 13, takođe možemo da izdvojimo kao karakteristične ponavaljajuće niske neuređenih regiona proteina. Takođe, kao karakteristični palindrom možemo da izdvojimo i palindrom GGGGGGGGGGGG zbog svoje dužine. Iako je podrška ovog pravila mala, ovakav palindrom se ne javlja u uređenim delovima.

8.2 Rezultati dobijeni klasifikacijom

Ulazni podatak, za sve modele klasifikacije, je leva strana palindroma, a ciljni atribut je kolona TIP koja ima vrednosti 'D' i 'O' što označava neuređeni i uređeni region, repspektivno. Trening i test podaci su podeljeni u odnosu 70 prema 30. Za klasifikaciju su korišćeni palindromi i ponovci kao kategorički atributi.

8.2.1 Klasifikacioni modeli zasnovani na palindromima

Prvi klasifikacioni model za ulazne podatke ima sve palindrome dužine manje od 5 i bez odsecanja grana drveta odlučivanja dobijen je model sa sledećim karakteristikama:

1. Rezultati nad trening podacima:
   (a) Veličina populacije za ovaj model je 13715 instanci, a procenat tačno klasifikovanih 75%.
   (b) Ukupna tačnost modela 0.67, tačnost klasifikacije palindroma neuređenih regiona je 0.803, dok je tačnost klasifikacije palindroma uređenih regiona 0.701.
   (c) Modalna vrednost klase palindroma neuređenih regiona je VE, a modalna vrednost palindroma uređenih regiona je palindrom LL.
   (d) Karakterističnih palindroma iz ovog modela za neuređene regione ima konačno mnogo i dati su na slici 8.
2. Rezultati nad test podacima:
   (a) Tačnost modela nad test podacima je 0.719 (74% tačno klasifikovanih od ukupno 373225 instance).

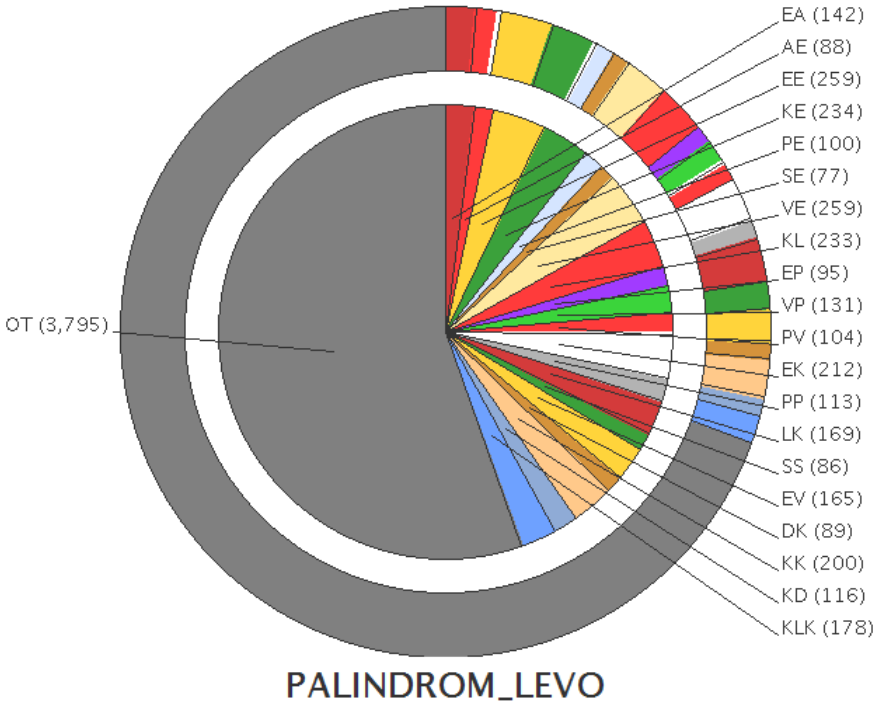
(b) Tačnost klasifikacije palindroma neuređenih regiona je 0.745.

3. Rezultati primene modela nad proteinima koji predstavljaju razliku verzije 7.03 i verzije 6.02:

   (a) Tačnost modela nad test podacima je 0.193 (11% tačno klasifikovanih od ukupno 3061 instance).

   (b) Tačnost klasifikacije palindroma uređenih regiona je 0.69, a tačnost klasifikacije palindrom neuređenih regiona je 0.745.

Dobri rezultati nad palindromima dužine manje od 5, možda govore o postojanju razlike između palindroma koji se javljaju u neuređenim i uređenim regionima, a time i potvrđuju mogućnost karakterizacije neuređenih regiona palindromima. Problem može da predstavlja jako loša tačnost modela nad proteinima koji su razlika verzije 7.03 i 6.02.
Klasifikacija po pojedinačnim dužinama palindroma, nije bila moguća zbog nedovoljnog broja palindroma u uređenim regionima. Kako je DisProt baza neuređenih proteina, prvobitna ideja da se naprave klasifikacioni modeli za svaku dužinu pojedinačno, pa i za celokupni skup palindroma i da se na kraju rezultati objedine, je, zbog nemogućnosti od strane samog alata, izmenjena i za pojedinačne dužine palindroma su napravljena pravila pridruživanja u odeljku 8.1. Da bi se izbeglo dobijanje samo kratkih palindroma (palindroma dužine manje od pet) kao rezultat klasifikacije i modalnih vrednosti za neuređene regione, moguće je bilo uraditi klasifikaciju za palindrome dužine veće od tri.



Slika 8: Palindromi dobijeni klasifikacijom palindroma dužine manje od 5 karakteristični za klasu D (*disorder*, odnosno klasa neuređenih delova proteina).

Klasifikacijom, čiji su ulazni podaci palindromi dužine veći od 3, dobijen je model sa sledećim karakteristikama:

1. Rezultati nad trening podacima:

   (a) Veličina populacije za ovaj model je 179 instanci, a procenat tačno klasifikovanih 97%.

   (b) Ukupan kvalitet modela (tačnost) je 0.946, tačnost klasifikacije palindroma neuređenih regiona je 0,926, a tačnost klasifikacije palindroma uređenih regiona je 1.00.

   (c) Modalna vrednost klase palindroma neuređenih regiona je palindrom KDDK, dok je modalna vrednost klase palindroma uređenih regiona palindrom EEEE. Osim ovog palindroma, palindromi koji se javljaju u ovom klasifikacionom modelu kao palindromi karakteristični za neuređene regione su palindromi: SKET(TEKS), PSYSP, TNTG(GTNT), VPVP(PVPV), SDSDSDSDSDSDS, PAAP, QKTKQ, ATTTA, QQQQ, DNRND, EQQE, TGGT, EDEDE, GPPPG, PVPV, TKKQKKT, EKRKE,...

   (d) Palindromi koji su najznačajniji za klasifikaciji dati su na slici 9.

2. Rezultati nad test podacima:

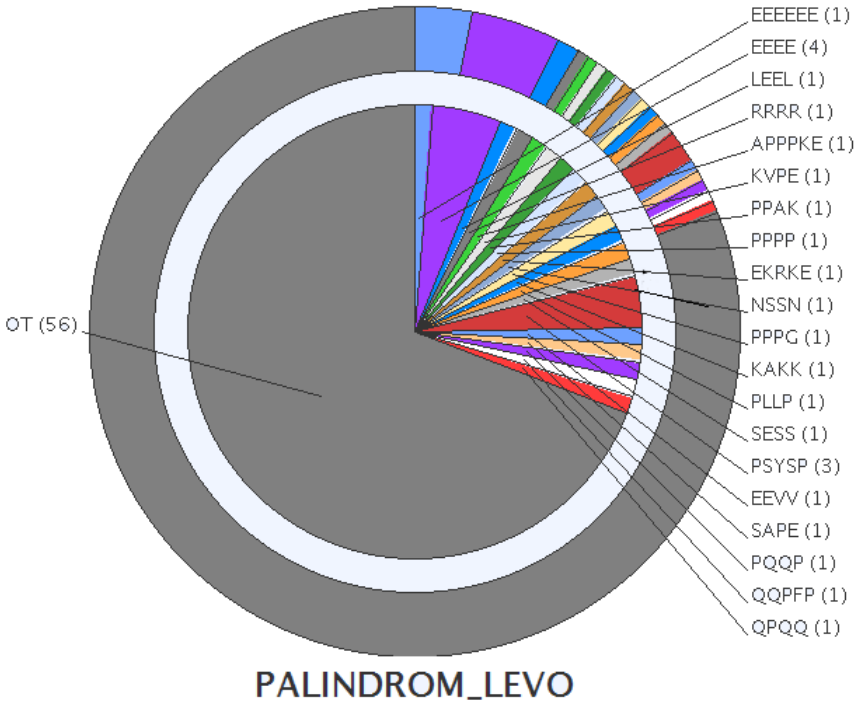
   (a) Tačnost modela nad test podacima je 0.245 (35% tačno klasifikovanih od ukupno 8817 instanci).

   (b) Tačnost klasifikacije palindroma uređenih regiona je 1.0, a tačnost klasifikacije neuređenog regiona je 0.351.

3. Rezultati primene modela nad proteinima koji predstavljaju razliku verzije 7.03 i verzije 6.02:

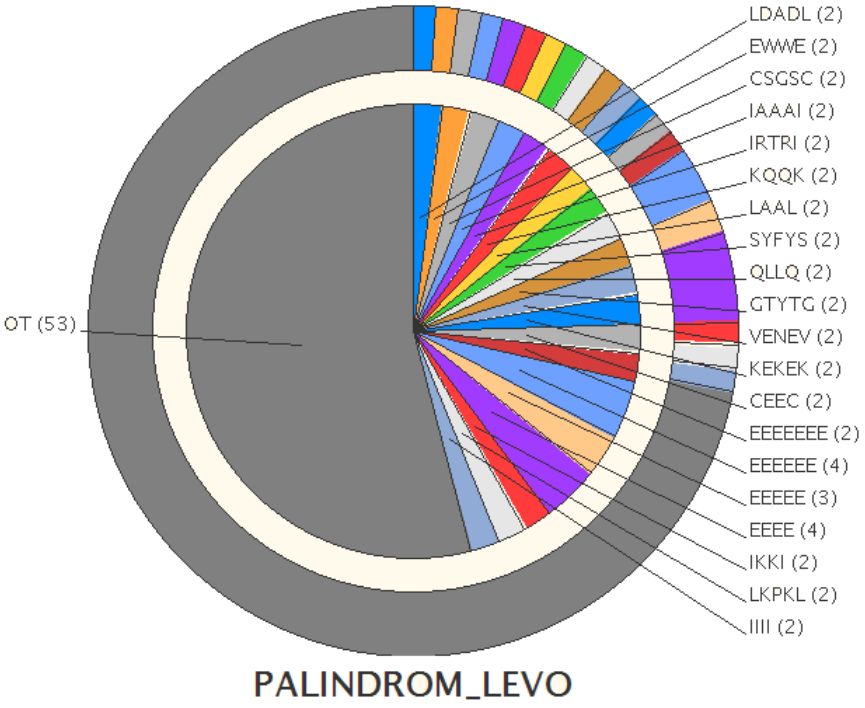
   (a) Tačnost modela nad test podacima je 0.019 (0% tačno klasifikovanih od ukupno 3.061 instanci).

   (b) Tačnost klasifikacije palindroma neuređenih regiona je 0.005, a tačnost klasifikacije uređenog regiona je nedostupna zbog nepostojanja informacija o uređenim regionima u verziji 7.03.



Slika 9: Palindromi dobijeni klasifikacijom palindroma dužine veće od 3 karakteristični za klasu D (eng. *disorder*, odnosno klasa neuređenih delova proteina).

Dok je tačnost klasifikacije nad trening podacima veoma dobra, tačnost nad test podacima je vrlo mala. Jedan od razloga je i mali broj uređenih regiona, što onemogućava modelu da bude dobro utreniran za dobru klasifikaciju i jednih i drugih. Primenjivanje modela nad proteinima koji predstavljaju razliku između verzije 7.03 i 6.02 daje jako loše rezultate gde uzrok može da bude isti. Kako je tačnost modela, i nad trening i nad test podacima za klasu uređenih regiona, 1.0, kao indikator neuređenog regiona može da bude odsustvo palindroma koji su karakteristični za uređene regione iz ovog modela. Takvi palindromi dati su na slici 10.



Slika 10: Palindromi dobijeni klasifikacijom palindroma dužine veće od 3 karakteristični za klasu O (eng. *order*, odnosno klasa uređenih delova proteina).

Na kraju je napravljen klasifikacioni model za sve palindrome uređenih i neuređenih regiona proteina DisProt baze i dobijen je model sa sledećim karakteristikama:

1. Rezultati nad trening podacima:

(a) Veličina populacije za ovaj model je 13714 instanci, a procenat tačno klasifikovanih 75%

(b) Ukupan kvalitet modela (tačnost) je 0.623, tačnost klasifikacije palindroma neuređenih regiona je 0,827, a tačnost klasifikacije palindroma uređenih regiona je 0.67.

(c) Modalna vrednost klase palindroma neuređenih regiona je palindrom EE, dok je modalna vrednost klase palindroma uređenih regiona palindrom LL. Osim ovog palindroma, palindromi koji se javljaju u ovom klasifikacionom modelu kao palindromi karakteristični za neuređene regione su palindromi: KLK, KK, AEA, PAP, PEE, EKK, PKK, VPP, EEV, VPKKPV, NAAAAAN, APA, PSYSP, EQQR, QGQ, EED, ILKLI, ADA, SKE, TVE, KDDK, VES, KKE, SKS, QQQ, PQQ, ESE, GKEEE, GGGGGGGG, DDK, ESK, ...

(d) Gore su izdvojeni neki palindromi iz modela klasifikacije koji se javljaju po nekoliko puta više u odnosu na broj pojavljivanja u uređenom regionu ili se javljaju veći broj puta u neuređenim regionima, a nema ih u uređenim regionima. Svi karakteristični palindromi klase neuređenih regiona dati su na slici 11.

2. Rezultati nad test podacima:

   (a) Tačnost modela nad test podacima je 0.743 (81% tačno klasifikovanih od ukupno 376230 instanci).

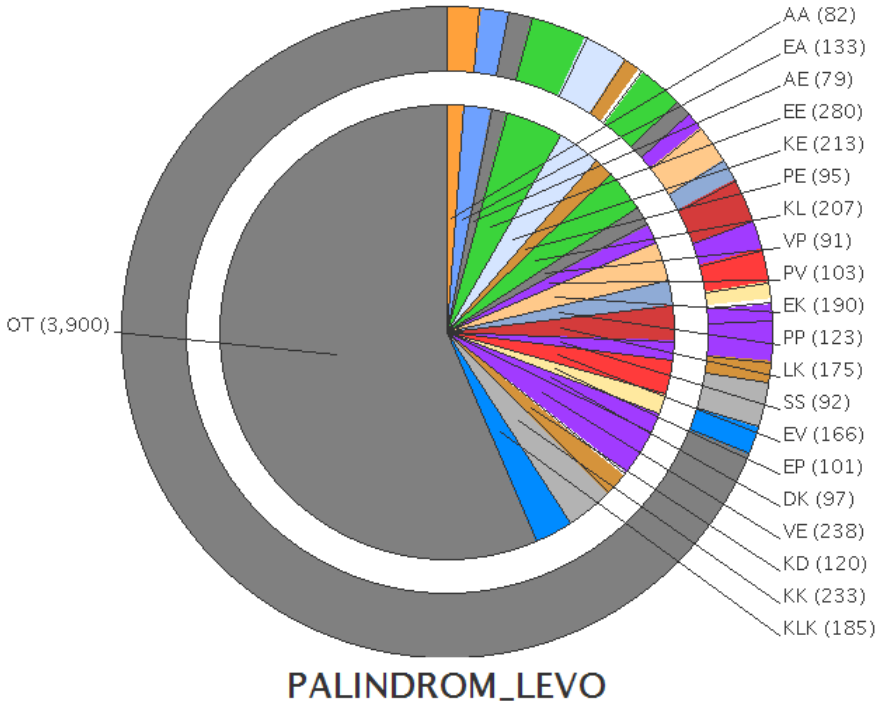
   (b) Tačnost klasifikacije palindroma uređenih regiona je 0.628, a tačnost klasifikacije neuređenog regiona je 0.809.

3. Rezultati primene modela nad proteinima koji predstavljaju razliku verzije 7.03 i verzije 6.02:

   (a) Tačnost modela nad test podacima je 0.915 (83% tačno klasifikovanih od ukupno 3061 instanci).

   (b) Tačnost klasifikacije palindroma neuređenih regiona je 0.831, a tačnost klasifikacije uređenog regiona nije dostupna zbog ne postojanja informacije o order regionima u verziji 7.03.

Po rezultatima nad test podacima može se videti da se dobro klasifikuju palindromi i uređenih i neuređenih regiona, pa samim tim se može reći da zaista postoje



Slika 11: Palindromi dobijeni klasifikacijom palindroma karakteristični za klasu D (eng. *disorder*, odnosno klasa neuređenih delova proteina).

kraći palindromi koji su karakteristični za uređene, odnosno neuređene regione proteina. Model jako dobro klasifikuje i neuređene regione proteina koji predstavljaju razliku verzija 7.03 i 6.02 što dokazuje postojanje veze između neuređenih delova proteina i palindroma. Kako model, zasnovan na svim palindromima iz baze gde u najvećoj meri učestvuju kratki palindromi, pokazuje najbolje rezultate nad ovim proteinima, dodatno govori da kratki palindromi mogu da budu indikatori neuređenosti proteina.
Takođe, ovakav model nam ne daje nove informacije o palindromima u neuređenim regionima od modela koji je rađen nad palindromima dužine manje od pet. Najviše iz razloga što kraći palindromi jesu zastupljeniji od dužih palindroma, pa tako sam model ne može da bude zasnovan nad vrednostima atributa koji su retki. Da bi se dobili duži palindromi koji su karakteristični samo za neuređene regione, SQL upitima su izdvojeni neki karakteristični palindromi. SQL upitom su izdvojeni palindromi dužine veće od 9 amino kiselina sa njihovim dužinama i brojem ponavljanja. Ovakvi palindromi su uzeti kao interesantni zbog dovoljno velike dužine i broja ponavljanja u proteinima. Palindromi dužine veće od devet amino kiselina su predstavljeni u tabeli 14.

U tabeli 14 su dati palindromi koji se javljaju više od jednog puta čime imaju veći značaj zbog veće verovatnoće neslučajnog pojavljivanja. Ostali palindromi dužine veće od devet koji se javljaju, takođe, mogu da okarakterišu neuređene regione proteina, naročito zbog svoje dužine (najduži palindrom se sastoji od 22 amino kiseline). Ovakvi palindromi velikih dužina zasigurno imaju značajne uloge u funkciji proteina i oni su dati u tabeli 15.

Nad dobijenim modalnim vrednostima u klasifikacionim modelima, koji karakterišu neuređene regione, SQL upitima je prebrojano koliko se puta ti palindromi javljaju i u koliko različitih proteina. Kako bi okarakterisali neuređene regione sa manjim ali pouzdanijim skupom palindroma, na palindrome je primenjen dodatni uslov - takav palindrom ne sme da se nalazi i u uređenom regionu nekog proteina, kao i da je broj pojavljivanja tog palindroma veći od nekog postavljenog mininmalnog praga. Programom, napisanom u Java programskom jezuku dobijeni su rezultati predstavljeni u tabelama 16, 17 i 19.

Palindrome iz tabele 17 koji se javljaju u više različitih proteina možemo da smatramo palindromima koji bliže karakterišu neuređene regione nego oni koji se javljaju samo u jednom proteinu zbog postojanja veće verovatnoće slučajnog pojavljivanja. Iako i takvi palindromi mogu da imaju značajne biološku i strukturalnu ulogu.

Tabela 14: Palindromi čija je dužina veća od 9 amino kiselina čiji je ukupan broj ponavljanja veći od 1.

Palindrom	Dužina	Broj ponavljanja
EEEEEEEEEE	10	2
VAGAAAAGAV	10	2
EEAEAEAEAEE	11	2
EEEEEDEEEEE	11	2
NNNNSSSNNNN	11	2
SSSSFQFSSSS	11	2
EDEEEEEEEEDE	12	2
PSYSPSSPSYSP	12	2
QGYAQTTQAYGQ	12	2
ATTTAAAAATTTA	13	2
GGGGGGGGGGGGG	13	2
EDEEEEEEEEEEDE	14	2
DDDDDDDEDDDDDDD	15	2
APAPAPAPAPAPAPAPAPA	19	2
PAPAPAPAPAPAPAPAPAP	19	2
GGGGGGGGGGGGGGGGGGGGGGGG	24	2
SDSDSDSHSDSDSDSHSDSHSDSDSDSHSDSDSDS	35	2

Tabela 15: Palindromi čija je dužina veća od 9 amino kiselina sa ukupno jednim pojavljivanjem.

Palindrom	Dužina
AAAGAAGAAA	10
EVGAEEAGVE	10
EEKEEAEEKEE	11
EKKPVPVPKKE	11
EEEEEGGEEEEE	12
GAAAGSASGAAAG	13
APAPAAPTPAAPAPA	15
QPQQPFPQQPQQPFPQQPQ	19
QQQQQQQQQQQQQQQQQQQQQQ	22

8.2.2 Klasifikacioni modeli zasnovani na ponovcima

Ista vrsta istraživanja urađena je i nad ponovcima.

Tabela 16: Palindromi dobijeni klasifikacijom karakteristični za neuređene regione sa brojem pojavljivanja većim od 50 koji se javljaju i u uređenim delovima.

Palindrom	Dužina	Pojavljivanja u neuređenom delu	Pojavljivanja u uređenom delu	Broj proteina
KLK	3	16694	5	109
EEEE	4	336	2	69
EEEEE	5	73	2	29

Tabela 17: Palindromi dobijeni klasifikacijom karakteristični za neuređene regione sa brojem pojavljivanja većim od 50 koji se ne javljaju u uređenim delovima.

Palindrom	Dužina	Broj pojavljivanja	Broj proteina
KDDK	4	2844	3
PSYSP	5	599	1
QQGQQ	5	230	1
QGQQQGQ	6	182	1
SSSS	4	165	18
PPPP	4	125	14
QQQQQ	5	102	10
KKKK	4	52	20
APPA	4	51	9

Klasifikacijom, čiji su ulazni podaci ponovci dužine manje od pet, dobijen je model sa sledećim karakteristikama:

1. Rezultati nad trening podacima:

   (a) Veličina populacije za ovaj model je 12452 instance, a procenat tačno klasifikovanih 76%

   (b) Ukupan kvalitet modela (tačnost) je 0.647, tačnost klasifikacije ripita neuređenih regiona je 0.803, a tačnost klasifikacije ripita uređenih regiona je 0.726.

   (c) Modalna vrednost klase ripita neuređenih regiona je ripit VE, dok je modalna vrednost klase ripita uređenih regiona ripit AL. Ostali ponovci koji su karakteristični neuređene regione su dati na slikama 12.

2. Rezultati nad test podacima:

   (a) Tačnost modela nad test podacima je 0.704 (74% tačno klasifikovanih od ukupno 350659 instanci).

   (b) Tačnost klasifikacije ripita uređenih regiona je 0.713, a tačnost klasifikacije ripita neuređenog regiona je 0.744.

3. Rezultati primene modela nad proteinima koji predstavljaju razliku verzije 7.03 i verzije 6.02:

   (a) Tačnost modela nad test podacima je 0.192 (12% tačno klasifikovanih od ukupno 2546 instanci).

   (b) Tačnost klasifikacije ripita neuređenih regiona je 0.117, a tačnost klasifikacije ripita uređenog regiona nije dostupna.

Model dobro klasifikuje ripite i uređenih i neuređenih regiona. Kao i kod palindroma, ripiti malih dužina su najzastupljeniji, pa je kvalitet modela isti kao i kod modela palindroma dužine manje od pet amino kiselina, Rezultati modela nad proteinima, koji predstavljaju razliku verzija 7.03 i 6.02, su loši.

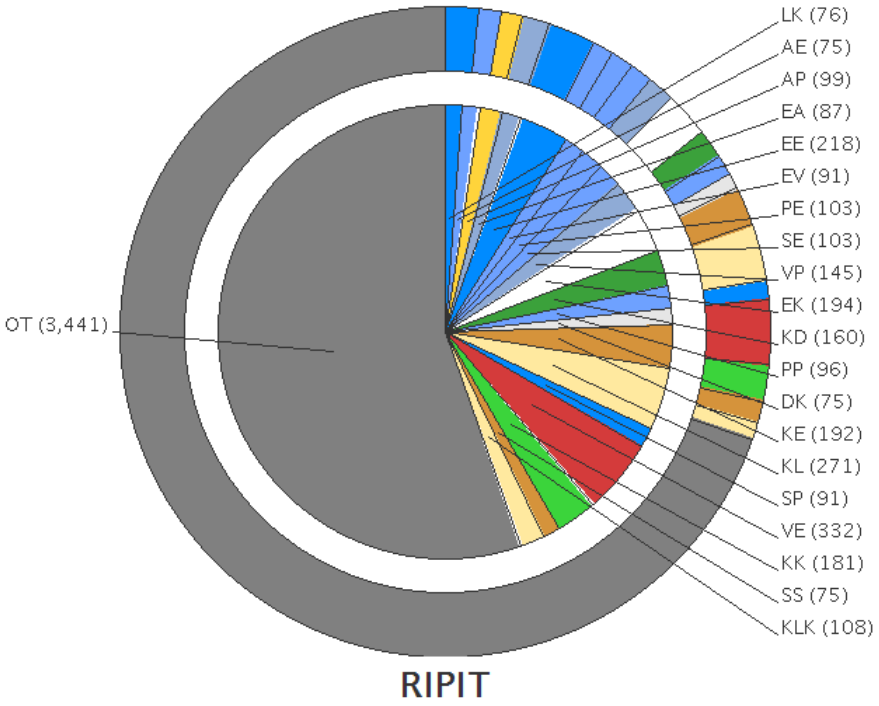
Kao i kod palindroma, klasifikacija palindroma pojedinačnih dužina nije bila moguća. Kako bi se dobili duži ponovci karakteristični za neuređene regione, urađena je klasifikacija ponovaka dužine veće od dve amino kiseline. Sužavanje klasifikacije, tako da u klasifikaciji učestvuju ponovci duži od tri amino kiseline, nije bilo moguće zbog nedostatka takvih ponovaka u uređenim regionima. Klasifikacijom, čiji su ulazni podaci ponovci dužine veće od dva, dobijen je model sa sledećim karakteristikama:

1. Rezultati nad trening podacima:

   (a) Veličina populacije za ovaj model je 720 instance, a procenat tačno klasifikovanih 97%

   (b) Ukupan kvalitet modela (tačnost) je 0.954, tačnost klasifikacije ripita neuređenih regiona je 0.953, a tačnost klasifikacije ripita uređenih regiona je 0.986.

   (c) Modalna vrednost klase ripita neuređenih regiona je ripit KLK, dok je modalna vrednost klase ripita uređenih regiona ripit AAA. Ostali ponovci koji su karakteristični za neuređene regione su dati na slici 13.



Slika 12: Ponovci dobijeni klasifikacijom ripita dužine manje od pet amino kiselina karakteristični za klasu D (*disorder*, odnosno klasa neuređenih delova proteina).

2. Rezultati nad test podacima:

   (a) Tačnost modela nad test podacima je 0.591 (59% tačno klasifikovanih od ukupno 51375 instanci).

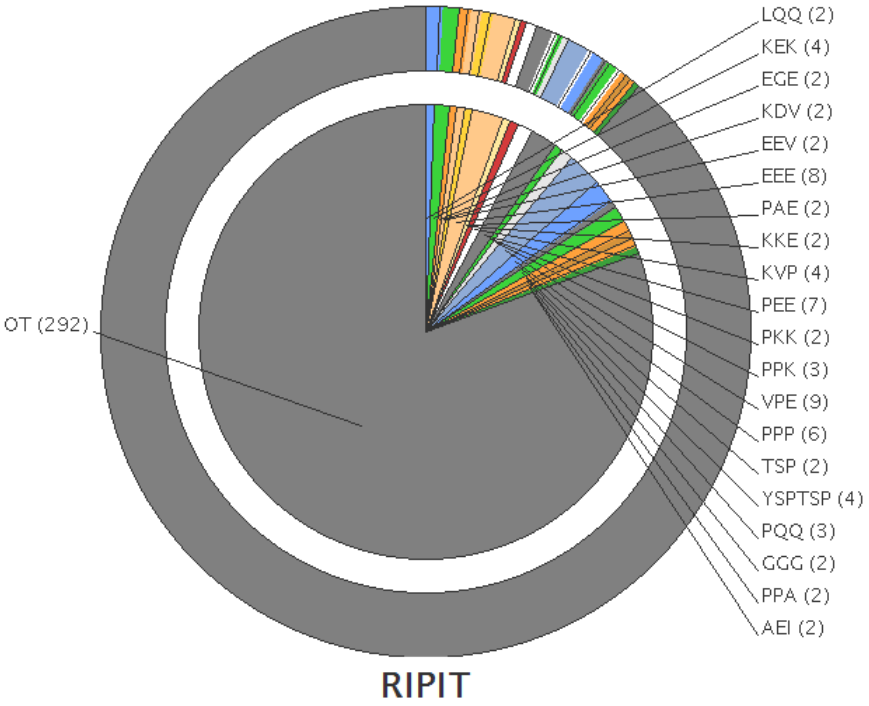
   (b) Tačnost klasifikacije ripita uređenih regiona je 0.518, a tačnost klasifikacije ripita neuređenog regiona je 0.942.

3. Rezultati primene modela nad proteinima koji predstavljaju razliku verzije 7.03 i verzije 6.02:

   (a) Tačnost modela nad test podacima je 0.726 (73% tačno klasifikovanih od ukupno 2546 instanci).

   (b) Tačnost klasifikacije ripita neuređenog regiona je 0.94, a tačnost klasifikacije ripita uređenog regiona nije dostupna.

Model dobro klasifikuje ripite neuređenih regiona, a lošije ripite uređenih delova, opet zbog nedostatka informacija o uređenim regionima. Model dobro klasifikuje i neuređene regione proteina koji predstavljaju razliku verzija 7.03 i 6.02 sa 94% tačnosti klasifikacije neuređenih regiona.

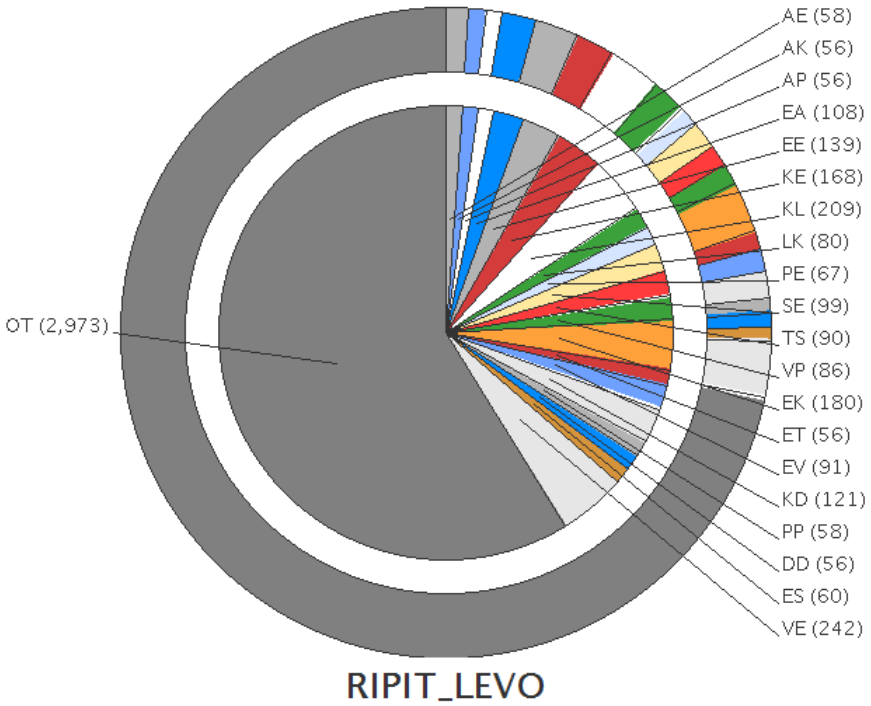


Slika 13: Ponovci dobijeni klasifikacijom dužine veća od dve amino kiselina karakteristični za klasu D (*disorder*, odnosno klasa neuređenih delova proteina).

Klasifikacijom, čiji su ulazni podaci svi ponovci, dobijen je model sa sledećim karakteristikama:

1. Rezultati nad trening podacima:
   (a) Veličina populacije za ovaj model je 12505 instanci, a procenat tačno klasifikovanih 77%
   (b) Ukupan kvalitet modela (tačnost) je 0.648, tačnost klasifikacije ripita u neuređenim regionima je 0.783, a tačnost klasifikacije ripita u uređenim regionima je 0.748.
   (c) Modalna vrednost klase ripita neuređenih regiona je ripit VE, dok je modalna vrednost klase ripita uređenih regiona ripit AL. Ostali ponovci koji su karakteristični za uređene, odnosno neuređene regione su dati na slici 14.
2. Rezultati nad test podacima:
   (a) Tačnost modela nad test podacima je 0.663 (72% tačno klasifikovanih od ukupno 358678 instanci).
   (b) Tačnost klasifikacije ripita uređenih regiona je 0.72, a tačnost klasifikacije ripita neuređenog regiona je 0.719.
3. Rezultati primene modela nad proteinima koji predstavljaju razliku verzije 7.03 i verzije 6.02:
   (a) Tačnost modela nad test podacima je 0.194 (12% tačno klasifikovanih od ukupno 2546 instanci).
   (b) Tačnost klasifikacije ripita neuređenih regiona je 0.124, a tačnost klasifikacije ripita uređenog regiona je nedostupna.

Model zasnovan na svim ponovcima lošije klasifikuje neuređene regione proteina koji su razlika proteina iz verzije 7.03 i 6.02 nego model koji je zasnovan na ponovcima koji su dužine veće od dva. Kako su u skupu, koji je dobijen izostavljanjem palindroma dužine manje od dve amino kiseline, najzastupljeniji ponovci dužine 3 i 4, sledi da model kao modalne vrednosti koristi baš te ripite, dok model zasnovan na svim ponovcima kao modalne vrednosti najviše koristi ripite dužine 2. To govori da najjača veza postoji među ponovcima dužine 3 i 4 sa neuređenim regionima proteina.



Slika 14: Ripiti dobijeni klasifikacijom rađenom na celim skupom karakteristični za klasu D (disorder, odnosno klasa neuređenih delova proteina).

Klasifikacijom ripita nisu dobijeni ponovci većih dužina kao pri klasifikaciji palindroma, stoga su, među ponovcima, SQL upitima izvojeni duži ponovci čiji je broj pojavljivanja veći od 50, dužine veće od 9 amino kiseline i razdvojeni su na one koji se nalaze i u uređenim delovima i oni koji se ne nalaze u uređenim delovima proteina. Ponovci, koji su duži od 4 amino kiseline sa brojem pojavljivanja većim od 50, a koji se javjaju i u uređenim delovima, nema. Ovo može da ukazuje na specifičnost ovih ripita da se javljaju samo u neuređenim regionima, te tako neuređeni regioni mogu da se okarakterišu ovakvim ponovcima. Rezultati su predstavljeni u tabeli 18.

Tabela 18: Ponovci karakteristični za neuređene regione sa brojem pojavljivanja većim od 50 koji se ne javljaju u uređenim delovima proteina.

Ripit	Dužina	Broj pojavljivanja	Broj proteina
QGQQPGQGQ	9	78	1
SPSYSPTSP	9	73	1
KEKDDKLKQE	10	114	1
PAEIVEQKDV	10	90	1
KDDKLKQEAD	10	76	1
PVESKETSEV	10	63	1
DKLKQEADAKL	11	68	1
DDKLKQEADAKL	12	372	1
VPETSAPTVEPT	12	113	1
KDDKLKQEADAKL	13	96	1
EKLAPVESKETSE	13	70	1
DDKLKQEADAKLKK	14	102	1
EKLAPVESKETSEV	14	102	1
KDDKLKQEADAKLKK	15	244	1
DDKLKQEADAKLKKE	15	155	1
KEKDDKLKQEADAKL	15	74	1
KEKDDKLKQEADAKLKK	17	110	1
EPTVEKLAPVESKETSEV	18	73	1
EKLAPVESKETSEVEPAEIVEQKDV	25	75	1

Osnovne razlike rezultata dobijeni klasifikacijom palindroma i ripita jeste da postoji mnogo veći broj dužih palindroma koji se javljaju u neuređenim delovima

u više različitih proteina nego što je to slučaj kod ripita. Međutim, kod ripita je karakteristično da se ni jedan ripit koji se nalazi u neuređenim regionima, a njegov ukupan broj pojavljivanja je veći od 50, se ne nalazi u uređenim regionima. Takođe, u slučaju palindroma, isti palindromi se javljaju u više različitih proteina, dok su ponovci, uglavnom, karakteristični za određen protein. Na kraju su, u tabelama 19 i 20, dati svi karakteristični palindromi i ponovci za neuređene regione proteina koji se ne javljaju u uređenim regionima proteina.

Tabela 19: Svi karakteristični palindromi neuređenih delova koji se ne javljaju u uređenim delovima sa ukupnim brojem ponavljanja većim od dva sa brojem različitih proteina u kojima se javljaju.

Palindrom	Dužina	Broj pojavljivanja	Broj proteina
EKDDK	5	208	1
SPSYSP	6	138	1
QGQQGQ	6	136	1
EATAEA	6	24	1
EVQQVE	6	24	1
GVVVVG	6	24	1
AAAAAA	6	23	7
KKSAAE	6	20	1
DDDDDD	6	16	3
ASKKAA	6	16	1
VPKKPV	6	15	1
DSKETS	6	12	1
ANEENA	6	11	1
QKLLKQ	6	11	1
QPLLPQ	6	10	1
PLVVLP	6	9	1
EDEEDE	6	8	3
RGRRGR	6	8	1
KEKKEK	6	5	2
DDLLDD	6	5	1
SDEEDS	6	4	1
LEKDDK	6	4	1
TSNNST	6	3	1
IIEEII	6	2	1
GPPGPPG	7	19	5

AEATAEA	7	18	1
SKKNKKS	7	11	1
EDPGPDE	7	9	1
EALTLAE	7	5	1
GGNGNGG	7	4	1
KDDGDDK	7	4	1
PKPEPKP	7	2	2
GQPGPAG	7	2	1
DDIHIDD	7	2	1
KPKAAKPK	8	9	1
KPKAAKPK	8	9	1
NGGCCGGN	8	6	1
NGGCCGGN	8	6	1
PQQPQQPF	8	5	1
PGPPGPPG	8	5	1
PQQPQQPF	8	5	1
PGPPGPPG	8	5	1
QQLQQLQQ	8	4	1
QQLQQLQQ	8	4	1
QQQAAQQQ	8	3	1
QQQAAQQQ	8	3	1
GPPGPPGPPG	10	7	3
AAAGAAGAAA	10	1	1
KKAAEVEAAKK	11	24	1
DDDDDDDDDDD	11	3	2
PPPPPPPPPPP	11	3	1
EDEEEEEEEEDE	12	4	1
PSYSPSSPSYSP	12	2	1
EDEEEEEEEEEEDE	14	2	1
APAPAAPTPAAPAPA	15	9	1
APAPAPAPAPAPAPAPAPA	19	2	1
PAPAPAPAPAPAPAPAPAP	19	2	1
QQQQQQQQQQQQ...QQQQQQ	22	8	2
QQQQQQQQQQQQQQ..QQQQQQ	26	3	1

Tabela 20: Svi karakteristični ponovci za neuređene delova koji se ne javljaju u uređenim delovima sa ukupnim brojem ponavljanja većim od dva sa brojem različitih proteina u kojima se javljaju.

Ripit	Dužina	Broj pojavljivanja	Broj proteina
ADAKL	5	213	1
QQPGQ	5	174	1
GQGQQ	5	172	1
VPETS	5	172	1
EAKSP	5	170	1
TVEPT	5	164	1
EQKDV	5	128	1
ESKET	5	108	1
EKLAP	5	90	1
KEKDD	5	87	1
KQEAD	5	80	1
QKEKD	5	68	1
QEGQL	5	65	1
SDSDS	5	62	1
DDKLK	5	56	1
EKDDK	5	54	1
YSPTSP	6	700	1
EADAKL	6	186	1
PAEIVE	6	144	1
DAKLKK	6	86	1
KDDKLK	6	76	1
VPETSA	6	58	1
SPSYSP	6	55	1
GYYPTS	6	52	1
KLKKEKD	7	144	1
KLKQEAD	7	142	1
QGQQPGQ	7	122	1
QQPGQGQ	7	92	1
DDKLKQE	7	84	1
SKETSEV	7	71	1
IVEQKDV	7	61	1
KDDKLKQ	7	57	1

KDDKLKQE	8	114	1
QQPGQGQQ	8	74	1
DDKLKQEA	8	70	1
ESKETSEV	8	63	1
KEKDDKLK	8	60	1
KQEADAKL	8	57	1
QGQQPGQGQ	9	78	1
SPSYSPTSP	9	73	1
KEKDDKLKQE	10	114	1
PAEIVEQKDV	10	90	1
KDDKLKQEAD	10	76	1
PVESKETSEV	10	63	1
DKLKQEADAKL	11	68	1
DDKLKQEADAKL	12	372	1
VPETSAPTVEPT	12	113	1
KDDKLKQEADAKL	13	96	1
EKLAPVESKETSE	13	70	1
DDKLKQEADAKLKK	14	102	1
EKLAPVESKETSEV	14	102	1
KDDKLKQEADAKLKK	15	244	1
DDKLKQEADAKLKKE	15	155	1
KEKDDKLKQEADAKL	15	74	1
KEKDDKLKQEADAKLKK	17	110	1
EPTVEKLAPVESKETSEV	18	73	1
EKLAPVESKETSEVEPAEIVEQKDV	25	75	1

9 Zaključak

Istraživanjem podataka nad DisProt bazom proteina, dobijeni su palindromi i ponovci koji karakterišu neuređene delove proteine, odnosno mogu da budu reprezentativan model za pretpostavku o neuređenosti proteina. Istraživanjem baze proteina došlo se do sledećih zaključaka:

a) Palindromi većih dužina se ne javljaju u uređenim delovima, pa se palindromima većih dužina mogu okarakteristati neuređeni regioni proteina. Na primer, palindromi dužine deset su u uređenim delovima proteina u značajnoj meri ređi nego što je to slučaj kod neuređenih regiona. Naročito ukoliko se posmatra broj palindroma manjih dužina u uređenim regionima. U neuređenim delovima, palindromi dostižu dužinu i do 87 amino kiselina.

b) U dobijenim klasifikacionim modelima najzastupljeniji su kratki palindromi. Na osnovu rezultata testiranja takvih modela nad novim proteinima, zaključuje se da dobijeni kratki palindromi mogu da okarakterišu neuređene regione.

c) Istraživanjem proteina i palindroma i ponovaka unutar proteina, zaključeno je da su palindromi karakterističniji za neuređene regione nego ponovci. Jedan isti palindrom se u proteinima javlja mnogo veći broj puta nego što je to slučaj kod ponovaka.

d) Kao sporedni rezultat, dobijeno je da se unutar jednog palindroma javlja mnogo manji broj različitih amino kiselina, dok to nije slučaj kod ponovaka. Ponovci su uglavnom raznovrsni u broju različitih amino kiselina.

Kao dodatno istraživanje, pokušano je sa idvajanjem ponavljajućih niski koje se javljaju u više od 90% proteina iz DisProt baze. Nažalost, takvih ponavljajućih niski nema. Najzastupljeniji palindrom jeste palindrom EE koji se javlja u 307 različitih proteina što je približno 44% od ukupnog broja proteina, ali mu se zbog njegove kratke dužine ne daje mnogo na značaju. Njegov ukupan broj pojavljivanja u svim proteinima je 22132 puta.

Klasifikacijom su izdvojeni palindromi koji mogu da okarakterišu neuređene regione i kao takvi, se u daljem radu mogu koristiti za naprednija istraživanja u hemijskim i biološkim oblastima. Ono što dodatno može da bude izdvojeno, kao značajno za dalja istraživanja, jesu hemijske osobine zastupljenih amino kiselina u neuređenim delovima. Posmatranjem palindroma i ponovaka, koji su dobijeni kao

modalne vrednosti koje opisuju neuređene regione, pronađene su osobine koje su zajedničke za amino kisline koje preovlađuju u izdvojenim ponavljajućim niskama. Na osnovu Venovog dijagrama, zajedničko za izdvojene amino kiseline jeste da pretežno spadaju u grupu malih i hidrofobnih ili malih i negativnih amino kiselina.

Literatura

[1] Vladimir N Uversky A Keith Dunker, Israel Silman and Joel L Sussman. Function and structure of inherently disordered proteins. *Current Opinion in Structural Biology*, 18:756–764, 2008.

[2] Arthur W. Adamson. *A Textbook of Polymer Chemistry*. A Subsidiary of Harcourt Brace Jovanovich, 1979.

[3] Essential and non-essential amino acids and derivates present in the bap network and determination of the putative importers and exporters required for their biosynthesis. http://journals.plos.org/plosone/article/file?type=supplementary&id=info:doi/10.1371/journal.pone.0029096.s004.

[4] Samira Eshafah Milos Beljanski Ana Jelovic, Nenad Mitic. Finding statistically significant repeats in nucleic acids and proteins. *poslato u časopis*, 2017.

[5] Ali Katanforoush Shahriar Arab Mehdi Sadeghi Hamid Pezeshk Changiz Eslahchi Armita Sheari, Mehdi Kargar and Sayed-Amir Marashi. A tale of two symmetrical tails: Structural and functional characteristics of palindromes in proteins. *BMC Bioinformatics*, 9:2008, 274.

[6] Florin Fulga Dan V. Nicolau Dan V. Nicolau Jr., Ewa Paszek. Mapping hydrophobicity on the protein molecular surface at atom-level resolution. *PLOS ONE*, 9:12, 2014.

[7] Disprot baza podataka. http://www.disprot.org/.

[8] Lina L. Faller. An Investigation of Palindromic Sequences in the Pseudomonas fluorescens SBW25 Genome. 2008.

[9] Craig Jankowski Farooq Nasar and Dilip K. Nag. Long Palindromic Sequences Induce Double-Strand Breaks during Meiosis in Yeast. *Molecular and Cellular Biology*, 20:3449–3458, 2000.

[10] Yu Feng 1 † Xiaoyun Wang 1 Jing Li 1 2 Wen Liu 1 Li Rong 1 Jianzong Li 1, † and Jinku Bao. An Overview of Predictors for Intrinsically Disordered Proteins over 2010–2014. *International Journal of Molecular Sciences*, 16:23446–23462, 2015.

[11] C. Kissinger J. E. Villafranca P. Romero, Z. Obradović and A. K. Dunker. Identifying disordered regions in proteins from amino acid sequence. In *Identifying Disordered Regions in Proteins from Amino Acid Sequence*. Proc. IEEE International Conference on Neural Networks, Huston, 1997.

[12] Promponas et al., 2000. http://www.sciencedirect.com/science/article/pii/S0969212603002351#BIB28.

[13] John D. Roberts and Marjorie C. Caserio. *Basic Principles of Organic Chemistry*. Benjamin, Inc., Menlo Park, CA, 1997.

[14] Ian M. Rosenberg. *Protein Structure*. Birkhäuser, Boston, MA, 1996.

[15] Francesca Diella Peer Bork Toby J Gibson Robert B Russell Rune Lindingm, Lars Juhl jensen. Protein Disorder Prediction: Implications for Structural Proteomics. *Elsevier Journal Finder*, 11:1453–1459, 2003.

[16] Kunchur Guruprasad Settu Sridhar, Mallapragada Nagamruta. Analyses of the Sequence and Structural Properties Corresponding to Pentapeptide and Large Palindromes in Proteins. *PLoS ONE*, 10, 2015.

[17] Swiss-prot baza podataka. http://www.uniprot.org/.

[18] Christopher J. Oldfield Vladimir N. Uversky and A. Keith Dunker. Intrinsically Disordered Proteins in Human Diseases:Introducing the D2 Concept. *Annual Review of Biophysics*, 37:215–246, 2008.

[19] Wootton, 1994. http://www.sciencedirect.com/science/article/pii/S0969212603002351#BIB41.