Univerzitet u Beogradu
Matematički fakultet

Master Rad

Razvoj metaprediktora za utvrđivanje neuređenosti proteina

Autor:
Una Stanković

Mentor:
dr Jovana Kovačević

Članovi komsije:

dr Jovana Kovačević
prof. dr Gordana Pavlović-Lažetić
prof. Nevena Veljković

Beograd, 2019.

Sažetak

Neuređenost proteina se utvrđuje eksperimentalno, laboratorijskim analizama, ili uz pomoć prediktora za automatsko predviđanje neuređenosti proteina. Laboratorijske analize spadaju u spore, veoma skupe metode, koje ne mogu da odgovore na potrebe akademske zajednice i industrije. Iz tog razloga, poslednjih godina, došlo je do razvoja velikog broja alata za automatsko predviđanje neuređenosti proteina. Zbog velike brojnosti ovih alata, razvijaju se metaprediktori koji predstavljaju njihove kombinacije. Specifičan cilj ovog rada je razvoj jednog metaprediktora za određivanje neuređenosti proteina koji bi konsenzusom objedinio rezultate najnovijih prediktivnih alata. Alat će biti testiran na skupu proteina sa eksperimentalno utvrđenom neuređenošću DisProt.

Zahvalnica

Zahvaljujem se, pre svega, svojoj porodici: Olgi i mami, za neizmernu podršku, ljubav i strpljenje pruženo tokom svih ovih godina. Nemoguće je rečima opisati zahvalnost koju osećam prema njima za sve što su mi pružile. Zahvaljujem se i svojim prijateljima koji su mi uvek bili podrška. Posebno bih se zahvalila i Anji Bukurov, sa kojom sam zajedno prošla kroz sve faze izrade naših master radova, Nikoli Ajzenhameru za sve LaTeXdetalje, i dr Tamari Vasić na pomoći oko prikupljanja literature i razumevanja bioloških osnova.

Sadržaj

Slike

Tabele

Olgi, tati i mami...

Uvod

Proteini su biološki makromolekuli neophodni za izgradnju i pravilno funkcionisanje ćelija i igraju mnogobrojne uloge u različitim procesima koji se odvijaju unutar organizma. Struktura proteina zavisi od redosleda aminokiselina i utiče na njegovu funckiju. Primarna struktura podrazumeva niz aminokiselina koje učestvuju u izgradnji proteina, dok se sekundarna odnosi na oblik koji protein zauzima u prostoru (spirala ili traka). Proteine sa nestabilnom sekundarnom strukturom nazivamo neuređenim. Pored značajne uloge u obavljanju brojnih bioloških funkcija, otkriveno je i postojanje veze između ovih proteina i razvoja neizlečivih bolesti i zbog toga su oni u fokusu bioinformatičke zajednice.

S obzirom na to da je priroda ovog rada multidisciplinarna, odnosno da pripada oblasti bioinformatike, neophodno je dati prigodan uvod koji bi čitaocu približio materiju. U prvom poglavlju date su biološke osnove bez kojih razumevanje motivacije, cilja i samog rada ne bi bilo moguće. Najpre je opisan protein, njegova struktura, značaj, funkcija i uloga u organizmu. Potom, poseban akcenat je stavljen na moguće strukture koje protein zauzima u prostoru i njihov izgled i uticaj. Na samom kraju poglavlja govori se o neuređenosti proteina iz biološkog ugla, kao uvodu u naredno poglavlje.

U drugom poglavlju, govori se o predikciji neuređenosti proteina iz ugla računarstva. Navedeno je nekoliko poznatijih prediktora, od kojih su neki korišćeni pri razvoju metaprediktora, kao i o bazama podataka DisProt i UniProt. Uloga DisProt baze, u ovom radu, leži pri testiranju preciznosti predikcije nad rezultatima koje su vratili prediktori, dok je značaj UniProt baze u prikupljanju niski potrebnih za rad prediktora.

Naredno poglavlje govori o samom metaprediktoru. Aplikacija koja je razvijena se posmatra iz više uglova, to su ugao arhitekture i organizacije, funkcionalnosti i iz ugla korisnika. Poslednje poglavlje opisuje postignute rezultate i moguća unapređenja.

Glava 1

Biološke osnove

U ovoj sekciji biće ukratko predstavljene biološke osnove neophodne za razumevanje rada i motivacije koja stoji iza određenih njegovih elemenata. Najpre, biće opisano šta su proteini, koje su njihove osnovne funkcije i kakva im je struktura, a zatim će posebno biti opisani neuređeni proteini, njihova uloga i uzroci koji mogu dovesti do njihove pojave.

1.1 Proteini

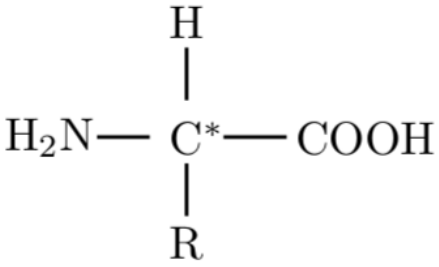
Proteini (grč. *protos* - "zauzimam prvo mesto") su biološki makromolekuli, koji čine 70% suve materije ćelija i neophodni su za njihovu izgradnju i pravilno funkcionisanje. Osim uloge u izgradnji ćelija, učestvuju u mnogobrojnim procesima koji se odvijaju unutar organizma. Predstavljaju najvažniji sastojak žive materije i utiču na brojnost i raznolikost živih bića. Specifičnost proteina je tolika da svaka biljna i životinjska vrsta ima svoje proteine, dok se, kod viših organizama, razlikovanje može uočiti i na individualnom nivou. Broj proteina u živim bićima je ogroman, na primer *E.coli* sa 3000 i čoveka sa 5 miliona proteina [1, 2].

Proteini su jedinjenja sačinjena od 100 ili više aminokiselina. Na osnovu broja aminokiselina koji ih čine peptidi se dele na:

- oligopeptide - sastoje se od 10 ili manje aminokiselina, među njih spadaju dipeptidi, tripeptidi, itd. i
- polipeptide - sastoje se od 100 ili manje aminokiselina.

Proteini se mogu posmatrati i kao nizovi nadovezanih polipeptida. Proteini i peptidi su izgrađeni od 22 aminokiseline [1] Sve proteinske aminokiseline su α-aminokiseline. Njih karakteriše to da su primarna amino i karboksilna grupa vezne za α-ugljenikov atom. Aminokiseline se međusobno razlikuju po strukturi bočnog R-ostatka, koji utiče na strukturu proteina. Opšta strukturna formula aminokiselina može se videti na slici 1.1 [2]. Proteinske aminokiseline (osim glicina) imaju asimetričan α-ugnjenikov atom i shodno tome mogu da se jave u dva oblika (prema Fišerovoj konvenciji) L i D. Sve standardne aminokiseline imaju L-konfiguraciju. Grafički prikaz može se videti na slici 1.2 [2].

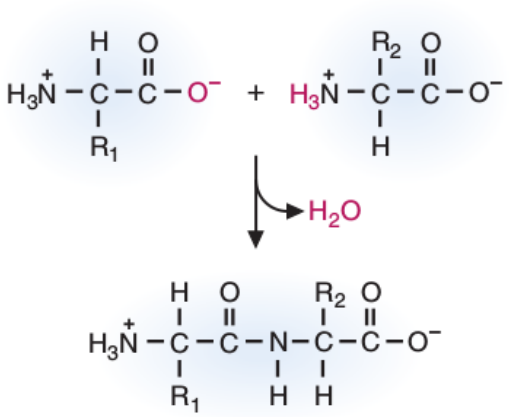
[1] Neki proteini u svom sastavu mogu da imaju 22 različite aminokiseline. Pored 20 standardnih aminokiselina (koji grade prirodne proteine), postoje i 2 nestandardne i to su Selenocistein (eng. *Selenocysteine*, simboli *Sec*, *U*) i Pirolizin (eng. *Pyrrolysine*, simboli *Pyl*, *O*). Ove dve aminokiseline se ređe javljaju [3].



Slika 1.1: Opšta strukturna formula aminokiselina [2].

Slika 1.2: Prikaz L i D prostorne konfiguracije [2].

U prirodi se pojavljuju $L-aminokiseline$ [2] i međusobno su povezane peptidnim vezama. Peptidne veze nastaju između α-karboksilne grupe jedne aminokiseline i α-amino grupe druge aminokiseline, pri čemu se oslobađa molekul vode (grafički prikaz se može videti na slici 1.3). Ovim postupkom nastaje nerazgranati polipeptidni lanac koji se sastoji od polipeptidne kičme i bočnih ostataka. Standardna grupa aminokiselina se može podeliti na esencijalne i neesencijalne, čiji spisak se može videti u tabeli 1.1 [3, 4].



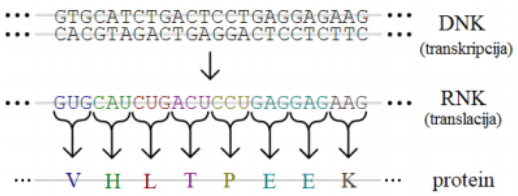
Slika 1.3: Prikaz spajanja α-karboksilne grupe jedne aminokiseline i α-amino grupe druge aminokiseline, pri čemu se oslobađa molekul vode [5].

[2] $L-aminokiseline$ su one aminokiseline sa levom prostornom konfiguracijom, analogno, postoje i $D-aminokiseline$, sa desnom

Esencijalne	Neesencijalne
Arginin	Alanin
Histidin	Asparagin
Leucin	Asparaginska kiselina
Izoleucin	Cistein
Lizin	Glutaminska kiselina
Metionin	Glutamin
Fenilalanin	Glicin
Treonin	Prolin
Triptofan	Serin
Valin	Tirozin

Tabela 1.1: Spisak esencijalnih i neesencijalnih aminokiselina

Svaki molekul proteina nastaje u ćeliji živog organizma. Redosled aminokiselina u proteinskoj sekvenci određen je redosledom aminokiselina u dezoksiribonukleinskoj kiselini (DNK), odnosno gena, koji predstavljaju trojke nukleotida. Svakoj takvoj trojci jedinstveno je pridružena po jedna aminokiselina na osnovu genetskog koda. Proces kojim se enkodirana informacija prevodi iz DNK u niz aminokiselina u proteinskom lancu, posredstvom glasničke (eng. *messenger*) ribonukleinske kiseline (RNK) i transportne (eng. *transfer*) RNK, naziva se genska ekspresija. Proces sinteze proteina predstavlja *centralnu dogmu molekularne biologije*, čiji se prikaz može videti na slici 1.4 [6].



Slika 1.4: Prikaz centralne dogme molekularne biologije [6]

Pri deobi ćelije dolazi do replikacije DNK, čime je obezbeđeno da ćelija nove generacije primi ceo skup informacija neophodnih za njeno normalno funkcionisanje i razvoj. U procesu replikacije može doći do greške, odnosno mutacije, a kao posledica toga javljaju se dva problema:

- novonastala greška se prenosi na sve buduće generacije
- mutacija u genu proteina dovodi do izmena na aminokiselinskoj sekvenci što može dovesti do smrti ćelije. Vrlo retko se dešava da mutacija dovodi do poboljšanja, ali kada se to desi najčešće se to dešava na nivou populacije, čime se prirodnom selekcijom "stari" gen menja u potpunosti.

[2]

1.1.1 Funkcije i osobine proteina

Pri istraživanju bioloških procesa neophodno je znati i dobro razumeti funkcije proteina. To se posebno može uočiti kod proučavanja oboljenja ljudi, ako se u obzir uzme činjenica da se mnoga oboljenja pojavljuju kao posledica funkcionalnih mutacija. Proteini su biološki najaktivniji molekuli sa velikim brojem esencijalnih funkcija koje se dele na:

- *dinamičke*, od kojih su najvažnije:
  1. transportna - prenos molekula (poput kiseonika, gvožđa, lipida) i hormona od mesta sinteze do mesta delovanja,
  2. biološka - regulacija metaboličkih procesa u ćeliji, kontrola i regulacija transkripcije gena i translacija,
  3. katalizatorska - biološka katalizacija [3],
  4. zaštitna - keratin, koagulacija krvi,
  5. održavanje zapremine tečnosti u organizmu,
- *strukturne*, od kojih su najvažnije:
  1. obezbeđivanje čvrstine i elastičnosti organa,
  2. davanje oblika organizmu,
  3. izgradnja strukturnih elemenata ćelije i
  4. bitna uloga u kontraktilnim i pokretnim elementima organizma.

Na slici 1.5 mogu se videti primeri nekih proteina i njihovih uloga.

Naziv proteina	Aktivnost / Funkcija / Nalaženje
Enzimi	kataliza svih reakcija u živim sistemima
Transportni proteini	
hemoglobin	transport kiseonika i ugljendioksida
mioglobin	transport kiseonika u mišićima
serum albumin	transport masnih kiselina, steroida, lekova...
Kontraktilni proteini	
miozin	pokretljivost mišića
aktin	
Zaštitni proteini	
imunoglobulini	stvaraju kompleks sa stranim telom
fibrinogen	prekursor fibrina pri zgrušavanju krvi
trombin	komponenta u zgrušavanju krvi
Hormoni	
insulin	reguliše metabolizam glukoze
hormoni rasta	stimulišu rast
Rezervni proteini	rezerva aminokiselina za mladu jedinku
ovalbumin	jaje
kazein	mleko
gliadin	pšenica
Strukturni proteini	
α-keratin	kosa, koža, krzno, nokti
fibroin	svila, paukova mreža
kolagen	vezivno tkivo

Slika 1.5: Primeri proteina [2]

Neke od karakteristika proteina koje su bitne u kontekstu strukture su:

- proteini grade kompleksna jedinjenja sa različitim supstancama po principu strukturne komplementarnosti i

[3] Katalizacija predstavlja proces povećavanja brzina reakcija

- proteini poseduju visoku osetljivost na različite agense koji ih denaturišu [4]. Neki od najčešćih agenasa su: visoka temperatura, pritisak, mehaničko tretiranje, dejstvo kiselina, baza, organskih rastvarača, materija, itd. [1, 6].

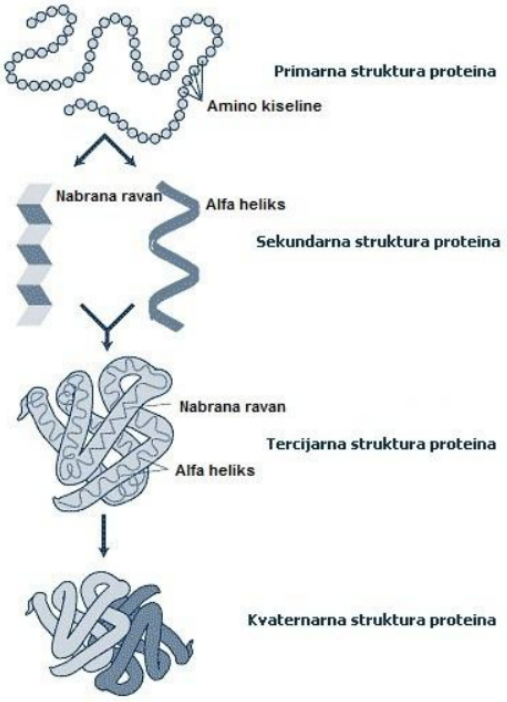
1.1.2 Struktura proteina

Osnovna struktura proteinskog molekula sastoji se od polipeptidnog niza aminokiselina povezanih peptidnom vezom. *Aminokiselinska* sekvenca je redosled kojim su povezane aminokiseline. Polipeptidni niz se spontano na različite načine uvija u kompleksnu trodimenzionalnu strukturu, koja se smatra najstabilnijom. Struktura proteina zavisi od redosleda aminokiselina i utiče na njegovu funkciju. Unutrašnjost takve strukture ima visoku gustinu, pa polipeptidni lanac ne dopušta promene u sastavu i zahteva prisustvo aminokiselina tačno određene veličine. Uobičajena raspodela aminokiselina u proteinima je daleko od ravnomerne. Neke aminokiseline se javljaju mnogo češće od ostalih, na primer, leucin se pojavljuje devet puta više od triptofana. [2, 4, 1]
Proteinsku strukturu održavaju različite vrste kovalentnih i nekovalentnih interakcija između hemijskih jedinjenja, na primer: vodonične, jonske, elektrostatičke, dipolne, itd.. Nabiranjem i uvijanjem lanaca kreiraju se različiti oblici proteina: vlaknasti, globularni ili eliptični. Strukturni proteini su vlaknasti, dok su oni koji pokazuju određenu aktivnost globularni. Ako mutacija dovede do toga da aminokiselina sa malim bočnim lancem bude zamenjena aminokiselinom sa velikim, pojaviće se problem u formiranju trodimenzionalne strukture. Ako bi se, pak, velika aminokiselina zamenila sa malom, pojavio bi se prazan prostor, što bi moglo dovesti do destabilizacije molekula proteina [2, 1, 4, 7].
Obično se struktura proteina posmatra u nivoima, pa tako postoji hijerarhijska strukturalna organizacija u četiri nivoa:

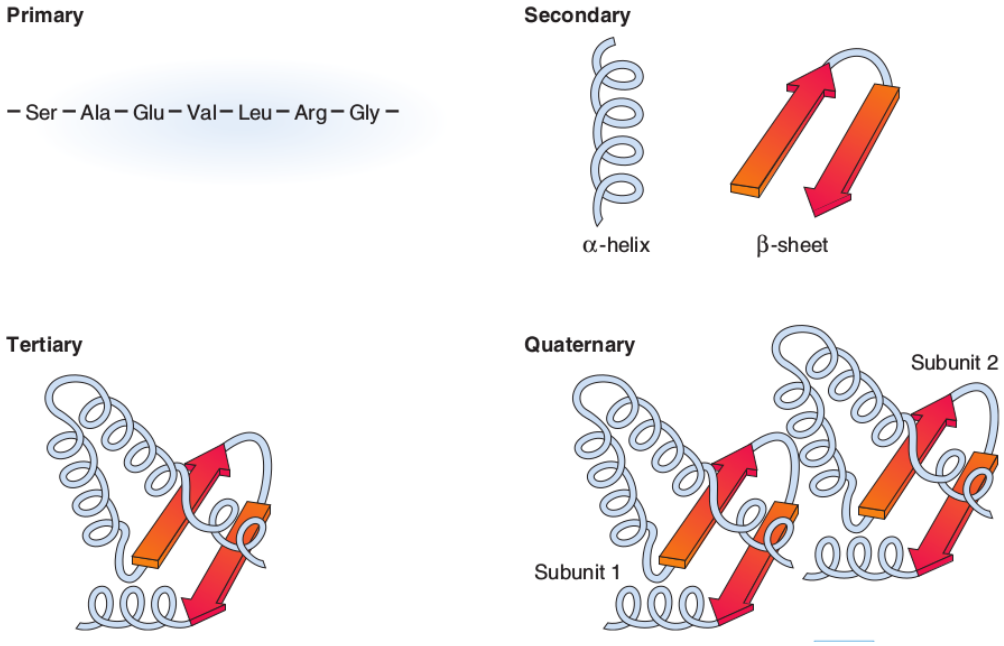
1. primarna,
2. sekundarna,
3. tercijarna i
4. kvaternarna.

Na slici 1.6 se može videti opšti prikaz mogućih struktura proteina, a na drugoj slici 1.7 šematski prikaz [1].

[4] Denaturacija proteina je proces koji izaziva promene u strukturi proteina, čime se menja i njihov fiziološki uticaj.



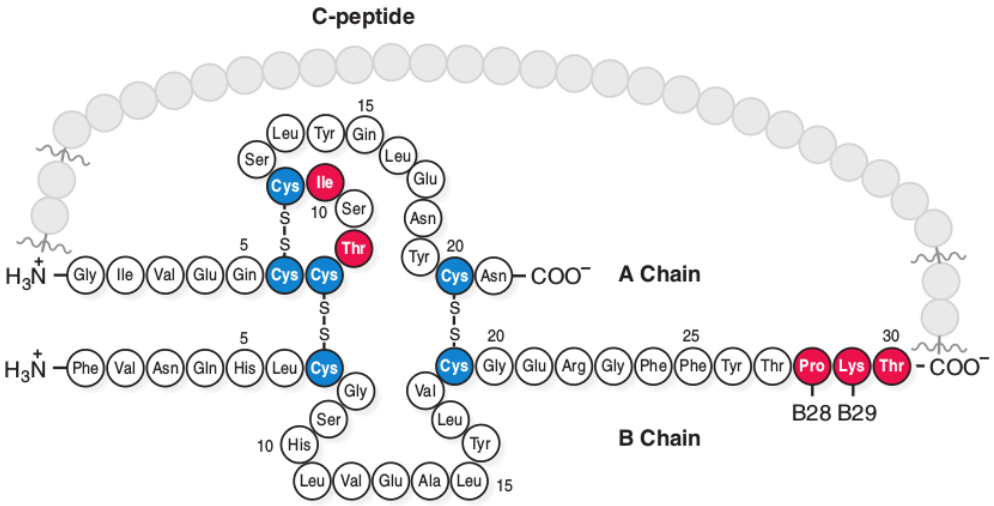
Slika 1.6: Prikaz struktura proteina



Slika 1.7: Šematski prikaz struktura proteina [5]

Određivanje sastava proteina u vidu aminokiselina je relativno jednostavno, dok je određivanje odnosa između sastava aminokiselina i strukture proteina komplikovano. Uprkos tome, često se mogu izvući korisni zaključci o strukturi proteina na osnovu aminokiselinskog sastava. [2]

Primarna struktura Predstavlja sâmu sekvencu aminokiselina[5] koje učestvuju u izgradnji proteina. Ova struktura ima ključni značaj za određivanje funkcije proteina zbog interakcija koje se javljaju između bočnih lanaca aminokiselina, a koji utiču na trodimenzionalnu strukturu. Proteini koji poseduju sličnu sekvencu aminokiselina nazivaju se *homologi*, a poređenje sekvenci među takvim proteinima može ukazati na genetsku relaciju između različitih vrsta.Prikaz izgleda primarne strukture na primeru insulina kod čoveka se vidi na slici 1.8 [1].

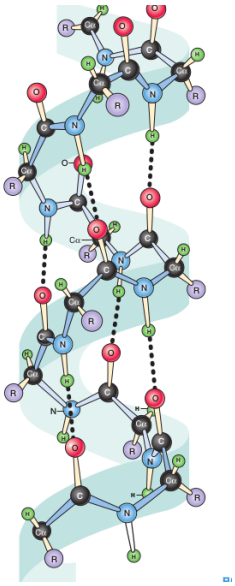


Slika 1.8: Prikaz primarne strukture [5]

Mnoge genetske bolesti rezultuju u proteinima sa poremećenim redosledom aminokiselina, što uzrokuje nepravilno presavijanje i gubitak ili nemogućnost normalnog funkcionisanja. Ukoliko su nam poznate strukture normalnih i mutiranih proteina, te informacije možemo iskoristiti za dijagnostikovanje ili proučavanje bolesti. Promene u primarnoj strukturi mogu imati uticaja i na više nivoe proteinskih struktura. Takve promene često dovode do lošeg presavijanja proteina i mogu dovesti do njegovog gubitka funkcije [8, 9].

Sekundarna struktura Odnosi se na oblik koji protein zauzima u prostoru i označava pravilno pojavljivanje ponavljanog prostornog rasporeda primarne strukture, u jednoj dimenziji. Ovu strukturu čini nekoliko različitih oblika, od kojih su najčešći α-heliks i β-presavijena traka (ili β-struktura), a čest je i tzv. β-okret [1, 7].
α-heliks - tip sekundarne strukture kod kog se gusto pakovani polipeptidni lanac spiralno uvrće. Karakteriše se brojem peptidnih jedinica po okretu i rastojanjem između dva okreta. Predstavlja najrasprostranjeniju sekundarnu strukturu i energetski je veoma siromašan iz čega se može zaključiti da je dosta stabilan. Javlja se kod globularnih i fibrilnih proteina. Heliks mogu obrazovati i $L-$ i $D-$ aminokiseline, pa postoje i dva tipa heliksa: levi i desni (u zavisnosti na koju stranu se navija, desni se navija u pravcu prstiju desne ruke kada se palac postavi u pravcu ose heliksa). Prikaz izgleda α-heliksa se vidi na slici 1.9 [1, 2].

[5]Redosled kojim su aminokiseline poređane u nekom polipeptidu se zove sekvenca aminokiselina [1].



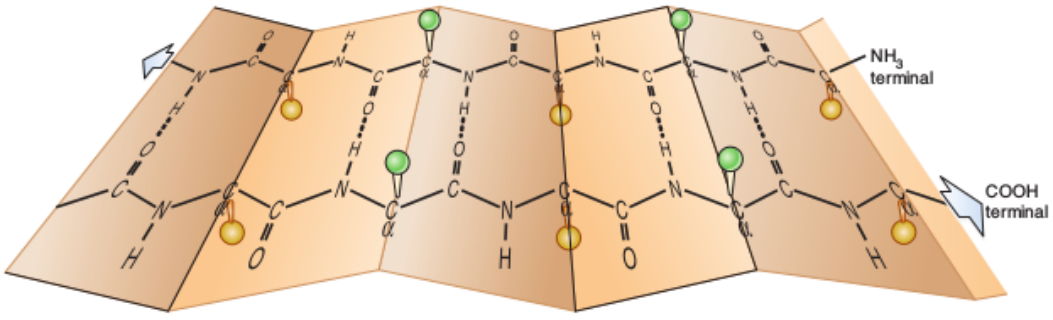
Slika 1.9: Prikaz α-heliksa [5]

β-struktura - Za razliku od α-heliksa, sastoji se od dva ili više peptidnih lanaca, ili segmenata polipeptidnih lanaca, a obrazuje se kada se ovakvi tipovi lanca povežu uzdužno vodoničnim vezama. Razlika između polipeptidnog niza u β-strukturi i potpuno istegnutog polipeptidnog niza je u tome što je kod β-strukture taj polipeptidni niz nabrane strukture. Postoje dva tipa β-struktura:

- paralelna - vodonično su vezani susedni polipeptidni nizovi istih smerova i
- antiparalelna - vodonično su vezani susedni polipeptdni nizovi suprotnih smerova.

Moguće su i mešovite paralelne-antiparalelne strukture. β-strukture se često javljaju u proteinima, a u globularnim se podjednako često javljaju i paralelne i antiparalelne. Sekundarna struktura se eksperimentalno utvrđuje na osnovu kristalne strukture proteina. Prikaz izgleda β-strukture se vidi na slici 1.10 [1, 2].

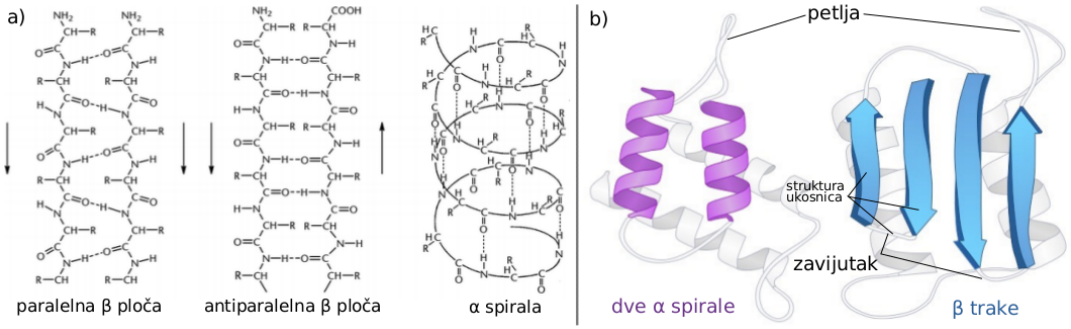
β-okreti - obrću pravac polipeptidnog lanca praveći kompaktan globularan oblik [9].



Slika 1.10: Prikaz β-strukture [5]

Prikaz izgleda sekundarnih struktura se nalazi na slici 1.11.

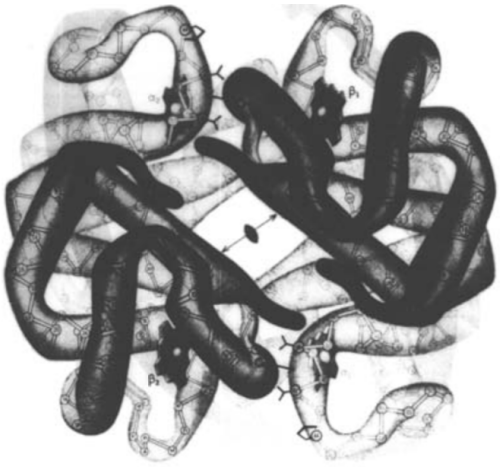
Tercijarna struktura Kod globularnih proteina se polipeptidni niz uvija u kompaktnu globulu. Tercijarna struktura proteina predstavlja unutarmolekularno slaganje polipeptidnog lanca u kompaktnu trodimenzionalnu strukturu specifičnog oblika



Slika 1.11: Prikaz sekundarnih struktura [10]

(globule), koja nastaje prostornim organizovanjem polipeptidnog lanca, sa sekundarnom strukturom. Na taj način se približavaju ostaci aminokiselina koji su udaljeni u primarnoj strukturi. Tercijarna struktura predstavlja način organizacije, odnosno rasporeda, sekundarnih struktura i položaj bočnih ostataka aminokiselina. Proteini ove strukture su globularni i kompaktni sa velikom gustinom u središtu. Poznavanje ove strukture proteina predstavlja osnovu za izučavanje funkcije i aktivnosti proteina. Kako bi se eksperimentalno utvrdila ova struktura vrši se rendgenska strukturna analiza. [1, 7, 2].

Kvaternarna struktura Predstavlja agregaciju više peptidnih lanaca u molekulu proteina. Mnogi proteini, posebno oni velike mase, izgrađeni su od nekoliko polipeptidnih lanaca. Svaka takva komponenta naziva se *podjedinica* ili *protomer*. Oni mogu biti identični[6] ili se razlikovati prema strukturi. Ovakav raspored dovodi do brzog i efikasnog transfera supstrata od jednog aktivnog centra enzima do drugog. Prikaz proteina sa kvatenarnom strukturom može se videti na 1.12 [1, 7]. Postoji



Slika 1.12: Prikaz hemoglobina, predstavnika globularnih proteina sa kvatenarnom strukturom [2]

nekoliko razloga iz kojih se kvaternarna struktura javila:

- Kompleksnija uloga zahteva kompleksniju strukturu

[6]Tada takve proteine nazivamo *oligomerima*

- Veća efikasnost katalitičkih procesa - katalizacija niza reakcija u metaboličkom procesu vrši se enzimima spojenih u multienzimske komplekse
- Viši nivo može da utiče na niži nivo strukture, pa tako kvaternarna struktura može da utiče na tercijarnu strukturu što se ogleda u njihovoj aktivnosti. Time se uvodi kooperativnost među subjedinicama. Posledica ovoga je regulacija i kontrola važnih biohemijskih procesa u ćeliji.
- Efikasnija biosinteza i lakše odstranjivanje grešaka pri procesu biosinteze.

Da bismo mogli da izučavamo kvaternarnu strukturu neophodno je obratiti pažnju na nekoliko bitnih aspekata. Prvi se odnosi na *stehiometriju*, odnosno, tip i broj podjedinica koje čine kvaternarnu strukturu. Drugi se odnosi na *geometriju*, odnosno, raspored podjedinica geometrijski, kao i tipove simetrije. Treći aspekt je *stabilnost kvaternarne strukture*. Stabilnost se odnosi na energetske aspekte interakcija i prirode kontakata između podjedinica. Naredni aspekt je *funkcionalni*, odnosno kako komunikacija među podjedinicama utiče na biološku funkciju. Poslednji aspekt odnosi se na *komunikaciju među podjedinicama*. [2]

1.1.3 Savijanje proteina

Izučavanje uvijanja i razvijanja proteina doprinosi razumevanju nastanka određene strukture proteina. Interakcije između lanaca aminokiselina koji se nalaze sa strane, određuju kako se dugački polipeptidni lanac presavija u trodimenzionalni oblik funkcionalnog proteina. Presavijanje proteina koje se događa u ćeliji traje od nekoliko sekundi do nekoliko minuta. Na slici 1.13 se može videti opšti prikaz savijanja proteina.

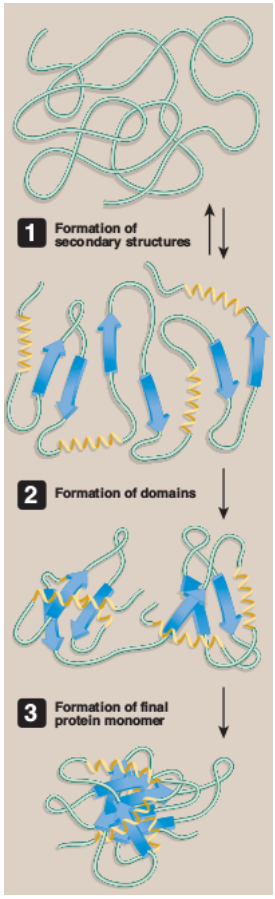
1.1.4 Denaturacija proteina

Denaturisanje proteina rezultuje u odvijanju i dezorganizaciji proteinske sekundarne i tercijarne strukture. U idealnim uslovima, denaturisanje proteina može biti *reverzibilno*. To znači da bi se protein, pri prestanku delovanja agenasa, vratio u normalno stanje. Međutim, većina proteina ostaje trajno neuređena. O neuređenosti proteina biće više reči u nastavku.
Jedno od objašnjenja zašto se protein ne vraća u originalno stanje se sastoji u tome da protein počinje sa savijanjem pre nego što se izvrši sinteza celog lanca. Osim toga, specijalizovana grupa pomoćnih proteina (engl. *chaperones*) je neophodna za pravilno savijanje mnogih vrsta proteina. Ovi pomoćni proteini interaguju sa polipeptidima u nekoliko faza tokom procesa savijanja, imaju ulogu u tome da održavaju protein nesavijenim dok sinteza nije gotova, ili imaju ulogu katalizatora. Loše savijanje proteina može dovesti do različitih bolesti kao što su: amiloidna bolest ili Prionova bolest [9].

1.2 Neuređenost proteina

Eksperimentalnim utvrđivanjem sekundarne strukture proteina (koje će biti detaljnije opisano u 1.2.1) uočeno je da se neretko, pod određenim fiziološkim uslovima, javljaju proteini sa trodimenzionalnom strukturom koja nije dobro definisana. Neuređenost predstavlja inherentno[7] svojstvo sekvence. Neuređen može biti ceo protein, a mogu biti neuređeni određeni regioni proteina različitih dužina. Kao posledica,

[7]Inherentno = nasleđeno



Slika 1.13: Prikaz savijanja proteina [9]

ovakve proteine nazivamo inherentno neuređenim proteinima, skraćeno IDP[8], a ako su u pitanju neuređeni, ali funkcionalni, regioni, onda je skraćenica IDPr[9]. Strukturalni poremećaji su česti kod viših eukariota. Kod ljudi, čak trećina svih proteina ima neuređenu strukturu. Neuređeni proteini su uključeni u procese stvaranja mnogih bolesti poput raka, neurodegenerativnih i kardiovaskulatnih bolesti, dijabetesa, brojnih neuronskih oboljenja i drugih. Statističkom analizom došlo se do zaključka da se aminokiseline mogu klasterovati na dve grupe:

1. aminokiseline koje promovišu uređenost (eng. *order promoting*) i
2. aminokiseline koje promovišu neuređenost (eng. *disorder promoting*).

Neuređene proteine ili neuređene regione je teško kategorizovati, a jedan od opštih opisa strukture dat je kao kombinacija više tipova foldona[10]:

- foldon (eng. *foldon*) je nezavisno organizujuća jedinica(region) proteina,
- indukativni foldon (eng. *inducible foldon*) je neuređeni region proteina koji savijanje lanca postiže barem delom vezivajući se za partnera,
- ne-foldon (eng. *non-foldon*) je neuređeni region proteina koji nikada ne postiže uređenost,

[8] eng. *Intrinsically Disordered Proteins*
[9] eng. *Intrinsically Disordered Protein Regions*
[10] Foldon ostaje u originalnom nazivu, kao posledica manjka literature. [10]

- polu-foldon (eng. *semi-foldon*) je neuređeni region proteina koji ostaje polovično neuređen i nakon vezivanja za partnera, i
- anti-foldon (eng. *unfoldon*) je region proteina koji iz uređenog prelazi u neuređeno stanje u cilju izvršavanja neke funkcije.

Postoji nekoliko mogućih stanja (oblika) u kojima se protein može naći. Ova stanja i prelazi između njih(neki proteini mogu prelaziti iz neuređenog u uređeno stanje, i obratno), prema *hipotezi proteinskog trojstva*, utiču na funkciju proteina. Svaki od mogućih oblika proteina može biti njegovo prirodno stanje i imati uticaja na njegovu ulogu u ćeliji. Proteini se mogu pojavljivati u raznim oblicima:

1. uređen protein,
2. topljiva globula (eng. *molten globule*),
3. pre-topljiva globula (eng. *pre-molten globule*) i
4. nasumično klupko (eng. *random coil*).

Neuređenost proteina se utvrđuje eksperimentalno, laboratorijskim analizama, ili uz pomoć prediktora za automatsko utvrđivanje neuređenosti [6, 11, 12, 13, 14, 15].

1.2.1 Eksperimentalno ispitivanje neuređenosti proteina

Eksperimentalno utvrđivanje neuređenosti proteina podrazumeva laboratorijsko utvrđivanje neuređenosti korišćenjem raznih biofizičkih i biohemijskih tehnika i njihovih kombinacija. Ono spada u veoma skupe i spore metode koje ne mogu da odgovore na izazove akademije i industrije. Uprkos tome, razvijen je veliki broj metoda za karakterizaciju strukture i osobina proteina. Svaka eksperimentalna metoda karakteriše se raznim prednostima manama i nivoom pouzdanosti, zbog čega je najbolje kombinovati dobijene rezultate. Naredne eksperimentalne, biofizičke i biohemijske, tehnike su najčešće u ispitivanju neuređenosti proteina [11, 6]:

- Kristalografija X-zracima(eng. *X-ray crystallography*),
- Spektroskopija nuklearnom magnetnom rezonancom (eng. *NMR spectroscopy*),
- Cirkularni dihroizam (eng. *Circular dichroism (CD) spectroscopy*),
- Osetljivost na proteolizu (eng. *Sensitivity to proteolysis*),
- Ramanova optička aktivnost, itd.

Navedene metode neće biti detaljnije obrazlagane jer prevazilaze domene ovog rada.

1.2.2 Računarsko ispitivanje neuređenosti proteina

Kao posledica osobina eksperimentalnog ispitivanja neuređenosti, veliki napori su uloženi u razvoj prediktora za računarsko utvrđivanje neuređenosti proteina. Ovi prediktori uz pomoć računara, korišćenjem tehnika mašinskog učenja, vrše utvrđivanje neuređenosti proteina. Iz godine u godinu, broj ovih prediktora je sve veći, a u poslednje vreme se radi i na kreiranju metaprediktora, koji predviđanje vrše kombinovanjem više tehnika. O ovoj vrsti predikcije biće više reči u narednom poglavlju.

Glava 2

Predikcija neuređenosti proteina

Razvitak istraživanja o neuređenim proteinima počinje oko 1978. godine, kada sa razvojem kristalografije X-zracima i spektroskopije nuklearnom magnetnom rezonancom, uspešno ukazuje na funkcionalne poremećaje u proteinima, čime ova oblast dobija na značaju. Tokom prvih godina, pojavljuju se mnogobrojni nazivi "osetljivi", "reomorfični", "mobilni", "kameleonski", "igrajući" i drugi. Usled velikog broja termina koji se, i kasnije, koriste za opisivanje ovakvih proteina: suštinski neuređeni, nesavijeni, denaturisani ili reomorfni proteini (eng. *intrinsically disordered/ unfolded/ unstructured*), u ovom radu, biće korišćen samo kraći termin - neuređeni proteini. Neuređeni proteini imaju bitnu ulogu u određivanju ćelijskog odgovora na spoljašnje uticaje, transkripciju i translaciju, kao i savijanje i odvijanje ćelijskih makromolekula. Kao što je navedeno u prethodnom poglavlju, neuređenost proteina se može, osim eksperimentalnog, određivati i računarski. Upravo o tom vidu predikcije, neuređenosti proteina govori ovo poglavlje. Najpre, biće detaljnije opisan računarski postupak kojim se vrši predikcija. Nakon toga, biće reči o prediktorima, od kojih će pojedini biti detaljnije objašnjeni. Potom ukratko će biti predstavljena baza podataka *DisProt* i njen značaj u ovom radu. Na kraju biće predstavljene statističke mere za procenu kvaliteta [6, 16, 17].

2.1 Prediktori

Više od sedamdeset prediktora razvijeno je od 1997. godine, od čega čak sedamnaest u periodu između 2010. i 2014.. Ovi prediktori se mogu ugrubo podeliti u nekoliko kategorija [18]:

1. prediktori zasnovani na metodama mašinskog učenja,
2. prediktori zasnovani na meta-pristupu (kombinovanjem predikcija više prediktora) i
3. prediktori zasnovani na fizičko-hemijskim karakteristikama.

Kada je reč o prediktorima, možemo govoriti o njihovoj nepouzdanosti (eng. *uncertainty*). Nepouzdanost, odnosno pouzdanost, prediktora odnosi se na meru poverenja koju imamo u rezultat dobijen korišćenjem prediktora. Postoje dva glavna izvora nepouzdanosti predikcije neuređenosti koji dolaze iz:

- nepouzdanosti modela i
- nepouzdanosti podataka.

Pouzdanost (ili nepouzdanost) modela zavisi od odabranog modela. Odabir modela se vrši tako što iz skupa dostupnih modela bira onaj čija je preciznost veća u odnosu na ostale dostupne modele, testiranjem na zadatom skupu sekvenci. Pouzdanost podataka se odnosi na pouzdanost u sekvence koje predstavljaju ulaze u prediktore. [19]

U nastavku biće opisani neki od najpoznatijih prediktora, od kojih su: *PONDR*, *DISEMBL* i *IUPRED* korišćeni za konsenzus u aplikaciji. Osim prediktora, na samom kraju poglavlja navode se i dve veoma korišćene baze proteina.

2.1.1 SPOT-D

SPOT-Disorder Predictor je razvijen da ima visoku efikasnost u predviđanju i kratkih i dugih neuređenih regiona bez odvojenog treninga, bez obzira na činjenicu da neuređeni regioni različitih veličina imaju različite sastave aminokiselina. SPOT-D je metod koji je nastao unapređivanjem metoda koji koristi tradicionalne neuronske mreže bazirane na prozorima nad svim testiranim skupovima bez odvajanja trening skupa na kratkim i dugim regionima. Utvrđeno je da je preciznost metoda SPOT-D uporediva u odnosu na ostale metode. Ovaj metod oslikava prednosti kombinovanja LSTM (eng. *Long Short Term Memory*) neuronskih mreža sa dubokim dvosmernim rekurentnim neuronskim mrežama, kako bi se uočile interakcije između proteina [20].

2.1.2 PONDR

PONDR prediktor vrši predikciju nad pojedinačnim sekvencama korišćenjem neuronskih mreža sa propagacijom unapred (eng. feedforward neural networks) koje koriste različite atribute nad prozorima dužine od 9 do 21 aminokiselina. Uzima se prosek vrednosti atributa nad ovim prozorima, a potom se te vrednosti koriste pri treniranju neuronskih mreža tokom konstrukcije prediktora. Prediktori neuronskih mreža se treniraju nad neredundantnim skupovima uređenih i neuređenih sekvenci, a izlazne vrednosti su brojevi između 0 i 1, koji se određuju za svaki region od po 9 aminokiselina. Ako vrednost dodeljena nekom regionu prevazilazi prag od 0.5 smatra se da je region neuređen.

2.1.3 IUPred

IUPred vrši previđanje neuređenosti proteina sa loše definisanom tercijarnom strukturom (eng. Intrinsically unstructured/disordered proteins - IUPs) na osnovu sekvenci aminokiselina procenjujući njihovu energiju prilikom interakcija. Metod se bazira na neuređenosti proteina. Naime, globularni proteini učestvuju u velikom broju interakcija, čime se obezbeđuje stabilizujuća energija koja nadoknađuje određene gubitke prilikom savijanja proteina. Nasuprot njima, neuređeni proteini imaju specijalne regione koji ne učestvuju u interakcijama.

Pristup korišćen pri razvoju ovog prediktora se zasniva na statističkoj proceni mogućnosti polipeptida da formiraju takve stabilne veze (interakcije). Pretpostavka koja postoji je da se neuređene sekvence ne savijaju zbog nemogućnosti da ostvare dovoljno stabilne veze prilikom interakcija. Pokazano je da energija prilikom interakcija može da se proceni numerički na osnovu sastava aminokiselina, uzimajući u obzir da koliko će aminokiselinski sastav doprineti uređenosti zavisi od hemijskog tipa aminokiseline i njene sposobnosti da interaguje sa drugima. Prilikom predikcije, mogu se

koristiti ugrađeni parametri koji su optimizovani za predviđanje kratkih ili dugačkih neuređenih regiona [21, 22, 23, 24].

2.1.4 ESpritz

ESpritz detektuje neuređene regione primarne strukture i bazira se na efikasnom sistemu za predviđanje koji ih pronalazi. Određivanje neuređenosti, na osnovu niza aminokiselina koje čine protein, je težak problem, ali ova metoda daje obećavajuće rezultate. Postoje dva razloga za to:

- ako niz aminokiselina određuje strukturu onda nestrukturirani regioni aminokiselina mogu imati drugačije osobine,
- neuređenost je bitna za mnogobrojne biološke funkcije, pa je prisutna očuvanost neuređenih proteina tokom evolucije.

ESpritz, pri svom radu, koristi dvosmerne rekurentne neuronske mreže (engl. BRNN - Bidirectional recursive neural network) i treniran je na više različitih tipova neuređenosti. Algoritam uči kontekst informacija kroz rekurzivnu dinamiku mreže, smanjujući time broj parametara i implicitno izvlačeći informacije iz sekvence. Ovo je efikasan metod za pojedinačne sekvence i bazira se na informacijama iz primarne sekvence proteina. Tipovi predviđanja neuređenosti nad kojima je ESpritz treniran su:

- Kratki x-zraci (eng. short x-ray): bazirano na nedostajućim atomima u strukturama koje su rešene sa X-zracima i nalaze se u PDB-a (eng. PDB - Protein Data Bank), ovaj tip predviđanja koristi se kod kraćih proteina.
- Duži disprot: skup podataka koji se koristi za ovaj tip sadrži duže neuređene segmente u odnosu na prethodni tip. Bazira se na funkcionalnim atributima neuređenih regiona. Smatra se da je pronađen neuređeni region ako se utvrdilo barem jednom da je neki region neuređen. Svi ostali regioni se smatraju uređenim.

ESpritz određuje verovatnoću neuređenosti za svaki region u sekvenci [25, 26, 27, 28].

2.1.5 DisEMBL

$DisEMBL$ predstavlja alat za računarsko određivanje neuređenosti proteina koje se oslanja na nekoliko novih koncepata. $DisEMBL$ je metod zasnovan na neuronskim mrežama. S obzirom da precizna definicija neuređenosti ne postoji, neuronske mreže su trenirane tako da zadovolje više definicija neuređenosti. Prediktor vrši procenu i prikazuje verovatnoću neuređenosti određenih regiona u sekvenci. Neuređenost proteina kod ovog prediktora posmatra se kroz model sa dva stanja - uređeno i neuređeno. Korišćeni kriterijumi su:

- Petlje / navoji - posmatranje petlji odnosno navoja, iako nije najpouzdaniji vid određivanja neuređenosti, može ukazati na njegovo postojanje u smislu da se neuređenost javlja baš u petljama (navojima).
- Vruće petlje - predstavljaju finije odabran podskup za posmatranje od prethodno navedenog. Ove petlje imaju visok stepen pokretljivosti i smatraju se same po sebi neuređenošću.

- Nedostajuće kordinate na X zracima, definisane u skupu podataka $remark465$. Ovo je jedna od prvih metoda korišćenih za određivanje neuređenosti, koja se ogleda u tome da ako postoje guste skupine elektrona koje se nalaze na mestu na kom ne bi trebalo to čvrsto ukazuje na postojanje neuređenosti.

[29]

2.1.6 Disopred2

DISOPRED2 je treniran na skupu od 750 neredundantnih sekvenci struktura dobijenih na osnovu X-zraka. Neuređenost se prepoznaje kod onih regiona koji se pojavljuju u primarnoj sekvenci proteina, ali se ne pojavljuju na elektronskoj mapi gustine. Ovaj način ima svoje mane koje se ogledaju u nesavršenosti metode kristalografije X-zracima kod koje se mogu javiti nedostaci. Iako nije idealan, ovaj način je najjednostavniji u nedostatku daljih eksperimentalnih analiza proteina. Ulazni vektor za svaki region se konstruiše iz profila sekvence simetričnim prozorom od po 15 pozicija. Podaci se koriste za treniranje metodom potpornih vektora (eng. SVM - support vector machine) [30].

2.2 Baze podataka u bioinformatici

Baze podataka imaju veliku ulogu u bioinformatici. Pre svega, njihova glavna uloga je prikupljanje podataka iz proverenih izvora i objedinjavanje dobijenih rezultata. Jedna od najvećih prednosti korišćenja ovih baza je to što su često javno dostupne tako da im mogu pristupiti svi, i koristiti dostupne informacije za različite vidove istraživanja.

Baza podataka DisProt Baza neuređenih proteina (eng. *the Database of Protein Disorder*), DisProt, sadrži podatke o neuređenim regionima proteina, za koje je ona određena eksperimentalnim putem ili na osnovu literature. Svaki region koji se smatra neuređenim poseduje dokaz koji to potvrđuje. Dokaz sadrži nekoliko informacija:

- eksperimentalnu metodu kojom je dobijen rezultat,
- odgovarajuću literaturu u kojoj se dokaz nalazi i
- pozicijom ili intervalom na kom se neuređenost javila.

eksperiment, odgovarajućim radom i pozicijom u sekvenci tog proteina. DisProt baza predstavlja jednu od najpoznatijih baza neuređenih proteina. [1]

Baza podataka UniProt *UniProt* je baza podataka otvorenog pristupa koja sadrži informacije o sekvencama proteina i njihovim funkcijama. Veliki broj informacija sadržanih u ovoj bazi su poreklom iz različitih projekata sekvenciranja genoma, kao i iz literature.[2]

[1] Veza ka zvaničnoj stranici: http://www.disprot.org/
[2] Veza ka zvaničnoj stranici: https://www.uniprot.org/

2.3 Procena kvaliteta

Da bi se procenio kvalitet donetih odluka koristi se nekoliko statističkih mera:

- Preciznost (eng. *precision) - Preciznost predstavlja udeo onih instanci koje su tačno klasifikovane kao pozitivne u odnosu na sve instance koje su klasifikovane kao pozitivne.*

$$Preciznost = \frac{SP + SN}{SP + LP + SN + LN}$$

- *Odziv (eng.* recall) - Odziv predstavlja udeo onih instanci koje su tačno klasifikovane kao pozitivne od svih instanci koje jesu pozitivne.

$$Odziv = \frac{SP}{SP + LN}$$

- Tačnost (eng. *accuracy) - Tačnost predstavlja udeo tačno klasifikovanih instanci u ukupnom broju instanci.*

$$Tanost = \frac{SP}{SP + LP}$$

- *F-mera (eng.* F-measure) - F-mera predstavlja harmonijsku sredinu između preciznosti i odziva.

$$F = \frac{2SP}{2SP + LP + LN}$$

Ove mere će biti korišćene za procenu kvaliteta metaprediktora.

Glava 3

Aplikacija

Jedan od glavnih ciljeva ovog rada je kreiranje aplikacije, koja bi vršila poređenje izlaza iz prediktora i metodom konsenzusa dala procenu, koja bi pomogla u utvrđivanju neuređenih regiona sekvence.

U nastavku je detaljno opisana izrađena aplikacija sa svim važnim aspektima koji je čine. Programski jezici korišćeni za izradu ovog projekta su **JavaScript** i **Python**, a za izradu samôg interfejsa korišćene su veb tehnologije **HTML** i **CSS**.

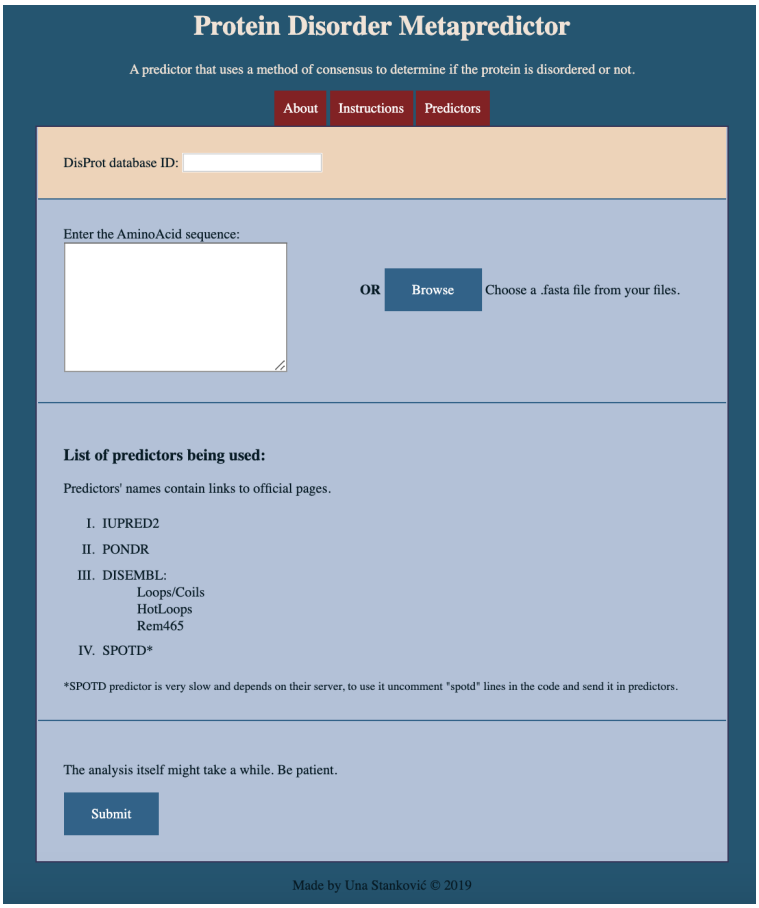
3.1 Arhitektura

Arhitekturalna organizacija aplikacije je *klijent − server* arhitektura. Klijent-server arhitektura podrazumeva odvajanje uloga: klijenta, koji zahteva neku uslugu od servera, i servera, koji tu uslugu pruža i šalje odgovor klijentu. Ovaj vid arhitekture je odabran kako bi se jasno razdvojile uloge između klijenta koji komunicira sa korisnikom i servera koji komunicira sa prediktorima.

3.1.1 Klijent

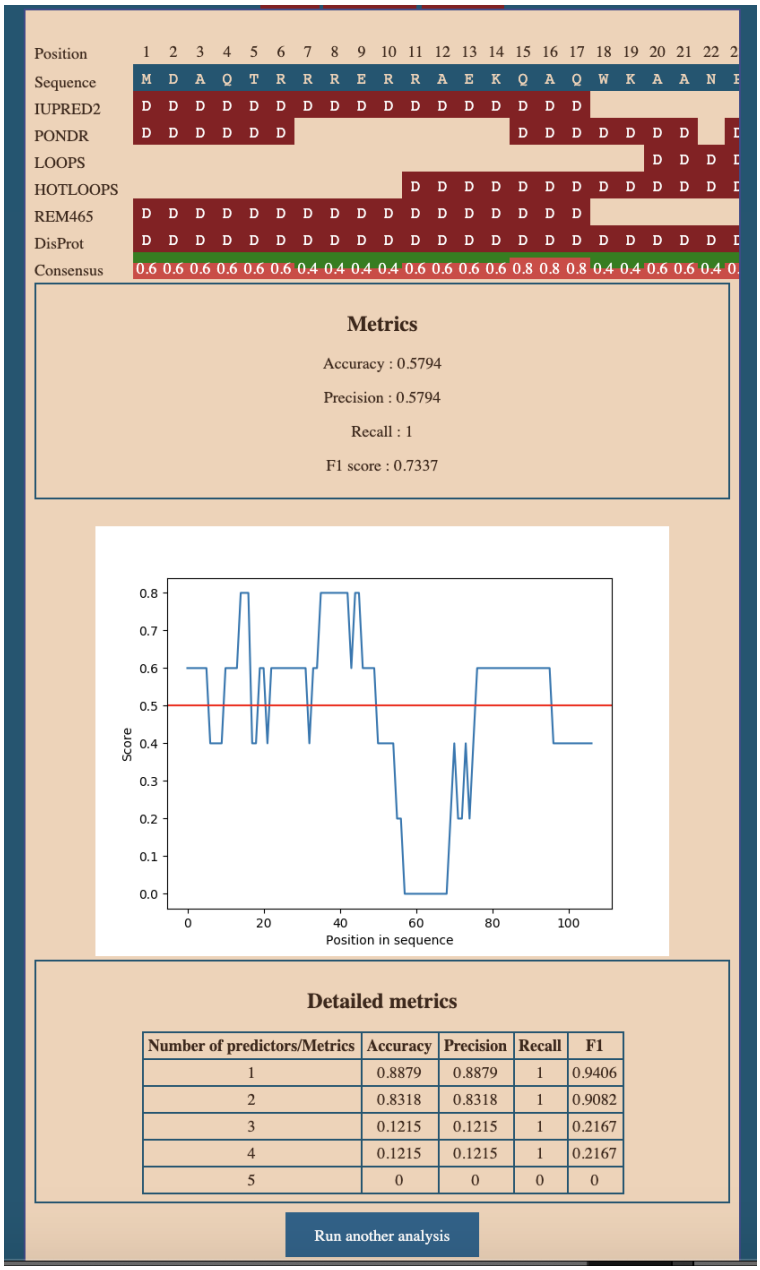
Klijent je implementiran u vidu korisničkog interfejsa, koji od korisnika aplikacije zahteva određene unose. Ti unosi se, potom, šalju serveru na obradu.

Od unosa na klijentskoj strani zahtevaju se *identifikator* u *DisProt* bazi i sekvenca, koja je opciona. Na osnovu identifikatora, sekvenca se, ukoliko nije uneta, povlači iz baze, a ukoliko identifikator nije unet, program izbacuje obaveštenje da ga je obavezno uneti. Sekvenca se unosi u *.fasta* formatu, klikom na dugme "browse" ili kao sirova niska kroz tekstualno polje. Klikom na dugme "submit" uneti podaci se šalju na server gde se obrađuju. Klijent ostatak vremena izvršavanja programa osluškuje i čeka na odgovor od servera. Prikaz izgleda korisničkog interfejsa može se videti na slici 3.1.



Slika 3.1: Prikaz korisničkog interfejsa.

Nakon što server obradi podatke, rezultati se vraćaju klijentu i prikazuju se u vidu nekoliko nizova. Ti nizovi predstavljaju izlaze prediktora, odnosno, njihove odluke o uređenosti (neuređenosti) svake od aminokiselina u sekvenci. U prvom redu predstavljena je sekvenca, aminokiselina po aminokiselina. U ostalim redovima predstavljeni su slovom "D" neuređeni regioni sekvence. Na osnovu *identifikatora* u *DisProt* bazi određuju se eksperimentalno neuređeni regioni, koji se nalaze u pretposlednjem redu, obeleženi, takođe, slovom "D" . U poslednjem redu se nalazi konsenzus prediktora (proračunat na osnovu odgovora prediktora, podrazumevano bez korišćenja informacija iz *DisProt* baze.). Potom, prikazane su vrednosti metrika ispod kojih se nalazi grafički prikaz konsenzusa nad niskom sa pragom od 0.5 i skalom datih rezultata na y-osi.
Na samom kraju, nalazi se tabela koja sadrži podatke o metrikama u zavisnosti od broja prediktora koji su rekli da je region neuređen. Posmatra se sledeće tvrđenje: "*Aminokiselina je neuređena ukoliko je k prediktora reklo da je neuređena.*"
Prikaz stranice sa rezultatima može se videti na slici 3.2.



Number of predictors/Metrics	Accuracy	Precision	Recall	F1
1	0.8879	0.8879	1	0.9406
2	0.8318	0.8318	1	0.9082
3	0.1215	0.1215	1	0.2167
4	0.1215	0.1215	1	0.2167
5	0	0	0	0

Slika 3.2: Prikaz korisničkog interfejsa pri povratku rezultata sa servera za protein sa identifikatorom $DP00005$.

Implementacija klijenta

Klijent je implementiran kroz skriptni programski jezik "JavaScirpt" i korišćenjem veb tehnologija $HTML5$ i $CSS3$. Ceo kôd klijenta sastoji se iz narednih datoteka:

- **index.html** - glavna strana koja od korisnika zahteva unos i sadrži informacije o projektu, uputstvo za korišćenje aplikacije, kao i informacije o prediktorima,
- **style.css** - datoteka sadrži sve stilove korišćene kroz aplikaciju,
- **sock_cli.js** - datoteka prikuplja informacije unete od strane klijenta i prosleđuje ih serveru i prima informacije od servera i poziva funkcije iz *results.js* kako bi se izvršio prikaz dobijenih rezultata na stranici,

- **results.js** - datoteka sadrži funkcije za kreiranje prikaza i vrši prikaz dobijenih rezultata na stranici dinamički, kreiranjem novih elemenata,
- **loader.js** - datoteka menja prikaz stranice prilikom čekanja na server.

Za komunikaciju sa serverom koristi se *socketio* biblioteka. Podaci se šalju emitovanjem određenog događaja čiji se naziv navodi kao prvi argument funkcije *emit*.

```
const socket = io('http://localhost:5000');
socket.on('connect', () => {
    console.log('Connected to server');
});

function sendData() {
    if (txtDisprotId.value == ""){
        alert("You must enter the DisProt id!");
    }
    else{
        showLoader();
        const json = generateData();
        console.log('Sending to server.');
        socket.emit('message', json);
    }
}
```

Primanje podataka vrši se reagovanjem na događaj koji se navodi kao prvi argument funkcije *on*. Kako bi se podaci pravilno prikazali više funkcija za prikaz različitih delova rezultata moraju biti pozvane.

```
socket.on('predicted', data => {
    console.log('Got from server.');
    console.log(data);
    createScale(data.sequence);
    for (const pred of data.predictors) {
        createPredictor(pred.result, pred.name);
    }
    hideLoader();
    createConsensus(data.consensus.result);
    addMetrics(data.metrics);
    addDetailedMetrics(data.partial_metrics);
});
```

3.1.2 Server

Serverska obrada se sastoji iz komunikacije sa lokalno čuvanim aplikacijama (programima za prediktore), kao i postojećim veb stranicama kako bi se obezbedili podaci neophodni za analizu.

Serveska strana komunicira sa klijentskom stranom korišćenjem soketa, odnosno *socketio* biblioteke. Ovaj vid komunikacije u kombinaciji sa nitima, je odabran zbog čekanja na izlaze iz predikotra. Ukoliko bi se radilo o nekom drugom vidu komunikacije moglo bi da se desi da se rezultati ne prikažu ili izmešaju zbog neujednačenog vremena povratka informacija. Server slušanjem čeka na informacije od klijenta i po primanju informacija vrši nekoliko radnji:

- *DisProt identifikator* šalje se zahtevom preko omogućenog *API*-ja[1]. Odgovor se potom parsira regularnim izrazima kako bi se dobili eksperimentalno neuređeni

[1] API - skraćeno od engl. *Application Programming Interface, predstavlja interfejs za komunikaciju sa, u ovom konkretnom slučaju, DisProt bazom.*

regioni - intervali regiona u sekvenci i sekvenca. Potom se ti intervali popunjavaju u listi i ona se vraća kao niz karaktera "D" i "-", gde "D" predstavlja neuređenost.

- Sekvenca se smešta u zasebnu datoteku, za potrebe lokalno čuvanih prediktora. Osim toga, sekvenca se dodatno parsira (ukoliko je u $.fasta$ formatu onda se iz nje uklanja prva linija, kako bi bila prilagodjena za ulaz prediktorima koji zahtevaju samo sirovu sekvencu).

- Sekvenca se šalje prediktorima. Prediktori su kreirani kao elementi klase *Predictor* i za svaki od njih se računa *calculate* funkcija koja vraća listu karaktera "D" i "-".

- Svaki od prediktora prima sekvencu i na poseban način je obrađuje:

  – *IUPRED2* - Prediktor se nalazi lokalno sačuvan i prosleđuje mu se naredba za pokretanje, kao i naziv datoteke u kojoj je sačuvana niska koja se obrađuje.

  – *PONDR* - Prediktor se nalazi na *web* lokaciji na kojoj se nalazi forma koja se dinamički popunjava na osnovu korisničkog unosa uz pomoć *mechanize* modula. Potom se regularnim izrazom parsira dobijeni rezultat.

  – *SPOTD* - Prediktor se nalazi na *web* lokaciji i postupak je poput prethodnog prediktora, međutim, problem koji se javlja sa ovim prediktorom je taj što je server poprilično spor (izvršavanje preko 10 minuta, server često nije dostupan). Za korišćenje ovog prediktora neophodno je skinuti komentare iz koda i dodati u *predictors* objekat rezultat *spotd* predikcije.

  – *DISEMBL* - Prediktor predstavlja spoj tri prediktora *coils/hotcoils*, *loops* i *rem*465 . Prediktor se nalazi na *web* lokaciji. Ovaj prediktor vraća rezultate na osnovu tri vida predikcije.

Implementacija servera

Serverska strana je u potpunosti implementirana u programskom jeziku *Python* i organizovana je kroz sledeće datoteke:

- **sock_serv.py** - Datoteka sadrži srž serverske strane programa. U datoteci su organizovani primanje informacija sa servera i emitovanje povratnih informacija.

- **predictor_service.py** - Datoteka sadrži kreiranje elemenata klasa i vraća izračunate vrednosti predikcije.

- **disprot_service.py** -U ovoj datoteci se nalazi povezivanje na *DisProt* bazu i parsiranje dobijenih rezultata. Iz dobijenih informacija iz baze zanimaju nas intervali koji sadrže neuređene regione i sekvenca. Iako su pored regiona obezbeđene informacije o metodu kojim su oni određeni, za ovaj rad ti podaci nisu relevantni.

- **predictor.py** - Datoteka sadrži kreiranu klasu prediktor, i metode koje su neophodne za svaki od prediktora.

- **consensus.py** - Datoteka sadrži funkcije vezane za kreiranje konsenzusa i iscrtavanje rezultata.

- **measures.py** - Datoteka sadrži funkcije vezane za formiranje metrika.

- **pondr.py**
- **spotd.py**
- **iupred2.py**
- **disembl.py**

Poslednje navedene datoteke sadrže informacije opisane u prethodnoj podeli, svaki rezultat je predstavljen u vidu liste.

Za komunikaciju sa klijentom, koristi se *socketio* biblioteka, koja funkcioniše po istom principu kao na klijentskoj strani. Konekcija sa klijentom:

```
sio = socketio.Server(cors_allowed_origins=['http://localhost:9004'])
app = socketio.WSGIApp(sio, static_files={
    '/': {'content_type': 'text/html', 'filename': 'index.html'}
})

@sio.event
def connect(sid, environ):
    print('connect ', sid)

if __name__ == '__main__':
    eventlet.wsgi.server(eventlet.listen(('', 5000)), app)
```

Primanje informacija od klijenta funkcioniše tako što server sluša čekajući na poruku i po prijemu poruke poziva funkciju za pozivanje prediktora.

```
@sio.on("message")
def message(sid, data):
    print('message ', data)

    # Calling all the predictors to give their opinion on the sequence
    in the same thread!
    x = threading.Thread(target=prediction_calls(data), args=(data))
    x.start()
```

Funkcija za pozivanje prediktora se sastoji iz nekoliko celina:

1. Najpre se iz prosleđenih podataka izvlače informacije o *identifikatoru* i *sekvenci* ako postoji.

   ```
   def prediction_calls(data):
       id_disp = data["disprot_id"]
       s = data["sequence"]
       # DisProt database call
       [disprot, seq] = disprot_service.get_sequence_info(id_disp)
       print(seq, len(seq))
       if (s == ""):
           s = seq
   ```

   Oni se prosleđuju *get_sequence_info* u datoteci *disprot_service.py* koja dohvata informacije iz baze. Pripremanje sekvence sastoji se iz parsiranja povratnih informacija sa servera, izvlačenja intervala i popunjavanja pozicija tih intervala sa "D".

   ```
   def get_sequence_info(id_disp):
      # response = requests.get('http://www.disprot.org/ws/get/' +
       id_disp) # Old API changed on 13.09.2019.
       response = requests.get('http://www.disprot.org/api/' +
       id_disp)
       data = json.loads(response.content)
       if re.search("200", str(response)) == None:
   ```

```
        data = "Not found."
    return prepare_sequence(data)

def prepare_sequence(data):
    if data == "Not found.":
        return data
    sequence = data['sequence']
    seq_len = len(shortened_sequence(sequence))
    intervals = []
    disprot = []
    disprot = ['-'] * seq_len
    for record in data['regions']:
        start_interval = record["start"]
        end_interval = record["end"]
        intervals.append((start_interval, end_interval))
    for interval in intervals:
        begin, end = int(interval[0]), int(interval[1])
        for i in range(begin-1, end):
            disprot[i] = "D"
    return disprot, sequence
```

2. Potom se sekvenca čuva lokalno u fajlu.

```
 # Storing the sequence in a file in order to locally use the
    predictors
    f = open("sequence.txt", "w")
    f.write(s)
    f.close()
```

3. Nakon toga pozivaju se prediktori da pojedinačno daju rezultat. Prediktor *SPOTD* je izuzet iz računanja, jer zavisi od njihovog servera koji je često nedostupan ili veoma spor (izračunavanje traje više od 15 minuta).

```
# Calling predictors
    iupred2 = predictor_service.iupred2_predict(s)
    #SPOTD: uncomment the function call
    #spotd = predictor_service.spotd_predict(s)
    pondr = predictor_service.pondr_predict(s)
    ss = shortened_sequence(s)
    [loops, hotloops, rem465] = predictor_service.disembl_predict(
    ss)
```

Prilikom poziva prediktora, unutar funkcija oblika *nazivprediktora_predict*, kreira se instanca klase *Predictor* za datu sekvencu, a vraća se izračunat rezultat predikcije. Navedene funkcije nalaze se u datoteci *predictor_service.py*.

```
def iupred2_predict(sequence):
    p = iupred2(sequence)
    return p.calculate()

def spotd_predict(sequence):
    p = spotd(sequence)
    return p.calculate()

def pondr_predict(sequence):
    p = pondr(sequence)
    return p.calculate()

def disembl_predict(sequence):
    p = disembl(sequence)
    return p.calculate()
```

Izgled klase *Predictor* iz datoteke *predictor.py*.

```
class Predictor:
    def __init__(self, sequence):
        self.sequence = sequence
        self.calculated = []

    def calculate(self):
        pass # Store result in self.calculated in form [
    aa1_prediction, aa2_prediction, aa3_prediction, ...]
```

Za prediktore kojima se pristupa preko njihovih veb stranica koristi se biblioteka *mechanize* za popunjavanje odgovarajućih formi i biblioteka *re* kojom se uvode regularni izrazi, neophodni za parsiranje dobijenih rezultata. Primer jednog od prediktora kod koga se predikcija vrši preko veb stranice:

```
class pondr(Predictor):
    def calculate(self):
      global br
      url = "http://www.pondr.com/cgi-bin/pondr.cgi"
      br.set_handle_robots(False)
      br.open(url)
      br.form = list(br.forms())[1]
      br['ProteinName'] = "test"
      br['Sequence'] = self.sequence
      response = br.submit()
      soup = BeautifulSoup(response.read(), features='html5lib')
      soup = soup.prettify()
      result = re.findall("VLXT\t\s.*", soup)
      # We don't need the first one because it is not sequence
      result = result[1:]
      predicted = []
      for r in result:
          pred = r[6:]
          predicted.append(pred)
      total = []
      # Getting everything in one sequence
      for i in predicted:
        total += i
      old = len(total)

      pom = shortened_sequence(self.sequence)
      if(old < len(self.sequence)):
        for i in range(0,len(pom) - old):
          total.append(" ")
      total = [x if x != ' ' else '-' for x in total ]

      self.calculated = total
      return self.calculated
```

4. Računa se konsenzus, a potom se računaju metrike uz pomoć funkcije *measure_all* koja se nalazi u datoteci *measure.py*.

```
 #SPOTD: add spotd to the call
    predictors_all = [iupred2, pondr, loops, hotloops, rem465]
    cons = consensus(ss, predictors_all)
    bin_disprot = binarize_values(disprot)
    m = measure_all(consensus_treshold(cons), bin_disprot)

    # Doing consensus and metrics for when one, two or more
    predictors said "D"
    consensus_values = {}
    for i in range(1,len(predictors_all)+1):
```

```
10          cons_tres = consensus_treshold(cons,1/len(predictors_all)*
    i)
11          consensus_values[i] = measure_all(cons_tres, bin_disprot)
```

Za računanje konsenzusa koristi se funkcija *consensus* iz datoteke *consensus.py*, koja prima dva argumenta: sekvencu i listu rezultata prediktora. Ona za svaku od pozicija u rezultatu prediktora i za svaki od prediktora, posmatra da li je aminokiselina neuređena na toj poziciji i ako jeste dodaje jedinicu na ukupan skor za tu poziciju. Da bi se dobio konsenzus, nakon što su svi prediktori dali odluku o nekoj poziciji skupljena vrednost u *val* se deli sa brojem prediktora. Na kraju se poziva funkcija za iscrtavanje grafika konsenzusa.

```
def consensus(sequence, predictors):
    num_of_preds = len(predictors)
    score = []
    for i in range(0, len(sequence)):
        val = 0
        for pred in predictors:
            if pred[i] == 'D':
                val += 1
        score.insert(i, val / num_of_preds)
    plot_consensus(score, sequence)
    return score
```

Da bi se odredio konsenzus i metrike kada treba da se posmatra slučaj da je jedan, dva ili više prediktora reklo da je region neuređen vrši se nekoliko poziva:

- funkcija *consensus_treshold* iz datoteke *consensus.py*, vraća binarizovani konsenzus u odnosu na neki prag (eng. *treshold). Prag se računa tako što se broj prediktora koje želimo da uračunamo deli sa ukupnim brojem prediktora. Podrazumevani prag je* 0.5.

```
def consensus_treshold(cons, treshold=0.5):
    new = []
    print(treshold)
    for c in cons:
        if c >= treshold:
            new.append(1)
        else:
            new.append(0)
    return new
```

- *funkcija measure_all prima dva argumenta binarizovani niz konsenzusa i binarizovani niz vrednosti iz DisProt baze. Binarizaciju je neophodno izvršiti kako bi ulazi bili pogodni za algoritme za računanje metrika, ona se vrši funkcijom binarize_values iz datoteke measures.py.*

```
def measure_all(x,y):
    a = round(m.accuracy_score(x,y),4)
    p = round(m.precision_score(x,y),4)
    f = round(m.f1_score(x,y),4)
    r = round(m.recall_score(x,y),4)
    return [a,p,r,f]

def binarize_values(x):
    new = []
    for el in x:
        if el == "D":
            new.append(1)
        else:
            new.append(0)
    return new
```

5. Na kraju se dobijeni rezultati pakuju u objekat koji se potom emituje.

```
result = {
                "predictors": [
                {
                    'name': 'IUPRED2',
                    'result': iupred2
                },
                {
                    'name': 'PONDR',
                    'result': pondr
                },
                {
                    'name': 'LOOPS',
                    'result': loops
                },
                {
                    'name': 'HOTLOOPS',
                    'result': hotloops
                },
                {
                    'name': 'REM465',
                    'result': rem465
                },
                {
                    'name': 'DisProt',
                    'result': disprot
                }
                ],
                "consensus": {
                    'name': 'cons',
                    'result': cons
                },
                "sequence": ss,
                "metrics" : [
                    {
                        'name': 'Accuracy',
                        'result' : m[0]
                    },
                    {
                        'name': 'Precision',
                        'result': m[1]
                    },
                    {
                        'name': 'Recall',
                        'result': m[2]
                    },
                    {
                        'name': 'F1 score',
                        'result': m[3]
                    }
                ],
                "partial_metrics": consensus_values
            }

    sio.emit('predicted', result)
```

3.2 Korišćenje aplikacije

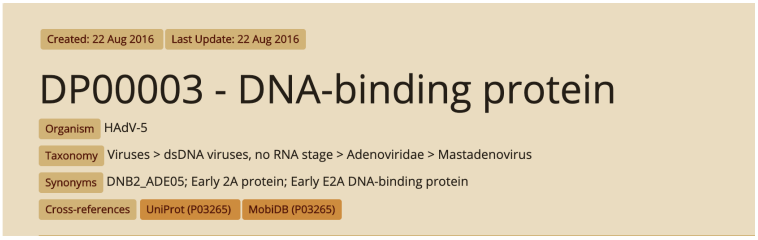
Korišćenje aplikacije je veoma jednostavno i linearno. Od korisnika se očekuje da obezbedi *DisProt* identifikator (obavezno) i sekvence (*.fasta* datoteke), ukoliko to

želi . Preporučeni način je odabirom proteina iz *DisProt* baze i unošenjem njegovog identifikatora. Drugi mogući način je davanjem samo *DisProt* identifikatora, a potom preuzimanjem *.fasta* datoteka iz *UniProt* baze. Za svaki unos u *DisProt* bazi, postoji veza ka unosu u *UniProt* bazi.

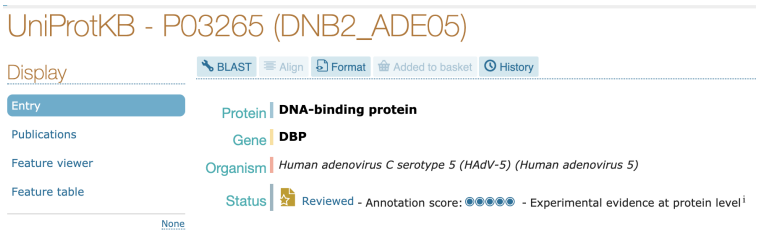
1. Unos *DisProt identifikatora*.
2. Odabir datoteke/ unos sekvence (opciono) klikom na *Browse* dugme.
3. Klik na *Submit* dugme.
4. Pregled rezultata.

3.2.1 Primer upotrebe

Za primer upotrebe biće uzeta niska sa *identifikatorom DP*00003 u *DisProt* bazi, odnosno, *DNK* vezujući protein [2]. Niska je odabrana iz *DisProt* baze, kao što se može videti na slici 3.3, a odgovarajuća *.fasta* datoteka je preuzeta iz *Uniprot* baze podataka. Prikaz dela veb stranice navedene baze i identifikatora niske nalazi se na slici 3.4. Za primer upotrebe odabran je unos i *DIsProt* identifikatora i niske kako bi se prikazale obe funkcionalnosti.



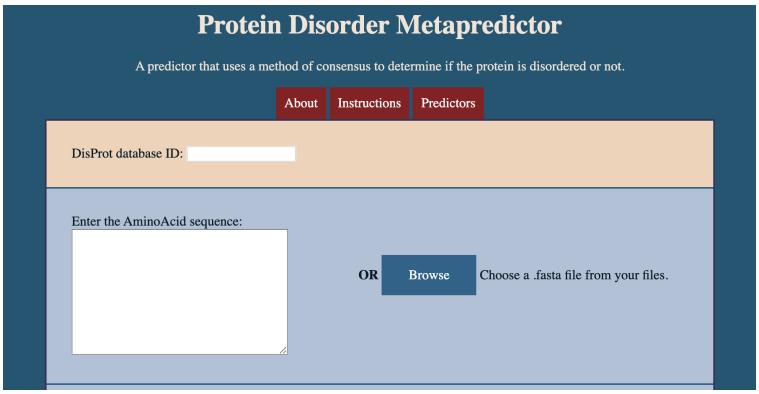
Slika 3.3: Prikaz *DisProt* baze.



Slika 3.4: Prikaz *UniProt* baze.

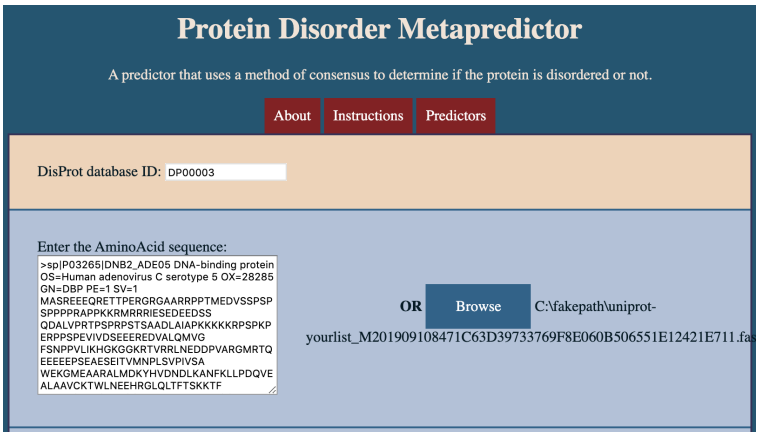
Najpre unose se *identifikator* i *.fasta* datoteka u odgovarajuća polja za unos, kao što se može videti na slici 3.5.

[2] eng. *DNA-binding protein*



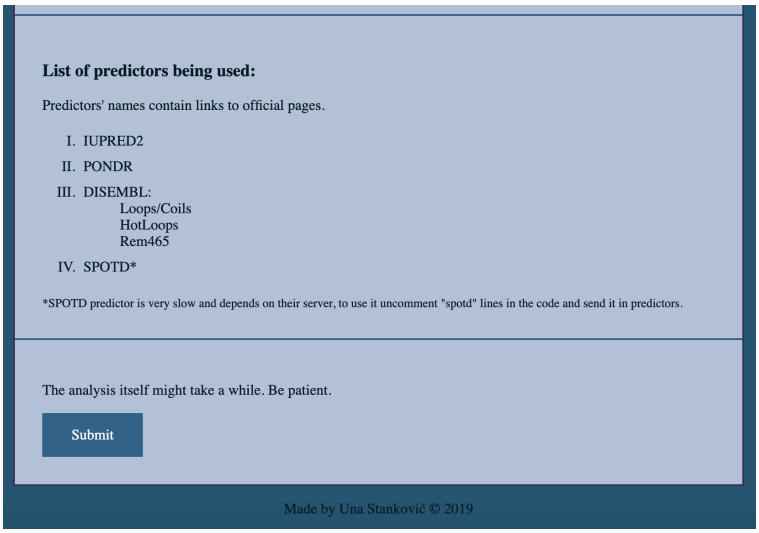
Slika 3.5: Prikaz polja za unos.

Bira se datoteka sa računara (ili unosi nisku), nakog čega se, ukoliko je niska uneta preko datoteke njen sadržaj prikazuje u tekstualnom polju čime se dobija izgled kao na slici 3.6.



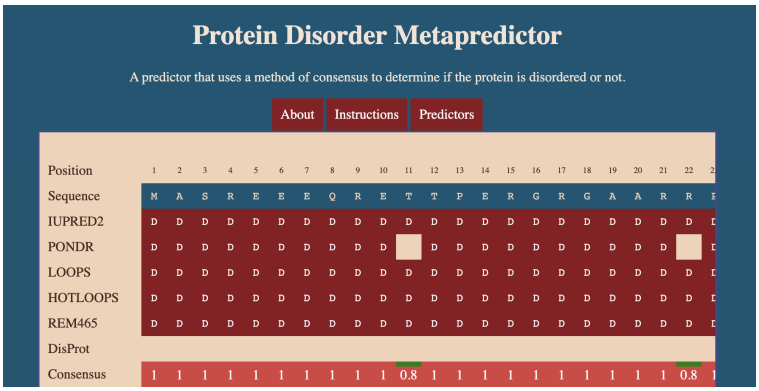
Slika 3.6: Prikaz polja za unos nakon popunjavanja.

Nakon toga, kretanjem ka dnu strane (prikazano na slici 3.7) nailazi se na spisak dostupnih prediktora, kao i veze ka njihovim zvaničnim stranicama. Klikom na dugme "Submit" podaci se šalju na server.



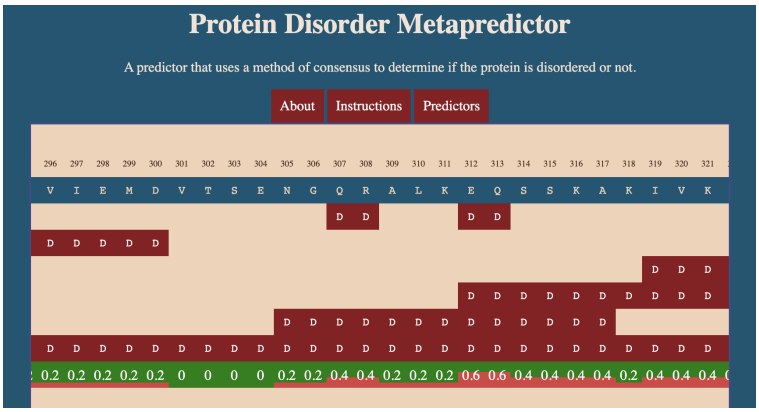
Slika 3.7: Prikaz liste dostupnih prediktora i slanja informacija ka serveru.

Potrebno je neko vreme da se sva izračunavanja izvrše, a potom se pojavljuje strana sa rezultatima. Sa leve strane navedeni su prediktori koji su učestvovali u davanju odluka, neuređeni regioni iz *DisProt* baze, kao i konsenzus u vidu brojeva. Procentualno, shodno visini konsenzusa, polje je obojeno u crveno.Prikaz stranice sa opisanim rezultatima nalazi se na slici 3.8.



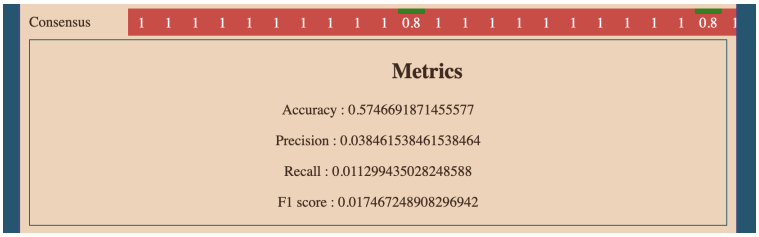
Slika 3.8: Prikaz dobijenih rezultata.

Uzimajući u obzir da je dužina sekvenci koja se posmatra obimna i da je nemoguće izvršiti prikaz na samo jednoj dužini stranice, omogućen je pregled cele sekvence, pomeranjem na levo sa dužinom kao što se može videti na slici 3.9.



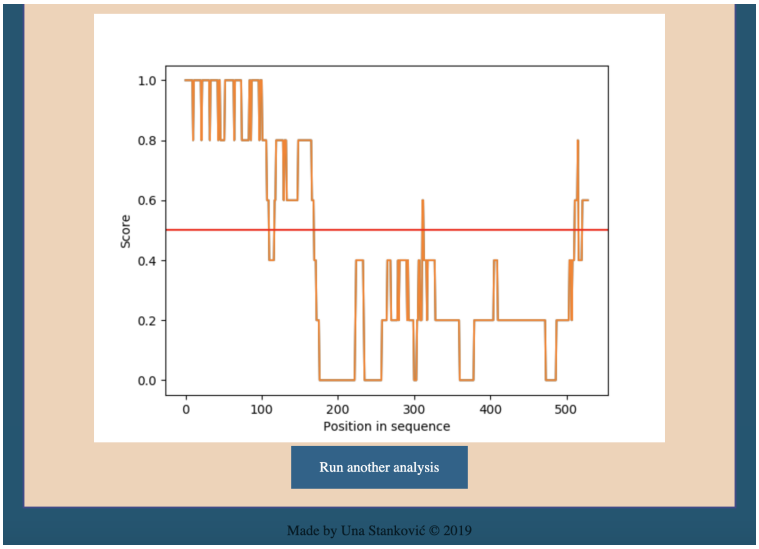
Slika 3.9: Prikaz dobijenih rezultata na kom se vidi i izlaz iz *DisProt* baze.

Ispod navedenih rezultata nalazi se prikaz vrednosti metrika za dati konsenzus i rezultate iz *DisProt* baze. U obzir su uzete četiri statističke mere: tačnost, preciznost, odziv i F1 mera. Izgled prikaza ovih metrika vidi se na slici 3.11



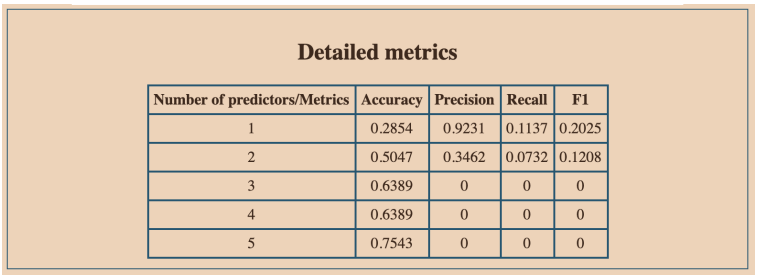
Slika 3.10: Prikaz vrednosti metrika.

Nakon metrika, sledeća stvar od interesa je grafik neuređenosti na osnovu konsenzusa. Na grafiku se nalazi visina konsenzusa u odnosu na poziciju u sekvenci. Analizom dobijenih rezultata možemo uvideti oko kojih regiona se prediktori slažu za formirani prag od 0.5 koji je obeležen crvenom linijom. Prikaz grafika može se videti na slici 3.11.



Slika 3.11: Prikaz dobijenih rezultata na grafiku.

S obzirom da su analizirane i detaljnije metrike, njihov prikaz se vrši na samom dnu strane u vidu tabele prikazane na 3.12. U tabeli su navedene mere i broj prediktora koji je korišćen.

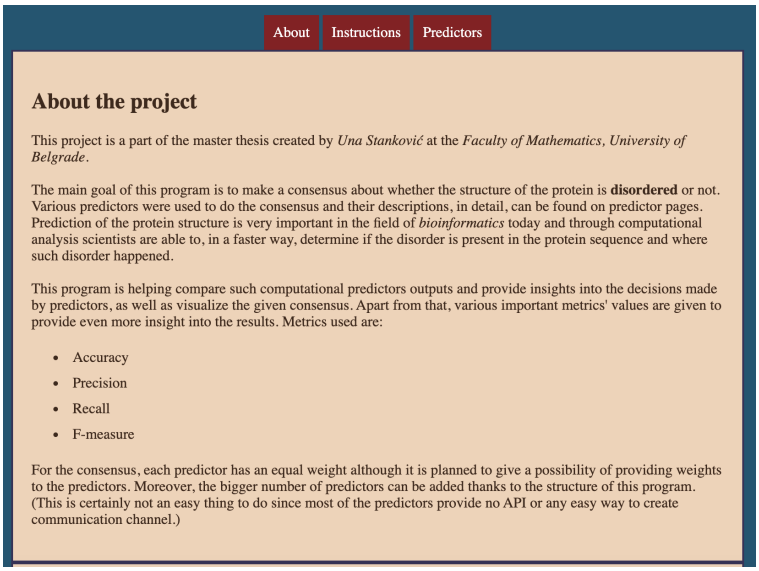


Detailed metrics

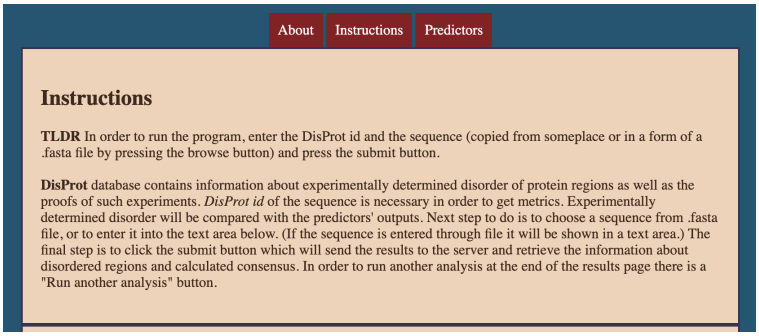
Number of predictors/Metrics	Accuracy	Precision	Recall	F1
1	0.2854	0.9231	0.1137	0.2025
2	0.5047	0.3462	0.0732	0.1208
3	0.6389	0	0	0
4	0.6389	0	0	0
5	0.7543	0	0	0

Slika 3.12: Prikaz dobijenih rezultata u tabeli.

Osim svega navedenog, u vrhu strane nalaze se dugmići koji se odnose na dodatne informacije, kao što su o projektu i uputstvo. Klikom na dugme *About* dobijaju se detaljnije informacije o autoru, motivacija, cilj i korišćene mere, kao što se može videti na slici 3.13. Dok se detaljno uputstvo za korišćenje aplikacije nalazi na stranici koja je prikazana na slici 3.14.



Slika 3.13: Prikaz stranice koja se odnosi na dodatne informacije.



Slika 3.14: Prikaz stranice sa uputstvom za korišćenje aplikacije.

3.3 Procena kvaliteta

Kao što je ranije navedeno, funkcionalnosti ove aplikacije ogledaju se u uporednoj analizi izlaza više prediktora. Osim što se izlazi prediktora mogu uporediti vizuelnim putem, posmatranjem tabele ili grafika, dati su i rezultati nekoliko statističkih mera koje pružaju formalniji uvid u rezultate metaprediktora.

3.3.1 Merenje pouzdanosti metaprediktora

Kako bi se stekla jasnija slika stanja proteina koje su vratili prediktori neophodno je da osim vizuelnog prikaza postoji i formalno merilo kvaliteta metaprediktora. Da bi se to postiglo korišćene su naredne mere kvaliteta:

- Preciznost (eng. *precision),*
- *Odziv (eng.* recall),
- Tačnost (eng. *accuracy) i*
- *F-mera (eng.* F-measure).

Ove mere su veoma pogodne kako bi se utvrdilo koliko dobro metaprediktor predviđa u odnosu na eksperimentalno utvrđenu neuređenost. Poređenje se radi u odnosu na povratne informacije iz $DisProt$ baze.

U aplikaciji je prikazano dva pristupa korišćenja metrika:

1. Prvi vid predstavlja prikaz metrika za vrednosti dobijene odnosom onoga što su svi prediktori rekli da je neuređenost (odnosno, konsenzusa koji su dali sa pragom 0.5) i onoga što je dobijeno iz $DisProt$ baze. Ovim vidom merimo rezultate koje je dao metaprediktor. S obzirom na to da se uočava da rezultati nisu uvek sjajni i da se veoma često pojavljuje situacija da se čak i svi prediktori slože da je neki region neuređen, a da to nije potvrđeno eksperimentalnim putem dolazi se do sledećeg. Prvo, moguće je da regioni za koje su prediktori utvrdili da su neuređeni nisu eksperimentom ni posmatrani. Drugo, moguće je da prediktori sami po sebi nisu dovoljno pouzdani.

2. Drugi vid predstavlja prikaz metrika za vrednosti dobijene odnosom konsenzusa za k prediktora (može ih biti od 1 do 5) i onoga što je dobijeno eksperimentalnim putem. Ovim vidom merimo rezultate dobijene posmatranjem manjim ili većim brojem prediktora i konsenzusa njihovih odluka. Vrednosti metrika koje se dobijaju ovim putem vrlo često se, posebno za *odziv* i F-meru, svode na 0 što je i razumljivo s obzirom na to da treba da se uklopi nekoliko faktora. Najpre više prediktora treba da kaže za neki region da je neuređen. Potom da bi taj region bio uzet u obzir treba da se poklopi i da je $DisProt$ za taj region rekao da neuređen.

U tabeli 3.1 nalazi se prikaz metrika za sekvencu sa identifikatorom $DP00003$, dok se u tabeli 3.2 nalazi se prikaz metrika za sekvencu sa identifikatorom $DP00005$. Iz prikaza ovih vrednosti vidi se da se izlaz rezultata metaprediktora poklapa poprilično sa onime što je $DisProt$ vratio za sekvencu $DP00005$ u odnosu na sekvencu $DP00003$. Međutim, ako bi se malo bolje obratila pažnja na ono što je vraćeno iz baze i na neuređenost regiona, primećuje se da je sekvenca $DP00005$ cela neuređena. Time dolazi do poprilično zanimljivih rezultata kada se posmatraju metaprediktori, odnosno, dolazi

se do toga da se vidi rezultat onoga što su metaprediktori dali, praktično nezavisno od *DisProt* baze.

Broj prediktora/Metrika	Tačnost	Preciznost	Odziv	Fmera
1	0.2854	0.9231	0.1137	0.2025
2	0.5047	0.3462	0.0732	0.1208
3	0.6389	0	0	0
4	0.6389	0	0	0
5	0.7543	0	0	0

Tabela 3.1: Prikaz detaljnih metrika za sekvencu DP00003.

Broj prediktora/Metrika	Tačnost	Preciznost	Odziv	Fmera
1	0.8879	0.8879	1	0.9406
2	0.8318	0.8318	1	0.9082
3	0.1215	0.1215	1	0.2167
4	0.1215	01215	1	0.2167
5	0	0	0	0

Tabela 3.2: Prikaz detaljnih metrika za sekvencu DP00005.

Glava 4

Zaključak

Aplikacija kreirana ovim radom omogućava korisniku da na veoma jednostavan način dobije pregled neuređenih regiona proteina kroz jasan i nedvosmislen interfejs. Najveći doprinos ovog rada leži u tome što objedinjuje odluke više prediktora i stavlja ih praktično pod okrilje jedne aplikacije, čime se ističu jednostavnost i pouzdanost pri donošenju odluka. Osim toga, time što nije dat neki konkretan prag za konsenzus ostavlja se na korisniku da sam odredi koji prag mu je dopustiv. Moguća unapređenja rada leže u parametrizaciji prediktora, povećanju njihove brojnosti, kao i u dodeljivanju težina prediktorima.

Bibliografija

[1] Vesna Spasojević-Kalimanovska Slavica Spasić Zorana Jelić-Ivanović. *Opšta biohemija*. 2002.

[2] *Principi strukture i aktivnosti proteina*. Hemijski fakultet, 2015.

[3] Marija Jeličić. *Povezanost dužine epitopa i uređenostidelova proteina*. on-line na: `http://elibrary.matf.bg.ac.rs/bitstream/handle/123456789/2428/Marijapdf?sequence=1`. 2012.

[4] Gerhard Michal Dietmar Schomburg. *Biochemical Pathways: An Atlas of Biochemistry and Molecular Biology*. Hoboken, New Jersey: John Wiley & Sons, Inc., 2012. ISBN: 9780470146842.

[5] Michael Lieberman. *Biochemistry, molecular biology, and genetics. — 6th edition*. 351 West Camden Street, Baltimore, MD 21201, Two Commerce Square, 2001 Market Street, Philadelphia, PA 19103: Lippincott Williams & Wilkins, a Wolters Kluwer business, 2014. ISBN: 978-1-4511-7536-3.

[6] Jovana Kovačević. *Strukturna predikcija funkcije proteina i odnos funkcionalnih kategorija proteina i njihove neuređenosti*. on-line na: `http://www.math.rs/files/DoktoratJK2015.pdf`. 2015.

[7] Ivana Čepelak Dubravka Čvorišćec. *Štrausova medicinska biokemija*. Medicinska naklada, 2009.

[8] Bradford A. Jameson Denise R. Ferrier. *Lippincott's Illustrated Reviews Flash Cards*. 2001 Market Street, Philadelphia, PA 19103: Wolters Kluwer Health, 2015. ISBN: 978-1-4511-9111-0.

[9] Denise R. Ferrier Richard A. Harvey. *Lippincott Illustrated Reviews: Biochemistry, 5th edition*. 351 West Camden Street, Baltimore, MD 21201, Two Commerce Square, 2001 Market Street, Philadelphia, PA 19103: Lippincott Williams & Wilkins, a Wolters Kluwer business, 2011. ISBN: 978-1-60831-412-6.

[10] Goran Vinterhalter. *Bioinformatička analiza povezanosti funkcije i neuređenosti proteina*. on-line na: `http://www.racunarstvo.matf.bg.ac.rs/MasterRadovi/2017_08_23_Goran_Vinterhalter/rad.pdf`. 2018.

[11] A.Keith Dunker et al. *Intrinsically disordered protein*. on-line na: `https://doi.org/10.1016/s1093-3263(00)00138-8`. 2001.

[12] A. Keith Dunker Christopher J. Oldfield. *Intrinsically Disordered Proteins and Intrinsically Disordered Protein Regions*. on-line na: `https://doi.org/10.1146/annurev-biochem-072711-164947`. 2014.

[13] Vladimir N. Uversky. *Dancing Protein Clouds: The Strange Biology and Chaotic Physics of Intrinsically Disordered Proteins*. on-line na: `https://doi.org/10.1074/jbc.r115.685859`. 2016.

[14] C.J.; Dunker A.K Uversky V.N.; Oldfield. *Intrinsically disordered proteins in human diseases: Introducing the D2 concept*. 2008.

[15] K.; Homma K.; Gojobori T.; Nishikawa K. Fukuchi S.; Hosoda. *Binary classification of protein molecules into intrinsically disordered and ordered segments.* 2011. DOI: doi:10.1186/1472-6807-11-29.

[16] V.N. Uversky. *Intrinsically disordered proteins from A to Z.* 2011.

[17] Han K.H. Tompa P. *Intrinsically disordered proteins.* 2012.

[18] Xiaoyun Wang Jing Li Wen Liu Li Rong i Jinku Bao Jianzong Li Yu Feng. *An Overview of Predictors for Intrinsically Disordered Proteins over 2010–2014.* on-line na: http://pubs.rsc.org/en/Content/ArticleLanding/2012/MB/C1MB05373F#!divAbstract. 2015. DOI: doi:10.3390/ijms161023446.

[19] Zoran Obradovic Mohamed F. Ghalwash A. Keith Dunker. *Uncertainty analysis in protein disorder prediction.* on-line na: http://pubs.rsc.org/en/Content/ArticleLanding/2012/MB/C1MB05373F#!divAbstract. 2012.

[20] Paliwal K1 Zhou Y2. Hanson J1 Yang Y2. *Improving protein disorder prediction by deep bidirectional long short-term memory recurrent neural networks.* online na: https://www.ncbi.nlm.nih.gov/pubmed/28011771. 2017. DOI: 10.1093/bioinformatics/btw678.

[21] Peter Tompa i István Simon Zsuzsanna Dosztányi Veronika Csizmok. *IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content.* on-line na: https://academic.oup.com/bioinformatics/article/21/16/3433/215919. 2005.

[22] et al. Garbuzynskiy S.O. *To be folded or to be unfolded?* 2004.

[23] K.A. Thomas P.D. i Dill. *An iterative method for extracting energy-like quantities from protein structures.* 1996.

[24] et al. Dosztányi Z. *The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins.* 2005.

[25] Di Domenico T Tosatto SC. Walsh I Martin AJ. *ESpritz: accurate and fast prediction of protein disorder.* on-line na: https://www.ncbi.nlm.nih.gov/pubmed/22190692. 2012. DOI: 10.1093/bioinformatics/btr682.

[26] S.; Frasconi P.; Soda G.; Pollastri G. Baldi P.; Brunak. *Exploiting the past and the future in protein secondary structure prediction.* 1999.

[27] C. Mooney A. Vullo G.Pollastri A. J. M. Martin. *Accurate prediction of protein secondary structure and solvent accessibility by consensus combiners of sequence and structure information.* 2007.

[28] J.L. Sussman O. Noivirt-Brik J. Prilusky. *Assessment of disorder predictions in CASP8.* 2009.

[29] Francesca Diella Peer Bork Toby J Gibson Robert B Rusell Rune Linding Lars Juhl Jensen. *Protein Disorder Prediction - Implications for Structural Proteomics.* 2003.

[30] McGuffin LJ Buxton BF Ward JJ Sodhi JS and Jones DT. *Prediction and functional analysis of native disorder in proteins from the three kingdoms of life.* 2004.