

UNIVERZITET U BEOGRADU
MATEMATIČKI FAKULTET

MASTER RAD

Razvoj metaprediktora za utvrđivanje
neuređenosti proteina

Autor:
Una STANKOVIĆ

Mentor:
dr Jovana KOVAČEVIĆ

ČLANOVI KOMISIJE:

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



Beograd, 2019.

Sažetak

Neuređenost proteina ogleda se u nestabilnoj trodimenzionalnoj strukturi proteina koja direktno utiče na njegovu funkciju. Neuređenost proteina se utvrđuje eksperimentalno, laboratorijskim analizama. 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. Kako bi se dobili što veća tačnost i preciznost predviđanja, razvijaju se metaprediktori koji predstavljaju kombinacije postojećih prediktora. Specifičan cilj ovog rada je razvoj jednog novog 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 \LaTeX detalje, i Tamari Vasić na pomoći oko prikupljanja literature i razumevanja bioloških osnova. Posebnu zahvalnost dugujem i svom mentoru, dr Jovani Kovačević na uvek zanimljivim sugestijama, podršci i divnoj saradnji. Na kraju, zahvalila bih se i članovima komisije prof dr Gordani Pavlović Lažetić i prof dr Neveni Veljković na izdvojenom vremenu i saradnji.

Sadržaj

Sažetak	ii
Zahvalnica	iii
1 Biološke osnove	3
1.1 Proteini	3
1.1.1 Funkcije i osobine proteina	5
1.1.2 Struktura proteina	7
1.1.3 Savijanje proteina	12
1.1.4 Denaturacija proteina	12
1.2 Neuređenost proteina	12
1.2.1 Eksperimentalno ispitivanje neuređenosti proteina	14
1.2.2 Računarsko predviđanje neuređenosti proteina	14
2 Predikcija neuređenosti proteina	15
2.1 Prediktori	15
2.1.1 SPOT-D	16
2.1.2 PONDR	16
2.1.3 IUPred	16
2.1.4 ESpritz	17
2.1.5 DisEMBL	18
2.1.6 Disopred2	18
2.2 Metaprediktori	18
2.2.1 FloatCons	19
2.2.2 GSmetaDisorderMD	19
2.2.3 PONDR-FIT	20
2.2.4 Positive Voter x	20
2.3 Baze podataka u bioinformatici	20
2.4 Procena kvaliteta	21
3 Aplikacija: razvoj, korišćenje i evaluacija	22
3.1 Arhitektura	22
3.1.1 Klijent	22
Implementacija klijenta	24
3.1.2 Server	26
Implementacija servera	26
3.2 Korišćenje aplikacije	32
3.2.1 Primer upotrebe	33
Prvi primer - zadavanje DisProt identifikatora	33
Drugi primer - zadavanje sekvence iz datoteke i izbor prediktora	38
3.3 Evaluacija metaprediktora	39
4 Zaključak	42
Bibliografija	43

Slike

1.1	Opšta strukturna formula aminokiselina [2]	4
1.2	Prikaz <i>L</i> i <i>D</i> prostorne konfiguracije [2]	4
1.3	Prikaz spajanja α -karboksilne grupe jedne aminokiseline i α -amino grupe druge aminokiseline, pri čemu se oslobađa molekul vode [5]	4
1.4	Prikaz centralne dogme molekularne biologije [6]	5
1.5	Prikaz struktura proteina [8]	8
1.6	Šematski prikaz struktura proteina [5]	8
1.7	Prikaz primarne strukture [5]	9
1.8	Prikaz α -heliksa [5]	10
1.9	Prikaz β -niza [5]	10
1.10	Prikaz sekundarnih struktura [11]	11
1.11	Prikaz hemoglobina, predstavnika globularnih proteina sa kvatenarnom strukturom [2]	11
1.12	Prikaz savijanja proteina [10]	13
3.1	Prikaz korisničkog interfejsa	23
3.2	Prikaz korisničkog interfejsa pri povratku rezultata sa servera za protein sa identifikatorom <i>DP00003</i>	24
3.3	Prikaz <i>DisProt</i> baze	33
3.4	Prikaz <i>UniProt</i> baze	33
3.5	Prikaz polja za unos	34
3.6	Prikaz polja za unos nakon popunjavanja	34
3.7	Prikaz liste dostupnih prediktora i slanja informacija ka serveru	34
3.8	Prikaz dobijenih rezultata	35
3.9	Prikaz dobijenih rezultata na kom se vidi i izlaz iz <i>DisProt</i> baze	35
3.10	Prikaz vrednosti mera kvaliteta metaprediktora	36
3.11	Grafički prikaz dobijenih rezultata	36
3.12	Tabelarni prikaz dobijenih rezultata	37
3.13	Prikaz stranice koja se odnosi na dodatne informacije	37
3.14	Prikaz stranice sa uputstvom za korišćenje aplikacije	37
3.15	Prikaz stranice sa unosom datoteke preko <i>Browse</i> dugmeta i odabirom prediktora	38
3.16	Prikaz stranice sa rezultatima koje je vratio server	38
3.17	Prikaz grafika sa rezultatima koje je vratio server	39

Tabele

1.1	Spisak esencijalnih i neesencijalnih aminokiselina	5
1.2	Primeri proteina [2]	6
3.1	Prikaz mera kvaliteta metaprediktora za sekvencu DP00003	40
3.2	Prikaz mera kvaliteta metaprediktora za sekvencu DP00004	40
3.3	Prikaz mera kvaliteta metaprediktora za sekvencu DP00005	40
3.4	Prikaz mera kvaliteta metaprediktora za sve sekvence u <i>DisProt</i> bazi	41

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 funkciju. Primarna struktura podrazumeva niz aminokiselina koje učestvuju u izgradnji proteina, dok je sekundarna struktura opšti trodimenzionalni oblik lokalnih segmenata. Tercijarna struktura se odnosi na trodimenzionalnu strukturu proteina. Klasa proteina koja ima delimično ili potpuno nedefinisanu trodimenzionalnu strukturu naziva se neuređenim proteinima. 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 proteine i njihovu strukturu. 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 i funkcija 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 opisanog u ovom radu, kao i o bazama podataka DisProt i UniProt. Uloga DisProt baze, u ovom radu, leži u proceni kvaliteta predikcije, dok je značaj UniProt baze u prikupljanju niski potrebnih za rad prediktora. Osim toga, opisano je i nekoliko metaprediktora sa metodima koje primenjuju.

Naredno poglavlje govori o samom metaprediktoru koji se u ovom radu razvija. Aplikacija koja je razvijena se posmatra iz više uglova i to iz ugla arhitekture i organizacije, iz ugla 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

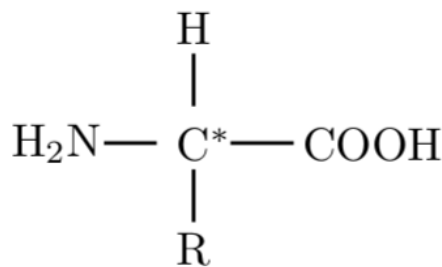
Proteini (grč. *protos* - "zauzimam prvo mesto") su biološki makromolekuli, koji čine 70% suve materije ćelija i neophodni su za njihovu izgradnju i pravilno funkcionisanje. Osim uloge u izgradnji ćelija, učestvuju u mnogobrojnim procesima koji se odvijaju unutar organizma. Predstavljaju najvažniji sastojak žive materije i utiču na brojnost i raznolikost živih bića. Broj proteina koji nastaju u živim bićima je ogroman, na primer *E.coli* sa 3000 i čoveka sa 5 miliona proteina [1, 2].

Proteini su jedinjenja sačinjena od aminokiselina. Na osnovu broja aminokiselina koji ih čine peptidi, koji predstavljaju kraće nizove aminokiselina u odnosu na proteine, dele se 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 aminokiseline¹ 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. Proteinske aminokiseline (osim glicina) imaju asimetričan α -ugljenikov 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.

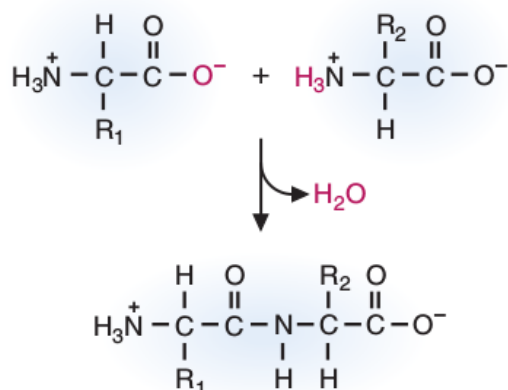
¹Neki proteini u svom sastavu mogu da imaju 22 različite aminokiseline. Pored 20 standardnih aminokiselina (koji grade prirodne proteine), postoje i 2 nestandardne i to su Selenocistein (eng. *Selenocysteine*, simboli *Sec*, *U*) i Pirolozin (eng. *Pyrrolysine*, simboli *Pyl*, *O*). Ove dve aminokiseline 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, što je grafički prikazano na slici 1.3. Ovim postupkom nastaje nerazgranati polipeptidni lanac koji se sastoji od polipeptidne kičme i bočnih ostataka. Standardna grupa aminokiseline 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]

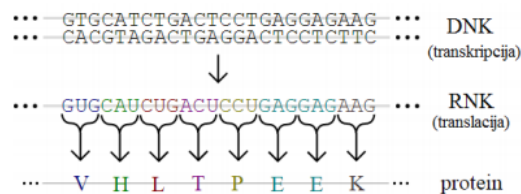
Svaki molekul proteina nastaje u ćeliji živog organizma. Redosled aminokiselina u proteinskoj sekvenci određen je redosledom aminokiselina u dezoksiribonukleinskoj

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

kiselini (DNK), odnosno gena, koji predstavljaju trojke nukleotida. Svakoj takvoj trojci jedinstveno je pridružena po jedna aminokiselina na osnovu genetskog koda izuzev tri koje označavaju *STOP* kodon. Proces kojim se kodirana 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 mogu se javiti 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.

U tabeli 1.2 mogu se videti primeri nekih proteina i njihovih uloga.

Naziv proteina	Aktivnost/Funkcija/Nalaženje
Enzimi	Kataliza svih reakcija u živim sistemima
Transportni proteini Hemoglobin Mioglobin Serum albumin	transport kiseonika i ugljendioksida transport kiseonika u mišićima transport masnih kiselina, lekova,...
Kontraktilni proteini Miozin Aktin	pokretljivost mišića
Zaštitni proteini Imunoglobulini Fibrinogen Trombin	stvaraju kompleks sa stranim telom prekursor fibrina pri zgrušavanju krvi komponenta u zgrušavanju krvi
Hormoni Insulin Hormoni rasta	reguliše metabolizam glukoze stimulišu rast
Rezervni proteini Ovalbumin Kazein Gliadin	rezerva aminokiselina za mladu jedinku jaje mleko pšenica
Strukturni proteini α -keratin Fibroin Kolagen	kosa, koža, krzno, nokti svila, paukova mreža vezivno tkivo

TABELA 1.2: Primeri proteina [2]

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

³Katalizacija predstavlja proces povećavanja brzina reakcija

- proteini grade kompleksna jedinjenja sa različitim supstancama po principu strukturne komplementarnosti i
- proteini poseduju visoku osetljivost na različite agense koji ih denaturišu⁴. 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 [1, 2, 4].

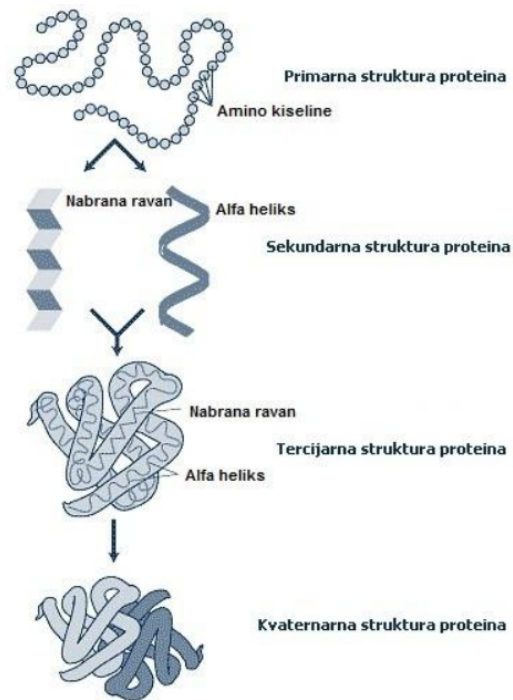
Proteinsku strukturu održavaju različite vrste kovalentnih i nekovalentnih interakcija između hemijskih jedinjenja, na primer: vodonične, jonske, elektrostatičke, dipolne, itd.. Nabiranjem i uvijanjem lanaca kreiraju se različiti oblici proteina: vlaknasti, globularni ili eliptični. Strukturni proteini 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 [1, 2, 4, 7].

Obično se struktura proteina posmatra u nivoima, pa tako postoji hijerarhijska strukturalna organizacija u četiri nivoa:

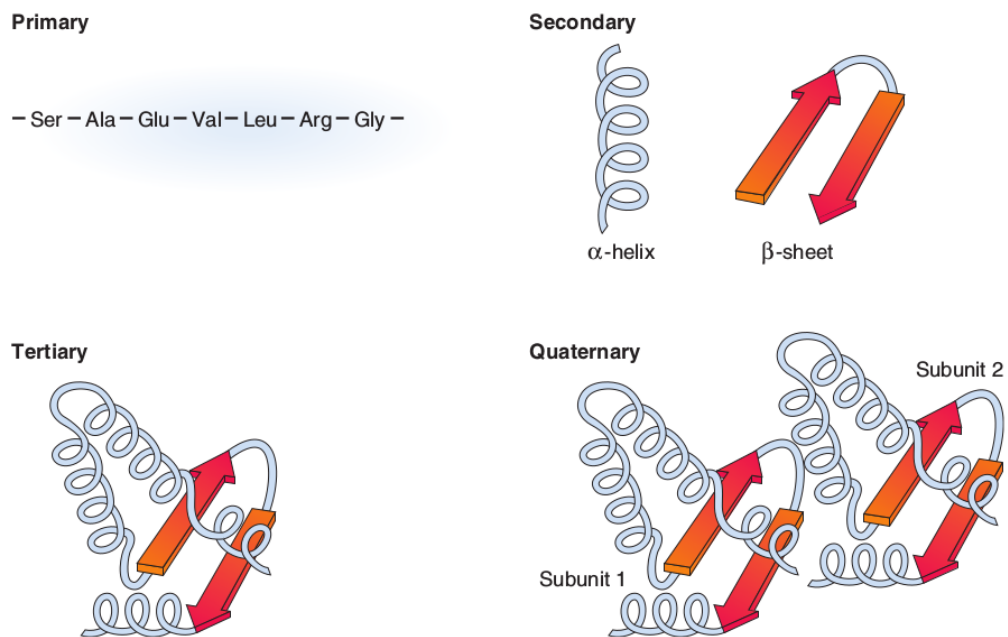
1. primarna,
2. sekundarna,
3. tercijarna i
4. kvaternarna.

Na slici 1.5 se može videti opšti prikaz mogućih struktura proteina, a na drugoj slici 1.6 šematski prikaz [1].

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



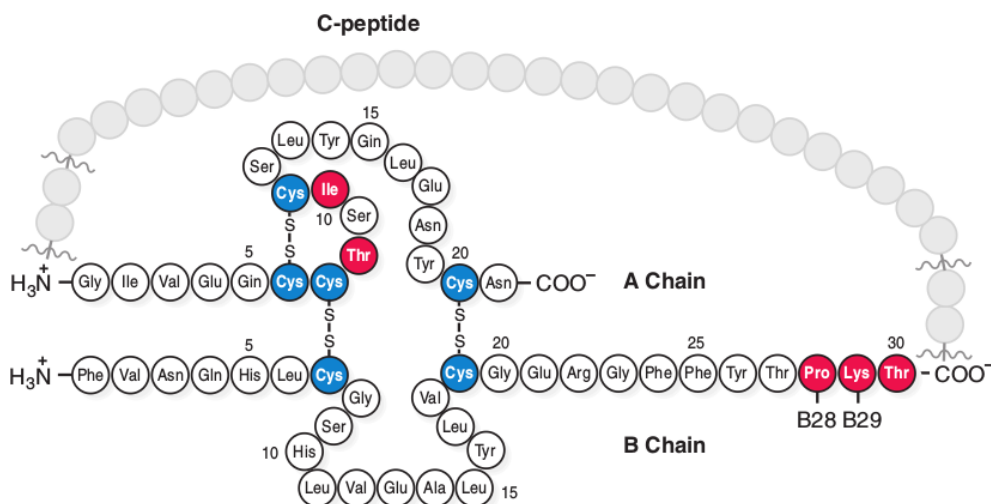
SLIKA 1.5: Prikaz struktura proteina [8]



SLIKA 1.6: Šematski prikaz struktura proteina [5]

Često se mogu izvući korisni zaključci o strukturi proteina na osnovu redosleda aminokiselina [2].

Primarna struktura Predstavlja sâmu sekvencu aminokiselina⁵ 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 1.7 [1].

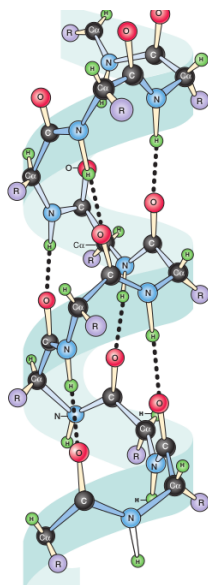


SLIKA 1.7: Prikaz primarne strukture [5]

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 sekvenci mogu imati uticaja i na više nivoje proteinskih struktura. Takve promene često dovode do lošeg presavijanja proteina i mogu dovesti do njegovog gubitka funkcije [9, 10].

Sekundarna struktura Sekundarna struktura predstavlja opšti trodimenzionalni oblik lokalnih segmenata proteina. Ovu strukturu čini nekoliko različitih oblika, od kojih su najčešći α -heliks i β -niz(ili β -struktura), a čest je i tzv. β -zavoj [1, 7]. α -heliks je 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 1.8 [1, 2].

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

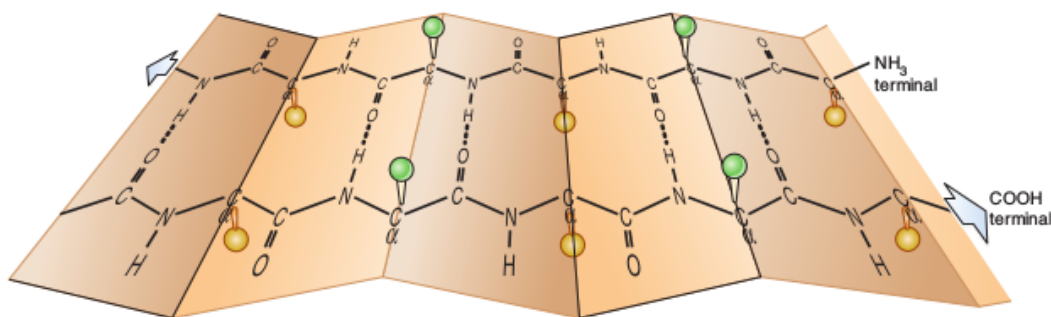
SLIKA 1.8: Prikaz α -heliksa [5]

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

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

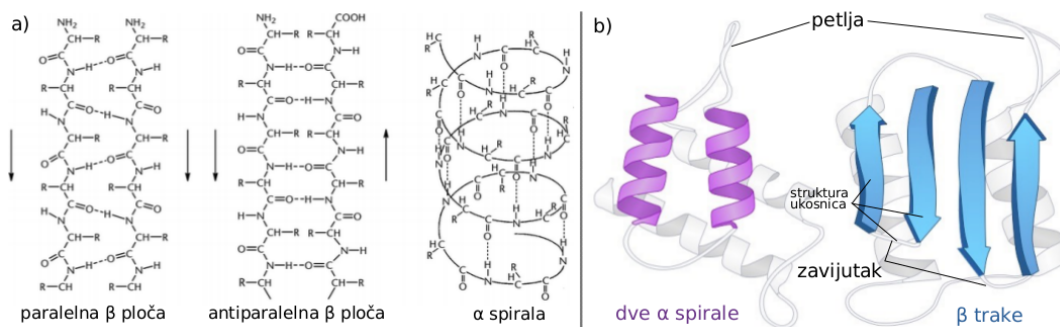
Moguće su i mešovite paralelne-antiparalelne strukture. β -nizovi se često javljaju u proteinima, a u globularnim proteinima se podjednako često javljaju i paralelne i antiparalelne. Sekundarna struktura se eksperimentalno utvrđuje na osnovu kristalne strukture proteina. Prikaz izgleda β -niza se vidi na slici 1.9 [1, 2].

β -**zavoji** - obrću pravac polipeptidnog lanca praveći kompaktan globularan oblik [10].

SLIKA 1.9: Prikaz β -niza [5]

Prikaz izgleda sekundarnih struktura se nalazi na slici 1.10.

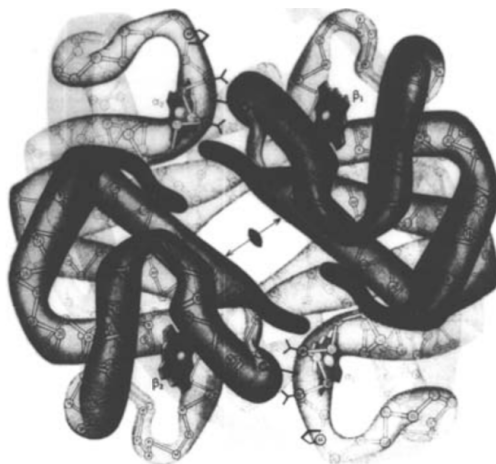
Tercijarna struktura Kod globularnih proteina se polipeptidni niz uvija u kompaktnu globulu. Tercijarna struktura proteina predstavlja unutarmolekularno slaganje polipeptidnog lanca u kompaktnu trodimenzionalnu strukturu specifičnog oblika



SLIKA 1.10: Prikaz sekundarnih struktura [11]

(globule), koja nastaje prostornim organizovanjem polipeptidnog lanca, sa sekundarnom strukturom. Na taj način se približavaju ostaci aminokiselina koji su udaljeni u sekvenci. 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 jedan od osnova za izučavanje funkcije i aktivnosti proteina. Kako bi se eksperimentalno utvrdila ova struktura vrši se rendgenska strukturna analiza. [1, 2, 7].

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 identični⁶ ili se razlikovati prema strukturi. Ovakav raspored dovodi do brzog i efikasnog transfera supstrata od jednog aktivnog centra enzima do drugog. Prikaz proteina sa kvaternarnom strukturom može se videti na 1.11 [1, 7]. Postoji



SLIKA 1.11: Prikaz hemoglobina, predstavnika globularnih proteina sa kvaternarnom strukturom [2]

nekoliko razloga iz kojih se kvaternarna struktura javila:

- Kompleksnija uloga zahteva kompleksniju strukturu

⁶Tada 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 1.12 se može videti opšti prikaz savijanja proteina.

1.1.4 Denaturacija proteina

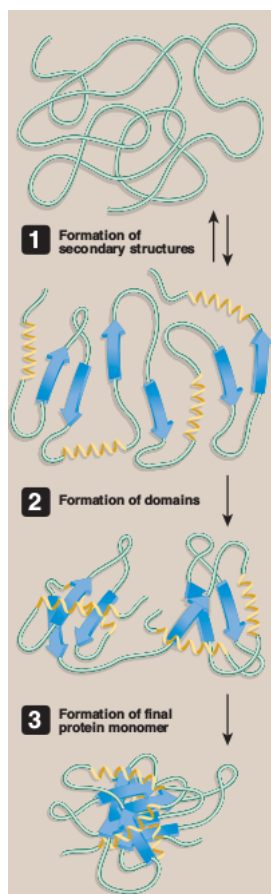
Denaturisanje proteina rezultuje u odvijanju i dezorganizaciji proteinske sekundarne i tercijarne strukture. U idealnim uslovima, denaturisanje proteina može biti *reverzibilno*. To znači da bi se protein, pri prestanku delovanja agensa, vratio u normalno stanje. Međutim, većina proteina ostaje trajno 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 [10].

1.2 Neuređenost proteina

Eksperimentalnim utvrđivanjem tercijarne 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 inherentno⁷ svojstvo sekvence. Neuređen može biti ceo protein,

⁷Inherentno = nasledeno



SLIKA 1.12: Prikaz savijanja proteina [10]

a mogu biti neuređeni određeni regioni proteina različitih dužina. Kao posledica, ovakve proteine nazivamo inherentno neuređenim proteinima, skraćeno IDP (eng. *Intrinsically Disordered Proteins*), a ako su u pitanju neuređeni, ali funkcionalni, regioni, onda je skraćenica IDPr (eng. *Intrinsically Disordered Protein Regions*). Kod ljudi, čak trećina svih proteina ima neuređenu strukturu. Neuređeni proteini su uključeni u procese stvaranja mnogih bolesti poput raka, neurodegenerativnih i kardiovaskularnih 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 foldona⁸:

- 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,

⁸Foldon ostaje u originalnom nazivu, kao posledica manjka literature [11].

- 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, 12, 13, 14, 15, 16].

1.2.1 Eksperimentalno ispitivanje neuređenosti proteina

Eksperimentalno utvrđivanje neuređenosti proteina podrazumeva laboratorijsko utvrđivanje neuređenosti korišćenjem raznih biofizičkih i biohemijskih tehnika i njihovih kombinacija. Eksperimentalno utvrđivanje neuređenosti je veoma skupo i nedovoljno brzo, zbog čega ne može da ispuni sve zahteve koje postavljaju akademija i industrija. 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 [6, 12]:

- Kristalografija X-zracima (eng. *X-ray crystallography*),
- Spektroskopija nuklearnom magnetnom rezonancom (eng. *NMR spectroscopy*),
- Cirkularni dihiroizam (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 predviđanje neuređenosti proteina

Kao posledica osobina eksperimentalnog ispitivanja neuređenosti, veliki naponi su uloženi u razvoj prediktora za računarsko predviđanje 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 reomorfnii 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 predvideti 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 će ukratko biti reči o metaprediktorima uz nekoliko primera, a zatim biće predstavljena baza podataka *DisProt* i njen značaj u ovom radu. Na kraju biće predstavljene statističke mere za procenu kvaliteta [6, 17, 18].

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 [19]:

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

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 ulazne podatke koje prediktorima dajemo na analizu, odnosno, proteinske sekvence [20].

U nastavku biće opisani neki od najpoznatijih prediktora, od kojih su *PONDR*, *DisEMBL* i *IUPRED* korišćeni za izgradnju metaprediktora u ovoj aplikaciji. Ovi prediktori su odabrani jer pokrivaju više različitih pristupa utvrđivanju neuređenosti. *PONDR* koristi neuronske mreže sa propagacijom unapred (eng. *feed-forward neural networks*) kako bi vršio predikciju nad pojedinačnim sekvencama. *DisEMBL* je neuronska mreža (eng. *Artificial Neural Network*) trenirana da predviđa zavoje, fleksibilne zavoje, nedostajuće koordinate na strukturama dobijenih pomoću X-zraka. *IUPRED* procenjuje energetska interakciju između sekvenci aminokiselina koristeći statističku procenu [22, 34, 35].

Osim prediktora, na samom kraju poglavlja navode se i dve javno dostupne baze proteina.

2.1.1 SPOT-D

SPOT-Disorder Predictor razvijen je od strane Laboratorije za strukturalnu bioinformatiku, Griffith univerziteta (eng. *Griffith University*) u Australiji. 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. 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 [21].

2.1.2 PONDR

PONDR prediktor je razvijen od strane Univerziteta države Vašington (eng. *Washington State University*) i firme Molekularna kinetika Inc. (eng. *Molecular Kinetics Inc.*). PONDR prediktor vrši predikciju nad pojedinačnim sekvencama korišćenjem neuronskih mreža sa propagacijom unapred (eng. *feedforward neural networks*) koje koriste različite attribute 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 [22].

2.1.3 IUPred

IUPred prediktor razvijen je od strane istraživačke grupe za bioinformatiku MTA ELTE Momentum (eng. *MTA-ELTE Momentum Bioinformatics Research Group*) iz

Mađarske pri Dostani laboratoriji (eng. *Dosztányi lab*). IUPred vrši previđanje neuređenosti proteina sa loše definisanom tercijskom 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 [23, 24, 25, 26].

2.1.4 ESpritz

ESpritz je prediktor razvijen od strane Iana Valša (eng. *Ian Walsh*) i Silvia Tosata (eng. *Silvio Tosatto*) pri laboratoriji za bioračunarstvo (eng. *Biocomputing*) na departmanu za Biomedicinske nauke, Univerzitet u Padovi (eng. *Department of Biomedical Sciences of the University of Padua*). ESpritz detektuje neuređene regione sekvence i bazira se na efikasnom sistemu za predviđanje koji ih pronalazi. Određivanje neuređenosti, na osnovu niza aminokiselina koje čine protein, predstavlja 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 sekvence proteina. Tipovi predviđanja neuređenosti nad kojima je ESpritz treniran su:

- Kratki x-zraci (eng. *short x-ray*): bazirano na nedostajućim atomima u strukturu koja su rešena 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 predvidelo barem jednom da je neki region neuređen. Svi ostali regioni se smatraju uređenim.

ESpritz određuje verovatnoću neuređenosti za svaki region u sekvenci [27, 28, 29, 30].

2.1.5 DisEMBL

DisEMBL je prediktor razvijen od strane Runi Lindinga (eng. *Rune Linding*) i Larsa Džula Džensena (eng. *Lars Juhl Jensen*) pri Evropskoj Molekularnoj Biološkoj Laboratoriji - EMBL (eng. *European Molecular Biology Laboratory*). *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 vjerovatnoću neuređenosti određenih regiona u sekvenci. Neuređenost proteina kod ovog prediktora posmatra se kroz model sa dva stanja - uređeno i neuređeno. Korišćeni kriterijumi su:

- Petlje / zavoji - posmatranje petlji odnosno zavoja, iako nije najpouzdaniji vid određivanja neuređenosti, može ukazati na njegovo postojanje u smislu da se neuređenost javlja baš u petljama (zavojima).
- Vruće petlje - predstavljaju finije odabran podskup za posmatranje od prethodno navedenog. Ove petlje imaju visok stepen pokretljivosti i smatraju se same po sebi neuređenim.
- Nedostajuće koordinate 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 [31].

2.1.6 Disopred2

DISOPRED2 je prediktor razvijen od strane departmana za računarske nauke Londonskog univerzitetskog koledža - UCL (eng. *University College London*). 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 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*) [32].

2.2 Metaprediktori

Predikcija strukture proteina predstavlja veliki izazov u današnjoj proteomici¹. Predikcija neuređenosti kod proteina na osnovu sekvence aminokiselina spada u veoma korisne strategije prilikom istraživanja u proteomici. Iako je u poslednjih deset godina razvijen veliki broj prediktora, nijedan od tih prediktora nije dovoljno pouzdan sam po sebi. Korišćenjem samo jednog modela za predikciju zanemaruje se nepouzdanost modela. Prediktori koriste različite pristupe utvrđivanju neuređenosti kao što su evolutivne informacije, energetske funkcije, različiti statistički metodi i metodi zasnovani na mašinskom učenju. Kombinovanjem više prediktora (a samim time i metoda), nastaju *metaprediktori* koji, veoma često metodom konsenzusa sa

¹Proteomika je oblast koja se bavi izučavanjem proteina sa fokusom na njihovu strukturu i funkcije [33].

težinama, daju pouzdanije rezultate [20, 34, 35]. U nastavku biće ukratko opisano nekoliko metaprediktora sa informacijama o tome koje prediktore koriste i koje metode primenjuju.

2.2.1 FloatCons

FloatCons metaprediktor je razvijen od strane Lukasa P Kozlovskog (eng. *Lukasz P Kozlowski*) i Janusa M Bujnickog (eng. *Janus M Bujnicki*), pri Laboratoriji za bioinformatiku i proteinsko inženjerstvo, na Međunarodnom institutu za molekularnu i ćelijsku biologiju u Poljskoj (eng. *Laboratory of Bioinformatics and Protein Engineering, International Institute of Molecular and Cell Biology*). *FloatCons* je metaprediktor koji predstavlja konsenzus trinaest proizvoljno odabranih prediktora i kao rezultat vraća skor. Pristup *meta-metoda* zasniva se na tome da različiti algoritmi imaju različite prednosti i mane, a kombinacija metoda može biti korišćena kako bi se povećala tačnost predikcije. Prediktori dobijaju sekvencu, primenjuju svoje metode predikcije, a potom se računa završni konsenzus predikcije neuređenosti. Kod računanja završnog konsenzusa dodeljene su težine prediktorima na osnovu tačnosti njihovih metoda. Prediktori korišćeni za računanje konsenzusa su:

- DisEMBL,
- DISOPRED2,
- DISpro,
- Globprot,
- iPDA,
- IUPred,
- Pdisorder,
- Poodle-s,
- Poodle-l,
- PrDOS,
- Spritz,
- DisPSSMP i
- RONN [35].

2.2.2 GSmetaDisorderMD

GSmetaDisorderMD metaprediktor je razvijen od strane Lukasa P Kozlovskog (eng. *Lukasz P Kozlowski*) i Janusa M Bujnickog (eng. *Janus M Bujnicki*), pri Laboratoriji za bioinformatiku i proteinsko inženjerstvo, na Međunarodnom institutu za molekularnu i ćelijsku biologiju u Poljskoj (eng. *Laboratory of Bioinformatics and Protein Engineering, International Institute of Molecular and Cell Biology*). *GSmetaDisorderMD* metaprediktor je zasnovan na konsenzusu rezultata prethodno navedenog metaprediktora *FloatCons* i prediktora *GSmetaDisorder3D*, zasnovanog na analizi praznina u poravnanju tercijarnih struktura proteina [35].

2.2.3 PONDR-FIT

PONDR – FIT metaprediktor su razvili Bin Ksu (eng. *Bin Xue*) et al. [34]. *PONDR – FIT* je metaprediktor zasnovan na konsenzusu rezultata šest pojedinačnih prediktora. Konsenzus se računa pomoću jednoslojne neuronske mreže. Prediktori korišćeni za računanje konsenzusa su:

- PONDR-VLXT,
- PONDR-VSL2,
- PONDR-VL3,
- FoldIndex,
- IUPRED i
- TopIDP [20, 34].

2.2.4 Positive Voter x

Positive Voter x je metaprediktor koji su razvili Mohamed F. Galvaš (eng. *Mohamed F. Ghalwash*), A. Kit Dunker (eng. *A. Keith Dunker*) i Zoran Obradović. *Positive Voter x* je metaprediktor zasnovan na metodi glasanja. Glasove o neuređenosti daju prediktor *VSL2B* i četiri klasifikaciona modela koja vrše predikcije neuređenosti na osnovu fizičkih osobina proteinske sekvence i pojedinačnih aminokiselina. Korišćeni klasifikacioni modeli su:

- SVM - predikcija metodom potpornih vektora,
- NN - predikcija neuronskom mrežom,
- CRF - predikcija uslovnim slučajnim poljem i
- Logit - predikcija logističkom regresijom.

Metaprediktor aminokiselinu klasifikuje kao neuređenu ako je bar x glasača (gde je x vrednost iz skupa $\{1, 2, 3, 4, 5\}$) klasifikovalo aminokiselinu kao neuređenu [20].

2.3 Baze podataka u bioinformatiči

Baze podataka imaju veliku ulogu u bioinformatiči. 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.²

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.³

2.4 Procena kvaliteta

Da bi se procenio kvalitet predikcija 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.

$$\text{Preciznost} = \frac{SP}{SP + LP}$$

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

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

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

$$\text{Tačnost} = \frac{SP + SN}{SP + LP + SN + LN}$$

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

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

Stvarno pozitivni, u oznaci *SP*, predstavlja broj instanci koje su klasifikovane kao pozitivne i one zaista jesu pozitivne. Stvarno negativni, u oznaci *SN*, broj je instanci koje su klasifikovane kao negativne i one zaista jesu negativne. Lažno pozitivni, u oznaci *LP*, broj je instanci koje su klasifikovane kao pozitivne, ali su one zapravo negativne. Lažno negativni, u oznaci *LN*, broj je instanci koje su klasifikovane kao negativne, ali su one zapravo pozitivne.

Opisane mere se koriste za ocenu kvaliteta svih prediktora, klasifikatora i sličnih alata i biće korišćeni i u ovom istraživanju.

²Veza ka zvaničnoj stranici: <http://www.disprot.org/>

³Veza ka zvaničnoj stranici: <https://www.uniprot.org/>

Glava 3

Aplikacija: razvoj, korišćenje i evaluacija

Jedan od glavnih ciljeva ovog rada je kreiranje aplikacije koja prikazuje izlaze iz prediktora i daje nove predikcije, kombinujući postojeće rezultate na nekoliko načina.

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

3.1 Arhitektura

Arhitekturna 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, sekvenca, koja je opcionalna, i obeležavanje prediktora koji će služiti za analizu. Na osnovu identifikatora sekvenca, ukoliko nije uneta, biva povučena iz baze, a ukoliko identifikator nije unet, program prikazuje rezultate bez poređenja sa *DisProt* bazom. Sekvenca se unosi u *.fasta* formatu, klikom na dugme "browse" ili kao sirova niska kroz tekstualno polje. Nakon toga, bira se skup prediktora koji će biti korišćeni. Ukoliko nije odabran nijedan prediktor podrazumeva se da su odabrani svi prediktori. 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.

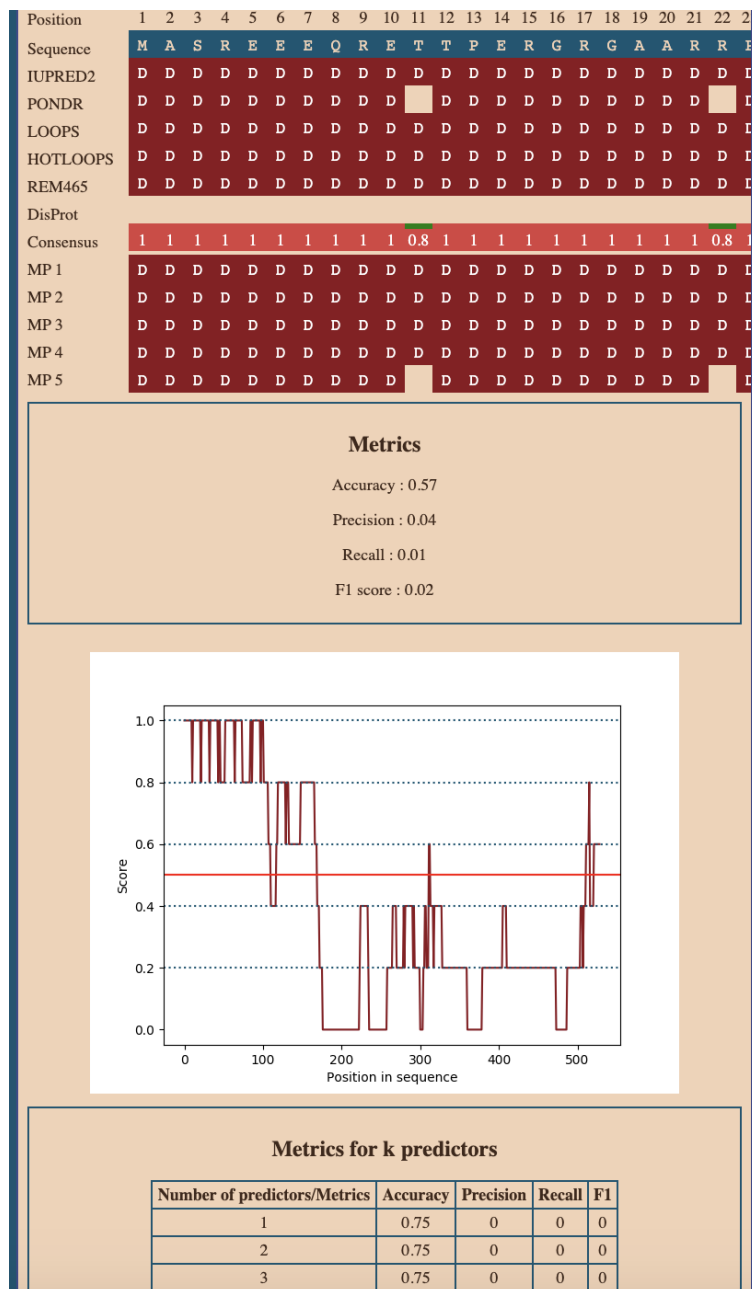
SLIKA 3.1: Prikaz korisničkog interfejsa

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 slovom "D" su predstavljene aminokiseline koje pripadaju neuređenim regionima sekvence. Ukoliko je identifikator u *DisProt* bazi unet, na osnovu njega se očitavaju i prikazuju eksperimentalno neuređeni regioni, koji se nalaze u redu označenom rečju *DisProt*, obeleženi, takođe, slovom "D" .

U narednom redu se nalazi prosečni skor prediktora (proračunat na osnovu odgovora prediktora, podrazumeva se bez korišćenja informacija iz *DisProt* baze). Ispod navedenog skora, nalaze se odluke o neuređenosti za sve metaprediktore, kada je bar k , $1 \leq k \leq \text{brojprediktora}$, prediktora predvidelo neuređenost neke aminokiseline - obeleženo sa *D*. Potom, ukoliko je dat *DisProt* identifikator, prikazane su vrednosti mera kvaliteta metaprediktora ispod kojih se nalazi grafički prikaz prosečne vrednosti skora nad sekvencom i skalom datih rezultata na *y*-osi.

Na samom kraju, nalazi se tabela koja sadrži podatke o merama kvaliteta metaprediktora u zavisnosti od broja prediktora koji su predvideli da je region neuređen. Posmatra se sledeće tvrđenje: "Aminokiselina je neuređena ukoliko je bar *k* prediktora predvidelo da je neuređena." Prikaz stranice sa rezultatima može se videti na slici 3.2.



SLIKA 3.2: Prikaz korisničkog interfejsa pri povratku rezultata sa servera za protein sa identifikatorom *DP00003*

Implementacija klijenta

Klijent je implementiran kroz programski jezik JavaScript 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, 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*. Ukoliko nisu uneti ni sekvenca ni *DisProt* identifikator, korisnik se obaveštava i zahteva se unos.

```

1 const socket = io('http://localhost:5000');
2 socket.on('connect', () => {
3   console.log('Connected to server');
4 });
5
6 function sendData() {
7   if (txtDisprotId.value == "" && txtSequence.value == ""){
8     alert("You must enter the DisProt id or the sequence!");
9   }
10  else{
11    showLoader();
12    const json = generateData();
13    console.log('Sending to server. ');
14    socket.emit('message', json);
15  }
16 }

```

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.

```

1 socket.on('predicted', data => {
2   console.log('Got from server. ');
3   console.log(data);
4   hideLoader();
5   createScale(data.sequence);
6   for (const pred of data.predictors) {
7     if (pred.result != "")
8       createPredictor(pred.result, pred.name);
9   }
10  createConsensus(data.consensus.result);
11  if(data.partial_metrics_cons != ""){
12    for (i=1; i < data.num_of_preds+1; i++){
13      if(data.partial_metrics_cons[i] != ""){
14        const name_pred = "MP " + i;
15        createPredictor(data.partial_metrics_cons[i], name_pred
16      );
17    }
18  }
19  if (txtDisprotId.value != ""){
20    addMetrics(data.metrics);
21    addDetailedMetrics(data.partial_metrics);
22  }
23  else{
24    addConsensusImage();
25    addReloadButton();
26  }
27
28 });

```


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.

Serverska strana komunicira sa klijentskom stranom korišćenjem soketa, odnosno *socketio* biblioteke. Ovaj vid komunikacije u kombinaciji sa nitima, je odabran zbog čekanja na izlaze iz prediktora. 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:

- Ukoliko je dostupan *DisProt identifikator* šalje se zahtevom preko omogućenog *API*-ja¹. Odgovor se, potom, parsira regularnim izrazima kako bi se dobili eksperimentalno neuređeni regioni - intervali regiona u sekvenci i sekvenca. Potom se ti intervali popunjavaju u listi i koja 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 zadata u *.fasta* formatu iz nje se 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 *veb* 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 *veb* 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 *rem465*. Prediktor se nalazi na *veb* 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:

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

- **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 *Predictor*, 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 mera kvaliteta metaprediktora.
- **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 je realizovana na sledeći način:

```

1 sio = socketio.Server(cors_allowed_origins=['http://localhost:9004'])
2 app = socketio.WSGIApp(sio, static_files={
3     '/': {'content_type': 'text/html', 'filename': 'index.html'}
4 })
5
6 @sio.event
7 def connect(sid, environ):
8     print('connect ', sid)
9
10 if __name__ == '__main__':
11     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.

```

1 @sio.on("message")
2 def message(sid, data):
3     print('message ', data)
4
5     # Calling all the predictors to give their opinion on the sequence
6     # in the same thread!
7     x = threading.Thread(target=prediction_calls(data), args=(data))
8     x.start()

```

Funkcija za pozivanje prediktora se sastoji iz nekoliko celina:

1. Najpre se iz prosleđenih podataka izvlače informacije o *prediktorima*, *identifikatoru* i *sekvenci* (u zavisnosti od toga šta je zadao korisnik).

```

1 def prediction_calls(data):
2     id_disp = data["disprot_id"]
3     s = data["sequence"]
4     # DisProt database call
5     [disprot, seq] = disprot_service.get_sequence_info(id_disp)
6     print(seq, len(seq))
7     if (s == ""):
8         s = seq
9     predictors_all = []
10    predictors_object = []
11
12    for i in range(0, len(data["predictors"])):
13        p = data["predictors"][i]
14        pred = {}
15        pred["name"] = p.upper()
16        if p == "iupred2":
17            pred["result"] = iupred2
18            predictors_all.append(iupred2)
19        if p == "pondr":
20            pred["result"] = pondr
21            predictors_all.append(pondr)
22        if p == "loops":
23            pred["result"] = loops
24            predictors_all.append(loops)
25        if p == "hotloops":
26            pred["result"] = hotloops
27            predictors_all.append(hotloops)
28        if p == 'rem465':
29            pred["result"] = rem465
30            predictors_all.append(rem465)
31        predictors_object.append(pred)

```

Identifikator se prosleđuje *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".

```

1 def get_sequence_info(id_disp):
2     # response = requests.get('http://www.disprot.org/ws/get/' +
3     # Old API changed on 13.09.2019.
4     response = requests.get('http://www.disprot.org/api/' +
5     id_disp)
6     data = json.loads(response.content)
7     if re.search("200", str(response)) == None:
8         data = "Not found."
9     return prepare_sequence(data)
10
11 def prepare_sequence(data):
12     if data == "Not found.":
13         return data
14     sequence = data['sequence']
15     seq_len = len(shortened_sequence(sequence))
16     intervals = []
17     disprot = []
18     disprot = ['-'] * seq_len
19     for record in data['regions']:
20         start_interval = record["start"]
21         end_interval = record["end"]
22         intervals.append((start_interval, end_interval))
23     for interval in intervals:

```

```

22         begin, end = int(interval[0]), int(interval[1])
23         for i in range(begin-1, end):
24             disprot[i] = "D"
25     return disprot, sequence

```

2. Potom se sekvenca (bez obzira na to da li je iz *DisProt* baze ili je zadata drugačije) čuva lokalno u fajlu. Ovo je neophodno kako bi mogla da bude ulaz za lokalno dostupni prediktor *IUPRED*.

```

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

```

3. Nakon toga pozivaju se prediktori da pojedinačno daju rezultat. Prediktor *SPOTD* je izuzet iz računanja, jer zavisi od servera na kom je prediktor smešten, a koji je često nedostupan ili veoma spor .

```

1  # Calling predictors
2  iupred2 = predictor_service.iupred2_predict(s)
3  #SPOTD: uncomment the function call
4  #spotd = predictor_service.spotd_predict(s)
5  pondr = predictor_service.pondr_predict(s)
6  ss = shortened_sequence(s)
7  [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 kroz polje *self.calculated*. Navedene funkcije nalaze se u datoteci *predictor_service.py*. Metod *calculate()* za svaku instancu klase *Predictor* izračunava neuređene regione i smešta ih u *self.calculated*.

```

1  def iupred2_predict(sequence):
2      p = iupred2(sequence)
3      return p.calculate()
4
5  def spotd_predict(sequence):
6      p = spotd(sequence)
7      return p.calculate()
8
9  def pondr_predict(sequence):
10     p = pondr(sequence)
11     return p.calculate()
12
13  def disembl_predict(sequence):
14     p = disembl(sequence)
15     return p.calculate()

```

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

```

1  class Predictor:
2      def __init__(self, sequence):
3          self.sequence = sequence
4          self.calculated = []
5
6      def calculate(self):
7          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 izgleda jednog od prediktora kod koga se predikcija vrši preko veb stranice:

```

1 class pondr(Predictor):
2     def calculate(self):
3         global br
4         url = "http://www.pondr.com/cgi-bin/pondr.cgi"
5         br.set_handle_robots(False)
6         br.open(url)
7         br.form = list(br.forms())[1]
8         br['ProteinName'] = "test"
9         br['Sequence'] = self.sequence
10        response = br.submit()
11        soup = BeautifulSoup(response.read(), features='html5lib')
12        soup = soup.prettify()
13        result = re.findall("VLXT\t\s.*", soup)
14        # We don't need the first one because it is not sequence
15        result = result[1:]
16        predicted = []
17        for r in result:
18            pred = r[6:]
19            predicted.append(pred)
20        total = []
21        # Getting everything in one sequence
22        for i in predicted:
23            total += i
24        old = len(total)
25
26        pom = shortened_sequence(self.sequence)
27        if(old < len(self.sequence)):
28            for i in range(0, len(pom) - old):
29                total.append(" ")
30        total = [x if x != ' ' else '-' for x in total ]
31
32        self.calculated = total
33        return self.calculated

```

4. Računa se prosečna vrednost skora predikotra, a potom se računaju procene mera kvaliteta prediktora uz pomoć funkcije *measure_all* koja se nalazi u datoteci *measure.py*. Prvo se ovaj postupak radi za ceo metaprediktor, a potom za svaki koji koristi *k* prediktora pojedinačno. U slučaju da *DisProt* identifikator nije unet, neophodno je vratiti prazne niske za rezultate.

```

1     cons = consensus(ss, predictors_all)
2     cons_decs = {}
3     for i in range(1, len(predictors_all)+1):
4         cons_tres = consensus_treshold(cons, 1/len(predictors_all)*
5 i)
6         cons_decs[i] = cons_tres
7     if (id_disp != ""):
8         bin_disprot = binarize_values(disprot)
9         m = measure_all(consensus_treshold(cons), bin_disprot)
10        # Doing consensus and metrics for when one, two or more
11        predictors said "D"
12        consensus_values = {}
13        for i in range(1, len(predictors_all)+1):
14            consensus_values[i] = measure_all(cons_tres,
15 bin_disprot)
16        pred = {}
17        pred['name'] = "DisProt"

```

```

15     pred['result'] = disprot
16     predictors_object.append(pred)
17     print(predictors_object)
18     else:
19         # If there is no DisProt id, only predictors results
           should be displayed.
20         m = []
21         for i in range(0,4):
22             m.append("")
23         consensus_values = ""
24         disprot = ""

```

Za računanje prosečne vrednosti skora prediktora 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 akumulirana vrednost u *val* se deli sa brojem prediktora. Na kraju se poziva funkcija za iscrtavanje grafika konsenzusa.

```

1 def consensus(sequence, predictors):
2     num_of_preds = len(predictors)
3     score = []
4     for i in range(0, len(sequence)):
5         val = 0
6         for pred in predictors:
7             if pred[i] == 'D':
8                 val += 1
9         score.insert(i, val / num_of_preds)
10    plot_consensus(score, sequence)
11    return score

```

Kako bi se odredili prosečna vrednost skora i mere kvaliteta metaprediktora kada treba da se posmatra slučaj da je jedan, dva ili više prediktora predvidelo da je region neuređen, vrši se nekoliko poziva:

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

```

1 def consensus_threshold(cons, threshold=0.5):
2     new = []
3     print(threshold)
4     for c in cons:
5         if c >= threshold:
6             new.append(1)
7         else:
8             new.append(0)
9     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 mera kvaliteta, ona se vrši funkcijom *binarize_values* iz datoteke *measures.py*.

```

1 def measure_all(x,y):
2     a = round(m.accuracy_score(x, y), 2)
3     p = round(m.precision_score(x, y), 2)

```

```

4     f = round(m.f1_score(x, y), 2)
5     r = round(m.recall_score(x, y), 2)
6     return [a, p, r, f]
7
8     def binarize_values(x):
9         new = []
10        for el in x:
11            if el == "D":
12                new.append(1)
13            else:
14                new.append(0)
15        return new

```

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

```

1 result = {
2     "predictors": predictors_object,
3     "consensus": {
4         'name': 'cons',
5         'result': cons
6     },
7     "sequence": ss,
8     "metrics" : [
9         {
10            'name': 'Accuracy',
11            'result' : m[0]
12        },
13        {
14            'name': 'Precision',
15            'result': m[1]
16        },
17        {
18            'name': 'Recall',
19            'result': m[2]
20        },
21        {
22            'name': 'F1 score',
23            'result': m[3]
24        }
25    ],
26     "partial_metrics": consensus_values,
27     "partial_metrics_cons": cons_decs,
28     "num_of_preds": predictors_num
29 }
30
31 sio.emit('predicted', result)

```

3.2 Korišćenje aplikacije

Korišćenje aplikacije je veoma jednostavno i linearno. Od korisnika se očekuje da obezbedi *DisProt* identifikator i/ili sekvencu (*.fasta* datoteke). Postoji više načina unosa podataka:

- Prvi način je unošenjem identifikatora proteina iz *DisProt* baze. Ovaj način ne zahteva i unos sekvence, jer se ona može uzeti iz baze.
- Drugi način je davanjem samo *DisProt* identifikatora i zadavanjem sekvence. Tada se koristi sekvenca koju je zadao korisnik, a ne ona iz *DisProt* baze.

- Treći način je unosom samo sekvence, bez davanja *DisProt* identifikatora. Ovaj način se razlikuje od prethodna dva zbog toga što se ne dobijaju mere procene kvaliteta.

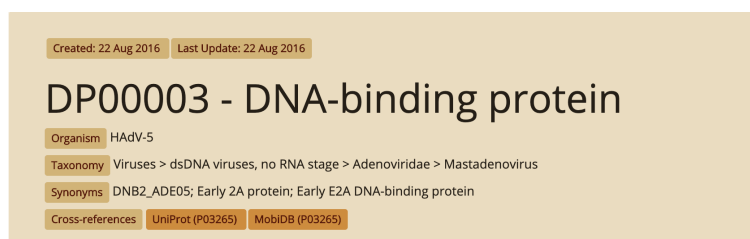
Korišćenje aplikacije se svodi na naredni skup koraka:

1. Unos *DisProt* identifikator i/ili sekvencu. Ukoliko se sekvenca unosi iz datoteke, onda se klikom na *Browse* dugme vrši odabir datoteke.
2. Klik na *Submit* dugme.
3. Pregled rezultata.

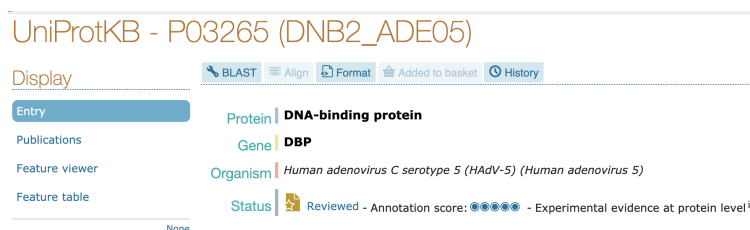
3.2.1 Primer upotrebe

Prvi primer - zadavanje *DisProt* identifikatora

Za prvi primer upotrebe biće uzeta niska sa identifikatorom *DP00003* u *DisProt* bazi. Niska je odabrana iz *DisProt* baze, kao što se može videti na slici 3.3. Odgovarajuća *.fasta* datoteka je dostupna u *Uniprot* bazi podataka. Prikaz dela veb stranice navedene baze i identifikatora niske nalazi se na slici 3.4. Za prvi primer upotrebe odabran je unos samo *DisProt* identifikatora.



SLIKA 3.3: Prikaz *DisProt* baze



SLIKA 3.4: Prikaz *UniProt* baze

Aplikacija nudi korisniku mogućnost da unese identifikator ili sekvencu proteina u odgovarajuća polja za unos teksta što je prikazano na slici 3.5.

Protein Disorder Metapredictor
A predictor that uses a method of consensus to determine if the protein is disordered or not.

[About](#)
[Instructions](#)
[Predictors](#)

DisProt database ID:

Enter the AminoAcid sequence:

OR Choose a .fasta file from your files.

SLIKA 3.5: Prikaz polja za unos

Nakon zadavanja ulaznih podataka, proteinska sekvenca se prikazuje u tekstualnom polju čime se dobija izgled kao na slici 3.6.

Protein Disorder Metapredictor
A predictor that uses a method of consensus to determine if the protein is disordered or not.

[About](#)
[Instructions](#)
[Predictors](#)

DisProt database ID:

Enter the AminoAcid sequence:

OR Choose a .fasta file from your files.

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, gde je moguće izvršiti odabir prediktora, kao i veze ka njihovim zvaničnim stranicama. Klikom na dugme "Submit" podaci se šalju na server. U ovom primeru odabrani su svi prediktori.

List of predictors being used:
Predictors' names contain links to official pages. Choose predictors to be used, if none selected all will be applied.

I. IUPRED2

II. PONDR

III. DISEMBL:
 Loops/Coils
 HotLoops
 Rem465

IV. SPOTD*

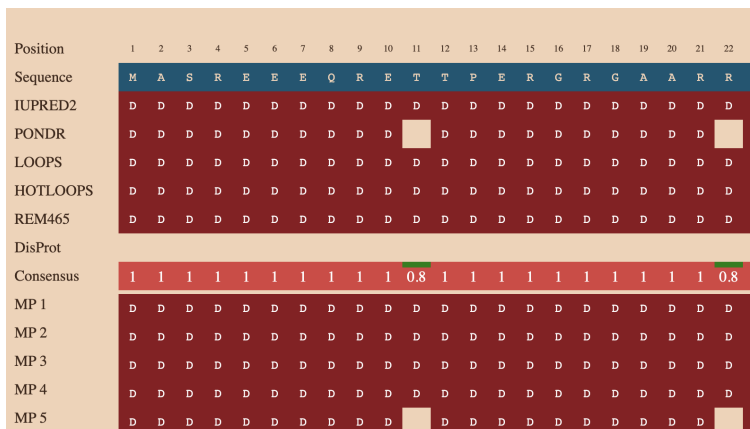
*SPOTD predictor is very slow and depends on their server, to use it uncomment "spotd" lines in the code and send it in predictors.

The analysis itself might take a while. Be patient.

Made by Una Stanković © 2019

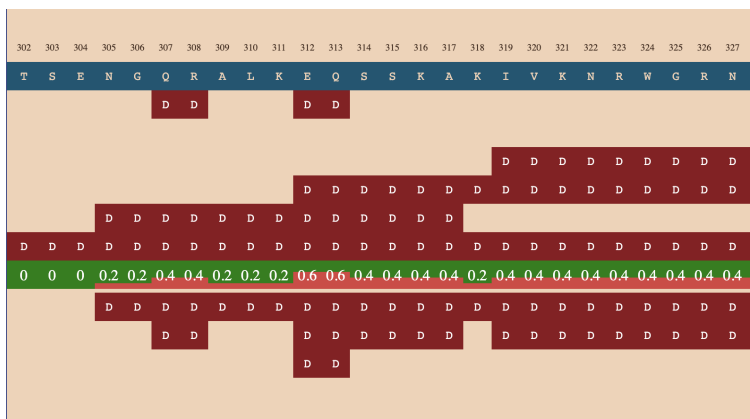
SLIKA 3.7: Prikaz liste dostupnih prediktora i slanja informacija ka serveru

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 (ukoliko je zadat identifikator), kao i prosečna vrednost skorova za svaku aminokiselinu po svim prediktorima (Consensus), pri čemu se podrazumeva da prediktor dodeljuje određenoj aminokiselini skor 1 ako je predvideo da je neuređena, a 0 u suprotnom. Polje je obojeno u crveno u delu veličine koji odgovara prosečnoj vrednosti skorova. Ispod skorova, nalazi se prikaz neuređenih aminokiselina za metaprediktor koji koristi k prediktora. Prikaz stranice sa opisanim rezultatima nalazi se na slici 3.8.

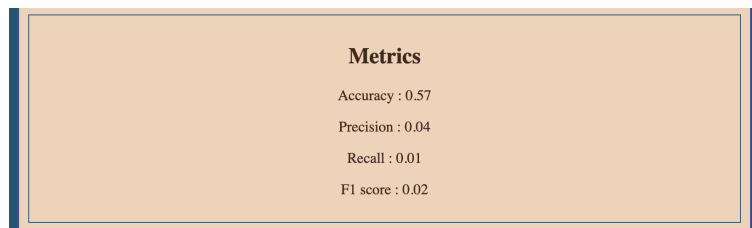


SLIKA 3.8: Prikaz dobijenih rezultata

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

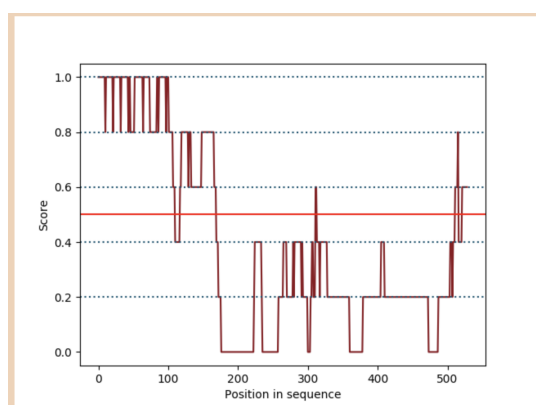
SLIKA 3.9: Prikaz dobijenih rezultata na kom se vidi i izlaz iz *DisProt* baze

Ispod navedenih rezultata, ukoliko je dat *DisProt* identifikator, nalazi se prikaz vrednosti mera kvaliteta metaprediktora za datu prosečnu vrednost skora (koja odgovara situaciji kada su barem 3 prediktora predvidela neuređenost) i rezultate iz *DisProt* baze. U obzir su uzete četiri statističke mere: tačnost, preciznost, odziv i F1 mera. Izgled prikaza procenjenih mera vidi se na slici 3.10



SLIKA 3.10: Prikaz vrednosti mera kvaliteta metaprediktora

Sledeća stvar od interesa je grafik neuređenosti na osnovu prosečne vrednosti skora. Na grafiku se nalazi visina prosečne vrednosti skora u odnosu na poziciju u sekvenci. Analizom dobijenih rezultata možemo uvideti oko kojih regiona se prediktori slažu. Prikaz grafika može se videti na slici 3.11.



SLIKA 3.11: Grafički prikaz dobijenih rezultata

S obzirom da su analizirane i mere kvaliteta za metaprediktore za koje je bar k prediktora predvidelo neuređenost, njihov prikaz se vrši na samom dnu strane u vidu tabele prikazane na 3.12. U tabeli su navedene mere kvaliteta i broj k koji je korišćen.

Da se podsetimo, kao što je navedeno u podsekciji 3.1.1, aminokiselina se smatra neuređenom ukoliko je barem k prediktora predvidelo da je neuređena. Preciznije, traži se da je barem k prediktora (od svih dostupnih prediktora) predvidelo neuređenost i na osnovu njihovih predviđanja se vrši procena kvaliteta. Broj k u prikazanoj tabeli je 5, jer su iskorišćeni svi dostupni prediktori. Kao što je navedeno u podsekciji 3.1.2, prediktor *DisEMBL* čini skup tri prediktora, za koje, svaki pojedinačno, vraća rezultat, pa se on može posmatrati kao tri prediktora, a osim njega korišćeni su *IUPRED2* i *PONDR*. Napominje se još jednom da su mere kvaliteta vidljive u rezultatima samo ukoliko postoji *DisProt* identifikator.

Nakon što su dobijene vrednosti mera kvaliteta, one se ispisuju u tabeli prikazanoj na slici 3.12. Iz tabele se vidi da su dobijene mere kvaliteta veoma niske, što će biti bliže razmatrano u podsekciji 3.3.

Number of predictors/Metrics	Accuracy	Precision	Recall	F1
1	0.29	0.92	0.11	0.2
2	0.5	0.35	0.07	0.12
3	0.57	0.04	0.01	0.02
4	0.64	0	0	0
5	0.75	0	0	0

SLIKA 3.12: Tabelarni prikaz dobijenih rezultata

Pored 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 instrukcije koja je prikazana na slici 3.14.

About the project

This project is a part of the master thesis created by *Una Stanković* at the *Faculty of Mathematics, University of Belgrade*.

The main goal of this program is to make a consensus about whether the structure of the protein is **disordered** or not. Various predictors were used to do the consensus and their descriptions, in detail, can be found on predictor pages. Prediction of the protein structure is very important in the field of *bioinformatics* today and through computational analysis scientists are able to, in a faster way, determine if the disorder is present in the protein sequence and where such disorder happened.

This program is helping compare such computational predictors outputs and provide insights into the decisions made by predictors, as well as visualize the given consensus. Apart from that, various important metrics' values are given to provide even more insight into the results. Metrics used are:

- Accuracy
- Precision
- Recall
- F-measure

For the consensus, each predictor has an equal weight although it is planned to give a possibility of providing weights to the predictors. Moreover, the bigger number of predictors can be added thanks to the structure of this program. (This is certainly not an easy thing to do since most of the predictors provide no API or any easy way to create communication channel.)

SLIKA 3.13: Prikaz stranice koja se odnosi na dodatne informacije

Protein Disorder Metapredictor

A predictor that uses a method of consensus to determine if the protein is disordered or not.

Instructions

TLDR In order to run the program, enter the DisProt id and/or the sequence (copied from someplace or in a form of a fasta file by pressing the browse button) and press the submit button.

DisProt database contains information about experimentally determined disorder of protein regions as well as the proofs of such experiments. *DisProt id* of the sequence is necessary in order to get metrics. Experimentally determined disorder will be compared with the predictors' outputs. Next step to do is optional, choose a sequence from fasta file, or enter it into the text area below. (If the sequence is entered through file it will be shown in a text area.) The final step is to click the submit button which will send the results to the server and retrieve the information about disordered regions and calculated consensus. In order to run another analysis at the end of the results page there is a "Run another analysis" button.

SLIKA 3.14: Prikaz stranice sa uputstvom za korišćenje aplikacije

Drugi primer - zadavanje sekvence iz datoteke i izbor prediktora

Naredni primer predstavlja primer upotrebe zadavanjem sekvence iz datoteke i odabirom nekih od prediktora iz liste dostupnih. Sekvenca je preuzeta iz *UniProt* baze i reč je o sekvenci sa *DisProt* identifikatorom *DP00005*, ali on neće biti unesen.

Klikom na dugme *Browse* bira se datoteka iz lokalnog direktorijuma. Odabir prediktora vrši se klikom na dugme za odabir pored imena prediktora, što se može videti na slici 3.15. Odabrani prediktori su *IUPRED* i *PONDR*.

DisProt database ID:

Enter the AminoAcid sequence:
 >sp|P03045|REGN_LAMBDA_Antitermination
 protein N OS=Escherichia phage lambda
 OX=10710 GN=N PE=1 SV=3
 MDAQTRRRERRRAEKDAQWKAANPLLGVSAKPV
 NRPLSLNRKPKSRVESALNPDITVL
 AEYHKQIESNLQRIERKNQRTWYKPGERGITCSG
 RQKIKGKSIPLI

OR C:\fakepath\DP00005.fasta

List of predictors being used:
 Predictors' names contain links to official pages. Choose predictors to be used, if none selected all will be applied.

I. IUPRED2
 II. PONDR
 III. DISEMBL:
 Loops/Coils
 HotLoops
 Rem465
 IV. SPOTD*

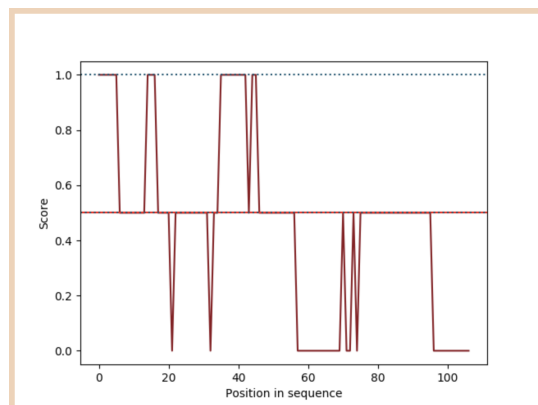
*SPOTD predictor is very slow and depends on their server, to use it uncomment "spotd" lines in the code and send it in predictors.

SLIKA 3.15: Prikaz stranice sa unosom datoteke preko *Browse* dugmeta i odabirom prediktora

Klikom na *Submit* dugme podaci o odabranoj sekvenci i prediktorima šalju se na server, koji potom vraća odgovor. Odgovor ne sadrži mere kvaliteta, kao ni sekvencu u *DisProt* bazi, zato što njen identifikator nije zadat. Na slikama 3.16 i 3.17 prikazan je celokupni odgovor servera.

Position	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	2
Sequence	M	D	A	Q	T	R	R	R	E	R	R	A	E	K	Q	A	Q	W	K	A	A	N	E
IUPRED2	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D
PONDR	D	D	D	D	D												D	D	D	D	D	D	E
Consensus	1	1	1	1	1	1	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	1	1	1	0.5	0.5	0.5	0.5	0	0
MP 1	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	E
MP 2	D	D	D	D	D												D	D	D				

SLIKA 3.16: Prikaz stranice sa rezultatima koje je vratio server



SLIKA 3.17: Prikaz grafika sa rezultatima koje je vratio server

Odgovor od servera sadrži samo prediktore koji su zadati od strane korisnika, na osnovu kojih se računa prosečna vrednost skora prediktora. Na grafiku se iscrtavaju pragovi na osnovu broja prediktora.

3.3 Evaluacija metaprediktora

Razvijena aplikacija pruža mogućnost za uporedni prikaz rezultata nekoliko postojećih prediktora kao i razvijenog metaprediktora. Rezultati su prikazani tabelarno i grafički. Pored toga, izvršena je evaluacija razvijenog metaprediktora.

Za potrebe procene kvaliteta dobijenih rezultata predikcije korišćene su statističke mere preciznost, odziv, tačnost i F-mera. Procena kvaliteta predikcije vrši se u odnosu na informacije o eksperimentalno utvrđenim neuređenim regionima dobijenim iz *DisProt* baze.

Pristup korišćen za prikaz vrednosti mera kvaliteta metaprediktora dobijenih poređenjem rezultata prosečne vrednosti skora svih prediktora sa pragom 0.5 i eksperimentalno utvrđenih neuređenih regiona. Ovim pristupom, procenjuje se kvalitet rezultata kompletnog metaprediktora. U pojedinim situacijama, primećuje se da rezultati metaprediktora nisu uvek u skladu sa eksperimentalno utvrđenim rezultatima, već se u značajnoj meri razlikuju. Razlog ove pojave može se tražiti u nesavršenosti samih prediktora koji učestvuju u konsenzusu, kao i u potencijalno niskoj pokrivenosti rezultata eksperimenata nad određenim proteinima.

Za svaki prediktor su izračunate mere kvaliteta metaprediktora za vrednosti dobijene odnosom konsenzusa k prediktora ($1 \leq k \leq n, n = 5$) i eksperimentalno utvrđenih neuređenih regiona. Ovim pristupom merimo rezultate dobijene posmatranjem manjeg ili većeg broja prediktora i konsenzusa njihovih odluka. Vrednosti metrika koje se dobijaju ovim putem vrlo često se, posebno za *odziv* i *F*-meru, svode na niske vrednosti što je i razumljivo s obzirom na to da treba da se uklopi nekoliko faktora:

1. više od k prediktora treba da predvidi za neki region da je neuređen.
2. da bi taj region bio uzet u obzir, potrebno je da se rezultat prediktora poklopi sa rezultatima iz *DisProt* baze.

U tabeli 3.1 nalazi se prikaz metrika za sekvencu sa identifikatorom *DP00003*, dok se u tabelama 3.2 i 3.3 nalaze se prikazi metrika za sekvence sa identifikatorima

DP00004 i *DP00005*. Ove sekvence su odabrane kako bi se pokazala razlika u dobijenim metrikama u odnosu na to koliko aminokiselina je vraćeno iz baze kao neuređeno. *DP00003* predstavlja jednu reprezentativnu sekvencu - ima i uređene i neuređene regione. Sa druge strane, sekvence *DP00004* i *DP00005* predstavljaju dve krajnosti. Naime, prva sekvenca je u potpunosti eksperimentalno uređena, dok je druga eksperimentalno u potpunosti neuređena. Nakon što program prikaže rezultate predikcije za sve tri sekvence može se videti da je broj mera sa ne-nula vrednostima značajno veći kod sekvence *DP00005* što se može videti u tabeli 3.3. To se može objasniti time što je cela sekvenca *DP00005* neuređena, čime se praktično posmatra samo izlaz iz metaprediktora. Sa druge strane, kod sekvence *DP00004* prisutne su mere kvaliteta samo za prvu kolonu - tačnost, gde vrednosti značajno rastu kada se u analizu uključi više od dva prediktora. Vrednost za ostale mere su 0.

Broj prediktora/Metrika	Tačnost	Preciznost	Odziv	Fmera
1	0.29	0.92	0.11	0.2
2	0.5	0.35	0.07	0.12
3	0.64	0	0	0
4	0.64	0	0	0
5	0.75	0	0	0

TABELA 3.1: Prikaz mera kvaliteta metaprediktora za sekvencu *DP00003*

Broj prediktora/Metrika	Tačnost	Preciznost	Odziv	Fmera
1	0.18	0	0	0
2	0.45	0	0	0
3	0.94	0	0	0
4	0.94	0	0	0
5	0.99	0	0	0

TABELA 3.2: Prikaz mera kvaliteta metaprediktora za sekvencu *DP00004*

Broj prediktora/Metrika	Tačnost	Preciznost	Odziv	Fmera
1	0.89	0.89	1	0.94
2	0.83	0.83	1	0.91
3	0.58	0.58	1	0.73
4	0.12	0.12	1	0.22
5	0	0	0	0

TABELA 3.3: Prikaz mera kvaliteta metaprediktora za sekvencu *DP00005*

Kako bi se procenio kvalitet metaprediktora nad većem skupu sekvenci, metaprediktor je pušten nad svim sekvencama u *DisProt* bazi (652 sekvence). Preciznije, urađen je prosek procena mera kvaliteta dobijenih rezultata za svaki od metaprediktora koji koristi k prediktora. Dobijeni rezultati predstavljeni su u tabeli 3.4.

Broj prediktora/Metrika	Tačnost	Preciznost	Odziv	Fmera
1	0.496	0.89	0.394	0.478
2	0.647	0.678	0.452	0.477
3	0.697	0.265	0.491	0.299
4	0.697	0.265	0.491	0.299
5	0.653	0.072	0.314	0.103

TABELA 3.4: Prikaz mera kvaliteta metaprediktora za sve sekvence u *DisProt* bazi

Glava 4

Zaključak

Aplikacija kreirana u okviru ovog istraživanja omogućava korisniku da na veoma jednostavan način dobije pregled neuređenih regiona proteina, kroz jasan interfejs. Značaj rada je u sticanju iskustva u izgradnji aplikacije koja objedinjuje odluke više prediktora i stavlja ih, praktično, pod okrilje jedne aplikacije. Moguća unapređenja rada ogledaju se u povećanju brojnosti dostupnih prediktora, kao i u dodeljivanju težina prediktorima prilikom računanja konsenzusa u zavisnosti od tipa proteina koji se analizira.

Bibliografija

- [1] Vesna Spasojević-Kalimanovska Slavica Spasić Zorana Jelić-Ivanović. *Opšta biohemija*. 2002.
- [2] Vesna Niketić. *Principi strukture i aktivnosti proteina*. Hemijski fakultet, 1995.
- [3] Marija Jeličić. *Povezanost dužine epitopa i uredenostidelova 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] *Proteinska struktura*. https://sr.wikipedia.org/sr-el/Proteinska_struktura.
- [9] 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.
- [10] 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.
- [11] 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.
- [12] A.Keith Dunker et al. *Intrinsically disordered protein*. on-line na: [https://doi.org/10.1016/s1093-3263\(00\)00138-8](https://doi.org/10.1016/s1093-3263(00)00138-8). 2001.
- [13] 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.
- [14] 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.

- [15] C.J.; Dunker A.K Uversky V.N.; Oldfield. *Intrinsically disordered proteins in human diseases: Introducing the D2 concept*. 2008.
- [16] 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.
- [17] V.N. Uversky. *Intrinsically disordered proteins from A to Z*. 2011.
- [18] Han K.H. Tompa P. *Intrinsically disordered proteins*. 2012.
- [19] 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.
- [20] 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.
- [21] Paliwal K1 Zhou Y2. Hanson J1 Yang Y2. *Improving protein disorder prediction by deep bidirectional long short-term memory recurrent neural networks*. on-line na: <https://www.ncbi.nlm.nih.gov/pubmed/28011771>. 2017. DOI: 10.1093/bioinformatics/btw678.
- [22] *Predictor Of Naturally Disordered Regions*. <http://www.pondr.com/pondr-tut2.html>.
- [23] 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.
- [24] S.O. et al. Garbuzynskiy. *To be folded or to be unfolded?* 2004.
- [25] K.A. Thomas P.D. i Dill. *An iterative method for extracting energy-like quantities from protein structures*. 1996.
- [26] et al. Dosztányi Z. *The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins*. 2005.
- [27] 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.
- [28] S.; Frasconi P.; Soda G.; Pollastri G. Baldi P.; Brunak. *Exploiting the past and the future in protein secondary structure prediction*. 1999.
- [29] 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.
- [30] J.L. Sussman O. Noivirt-Brik J. Prilusky. *Assessment of disorder predictions in CASP8*. 2009.
- [31] Francesca Diella Peer Bork Toby J Gibson Robert B Rusell Rune Linding Lars Juhl Jensen. *Protein Disorder Prediction - Implications for Structural Proteomics*. 2003.
- [32] 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.

-
- [33] Anderson NG Anderson NL. *Proteome and proteomics: new technologies, new concepts, and new words*. 1998.
- [34] Robert W. Williams A. Keith Dunker Vladimir N. Uversky Bin Xue Roland L. Dunbrack. *PONDR-FIT: A Meta-Predictor of Intrinsically Disordered Amino Acids*. 2010. DOI: 10.1016/j.bbapap.2010.01.011.
- [35] Janusz M Bujnicki Lukasz P Kozlowski. *MetaDisorder: a meta-server for the prediction of intrinsic disorder in proteins*. 2012. DOI: 10.1186/1471-2105-13-111.