

UNIVERZITET U BEOGRADU
MATEMATIČKI FAKULTET

MASTER RAD

Predviđanje funkcije proteina

Autor:
Anja BUKUROV

Mentor:
dr Jovana KOVAČEVIĆ

ČLANOVI KOMISIJE:

dr Jovana Kovačević

...

...



Beograd, 2019

UNIVERZITET U BEOGRADU

Sažetak

Matematički fakultet
Katedra za Računarstvo i informatiku

Informatičar

Predviđanje funkcije proteina

by Anja BUKUROV

Coming soon!

Zahvalnice

Sadržaj

Sažetak	iii
Zahvalnice	v
1 Uvod	1
2 Biološke osnove	3
2.1 Proteini	3
2.1.1 Funkcije i osobine proteina	4
2.1.2 Struktura proteina	5
2.1.3 Savijanje proteina	8
2.1.4 Denaturacija proteina	8
2.2 Neuređenost proteina	9
2.2.1 Eksperimentalno ispitivanje neuređenosti proteina	10
2.2.2 Računarsko ispitivanje neuređenosti proteina	11
3 Predikcija neuređenosti proteina	13
3.1 Prediktori	13
3.1.1 SPOT-D	14
3.1.2 PONDR	14
3.1.3 s2D	14
3.1.4 IUPred	14
3.1.5 ESpritz	15
3.1.6 SEG	15
3.1.7 Disopred2	15
3.2 Baza podataka DisProt	15
4 Aplikacija	17
4.1 Arhitektura	17
4.2 Funkcionalnosti	17
4.3 Korišćenje aplikacije	17
4.4 Primer upotrebe	17
5 Implementacija	19
6 Zaključak	21
Bibliografija	23

Spisak slika

2.1	Prikaz spajanja α -karboksilne grupe jedne aminokiseline i α -amino grupe druge aminokiseline, pri čemu se oslobađa molekul vode [4].	4
2.2	Prikaz centralne dogme molekularne biologije [5]	5
2.3	Prikaz struktura proteina	6
2.4	Šematski prikaz struktura proteina [4]	7
2.5	Prikaz primarne strukture [4]	8
2.6	Prikaz α -heliksa [4]	9
2.7	Prikaz β -strukture [4]	9
2.8	Prikaz sekundarnih struktura [9]	10
2.9	Prikaz savijanja proteina [8]	11

Spisak tabela

2.1 Spisak esencijalnih i neesencijalnih aminokiselina	4
--	---

Tati, mami i Olgi...

Glava 1

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 rasporeda aminokiselina i utiče na njegovu funkciju. Primarna struktura podrazumeva niz aminokiselina koje učestvuju u izgradnji proteina, dok se sekundarna odnosi na oblik koji protein zauzima u prostoru (spiralna ili traka). Proteine sa nestabilnom sekundarnom strukturom nazivamo neuređenim. Pored značajne uloge u obavljanju brojnih bioloških funkcija, otkriveno je i da postoji veza između ovih proteina i razvoja savremenih, neizlečivih bolesti i zbog toga su oni u fokusu bioinformatičke zajednice.

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

Glava 2

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. Potom, biće opisana svaka od struktura ponaosob, uz priložen grafički prikaz istih. Na kraju, posebno će biti opisani neuređeni proteini, njihova uloga i uzroci koji mogu dovesti do njihove pojave.

2.1 Proteini

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

Proteini su jedinjenja sačinjena od 100 ili više aminokiselina. Za razliku od njih, peptide čine manje skupine aminokiselina. Tako imamo:

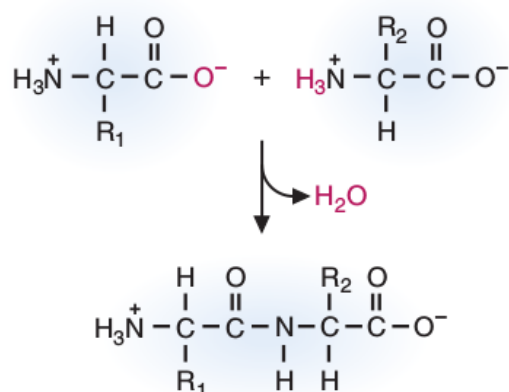
- oligopeptide - sastoje se od 10 ili manje aminokiselina, među njih spadaju dipeptidi, tripeptidi, itd. i
- polipeptide - sastoje se od 100 ili manje aminokiselina.

Iz navedenog lako se uočava da se proteini sastoje od više spojenih polipeptida. Proteini i peptidi su izgrađeni od 22¹ *L* – *aminokiseline* ² koje se javljaju u prirodi i povezani su peptidnim vezama, koje nastaju između α -karboksilne grupe jedne aminokiseline i α -amino grupe druge aminokiseline, pri čemu se oslobađa molekul vode (grafički prikaz se može videti na slici 2.1). Ovim postupkom nastaje nerazgranati polipeptidni lanac izgrađen od glavnog lanca, koji se pravilno ponavlja, i međusobno različitih ogranaka. Standardna grupa aminokiselina se može podeliti na esecijalne i neesecijalne, čiji spisak se može videti u tabeli 2.1 [2, 3].

Svaki molekul proteina nastaje u ćeliji živog organizma. Proteinski lanac se sastoji od određenog broja i vrsti aminokiselina poređanih po specifičnom redosledu.

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

²*L* – *aminokiseline* su one sa levom prostornom konfiguracijom, analogno, postoje i *D* – *aminokiseline*, sa desnom



SLIKA 2.1: Prikaz spajanja α -karboksilne grupe jedne aminokiseline i α -amino grupe druge aminokiseline, pri čemu se oslobađa molekul vode [4].

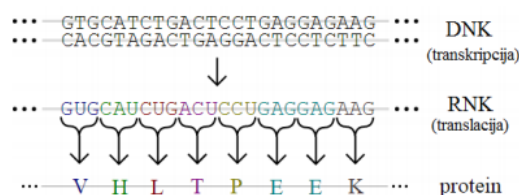
Esencijalne	Neesencijalne
Arginin	Alanin
Histidin	Asparagin
Leucin	Asparaginska kiselina
Izoleucin	Cistein
Lizin	Glutaminska kiselina
Metionin	Glutamin
Fenilalanin	Glicin
Treonin	Prolin
Triptofan	Serin
Valin	Tirozin

TABELA 2.1: Spisak esencijalnih i neesencijalnih aminokiselina

Ovaj redosled je unapred određen i zavisi od redosleda nukleotida u dezoksiribonukleinskoj kiselini (DNK), odnosno gena, koji predstavljaju trojke nukleotida. Svakoј takvoj trojci jedinstveno je pridružena po jedna aminokiselina na osnovu genetskog koda. Proces kojim se enkodirana informacija prevodi iz DNK u niz aminokiselina u proteinskom lancu, posredstvom glasničke (eng. *messenger*) ribonukleinske kiseline (RNK) i transportne (eng. *transfer*) RNK, naziva se genska ekspresija. Proces sinteze proteina predstavlja *centralnu dogmu molekularne biologije*, čiji se prikaz može videti na slici 2.2 [5].

2.1.1 Funkcije i osobine proteina

Pri istraživanju bioloških procesa neophodno je znati i dobro razumeti funkcije proteina, što 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:



SLIKA 2.2: Prikaz centralne dogme molekularne biologije [5]

- 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.

Postoji nekoliko karakteristika proteina:

- grade kompleksna jedinjenja sa različitim supstancama po principu strukturne komplementarnosti i
- poseduju visoku osetljivost na različite agense koji ih denaturišu ⁴. Neki od najčešćih agenasa su: visoka temperatura, pritisak, mehaničko tretiranje, dejstvo kiselina, baza, organskih rastvarača, materija, itd. [1, 5].

2.1.2 Struktura proteina

Struktura proteina zavisi od rasporeda aminokiselina i utiče na njegovu funkciju. Trodimenzionalna struktura, koja se smatra najstabilnijom, formira se presavijanjem polipeptidnog lanca na različite načine. 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. Proteinsku strukturu održavaju različite vrste kovalentnih i nekovalentnih interakcija između hemijskih jedinjenja, na primer: vodonične, jonske, elektrostatičke, dipolne, itd.. Nabiranjem i uvijanjem lanca kreiraju se različiti oblici proteina: vlaknasti, globularni ili eliptični. Ako mutacija

³Katalizacija predstavlja proces povećavanja brzina reakcija

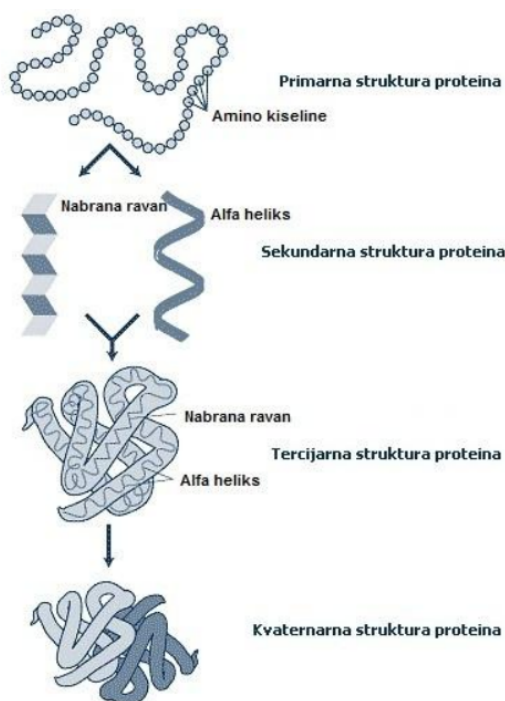
⁴Denaturacija proteina je proces koji izaziva promene u strukturi proteina, čime se menja i njihov fiziološki uticaj.

dovede do toga da aminokiselina sa malim bočnim lancem bude zamenjena aminokiselinom sa velikim, pojaviće se problem sa formiranjem 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, 3, 6].

U molekulima proteina postoji hijerarhijska strukturalna organizacija u četiri nivoa:

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

Na slici 2.3 se može videti opšti prikaz mogućih struktura proteina, a na drugoj 2.4 šematski prikaz [1].

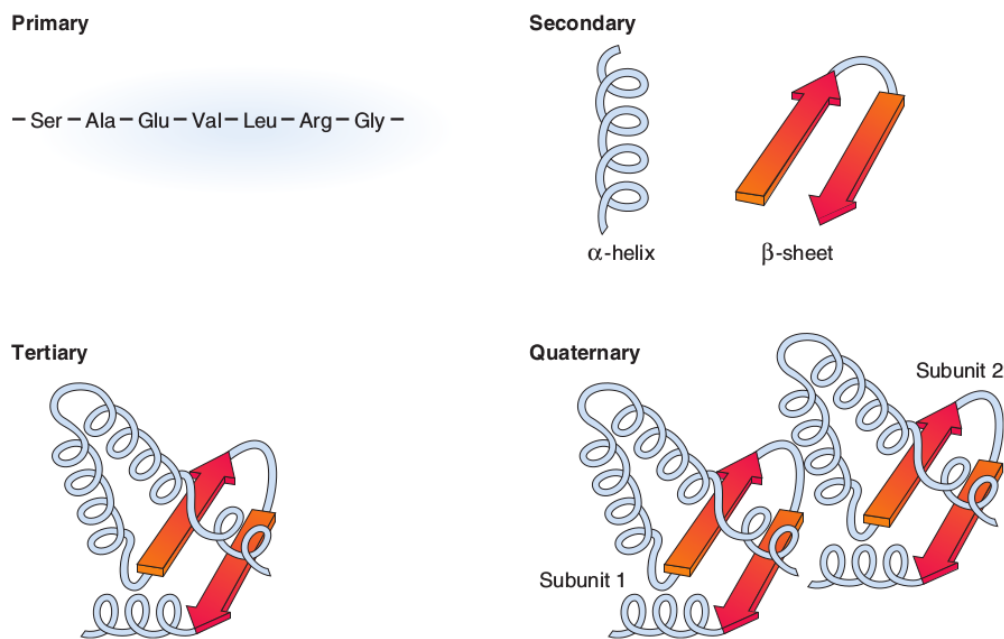


SLIKA 2.3: Prikaz struktura proteina

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 2.5 [1].

Mnoge genetske bolesti rezultuju u proteinima sa poremećenim redosledom aminokiselina, što uzrokuje nepravilno presavijanje i gubitak ili nemogućnost normalnog

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



SLIKA 2.4: Šematski prikaz struktura proteina [4]

funkcionisanja. Ukoliko su nam poznate strukture normalnih i mutiranih proteina, te informacije možemo iskoristiti za dijagnostikovanje ili proučavanje bolesti. Promene u primarnoj strukturi mogu imati uticaja i na više nivoje proteinskih struktura. Takve promene često dovode do lošeg presavijanja proteina i mogu dovesti do njegovog gubitka funkcije [7, 8].

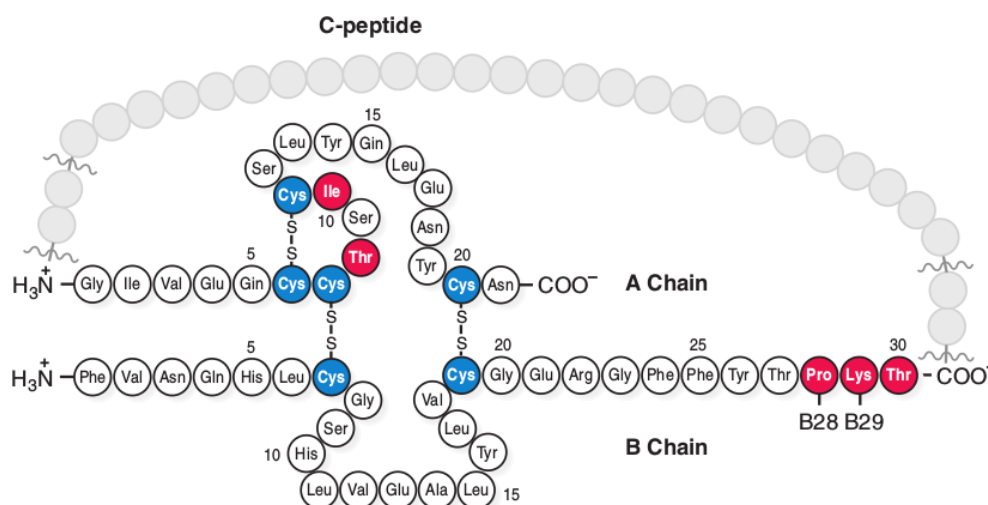
Sekundarna struktura Odnosi se na oblik koji protein zauzima u prostoru i označava pravilno pojavljivanje ponavljano prostornog rasporeda primarne strukture, u jednoj dimenziji. Ovu strukturu čini nekoliko različitih oblika, od kojih su najčešći α -heliks i β -presavijena traka (ili β -struktura), a čest je i tzv. β -okret [1, 6].

α -heliks - tip sekundarne strukture kod kog se gusto pakovani polipeptidni lanac spiralno uvrće. Karakteriše se brojem peptidnih jedinica po okretu i rastojanjem između dva okreta. Spada pod energetski najsiromašniji, a time, i najstabilnije strukture proteina. Heliks mogu obrazovati i *L*- i *D*- aminokiseline, pa postoje i dva tipa heliksa: levostrani i desnostrani. Prikaz izgleda α -heliksa se vidi na slici 2.6 [1].

β -struktura - Za razliku od α -heliksa, sastoji se od dva ili više peptidnih lanaca, ili segmenata polipeptidnih lanaca, a obrazuje se kada se ovakvi tipovi lanca povežu uzdužno. Postoje dva tipa β -struktura: paralelna i antiparalelna. Prikaz izgleda β -strukture se vidi na slici 2.7 [1].

β -okreti - obrću pravac polipeptidnog lanca praveći kompaktan globularan oblik [8]. Prikaz izgleda sekundarnih struktura se nalazi na slici 2.8.

Tercijarna struktura Podrazumeva unutarmolekularno slaganje polipeptidnog lanca u kompaktnu trodimenzionalnu strukturu specifičnog oblika, koja nastaje prostornim organizovanjem polipeptidnog lanca, koji već poseduje sekundarnu strukturu. Na taj način se približavaju ostaci aminokiselina koji su udaljeni u primarnoj strukturi. Proteini koji imaju ovakvu strukturu su globularni i kompaktni sa velikom gustinom u središtu [1, 6].



SLIKA 2.5: Prikaz primarne strukture [4]

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 [1, 6].

2.1.3 Savijanje proteina

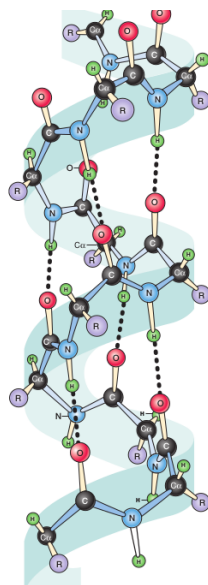
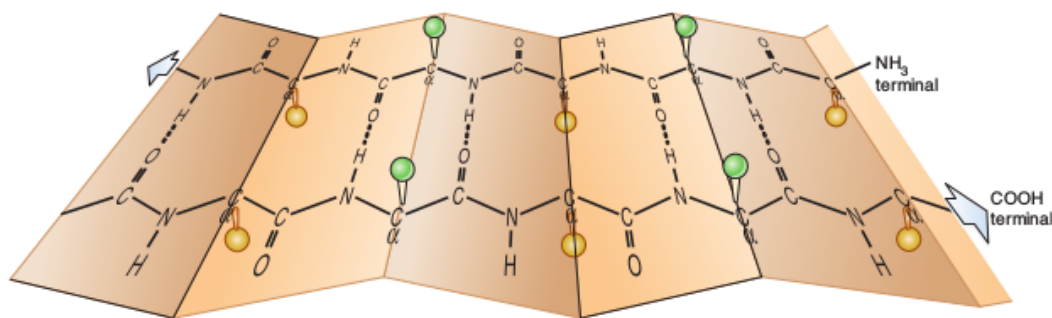
Interakcije između lanaca aminokiselina koji se nalaze sa strane, određuju kako se dugački polipeptidni lanac presavija u trodimenzionalni oblik funkcionalnog proteina. Presavijanje proteina koje se događa u ćeliji traje od nekoliko sekundi do nekoliko minuta. Na slici 2.9 se može videti opšti prikaz savijanja proteina.

2.1.4 Denaturacija proteina

Denaturisanje proteina rezultuje u odvijanju i dezorganizaciji proteinske sekundarne i tercijarne strukture. Pod idealnim uslovima, denaturisanje proteina može biti *reverzibilno*. To znači da bi se protein, pri prestanku delovanja agenasa, vratio u normalno stanje. Međutim, većina proteina ostaje trajno neuređena. O neuređenosti proteina biće više reči u nastavku.

Jedno od objašnjenja zašto se protein ne vraća u originalno stanje se sastoji u tome da protein počinje sa savijanjem pre nego što se izvrši sinteza celog lanca. Osim toga, specijalizovana grupa pomoćnih proteina (engl. *chaperones*) je neophodna za pravilno savijanje mnogih vrsta proteina. Ovi pomoćni proteini interaguju sa polipeptidima u nekoliko faza tokom procesa savijanja, imaju ulogu u tome da održavaju protein nesavijenim dok sinteza nije gotova, ili imaju ulogu katalizatora. Loše savijanje proteina može dovesti do različitih bolesti kao što su: amiloidna bolest ili Prionova bolest [8].

⁶Tada takve proteine nazivamo *oligomerima*

SLIKA 2.6: Prikaz α -heliksa [4]SLIKA 2.7: Prikaz β -strukture [4]

2.2 Neuređenost proteina

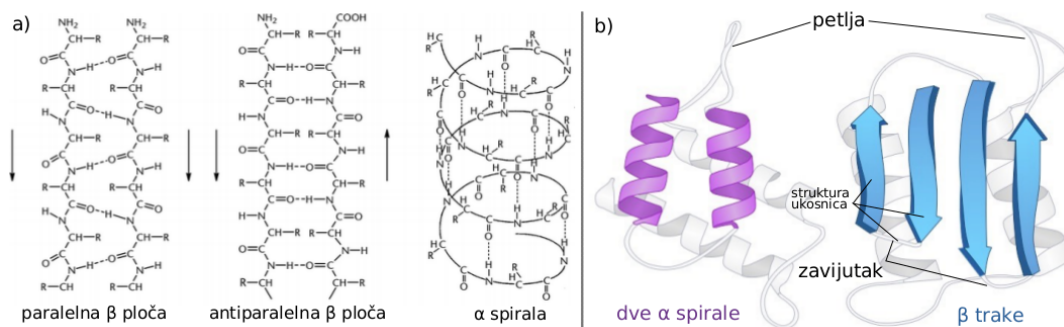
Eksperimentalnim utvrđivanjem sekundarne strukture proteina (koje će biti detaljnije opisano u 2.2.1) uočeno je da se neretko, pod određenim fiziološkim uslovima, javljaju proteini sa trodimenzionalnom strukturom koja nije dobro definisana. Neuređenost predstavlja inherentno⁷ svojstvo sekvence. Neuređen može biti ceo protein, a mogu biti neuređeni određeni regioni proteina različitih dužina. Kao posledica, ovakve proteine nazivamo inherentno neuređenim proteinima, skraćeno IDP⁸, a ako su u pitanju neuređeni, ali funkcionalni, regioni, onda je skraćenica IDPr⁹. Strukturalni poremećaji su česti kod viših eukariota. Kod ljudi, čak trećina svih proteina ima neuređenu strukturu. Neuređeni proteini su uključeni u procese stvaranja mnogih bolesti poput raka, neurodegenerativnih i kardiovaskulatnih bolesti, dijabetesa, brojnih neuronskih oboljenja i drugih. Statističkom analizom došlo se do zaključka da se aminokiseline mogu klasterovati na dve grupe:

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

⁷Inherentno = nasleđeno

⁸eng. *Intrinsically Disordered Proteins*

⁹eng. *Intrinsically Disordered Protein Regions*



SLIKA 2.8: Prikaz sekundarnih struktura [9]

Neuređene proteine ili neuređene regione je teško kategorizovati, a jedan od opštih opisa strukture dat je kao kombinacija više tipova foldona¹⁰:

- foldon (eng. *foldon*) je nezavisno organizujuća jedinica(region) proteina,
- induktivni 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,
- 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:

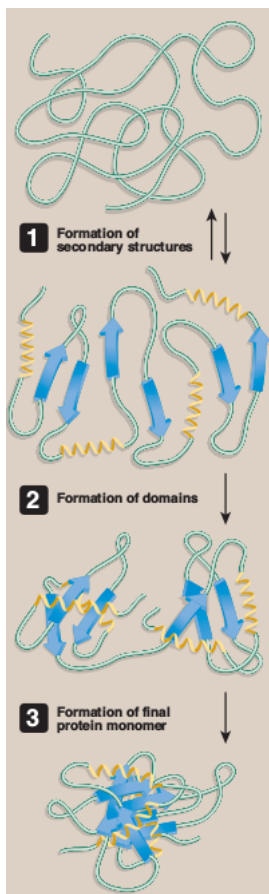
1. uređen protein,
2. topljiva globula (eng. *molten globule*),
3. pre-topljiva globula (eng. *pre-molten globule*) i
4. nasumično klupko (eng. *random coil*).

Neuređenost proteina se utvrđuje eksperimentalno, laboratorijskim analizama, ili uz pomoć prediktora za automatsko utvrđivanje neuređenosti [5, 10, 11, 12, 13, 14].

2.2.1 Eksperimentalno ispitivanje neuređenosti proteina

Eksperimentalno utvrđivanje neuređenosti proteina podrazumeva laboratorijsko utvrđivanje neuređenosti korišćenjem raznih biofizičkih i biohemijskih tehnika i njihovih kombinacija. Ono spada u veoma skupe i spore metode koje ne mogu da odgovore na izazove akademije i industrije. Uprkos tome, razvijen je veliki broj metoda za karakterizaciju strukture i osobina proteina. Svaka eksperimentalna metoda karakteriše

¹⁰Foldon ostaje u originalnom nazivu, kao posledica manjka literature. [9]



SLIKA 2.9: Prikaz savijanja proteina [8]

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 [10, 5]:

- 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.

2.2.2 Računarsko ispitivanje neuređenosti proteina

Kao posledica osobina eksperimentalnog ispitivanja neuređenosti, veliki napor su uloženi u razvoj prediktora za računarsko utvrđivanje neuređenosti proteina. Ovi prediktori uz pomoć računara, korišćenjem tehnike 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 3

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 istraživanje 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: prirodno/suštinski neuređeni, nesavijeni, denaturisani ili reomorfni proteini (eng. *intrinsically disordered/ unfolded/ unstructured*), u ovom radu, biće korišćen samo kraći termin - neuređeni proteini. Neuređeni proteini bitnu ulogu u određivanju ćelijskog odgovora na spoljašnje uticaje, transkripciju i translaciju, kao i savijanje i odvijanje ćelijskih makromolekula. Kao što je navedeno u prethodnom poglavlju, neuređenost proteina se može, osim eksperimentalnog, određivati i računarski. Upravo o tom vidu određivanja, odnosno, predikcije, neuređenosti proteina govori ovo poglavlje. Najpre, biće detaljnije opisan računarski postupak. Nakon toga, biće priče o prediktorima, od kojih će pojedini biti detaljnije objašnjeni. Na kraju, ukratko, će biti predstavljena baza podataka DisProt i njen značaj u ovom radu. [15, 5, 16]

Značaj pronalaska neuređenih proteina/regiona leži, pre svega, u tome što uočavanjem ovakvih regiona poboljšavamo analizu proteina i time izbegavamo poravnavanje uređenih i neuređenih proteinskih regiona čime se povećava preciznost analize sličnosti sekvenci. Još jedan bitan razlog je ušteda vremena pri upotrebi eksperimentalnih tehnika, jer dolazi do velikih gubitaka vremena na utvrđivanje strukture proteina koji je nema. [17, 18]

3.1 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, one bazirane na [19]:

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

Svaki prediktor koristi različite koncepte, fizičko-hemijske karakteristike ili različite algoritme mašinskog učenja. Međutim, ni ove metode nisu najpouzdanije. 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. Nepouzdanost podataka se odnosi na [20]

3.1.1 SPOT-D

SPOT-Disorder Predictor je razvijen da ima visoku efikasnost u predviđanju i kratkih i dugih neuređenih regiona bez odvojenog treninga, bez obrzira na činjenicu da neuređeni regioni različitih veličina imaju različite sastave aminokiselina. SPOT-D je metod koji je nastao unapređivanjem metoda koji koristi tradicionalne neuralne mreže bazirane na prozorima nad svim testiranim skupovima bez odvajanja trening skupa na kratkim i dugim regionima. Utvrđeno je da je SPOT-D jednako ili više precizan 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]

3.1.2 PONDR

PONDR prediktor vrši predikciju nad pojedinačnim sekvencama korišćenjem neuronskih mreža sa propagacijom unapred (eng. feedforward neural networks) koje koriste sekvence atributa nad prozorima od 9 do 21 aminokiseline. Uzima se prosek nad ovim prozorima, a potom se te vrednosti koriste pri treniranju neuronskih mreža tokom konstrukcije prediktora. Iste vrednosti se koriste za ulaze da bi se napravila predikcija. Prediktori neuronskih mreža se treniraju nad neponavljajućim skupovima uređenih i neuređenih sekvenci, a izlazi su brojevi između 0 i 1, koji se odvajaju na prozore od po 9 aminokiseline. Ako vrednost regiona prevazilazi prag od 0.5 smatra se da je region neuređen.

3.1.3 s2D

3.1.4 IUPred

IUPred vrši previđanje neuređenosti proteina sa loše definisanom tercijarnom strukturom (eng. Intrinsically unstructured/disordered proteins - IUPs) na osnovu sekvenci aminokiselina procenjujući njihovu energiju prilikom interakcija. Metod se bazira na fizičkim osnovama uređene/neuređene prirode proteina. Naime, globularni proteini prave veliki broj interakcija, čime se obezbeđuje stabilizujuća energija koja nadoknađuje određene gubitke prilikom savijanja proteina. Nasuprot njima, neuređeni proteini imaju specijalne regione koji nemaju sposobnost kreiranja interakcija.

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. Pokazalo se da je suma energije prilikom interakcija može da se proceni matematički na osnovu sastava aminokiselina, uzimajući u obzir da doprinos aminokiselina 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. [22, 23, 24, 25]

3.1.5 ESpritz

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

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

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

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

ESpritz određuje verovatnoću poremećaja za svaki region. [26, 27, 28, 29]

3.1.6 SEG

3.1.7 Disopred2

DISOPRED2 je treniran na skupu od 750 neponavljajućih sekvenci struktura dobijenih na osnovu X-zraka. Neuređenost se prepoznaje kod onih delova koji se pojavljuju u podacima o sekvenci, ali imaju koordinate koje nedostaju na elektronskoj mapi gustine. Ovaj način ima svoje nesavršenosti koje se ogledaju u nesavršenosti metode kristalografije X-zracima kod koje se mogu javiti nedostaci. Iako nesavršen, 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 linearnim potpornim vektorima (eng. SVM - support vector machine). [30]

3.2 Baza podataka DisProt

Glava 4

Aplikacija

4.1 Arhitektura

4.2 Funkcionalnosti

4.3 Korišćenje aplikacije

4.4 Primer upotrebe

Glava 5

Implementacija

Glava 6

Zaključak

Bibliografija

- [1] Vesna Spasojević-Kalimanovska Slavica Spasić Zorana Jelić-Ivanović. *Opšta biohemija*. 2002.
- [2] 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.
- [3] Gerhard Michal Dietmar Schomburg. *Biochemical Pathways: An Atlas of Biochemistry and Molecular Biology*. Hoboken, New Jersey: John Wiley & Sons, Inc., 2012. ISBN: 9780470146842.
- [4] 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.
- [5] Jovana Kovačević. *Strukturna predikcija funkcije proteina i odnos funkcionalnih kategorija proteina i njihove neuredenosti*. on-line na: <http://www.math.rs/files/DoktoratJK2015.pdf>. 2015.
- [6] Ivana Čepelak Dubravka Čvorišćec. *Štrausova medicinska biokemija*. Medicinska naklada, 2009.
- [7] 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.
- [8] 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.
- [9] Goran Vinterhalter. *Bioinformatička analiza povezanosti funkcije i neuredenosti proteina*. on-line na: http://www.racunarstvo.matf.bg.ac.rs/MasterRadovi/2017_08_23_Goran_Vinterhalter/rad.pdf. 2018.
- [10] 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.
- [11] 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.
- [12] 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.
- [13] C.J.; Dunker A.K Uversky V.N.; Oldfield. *Intrinsically disordered proteins in human diseases: Introducing the D2 concept*. 2008.
- [14] 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](https://doi.org/10.1186/1472-6807-11-29).

- [15] V.N. Uversky. *Intrinsically disordered proteins from A to Z*. 2011.
- [16] Han K.H. Tompa P. *Intrinsically disordered proteins*. 2012.
- [17] S.; Canard B.; Karlin D. Ferron F.; Longhi. *A practical overview of protein disorder prediction methods*. 2006.
- [18] J.; Cheng J. Deng X.; Eickholt. *A comprehensive overview of computational protein disorder prediction methods*. 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](https://doi.org/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](https://doi.org/10.1093/bioinformatics/btw678).
- [22] 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.
- [23] et al. Garbuzynskiy S.O. *To be folded or to be unfolded?* 2004.
- [24] K.A. Thomas P.D. i Dill. *An iterative method for extracting energy-like quantities from protein structures*. 1996.
- [25] et al. Dosztányi Z. *The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins*. 2005.
- [26] 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](https://doi.org/10.1093/bioinformatics/btr682).
- [27] S.; Frasconi P.; Soda G.; Pollastri G. Baldi P.; Brunak. *Exploiting the past and the future in protein secondary structure prediction*. 1999.
- [28] 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.
- [29] J.L. Sussman O. Noivirt-Brik J. Prilusky. *Assessment of disorder predictions in CASP8*. 2009.
- [30] McGuffin LJ Buxton BF Ward JJ Sodhi JS and Jones DT. *Prediction and functional analysis of native disorder in proteins from the three kingdoms of life*. 2004.