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
dr Nina Radojičić

Beograd, 2018

Sažetak

...

Zahvalnica

Sadržaj

Slike

Tabele

Olgi, tati i mami...

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

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, a zatim će posebno biti opisani neuređeni proteini, njihova uloga i uzroci koji mogu dovesti do njihove pojave.

2.1 Proteini

Proteini (grč. *protos* - "zauzimam prvo mesto") su biološki makromolekuli, koji čine 70% suve materije ćelija i neophodni su za njihovu izgradnju i pravilno funkcionisanje. Osim uloge u izgradnji ćelija, učestvuju u mnogobrojnim procesima koji se odvijaju unutar organizma. Predstavljaju najvažniji sastojak žive materije i utiču na brojnost i raznolikost živih bića. Specifičnost proteina je tolika da svaka biljna i životinjska vrsta ima svoje proteine, dok se, kod viših organizama, razlikovanje može uočiti i na individualnom nivou. Broj proteina u živim bićima je ogroman, na primer *E.coli* sa 3000 i čoveka sa 5 miliona proteina [1, 6].

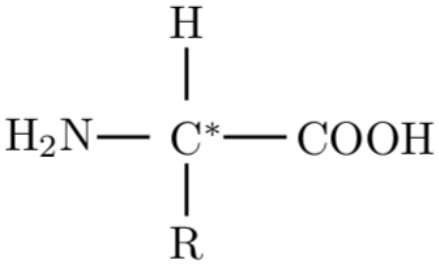
Proteini su jedinjenja sačinjena od 100 ili više aminokiselina. Na osnovu broja aminokiselina koji ih čine peptidi se dele na:

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

Proteini se mogu posmatrati i kao nizovi nadovezanih polipeptida. Proteini i peptidi su izgrađeni od 22 aminokiseline [1] Sve proteinske aminokiseline su α-aminokiseline. Njih karakteriše to da su primarna amino i karboksilna grupa vezne za α-ugljenikov atom. Aminokiseline se međusobno razlikuju po strukturi bočnog R-ostatka, koji utiče na strukturu proteina. Opšta strukturna formula aminokiselina može se videti na slici 2.1 [6]. Proteinske aminokiseline (osim glicina) imaju asimetričan α-ugnjenikov atom i shodno tome mogu da se jave u dva oblika (prema Fišerovoj konvenciji) L i D. Sve standardne aminokiseline imaju L-konfiguraciju. Grafički prikaz može se videti na slici 2.2 [6]. 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

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

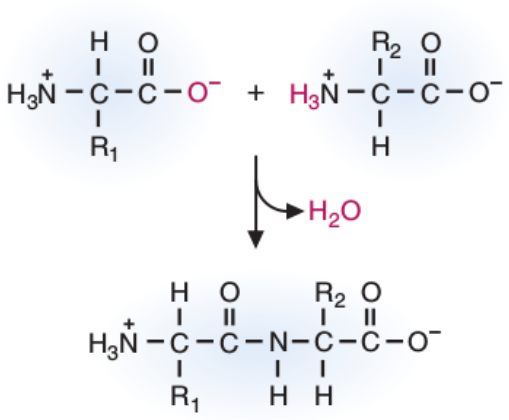
[2] $L-aminokiseline$ su one aminokiseline sa levom prostornom konfiguracijom, analogno, postoje i $D-aminokiseline$, sa desnom



Slika 2.1: Opšta strukturna formula aminokiselina [6].

Slika 2.2: Prikaz L i D prostorne konfiguracije [6].

(grafički prikaz se može videti na slici 2.3). Ovim postupkom nastaje nerazgranati polipeptidni lanac koji se sastoji od polipeptidne kičme i bočnih ostataka. Standardna grupa aminokiselina se može podeliti na esencijalne i neesencijalne, čiji spisak se može videti u tabeli 2.1 [2, 3].



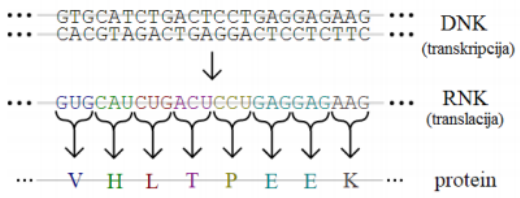
Slika 2.3: Prikaz spajanja α-karboksilne grupe jedne aminokiseline i α-amino grupe druge aminokiseline, pri čemu se oslobađa molekul vode [4].

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

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

predstavlja *centralnu dogmu molekularne biologije*, čiji se prikaz može videti na slici 2.4 [5].



Slika 2.4: Prikaz centralne dogme molekularne biologije [5]

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

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

[6]

2.1.1 Funkcije i osobine proteina

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

- *dinamičke*, od kojih su najvažnije:

1. transportna - prenos molekula (poput kiseonika, gvožđa, lipida) i hormona od mesta sinteze do mesta delovanja,
2. biološka - regulacija metaboličkih procesa u ćeliji, kontrola i regulacija transkripcije gena i translacija,
3. katalizatorska - biološka katalizacija [3],
4. zaštitna - keratin, koagulacija krvi,
5. održavanje zapremine tečnosti u organizmu,

- *strukturne*, od kojih su najvažnije:

1. obezbeđivanje čvrstine i elastičnosti organa,
2. davanje oblika organizmu,
3. izgradnja strukturnih elemenata ćelije i
4. bitna uloga u kontraktilnim i pokretnim elementima organizma.

Na slici 2.5 mogu se videti primeri nekih proteina i njihovih uloga.

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

Slika 2.5: Primeri proteina [6]

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

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

2.1.2 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

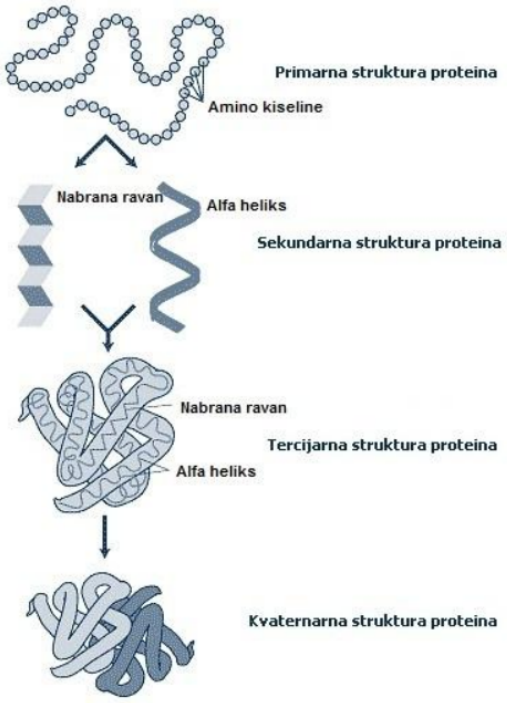
[3] Katalizacija predstavlja proces povećavanja brzina reakcija

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

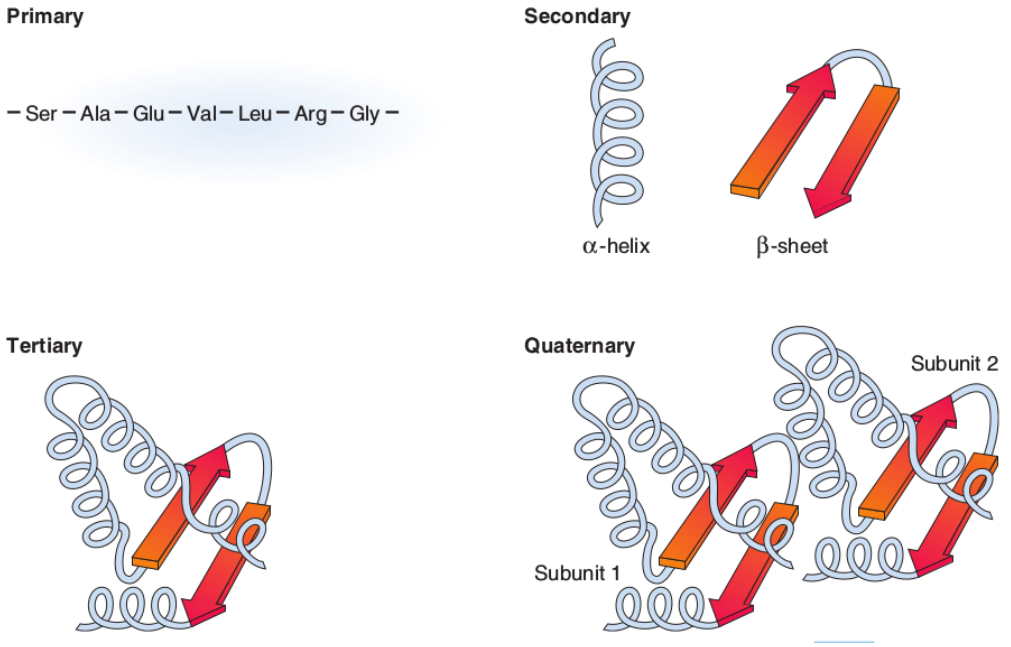
kompleksnu trodimenzionalnu strukturu, koja se smatra najstabilnijom. Struktura proteina zavisi od redosleda aminokiselina i utiče na njegovu funkciju. Unutrašnjost takve strukture ima visoku gustinu, pa polipeptidni lanac ne dopušta promene u sastavu i zahteva prisustvo aminokiselina tačno određene 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. [6, 3, 1]
Proteinsku strukturu održavaju različite vrste kovalentnih i nekovalentnih interakcija između hemijskih jedinjenja, na primer: vodonične, jonske, elektrostatičke, dipolne, itd.. Nabiranjem i uvijanjem lanaca kreiraju se različiti oblici proteina: vlaknasti, globularni ili eliptični. Strukturni proteini su vlaknasti, dok su oni koji pokazuju određenu aktivnost globularni. Ako mutacija dovede do toga da aminokiselina sa malim bočnim lancem bude zamenjena aminokiselinom sa velikim, pojaviće se problem u formiranju trodimenzionalne strukture. Ako bi se, pak, velika aminokiselina zamenila sa malom, pojavio bi se prazan prostor, što bi moglo dovesti do destabilizacije molekula proteina [6, 1, 3, 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 2.6 se može videti opšti prikaz mogućih struktura proteina, a na drugoj slici 2.7 šematski prikaz [1]. Određivanje sastava proteina u vidu aminokiselina je relativno



Slika 2.6: Prikaz struktura proteina



Slika 2.7: Šematski prikaz struktura proteina [4]

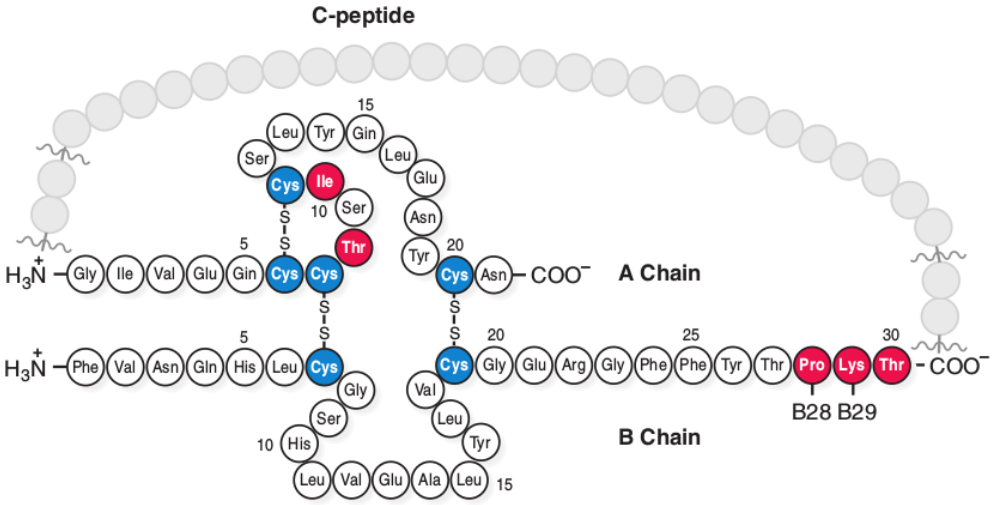
jednostavno, dok je određivanje odnosa između sastava aminokiselina i strukture proteina komplikovano. Uprkos tome, često se mogu izvući korisni zaključci o strukturi proteina na osnovu aminokiselinskog sastava. [6]

Primarna struktura Predstavlja sâmu sekvencu aminokiselina[5] koje učestvuju u izgradnji proteina. Ova struktura ima ključni značaj za određivanje funkcije proteina zbog interakcija koje se javljaju između bočnih lanaca aminokiselina, a koji utiču na trodimenzionalnu strukturu. Proteini koji poseduju sličnu sekvencu aminokiselina nazivaju se *homologi*, a poređenje sekvenci među takvim proteinima može ukazati na genetsku relaciju između različitih vrsta.Prikaz izgleda primarne strukture na primeru insulina kod čoveka se vidi na slici 2.8 [1].

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

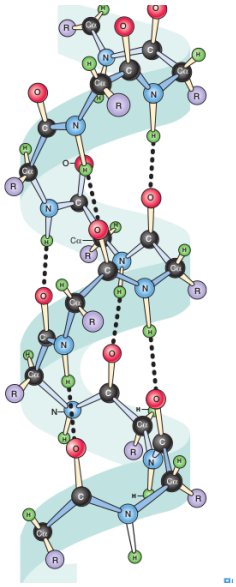
Sekundarna struktura Odnosi se na oblik koji protein zauzima u prostoru i označava pravilno pojavljivanje ponavljanog prostornog rasporeda primarne strukture, u jednoj dimenziji. Ovu strukturu čini nekoliko različitih oblika, od kojih su najčešći α-heliks i β-presavijena traka (ili β-struktura), a čest je i tzv. β-okret [1, 7]. **α-heliks** - tip sekundarne strukture kod kog se gusto pakovani polipeptidni lanac spiralno uvrće. Karakteriše se brojem peptidnih jedinica po okretu i rastojanjem između dva okreta. Predstavlja najrasprostranjeniju sekundarnu strukturu i energetski

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



Slika 2.8: Prikaz primarne strukture [4]

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 2.9 [1, 6].



Slika 2.9: Prikaz α-heliksa [4]

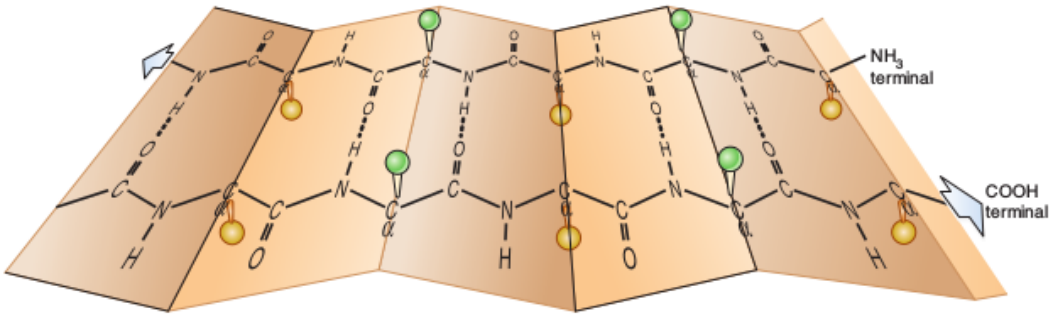
β-struktura - Za razliku od α-heliksa, sastoji se od dva ili više peptidnih lanaca, ili segmenata polipeptidnih lanaca, a obrazuje se kada se ovakvi tipovi lanca povežu uzdužno vodoničnim vezama. Razlika između polipeptidnog niza u β-strukturi i potpuno istegnutog polipeptidnog niza je u tome što je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

- paralelna - vodonično su vezani susedni polipeptidni nizovi istih smerova i

- antiparalelna - vodonično su vezani susedni polipeptdni nizovi suprotnih smerova.

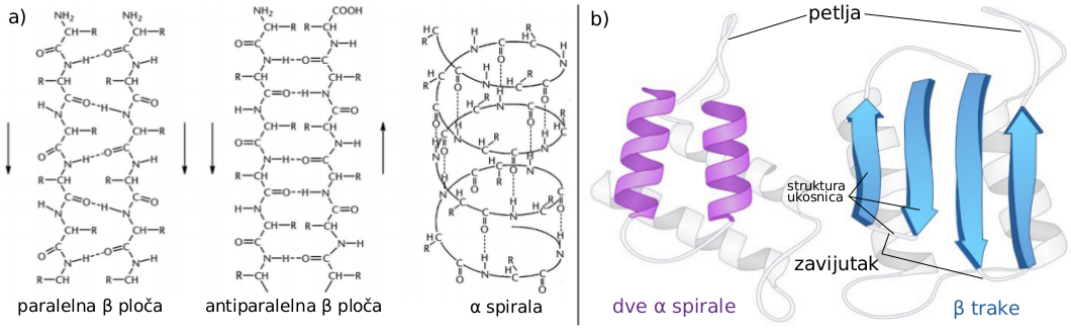
Moguće su i mešovite paralelne-antiparalelne strukture. β-strukture se često javljaju u proteinima, a u globularnim se podjednako često javljaju i paralelne i antiparalelne. Sekundarna struktura se eksperimentalno utvrđuje na osnovu kristalne strukture proteina. Prikaz izgleda β-strukture se vidi na slici 2.10 [1, 6].

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



Slika 2.10: Prikaz β-strukture [4]

Prikaz izgleda sekundarnih struktura se nalazi na slici 2.11.

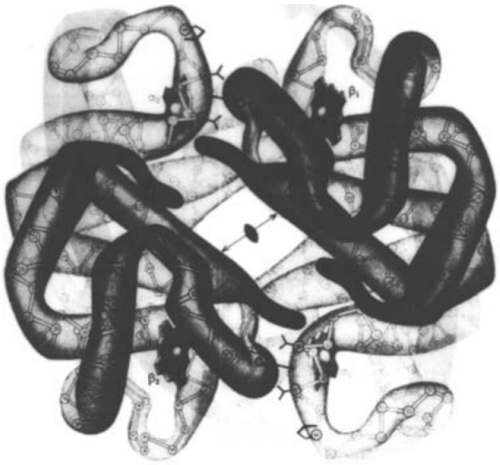


Slika 2.11: Prikaz sekundarnih struktura [10]

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 (globule), koja nastaje prostornim organizovanjem polipeptidnog lanca, sa sekundarnom strukturom. Na taj način se približavaju ostaci aminokiselina koji su udaljeni u primarnoj strukturi. Tercijarna struktura predstavlja način organizacije, odnosno rasporeda, sekundarnih struktura i položaj bočnih ostataka aminokiselina. Proteini ove strukture su globularni i kompaktni sa velikom gustinom u središtu. Poznavanje ove strukture proteina predstavlja osnovu za izučavanje funkcije i aktivnosti proteina. Kako bi se eksperimentalno utvrdila ova struktura vrši se rendgenska strukturna analiza. [1, 7, 6].

Kvaternarna struktura Predstavlja agregaciju više peptidnih lanaca u molekulu proteina. Mnogi proteini, posebno oni velike mase, izgrađeni su od nekoliko polipeptidnih lanaca. Svaka takva komponenta naziva se *podjedinica* ili *protomer*.

Oni mogu biti identični[6] ili se razlikovati prema strukturi. Ovakav raspored dovodi do brzog i efikasnog transfera supstrata od jednog aktivnog centra enzima do drugog. Prikaz proteina sa kvatenarnom strukturom može se videti na 2.12 [1, 7]. Postoji



Slika 2.12: Prikaz hemoglobina, predstavnika globularnih proteina sa kvatenarnom strukturom [6]

nekoliko razloga iz kojih se kvaternarna struktura javila:

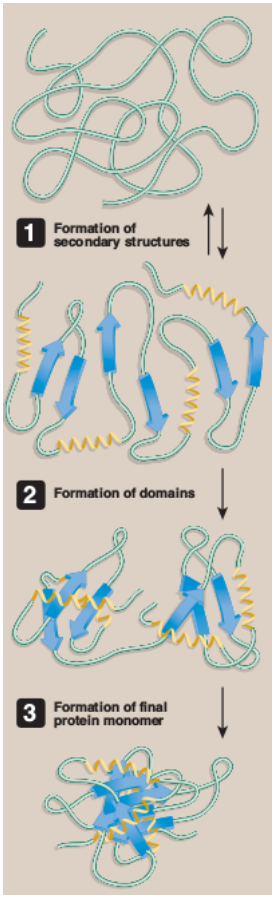
- Kompleksnija uloga zahteva kompleksniju strukturu
- Veća efikasnost katalitičkih procesa - katalizacija niza reakcija u metaboličkom procesu vrši se enzimima spojenih u multienzimske komplekse
- Viši nivo može da utiče na niži nivo strukture, pa tako kvaternarna struktura može da utiče na tercijarnu 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.* [6]

2.1.3 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 2.13 se može videti opšti prikaz savijanja proteina.

[6]Tada takve proteine nazivamo *oligomerima*



Slika 2.13: Prikaz savijanja proteina [9]

2.1.4 Denaturacija proteina

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

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 inherentno[7] svojstvo sekvence. Neuređen može biti ceo protein, a mogu biti neuređeni određeni regioni proteina različitih dužina. Kao posledica,

[7]Inherentno = nasleđeno

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

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

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

- foldon (eng. *foldon*) je nezavisno organizujuća jedinica(region) proteina,
- indukativni foldon (eng. *inducible foldon*) je neuređeni region proteina koji savijanje lanca postiže barem delom vezivajući se za partnera,
- ne-foldon (eng. *non-foldon*) je neuređeni region proteina koji nikada ne postiže uređenost,
- 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, 11, 12, 13, 14, 15].

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

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

se raznim prednostima manama i nivoom pouzdanosti, zbog čega je najbolje kombinovati dobijene rezultate. Naredne eksperimentalne, biofizičke i biohemijske, tehnike su najčešće u ispitivanju neuređenosti proteina [11, 5]:

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

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

2.2.2 Računarsko ispitivanje neuređenosti proteina

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

Glava 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. [16, 5, 17]

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. [18, 19]

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 [20]:

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 [21]

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. [22]

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 aminokiselina. 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. [23, 24, 25, 26]

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

ESpritz određuje verovatnoću poremećaja za svaki region. [27, 28, 29, 30]

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). [31]

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 uređenostidelova 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 neuređenosti*. on-line na: `http://www.math.rs/files/DoktoratJK2015.pdf`. 2015.

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

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

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

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

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

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

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

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

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

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

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

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

[18] S.; Canard B.; Karlin D. Ferron F.; Longhi. *A practical overview of protein disorder prediction methods.* 2006.

[19] J.; Cheng J. Deng X.; Eickholt. *A comprehensive overview of computational protein disorder prediction methods.* 2012.

[20] 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.

[21] 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.

[22] 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.

[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] et al. Garbuzynskiy S.O. *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] 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.