UNIVERZITET U BEOGRADU
MATEMATIČKI FAKULTET

MASTER RAD

Računarska analiza povezanosti funkcije i neuređenosti proteina

Autor:
Goran VINTERHALTER

Mentor:
dr. Jovana KOVAČEVIĆ

ČLANOVI KOMSIJE:

prof. dr Gordana Pavlović-Lažetić
prof. dr Saša Malkov
doc. dr Jovana Kovačević

Beograd, 2018

UNIVERZITET U BEOGRADU

Sažetak

Matematički fakultet
Katedra za Računarstvo i informatiku

Master informatičar

Računarska analiza povezanosti funkcije i neuređenosti proteina

Goran VINTERHALTER

The Thesis Abstract is written here (and usually kept to just this page). The page is kept centered vertically so can expand into the blank space above the title too...

Sadržaj

Glava 1

Osnovni pojmovi

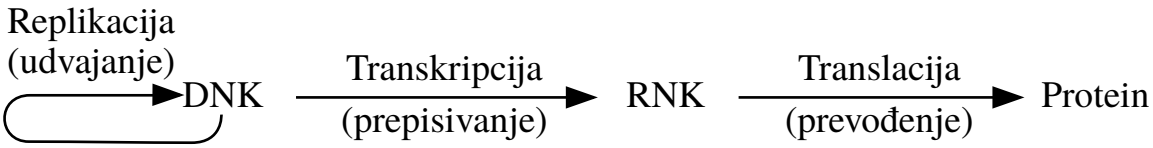
Svi živi organizmi sastoje se od jedne ili više ćelija, a svaka ćelija od molekula. Veliki[1] molekuli (makromolekuli) organskog porekla obično[2] su sačinjeni od ponavljajućih strukturnih jedinica **monomera** *(mono- = jedan, mer- = deo)* međusobno povezanih **kovalentnim** vezama. Takav molekul zovemo **polimer** *(poli- =mnogo, -mer= deo)*. Skup monomera možemo da smatramo azbukom koja gradi jezik polimera. Mali broj monomera je dovoljan za strukturnu kompleksnost bilo koje ćelije. Tri najznačajnija tipa bioloških polimera i njihovi monomeri prikazani su u Tabeli 1.1.

TABELA 1.1: Najznačajniji biološki polimeri

Polimer	**Monomer**
Ugljeni hidrati	Monosaharid (šećeri)
Nukleinska kiselina (DNK)	Nukleotid
Protein	Aminokiselina

1.1 Centralna dogma molekularne biologije

Centralna dogma molekularne biologije prikazana Slikom 1.1 objašnjava protok informacija kroz generacije i ćeliju.

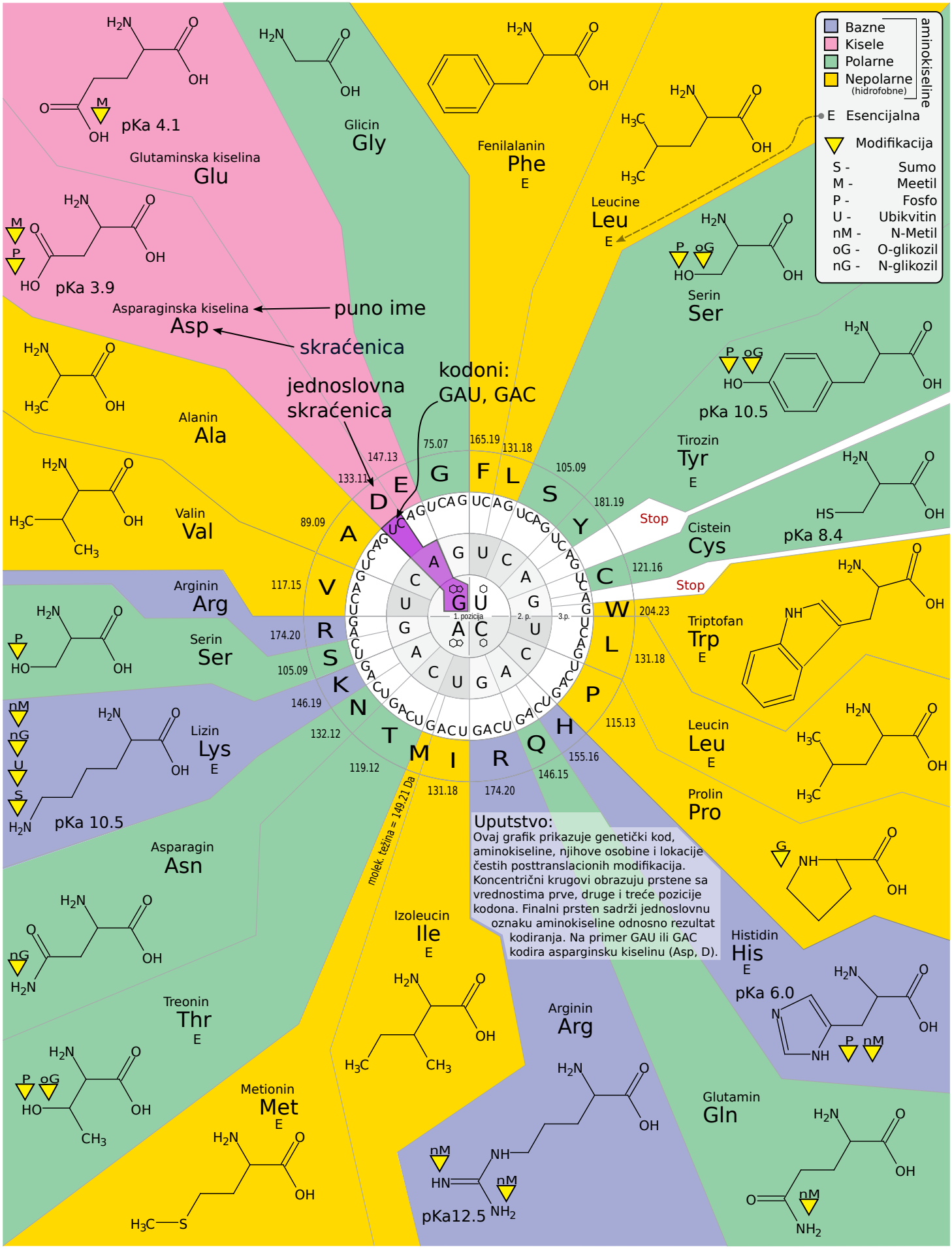


SLIKA 1.1: Centralna dogma molekularne biologije

Dezoksiribonukleinska kiselina, kraće **DNK** se procesom replikacije (tokom deobe ćelije) udvaja u dve kopije. Za života ćelije regioni DNK molekula tzv. **geni** bivaju transkribovani (prepisani) u oblik ribonukleinske kiseline, kraće **RNK**. Glasnička RNK, kraće **mRNK** dobijena transkripcijom tzv. **kodirajućeg gena** sadrži kodirane informacije za sintezu proteina i biva transportovana do molekulskih mašina ribozoma. Ribozom dekodira poruku mRNK odnosno translira (prevodi) mRNK u protein. Pomenuti kod prikazan centralnim delom Slike 1.2 zove se **genetički kod** i opisuje mapiranje uzastopnih trojki baza nukleinske kiseline tzv. **kodona** u aminokiseline ili stop oznaku. Na primer trojka baza: Guanin-Adenin-Uracil (GAU) ili Guanin-Adenin-Citozin (GAC) prevode su u Asparginsku kiselinu.

[1] Obično se molekulska masa od $1000Da$ (Daltona) uzima kao granica između malih molekula i makromolekula.

[2] Lipidi recimo nisu polimeri, ali su principijalno slični



SLIKA 1.2: Genetički kod, aminokiseline, osobine AK i lokacije PTM
(Originalna Slika pripada vikimedija javnom domenu, autor Bromomir)

Genetički kod smatra se univerzalnim za sva živa bića (poznat kao kanonski). Ipak sitne razlike u tumačenju nekih kodona javljaju se kod malog broja vrsta, obično nekih bakterija, mitohondrije, arhea ali i viših biljaka.

Gen je region DNK sekvence koji biva transkribovana u RNK. Neki geni su tzv. **nekodirajući geni** i transkribuju se u funkcionlane RNK molekule dok gore pomenuti **kodirajući geni** služe sintezi proteina. Proces prepisivanja gena u funkcionalnu jedinicu još se naziva **ekspresija gena**, a rezultat **genski produkt**. Kompletan DNK kod nekog organizma predstavlja njegov **genom**, dok svi transkribovani RNK molekuli čine **transkriptom**, a svi sintetisani proteini **proteom**. Tri velike oblasti bioinformatike koje prikupljaju i analiziraju ove podatke su: **genomika**, **transkriptomika** i **proteomika**. U bioinformatici gene i genske produkte najjednostavnije predstavljamo sekvencama njihovih respektivnih monomera.

Centralnu dogmu predložio je Francis Krik 1958. godine. Ipak vremenom razni specijalni slučajevi toka informacija su otkriveni, prvenstveno kod virusa i bakterija. U ovom radu nećemo ih razmatrati.

1.2 Homologija

Dve sekvence su **homologe** ako dele zajedničkog evolutivnog pretka (originalna sekvenca). Skup homologih proteina čine **proteinsku familiju**. Homologi proteini skoro uvek dele sličan trodimenzionalan oblik. Međutim, isto se ne može reći za sličnost samih sekvenci jer sadržaj sekvence brže divergira od oblika koji kodira. Razlikujemo dva tipa homologa:

- Dve sekvence su **ortologe** (orto - pravi) ako predstavljaju istu sekvencu (isti gen, protein) u različitim vrstama nastalu specijalizacijom. Na primer geni mioglobina kod čoveka i kod pacova su ortolozi. Ortolozi uvek obavljaju istu funkciju u ćeliji.
- Dve sekvence su **paraloge** ako su nastale duplikacijom originalne sekvence. U slučaju duplikacije, jedna sekvenca je slobodna da se promeni i služi različitoj funkciji. Na primer ljudski alfa globin i beta globin su paralozi.

Pošto je tačan trodimenzionalan oblik retko poznat, pronalaženje homologa se oslanja na sličnost sekvenci. Postoje razne metode za poređenje (poravnavanje) sekvenci od kojih se za pronalaženje homologa najčešće koriste PSI-BLAST (ili noviji, senzitivniji DELTA-BLAST). PSI-BLAST (engl. *position-specific iterated BLAST, ψ-BLAST*) prvo gradi PSSM matricu originalne sekvence (od bliskih, sličnih sekvenci) koju dalje koristi u pretrazi udaljenih proteinskih sekvenci. PSSM (engl. *Position-Specific Scoring Matrix*) je $20 \times n$ matrica koja predstavlja verovatnoću pojavljivanja aminokiselina za svaku od n pozicija sekvence. PSSM se generiše iz poravnanja nekoliko sličnih sekvenci. Rezultat pretrage može se agregirati u novu PSSM koja predstavlja evolutivni profil i opisuje familiju proteina.

1.3 Proteini

Proteini (belančevine) su najčešći biološki makromolekuli koji čine i do 80% suve mase organizma. Strukturno, protein je linearan polimer sačinjen od lanca **aminokiselina** (monomeri) skraćeno AK.

1.3.1 Aminokiseline

Aminokiseline koje translacijom mRNK ulaze u sastav proteina poznate su kao **proteinogene**[3]. Ukupno su poznate 22 proteinogene aminokiseline od kojih je 20 kodirano kanonskim genetičkim kodom tzv. **standardne** AK, dok se 21. AK (selenocistein) i 22. AK (pirolizin) prevode specijalnim tumačenjem STOP kodona i to samo kod nekih organizama[4].

Proteinogene aminokiseline imaju šablon strukturu predstavljenu Slikom 1.3 a). Centralni alfa ugljenikov atom (C_α) povezan je **amino grupom** ($-NH_2$), **karboksilnom grupom** ($-COOH$), atomom vodonika i tzv. **R grupom**. Vrsta aminokiseline određena je R grupom još poznatom kao **bočni niz** ili **bočni ostatak** (engl. *residue*).

Reakcijom kondenzacije prikazane na Slici 1.3b dve aminokiseline grade kovalentnu tzv. peptidnu vezu rezultujući peptidom (dipeptid na slici). Dakle, peptid je polimer aminokiselina koje su međusobno povezane peptidnim vezama. Peptid duži od 10 AK (Slika 1.3c) se smatra polipeptidom (skraćeno pp). Neki autori pod terminom protein podrazumevaju samo polipeptide duže od 50 AK, ali mi nećemo slediti taj primer. Ponavljajući elementi $N - C_\alpha - C$ $(-N - C_\alpha - C)^*$ $-N - C_\alpha - C$ čine tzv. **pp lanac** ili **kičmu peptida** sa koje štrče bočni nizovi. Polipeptidi se zapisuju u smeru u kome su sintetisani pa zato početak (levi kraj) nazivamo N-terminus (zbog amino grupe), a desni kraj C-terminus (zbog karboksilne grupe).

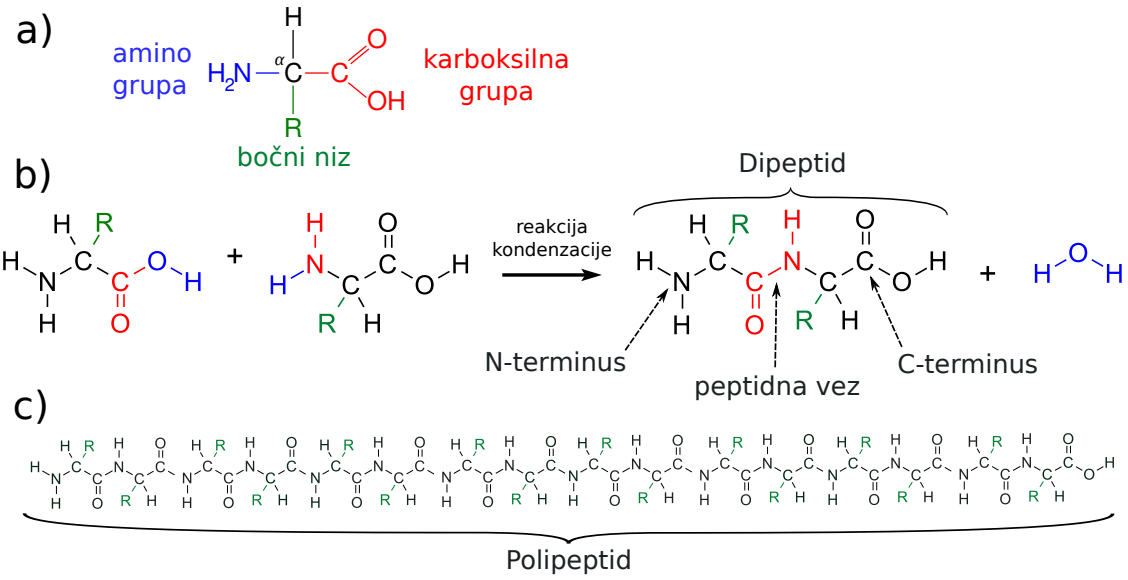
Prema fizičkim i hemijskim osobinama bočnog niza aminokiseline se mogu klasifikovati na nekoliko načina prikazanih Slikama 1.2 i 1.4 od čega izdvajamo sledeću klasifikaciju:

- Nepolarne AK - zbog manjka asimetrije u naelektrisanju R grupe ovi molekuli nisu rastvorljivi u vodi (koja je polaran molekul). Hidrofobni su (ne vole vodu). Obično se ove aminokiseline nalaze u unutrašnjosti savijenog proteina gde je kontakt sa vodom mimimalan.
- Nenaelektrisane polarne AK - su rastvorljive u vodi (vole vodu). Uglavnom se nalaze na spoljnim delovima proteina, često na hemijski aktivnim delovima.
- Naelektrisane polarne AK - su jako hidrofilne. Dodatno se dele na pozitivno i negativno naelektrisane.
- Aromatične AK - su najveće i najtežim jer u bočnom nizu sadrže aromatični ugljenikov prsten.

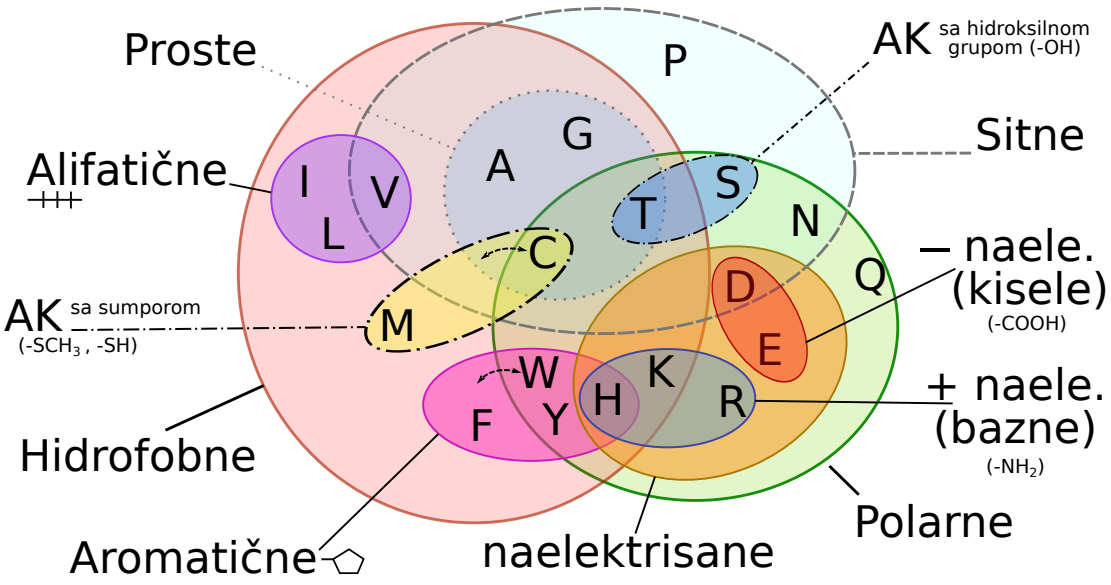
Za vreme života proteina (nakon translacije) aminokiseline koje ga čine mogu biti modifikovane od strane raznih enzima na primer pravljenjem kovalentnih veza sa novim funkcionalnim grupama. Ovaj proces poznat je kao **posttranslaciona modifikacija**, kraće **PTM**. Na Slici 1.2 prikazane su najčešće PTM i gde se javljaju.

[3] U prirodi se javljaju stotine različitih AK, ali one ne ulaze u sastav proteina

[4] selenocistein se javlja u svim domenima života (arhea, bakterija i eukariota), ali ne i kod svih vrsta organizama, dok se pirolizin javlja samo kod određenih bakterija i arhea



SLIKA 1.3: a) Šematski prikaz aminokiseline b) spajanje aminokiselina reakcijom kondenzacije c) Šematski prikaz polipeptida



SLIKA 1.4: Veneov dijagram osobina bočnih nizova aminokiselina

1.3.2 Struktura proteina

Protein je sačinjen iz jednog (monomerni) ili više (oligomerni) polipeptdinih lanaca. Proteinska struktura (oblik) opisuje se kroz četiri nivoa rastuće složenosti.

Primarna struktura opisana je redosledom peptidnih veza kičme peptida, odnosno redosledom aminokiselina. Primarna struktura predstavlja **sekvencu proteina** i kompaktno je zapisujemo azbukom od minimum 20 karaktera.

Sekundardna struktura najčešće nastaje formiranjem vodoničnih veza između atoma kiseonika i azota istog lanca peptida stvarajući na taj način dva različita oblika koja su prikazana na Slici 1.5:

- α**-spirala** - je spiralna struktura u kojoj R grupe štrče spolja
- β**-traka** - predstavlja spoj dva ili više (anti)paralelnih delova polipeptidnog lanca.

Delovi pp lanca koji povezuju navedene strukture (Slika 1.5b) često nemaju uređenje i po dužini ih delimo na kraći zaokretaj i duže petlje (engl. *short turn, long loop*). Specijalno zavijutak (engl. β*-hairpin*) je kratak zaokret lanca kod antiparalelne β-traka. Formiranje sekundarne strukture naziva se **lokalno savijanje**.

Usled delenja elektrona, peptidna veza ponaša se slično dvostrukoj kovalentnoj vezi rezultujući onemogućavanjem rotacije ograničavanjem mogućih konformacija pp lanca. Posledica ovog ponašanja omogućava kreiranje pravilnih sekundarnih struktura dodatnim ograničavanjem vrednosti tzv. dihedralnih uglova ψ i ϕ prikazanih na Slici 1.6a. Dijagram svih vrednosti ψ i ϕ uglova jednog polipeptida otkriva klasterovanje vrednosti prikazano tzv. Ramačandrovim dijagramom Slika 1.6b.

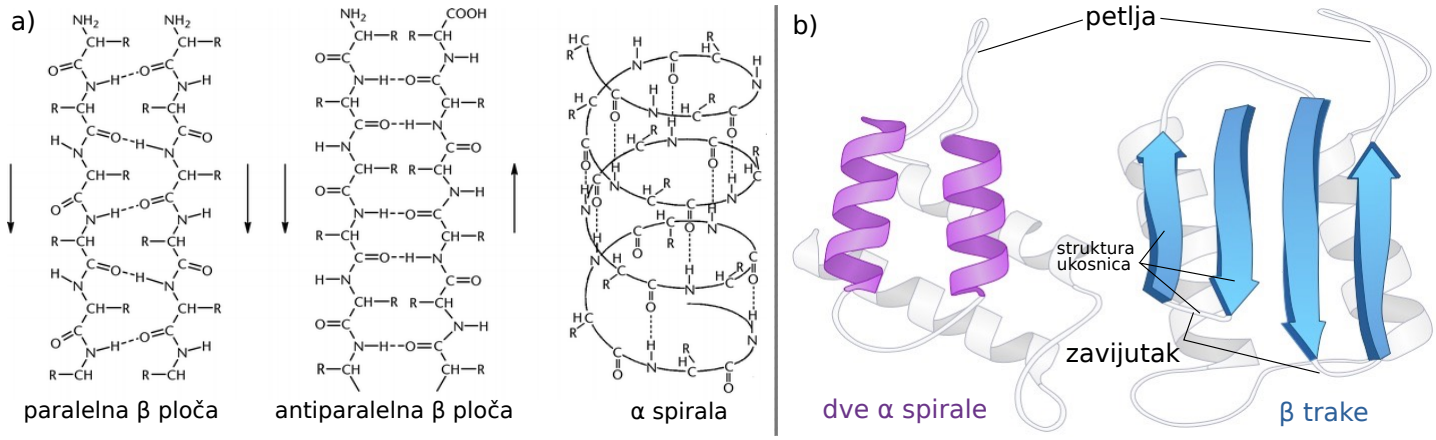
Pored gore navedenih, strukture: disulfidni most[5], zink prsti i ukosnica takođe se smatraju elementima sekundarne strukture. Postoje i drugi oblici spirala i traka, ali ih nećemo navoditi.

Tercijarna struktura predstavlja prostorni oblik koji protein zauzima indukovan interakcijama bočnih nizova aminokiselina. Ovaj korak poznat je kao **savijanje** (engl. *folding*). Pod poznavanjem tercijarne strukture misli se na poznavanje prostornih koordinata svih atoma polipeptida. Primer dijagrama tercijarne strukture dat je Slikom 1.5b. Najveći uticaj na formiranje sekundarne strukture ima polarnost aminokiselina. Tercijarna struktura proteina obično je podeljena u jedan ili više **domena** koji predstavljaju rigidne savijene regione pp lanca. Domeni su obično sačinjeni od nekoliko elemenata sekundarne strukture.

Kvaternerna struktura opisuje proteinske komplekse (oligomerne proteine) sastavljene iz nekoliko savijenih polipeptida. Na primer, hemoglobin je proteinski kompleks.

Savijanje proteina u funkcionalan oblik (oblik koji indukuje biološku funkciju) odnosno njegova tercijarna ili Kvaternerna struktura direktno zavise od primarne strukture tj. sekvence. Dovoljna je zamena jedne aminokiseline u sekvenci pa da protein izgubi sekundarnu, a time i tercijarnu strukturu gubeći mogućnost izvršavanja biološke funkcije. Familija proteina obično je okarakterisana sličnošću domena, pa još kažemo da su domeni **konzervirani** ili očuvani. Konzervirani elementi sekundarne strukture domena predstavljaju **strukturne motife**, kraće **motif** i okarakterisani su velikom sličnošću primarne strukture. Promena okruženja (ph, temperature) dovodi do promene tercijarne strukture ili potpunog gubitka strukture (denaturacija).

[5]Neki autori disulfidni most zbog kovalentne veze smatraju elementom primarne strukture



SLIKA 1.5: Alfa spirala i β-traka.
Levi deo slike preuzet je iz rada *Proteins: fundamental chemical properties. Cozzone, A. J. Encyclopedia of Life Sciences (Nature Publishing Group, London, 2001).* Desni deo slike preuzet je sa veb strane bioninja.com

1.3.3 Enzimi

Život na ćelijskom nivou tj. održavanje balansa (homeostaze) zahteva hemijske reakcije koje se pri fiziološkim uslovima[6] ne izvršavaju dovoljno brzo ili ne vrše. **Katalizator** je molekul koji ubrzava hemijsku reakciju bez da sam bude promenjen. **Enzim** je katalizator biološkog porekla, a najčešće je protein. Molekul koji biva **katalizovan** odnosno promenjen u interakciji sa enizmom zovemo **substrat**. Mesto enzima koje interreaguje sa substratom naziva se **aktivni region**. Skoro svaka reakcija u živom organizmu ubrzana je enzimskom katalizom. Ime enzima obično se završava na '-za' Karakteristika enzima je da katalizuje reakciju specifičnog molekula.

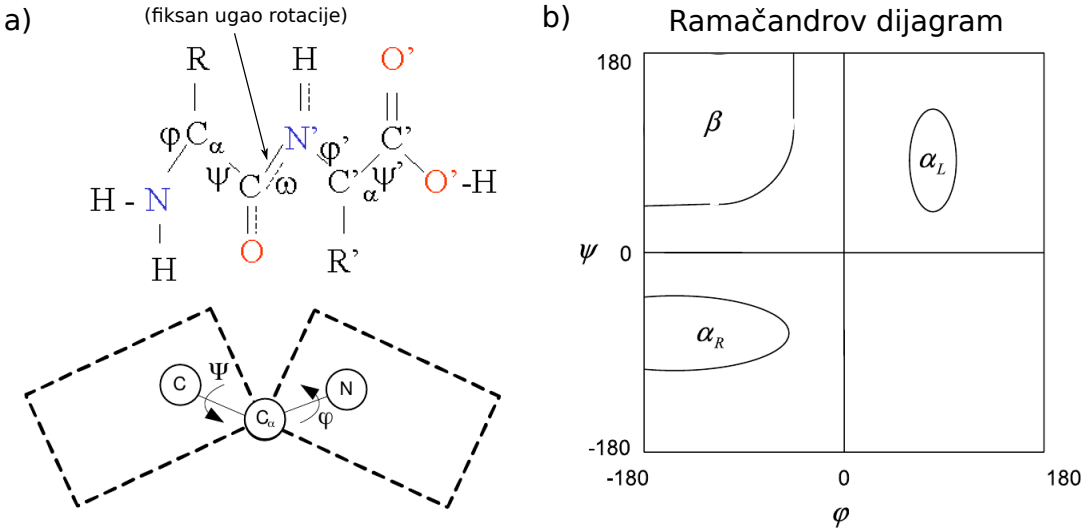
1.3.4 Funkcija proteina

Udžbenici iz biologije često nabrajaju uloge koje protein ima: formiranje strukture, zaštitna uloga (tzv. antigeni), transport, katalizacija hemijskih reakcija (enzimi), regulacija procesa u ćeliji itd. Postoji nekoliko sistema (nomenkulatura) za razdvajanje proteina po funkciji. Na primer, specijalno za enzime postoji numerička klasifikacija po hemijskoj reakciji koju katalizuju. Pomenuta klasifikacija klasteruje enzime u hijearhiju podskupova. Ovaj pristup ne radi dobro u opštem slučaju jer skupovi i podskupovi nisu adekvatni za reprezentaciju ponašanja svih proteina.

Ontologija predstavlja podobniju strukturu koja je oblika acikličkog usmerenog grafa. Ontologije su nastale u cilju opisivanja sveta, odnosno stvari koje ga čine, njihove međusobne relacije i predstavlja oblast filozofije. Gore navedena klasifikacija enzima može se predstaviti kao specijalan slučaj ontologije tj. stablo sa jednim tipom roditeljske veze (*is_a* veza).

Protein može biti sagledan iz nekoliko uglova. Na pirmer: kom procesu pripada, gde se taj proces odvija u ćeliji i koja je molekulska funkcija koju obavlja tokom izvršavanja pomenutog procesa. Sve tri stavke daju uvid u funkciju proteina i mogu biti predstavljene ontologijama. U ovom radu od interesa je isključivo molekulska funkcija proteina. Detaljan opis molekulske funkcije i ontologije koja je predstavlja dat je u Poglavlju 3.1 i Potpoglavlju 3.1.3.

[6] Pod fiziološkim uslovima podrazumevaju se normalni uslovi u živoj ćeliji (ph, temperatura, ...)



SLIKA 1.6: a) Prikazuje, ψ, ϕ i ω uglove. Ugao ω je fiksan zbog specifičnosti peptidne veze omogućavajući rotiacije samo oko ψ i ϕ uglova. b) Ramačandrov dijagram, prikazuje najčešću raspodelu vrednosti uglova i sekundarne strukture koje oni obrazuju.

Glava 2

Inherentno neuređeni proteini

Funkcionalni proteini sa delimičnim ili potpunim izostankom strukture (pri fiziološkim uslovima) nalaze se svuda u živom svetu, do te mere da ima više smisla upitati "gde se oni ne nalaze?" nego obratno [1]. Danas neuređenost proteina je uzrokovala nastanak velikog broj hipoteza, od D^2 koncepta bolesti [2] pa sve do evolucije višećelijskih organizama [3] i osobina prvih oblika života [4, 1]. Šta više, sa početkom 21. veka broj naučnih radova koji se bave ovom temom doživljava skoro eksponencijalan porast [5],ali da bi se razumela popularnost i perspektiva koju polje donosi neophodno je osvrnuti se na istoriju.

Fišerova[1] analogija o **bravi i ključu** ponovo otkrivena nezavisnim istraživanjima Hsien Wu, Mirski i Paulinga postavila je temelje "opšteprihvaćene" **struktura-funkcija** paradigme [6]. *"The characteristic specific properties of native*[2] *proteins we attribute to their uniquely defined configurations. The denatured protein molecule we consider to be characterized by the absence of a uniquely defined configuration"* [7]. Predloženi model prilagođen je funkcionisanju enzima, čija sposobnost da katalizuju zavisi od jasno definisanog geometrijskog oblika koji moraju da zauzmu odnosno u koji moraju da se saviju. Substrat (ključ ili funkcija) diktira oblik enzima (brave ili strukture) [8]. Kontrapozicijom sledi da nedostatak strukture vodi izostanku funkcije.

Prvi kontraprimer navedene teorije javio se još 1950. Protein krvne plazme, serum albumin pokazivao je veliku mogućnost vezivanja za različite partnere [6]. Ovo otkriće ukazivalo je da specifične zahteve enzima ne treba generalizovati na sve proteine. Ipak model brave i ključa i njena poboljšana varijanta, **teorija indukovanog fita**[3] (engl. *induced-fit theory*) dominirale su krajem prošlog veka, zanemarujuću konstantno rastući skup funkcionalnih "ne-nativnih" proteina čije postojanje nisu mogle da objasne. Sa druge strane tehnološki napretci u razlučivanju strukture proteina jasno su demonstrirali obimno postojanje funkcionalnih proteina bez uređene 3D strukture (pri fiziološkim uslovima) od kojih su neki bili neuređeni celom dužinom [6]. Nova paradigma je bila neophodan.

Hipoteza proteinskog trojstva [6] (nastala tek početkom 21. veka) predlaže da funkcija proteina može zavisiti od bilo kojeg od tri stanja ili tranzicije između tih stanja. Predložena stanja predstavljaju nativne oblike proteina i analogna su najčešćim stanjima materije na zemlji. Model je naknadno dopunjen još jednim stanjem:

1. **Uređen protein** - čvrsto stanje
2. **Topljiva globula** (engl. *molten globule*) - tečno stanje

[1] Emil Fišer bio je nemački hemičar koji je 1894. predložio analogiju brave i ključa opisujući karakteristike enzima pivske plesni [6].

[2] nativno stanje proteina je savijeno, operativno, funkcionalno stanje [6]. Ovaj termin bio je isprepleten sa paradigmom struktura-funkcija

[3] Teorija indukovanog fita omekšava rigidnost modela brava-ključ sugerišući da interakcija sa substratom indukuje konačni oblik enzima maksimizujući reakciju [8]

3. **Pre-topljiva globula** (engl. *pre-molten globule*) - međustanje
   Usled nejasne tranzicije između stanja topljivog globula i nasumičnog klupka (suprotno analogiji tečnog i gasovitog stanja) [6] model je dopunjen.

4. **Nasumično klupko** (engl. *random coil*) - gasovito stanje

Povezanost sekvence sa strukturom sugeriše da je neuređenost enkodirano inherentno[4] svojstvo [6] stoga ove proteine nazivamo **Inherentno**[5] **Neuređeni Proteini** (engl. *Intrinsically Disorderd Proteins*) skraćeno **IDP**, a njihove neuređene ali funkcionalne regione **IDPr** [1]. U ovom radu pod neuređenošću proteina podrazumevaćemo inherentnu neuređenost osim ako to nije drugačije naglašeno[6].

Današnje procene zastupljenosti pronašle su da 19% aminokiselina kod eukariota, 6% kod bakterija i 4% kod arhea pripadaju IDPr [9]. Čak 50% proteina eukariota ima bar jedan IDPr duži ili jednak 30 uzastopnih AK [10] dok je za 6% do 17% predviđeno da su neuređeni celom dužinom [11]. Ovi podaci bude veliko interesovanje naučnika da istraže funkciju i ponašanje IDP i IDPr.

2.1 Osobine i uticaj na funkciju

Detaljno opisivanje osobina i posledica neuređenosti proteina prevazilazi obim rada zalazeći u biohemiju i biofiziku. Sa druge strane, broj novih saznanja raste veoma brzo. Recimo, u časopisu *Nature* objavljen je rad [12] koji kratko sumira najnovija saznanja koja fundamentalno menjaju poglede na mogućnost jakog vezivanja potpuno neuređenih proteina u dinamične komplekse. Iz tih razloga navodimo samo globalne osobine IDP i IDPr kao i osobine relevantne za naš rad.

- Neuređenost je inherentno svojstvo sekvence [6]. Pokazano je da nisko očekivanje indeksa hidropatije[7] zajedno sa visokim ukupnim nabojem predstavlja bitan preduslov koji sprečava savijanje proteina u fiziološkim uslovima [1]. Statističkom analizom otkriveno je klasterovanje aminokiselina u one koje promivišu uređenost C, W, I, Y, F, L, M, H i N (engl. *order promoting*) i one koje promovišu neuređenost P, E, S, Q i K (engl. *disorder promoting*). [5, 1]. Opisane osobine daju validnost primeni mašinskog učenja u predviđanju neuređenih regiona proteina [5].

- Post translacione modifikacije proteina (PTM) značajno utiču na kontrolu i proširenje funkcije pogotovo neuređenih delova proteina. Postoji značajno preklapanje gore pomenute klasifikacije aminokiselina sa skupom AK koje su često modifikovane [1]. Iako je PTM povezano sa neuređenošću i sugeriše veliki uticaj na funkciju proteina [1] kompleksnost ove teme prevazilazi obime ovog istraživanja.

- IDP i IDPr su po zastupljenosti AK prostije[8] sekvence u poređenju sa domenima savijenih proteina. Ipak, usled manje restrikcija (obaveznog savijanja) mogućnost interakcije sa više partnera je mnogo veća što moguće funkcije čini raznolikim [1].

[4] Inherentno ili prirođeno, nasleđeno

[5] U nedostatku adekvatne domaće reči koristimo najbliži sinonim reči (engl. *intrinsic*) tj. (engl. *inherent*), koja čuva suštinu originalnog značenja.

[6] tumačenje neuređenosti zavisi od konteksta i može da označava denaturisane ili na drugi način dobijene nefunkcionalne proteine

[7] mera hidrofobnosti

[8] Prostije u smislu da sadrže manje informacija, manji Šenonov indeks.
Šenonov indeks $I = -\sum_{i=1}^{n} p_i \cdot log_2 p_i$ je mera kvaliteta informacije poistovećena sa brojem bita potrebnih da je kodiraju

Pomenuta interakcija kod nekih neuređenih proteina vodi do njihovog potpunog ili parcijalnog savijanja dok neki i dalje ostaju neuređeni [1]. Primena izreke *manje je više* proizvela je brave koje otključava nekoliko ključeva i ključeve koji otključavaju nekoliko brava.

- IDP i IDPr teško je strukturno kategorizovati [6, 5] iako su rani pokušaji napravljeni u radu [6]. Najopšteniji opis strukture ovih proteina dat je kao **kombinacija različitih tipova foldona**[9] [1]:
  - **foldon** (engl. *foldon*) je nezavisno organizujuća jedinica (region) proteina.
  - **indukativni foldon** (engl. *inducible foldon*) je IDPr koji savijanje lanca proteina postiže barem delom vezivajući se za partnera.
  - **ne-foldon** (engl. *non-foldon*) je IDPr koji nikad ne postiže uređenost.
  - **polu-foldon** (engl. *semi-foldon*) je IDPr koji ostaje polovično neuređen i nakon vezivanja za partnera.
  - **anti-foldon** (engl. *unfoldons*) je region proteina koji iz uređenog prelazi u neuređeno stanje u cilju vršenja funkcije.
- Gore pomenut opšti prikaz strukture nastao je iz raznih opažanja interakcije, prvestveno vezivanja proteina za partnere. Detaljan opis i iscrpna lista ovih i drugih pojava može se naći u radovima [13, 1] kao i poglavljima 10, 12 i 14 knjige *Structure and Function of Intrinsically Disordered Proteins by Peter Tompa, Alan Fersht*.

2.2 Eksperimentalno ispitivanje neuređenosti

- **Kristalografija X zracima** (engl. *X-ray crystallography*)
- Spektrosokopija Nuklearnom Magnetnom Rezonancom (NMR) (engl. *NMR spectroscopy*)
- (engl. *Circular dichroism (CD) spectroscopy*)
- (engl. *Protease digestion*)
- (engl. *Stoke's radius determination*)

2.3 Predikcija neuređenosti

Do danas napravljeno je preko 60 prediktora inherentno neuređenih proteina [14]. Prediktor u kontekstu proteina je program koji je tehnikama **mašinskog učenja** (skraćeno ML) terniran da predviđa osobine proteina. U radu [15] hronološkim redosledom prikazane su karakteristike i dostupnost tridesetak popularnih prediktora neuređenosti.

Istorijski posmatrano razlikujemo tri epohe razvoja: [15]

- Prva generacija (1979[10]-2001) Prvi prediktori oslanjali su se na razne fizičko-hemijske osobine proteina uključujući i osobine aminokiselina.

[9] Zbog nove prirode termina i manjka prevedene literature autor je odlučio da usvoji naziv u originalu.
[10] Nakon 1979. godine drugi prediktor nastao je tek 1997. [15]

- Druga generacija (2002-2006)
  Ovaj period okarakterisan je korišćenjem relativno jednostavnih ML modela treniranih nad sekvencama AK ili njihovim evolutivnim profilima.
- Treća generacija (2007-)
  Prediktori današnjice koriste komplikovanije ML modele. Uglavnom se podrazumeva meta-prediktor koji kombinuju rezultate nekoliko običnih ML modela. Na primer kombinacija NN, SVM i K-najbližih suseda tehnikom glasanja.

Po arhitekturi prediktore delimo u četiri kategorije: [15]

1. scorring function based
2. ML metode
3. Meta-prediktori
4. Predikcije na osnovu strukture[11].

2.3.1 Evaluacija ML modela

TODO, samo osnovne formule za preciznost i druge mere...

2.3.2 PONDR familija prediktora i VSL2b

PONDR familija (engl. *Predictors of Natural Disordered Regions*) je grupa prediktora druge generacije zasnovanih na neuronskim mrežama, kraće NN . Neuronske mreže sa propagacijom unapred (engl. *feed forward NN*) sa veličinom prozora između 9 i 21 AK trenirane su na različitim trening skupovima proteinskih sekvenci. Finalni prediktor predstavlja kombinaciju nekoliko neuronskih mreža od kojih je svaka specijalizovana za regione određene dužine ili položaja. PONDR familija sadrži nekoliko prediktora koji se razlikuju po izboru trening skupova. Oznaka "VSL" kodira tipove i poreklo atributa proteinskih trening skupova.

- V - Opisuje eksperimentalnu tehniku kojom je neuređenost utvrđena na trening skupu (engl. *X-ray, NMR, circular dichroism*)
- S - Prediktor je treniran na skupu proteina sa **kratkim** neuređenim regionima ($<$ 30 AK)
- L - Prediktor je treniran na skupu proteina sa **dugim** neuređenim regionima ($>$ 30 AK)

CASP (engl. *Critical Assessment of protein Structure Prediction*) je takmičenje u predikciji strukture proteina (ili neurđenosti) gde se objektivno ocenjuje kvalitet razvijenih prediktora i počev od 1994 održava se svake dve godine. Tokom CASP7 takmičenja 2006. VSL2b je evaluiran je kao prediktor sa ukupnim najtačnijim predviđanjima neuređenosti [16]. Međutim, po današnjim merilima [14] VSL2b ipak se smatra zastarelim. Ali, kako je VSL2b nezavistan paket koji se lako može pokrenuti na kućnom računaru i projektovan je da bude brz (visoko propustan) ovo istraživanje temelji se upravo na njemu.

VSL2b kao ulaz prima proteinsku sekvencu[12] minimalne dužine 9 AK kodiranih jednim karakterom i podržava azbuku od 20 standardnih AK. Rezultat predikcije je niz

[11] podrazumeva predviđanje strukturnih elemenata proteina čije odsustvo predviđa neuređenost

[12] Postoje varijante prediktora koje kao ulaz primaju evolutivni profil, ali zbog dodatne složenosti koraka PSI-BLAST pretrage ovaju pristup nije korišćen.

ocena (verovatnoća) za svaku poziciju sekvence koje govore da li je pozicija uređena ili neuređena. Pozicija sa vrednostima iznad 0.5 smatra se neuređenim, a suprotno uređenim.

Glava 3

Baze podataka u bioinformatici

Automatizacija bioloških i hemijskih analiza početkom 21. veka omogućila je ubrzanu i paralelnu analizu velikog broja uzoraka. Ove tehnologije žargonski su poznate kao **tehnologije velike propusnosti** (engl. *high throughput technology*). Primera radi, tehnologije **sekvenciranja nove generacije** (engl. *Next-Generation Sequencing*) ili skraćeno **NGS** neprekidno napreduju spuštajući cenu čitanja genoma i eksponencijalno povećavajući količinu dostupnih sekvenci. Da bi se razumeo uticaj NGS tehnologije navodimo sledeći primer. Od sveže sekvencionisanih nepoznatih genoma predviđaju se potencijalni geni, a od gena potencijalne proteinske sekvence. Dobijene proteinske sekvence mogu se dalje klasterovati u familije, automatski anotirati, predviđati im se struktura, osobine itd. Zatim, moguće je vršiti analize za oktrivanje novih bioloških znanja. Povezanost između funkcije i neuređenosti proteina je jedan primer biološkog znanja. Ovaj primer ilustruje dve bitne stvari:

1. Eksponencijalni rast podataka uvodi bioinformatiku u oblast Big Data, posebno njene discipline poznate pod nazivom omike (na primer, genomike, proteomika itd.)

2. Velika povezanost između bioloških podataka.

Povezanost podataka preslikava se na baze podataka. Većina baza je usko specijalizovana za jedan tip informacije ili jedan organizam, ali zato sadrži reference ka drugim (spoljnim) bazama, naučnim radovima ili manje formalnim, ali informativnim resursima (veb strane, vikipedija, itd...). Specijalne baze podataka kao što je *UniProtKB* , pored primarnog sadržaja održavaju i veliki broj referenci ka drugim bazama podataka (tzv. dbxref (engl. *database cross reference*)) pokušavajući da međusobno povežu sve dostupne informacije. Konkrentno *UniProtKB* (feb. 2018) održava reference ka čak 164 različite baze podataka[1]. Dakle, bioinformatika kao disciplina podrazumeva da će analize biti vršene kombinacijom informacija nekoliko različitih baza. Zbog raznovrsnosti i svrhe prikupljenih informacija postoji veliki broj kategorija[2](vrsta) baza. Na adresi [17] autori Čen, Huang i Vu kategorizovali su i prikazali novije, javno dostupne i visoko kvalitetne proteinski orijentisane baze podataka (prikazana lista nije iscrpna) [18]. Za temu ovog rada od značaja su naredne tri kategorije:

- Baze sekvenci.
  Ove baze podataka sadrže sve poznate javno dostupne sekvence i kontrolišu dodeljivanje identifikacionog broja sekvence.
  - Proteinske sekvence: *UniProtKB*
  - DNK sekvence: EMBL-Bank, GenBank, DDBJ, ...

[1] www.uniprot.org/docs/dbxref

[2] Baze podataka ne pripadaju ekskluzivno samo jednoj kategoriji

- Baze strukture: DisProt, D2P2, MobiDB, PDB, ...
- Baze ontologija: Gene Ontology, Protein Ontology

3.1 Ontologije gena

Ontologija Gena (engl. *Gene Ontology*) ili skraćeno **GO**, predstavlja znanje o funkciji gena odnosno genskog produkta (protein, nekodirajuća RNK ili makromolekulski kompleks) [19]. GO baza sačinjena je iz dve komponente:

1. **Ontologije gena**.
2. **GO anotacije** tj. anotacije genskog produkta **GO terminom**. U našoj analizi anotacije su preuzete iz *Swiss-Prot* baze podataka[3].

Ontologija gena definiše skup termina, takozvanih **GO termina** (engl. *GO terms*) i njihove međusobne relacije. GO termini predstavljaju biološke termine (koncepte) koji opisuju funkciju. Ontologija gena sagledava funkciju genskog produkta iz tri aspekta koji se u terminologiji ontologije nazivaju imenski prostori (engl. *namespace*):

- **Molekulska funkcija (MF)** je biohemijska aktivnost (uključujući specifično vezivanje za ligande[4] ili strukture) genskog produkta.
- **Ćelijske komponente (CC)** se odnosi na mesto u ćeliji gde je genski produkt aktivan.
- **Biološki procesi (BP)** se odnose na procese kome genski produkt doprinosi.

Inspirisani sličnošću prva tri sekvencirana eukariotska organizma, GO projekat nastao je sa ciljem da objedini biologiju pod jedan univerzum termina za opis genskih proizvoda svih vrsta organizama [20].

3.1.1 GO termin

GO termin može biti zastareo i tada se relacijom **replaced_by** pokazuje na noviji termin. Relacija **consider** ukazuju na postojanje mogućih ekvivalentnih termina. Pored glavnog univerzuma termina postoje i podskupovi[5] termina tzv. *GO slim*. U donjem desnom delu Slike 3.1 prikazani su *GO slim* podskupovi.

GO termini takođe sadrže informacije kao sto su definicija, komentar, autor, datum nastanka, sinonimi itd. Pored ovih informacija takođe postoje reference ka drugim veb stranam i bazama podataka često vezanih uz definiciju. Uz GO termin obično se navode sinonimi koji odgovaraju imenu termina, ali se razlikuju po opsegu:

- *exact* - sinonim je ekvivalentan imenu termina
- *broad* - sinonim ima širi smisao od imena termina
- *narrow* - sinonima ima uži smisao od imena termina
- *related* - sinonim i ime termina su na neki način povezani

[3] Ali *Swiss-Prot* koristi anotacije iz ontologije gena

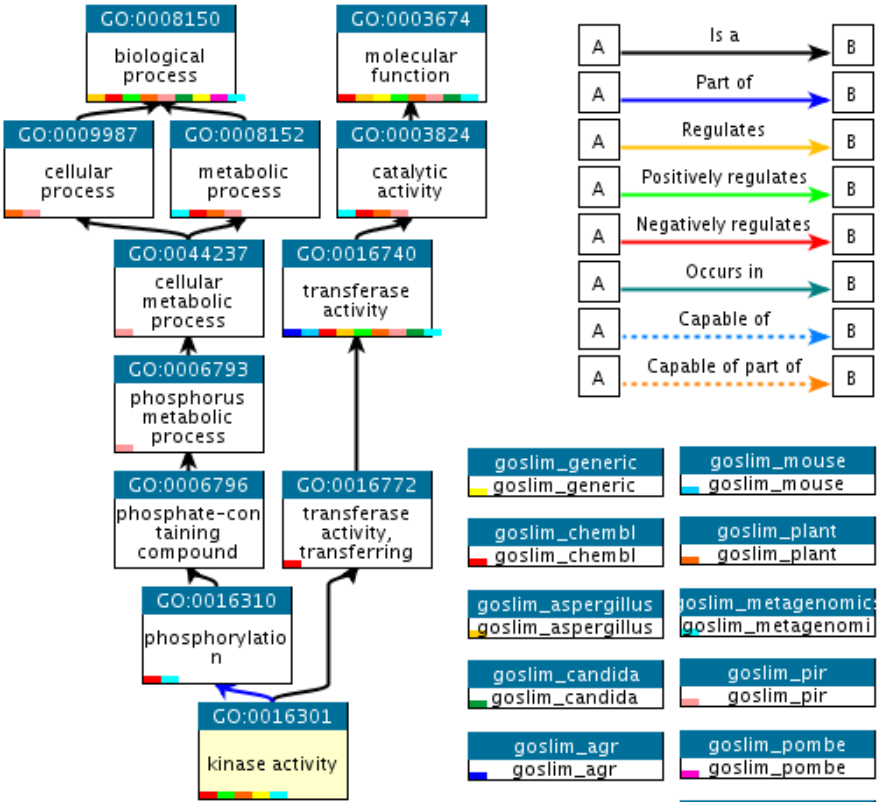
[4] U ovom kontekstu ligand je protein koji se vezuje za receptor u cilju izvršavanja biološke funkcije. Termin podrazumeva vezivanje (engl. *binding*)

[5] Uglavnom ovi podskupovi predstavljaju model organizme

3.1.2 GO relacije

Suštinu ontologije čine relacije između termina i pravila dedukcije koja se nad njima mogu primenjivati. Osnovnu strukturu ontologije čini usmereni aciklički graf (engl. *DAG*) obrazovan roditeljskom vezom (relacijom) **is_a**. Prateći ovu relaciju termini jednog imenskog prostora na primer MF neće nikad preći u druga dva CC i BP. Razlikujemo tri ontologija sa korenim čvorovima: MF, CC i BP [21]. Primer strukture (acikličkog usmerenog grafa) prikazan je na Slici 3.1. Pored relacije **is_a** postoje dodatne relacije od kojih su najčešće:

- **part_of** - je deo (ne znači da je uvek deo vezanog termina, relacija agregacije)
- **has_part** - sadrži (deo uvek postoji, relacija kompozicije)
- **regulates** - pozitivna ili negativna regulacija
- **positvely_regulates** - pozitivna regulacija (**is_a** termin koji reguliše)
- **negatively_regulates** - negativna regulacija (**is_a** termin koji reguliše)



SLIKA 3.1: Struktura ontologije
(preuzeto sa geneontology.org)

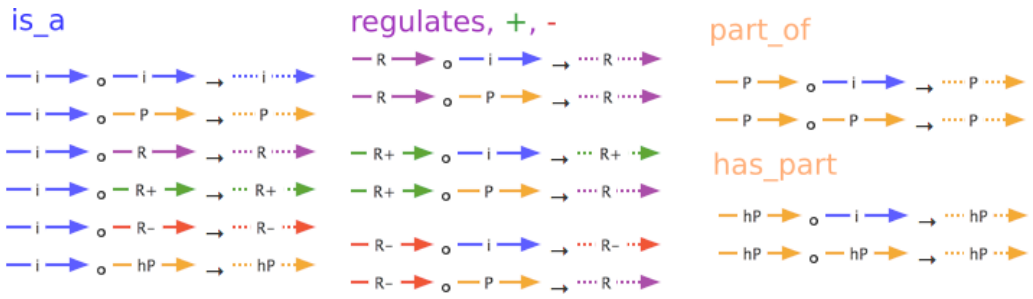
Vremenom se GO ontologija proširuje novim tipovima relacije koje su van okvira ovog rada. Svaka relacija ima strogo definisana pravila kompozicije koja omogućavaju automatsko rezonovanje. Na primer relacija **is_a** ima svojstvo tranzitivnosti [22] prikazano Slikom 3.2:

Siže pravila rezonovanja prikazan je na Slici 3.3.

Jedan od najčešće korišćenih formata je ravni `.obo` formata, a pored njega su u upotrebi `RDF/XML` i `OWL` formati. Poslednja dva formata namenjena su automatskom rezonovanju unutar specijalizovanih softvera i upitnih jezika (protégé, SPARQL, ...).

$$A\ \texttt{is\_a}\ B \ \wedge\ B\ \texttt{is\_A}\ C \implies A\ \texttt{is\_a}\ C$$
$$A\ \texttt{part\_of}\ B \ \wedge\ B\ \texttt{is\_A}\ C \implies A\ \texttt{part\_of}\ C$$

SLIKA 3.2: Tranzitivnost relacije **is_a**



SLIKA 3.3: Pravila rezonovanja (isprekidane relacije su rezultat)

3.1.3 Molekulska funkcija

Funkcija genskog produkta je posao koji obavlja ili "osobinu"koju ima. Razmotrimo narednu analogiju [23]. U kompaniji radnik (genski produkt) ima titulu (ime genskog produkta) i vrši poslove odnosno obavlja funkcije (molekulska funkcija) zarad izvršavanja nekog cilja tj. zadatka (biološki proces). Na primer, funkcije vozača bile bi: upravljanje volanom, stiskanje kvačila, utovarivanje prtljaga itd, ali ne bi bilo korektno reći da je funkcija vozača "vozačka funkcija", jer se zavisno od firme ili prevoznog sredstva menja skup radnji koje vozač obavlja. Vozač koji prevozi putnika na bicklu neće utovarivati prtljag niti stiskati kvačilo. Najbitnije karakteristike molekulske funkcije su:[23]

- MF nije specifična za jedan genski produkt već važi za sve organizme. Dakle, ne treba mešati MF sa imenom genskog produktom.
- MF nije biološki proces jer se BP sastoji od nekoliko MF.
- Granularnost staje na nivou molekula. MF ne opisuje rekacije na nivou atoma. Ako se reakcija može izvršavati na nekoliko načina, onda za svaki od njih postojaće poseban MF termin.

Postoji nekoliko standardnih definicija i šablon naziva koji se pripisuju genskom produktu *x* (ili nekom enzimu): [23]

- *x* **binding** - interreaguje selektivno i nekovalentno sa *x*
- **<enzyme> activity** - katalizuje reakciju (reakcija katalizovan od strane enzima)
- *x* **receptor activity** - vezuje se sa *x* zarad incijacije neke ćelijske aktivnosti
- *x* **transporter activity** - omogućava direktno pomeranje *x* u ćeliju, iz ćelije, unutar ćelije ili između ćelija

Osim korenskog MF čvora i *x* **binding** termina svi ostali MF termini sadrže sufiks *activity*. Ovo je uvedeno iz filozofskih razloga jer za razliku od entiteta, MF termini predstavljaju događaje, procese ili aktivnosti [23].

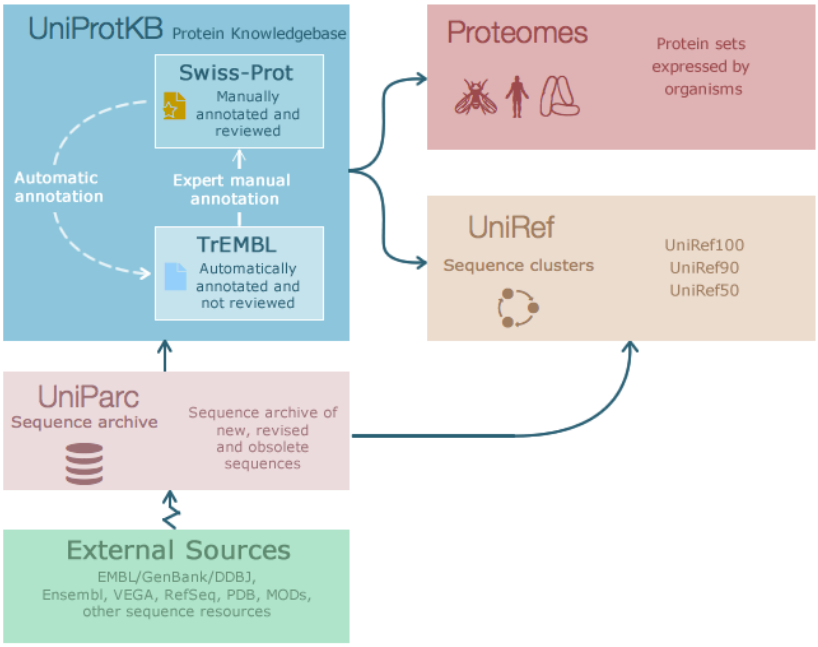
MF termini takođe imaju prepoznatljiv standardni sinonim za *x* **binding** [23]:

- *x* **receptor ligand**
- *x* **<name> binding**

3.2 UniProtKB/Swiss-Prot

UniProt skraćeno od (engl. *Universal Protein Resource*) je konzorcijum nastao 2002. izmedju tri organizacije: Evropski Bioinformatički Institut (EBI), Švajcarski Institut za Bioinformatiku (SIB) i Resurs Proteinskih Informacija (PIR).

UniProt obuhvata nekoliko baza i podbaza sa striktno definisanim tokom informacija Slika 3.4. Od prikazanih najbitnija je ***UniProtKB*** (engl. *UniProt Knowledge Base*) sačinjena od 2 podbaze.



SLIKA 3.4: Šematski prikaz povezanosti *UniProt* baze
(preuzeto sa uniprot.com)

1. ***Swiss-Prot*** sadrži visoko kvalitetne anotacije **neredundantnih** (objašnjeno u produžetku stavka6) proteinskih sekvenci. Informacije o sekvenci su dobijene iz postojeće literature, a kompjuterski predviđene anotacije su ručno proverene. *Swiss-Prot* kao baza postoji preko 30 godina.

2. *TrEMBL* (engl. *Translated EMBL*) je nadskup *Swiss-Prot* sekvenci dobijen prevođenjem EMBL i drugih nukleinskih sekvenci. Automatskom računarskom analizom anotirane su funkcijom i osobinama, ali zbog obimne količine sekvenci ti rezultati još nisu ručno provereni. Ove sekvence su redundantne, a njihova obimnost posledica je masovne primene NGS tehnologija. U februaru 2018. god *TrEMBL* sadržao je 107 627 435 sekvenci što je oko 200 puta više u poređenju sa 556 568 ručno proverenih *Swiss-Prot* sekvenci. Sve nove sekvence prvo ulaze u sastav *TrEMBL* da bi ručnom proverom napredovale u *Swiss-Prot* što se ogleda na Slici 3.4.

Distribucije *Swiss-Prot* baze dostupne su u nekoliko tekstualnih formata: ravna datoteka (engl. *flat file*), XML, RDF/XML. Ravni tekstualni format zbog standardizacije prati EMBL-Bank[**embl**] ravni format [24]. Unos u bazu se zove **slog** (engl. *record*) i sadrži sve informacije vezane za jedan protein. Jedan slog predstavljen u formatu ravne datoteke ilustrovan je uporšćenim prikazom na Slici 3.5. Ključne osobine slogova i *Swiss-Prot* baze podataka su:

1. Ime sloga **ID** (engl. *entery name*) je mnemonički zapis koji kodira taksonomske informacije o genu i proteinu. ID je podložan promenama i ne može se koristiti kao identifikator [26].

2. Identifikacioni broj, skraćeno **AC** (engl. *accession number*). Prvi u listi identifikatora naziva se **primarni** i služi da jednoznačno odredi slog. Ostatak identifikatora su tzv. **sekundarani AC** i nastaju iz dva moguća razloga [24, 26]:

   - Unifikacija postojećih proteina u jedan novi slog.
   - Specijalizacija jednog proteina u više različitih.

   U oba slučaja stari (primarni) AC se zadržava kao sekundarni AC u novom slogu.

3. Za razliku od *TrEMBL* , GO mapiranje za *Swiss-Prot* sekvence određuju se ručno [26].

4. **Ključne reči** (engl. *keywords*) označene **KW** opisuju hijearhisku strukturu kontrolisanog vokabulara namenjenog opisivanju funkcije proteina. Postoji 10 kategorija ključnih reči od kojih je za naše istraživanje bitna "Molekulska funkcija"[24]. Za razliku od GO čiji ideal je opis svih genskih produkta svih vrsta, termini ključnih reči prilagođeni su opisivanju sadržaja isključivo *Swiss-Prot* proteina [26].

5. Sekvenca **SQ** u slogu poznata je kao **kanonska** (engl. *canonical*) sekvenca. Kanonska sekvenca predstavlja konsenzus sekvencu produkta (protein) gena jedne vrste organizma. **FT** linije sadrže različite osobine kanonske sekvence uključujući i razlike u odnosu na izoforme[6] sekvence. U našoj analizi korišćena je isključivo kanonska sekvenca. Detaljan opis pravila za biranje kanonske sekvence može se naći na [26].

6. *Swiss-Prot* je **minimalno redundantna** u smislu da svi proteini kodirani jednim genom, jedne vrste su predstavljeni jednim slogom. Sve izoforme su grupisane pod jedan slog i jednu kanonsku sekvencu [27].

7. Postojnost proteina **PE** (engl. *Protein existance*) opisuje stepen sigurnosti da protein postoji **??**. Moguće vrednosti u rastućem poretku su: pronađeno na nivou proteina, pronađeno na nivou RNK, zaključeno iz homologije, predviđen i nesiguran.

[6] Izoforme su alternativni oblici sekvence nastali usled: alternativnog splajsovanja, upotrebe više promotera, alternativnih start kodona ili alternativnih okvira čitanja

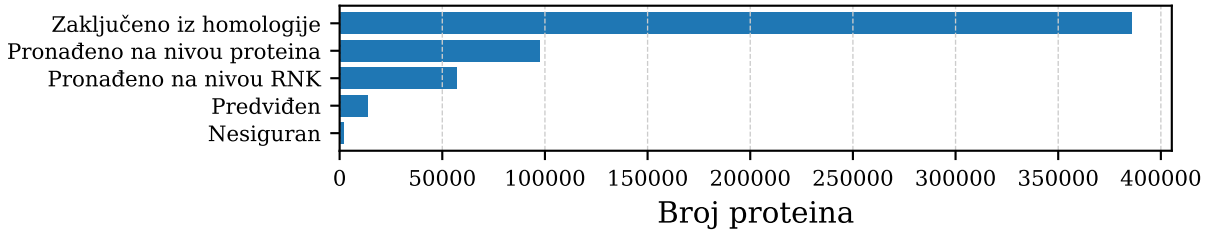
```
ID   ACSA_DROME              Reviewed;         670 AA.   | ime sloga, info
AC   Q9VP61; Q24226; Q8IH30; Q9VP60;                      | identifikacija
DT   19-SEP-2003, integrated into UniProtKB/Swiss-Prot. | ulazak u Svis-Prot
DT   01-MAY-2000, sequence version 1.                    | ulazak u TrEMBL
DT   25-OCT-2017, entry version 116.                     | poslednje
                                                         osvezavanje sloga

DE   RecName: Full=Acetyl-coenzyme A synthetase;         |
DE            EC=6.2.1.1;                                |
DE   AltName: Full=Acetyl-CoA synthetase;                |
DE            Short=ACS;                                 |
GN   Name=AcCoAS; ORFNames=CG9390;                       |
OS   Drosophila melanogaster (Fruit fly).                | Taksonomija
OC   Eukaryota; Metazoa; Ecdysozoa; Arthropoda; Hexap...|
OC   Pterygota; Neoptera; Holometabola; Diptera; Brac...|
OC   Ephydroidea; Drosophilidae; Drosophila; Sophopho...|
OX   NCBI_TaxID=7227 {ECO:0000312|EMBL:AAL90278.1};      |

RN   [1] {ECO:0000305}                                   | Prva referenca
RP   NUCLEOTIDE SEQUENCE (ISOFORM B).                    |
RA   Russell S.R., Heimbeck G.M., Carpenter A.T., Ash...| Autori
RT   "A Drosophila melanogaster acetyl-CoA-synthetase...| Naslov
RL   Submitted (NOV-1994) to the EMBL/GenBank/DDBJ da...|
RN   [2]                                                 | Druga referenca
...
CC   -!- FUNCTION: Activates acetate so that it can b...| Komentari
CC       synthesis or for energy generation.             |
CC       {ECO:0000250|UniProtKB:Q9NR19}.                 |
CC   -!- CATALYTIC ACTIVITY: ATP + acetate + CoA = AM...|
...

DR   EMBL; Z46786; CAA86738.1; ALT_SEQ; mRNA.            | reference ka
DR   EMBL; AE014296; AAF51695.2; -; Genomic_DNA.         | drugim bazama
...                                                      | (dbxref)
DR   ExpressionAtlas; Q9VP61; differential.              |
DR   Genevisible; Q9VP61; DM.                            |
DR   GO; GO:0005737; C:cytoplasm; IEA:UniProtKB-SubCell.| GO termin <----
DR   GO; GO:0003987; F:acetate-CoA ligase activity; I...| GO termin <----
...                                                      |

PE   2: Evidence at transcript level;
KW   Alternative splicing; ATP-binding; Complete proteome; Cytoplasm;
KW   Ligase; Nucleotide-binding; Reference proteome.
FT   CHAIN         1    670       Acetyl-coenzyme A synthetase.
FT                                /FTId=PRO_0000208425.
FT   VAR_SEQ       1    146       Missing (in isoform B).
FT                                {ECO:0000303|PubMed:12537569}.
FT                                /FTId=VSP_008310.
FT   CONFLICT    227    227       C -> S (in Ref. 1; CAA86738).
FT                                {ECO:0000305}.
SQ   SEQUENCE   670 AA;  75960 MW;  CE24364755CDBFFC CRC64;
     MPAEKSIYDP NPAISQNAYI SSFEEYQKFY QESLDNPAEF WSRVAKQFHW ETPADQDKFL
...
     KKMVRERIGP FAMPDVIQNA PGLPKTRSGK IMRRVLRKIA VNDRNVGDTS TLADEQIVEQ
     LFANRPVEAK
//   <--- oznacava kraj sloga
```

SLIKA 3.5: Uprošćen primer sloga, preuzet iz ravne datoteke `uniprot_sport.data` (preuzeto sa FTP servera [25])

8. *Swiss-Prot* takđe vrši predikcije neuređenih regiona: koristeći *DISOPRED2* and *CLADIST* prediktor [15]
   Međutim ove informacije postale su irelevantne pojavom baza neuređenja *MobiDB* i *D2P2*.

9. Zanimljiva zapažanja globalne statistike:

   - Najzastupljenije sekvence su kraće od 500 aminokiselina.
   - Postojnost oko 70% proteina potvrđeno je homologijom.
   - Zastupljeno je preko 1000 različitih organizama međutim većina *Swiss-Prot* sekvenci pripada malom broju model organizama.
   - Većina proteina ima dužinu između 100 i 500 AK.



SLIKA 3.6: Histogram nivoa pouzdanosti *Swiss-Prot* proteina

Glava 4

Podaci i metode

Cilj rada je ispitivanje veze između molekulske funkcije proteina i njegove (ne)uređenosti tj. da li molekulska funkcija zavisi više od uređenosti ili neuređenosti. Istraživanje je motivisano radom *"Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions"* [28]. Navedeni rad je prvi u seriji od tri rada i bavi se prvenstveno biološkim procesima i molekulskim funkcijama. U nastavku teksta pod terminima **originalni** rad, autori, podaci, metode i slično podrazumevaćemo navedeni rad njegve autore, metode, podatke itd.

Najveća razlika u našim pristupima je što su originalni rezultati izraženi u terminima **ključnih reči** dok su naši rezultati izraženi u **GO terminima**. Oba pristupa proizvode listu funkcija koje više obuhvataju (ne)uređene proteine, ali GO termini zbog granularnosti se prirodnije predstavljaju grafovski i sadrže mnogo više informacija. Konačni rezultat ovog istraživanja predstavlja poređenje ova dva pristupa.

4.1 Podaci

Za metode koje prezentujemo potrebne su tri vrste informacija:

1. Što više različitih proteina
2. Pouzdana anotacija funkcija
3. Informacije o funkcijama, prvenstveno međurelacije
   (Međurelacije između funkcija su bitne ako ih je potrebno grupisati)

4.1.1 Podaci iz originalnog rada

U originalnom radu [28] korišćena je baza podataka ***Swiss-Prot*** Poglavlje 3.2, verzija 48 iz 2005. Verzija 48 ima 201 560 proteina od kojih 196 326 imaju dužinu preko 40 aminokiselina (što je potrebno zbog Definicije 1 u nastavku). Funkcije pridružene proteinima izražene su **kontrolisanim vokabularom** (engl. *controlled vocabulary*) koga čine takozvane UniProtKB **ključne reči** (engl. *keywords*). U verziji 48, UniProtKB sadrži 874 ključnih reči. Zbog statističke značajnosti posmatrane su one ključne reči kojima je bilo anotirano barem 20 proteina, tj. 710 ključnih reči.

Kao što je pomenuto u Poglavlju 3.2 kanonske sekvence (proteini) u *Swiss-Prot* bazi podataka nisu redundantne u smislu da jedan unos u bazi podataka predstavlja produkt jednog gena iz jedne vrste organizma. Međutim za analizu funkcija *Swiss-Prot* **jeste statistički redundantna** [**proveriti**] jer sadrže veliku količinu **homologih** proteina (prvenstveno ortologa). Originalni autori izvršili su klasterovanje *Swiss-Prot* proteina u **proteinske familije** dobivši 27 217 familija. Pri klasterovanju svaki protein ima težinu kojom doprinosi daljoj analizi. Težina svakog proteina u preseku klastera sa datom

funkcijom je obrnuto proporcionalna veličini preseka tako da je zbir težina svih proteina jednaka veličini preseka.

4.1.2 Naši podaci

Ugledom na **CASP** takimčenja, **CAFA** (engl. *Critical Assessment of Functional Annotation*) takmičenje pokrenuto je zarad objektivnog ocenjivanja prediktora funkcije i usmerivanja budućeg razvoja prediktora funkcije [29]. U ovom radu korišćen je trening skup proteina preuzet sa **CAFA3** takmičenja korišćen za treniranje prediktora funkcije.

CAFA3 trening skup (u nastavku samo CAFA3 skup ili CAFA3 podaci) je podskup *Swiss-Prot* baze (iz 2016.) koji uključuje sve proteine iz model organizama: *Human, Mouse, Rat, S. cerevisiae, S. pombe, E. coli, A. thaliana, Dictyostelium discoideum, Zebrafish, Bacillus cereus.* sa izuzetokm sekvenci *Drosophila and Candida* koje su preuzete iz svojih respektivnih genomskih baza podataka. Informacije o sadržaju CAFA3 trening skupa izneo je Iddo Friedberg PhD iz CAFA3 tima. Naša analiza izvršena je pod pretpostavkom da CAFA3 skup nije statistički redundantan. Ova pretpostavka uporošćava metode analize.

TODO

Swiss-Prot proteini su kodirani jednim karakterom koristeći **IUPAC** kodove. U podacima javljaju se sekvence sa 21. i 22. aminokiselinom ('U' i 'O') kao i višeznačne oznake: 'B', 'J', 'X' i 'Z'. Ovakve sekvence nisu podržane od strane VSL2b prediktora i za nas predstavljaju nevalidne proteinske sekvence. Pod **validnom proteinskom sekvencom** smatraćemo sekvencu koja je validan ulaz za VSL2b, tj. čini je azbuka od 20 standardnih aminokiselina i ima minimalnu dužinu 9 AK.

CAFA3 Podaci se sastoje od dve datoteke:

1. `uniprot_sprot_exp.fasta` sadrži 66 841 protein od kojih 66 599 za našu analizu predstavljaju validnu proteinsku sekvencu. Od preostalih proteina 66 063 ima dužinu veću ili jednaku od 40 aminokiselina.

2. `uniprot_sprot_exp.txt` pridružuje funkcije u obliku **GO termina**. Zastupljeni su termini iz sva tri imenska prostora: 16 117 ćelijskih komponenti, 5 966 molekulskih funkcija i 16 117 bioloških procesa. Jednom proteinu može biti pridruženo više GO termina i obrnuto.

Naša analiza primarno je orijentisna na korišćenje GO termina za opis funkcije i razlikuje se od originalnog pristupa. Analiza sa GO terminima (grupisanje po funkciji) zahteva prvenstveno poznavanje *IS_A* roditeljske veze između termina. Takođe, tokom istraživanja bile su nam potrebne i ostale informacije o terminima. Pomenute informacije dobili smo iz datoteke [30] verzije 01.12.2017.

Radi poređenja dobijenih rezultata potrebno je poznavanje relacije između ključnih reči i GO termina. Postoje dva dostupna mapiranja:

- [31] verzija 20.12.2017 sadrži detaljan opis 1188 ključnih reči od kojih 195 pripada kategoriji **Molekulskih funkcija**. Od 195 samo 145 ima mapiranje na jedan ili više GO termina.

- [32] sadrži samo mapiranja i generiše ih **GOA projekat** [33]. Ipak ova mapiranja nisu korišćena jer su u našoj analizi proizvodila redundantna i nepoželjna mapiranja.

Pošto je originalni rad iz 2007. godine, postoje razlike u vokabularu ključnih reči, razlike u samim sekvencama proteina, broj proteina i anotacije ključnih reči na proteine.

Iz tog razloga, bilo je potrebno prvo ponoviti analizu sa vokabularom ključnih reči da bi se procenilo koliko ove razlike utiču na originalne rezultate.

Zbog toga CAFA3 podatke ne možemo da posmatramo kao crnu kutiju već je bilo neophodno povezati ih sa *Swiss-Prot* proteinima prvenstveno zbog pridruživanja ključnih reči. Kako postoje razlike između najnovije verzije *Swiss-Prot* baze i CAFA3 podataka bilo je neophodno izvršiti "korektno" spajanje i analizu razlika. Informacije o pridruženim ključnim rečima takođe su nam bile značajne za proveru validnosti mapiranja na GO termine i testiranje potencijalnih drugih metoda mapiranja. Naša očekivanja su da iste funkcije podrazumevaju anotaciju na iste proteine. Ovi koraci detaljno su opisani u Poglavnju 5.

4.2 Metod

Idealan slučaj. Pretpostavimo da za proizvoljnu molekulsku funkciju znamo sve strukturno različite proteine koji je obavljaju. Da bi dali korektan odgovor moramo da znamo kako neuređenost pojedinačnog proteina utiče na na njegovo ponašanje, i kako to ponašanje (tip neuređenosti) utiče na datu funkciju.

Realnost.

- Broj eksperimentalno određenih neuređenih regiona je veoma mali. Baza eksperimentalno utvrđenih neuređenih regiona **Disprot** ima svega 803 proteina sa opisanih 2167 neuređenih regiona [34]. Nažalost pouzdanost ovih regiona je diskutabilna jer različite eksperimentalne tehnike koje su korišćene imaju različitu pouzdanost. Najveću pouzdanost nose regioni koji su eksperimentalno utvrđeni sa većim brojem eksperimentalnih tehnika.

- Prediktori su trenirani na malom podskupu proteina iz Disprot i PDB baze. Čak i konsenzus nekoliko različitih prediktora ne daje dovoljno pouzdane rezultate o lokaciji neuređenog regiona (prof. Nenad Mitić, usmena komunikacija 2017.).

- Pozitivna strana je najnoviji napredak, razvoj prediktora koji direktno pokušavaju da predvide funkciju koju IDPr obavlja [15].

Jednostavna alternativa je da se pretpostavi da veći udeo neuređenih u odnosu na uređene proteine podrazumeva da funkcija zavisi više od neuređenosti. Međutim, prvo je potrebno definisati kada protein smatramo neuređenim. Definicija mora da ima biološkog smisla, da bude prilagođena analizi, ali pored takođe je ograničena sposobnostima i preciznošću prediktora koji se korist. Više o tome u narednom Podpoglavlju 4.2.1.

4.2.1 Predikcija neuređenosti proteina

Originalni autori koristili su **PONDR VL3E** prediktor koji postiže tačnost od 87% pri unakrsnoj validaciji nad uravnoteženim test skupom. Zbog ekonomičnosti i dostupnosti u ovom radu korišćen je noviji prediktor druge generacije **PONDR VSL2b**. Relevantne karakteristike VSL2b detaljno su opisane u 2.3.2. Za potrebe analize originalni autori uvode sledeću definiciju:

Definicija 1 *Protein je **putativno neuređen** (najverovatnije neuređen) (engl. putatively disordered) ako sadrži bar jedan region veći ili jednak od 40 uzastopnih aminokiselina takvih da imaju predviđenu neuređenost iznad 0.5.*

Onda definišemo operator d takav da za svaku proteinsku sekvencu s_i važi:

$$d(s_i) = \begin{cases} 1 & \text{ako je } s_i \quad \textbf{putativno neuređena} \\ 0 & \text{suprotno} \end{cases}$$

U nastavku pod neuređenošću podrazumevaće se **putativna neuređenost**. Uslov "≥ 40" u originalnom radu delom je posledica ograničenja VL3 prediktora koji je treniran na **dugim** sekvencama[1].

4.2.2 Zavisnost dužine proteina i predikcije dugačkog neuređenog regiona

Verovatnoća da po gornjoj definiciji protein bude klasifikovan kao verovatno neuređen raste sa porastom njegove dužine. Ova činjenica utiče na statističku značajnost rezultata. Autori [28] predlažu narednu formulu da se ta verovatnoća proceni:

Neka je S_L skup proteina sa dužinama iz intervala $[L-l, L+l]$ gde je $l = 0.1 \cdot L$[2]. Dobijamo sledeće formule:

$$S_L = \{s_i : \quad |L - |s_i|| \leq l \quad \}, \quad |s_i| \text{ je dužina sekvence}$$

$$P_L = \frac{\sum_{s_i \in S_L} d(s_i)}{|S_L|}, \quad |S_L| \text{ je kardinalnost skupa}$$

Grafik funkcije P_L u zavisnosti od promenljive L predstavljen je na Slici 4.1. Glatkoća rezulatata kontroliše se veličinom l koja predstavlja prozor uprosečavanja. Kako prozor uprosečavanja raste sa porastom dužine proteina ($l = 0.1 \cdot L$) tako P_L postaje glađa sa veličinom promenljive L. Konstantni prozor uprosečavanja je tehnika još poznata kao (engl. *rolling average*) ili (engl. *boxcar filter*) i predstavlja prostu vrstu konvolucije. Nije nam poznato zašto su originalni autori odlučili da veličina prozora raste sa dužinom proteina. Koliko nam je poznato u pitanju je heruistika.

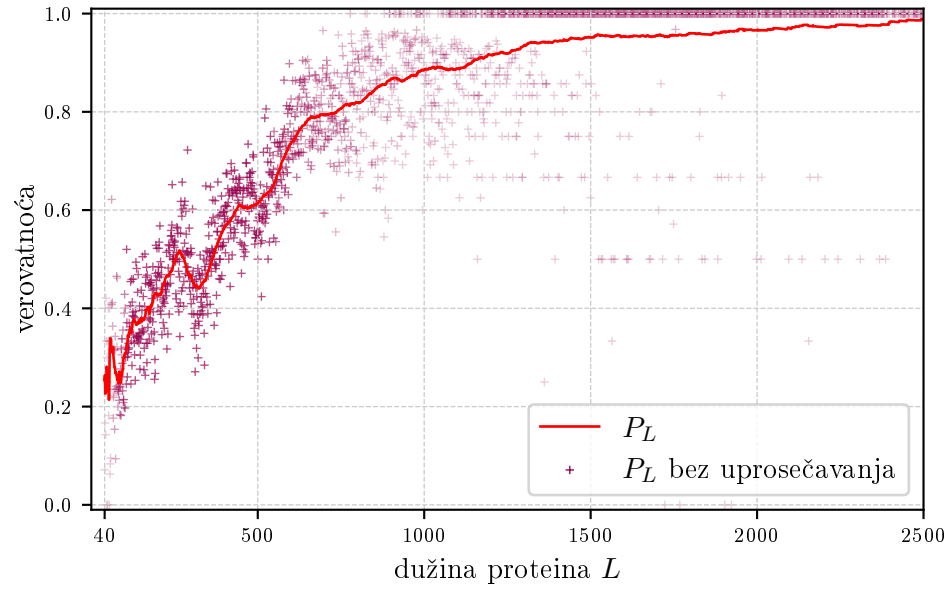
Pored gore prikazanog originalnog metoda biće predstavljen još jedan pristup **Slučajno generisani** (engl. *random generated*) proteini za procenu P_L. Razmotrićemo dva modela. Prvi je naivni model **uniformne verovatnoće** koji podrazumeva da se svaka aminokiselina javlja sa istom verovatnoćom odnosno $1/20$. U statistici ovo je još poznato kao (engl. *equiprobable model*). Drugi model koji ćemo zvati 'slučajni' ili 'random' model predstavlja slučajnu promenljivu čija verovatnoća zavisi od učestalosti aminokiselina iz CAFA3 skupa i prikazana je na Slici 4.2. Koristeći ova dva modela za svaki protein generisan je slučajan protein iste dužine koji se koristi za procenu P_L.

Poređenje predloženih pristupa sa originalnim P_L prikazano je na Slici 4.3. Jasno se vidi da slučajni model predstavlja vizuelno dobru aproksimaciju dok uniformni model znatno odstupa i dosta sporije raste (naizgled skoro linearno). Kako VSL2b prediktor prepoznaje neuređene regione na osnovu učestalosti aminokiselina, ovo ponašanje nije čudno jer je manja verovatnoća pojave aminokiselina koje promovišu neuređenost. Zbog suviše velikog odstupanja uniformni model nije korišćen u daljoj analizi.

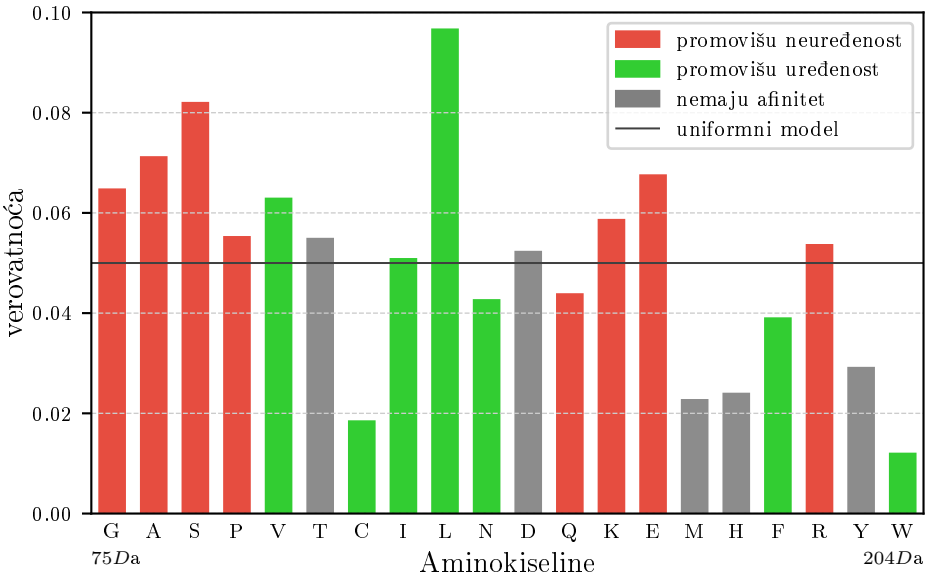
Jedno od objašnjenja zašto je uniformni model naivan i toliko odstupa od prvobitnog metoda proizilazi iz činjenice da aminokiseline imaju inherentno različite verovatnoće. Naime, aminokiseline ne mogu imati istu verovatnoću jer se broj njihovih kodona razlikuje. Neke aminokiseline su kodirane sa samo jednim, a druge i sa 6 kodona. Očekivano je da broj kodona povećava učestalost aminokiseline i ta korelacija uz izuzetke arginina se vidi na Slici 4.4 [35].

[1] L označava duge regione, ≥ 30 AK

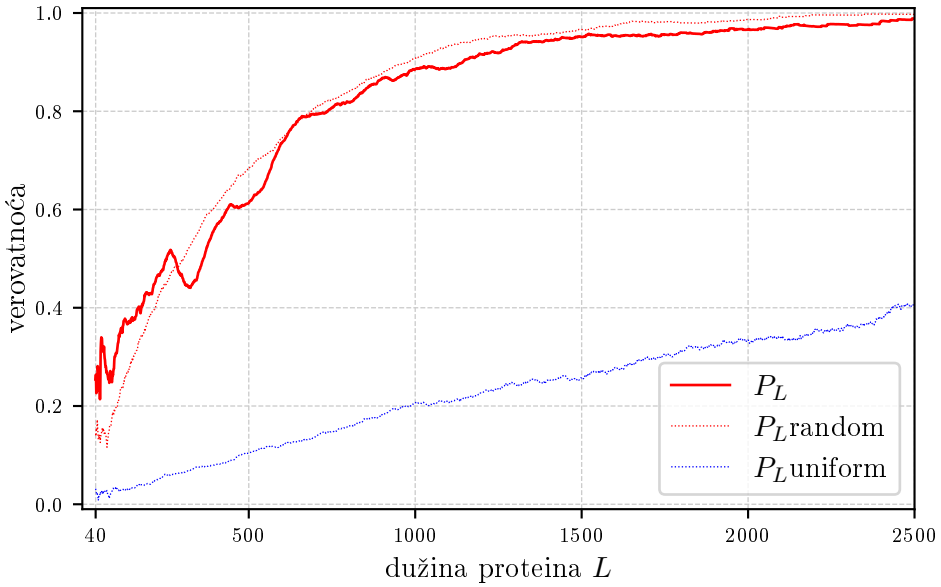
[2] Na primer, skup S_{100} sadrži proteine iz invervala $[90, 100]$, a S_{500} iz intervala $[450, 550]$



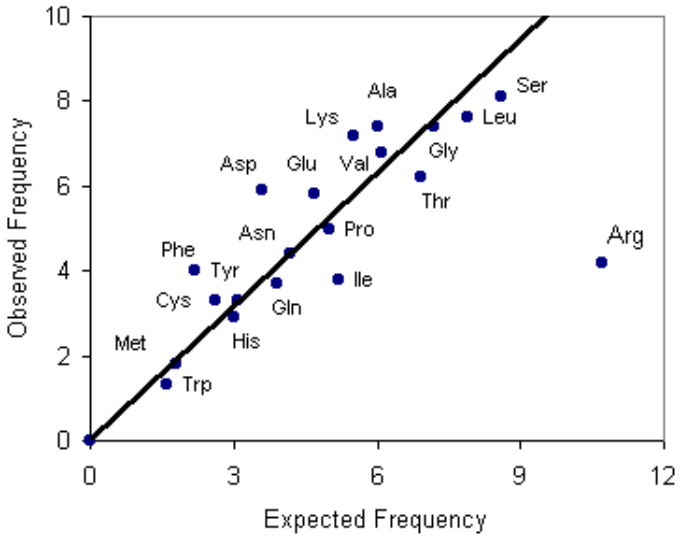
SLIKA 4.1: Zavisnot neuređenosti i dužine CAFA3 proteina minimalne dužine 40 AK
(P_L sa prozorom uprosečavanja podrazumeva $l = 0.1L$, dok krstići predstavljaju sirove vrednosti $l = 0$. Transparentnost krstrića služi da ilustruje brojnost proteina date dužine. Ako je krstić transparentan znači da sadrži manje od 100 proteina.)



SLIKA 4.2: Učestalost aminokiselina u CAFA3 podacima
(Učestalost prikazuje uniformni model, dok boja AK obeležava afinitet prema uređenosti ili neuređenosti. Aminokiseline su poređane u rastućem poretku od najlakše do najteže.



SLIKA 4.3: Upoređivanje modela za procenu P_L nad CAFA3 podacima



SLIKA 4.4: Očekivana i realna učestalost aminokiselina kod sisara
(Preuzeto sa: www.tiem.utk.edu/~gross/bioed/webmodules/aminoacid.htm)

4.2.3 Ocenjivanje zavisnosti funkcije od neuređenosti

Neka je S_j skup proteina koji imaju pridruženu funkciju j. Tada se procenat neuređenih proteina u oznaci F_j može izračunati kao:

$$F_j = \frac{\sum_{s_i \in S_j} d(s_i)}{\|S_j\|}$$

Nulta hipoteza za F_j je tvrdnja da F_j zavisi samo od dužine sekvence tj. P_L.
Neka je X_L Bernulijeva slučajna promenljiva oblika $X_L : \begin{pmatrix} 0 & 1 \\ P_L & 1 - P_L \end{pmatrix}$.
Tada nultu hipotezu modeliramo raspodelom Y_j, koja za razliku od F_j koristi slučajnu promenljivu X_L umesto $d(s_i)$, odnosno:

$$Y_j = \frac{\sum_{s_i \in S_j} X_{|s_j|}}{\|S_j\|}$$

Ako F_j izlazi iz intervala poverenja raspodele Y_j onda, funkcija j sadrži značajno mnogo predviđenih (ne)uređenih proteina. Preciznije, ako je p-vrednost (engl. *p-value*) < 0.05 funkcija j je povezana sa neuređenim proteinima a, ako je *p-value* > 0.95 funkcija j je povezana sa uređenim proteinima. Suprotno ne može ništa da se tvrdi za funkciju j. U kada se kaže da je funkcija, ključna reč ili GO termin (ne)uređen podrazumevaće se da je odgvorajauća p-vrednost manja od < 0.5 ili veća od > 0.95.

Zbog X_L teško je analitički proceniti Y_j, pa se se pribegava empirijskom računanju p-vrednosti. Empirijska p-vrednost određena je tako što je za 1000 realizacija Y_j izračunato očekivanje da je realizacija Y_j veća od F_j.

Preciznije, vektor[3] S_j sadrži k prteina $S_j = \{s_1, s_2, ..., s_k\}$. Protein s_i ima dužinu L_i za koju je izračunata verovatnoća P_{L_i}. Tada generatorom Bernulijevih slučajnih brojeva, za svaki protein p_i na osnovu P_{L_i} generišemo realizaciju X_L. Dobijen je vektor od k vrednosti nula ili jedan. Učestalost jedinica u dobijenom vektoru predstavlja prvu realizacija Y_j. Postupak se ponavlja hiljadu puta i broji se koliko puta je realizacija Y_j bila veća od F_j. Dobijeni zbir deli se sa 1000 i rezulatat je empirijska p-vrednost.

Originalni autori tvrde da se za veće skupove S_j, raspodela Y_j ponaša kao normalna. To znači da se ocena Z-skor može dobiti kao $Z_j = (F_j - \mu_j)/\delta_j$ gde je μ_j očekivanje, a δ_j standardna devijacija. Dodatno, p-vrednost može da se aproksimira kao $1/2(1 - erf(Z_j/2))$ [4] ako raspodela liči na normalnu. Ovo je nekad korisno jer sa 1000 realizacija Y_j nema dovoljnu preciznost za p vrednost manju od $1/1000 = 0.001$. Međutim u ovom radu to nije korišćeno jer su sva sortiranja (kao i u originalnom radu) izvršena po Z-skor oceni.

[3] zamenili smo skupa S_j za vektor S_j. Ovo je implementacioni detalj

[4] $erf()$ je gausova funkcija greške, $erf(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$

Glava 5

Priprema podataka

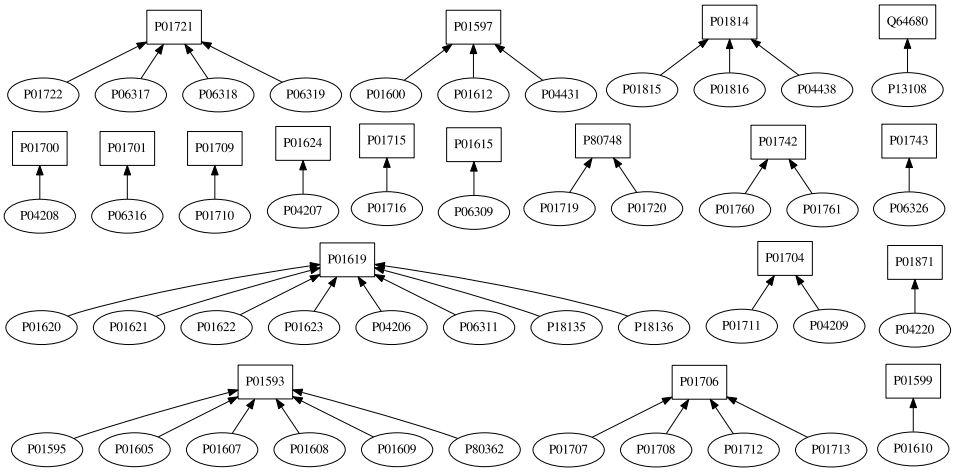
5.1 Objedinjavanje CAFA3 i novije *Swiss-Prot* verzije

Iz CAFA3 trening skupa izdvojeni su svi validni proteini (dužine barem 9 i azbukom od 20 standardnih aminokiselina). U ovom koraku ne izbacujemo proteine kraće od 40 AK.

Informacije o *Swiss-Prot* bazi podataka dobijene su iz verzije 2017_12, iz datoteke `uniprot_sprot-only2017_12.tar.gz` [25]. Navedena verzija sadrži 556 196 proteina.

Od 66 599 validnih CAFA3 proteina 66 530 ima nepromenjen **primarni identifikator**. Međutim 69 slogova u *Swiss-Prot* bazi sadrže nedostajuće CAFA3 identifikatore kao sekundarne kao što je navedeno u Potpoglavlju 3.2. Ovo je posledica dva moguća mehanizma:

1. Unifikacija nekoliko CAFA3 proteina pod novi slog. Rezultat unifikacije prikazan je na Slici 5.1. Analizom ovih promena uspešno su rekonstruisana svega četri nova *Swiss-Prot* sloga koja odgovaraju nedostajućim CAFA3 proteinima. Kako je četri suviše mali broj, zbog jednostavnosti nismo ih ubrajali u dalju analizu te koristimo samo 66 530 originalnih CAFA3 proteina.



SLIKA 5.1: Unifikacija starih(elipse) na nove slogove u *Swiss-Prot* bazi podataka

2. Specijalizacija jednog CAFA3 proteina u više različitih slogova. Zbog moguće statističke redundantnosti ovi slogovi su zanemareni.

Validini CAFA3 proteini anotirani su sa 5 957 različitih GO termina Molekulske Funkcije (MF) od kojih je 50 zastarelo i izbačeno iz `go.obo` datoteke. U *Swiss-Prot* bazi

podataka nismo bili u mogućnosti da proverimo samo za MF, ali ukupno je izbačeno 319 GO termina. CAFA3 sadrži 67 MF koje se ne javljaju u *Swiss-Prot* anotacijama dok *Swiss-Prot* sadrži 888 MF koje se ne javljaju u CAFA3 anotacijama. Pošto *Swiss-Prot* treba da sadrži svežije (tačnije) informacije, CAFA3 verzija anotacija je zanemarena u korist novijih *Swiss-Prot* anotacija. Ove informacije sumirane su Tabelom 5.1

Dodatno, *Swiss-Prot* sadrži 194 proteina čije se sekvence razlikuje u odnosu na CAFA3 verziju proteina. Odlučili smo da zadržimo originalne CAFA3 sekvence.

	CAFA3	*Swiss-Prot*
MF termini	5 957	-
fali u obo.go	60 MF	319 MF, CC i BP
MF, samo u	67	888

TABELA 5.1: Razlike u GO terminima između CAFA3 i *Swiss-Prot*

5.2 Grupisanje proteina po GO terminima

Ako za GO termin A važi *is_a* B i želimo da diskutujemo o funkciji B onda i svi proteini funkcije A trebaju biti pridruženi funkciji B. Primetili smo da anotacije ključnih reči ovo već podrazumevaju dok za GO to nije slučaj. Tek nakon što je ovo grupisanje odrađeno validno je odbaciti termine koji anotiraju manje od 20 proteina. Za predloženo grupisanje koristili smo algoritam topološkog sortiranja. Ovom metodom dobijeno je 1781[1] MF termin sa minimum 20 pridruženih proteina. Ovo uključuje i koreni[2] termin . U ovom koraku uračunati su samo proteini minimalne dužine 40 AK.

5.3 Ontologije gena i ključne reči

U nastavku razmatramo samo ključne reči kategorije MF i načine na koje ih možemo povezati sa ekvivalentnim MF terminima. Direktno mapiranje sa **ključnih reči** na GO termine nije uvek moguće. Neke ključne reči (*Represor, Cyclin, Activator, Superantigen, Tumorantigen...*) nemaju odgovarajući GO termin. Od 226 ključnih reči svega 175 ima odgovarajući GO termin. Od preostalih 175 ključnih reči postoji 105 mapiranja na MF termin, 59 na BP i 11 na CC. Dakle, za 70 MF ključnih reči ne postoji direktno mapiranje na MF termine. Slika 6.1 prikazuje moguća direktna mapiranja (dobijena iz `keywords.txt`) za 20 neuređenih MF ključnih reči pronađenih u radu [28]. U našoj verziji `keywords.txt` Antigen je izbačen i zamenjen specijalizovanijom podelom. Tri ključne reči nemaju mapiranje, dve se mapiraju na ćelijske komponente, osam na biološke procese i svega šest na molekulske funkcije. Statistički značajne ključne reči su ljubičaste dok radi kompletnosti navodimo neke njihove specijalizacije i generalizacije koje su obojene sivo. GO termini su predstavljeni manjim kružićima.

Za neke BP termine moguće je doći do molekulske funkcije praćenjem veza **je deo** ili **sadrži**. Međutim od 8 bioloških procesa prikazanih na Slici 6.1 ovo je moguće samo za Neuropeptid. **Cypher** upitom:

```
MATCH p=(:Keyword {name:"Neuropeptide"})--(:GOTerm)<-[*0..]-(:GOTerm)<--(:Protein)
RETURN p
```

[1] Bez ovog grupisanja imali bi samo 1146 MF termina

[2] koreni termin ili koreni čvor ontologije tj. termin molekulska funkcija

pronašli smo mapiranje na MF *neuropeptide receptor activity* predstavljeno Slikom 5.3. Nažalost uočava se da MF termini i ključna reč Neuropeptid sadrže svega 5 zajedničkih proteina. Ovo je očekivano s obzirom da pronađeni MF termin predstavlja proteine koji se vezuju za neuropeptide.

Mi pretpostavljamo da je neuspešnost mapiranja posledica veze **je deo** koja ne podrazumeva kompoziciju (**sadrži**) već samo agregaciju. Relacija agregacije podrazumeva da molekulska funkcija postoji nezavisno od biološkog procesa za razliku od relacije kompozicije. Kod pomenutih osam BP termina i njihovih specijalizacija ne javlja se veza **sadrži**. Dakle automatsko poređenje moguće je samo za šest od 20 originalnih neuređenih MF ključnih reči. Za uređene ključne reči postoji više direktnih mapiranja ali njih ovde nećemo navoditi.

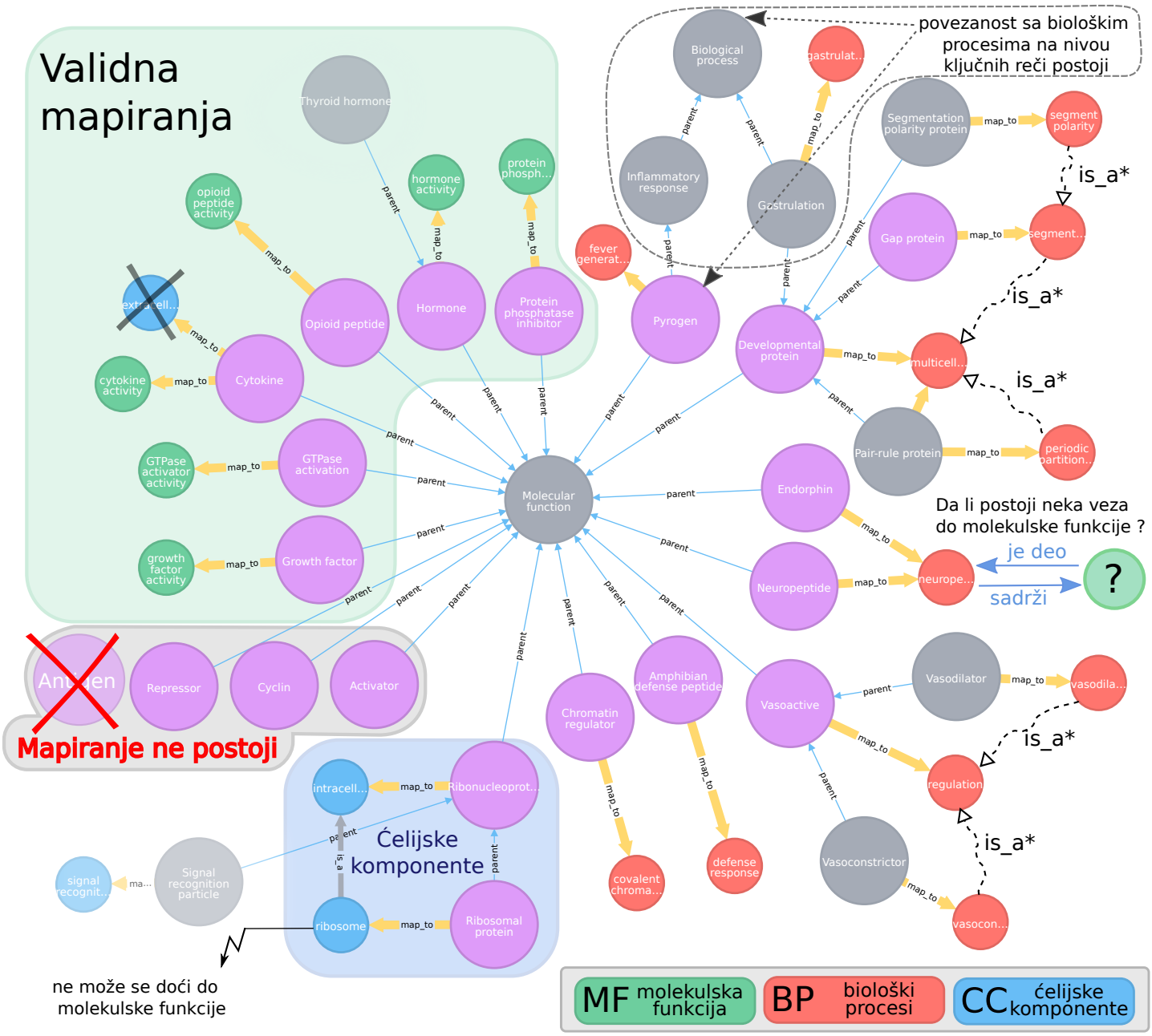
Konačni metod koji koristimo da dopunimo gore pomenuto mapiranje je poluautomatski. Ključne reči od intersa ili njihove delove slažemo u regularan izraz oblika:

```
# words je lista ključnih reči ili njihovih delova
# re.I znači izjednačavanje malih i velikih slova
expresion = re.compile( f"({'|'.join(words)})[^ ,.)]*", re.I )
```

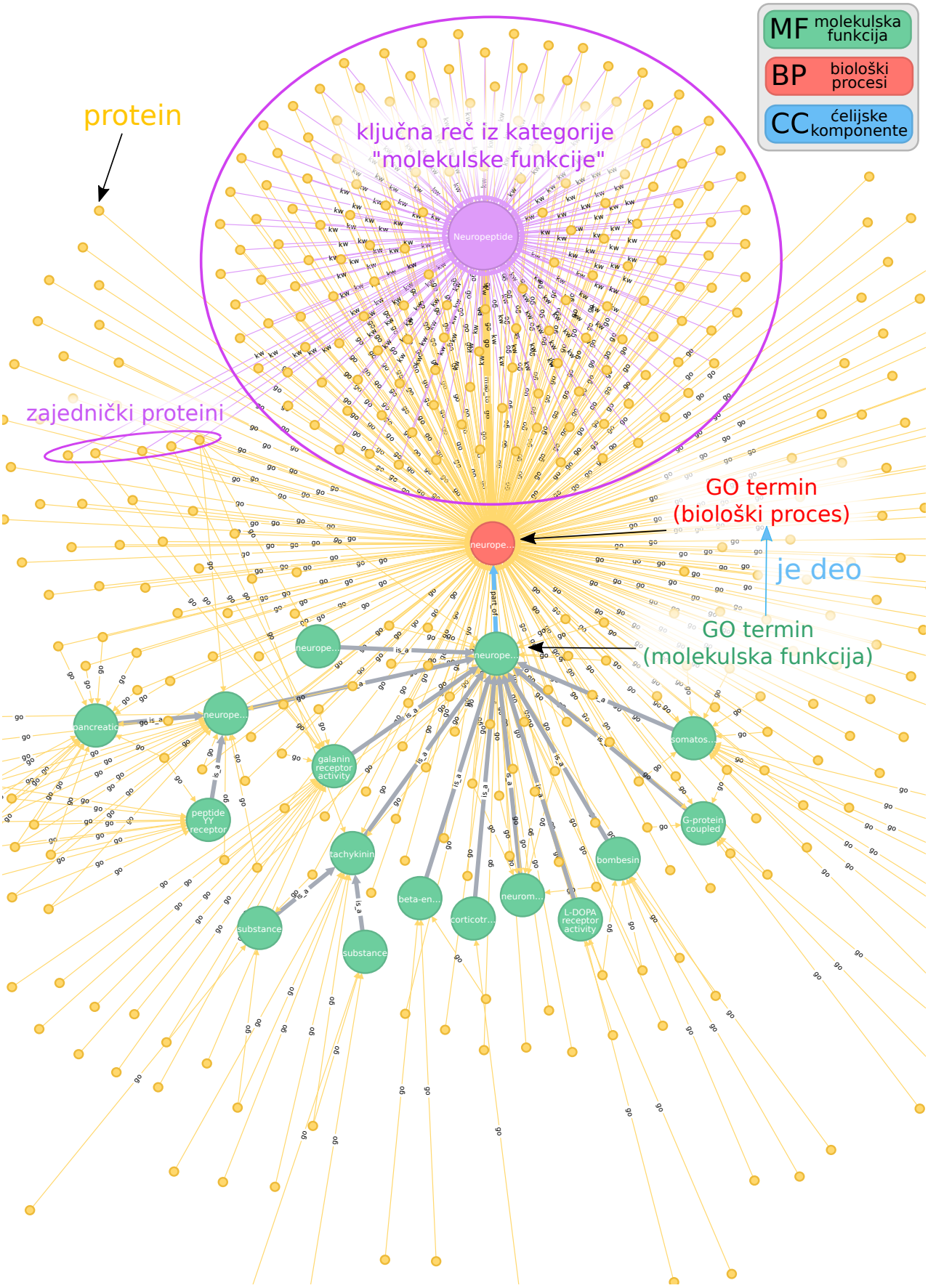
Konstruisani regularan izraz iskorišćen je na sledeći način. Prvo se pretražuje da li ime GO termina sadrži neku reč iz regularnog izraza. Ako ne, pretraga se proširuje na sinonime GO termina. Ako ni tu nije pronađena, pretraživanje se proširuje na definiciju termina. Rezultat su tri vrste relacije opadajuće pouzdanosti (ime, sinonim, definicija). Zajedno sa originalnim direktnim mapiranjima i pažljivom ručnom analizom moguće je dopuniti mapiranje i izvesti poređenje rezultata. Pri analizi rezultata neophodno je uzeti u obzir standardne definicije GO termina i sinonima navedenih u Potpoglavlju 3.1.3.

Pomenuto mapiranje vršimo nad statistički značajnim (ne)uređenim MF terminima. U slučaju da se ključna reč ne pronađe pretraga se može izvesti za deo imena ključne reči, međutim ovaj pristup može dovesti do netačnog mapiranja. Sasvim je moguće da ključna reč nije pronađena jer odgovarajući MF termini nisu statistički značajni što se mora zasebno proveriti. Smatramo da je opisan metod uspešan zbog konzervativnosti fraza koje tražimo kao i informacije o sinonimima koje GO termini sadrže.

Ključne reči često imaju karakter '-' koju GO termini izbegavaju. Zamena blanko oznakom je neophodna za pretraživanje imena, ali ne i za sinonime koji često sadrže '-' karakter.



SLIKA 5.2: Mapiranje 20 najznačajnijih neuređenih MF ključnih reči [28] na GO termine.



SLIKA 5.3: Mapiranje ključne reči **Neuropeptid** na MF termine preko veze ”je deo”. Ovako postignuto mapiranje rezultuje mali brojem zajedničkih proteina.

Glava 6

Rezultati

Od ukupno 186 MF ključnih reči sa preko 20 pridruženih proteina, 116 su statistički značajne od čega su 53 uređene (p<0.05), a 63 neuređene (p>0.95). Od ukupno 1781 MF termina sa preko 20 pridruženih proteina (dobijeno grupisanjem), 1315 su statistički značajni od čea su 699 uređeni, a 616 neuređeni. Tabela 6.1 prikazuje razlike u odnosu na originalne rezultate.

	Org. rez.	Naši rezultati	
	Xie2007 kw	MF kw	MF termini
ukupno	143	186	1781
p<0.05 (uređene)	37	53	699
p>0.95 (neuređene)	51	63	616

TABELA 6.1: Uopštena usporedba rezultata

Pošto su izvršene dve vrste grupisanja rezultati će biti predstavljeni na sledeći način. Tabelama 6.1 i 6.3 prikazano je poređenje originalnih rezultata sa našim grupisanjem po **ključnim rečima**. Navedene tabele sortirane su po Z-skor vrednosti. Za razliku od originalnih rezultata naše tabele ne sadrže broj klastera, ali dodatno sadrže informaciju o procentu neuređenih proteina F_j koju nazivamo **neuređenost funkcije**.

Grupisanje po ključnim rečima predstavlja se kao podgraf originalne MF ontologije pri čemu su zadržani samo statistički značajni termini. Isečak grafa neuređenih termina prikazan je na Slici 6.2. Pozadinska boja termina kodira pomenutu neuređenost termina. Korišćeno je viridis [36] mapiranje boja koje nulu predstavlja tamno ljubičastom a jedinicu svetlo žutom. Tamni, plavkasti termini su uređeni dok su svetli, zelenkasti neuređeni. Crvene elipse predstavljaju naj. 20 (ne)uređenih ključnih reči iz originalnog rada i povezani su sa terminima preko četiri tipa veze. Zelena veza predstavlja direktno mapirnaje, crvena preko imena, ljubičasta preko sinonima, a žuta pominjanje u definiciji. Na adresi [37] u `.svg` formatu dostupni su svi rezultati ovoga oblika. Pomenute slike sadrže dodatne informacije o terminima koje se otkrivaju ako strelicu miša držite iznad termina (samo ako sliku otvorite u veb pretraživaču). Svaki termin sadrži odgovarajući hiperlink ka *amigo* veb stranici.

Radi izbegavanja dvosmislenosti u daljem tekstu imena ključnih reči imaće dodeljen prefiks *KW:* dok će imena MF termina imati prefiks *MF:* .

#	*Keywords*	*Number of proteins*	*Number of families*	*Average sequence length*	*Z-score*	*P-value*
1	Ribonucleoprotein	12236	412	150.55	22.13	1
2	Ribosomal protein	11692	330	140.58	20.63	1
3	Developmental protein	3260	721	477.93	19.28	1
4	Hormone	1187	161	141.13	15.58	1
5	Growth factor	785	84	255.70	11.16	1
6	Cytokine	899	110	213.28	10.21	1
7	Neuropeptide	268	209	95.08	9.65	1
8	Activator	3086	573	428.47	9.04	1
9	GAP protein	47	2	232.96	7.42	1
10	Antigen	1113	455	437.48	6.99	1
11	Repressor	2309	449	374.46	6.92	1
12	Chromatin regulator	334	100	801.24	6.70	1
13	Pyrogen	37	2	262.59	6.44	1
14	Vasoactive	125	39	160.39	5.56	1
15	Amphibian defense peptide	123	148	50.64	5.44	1
16	GTPase activation	311	70	831.03	5.36	1
17	Endorphin	42	4	226.68	5.35	1
18	Opioid peptide	24	4	216.96	5.14	1
19	Protein phosphatase inhibitor	47	8	366.51	5.07	1
20	Cyclin	182	25	430.58	4.88	1

#	*name*	*n*	*avg_len*	*avg_dis*	*z*	*p*
1	DNA-binding	6518	546.53	0.87	46.90	1.0
2	Developmental protein	3897	655.21	0.86	31.10	1.0
3	Activator	2574	600.51	0.88	28.12	1.0
4	Repressor	1988	589.29	0.85	22.63	1.0
5	RNA-binding	2728	575.76	0.76	16.62	1.0
6	Chromatin regulator	1038	847.06	0.90	13.91	1.0
7	Ribonucleoprotein	1886	272.29	0.60	13.39	1.0
8	Serine/threonine-protein kinase	1782	802.24	0.84	11.56	1.0
9	Chaperone	937	430.43	0.71	10.02	1.0
10	Ribosomal protein	1408	186.38	0.53	9.34	1.0
11	Growth factor	398	299.63	0.70	8.98	1.0
12	Protein kinase inhibitor	49	337.20	0.96	8.34	1.0
13	Calmodulin-binding	520	1229.00	0.90	7.57	1.0
14	Hormone	338	221.13	0.59	7.24	1.0
15	Cyclin	133	422.71	0.87	7.18	1.0
16	Signal transduction inhibitor	115	408.43	0.84	6.76	1.0
17	Guanine-nucleotide releasing factor	319	1144.39	0.96	6.40	1.0
18	GTPase activation	424	867.35	0.88	6.28	1.0
19	Growth factor binding	50	593.98	1.00	6.09	1.0
20	Neuropeptide	105	234.96	0.68	5.99	1.0
21	Potassium channel	191	621.52	0.85	5.10	1.0
22	Calcium channel	193	1397.77	0.93	5.09	1.0
23	Protein phosphatase inhibitor	64	352.86	0.81	5.07	1.0
24	Tyrosine-protein kinase	376	863.00	0.89	5.04	1.0
25	Mitogen	137	286.03	0.68	4.80	1.0
26	Vasoactive	46	267.00	0.76	4.29	1.0
27	Heparin-binding	221	650.97	0.73	3.88	1.0
28	Muscle protein	193	920.01	0.73	3.81	1.0
29	Actin-binding	837	974.92	0.77	3.80	0.999
30	Endorphin	12	246.50	1.00	3.49	1.0
31	Amphibian defense peptide	49	85.80	0.53	3.34	1.0
32	Helicase	739	1086.05	0.87	3.29	1.0
33	Opioid peptide	9	228.67	1.00	3.14	1.0

SLIKA 6.1: 20 statistički najznačajnijih **neuređenih** ključnih reči iz rada [28] upoređeno sa našom analizom po ključnim rečima nad CAFA3 podacima.

6.0.1 Poređenje neuređenih funkcija

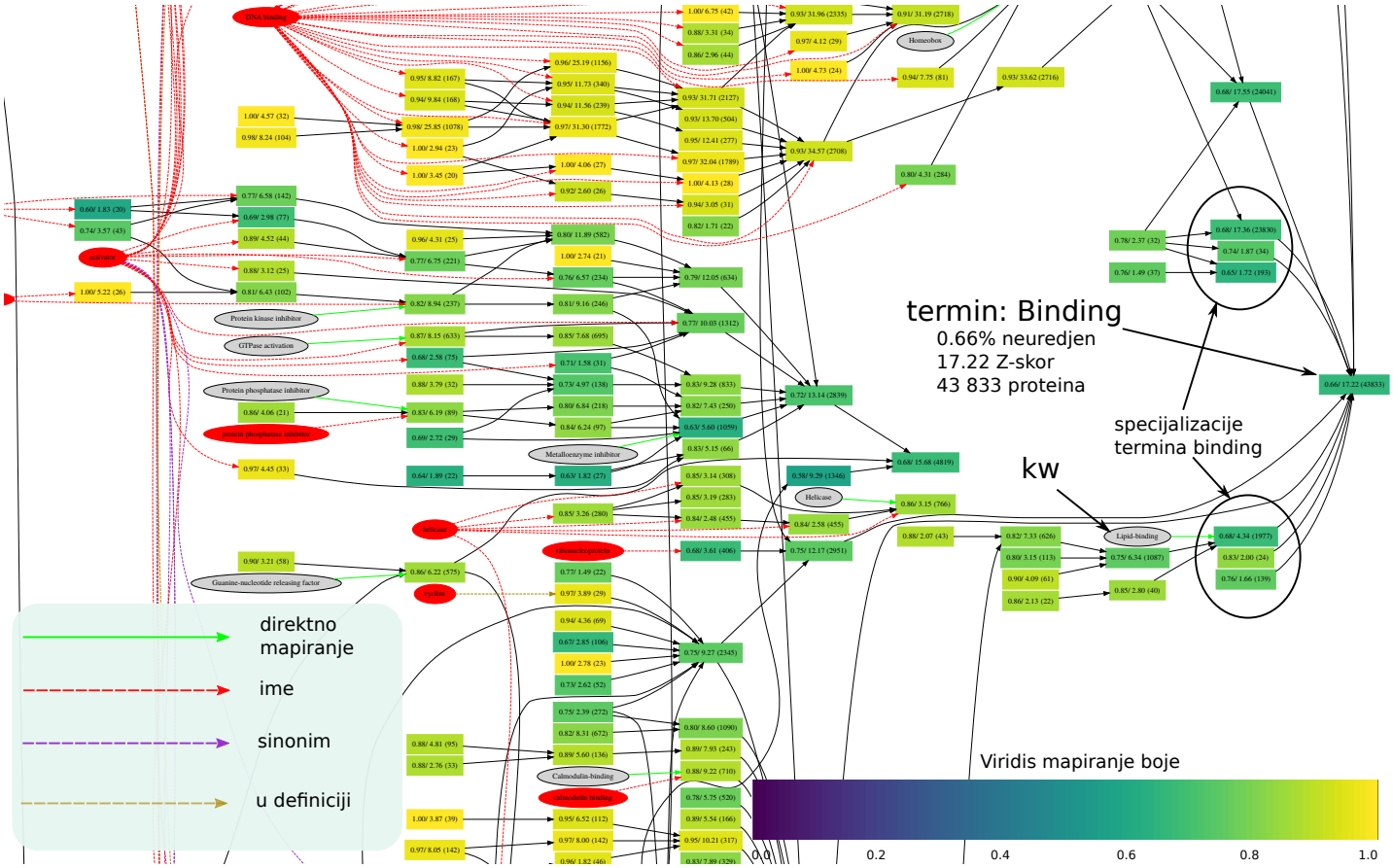
1. ***KW: Ribonucleoprotein*** nalazi se na 7. mestu, ima Z-skor 13.39 (skoro duplo manje) i neuređenosti 0.6. Direktno mapiranje nije nađeno. Pronađeni ***MF: Ribonucleoprotein binding complex*** ima Z-skor od svega 3.51, neuređenost 0.68 obuhvatajući svega 406 proteina.

2. ***KW: Ribosomal protein*** nalazi se na 10. mestu, ima Z-skor 9.34 (duplo manje) i neuređenost 0.53. Direkto mapiranje nije nađeno. Pronađen je sinonim u širem smislu, ***MF: structual costituent of ribosome*** sa skoro istim karakteristikam Z-skor 9.79 i neuređenost 0.53.

3. ***KW: Developmental protein*** nalazi se na 2. mestu, ima Z-skor 31.1 i neuređenost 0.87. **MF Mapiranje nije nađeno!**

4. ***KW: Hormone*** se nalazi na 14. mestu, ima Z-skor 7.24 (duplo manji) i neuređenost 0.59. Direktno mapiran ***MF: Hormone activity*** ima Z-skor 8.93 neuređenost 0.61. Dodatni relevantni MF termini uključuju:

- ***MF: Hormone receptor binding*** Z-skor 8.4, neuređenost 0.84
- ***MF: Steroid Hormone receptor binding*** Z-skor 9.4, neuređenost 0.92



SLIKA 6.2: Isečak grafa neuređenih termina

- ***MF: Nuclear Hormone receptor binding*** Z-skor 9.07, neuređenost 0.9
- ***MF: Thyroid Hormone receptor binding*** Z-skor 9.07, neuređenost 0.9
- ***MF: Nuclear Hormone receptor activity*** Z-skor 4.31, neuređenost 1.0

5. ***KW: Growth factor*** nalazi se na 11. mestu, ima Z-skor 8.98 i neuređenost 0.7. Direktno mapiran ***MF: Growth factor activity*** ima Z-skor 9.2 i neuređenost 0.7. Dodatna mapiranja uključuju:

- ***MF: Growth factor receptor binding*** ima Z-skor 3.76 i neuređenost 0.66. Postoji nekoliko specijalizacije ovog termina koje imaju veći Z-skor i neuređenost.
- ***MF: Growth factor binding*** ima Z-skor 4.42 i neuređenost 0.79. Postoji nekoliko specijalizacija visoke neuređenosti pogotovo za *insulin-like growth factor* vezivanje.

6. ***KW: Cytokine*** nije statistički značajan u našoj analizi ključnih reči. ***MF: Cytokine receptor binding*** ima Z-skor 2.57 (4 puta manje) i neuređenost 0.56

7. ***KW: Neuropeptide*** nalazi se na 20. mestu, ima Z-skor 6 i neređenost 0.68. Direktno mapiranje nije nađeno. ***MF: neuropeptide receptor binding*** ima Z-skor 6.71 i neuređenost 0.73 dok, ***MF: neuropeptide hormone activity*** ima Z-skor 6.06 i neuređenost 0.65.

8. ***KW: Activator*** nalazi se na 11. mestu, ima Z-skor 28.12 (3 put veći) i neuređenost 0.88.

Ekvivalentan MF termin ne postoji.

9. ***KW: GAP protein*** je izbačen iz analize jer sadrži manje od 20 proteina.

10. ***KW: Antigen*** je izbačena iz novih verzija ključnih reči. Analizom definicije ***MF: MHC class II protein binding*** zaključena je relavantnost u kontekstu antigena (iterakcija sa *histocompatibility* kompleksima). Z-skor 2.26, neuređenost 0.85. Ova funkcija sadrži svega 20 proteina.

11. ***KW: Repressor*** nalazi se na 4. mestu, ima Z-skor 22.63 (skoro 4 puta veća) i neuređenost 0.85.
Ekvivalentan MF termin ne postoji.

12. ***KW: Chromatin regulator*** nalazi se na 6. mestu, ima Z-skor 13.91 (duplo veća) i neuređenost 0.9.
Odgovarajući termin ***MF: covalent chromatin modification*** nema dovoljno pridruženih proteina.

13. ***KW: Pyrogen*** je izbačen iz analize jer sadrži manje od 20 proteina.

14. ***KW: Vasoactive*** nalazi se na 26. mestu, ima Z-skor 4.3 (orig. 5.56) i neuređenost 0.76.
Odgovarajući ***MF: regulation of blood vessel size*** nema dovoljno pridruženih proteina.

15. ***KW: Amphibian defense peptide*** nalazi se na 31. mestu, ima Z-skor 3.34 (orig. 5.44) i neuređenost 0.53.
Odgovarajući ***MF: defense response*** nema dovoljno pridruženih proteina.

16. ***KW: GTPase activation*** nalazi se na 18. mestu, ima Z-skor 6.28 (orig. 5.44) i neuređenost 0.88. Direktno mapiran ***MF: GTPase activation*** ima Z-skor 8.06, neuređenost 0.87.

17. ***KW: Endorphin*** nalazi se na 30. mestu, ima Z-skor 3.49 i neuređenost 1.0.
Odgovarajući ***MF: neuropeptide signaling pathway*** nema dovoljno pridruženih proteina. Sa druge strane vezivanje neuropeptida je pristuno

18. ***KW: Opioid peptide*** nalazi se na 33. mestu, ima Z-skor 3.14 i neuređenost 1.0. Direktno mapiran ***MF: Opioid peptide binding*** Nije statistički značajan.

19. ***KW: Protein phosphatase inhibitor*** nalazi se na 23. mestu, ima Z-skor 5.07 i neuređenost 0.81. Direktno mapiranje ***MF: Protein phosphatase inhibitor activity*** ima Z-skor 5.62 i neurđenost 0.83.

20. ***KW: Cyclin*** nalazi se na 15. mestu, ima Z-skor 7.18 i neuređenost 0.87. Direktno mapirnaje nije nađeno. ***MF: Cyclin binding*** ima Z-skor 2.042 i neuređenost 0.71.

6.0.2 Neuređene ključne reči značajne samo za CAFA3 podatke

1. ***KW: DNA-binding*** ima Z-skor 46.9 i neuređenost 0.87. Mapiran DNA binding ima Z-skor 46.13 i neurđenost 0.86. Pored ovog postoje mnogi specijalizovani neuređeni termini

5. ***KW: RNA-binding*** ima Z-skor 16.62 i neuređenost 0.76. Direktno mapiran RNA binding ima Z-skor 19.79 i neurđenost 0.73. Pored ovog postoje mnogi specijalizovani neuređeni termini

8. ***KW: Serine/threonine-protein kinase*** ima Z-skor 11.56 i neuređenost 0.84. Direktno mapiran proein serine/threonine kinase activity ima Z-skor 12.88 i neurđenost 0.84.

9. ***KW: Chaperone*** ima Z-skor 10.02 i neuređenost 0.71. Nema direktno mapiranje ali su pronađeni:

- ***MF: chaperone binding*** ima Z-skor 8.86 i neuređenost 0.84.
- ***MF: ufolded protein binding*** sinonima ima Z-skor 8.86 i neuređenost 0.84 (mapiran preko nekolio šest *related* sinonima).

12. ***KW: Protein kinase inhibitor*** ima Z-skor 8.34 i neuređenost 0.96. Direktno mapiran ***MF: protein kinase inhibitor activity*** ima Z-skor 8.94 i neuređenost 0.82.

13. ***KW: Calmodulin-binding*** ima Z-skor 7.57 i neuređenost 0.9. Direktno mapiran ***MF: calmodulin binding*** ima Z-skor 9.22 i neuređenost 0.88.

16. ***KW: Signal transduction inhibitor*** ima Z-skor 6.76 i neuređenost 0.84. Odgovarajući ***MF: negative regulation of signal transduction*** nema dovoljno pridruženih proteina.

17. ***KW: Guanine-nucleotide releasing factor*** ima Z-skor 6.4 i neuređenost 0.96. Direktno mapiran ***MF: guanyl-nucleotide exchange factor activity*** ima Z-skor 6.22 i neuređenost 0.86.

19. ***KW: Growth factor binding*** ima Z-skor 6.09 i neuređenost 1.00. ***MF: Growth factor binding*** ima Z-skor 4.42 i neuređenost 0.79. Postoji nekoliko specijalizacija visoke neuređenosti pogotovo za *insulin-like growth factor* vezivanje.

6.0.3 Poređenje uređenih funkcija

6.0.4 Neuređene ključne reči značajne samo za CAFA3 podatke

#	*Keywords*	*Number of proteins*	*Number of families*	*Average sequence length*	*Z-score*	*P-value*
1	Oxidoreductase	14995	992	376.63	−29.54	0
2	Transferase	26525	1606	445.17	−24.25	0
3	Lyase	7262	347	377.92	−22.64	0
4	Hydrolase	20464	1995	430.68	−21.75	0
5	Isomerase	4487	220	383.98	−14.18	0
6	Glycosidase	1826	244	444.73	−13.98	0
7	Glycosyltransferase	2950	261	437.53	−12.51	0
8	Acyltransferase	2239	179	402.83	−10.85	0
9	Methyltransferase	3524	224	349.60	−10.53	0
10	Kinase	7017	322	448.29	−10.22	0
11	Ligase	8010	230	529.41	−10.06	0
12	Decarboxylase	1293	63	345.26	−9.66	0
13	Monooxygenase	1668	73	444.87	−9.26	0
14	Metalloprotease	1100	109	553.73	−7.89	0
15	Aminopeptidase	452	39	509.17	−7.55	0
16	Dioxygenase	360	66	433.20	−7.32	0
17	Aminoacyl-tRNA synthetase	3402	37	571.83	−7.15	0
18	Protease	4423	380	549.70	−7.1	0
19	Aminotransferase	955	28	420.27	−6.02	0

#	*name*	*n*	*avg_len*	*avg_dis*	*z*	*p*
1	Oxidoreductase	4126	472.25	0.28	-41.13	0
2	Hydrolase	7564	614.81	0.51	-26.80	0
3	Lyase	1431	481.37	0.30	-23.83	0
4	Monooxygenase	555	503.36	0.20	-20.62	0
5	Transferase	8846	631.95	0.55	-19.54	0
6	Ligase	995	693.30	0.46	-17.42	0
7	Glycosyltransferase	1134	551.26	0.40	-16.83	0
8	Glycosidase	697	570.50	0.37	-16.62	0
9	Isomerase	931	422.72	0.35	-13.65	0
10	Transducer	1703	482.28	0.41	-12.99	0
11	Protease	1863	674.42	0.54	-12.90	0
12	G-protein coupled receptor	1385	465.62	0.39	-12.61	0
13	Acyltransferase	867	531.58	0.42	-11.72	0
14	Decarboxylase	195	488.21	0.25	-11.40	0
15	Aminotransferase	202	451.05	0.24	-10.36	0
16	Aminopeptidase	130	668.72	0.37	-9.24	0
17	Serine protease	460	700.07	0.50	-8.71	0
18	Methyltransferase	874	611.22	0.47	-8.65	0
19	Metalloprotease	507	688.25	0.56	-8.26	0
20	Carboxypeptidase	116	631.16	0.37	-8.11	0
21	Threonine protease	138	246.88	0.18	-7.35	0
22	Dioxygenase	366	622.32	0.48	-7.30	0

SLIKA 6.3: 20 statistički najznačajnijih **uređenih** ključnih reči iz rada [28] upoređeno sa našom analizom po ključnim rečima nad CAFA3 podacima.

6.0.5 P_L*random* model

#	*name*	*n*	*avg_len*	*avg_dis*	*z*	*p*
1	DNA-binding	6518	546.53	0.87	46.90	1
2	Developmental protein	3897	655.21	0.86	31.10	1
3	Activator	2574	600.51	0.88	28.12	1
4	Repressor	1988	589.29	0.85	22.63	1
5	RNA-binding	2728	575.76	0.76	16.62	1
6	Chromatin regulator	1038	847.06	0.90	13.91	1
7	Ribonucleoprotein	1886	272.29	0.60	13.39	1
8	Serine/threonine-protein ki...	1782	802.24	0.84	11.56	1
9	Chaperone	937	430.43	0.71	10.02	1
10	Ribosomal protein	1408	186.38	0.53	9.34	1
11	Growth factor	398	299.63	0.70	8.98	1
12	Protein kinase inhibitor	49	337.20	0.96	8.34	1
13	Calmodulin-binding	520	1229.00	0.90	7.57	1
14	Hormone	338	221.13	0.59	7.24	1
15	Cyclin	133	422.71	0.87	7.18	1
16	Signal transduction inhibitor	115	408.43	0.84	6.76	1
17	Guanine-nucleotide releasin...	319	1144.39	0.96	6.40	1
18	GTPase activation	424	867.35	0.88	6.28	1
19	Growth factor binding	50	593.98	1	6.09	1
20	Neuropeptide	105	234.96	0.68	5.99	1
21	Potassium channel	191	621.52	0.85	5.10	1
22	Calcium channel	193	1397.77	0.93	5.09	1
23	Protein phosphatase inhibitor	64	352.86	0.81	5.07	1
24	Tyrosine-protein kinase	376	863.00	0.89	5.04	1
25	Mitogen	137	286.03	0.68	4.80	1
26	Vasoactive	46	267.00	0.76	4.29	1
27	Heparin-binding	221	650.97	0.73	3.88	1
28	Muscle protein	193	920.01	0.73	3.81	1
29	Actin-binding	837	974.92	0.77	3.80	1
30	Amphibian defense peptide	49	85.80	0.53	3.34	1
31	Helicase	739	1086.05	0.87	3.29	1
32	Prion	22	497.05	0.91	3.06	1
33	Ion channel	1027	861.88	0.76	2.89	1
34	Voltage-gated channel	386	816.98	0.78	2.87	1
35	Viral nucleoprotein	39	1202.79	0.90	2.46	1
36	Tumor antigen	26	428.81	0.77	2.25	0.99
37	Exonuclease	239	725.99	0.75	2.19	0.99
38	Segmentation polarity protein	24	712.75	0.92	2.17	0.99
39	Motor protein	467	1227.67	0.83	1.92	0.97
40	Initiation factor	248	489.76	0.65	1.78	0.97
41	rRNA-binding	319	220.01	0.47	1.78	0.96
42	Topoisomerase	60	910.32	0.92	1.67	0.97
43	Protein synthesis inhibitor	36	419.33	0.69	1.64	0.97
44	RNA-directed DNA polymerase	62	1565.08	0.98	1.48	0.99
45	Aminoacyl-tRNA synthetase	243	673.71	0.67	-1.79	0.04
46	Antioxidant	111	203.40	0.34	-1.79	0.04
47	tRNA-binding	119	575.46	0.56	-2.01	0.03
48	Thiol protease inhibitor	63	411.06	0.35	-2.04	0.03
49	Hemostasis impairing toxin	55	297.55	0.35	-2.10	0.02
50	Aspartyl protease	120	942.30	0.64	-2.39	0.01
51	Voltage-gated sodium channe...	22	79.55	0.05	-2.42	0.01
52	Ion channel impairing toxin	60	74.92	0.15	-2.47	0
53	Nuclease	703	668.54	0.60	-2.49	0.01
54	Bacteriolytic enzyme	35	434.66	0.31	-2.53	0.01
55	Toxin	214	412.51	0.39	-2.56	0
56	Endonuclease	443	728.14	0.60	-2.59	0.01
57	Elongation factor	106	467.68	0.48	-2.59	0.01
58	Prenyltransferase	39	376.21	0.31	-2.59	0.01
59	Photoreceptor protein	77	540.68	0.47	-3.02	0.01
60	Myosin	185	1275.20	0.72	-3.29	0
61	Retinal protein	47	384.06	0.28	-3.37	0
62	Hemagglutinin	23	281.22	0.13	-3.37	0
63	Neurotoxin	66	209.62	0.17	-3.80	0
64	RNA-directed RNA polymerase	62	2506.11	0.87	-4.11	0
65	Protease inhibitor	284	446.29	0.41	-4.15	0
66	Integrin	71	1038.25	0.70	-4.29	0
67	Dipeptidase	22	522.27	0.23	-4.43	0
68	Porin	63	318.84	0.21	-4.55	0
69	Serine protease inhibitor	182	497.66	0.38	-4.92	0
70	Peroxidase	221	457.61	0.40	-5.50	0
71	Nucleotidyltransferase	600	969.68	0.63	-6.18	0
72	Receptor	3424	647.92	0.59	-7.09	0
73	Serine esterase	141	423.09	0.28	-7.21	0
74	Dioxygenase	366	622.32	0.48	-7.39	0
75	Threonine protease	138	246.88	0.18	-7.50	0
76	Carboxypeptidase	116	631.16	0.37	-8.25	0
77	Methyltransferase	874	611.22	0.47	-8.33	0
78	Metalloprotease	507	688.25	0.56	-8.43	0
79	Serine protease	460	700.07	0.50	-8.87	0
80	Aminopeptidase	130	668.72	0.37	-9.10	0
81	Aminotransferase	202	451.05	0.24	-10.23	0
82	Decarboxylase	195	488.21	0.25	-10.70	0
83	Acyltransferase	867	531.58	0.42	-11.28	0
84	G-protein coupled receptor	1385	465.62	0.39	-12.45	0
85	Transducer	1703	482.28	0.41	-12.56	0
86	Protease	1863	674.42	0.54	-13.20	0
87	Isomerase	931	422.72	0.35	-13.60	0
88	Glycosidase	697	570.50	0.37	-16.81	0
89	Glycosyltransferase	1134	551.26	0.40	-17.04	0
90	Ligase	995	693.30	0.46	-18.05	0
91	Transferase	8846	631.95	0.55	-19.72	0
92	Monooxygenase	555	503.36	0.20	-20.29	0
93	Lyase	1431	481.37	0.30	-23.62	0
94	Hydrolase	7564	614.81	0.51	-26.99	0
95	Oxidoreductase	4126	472.25	0.28	-41.35	0

#	*name*	*n*	*avg_len*	*avg_dis*	*z*	*p*
1	DNA-binding	6518	546.53	0.87	44.32	1
2	Developmental protein	3897	655.21	0.86	28.98	1
3	Activator	2574	600.51	0.88	27.75	1
4	Repressor	1988	589.29	0.85	22.04	1
5	Ribonucleoprotein	1886	272.29	0.60	18.90	1
6	RNA-binding	2728	575.76	0.76	17.90	1
7	Ribosomal protein	1408	186.38	0.53	15.70	1
8	Chromatin regulator	1038	847.06	0.90	12.76	1
9	Chaperone	937	430.43	0.71	10.68	1
10	Hormone	338	221.13	0.59	10.32	1
11	Growth factor	398	299.63	0.70	10.05	1
12	Protein kinase inhibitor	49	337.20	0.96	9.16	1
13	Serine/threonine-protein ki...	1782	802.24	0.84	9.16	1
14	Neuropeptide	105	234.96	0.68	8.31	1
15	Calmodulin-binding	520	1229.00	0.90	6.69	1
16	Amphibian defense peptide	49	85.80	0.53	6.53	1
17	Cyclin	133	422.71	0.87	6.48	1
18	Growth factor binding	50	593.98	1	6.46	1
19	Signal transduction inhibitor	115	408.43	0.84	6.45	1
20	Protein phosphatase inhibitor	64	352.86	0.81	6.10	1
21	Guanine-nucleotide releasin...	319	1144.39	0.96	5.62	1
22	Mitogen	137	286.03	0.68	5.56	1
23	GTPase activation	424	867.35	0.88	5.42	1
24	Vasoactive	46	267.00	0.76	5.21	1
25	Potassium channel	191	621.52	0.85	4.90	1
26	Antibiotic	270	152.33	0.39	4.82	1
27	Calcium channel	193	1397.77	0.93	4.52	1
28	Muscle protein	193	920.01	0.73	4.46	1
29	rRNA-binding	319	220.01	0.47	4.21	1
30	Heparin-binding	221	650.97	0.73	4.09	1
31	Fungicide	81	129.59	0.40	3.69	1
32	Tyrosine-protein kinase	376	863.00	0.89	3.68	1
33	Antimicrobial	349	191.07	0.38	3.55	1
34	Actin-binding	837	974.92	0.77	3.11	1
35	Prion	22	497.05	0.91	3.09	1
36	Defensin	53	81.09	0.30	2.60	0.99
37	Protein synthesis inhibitor	36	419.33	0.69	2.35	0.99
38	Viral nucleoprotein	39	1202.79	0.90	2.31	0.99
39	Voltage-gated channel	386	816.98	0.78	2.14	0.98
40	Tumor antigen	26	428.81	0.77	1.98	0.99
41	Motor protein	467	1227.67	0.83	1.85	0.97
42	Segmentation polarity protein	24	712.75	0.92	1.84	0.99
43	Cytokine	433	236.45	0.43	1.69	0.96
44	Metalloenzyme inhibitor	34	303.21	0.59	1.52	0.95
45	Metalloprotease inhibitor	32	264.50	0.56	1.44	0.95
46	Topoisomerase	60	910.32	0.92	1.39	0.95
47	Neurotoxin	66	209.62	0.17	-2.01	0.03
48	tRNA-binding	119	575.46	0.56	-2.21	0.02
49	Bacteriolytic enzyme	35	434.66	0.31	-2.24	0.02
50	Kinase	3072	730.20	0.71	-2.57	0.01
51	Endonuclease	443	728.14	0.60	-2.85	0
52	Nuclease	703	668.54	0.60	-2.97	0
53	Aminoacyl-tRNA synthetase	243	673.71	0.67	-3.11	0
54	Prenyltransferase	39	376.21	0.31	-3.22	0
55	Elongation factor	106	467.68	0.48	-3.24	0
56	Hemagglutinin	23	281.22	0.13	-3.28	0
57	Protease inhibitor	284	446.29	0.41	-3.43	0
58	Aspartyl protease	120	942.30	0.64	-3.56	0
59	Myosin	185	1275.20	0.72	-3.79	0
60	Photoreceptor protein	77	540.68	0.47	-4.01	0
61	Retinal protein	47	384.06	0.28	-4.24	0
62	Dipeptidase	22	522.27	0.23	-4.76	0
63	Serine protease inhibitor	182	497.66	0.38	-5.07	0
64	Porin	63	318.84	0.21	-5.21	0
65	Integrin	71	1038.25	0.70	-5.45	0
66	RNA-directed RNA polymerase	62	2506.11	0.87	-5.89	0
67	Peroxidase	221	457.61	0.40	-6.34	0
68	Threonine protease	138	246.88	0.18	-6.39	0
69	Serine esterase	141	423.09	0.28	-8.09	0
70	Nucleotidyltransferase	600	969.68	0.63	-8.50	0
71	Dioxygenase	366	622.32	0.48	-8.64	0
72	Carboxypeptidase	116	631.16	0.37	-9.18	0
73	Serine protease	460	700.07	0.50	-9.55	0
74	Metalloprotease	507	688.25	0.56	-10.40	0
75	Methyltransferase	874	611.22	0.47	-10.41	0
76	Aminopeptidase	130	668.72	0.37	-10.59	0
77	Receptor	3424	647.92	0.59	-11.81	0
78	Aminotransferase	202	451.05	0.24	-11.95	0
79	Acyltransferase	867	531.58	0.42	-13.00	0
80	Decarboxylase	195	488.21	0.25	-13.02	0
81	Isomerase	931	422.72	0.35	-15.23	0
82	Protease	1863	674.42	0.54	-16.00	0
83	Transducer	1703	482.28	0.41	-18.09	0
84	G-protein coupled receptor	1385	465.62	0.39	-18.15	0
85	Glycosidase	697	570.50	0.37	-19.18	0
86	Ligase	995	693.30	0.46	-20.02	0
87	Glycosyltransferase	1134	551.26	0.40	-21.11	0
88	Monooxygenase	555	503.36	0.20	-24.45	0
89	Lyase	1431	481.37	0.30	-25.90	0
90	Transferase	8846	631.95	0.55	-27.02	0
91	Hydrolase	7564	614.81	0.51	-32.98	0
92	Oxidoreductase	4126	472.25	0.28	-46.56	0

SLIKA 6.4: P_L levo, upoređen sa $P_L random$ desno dobijeno iz CAFA3 proteina grupisanih po ključnim rečima.

Bibliografija

[1] Vladimir N. Uversky. **?**Dancing Protein Clouds: The Strange Biology and Chaotic Physics of Intrinsically Disordered Proteins**?** U: *Journal of Biological Chemistry* 291.13 (2016), str. 6681–6688. DOI: 10.1074/jbc.r115.685859. URL: https://doi.org/10.1074/jbc.r115.685859.

[2] Vladimir N. Uversky, Christopher J. Oldfield i A. Keith Dunker. **?**Intrinsically Disordered Proteins in Human Diseases: Introducing the D2 Concept**?** U: *Annual Review of Biophysics* 37.1 (2008), str. 215–246. DOI: 10.1146/annurev.biophys.37.032807.125924.

[3] P. R. Romero i dr. **?**Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms**?** U: *Proceedings of the National Academy of Sciences* 103.22 (2006), str. 8390–8395. DOI: 10.1073/pnas.0507916103.

[4] E.N Trifonov. **?**Consensus temporal order of amino acids and evolution of the triplet code**?** U: *Gene* 261.1 (2000), str. 139–151. DOI: 10.1016/s0378-1119(00)00476-5. URL: https://doi.org/10.1016/s0378-1119(00)00476-5.

[5] Christopher J. Oldfield i A. Keith Dunker. **?**Intrinsically Disordered Proteins and Intrinsically Disordered Protein Regions**?** U: *Annual Review of Biochemistry* 83.1 (2014), str. 553–584. DOI: 10.1146/annurev-biochem-072711-164947. URL: https://doi.org/10.1146/annurev-biochem-072711-164947.

[6] A.Keith Dunker i dr. **?**Intrinsically disordered protein**?** U: *Journal of Molecular Graphics and Modelling* 19.1 (2001), str. 26–59. DOI: 10.1016/s1093-3263(00)00138-8. URL: https://doi.org/10.1016/s1093-3263(00)00138-8.

[7] Alfred Ezra Mirsky i Linus Pauling. **?**On the Structure of Native, Denatured, and Coagulated Proteins**?** U: *Proceedings of the National Academy of Sciences of the United States of America* 22.7 (1936), str. 439–447. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1076802/.

[8] Burton S. Guttman. **?**Biology**?** U: (1998), str. 66–107. URL: https://www.amazon.com/Biology-Burton-S-Guttman/dp/0697223663.

[9] Zhenling Peng i dr. **?**Exceptionally abundant exceptions: comprehensive characterization of intrinsic disorder in all domains of life**?** U: *Cellular and Molecular Life Sciences* 72.1 (2014), str. 137–151. DOI: 10.1007/s00018-014-1661-9. URL: https://doi.org/10.1007/s00018-014-1661-9.

[10] Bin Xue, A. Keith Dunker i Vladimir N. Uversky. **?**Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life**?** U: *Journal of Biomolecular Structure and Dynamics* 30.2 (2012), str. 137–149. DOI: 10.1080/07391102.2012.675145. URL: https://doi.org/10.1080/07391102.2012.675145.

[11] Peter Tompa. **?**Intrinsically unstructured proteins**?** U: *Trends in Biochemical Sciences* 27.10 (2002), str. 527–533. DOI: 10.1016/s0968-0004(02)02169-2. URL: https://doi.org/10.1016/s0968-0004(02)02169-2.

[12] Rebecca B. Berlow i Peter E. Wright. **?**Tight complexes from disordered proteins**?** U: (2018). DOI: doi:10.1038/d41586-018-01694-y. URL: https://www.nature.com/articles/d41586-018-01694-y.

[13] Vladimir N. Uversky. **?**Intrinsically disordered proteins from A to Z**?** U: *The International Journal of Biochemistry & Cell Biology* 43.8 (2011), str. 1090–1103. DOI: 10.1016/j.biocel.2011.04.001. URL: https://doi.org/10.1016/j.biocel.2011.04.001.

[14] Fanchi Meng, Vladimir N. Uversky i Lukasz Kurgan. **?**Computational Prediction of Intrinsic Disorder in Proteins**?** U: *Current Protocols in Protein Science* (2017), str. 2.16.1–2.16.14. DOI: 10.1002/cpps.28.

[15] Fanchi Meng, Vladimir N. Uversky i Lukasz Kurgan. **?**Comprehensive review of methods for prediction of intrinsic disorder and its molecular functions**?** U: *Cellular and Molecular Life Sciences* 74.17 (2017), str. 3069–3090. DOI: 10.1007/s00018-017-2555-4. URL: https://doi.org/10.1007/s00018-017-2555-4.

[16] Bo He i dr. **?**Predicting intrinsic disorder in proteins: an overview**?** U: *Cell Research* 19.8 (2009), str. 929–949. DOI: 10.1038/cr.2009.87. URL: https://doi.org/10.1038/cr.2009.87.

[17] *database summary 2015*. https://proteininformationresource.org/staff/chenc/MiMB/dbSummary2015.html. Pristupljeno: 13.02.2018.

[18] Chuming Chen, Hongzhan Huang i Cathy H. Wu. **?**Protein Bioinformatics Databases and Resources**?** U: (2017), str. 3–39. DOI: 10.1007/978-1-4939-6783-4_1.

[19] GO Consortium. **?**Expansion of the Gene Ontology knowledgebase and resources**?** U: *Nucleic Acids Research* 45.D1 (2016), str. D331–D338. DOI: 10.1093/nar/gkw1108. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5210579.

[20] Michael Ashburner i dr. **?**Gene Ontology: tool for the unification of biology**?** U: *Nature Genetics* 25.1 (2000), str. 25–29. DOI: 10.1038/75556. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3037419.

[21] *Ontology Structure*. http://geneontology.org/page/ontology-structure. Pristupljeno: 22.02.2018.

[22] *Ontology Relations*. http://geneontology.org/page/ontology-relations#isa_reas. Pristupljeno: 22.02.2018.

[23] *Molecular Function Ontology Guidelines*. http://geneontology.org/page/molecular-function-ontology-guidelines. Pristupljeno: 22.02.2018.

[24] B. Boeckmann. **?**The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003**?** U: *Nucleic Acids Research* 31.1 (2003), str. 365–370. DOI: 10.1093/nar/gkg095. URL: https://doi.org/10.1093/nar/gkg095.

[25] *link*. ftp://ftp.uniprot.org/pub/databases/uniprot/previous_releases/release-2017_12/knowledgebase.

[26] *UniProtKB manual*. https://www.uniprot.org/help/?query=*&fil=section:manual. Pristupljeno: 13.02.2018.

[27] *How redundant are the UniProt databases?* http://www.uniprot.org/help/redundancy. Pristupljeno: 13.12.2017.

[28] Hongbo Xie i dr. **?**Functional Anthology of Intrinsic Disorder. 1. Biological Processes and Functions of Proteins with Long Disordered Regions**?** U: *Journal of Proteome Research* 6.5 (2007), str. 1882–1898. DOI: 10.1021/pr060392u. URL: https://www.ncbi.nlm.nih.gov/pubmed/17391014.

[29] *CAFA*. http://biofunctionprediction.org/cafa/. Pristupljeno: 13.12.2017.

[30] *go.obo*. http://purl.obolibrary.org/obo/go.obo. Pristupljeno: 01.12.2017.

[31] *keywlist.txt*. www.uniprot.org/docs/keywlist.txt. Pristupljeno: 20.12.2017.

[32] *uniprotkb_kw2go*. http://geneontology.org/external2go/uniprotkb_kw2go. Pristupljeno: 20.12.2017.

[33] D. Barrell i dr. **?**The GOA database in 2009–an integrated Gene Ontology Annotation resource**?** U: *Nucleic Acids Research* 37.Database (2009), str. D396–D403. DOI: 10.1093/nar/gkn803.

[34] Damiano Piovesan i dr. **?**DisProt**?** U: 45.D1 (2016), str. D219–D227. DOI: 10.1093/nar/gkw1056. URL: https://doi.org/10.1093/nar/gkw1056.

[35] *Amino acid frequency*. http://www.tiem.utk.edu/~gross/bioed/webmodules/aminoacid.htm. Pristupljeno: 13.13.2017.

[36] *viridis*. https://matplotlib.org/users/colormaps.html. Pristupljeno: 18.03.2018.

[37] *rezultati*. https://github.com/gvinterhalter/MASTER2/tree/master/data/OUT/.