UNIVERZITET U BEOGRADU
MATEMATIČKI FAKULTET

MASTER RAD

Bioinformatička 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

Bioinformatička analiza povezanosti funkcije i neuređenosti proteina

Goran VINTERHALTER

Korelacija između molekulske funkcije proteina i inherentne neuređenosti predstavlja bitan aspekt izučavanja odnosa između funkcije, sekvence i strukture. Ovo istraživanje je inspirisano statističkom metodom za ocenu korelacije predloženom od strane *Xie et al.* [1] gde autori izučavaju odnos između strukture i funkcije proteina iz *Swiss-Prot* baze, a funkcije su opisane *Swiss-Prot* ključnim rečima. U ovom istraživanju, korišćen je drugi skup proteina (trening skup sa CAFA3 takmičenja), a funkcije su opisane terminima molekulske funkcije iz ontologije gena (GO). Rezultati su upoređeni sa originalnim radom tako što je analiza prvo ponovljena sa *Swiss-Prot* ključnim rečima, a zatim sa GO terminima. Prediktor PONDR VSL2b korišćen je za karakterizaciju preko 66000 CAFA3 proteina kao putativno uređenih ili neuređenih, dok su funkcijske anotacije (ključne reči i GO termini) preuzete iz *Swiss-Prot* baze. Od 186 ključnih reči kategorije molekulske funkcije sa minimum 20 anotiranih proteina, utvrđeno je da su 53 korelirane sa uređenim proteinima, a 44 sa neuređenim. Pod istim uslovima, od od 1781 GO termina molekulske funkcije 699 je korelirano sa uređenim proteinima, a 616 sa neuređenim. Rezultati GO termina predstavljeni su kao interaktivni grafovi koji prikazuju kompleksnu hijerarhijsku strukturu ontologije gena. Komparacija dve funkcijske nomenklature, GO i ključne reči, pokazala je konzistentne rezultate za adekvatna mapiranja. Međutim, komparacija originalnih i novih rezultata otkrila je da funkcije koje opisuju molekulsko vezivanje preovlađuju među novim rezultatima (neuređenih funkcija) dok se u starim rezultatima ne javljaju. Zbog malog broja poznatih veza u mapiranju između ključnih reči i GO termina, predložena je nova metoda za izvođenje najverovatnijih nedostajućih veza tako što se verovatnoća ocenjuje sličnošću (Žakardov indeks) između skupova proteina anotiranih različitim funkcijama. Takođe, pokazano je da se korelacija između dužine proteina i putativne neuređenosti može aproksimirati slučajno generisanim proteinskim sekvencama.

Sadržaj

Glava 1

Osnovni pojmovi

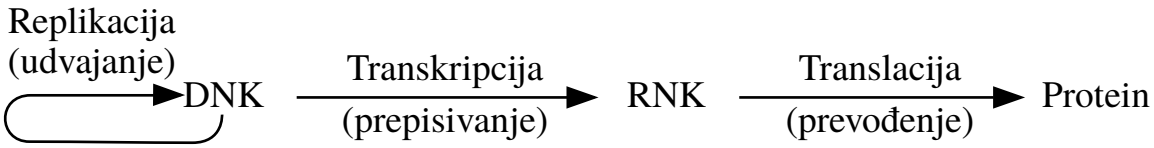
Svi živi organizmi sastoje se od jedne ili više ćelija, a svaka ćelija od molekula. Veliki[1] molekuli (makromolekuli) organskog porekla obično[2] su sačinjeni od ponavljajućih strukturnih jedinica **monomera** *(mono- = jedan, mer- = deo)* međusobno povezanih **kovalentnim** vezama. Takav molekul zovemo **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.

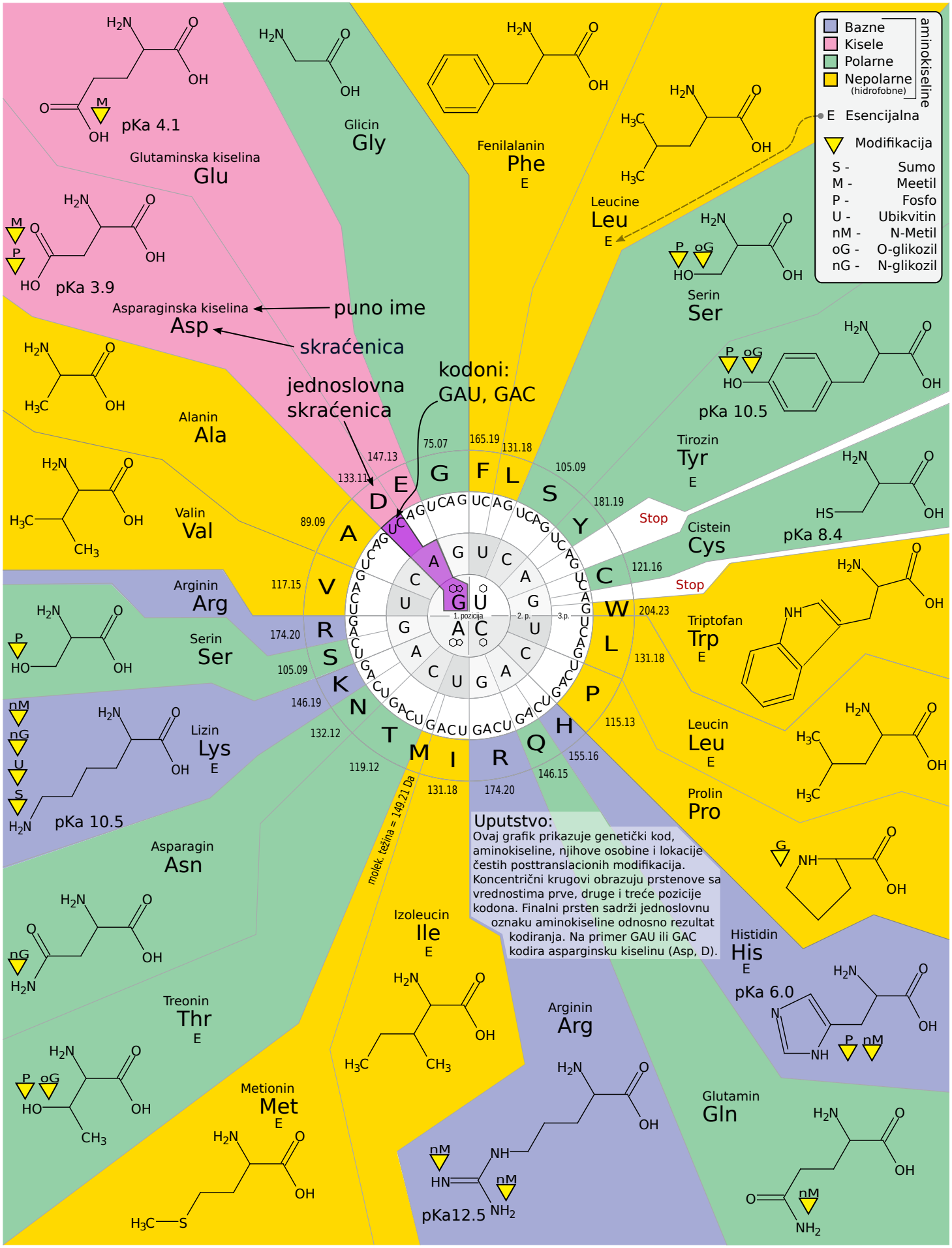


SLIKA 1.1: Centralna dogma molekularne biologije

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 arhea, bakterija ali i mitohondrija.

Gen je region DNK sekvence koji biva transkribovana u RNK. Neki geni su tzv. **nekodirajući geni** i transkribuju se u funkcionalne 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 Frensis 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 može 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 (poravnanje) 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 sadrži 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]. Do danas, otkrivene su 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). Pod proteinom se podrazumeva polipeptid velike dužine. Različiti izvori navode različite granice, na primer: 20-30 AK, 50 AK ili 100 AK, ali zapravo ne postoji jasna granica između polipeptida i proteina [2]. 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).

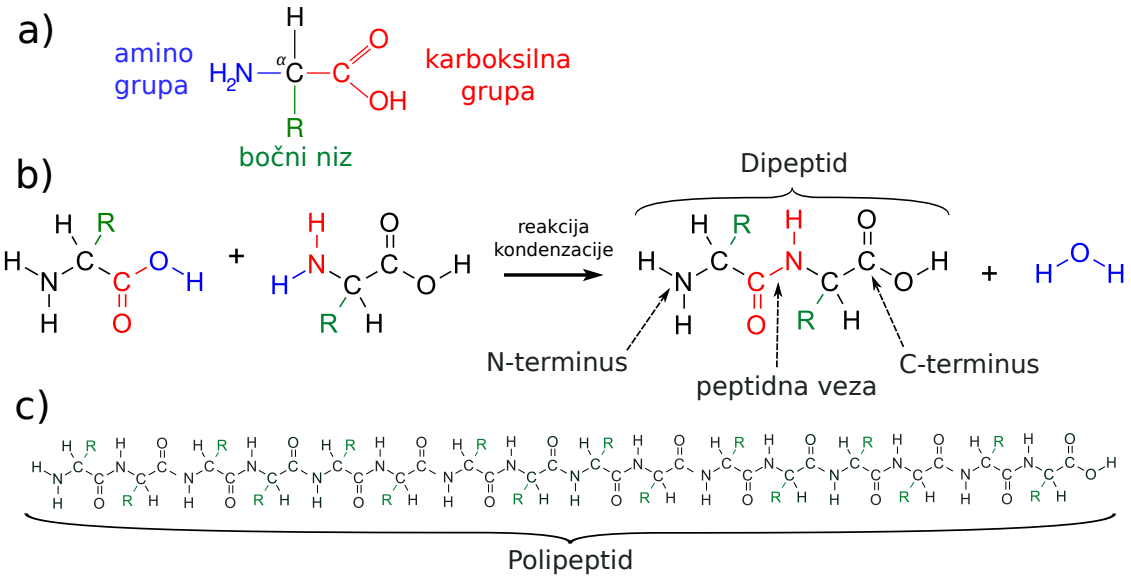
Prema fizičkim i hemijskim osobinama bočnog niza, aminokiseline se mogu klasifikovati na nekoliko načina prikazanih Slikama 1.2 i 1.4 od čega izdvajamo sledeću klasifikaciju:

- 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 - rastvorljive u vodi (vole vodu). Uglavnom se nalaze na spoljnim delovima proteina, često na hemijski aktivnim delovima.
- Naelektrisane polarne AK - veoma hidrofilne. Dodatno se dele na pozitivno i negativno naelektrisane.
- Aromatične AK - najveće i najtežim jer u bočnom nizu sadrže aromatični prsten (jedan ili više) od ugljenikovih atoma.

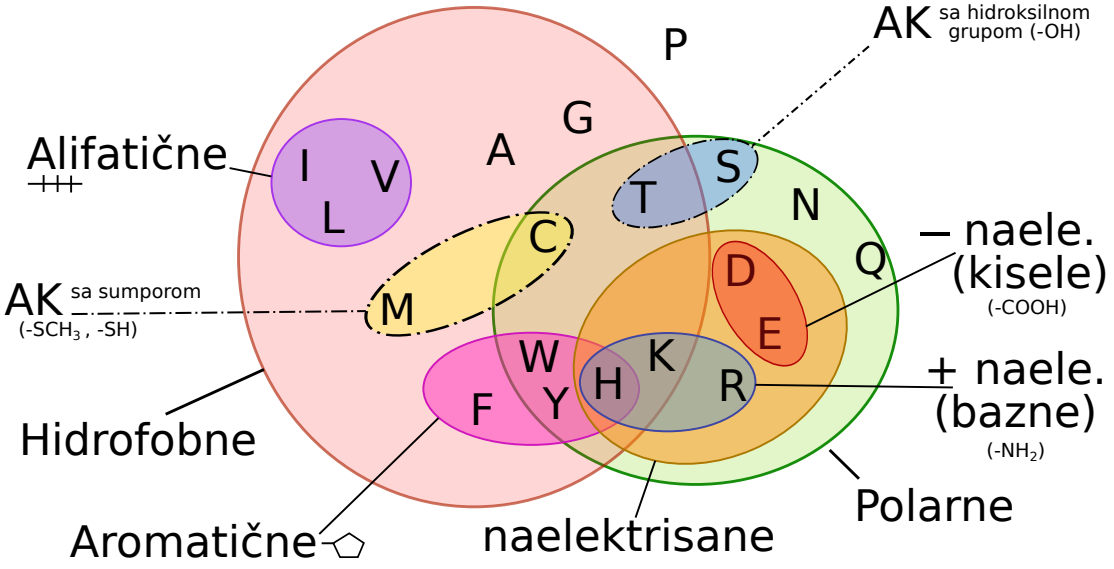
Za vreme života proteina (nakon translacije) aminokiseline koje ga čine mogu biti modifikovane od strane raznih enzima, na primer pravljenjem kovalentnih veza sa novim funkcionalnim grupama. Ovaj proces poznat je kao **posttranslaciona modifikacija**, kraće **PTM**. Na Slici 1.2 prikazane su najčešće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

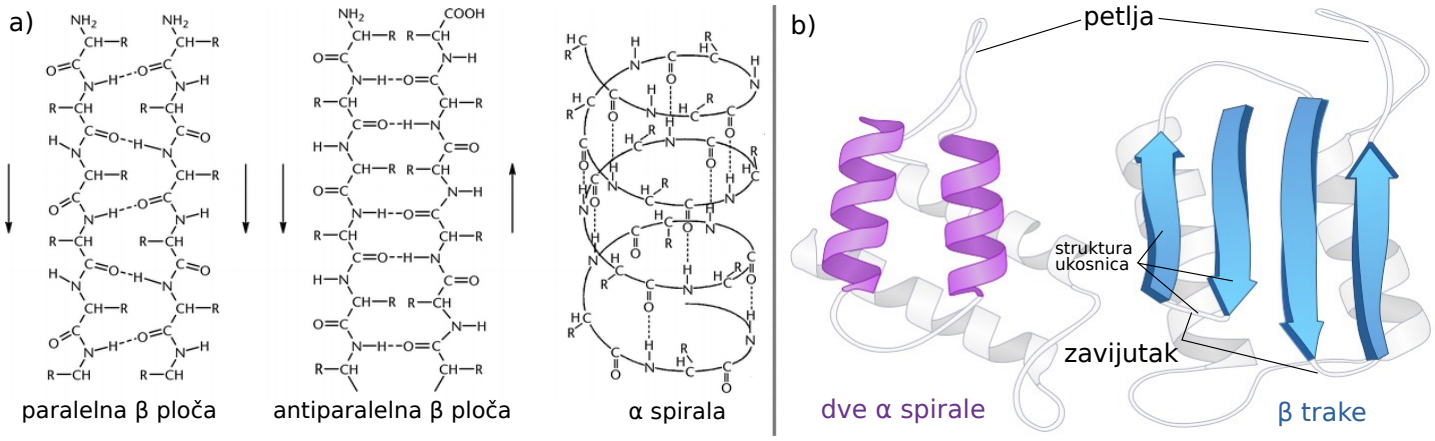
1.3.2 Struktura proteina

Protein je sačinjen iz jednog (monomerni) ili više (oligomerni) polipeptidnih 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.

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



SLIKA 1.5: Sekundarna struktura α-spirala i β-traka. Levi deo slike preuzet je sa [3]. Desni deo slike preuzet je sa [4].

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, β-ukosnica (engl. *β-hairpin*) je kratak zaokret lanca kod antiparalelne β-trake. Formiranje sekundarne strukture naziva se **lokalno savijanje**.

Usled deljenja elektrona, peptidna veza ponaša se slično dvostrukoj kovalentnoj vezi i onemogućava rotaciju ograničavanjem mogućih konformacija pp lanca. Ovakvo ponašanje dozvoljava kreiranje pravilnih sekundarnih struktura dodatnim ograničavanjem vrednosti tzv. dihedralnih uglova ψ i ϕ prikazanih na Slici 1.6a. Sekundarne strukture (α-spirala i β-traka) imaju specifične kombinacije vrednosti ψ i ϕ uglova. Ove kombinacije ilustrovane su skicom Ramačandranovog dijagrama (Slika 1.6b) koji se dobija kao šema raspršenih elemenata (engl. *scatter plot*) svih ψ i ϕ vrednosti jednog polipeptida.

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

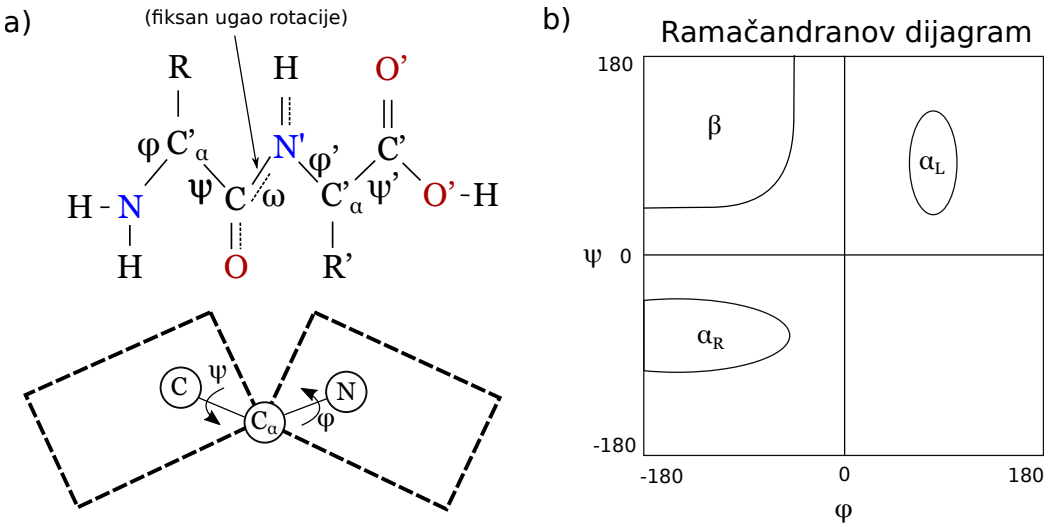
Tercijarna struktura predstavlja prostorni oblik koji protein zauzima indukovan interakcijama bočnih nizova aminokiselina. Ovaj korak poznat je kao **savijanje** (engl. *folding*). Pod poznavanjem tercijarne strukture podrazumeva se opisivanje

[5] Disulfidni most se formira izvan ćelije, nakon što je protein već savijen, i služi stabilizaciji 3D strukture.

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 **strukturni motivi**, kraće **motiv** i okarakterisani su velikom sličnošću primarne strukture. Promena okruženja (ph, temperature) dovodi do promene tercijarne strukture ili potpunog gubitka strukture (denaturacija).



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) Skica ramačandranovog dijagram, prikazuje najčešću raspodelu vrednosti uglova i sekundarne strukture koje oni obrazuju. (Dijagrami su preuzeti iz [5])

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 **supstrat**. Mesto enzima koje interaguje sa supstratom 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.

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

1.3.4 Funkcija proteina

Primeri uloga koje protein može da ima uključuju: formiranje strukture, zaštitna uloga (tzv. antigeni), transport, katalizacija hemijskih reakcija (enzimi), regulacija procesa u ćeliji itd. Postoji nekoliko sistema (nomenklatura) za razdvajanje proteina po funkciji. Na primer, specijalno za enzime postoji numerička klasifikacija po hemijskoj reakciji koju katalizuju. Pomenuta klasifikacija klasteruje enzime u hijerarhiju podskupova. Ovaj pristup nije adekvatan 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 primer 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.

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 [6]. Danas je neuređenost proteina uzrokovala nastanak velikog broj hipoteza, od D^2 koncepta bolesti [7] pa sve do evolucije višećelijskih organizama [8] i osobina prvih oblika života [6, 9]. Zajedno sa početkom 21. veka broj naučnih radova koji se bave ovom temom doživljava skoro eksponencijalan porast [10], 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 [11]. Ova paradigma predlaže da su funkcionalne (operativne) osobine proteina posledica njihovog jedinstvenog savijenog oblika (konformacije) tj. *"the characteristic specific properties of native[2] proteins we attribute to their uniquely defined configurations"* [12]. 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. Supstrat (ključ ili funkcija) diktira oblik enzima (brave ili strukture) [13]. 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 [11]. 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 napreci 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 [11]. Nova paradigma je bila neophodna.

Hipoteza proteinskog trojstva [11] (nastala tek početkom 21. veka) predlaže da funkcija proteina može zavisiti od bilo kog 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 [11].

[2] nativno stanje proteina je savijeno, operativno, funkcionalno stanje [11]. Ovaj termin bio je isprepleten sa paradigmom struktura-funkcija

[3] Teorija indukovanog fita omekšava rigidnost modela brava-ključ sugerišući da interakcija sa supstratom indukuje konačni oblik enzima maksimizujući reakciju [13]

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[11]) model je dopunjen.

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

Povezanost sekvence sa strukturom sugeriše da je neuređenost enkodirano inherentno[4] svojstvo [11] stoga ove proteine nazivamo **inherentno**[5] **neuređeni proteini** (engl. *Intrinsically Disorderd Proteins*) skraćeno **IDP**, a njihove neuređene ali funkcionalne regione **IDPr** [6]. U ovom radu pod neuređenošću proteina podrazumevaćemo inherentnu neuređenost osim ako to nije drugačije naglašeno[6].

Današnje procene zastupljenosti pokazuju da 19% aminokiselina kod eukariota, 6% kod bakterija i 4% kod arhea pripadaju IDPr [14]. Čak 50% proteina eukariota ima bar jedan IDPr duži ili jednak 30 uzastopnih AK [15] dok je za 6% do 17% predviđeno da su neuređeni celom dužinom [16]. 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, količina novih saznanja raste veoma brzo. Na primer, u časopisu *Nature* objavljen je rad [17] 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 ovo istraživanje.

- Neuređenost je inherentno svojstvo sekvence [11]. Pokazano je da nisko očekivanje indeksa hidropatije[7] zajedno sa visokim ukupnim naelektrisanjem predstavlja bitan preduslov koji sprečava savijanje proteina u fiziološkim uslovima [6]. 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*). [6, 10]. Opisane osobine daju validnost primeni mašinskog učenja u predviđanju neuređenih regiona proteina [10].

- 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 [6]. Iako su PTM povezane sa neuređenošću i sugeriše veliki uticaj na funkciju proteina [6], 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

[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 (mera kvaliteta informacije poistovećena sa brojem bita potrebnih da je kodiraju)

raznolikim [6]. Pomenuta interakcija kod nekih neuređenih proteina vodi do njihovog potpunog ili parcijalnog savijanja dok neki i dalje ostaju neuređeni [6]. 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 [10, 11] iako su rani pokušaji napravljeni u radu [11]. Najopštiji opis strukture ovih proteina dat je kao **kombinacija različitih tipova foldona**[9] [6]:
  - **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 izvršavanja funkcije.
- Gore pomenut opšti prikaz strukture nastao je iz raznih opažanja interakcije, prvenstveno vezivanja proteina za partnere. Detaljan opis i iscrpna lista ovih i drugih pojava može se naći u radovima [6, 18, 19].

2.2 Eksperimentalno ispitivanje neuređenosti

Postoji veliki broj eksperimentalnih metoda za karakterizaciju strukture i osobina proteina. Svaka od njih ima prednosti, mane i nivo pouzdanosti. Da bi se protein potpuno okarakterisao korisno je sagledati rezultate nekoliko eksperimentalnih metoda. Isto važi i za karakterizaciju neuređenih regiona proteina. ***DisProt*** [20] baza neuređenih proteina verzija 7 na svojoj veb stranici nabraja čak 36 eksperimentalnih tehnika. Eksperimentalne tehnike koje se najčešće koriste za karakterizaciju neuređenosti proteina su [11]:

- Kristalografija X zracima (engl. *X-ray crystallography*)
- Spektrosokopija Nuklearnom Magnetnom Rezonancom (NMR) (engl. *NMR spectroscopy*)
- Cirkularni dihroizam (engl. *Circular dichroism (CD) spectroscopy*)
- Senzitivnost na proteolizu (engl. *Sensitivity to proteolysis*)

2.3 Predikcija neuređenosti

Do danas napravljeno je preko 60 prediktora inherentno neuređenih proteina [21]. Prediktor u kontekstu proteina je program koji računarskim metodama predviđa osobine proteina. Primer računarske metode predstavljaju tehnike **mašinskog učenja** (skraćeno ML). U radu [22] hronološkim redosledom prikazane su karakteristike i dostupnost tridesetak popularnih prediktora neuređenosti.

Istorijski posmatrano razlikujemo tri epohe razvoja [22]:

[9] Zbog nove prirode termina i manjka prevedene literature autor je odlučio da usvoji naziv u originalu.

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

- Druga generacija (2002-2006)
  Ovaj period okarakterisan je korišćenjem relativno jednostavnih prediktivnih modela: prediktore isključivo zasnovane na osobinama AK sekvence ali i popularne ML metode. Kao ulaz, pored sekvence, neke metode su podržavale i evolutivne profile.

- 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.

2.3.1 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 neuređenosti) 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 [23]. Međutim, po današnjim merilima [21] 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[11] minimalne dužine 9 AK kodiranih jednim karakterom i podržava azbuku od 20 standardnih AK. Rezultat predikcije je niz 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.

[10] Nakon 1979. godine drugi prediktor nastao je tek 1997. [22]

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

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 uraditi 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. Veliku 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 obavljena kombinacijom informacija nekoliko različitih baza. Zbog raznovrsnosti i svrhe prikupljenih informacija postoji veliki broj kategorija[2](vrsta) baza. Na adresi [24] autori Čen, Huang i Vu kategorizovali su i prikazali novije, javno dostupne i visoko kvalitetne proteinski orijentisane baze podataka (prikazana lista nije iscrpna) [25]. 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*

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

[2] Baze podataka ne pripadaju ekskluzivno samo jednoj kategoriji

  - DNK sekvence: EMBL-Bank, GenBank, DDBJ, ...

- 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) [26]. 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.

Tvorci ontologije gena zasnovali su ovu nomenklatoru na opažanju da različiti organizmi (pogotovo eukarioti) dele veliki broj ortologih gena. Većina uočenih konzerviranih funkcija (ortologih gena) ispostavila se neophodnom za osnovne biološke procese bilo kog živog organizma. Iz ovog opažanja rodila se ideja o definisanju jednog skupa termina koji će opisivati genske proizvode svih vrsta organizama, ontologija gena [27].

3.1.1 GO termin

Skup GO termina se stalno menja. 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 skupa 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 stranama i bazama podataka često vezanih uz definiciju. Uz GO termin obično se navode sinonimi koji odgovaraju imenu termina, ali se razlikuju po opsegu:

[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*)

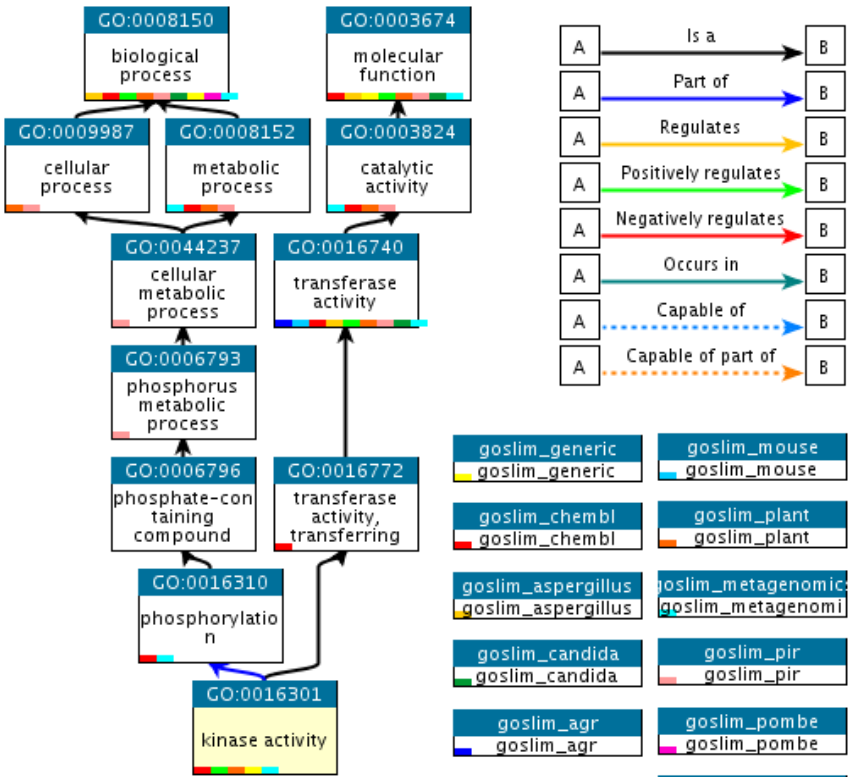
[5] Uglavnom ovi podskupovi predstavljaju model organizme

- *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

3.1.2 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. *Directed Acyclic Graph, 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 ontologije sa korenim čvorovima: MF, CC i BP [28]. 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 [29])

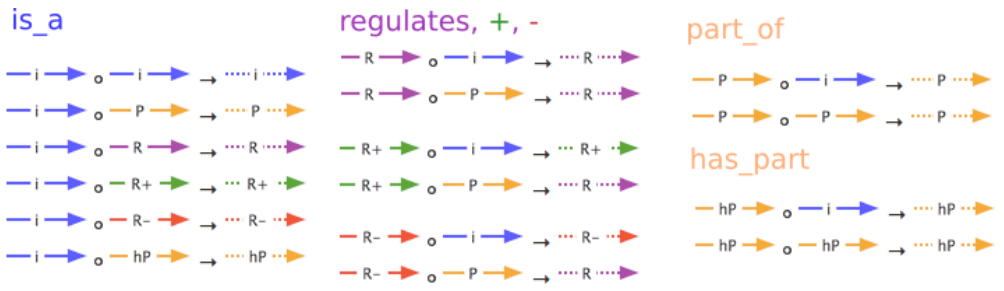
Vremenom se GO ontologija proširuje novim tipovima relacije koje su van okvira ovog rada. Svaka relacija ima strogo definisana pravila kompozicije koja omogućavaju automatsko rezonovanje. Na primer, relacija **is_a** ima svojstvo tranzitivnosti [30] prikazano Slikom 3.2:

$$\begin{aligned} &A\ \texttt{is\_a}\ B \ \wedge\ B\ \texttt{is\_A}\ C \implies A\ \texttt{is\_a}\ C \\ &A\ \texttt{part\_of}\ B \ \wedge\ B\ \texttt{is\_A}\ C \implies A\ \texttt{part\_of}\ C \end{aligned}$$

SLIKA 3.2: Tranzitivnost relacije **is_a**

Siže pravila rezonovanja prikazan je na Slici 3.3.

Jedan od najčešće korišćenih formata je ravni `.obo` format, 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, ...).



SLIKA 3.3: Pravila rezonovanja (isprekidane relacije su rezultat).
Elementi slike su preuzeti sa [29].

3.1.3 Molekulska funkcija

Funkcija genskog produkta predstavlja posao koji on obavlja ili osobinu koju on ima. Razmotrimo narednu analogiju [31]. 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 [31]:

- 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 reakcije 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) [31]:

- x **binding** - interaguje 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 [31].

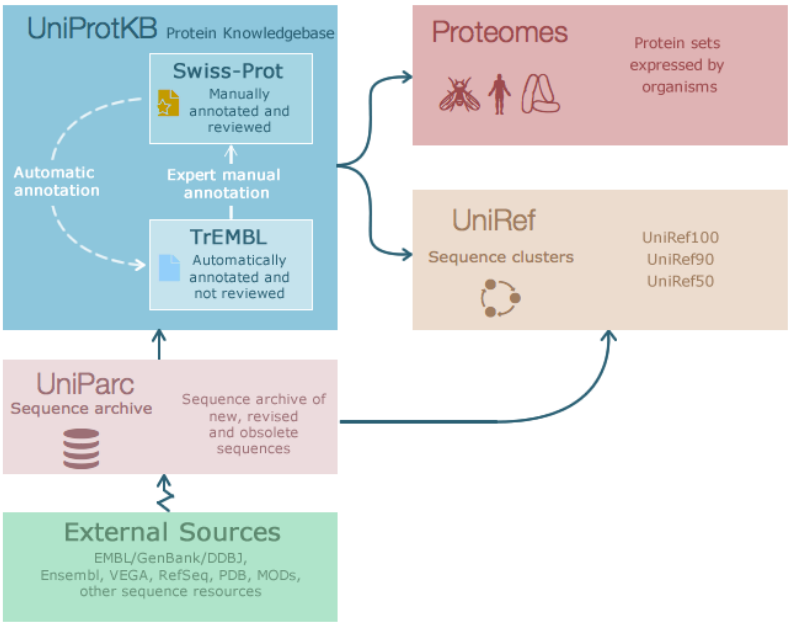
MF termini takođe imaju prepoznatljiv standardni sinonim za *x* **binding** [31]:

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

3.2 UniProtKB/Swiss-Prot

UniProt skraćeno od *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 [32])

1. ***Swiss-Prot*** sadrži visoko kvalitetne anotacije **neredundantnih** (termin je definisan u nastavku liste, stavka 6) proteinskih sekvenci. Informacije o sekvenci su dobijene iz postojeće literature, a automatski predviđene anotacije su ručno proverene i verifikovane od strane biokuratora (stručnjaci koji se brinu da podaci koji postanu deo *Swiss-Prot* baze budu kvalitetni). *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 ravni format [33]. 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 uprošćenim 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 [35].

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 [33, 35]:

   - 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 [35].

4. **Ključne reči** (engl. *keywords*) označene sa **KW** opisuju hijerarhisku strukturu kontrolisanog vokabulara namenjenog opisivanju funkcije proteina. Postoji 10 kategorija ključnih reči od kojih je za naše istraživanje bitna "Molekulska funkcija"[33]. Za razliku od GO čiji ideal je opis svih genskih produkta svih vrsta, termini ključnih reči prilagođeni su opisivanju sadržaja isključivo *Swiss-Prot* proteina [35].

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 [35].

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 [36].

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

7. Postojnost proteina **PE** (engl. *Protein existance*) opisuje stepen sigurnosti da protein postoji (Slika 3.6). Moguće vrednosti za pouzdanost su (u opadajućem poretku): potvrđeno na nivou proteina, potvrđeno na nivou RNK, zaključeno iz homologije, predviđen i nesiguran.

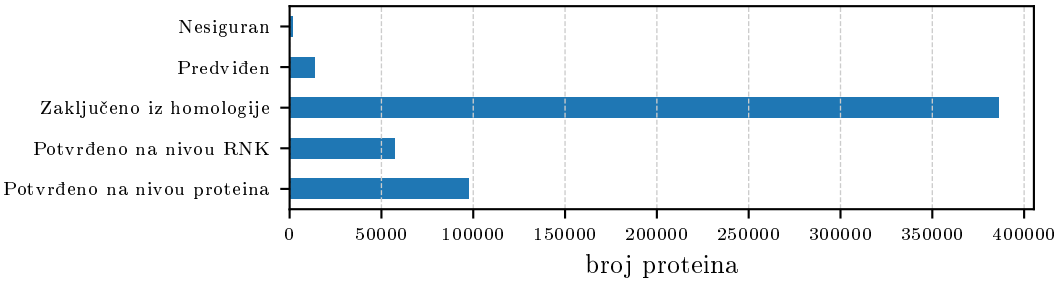
```
1  ID   ACSA_DROME              Reviewed;         670 AA.
   | ime sloga, info
2  AC   Q9VP61; Q24226; Q8IH30; Q9VP60;
   | identifikacija
3  DT   19-SEP-2003, integrated into UniProtKB/Swiss-Prot. | ulazak u Svis-Prot
4  DT   01-MAY-2000, sequence version 1.
   | ulazak u TrEMBL
5  DT   25-OCT-2017,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

7  DE   RecName: Full=Acetyl-coenzyme A synthetase;        |
8  DE            EC=6.2.1.1;                              |
9  DE   AltName: Full=Acetyl-CoA synthetase;              |
10 DE            Short=ACS;                               |
11 GN   Name=AcCoAS; ORFNames=CG9390;                     |
12 OS   Drosophila melanogaster (Fruit fly).              | Taksonomija
13 OC   Eukaryota; Metazoa; Ecdysozoa; Arthropoda; Hexap...|
14 OC   Pterygota; Neoptera; Holometabola; Diptera; Brac...|
15 OC   Ephydroidea; Drosophilidae; Drosophila; Sophopho...|
16 OX   NCBI_TaxID=7227 {ECO:0000312|EMBL:AAL90278.1};     |
17
18 RN   [1] {ECO:0000305}
   | Prva referenca
19 RP   NUCLEOTIDE SEQUENCE (ISOFORM B).                  |
20 RA   Russell S.R., Heimbeck G.M., Carpenter A.T., Ash...| Autori
21 RT   "A Drosophila melanogaster acetyl-CoA-synthetase...| Naslov
22 RL   Submitted (NOV-1994) to the EMBL/GenBank/DDBJ da...|
23 RN   [2]
   | Druga referenca
24 ...
25 CC   -!- FUNCTION: Activates acetate so that it can b...| Komentari
26 CC       synthesis or for energy generation.           |
27 CC       {ECO:0000250|UniProtKB:Q9NR19}.               |
28 CC   -!- CATALYTIC ACTIVITY: ATP + acetate + CoA = AM...|
29 ...

30 DR   EMBL; Z46786; CAA86738.1; ALT_SEQ; mRNA.          | reference ka
31 DR   EMBL; AE014296; AAF51695.2; -; Genomic_DNA.       | drugim bazama
32 ...                                                    | (dbxref)
33 DR   ExpressionAtlas; Q9VP61; differential.            |
34 DR   Genevisible; Q9VP61; DM.                          |
35 DR   GO; GO:0005737; C:cytoplasm; IEA:UniProtKB-SubCell.| GO termin <----
36 DR   GO; GO:0003987; F:acetate-CoA ligase activity; I...| GO termin <----
37 ...                                                    |

38 PE   2: Evidence at transcript level;
39 KW   Alternative splicing; ATP-binding; Complete proteome; Cytoplasm;
40 KW   Ligase; Nucleotide-binding; Reference proteome.
41 FT   CHAIN         1    670       Acetyl-coenzyme A synthetase.
42 FT                               /FTId=PRO_0000208425.
43 FT   VAR_SEQ       1    146       Missing (in isoform B).
44 FT                               {ECO:0000303|PubMed:12537569}.
45 FT                               /FTId=VSP_008310.
46 FT   CONFLICT    227    227       C -> S (in Ref. 1; CAA86738).
47 FT                               {ECO:0000305}.
48 SQ   SEQUENCE   670 AA;  75960 MW;  CE24364755CDBFFC CRC64;
49      MPAEKSIYDP NPAISQNAYI SSFEEYQKFY QESLDNPAEF WSRVAKQFHW ETPADQDKFL
50 ...
51      KKMVRERIGP FAMPDVIQNA PGLPKTRSGK IMRRVLRKIA VNDRNVGDTS TLADEQIVEQ
52      LFANRPVEAK
53 //  <--- oznacava kraj sloga
```

SLIKA 3.5: Uprošćen primer sloga, preuzet iz ravne datoteke uniprot_sport.data (preuzeto sa FTP servera [34])

8. *Swiss-Prot* takđe vrši predikcije neuređenih regiona koristeći *DISOPRED2* i *CLADIST* prediktore. [22]. Međutim, ove informacije postale su irelevantne pojavom baza neuređenja *MobiDB*[37] i *D2P2*[38].

9. Zanimljiva zapažanja iz globalne statistike o *Swiss-Prot* bazi:

   - Većina proteina ima dužinu između 100 i 500 AK.
   - Postojnost oko 70% proteina potvrđeno je homologijom (Slika 3.6).
   - Zastupljeno je preko 1000 različitih organizama, međutim većina *Swiss-Prot* sekvenci pripada malom broju organizama.



SLIKA 3.6: Histogram nivoa pouzdanosti postojanja *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[1]. 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, njegove autore, metode, podatke itd.

Najveća razlika u pristupu između originalnog i ovog istraživanja je što su originalni rezultati izraženi u terminima **ključnih reči** dok su ovde 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 funkcija. Jedan od rezultata 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 [1] 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.

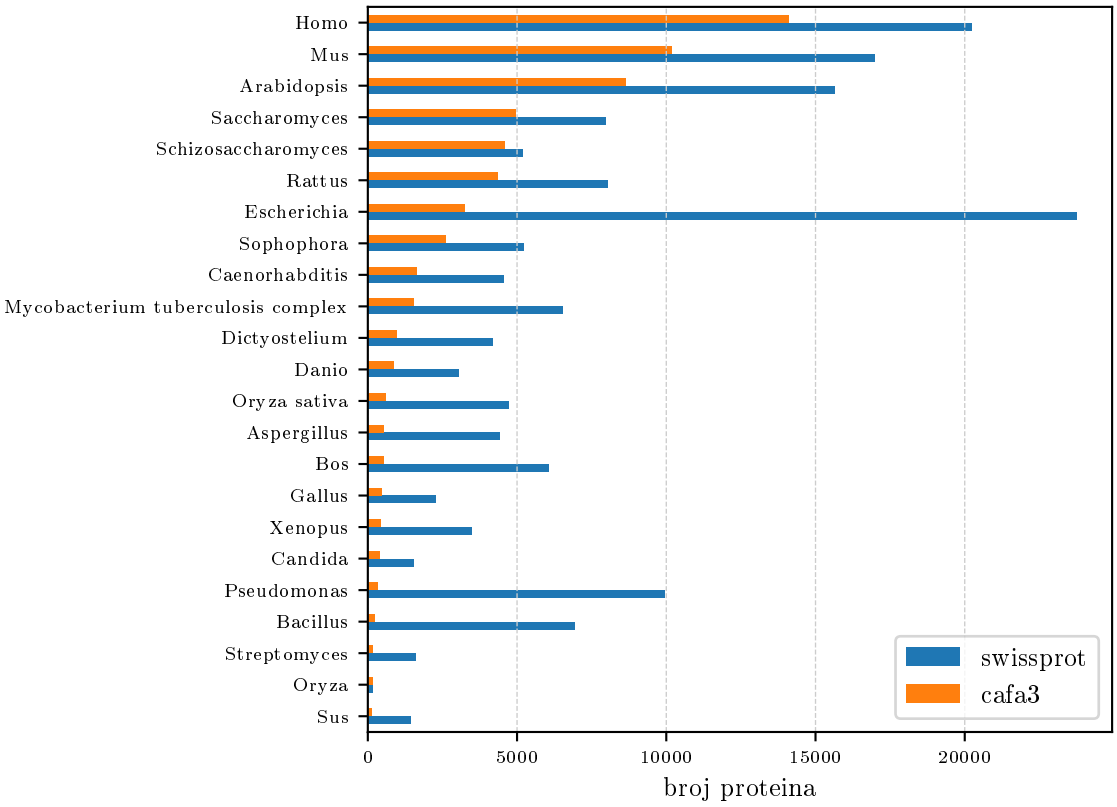
Kao što je pomenuto u Poglavlju 3.2, kanonske sekvence (proteini) u *Swiss-Prot* bazi podataka nisu redundantne u smislu da jedan unos u bazi podataka predstavlja produkt jednog gena iz jedne vrste organizma. Međutim, za analizu funkcija *Swiss-Prot* **jeste statistički redundantna** [1] jer sadrže veliku količinu **homologih** proteina (prvenstveno ortologa). Zbog statističke redundantnosti originalni autori su izvršili klasterovanje *Swiss-Prot* proteina u **proteinske familije** dobivši 27 217 familija. Posledica klasterovanja je da svaki protein dobija 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 Podaci korišćeni u ovom istraživanju

Ugledom na **CASP** takmičenja, **CAFA** (engl. *Critical Assessment of Functional Annotation*) takmičenje pokrenuto je zarad objektivnog ocenjivanja prediktora funkcije proteina i usmeravanja budućeg razvoja ove oblasti [39]. U ovom radu korišćen je trening skup proteina preuzet sa **CAFA3** takmičenja, održanog 2017. Trening skupovi su podaci na osnovu kojih prediktor uči, pa shodno tome ovaj skup treba da predstavlja dobar uzorak proteina odnosno njihovih funkcija.

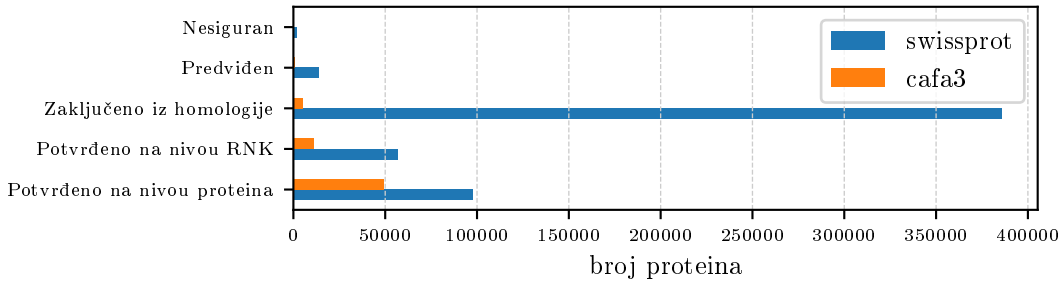
CAFA3 trening skup (u nastavku samo CAFA3 skup ili CAFA3 podaci) je podskup *Swiss-Prot* baze (iz 2016.) koji uključuje proteine iz model organizama: *Human, Mouse, Rat, S. cerevisiae, S. pombe, E. coli, A. thaliana, Dictyostelium discoideum, Zebrafish, Bacillus cereus* sa izuzetokm sekvenci *Drosophila* i *Candida* koje su preuzete iz svojih genomskih baza podataka, respektivno. Slika 4.1 pruža detaljan taksonomski uvid o poreklu CAFA3 proteina. Na primer, *Swiss-Prot* baza sadrži oko 20 000 ljudskih (*Homo*) proteinskih sekvenci dok CAFA3 podskup sadrži malo manje od 15 000. Vrste koje doprinose sa manje od 100 proteina nisu prikazane radi kompaktnosti.



SLIKA 4.1: Taksonomsko poreklo CAFA3 proteina.

Za razliku od proteina u *Swiss-Prot* bazi čije je postojanje pretežno utvrđeno iz homologije, 74% proteina iz CAFA3 podskupa identifikovani su na nivou nivou proteina što je najveći stupanj sigurnosti da proteini zaista postoji. Na Slici 4.2 ilustrovana je razlika u odnosu na pouzdanost postojanja proteina.

Važno je naglasiti da je dalja analiza izvršena pod pretpostavkom da CAFA3 skup nije statistički redundantan, što znači da klasterovanje u proteinske familije



SLIKA 4.2: Razlika izmedju pouzdanosti postojanja proteina iz *Swiss-Prot* baze i njenog CAFA3 podskupa.

nije potrebno. Isključivanje ove pretpostavke bi dovelo do komplikovanijih računarskih metoda za analizu što je van opsega ovog rada.

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` proteinima 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.

Analiza u ovom radu je primarno orijentisna ka korišćenju GO termina za opis funkcije što je razlikuje od originalnog pristupa korišćenja ključnih reči. Analiza GO termina zahteva poznavanje prvenstveno *is_a* roditeljske veze između termina. Takođe, tokom istraživanja bile su nam potrebne i ostale informacije o terminima. Pomenute informacije dobili smo iz datoteke `go.obo` [40] verzije 01.12.2017.

Odnos tj. mapiranje između ključnih reči i GO termina dostupno je sa dva izvora:

- `keywlist.txt`[41] verzija 20.12.2017 sadrži informacije o 1188 ključnih reči od kojih 195 pripada kategoriji **Molekulskih funkcija**. Više o mapiranju biće izloženo u Potpoglavlju 5.3.

- `uniprotkb_kw2go`[42] sadrži samo mapiranja a generiše ih **GOA projekat** [43].

U ovom radu korišćena je isključivo `keywlist.txt` datoteka. Iako je datoteka `uniprot_kw2go` sadržala veći broj mapiranja, unosila je nepoželjnu višeznačnost.

Pošto je originalni rad iz 2007. godine, postoje razlike u broju proteina, anotacijama ključnih reči, broju ključnih reči ali i primarnoj strukturi proteinskih sekvenci. Radi procene uticaja navedenih razlika na rezultat, odlučeno je da se analiza prvo ponovi originalnim pristupom, koristeći vokabular ključnih reči. Iz tog razloga, CAFA3 podaci nisu mogli da se posmatraju kao crne kutije, već je bilo neophodno mapirati ih nazad na *Swiss-Prot* proteine čije su anotacije ključnim rečima poznate.

Ovaj korak objedinjavanja baza podataka opisan je u Potpoglavlju 5.1. Dobijene anotacije ključnim rečima takođe su iskorišćene za proveru validnosti mapiranja ključnih reči na GO termine. Mapiranja su detaljno opisana u Potpoglavlju 5.3

4.2 Metod

Pod **idealnim slučajem**, 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.

Nažalost, ograničenja današnjih podataka i razvijenih metoda su brojna:

- Baza eksperimentalno utvrđenih neuređenih regiona ***DisProt*** ima svega 803 proteina sa opisanih 2167 neuređenih regiona [20]. Uz to, kvalitet navedenih informacija je diskutabilan zbog razlika u pouzdanosti među eksperimentalnim tehnikama koje su korišćene. Najveću pouzdanost nose regioni koji su eksperimentalno okarakterisani sa što većim brojem laboratorijskih tehnika.
- Prediktori se treniraju na malom podskupu proteina iz *DisProt* i PDB baze. Čak i konsenzus nekoliko različitih prediktora ne daje dovoljno pouzdane rezultate o lokaciji neuređenog regiona.
- Pozitivna strana je najnoviji napredak, razvoj prediktora koji direktno pokušavaju da predvide funkciju koju IDPr obavlja [22].

Jednostavna alternativa idealnom slučaju je da se pretpostavi da veći udeo neuređenih u odnosu na uređene proteine podrazumeva da funkcija zavisi više od neuređenosti.

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 prediktora detaljno su opisane u 2.3.1. Za potrebe analize originalni autori uvode sledeću definiciju:

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

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

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

Uslov "≥ 40" u originalnom radu delom je posledica ograničenja VL3 prediktora koji je treniran nad skupom sekvenci sa dugim neuređenim regionima [1].

[1] L označava duge regione, ≥ 30 AK

4.2.2 Zavisnost dužine proteina i predikcije dugačkog neuređenog regiona

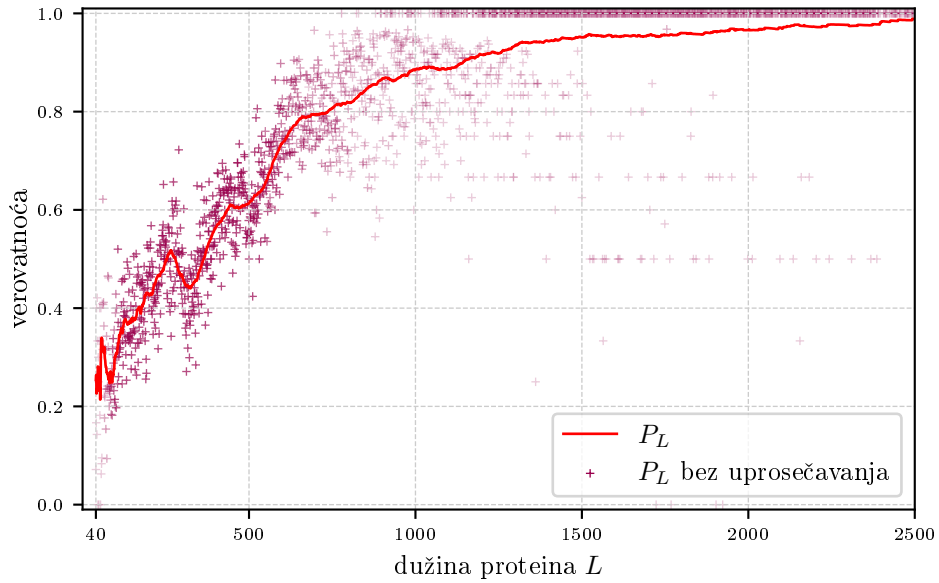
Verovatnoća da po gornjoj definiciji protein bude klasifikovan kao neuređen raste sa porastom njegove dužine. Ovo ponašanje utiče na statističku značajnost rezultata. Originalni autori predlažu narednu formulu za procenu pomenute verovatnoće.

Neka je S_L skup proteina sa dužinama iz intervala $[L-l, L+l]$[2] gde je $l = 0.1 \cdot L$. 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 4.3. 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. Ovo je suprotno konstantnom prozoru uprosečavanja koji je tehnika još poznata kao *rolling average* ili *boxcar filter* i predstavlja prostu vrstu konvolucije.

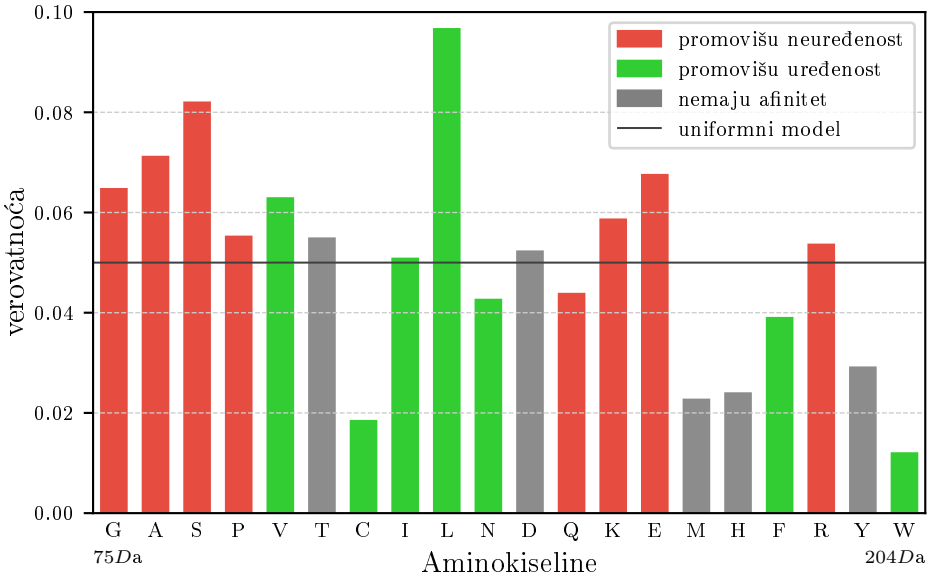


SLIKA 4.3: Zavisnot neuređenosti i dužine CAFA3 proteina minimalne dužine 40 AK (P_L sa prozorom uprosečavanja podrazumeva $l = 0.1L$, dok krstići predstavljaju diskretne vrednosti L za $l = 0$. Transparentnost krstića ilustruje brojnost proteina dužine L. Ako je krstić transparentan znači da sadrži manje od 100 proteina.)

Pored gore prikazanog originalnog metoda predložićemo još jedan pristup procenjivanju veličine P_L. Razmotrićemo dva modela zasnovana na slučajno generisanim proteinima. Prvi, naivni model **uniformne verovatnoće** podrazumeva da se svaka aminokiselina javlja sa istom verovatnoćom, odnosno $p = 1/20$. U statistici je ovaj model još poznat kao model jednakih verovatnoća (engl. *equiprobable model (EPM)*). Drugi model koji ćemo zvati slučajni ili *random* model RM) predstavlja

[2]Na primer, skup S_{100} sadrži proteine iz intervala $[90, 110]$, a S_{500} iz intervala $[450, 550]$

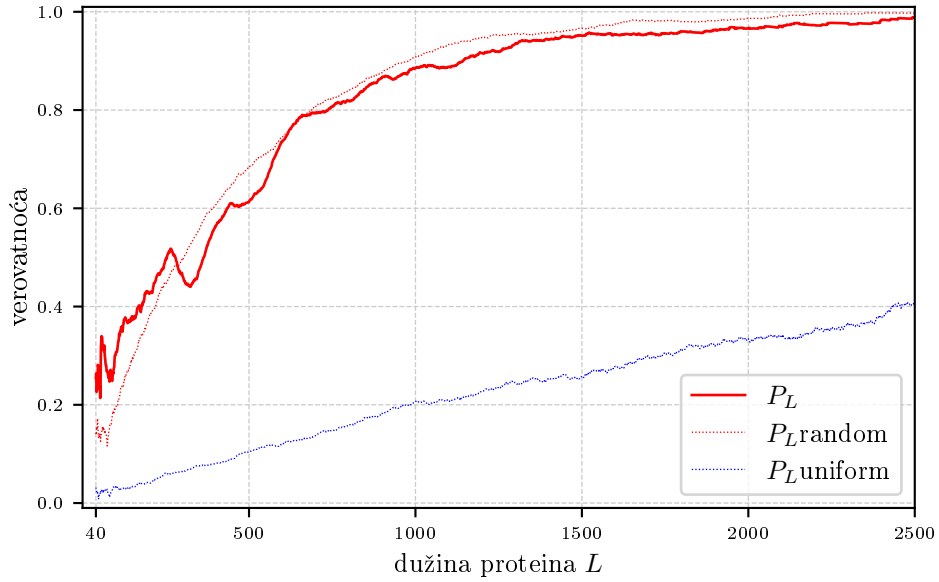
slučajnu promenljivu čija verovatnoća zavisi od učestalosti aminokiselina iz CAFA3 skupa i prikazana je na Slici 4.4. Na osnovu predstavljenih modela definisana su dva nova skupa sekvenci (EPMS i RMS) iste kardinalnosti kao polazni CAFA3 skup, u kojima su sekvence bile istih dužina kao u polaznom skupu. U EPMS, sekvence su generisane na osnovu uniformne raspodele nukleotida, dok su u RMS sekvence definisane na osnovu učestalosti aminokiselina iz CAFA3 skupa.



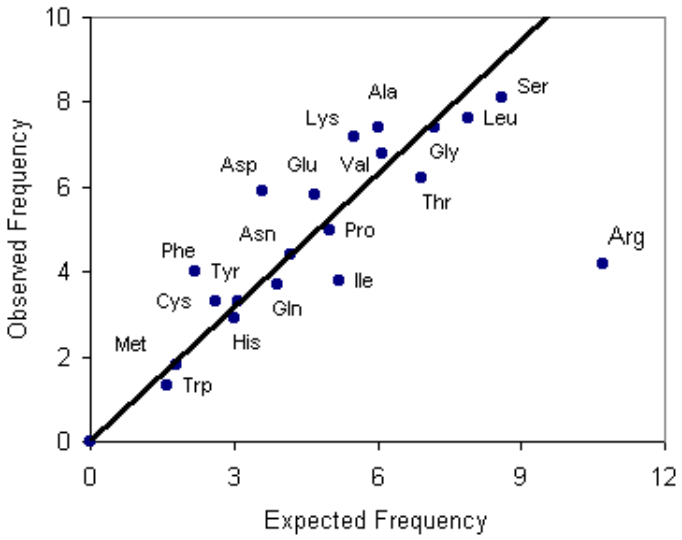
SLIKA 4.4: Učestalost aminokiselina u CAFA3 podacima
(Učestalost prikazuje uniformni model, dok boja AK obeležava afinitet prema uređenosti ili neuređenosti. Aminokiseline su poređane u rastućem poretku od najlakše do najteže.

Poređenje predloženih modela sa originalnim P_L prikazano je na Slici 4.5. Jasno se vidi da slučajni model predstavlja vizuelno dobru aproksimaciju dok uniformni model znatno odstupa rastući sporije (naizgled skoro linearno). Zbog znatnog vizuelnog 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 različitu učestalost u živom svetu. Naime, aminokiseline ne mogu imati istu verovatnoću pojavljivanja jer se broj njihovih kodona razlikuje. Neke aminokiseline su kodirane sa samo jednim, a druge i sa 6 kodona. Očekivano je da veći broj kodona povećava učestalost aminokiseline i ta korelacija uz izuzetke arginina predstavljena je Slikom 4.6 [44].



SLIKA 4.5: Upoređivanje P_L, $P_L random$ i $P_L uniform$ modela nad CAFA3 podacima



SLIKA 4.6: Očekivana i izmerena učestalost aminokiselina kod kičmenjaka (preuzeto sa [45]). Izmerena učestalost dobijena je analizom 53 kompletno sekvencionirana proteina iz kičmenjaka [46]. Očekivana učestalost kodona izračunata je kao proizvod učestalosti nukleinskih baza koje ga čine. Očekivana učestalost aminokiseline je zbir očekivanih učestalosti njenih kodona.

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*) manja od 0.05 onda je funkcija j povezana sa neuređenim proteinima, a ako je *p-value* veća od 0.95 onda je funkcija j povezana sa uređenim proteinima. Suprotno, povezanost funkcije j sa (ne)uređenošću nije statistički značajna.

U nastavku teksta, pod neuređenošću funkcije, ključne reči ili GO termina, biće podrazumevano da je odgovarajuća p vrednost manja od 0.05. Analogno, pod uređenošću funkcije, ključne reči ili GO termina, biće podrazumevano da je odgovarajuća p vrednost veća od 0.95.

Zbog matematičkog oblika 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 proteina $S_j = \{s_1, s_2, ..., s_k\}$. 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. Rezultat je vektor od k vrednosti nula ili jedan. Učestalost jedinica u rezultujućem vektoru predstavlja prvu realizacija Y_j. Postupak se ponavlja 1000 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.

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

Glava 5

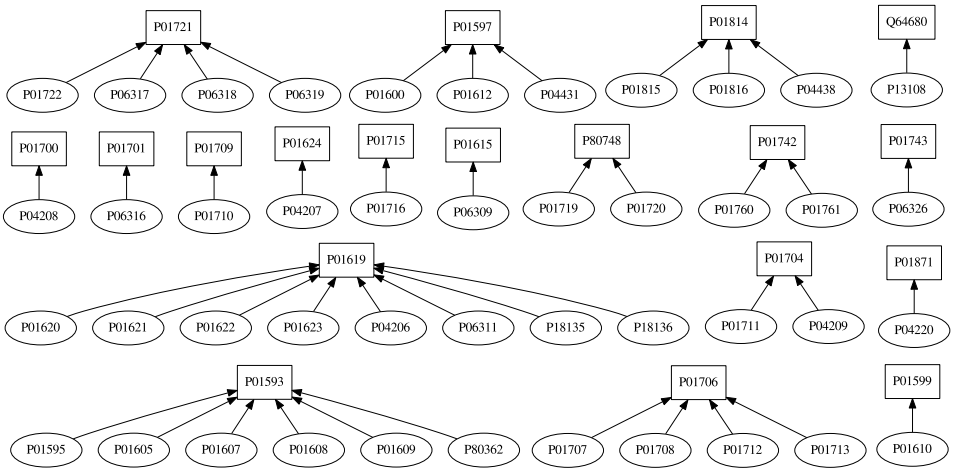
Priprema podataka

5.1 Objedinjavanje CAFA3 i novijih *Swiss-Prot* proteina

Povezivanje CAFA3 proteina sa *Swiss-Prot* unosima je jedini način da se dođe do anotacija ključnim rečima. Takođe, dobijamo najnovije anotacije GO termina ali i dodatne informacije, na primer taksonomsko poreklo proteina koje pomaže u razumevanju skupa podataka.

Iz CAFA3 trening skupa izdvojeni su svi validni proteini (dužine barem 9 i azbuke od 20 standardnih aminokiselina). U ovom koraku ne izbacujemo proteine kraće od 40 AK. Sadržaj *Swiss-Prot* baze dobijen je iz datoteke `uniprot_sprot-only2017_12.tar.gz` [34]. Navedena verzija sadrži 556 196 proteina. Od 66 599 validnih CAFA3 proteina 66 530 ima nepromenjen **primarni identifikator**, ali 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 četiri nova *Swiss-Prot* sloga koja odgovaraju nedostajućim CAFA3 proteinima. Kako je četiri suviše mali broj, zbog jednostavnosti nismo ih ubrajali u dalju analizu te koristimo samo 66 530 slogova čiji primarni identifikatori su nepromenjeni.



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

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 je korišćena verzija *Swiss-Prot* baze novijeg datuma od CAFA3 podskupa, CAFA3 verzija anotacija je zanemarena u korist novijih *Swiss-Prot* anotacija. Ove informacije sumirane su Tabelom 5.1

	CAFA3	*Swiss-Prot*
MF termini	5 957	7 471
nedostaje u go.obo	60 MF	319 MF, CC i BP
MF, samo u	67	888

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

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

5.2 Grupisanje proteina po GO terminima

Kao što je bilo reči u Sekciji 4.2.2 sa S_A označavamo skup proteina koji obavljaju funkciju A, odnosno funkcija A anotira proteine grupisane skupom S_A. Ako je GO termin A potomak GO termina B tj. važi A *is_a* B i želimo da diskutujemo o funkciji B, onda i svi proteini koji imaju funkciju A (i time pripadaju skupu S_A) treba da budu sadržani i u skupu S_B. Primetili smo da anotacije ključnih reči već podrazumevaju predloženo grupisanje. Na primer, ključna reč *Ribosomalprotein* anotira 1420 proteina, a njen predak (uopštenje) *Ribonucleoprotein* takođe anotira pomenutih 1420 proteina. Anotacije GO terminima ne podrazumevaju ovako grupisanje već anotacije koje preciznije odgovaraju funkciji proteina.

Za implementaciju predloženog grupisanja koristili smo algoritam topološkog sortiranja. Ovako formirani skupovi koji sadrže manje od 20 proteina nisu bili od značaja za dalju analize zbog čega su odbačeni. Ovom metodom dobijeno je 1781[1] MF termin (od ukupno 11 135 validnih MF termina) sa minimum 20 pridruženih proteina uključujući i koreni[2] termin. U ovom koraku uračunati su samo proteini minimalne dužine 40 AK.

5.3 Mapiranje između GO termina i *Swiss-Prot* ključnih reči

Kako je ovo istraživanje ograničeno na molekulske funkcije, u daljem tekstu biće razmatrana mapiranja isključivo između ključnih reči kategorije MF koje zovemo **MF ključne reči** i GO termina imenskog prostora MF koje zovemo **MF termini**. Pokazaćemo da ovo nije trivijalan zadatak i da u opštem slučaju zbog razlika u nomenklaturi mapiranje ne postoji ili nije ekvivalentno originalnoj funkciji.

U originalnom radu navodi se da je *Swiss-Prot* baza sadržala 143 MF ključne reči. Datoteka `keywlist.txt` [41] iz 20.12.2017 sadrži 195 MF ključnih reči. Nažalost, nismo pronašli originalnu `keywlist.txt` datoteku pa ne znamo egzaktnu razliku, ali jasno je da su neke ključne reči izbačene ili zamenjene a neke samo dodate. Datoteka

[1] Bez ovog grupisanja imali bi samo 1146 MF termina koji zadovoljavaju limit od min. 20 proteina
[2] koreni termin ili koreni čvor ontologije tj. termin molekulska funkcija

`keywlist.txt` opisuje ključne reči, a sadrži i pridruživanja (relaciju) odgovarajućim GO terminima. Pomenuta relacija pridružuje MF ključne reči ne samo MF terminima već i BP i CC terminima. Strogo posmatrano pridruživanja ne čine funkciju (mapiranje), jer se neke ključne reči preslikavaju u nekoliko GO termina čak i ako ograničimo sliku preslikavanja na samo MF[3] ili samo BP termine. Ipak, opisana pridruživanja nazivaćemo mapiranja ili direktna mapiranja. Direktna mapiranja za MF ključne reči opisana su Tabelom 5.2. Za 20 ključnih reči uopšte ne postoji mapiranje dok je broj mapiranja ka MF, BP i CC terminima, redom 104, 54 i 11. Dakle, veliki broj mapiranja ka MF terminima nedostaje.

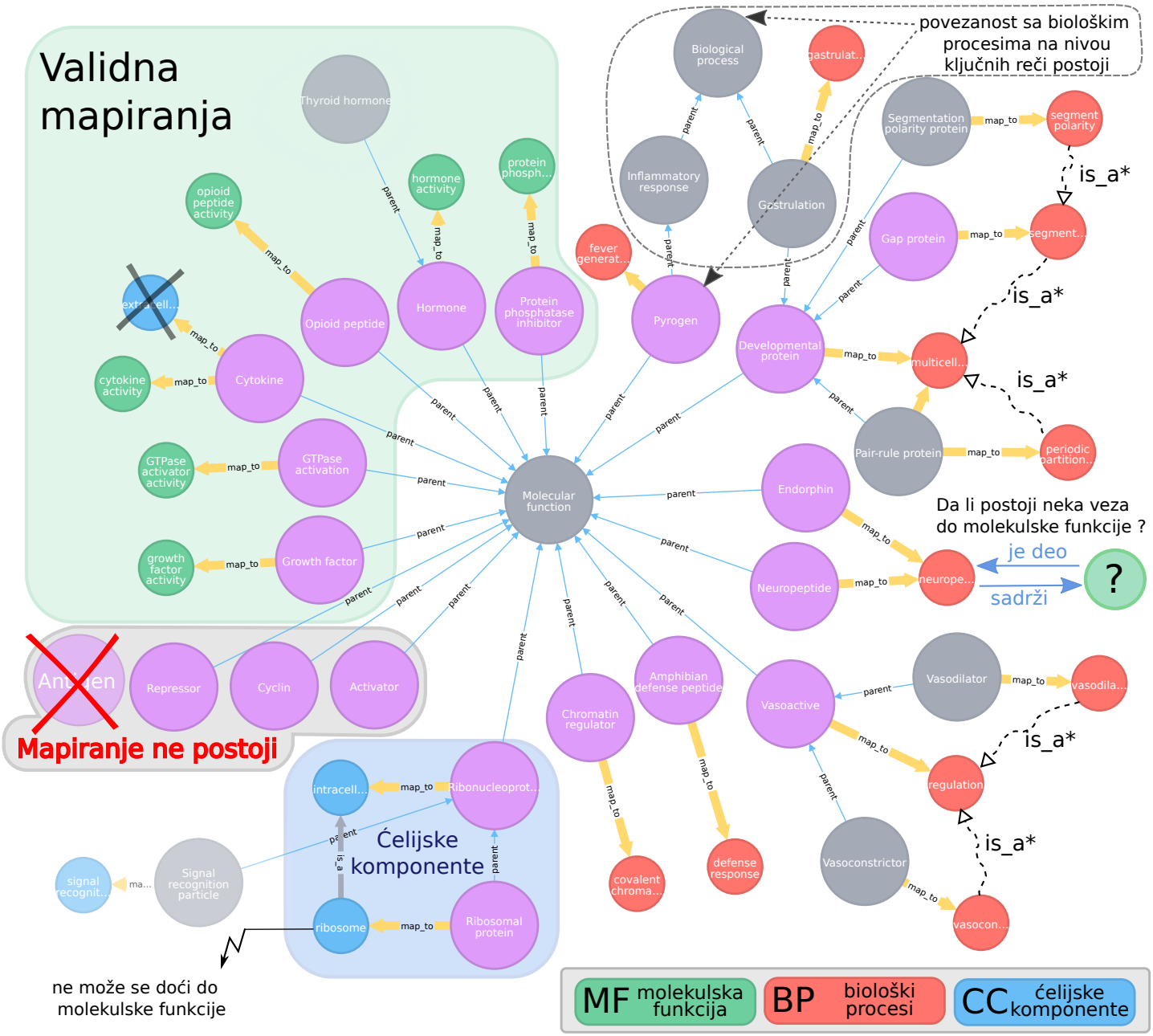
	ukupno	nema map.	MF map.	BP map.	CC map.
MF ključne reči	195	20	104	54	11

TABELA 5.2: Direktna mapiranja za MF ključne reči

Nedostajuća direktna MF mapiranja pogotovo dolaze do izražaja za neuređene ključne reči. Slika 5.2 prikazuje direktna mapiranja za 20 najznačajnijih neuređenih MF ključnih reči preuzetih iz originalnog rada [1]. Tri ključne reči nemaju direktno mapiranje, dve se mapiraju na ćelijske komponente, osam na biološke procese i svega šest na molekulske funkcije dok ključna reč *Antigen* više ne postojii. Sa druge strane, od 20 uređenih MF ključnih reči samo jednoj fali direktno mapiranje.

Nedostajuća MF mapiranja (skoro polovina MF ključnih reči) ne samo da otežavaju poređenje pojedinačnih ključnih reči već predstavljaju metodološki problem za poređenje nomenklatura. Originalni rezultati su sortirani prema statističkoj značajnosti (Z-skor za F_j) i predstavljaju dve tabele od 20 statistički najznačajnijih neuređenih odnosno 20 uređenih MF ključnih reči zanemarujući roditeljski odnos između njih. Na nivou ključnih reči ovo nije problem, ali ako bi se isti postupak primenio na MF termine poredili bi manje od 200 ključnih reči sa potencijalno preko 11 000 MF termina. Potrebno je odabrati dovoljno opšte MF termine, takve da čine reprezentativan uzorak molekulskih funkcija i da njihovo sortiranje po Z-skoru (kao u originalnom radu) ima smisla. Biće predložena dva pristupa za automatsko biranje opisanih MF termina dok je ručni odabir izvan obima ovoga rada. Oba pristupa imaju sličnu ideju koja podrazumeva izdvajanje samo onih MF termina koji su pridruženi MF ključnim rečima. Međutim, kako polovina MF mapiranja nedostaje (Tabela 5.2), neophodno je dopuniti nedostajuća mapiranja zarad smislenosti poređenja. Automatski metodi koje ćemo predložiti razlikuju se u načinu pronalaženja ovih nedostajućih mapiranja. Treba primetiti da, čak i da nema nedostajućih mapiranja, ovaj pristup zanemaruje statistički značajne MF termine koji nemaju ekvivalentne MF ključne reči.

[3]Samo *DNA invertase* se preslikava u dva različita MF termina (*DNA binding* i *recombinase activity*)



SLIKA 5.2: Direktno mapiranje 20 najznačajnijih neuređenih MF ključnih reči [1] na GO termine. 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.

5.3.1 Metod indirektnih mapiranja

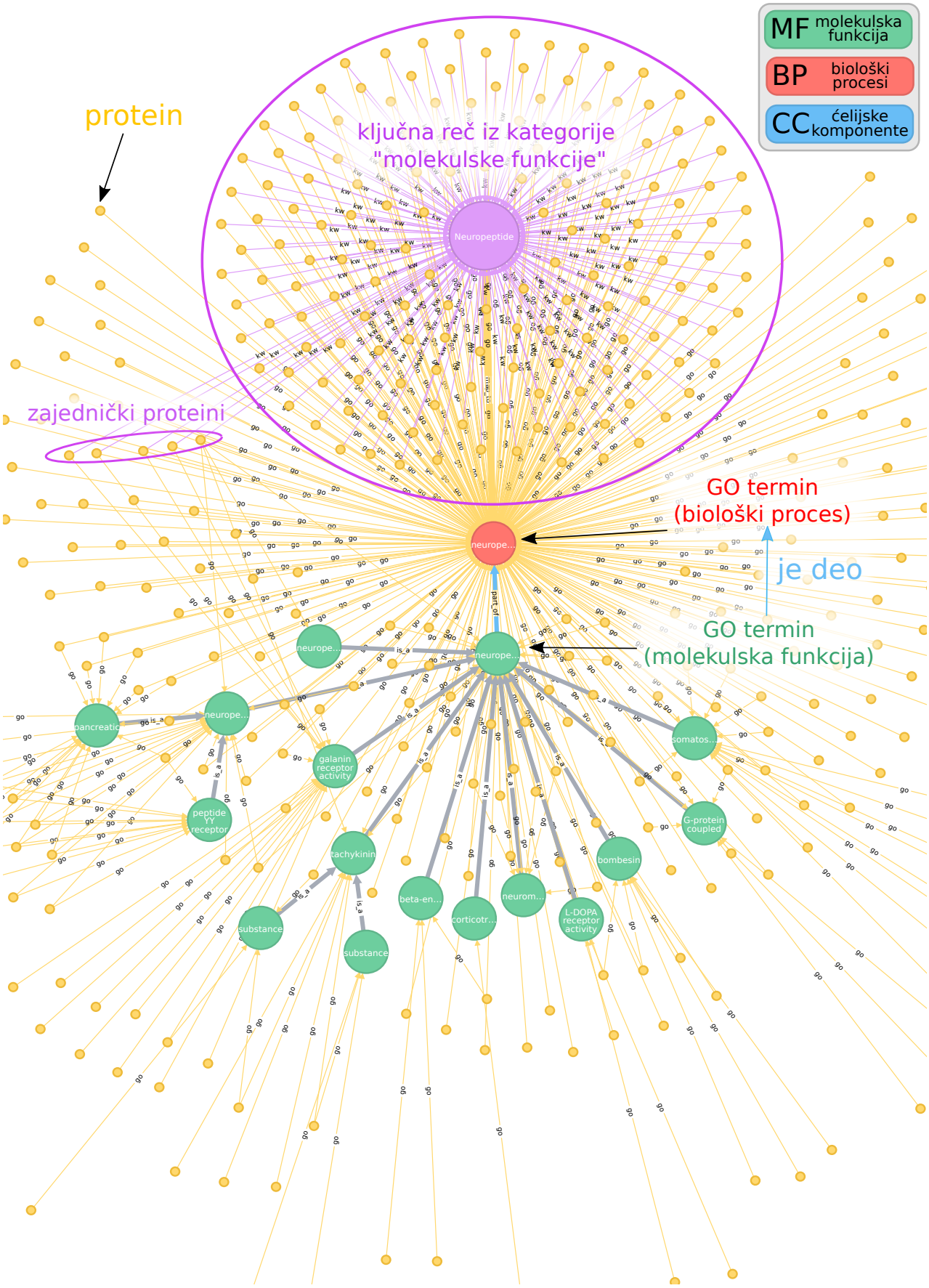
Rezonujući nad relacijama ontologije gena, moguće je doći do **indirektnih mapiranja** koja preko BP ili CC termina vode ka MF terminima. Neki MF termini su relacijama **part_of** i **has_part** povezani sa BP ili CC terminima, mada postoje MF termini koji nikako nisu povezani sa terminima ostalih ontologija. Traženje indirektnih mapiranja izvršeno je korišćenjem **Neo4j** grafovkse baze. Za sve ljubičaste BP termine sa Slike 5.2 i sve njihove specijalizacije provereno je da li su nekom (bilo kojom) relacijom u vezi sa nekim (bilo kojim) MF terminom. Pronađena je samo jedno indirektno mapiranje, zajedničko za MF ključne reči *Neuropeptide* i *Endorphin*. Za tri pomenuta CC termina indirektna mapiranja nisu pronađena.

Razmotrimo pronađeno mapiranje za MF ključnu reč *Neuropeptide* i proteine koji su anotirani razmatranim funkcijama. Slika 5.3 generisana je **Cypher**[4] upitom koji koji pored GO termina takođe vraća anotirane proteine. Preciznije, upit prvo pronalazi direktno mapirnje (BP termin) a zatim sve potomke (indirektne MF termine) i na kraju anotirane proteine. Proteini su preuzeti iz CAFA3 skupa.

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

Sa Slike 5.3 jasno se uočava da ključna reč *Neuropeptide* deli svega 5 anotacija sa svim prikazanim MF terminima. Smatramo da je zbog male sličnosti u anotacijama pronađeno indirektno mapiranje nevalidno, jer se očigledno MF ključna reč *Neuropeptide* i MF termin *Neuropeptide receptor activity* koriste u različitim kontekstima. Zapravo, šablonski naziv pronađenog MF termina (po pravilima iz Potpoglavlja 3.1.3) označava da se termin koristi za anotiranje proteina koji se *vezuju za neuropeptid zarad inicijacije neke ćelijske funkcije*, a ne samih neuropeptida. Takođe, treba primetiti da je pronađeni MF termin povezan na BP termin relacijom **part_of**. Ovo može da predstavlja još jedan razlog zašto je pronađeno mapiranje nevalidno. Veza **part_of** podrazumeva agregaciju a ne kompoziciju. Agregacija podrazumeva da molekulska funkcija postoji nezavisno od biološkog procesa što je suprotno od kompozicije koja predstavlja relaciju **has**. Nažalost, nijedan od osam pomenutih BP termina nema relaciju **has** ka nekom MF terminu.

[4] *Cypher* je upitni jezik za Neo4j grafovsku bazu



SLIKA 5.3: Mapiranje ključne reči **Neuropeptide** na MF termin *Neuropeptide receptor activity* preko relacije **part_of**. Dobijeno mapiranje rezultuje malim brojem zajednički anotiranih proteina.

Kao što je bilo reči u Potpoglavlju 3.1.3, molekulske funkcije ne mogu da predstavljaju gene ili genske produkte već funkcije koje oni obavljaju. Ključna reč **Neuropeptide** po definiciji predstavlja peptide koje neuronske ćelije oslobađaju ili hormone koji se oslobađaju iz drugih tipova ćelije. Jasno je da ekvivalentan MF termin ne može da postoji. Nažalost, postoji veliki broj MF ključnih reči koje iz ovog razloga ne mogu da imaju ekvivalentan MF termin, na primer: *Neurotoxin, Endorphin, Pyrogen, Milk protein, GAP protein, Motor protein, Myosin, ...* Veliki broj ovakvih ključnih reči objašnjava zašto skoro pola direktnih mapiranja na MF termine ne postoji. Zbog svih navedenih problema, zaključili smo da indirektna mapiranja nisu adekvatna za primene u ovom radu.

Ipak, treba pretpostaviti da je za neke ključne reči moguće pronaći MF termin koji predstavlja zajedničku molekulsku funkciju za skup proteina anotiran polaznom MF ključnom reči. U cilju pronalaženja takvih MF termina predlažemo jednostavan metod zasnovan na sličnosti skupova.

5.4 Metod sličnih anotacija

Metod sličnih anotacija pretpostavlja da dve ekvivalentne funkcije (iz dve različite nomenklature) anotiraju sličan skup proteina. Jedan način da se definiše sličnost između dva skupa A i B je preko **Žakardovog indeksa** (engl. *Jaccard index*), kraće **Ji** definisanog sledećom formulom:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

$$J(A,B) \in [0,1]$$

$$J(A,B) = \begin{cases} 1, & A = B \\ 0, & A \cap B = 0 \end{cases}$$

Za realizaciju ove metode korišćeni su svi proteini iz *Swiss-Prot* baze, ne samo CAFA3 podskup. Neka skup A predstavlja proteine koje MF ključna reč kw_A anotira a skup B proteine koje anotira MF termin t_B. Zbog jednostavnosti, skup B nije dobijen grupisanjem opisanim u Potpoglavlju 5.2, dakle ne sadrži proteine specifične za potomke termina t_B već samo sirove anotacije iz *Swiss-Prot* baze. Za kw_A najverovatnije mapiranje predstavlja MF termin t_B, takav da je Žakardov indeks $J(A,B)$ najveći. Ipak, ne treba očekivati da će najveći Žakardov indeks nužno značiti i najpogodnije mapiranje. Iz tog razloga potrebno je sagledavati nekoliko najboljih predloga za mapiranje, sortiranih opadajuće po veličini Žakardovog indeksa. Mapiranja dobijena ovim postupkom zvaćemo **izvedena mapiranja**.

Pouzdanost predloženog metoda testirana je nad 104 MF ključne reči sa poznatim direktnim mapiranjem. Za svaku ključnu reč izdvojili smo maksimum pet najverovatnijih izvedenih mapiranja. Izvedna mapiranja sa najvećim Žakardovim indeksom poklapaju se sa poznatim direktnim mapiranjima za 61 ključnu reč (0.59%). Ručnom validacijom tj. razmatranjem izvedenih mapiranja nižeg Žakardovog indeksa moguće je povećati broj korektnih mapiranja na 90 (0.85%).

Opisani metod uz razmatranje maksimum 5 najpogodnijih izvedenih mapiranja primenili smo nad 91 MF ključnu reč sa nedostajućim direktnim mapiranjem. Izvedena mapiranja sa Žakardovim indeksom manjim od 0.1 nisu razmatrana što je automatski eliminisalo 25 ključnih reči iz daljeg razmatranja. Nakon ručne validacije dobijena su 64 izvedena mapiranja. Tabela 5.3 prikazuje izvedena mapiranja sa

minimalnim Žakardovim indeksom 0.2 izuzev poslednja tri reda. Poslednja tri reda i ostali redovi sa zadebljanim tekstom (engl. *bold text*) predstavljaju nedostajuća mapiranja sa Slike 5.2.

Dakle, od ukupno 195 MF ključnih reči za 104 postoje direktna mapiranja ka MF terminima (ukupno 105 pridruživanja) što je dopunjeno sa još 64 izvedena mapiranja. Ukupno je dobijeno 169 pridruživanja između 168 MF ključnih reči i 138 MF termina.

MF keyword	**n_kw**	**Ji**	**n_go**	*MF name*
Dermonecrotic toxin	148	0.96	142	phospholipase D activity
Ribosomal protein	49054	0.91	48096	structural constituent of ribosome
Complement system impairing toxin	160	0.81	142	phospholipase D activity
Hemagglutinin	397	0.75	299	host cell surface receptor binding
Mutator protein	255	0.75	288	damaged DNA binding
Antifreeze protein	10	0.7	7	ice binding
Light-harvesting polypeptide	90	0.68	61	bacteriochlorophyll binding
Cyclin	197	0.61	124	cyclin-dependent protein serine/threonine kinase regulator activity
Defensin	55	0.55	32	CCR6 chemokine receptor binding
Ribonucleoprotein	50698	0.54	28317	rRNA binding
Neurotoxin	2734	0.53	4145	toxin activity
Photoprotein	40	0.48	19	alkanal monooxygenase (FMN-linked) activity
Endorphin	48	0.45	32	opioid peptide activity
Mobility protein	7	0.43	3	DNA topoisomerase type I activity
Protein synthesis inhibitor	150	0.43	67	rRNA N-glycosylase activity
Neuropeptide	561	0.42	267	neuropeptide hormone activity
Signal transduction inhibitor	157	0.38	158	GTPase activator activity
Mitogen	282	0.37	284	growth factor activity
Repressor	8177	0.33	7798	DNA binding
Chaperone	11245	0.31	7412	ATP binding
Myosin	372	0.31	275	motor activity
Viral nucleoprotein	727	0.31	486	structural molecule activity
Pair-rule protein	24	0.29	16	RNA polymerase II sequence-specific DNA binding transcription factor binding
Prion	91	0.28	217	copper ion binding
Milk protein	96	0.27	83	transporter activity
Motor protein	919	0.27	251	microtubule motor activity
Pyrogen	43	0.27	41	interleukin-1 receptor binding
Activator	7081	0.26	7798	DNA binding
Bence-Jones protein	8	0.26	26	antigen binding
Serine protease homolog	57	0.26	21	hemoglobin binding
Thyroid hormone	28	0.25	32	thyroid hormone binding
Antiviral protein	40	0.23	25	ribonuclease III activity
Ligand-gated ion channel	460	0.23	111	acetylcholine-gated cation-selective channel activity
Actin capping	168	0.22	367	actin binding
Presynaptic neurotoxin	307	0.22	251	phospholipase A2 activity (consuming 1 & 2-dipalmitoylphosphatidylcholine)
Retinal protein	267	0.22	64	G-protein coupled photoreceptor activity
Fungicide	157	0.21	76	chitin binding
Receptor	6753	0.21	1565	G-protein coupled receptor activity
Neurotransmitter	34	0.2	32	opioid peptide activity
Vasoactive	243	0.17	489	hormone activity
Chromatin regulator	1939	0.12	861	chromatin binding
Developmental protein	6285	0.12	2464	sequence-specific DNA binding

TABELA 5.3: Izvedena mapiranja (*n_kw* i *n_go* obeležavaju arnost skupa proteina koje anotira MF ključna reč i skupa proteina koje anotira MF termin)

Glava 6

Rezultati

Računarska analiza implementirana je u jeziku Pajton (engl. *Python*) verzija 3.6. Kompletan projekat može se naći na *github* adresi [47]. Za potrebe projekta napravljene su dve baze podataka: relaciona baza podataka (*PostgreSQL* v9.5) i grafovska baza podataka (*Neo4j* v3.1).

Od 186 MF ključnih reči koje anotiraju bar 20 proteina, 97 je statistički značajno od čega su 53 predviđeno uređene ($p<0.05$), a 44 predviđeno neuređene ($p>0.95$). Od 1781 MF termina sa preko 20 pridruženih proteina (dobijeno grupisanjem, Potpoglavnje 5.2), 1315 je statistički značajno od čega su 699 predviđeno uređeni, a 616 predviđeno neuređeni. Tabela 6.1 prikazuje razlike u odnosu na originalne rezultate.

	originalni rezultati	novi rezultati	
	Xie2007 kw	MF kw	MF termini
ukupno ($br.prot \geq 20$)	143	186	1781
p<0.05 (uređene)	37	53	699
p>0.95 (neuređene)	51	44	616

TABELA 6.1: Uopšteno poređenje rezultata

Slika 6.1 sadrži tri tabele predviđeno neuređenih funkcija koje sadrže: rezultate iz originalnog rada [1] (levo), nove rezultate za MF ključne reči (sredina) i rezultate za mapirane MF termine (desno). Sve tabele su sortirane po z-skoru opadajuće (statistička značajnost predviđene neuređenosti opada). Leva tabela sadrži samo 20 statistički najznačajnije neuređenih MF ključnih reči dok je srednja ograničena na z-skor veći od dva. Redovi leve tabele su crveni ako se odgovarajuća ključna reč ne nalazi u srednjoj tabeli. Redovi srednje tabele su zeleni ako se nalaze među prvih 20 a ključna reč se ne javlja u levoj tabeli. Desna tabela MF termina sadrži samo one termine koji su rezultat direktnog ili izvedenog mapiranja. Punim strelicama prikazana su direktna mapiranja, a isprekidanim izvedena. Nedostatak veze između srednje i desne tabele sugeriše da suprotna funkcija nije statistički značajna, a crvena boja redova u srednjoj tabeli označava da ni direktno ni indirektno mapiranje ne postoji. Srednja i desna tabela pored kolona z-skor i p vrednosti sadrže i procenat neuređenih proteina (avg_dis) na koji se ubuduće referišemo kao **neuređenost funkcije**. Slike 6.2a, 6.2b pojednostavljuju sagledavanje pojedinačnih razlika među rezultatima, međutim zbog ograničenja formata slike neke tabele su skraćene.

Slike 6.3 i 6.4 imaju ekvivalentan format poređenja kao i Slike 6.1 i 6.2, ali predstavljaju sortirane uređene funkcije.

20 MF ključnih reči sa najznačajnijom predviđenom neuređenošću (org. 2007)						
#	*name*	*n prot*	*n families*	*avg_len*	*Z-score*	*P-value*
0	Ribonucleoprotein	12236	412	150.55	22.13	1
1	Ribosomal protein	11692	330	140.58	20.63	1
2	Developmental protein	3260	721	477.93	19.28	1
3	Hormone	1187	161	141.13	15.58	1
4	Growth factor	785	84	255.70	11.16	1
5	Cytokine	899	110	213.28	10.21	1
6	Neuropeptide	268	209	95.08	9.65	1
7	Activator	3086	573	428.47	9.04	1
8	GAP protein	47	2	232.96	7.42	1
9	Antigen	1113	455	437.48	6.99	1
10	Repressor	2309	449	374.46	6.92	1
11	Chromatin regulator	334	100	801.24	6.70	1
12	Pyrogen	37	2	262.59	6.44	1
13	Vasoactive	125	39	160.39	5.56	1
14	Amphibian defense peptide	123	148	50.64	5.44	1
15	GTPase activation	311	70	831.03	5.36	1
16	Endorphin	42	4	226.68	5.35	1
17	Opioid peptide	24	4	216.96	5.14	1
18	Protein phosphatase inhibitor	47	8	366.51	5.07	1
19	Cyclin	182	25	430.58	4.88	1

Predviđeno neuređene MF ključne reči iz ovog rada (*z*-skor > 0.2)						
#	*name*	*n*	*avg_len*	*avg_dis*	*z*	*p*
0	DNA-binding	6518	546.53	0.87	46.90	1
1	Developmental protein	3897	655.21	0.86	31.10	1
2	Activator	2574	600.51	0.88	28.12	1
3	Repressor	1988	589.29	0.85	22.63	1
4	RNA-binding	2728	575.76	0.76	16.62	1
5	Chromatin regulator	1038	847.06	0.90	13.91	1
6	Ribonucleoprotein	1886	272.29	0.60	13.39	1
7	Serine/threonine-protein ki...	1782	802.24	0.84	11.56	1
8	Chaperone	937	430.43	0.71	10.02	1
9	Ribosomal protein	1408	186.38	0.53	9.34	1
10	Growth factor	398	299.63	0.70	8.98	1
11	Protein kinase inhibitor	49	337.20	0.96	8.34	1
12	Calmodulin-binding	520	1229.00	0.90	7.57	1
13	Hormone	338	221.13	0.59	7.24	1
14	Cyclin	133	422.71	0.87	7.18	1
15	Signal transduction inhibitor	115	408.43	0.84	6.76	1
16	Guanine-nucleotide releasin...	319	1144.39	0.96	6.40	1
17	GTPase activation	424	867.35	0.88	6.28	1
18	Growth factor binding	50	593.98	1	6.09	1
19	Neuropeptide	105	234.96	0.68	5.99	1
20	Potassium channel	191	621.52	0.85	5.10	1
21	Calcium channel	193	1397.77	0.93	5.09	1
22	Protein phosphatase inhibitor	64	352.86	0.81	5.07	1
23	Tyrosine-protein kinase	376	863.00	0.89	5.04	1
24	Mitogen	137	286.03	0.68	4.80	1
25	Vasoactive	46	267.00	0.76	4.29	1
26	Heparin-binding	221	650.97	0.73	3.88	1
27	Muscle protein	193	920.01	0.73	3.81	1
28	Actin-binding	837	974.92	0.77	3.80	1
29	Amphibian defense peptide	49	85.80	0.53	3.34	1
30	Helicase	739	1086.05	0.87	3.29	1
31	Prion	22	497.05	0.91	3.06	1
32	Ion channel	1027	861.88	0.76	2.89	1
33	Voltage-gated channel	386	816.98	0.78	2.87	1
34	Viral nucleoprotein	39	1202.79	0.90	2.46	1
35	Tumor antigen	26	428.81	0.77	2.25	0.99
36	Exonuclease	239	725.99	0.75	2.19	0.99
37	Segmentation polarity protein	24	712.75	0.92	2.17	0.99

Predviđeno neuređeni MF termini dobijeni iz direktnog i izvedenog mapiranja						
#	*name*	*n*	*avg_len*	*avg_dis*	*z*	*p*
0	DNA binding	8254	574.89	0.86	46.13	1
1	sequence-specific DNA binding	3836	536.95	0.92	41.77	1
2	chromatin binding	1538	713.64	0.93	23.54	1
3	RNA binding	5797	565.80	0.73	19.79	1
4	receptor binding	4109	529.68	0.68	14.04	1
5	protein serine/threonine ki...	1942	784.87	0.84	12.88	1
6	structural molecule activity	2884	558.88	0.64	10.84	1
7	structural constituent of r...	1350	182.95	0.53	9.25	1
8	calmodulin binding	710	1079.79	0.88	9.22	1
9	growth factor activity	458	312.38	0.70	8.99	1
10	protein kinase inhibitor ac...	237	401.60	0.82	8.94	1
11	hormone activity	452	218.09	0.61	8.84	1
12	cell adhesion molecule binding	1090	805.74	0.80	8.60	1
13	GTPase activator activity	633	820.60	0.87	8.15	1
14	actin binding	1228	1000.47	0.80	7.73	1
15	cyclin-dependent protein se...	142	357.65	0.77	6.58	1
16	protein tyrosine kinase act...	504	933.02	0.86	6.26	1
17	guanyl-nucleotide exchange ...	575	990.55	0.86	6.22	1
18	protein phosphatase inhibit...	89	425.83	0.83	6.19	1
19	heparin binding	403	652.76	0.76	6.01	1
20	neuropeptide hormone activity	117	158.64	0.65	5.88	1
21	microtubule motor activity	225	1394.78	0.97	5.65	1
22	enzyme inhibitor activity	1059	462.66	0.63	5.60	1
23	potassium channel activity	339	612.39	0.79	4.88	1
24	damaged DNA binding	284	680.80	0.80	4.31	1
25	growth factor binding	351	737.85	0.79	4.20	1
26	chloride channel inhibitor ...	25	827.64	0.96	3.96	1
27	calcium channel activity	327	1205.13	0.86	3.54	1
28	morphogen activity	23	331.57	0.83	3.20	1
29	helicase activity	766	1068.02	0.86	3.15	1
30	motor activity	447	1275.57	0.86	3.10	1
31	voltage-gated ion channel a...	501	767.73	0.76	3.06	1
32	rRNA binding	541	266.55	0.52	3.00	1
33	calcium ion binding	1878	859.00	0.70	2.80	1
34	RNA-directed DNA polymerase...	75	1427.97	0.97	2.72	1
35	kinase activity	3593	726.02	0.71	2.70	1
36	ATP-dependent helicase acti...	455	926.44	0.84	2.48	1
37	exonuclease activity	371	749.41	0.75	2.40	0.99
38	translation initiation fact...	268	482.02	0.65	1.99	0.97
39	acetylcholine-gated cation-...	60	504.17	0.75	1.99	0.99
40	metalloendopeptidase inhibi...	40	323.35	0.65	1.86	0.98
41	cysteine-type endopeptidase...	153	424.63	0.58	1.48	0.95

SLIKA 6.1: Poređenje predviđenih neuređenih funkcija

a)

20 MF ključnih reči sa najznačajnijom predviđenom neuređenošću (org. 2007)

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

Predviđeno neuređene MF ključne reči iz ovog rada (z-skor > 0.2)

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

b)

Predviđeno neuređene MF ključne reči iz ovog rada (z-skor > 0.2)

#	name	n	avg_len	avg_dis	z	p
0	DNA-binding	6518	546.53	0.87	46.90	1
1	Developmental protein	3897	655.21	0.86	31.10	1
2	Activator	2574	600.51	0.88	28.12	1
3	Repressor	1988	589.29	0.85	22.63	1
4	RNA-binding	2728	575.76	0.76	16.62	1
5	Chromatin regulator	1038	847.06	0.90	13.91	1
6	Ribonucleoprotein	1886	272.29	0.60	13.39	1
7	Serine/threonine-protein ki...	1782	802.24	0.84	11.56	1
8	Chaperone	937	430.43	0.71	10.02	1
9	Ribosomal protein	1408	186.38	0.53	9.34	1
10	Growth factor	398	299.63	0.70	8.98	1
11	Protein kinase inhibitor	49	337.20	0.96	8.34	1
12	Calmodulin-binding	520	1229.00	0.90	7.57	1
13	Hormone	338	221.13	0.59	7.24	1
14	Cyclin	133	422.71	0.87	7.18	1
15	Signal transduction inhibitor	115	408.43	0.84	6.76	1
16	Guanine-nucleotide releasin...	319	1144.39	0.96	6.40	1
17	GTPase activation	424	867.35	0.88	6.28	1
18	Growth factor binding	50	593.98	1	6.09	1
19	Neuropeptide	105	234.96	0.68	5.99	1
20	Potassium channel	191	621.52	0.85	5.10	1
21	Calcium channel	193	1397.77	0.93	5.09	1
22	Protein phosphatase inhibitor	64	352.86	0.81	5.07	1
23	Tyrosine-protein kinase	376	863.00	0.89	5.04	1
24	Mitogen	137	286.03	0.68	4.80	1
25	Vasoactive	46	267.00	0.76	4.29	1
26	Heparin-binding	221	650.97	0.73	3.88	1
27	Muscle protein	193	920.01	0.73	3.81	1
28	Actin-binding	837	974.92	0.77	3.80	1
29	Amphibian defense peptide	49	85.80	0.53	3.34	1
30	Helicase	739	1086.05	0.87	3.29	1
31	Prion	22	497.05	0.91	3.06	1
32	Ion channel	1027	861.88	0.76	2.89	1
33	Voltage-gated channel	386	816.98	0.78	2.87	1
34	Viral nucleoprotein	39	1202.79	0.90	2.46	1
35	Tumor antigen	26	428.81	0.77	2.25	0.99

Predviđeno neuređeni MF termini dobijeni iz direktnog i izvedenog mapiranja

#	name	n	avg_len	avg_dis	z	p
0	DNA binding	8254	574.89	0.86	46.13	1
1	sequence-specific DNA binding	3836	536.95	0.92	41.77	1
2	chromatin binding	1538	713.64	0.93	23.54	1
3	RNA binding	5797	565.80	0.73	19.79	1
4	receptor binding	4109	529.68	0.68	14.04	1
5	protein serine/threonine ki...	1942	784.87	0.84	12.88	1
6	structural molecule activity	2884	558.88	0.64	10.84	1
7	structural constituent of r...	1350	182.95	0.53	9.25	1
8	calmodulin binding	710	1079.79	0.88	9.22	1
9	growth factor activity	458	312.38	0.70	8.99	1
10	protein kinase inhibitor ac...	237	401.60	0.82	8.94	1
11	hormone activity	452	218.09	0.61	8.84	1
12	cell adhesion molecule binding	1090	805.74	0.80	8.60	1
13	GTPase activator activity	633	820.60	0.87	8.15	1
14	actin binding	1228	1000.47	0.80	7.73	1
15	cyclin-dependent protein se...	142	357.65	0.77	6.58	1
16	protein tyrosine kinase act...	504	933.02	0.86	6.26	1
17	guanyl-nucleotide exchange ...	575	990.55	0.86	6.22	1
18	protein phosphatase inhibit...	89	425.83	0.83	6.19	1
19	heparin binding	403	652.76	0.76	6.01	1
20	neuropeptide hormone activity	117	158.64	0.65	5.88	1
21	microtubule motor activity	225	1394.78	0.97	5.65	1
22	enzyme inhibitor activity	1059	462.66	0.63	5.60	1
23	potassium channel activity	339	612.39	0.79	4.88	1
24	damaged DNA binding	284	680.80	0.80	4.31	1
25	growth factor binding	351	737.85	0.79	4.20	1
26	chloride channel inhibitor ...	25	827.64	0.96	3.96	1
27	calcium channel activity	327	1205.13	0.86	3.54	1
28	morphogen activity	23	331.57	0.83	3.20	1
29	helicase activity	766	1068.02	0.86	3.15	1
30	motor activity	447	1275.57	0.86	3.10	1
31	voltage-gated ion channel a...	501	767.73	0.76	3.06	1
32	rRNA binding	541	266.55	0.52	3.00	1
33	calcium ion binding	1878	859.00	0.70	2.80	1

SLIKA 6.2: Razdvojeno poređenje predviđenih neuređenih funkcija.
Neke tabele su skraćene.

19 MF ključnih reči sa najznačajnijom predviđenom uređenošću (org. 2007)					
#	*name*	*n_prot*	*n_families*	*avg_len*	*Z-score*	*P-value*
0	Oxidoreductase	14995	992	376.63	−29.54	0
1	Transferase	26525	1606	445.17	−24.25	0
2	Lyase	7262	347	377.92	−22.64	0
3	Hydrolase	20464	1995	430.68	−21.75	0
4	Isomerase	4487	220	383.98	−14.18	0
5	Glycosidase	1826	244	444.73	−13.98	0
6	Glycosyltransferase	2950	261	437.53	−12.51	0
7	Acyltransferase	2239	179	402.83	−10.85	0
8	Methyltransferase	3524	224	349.60	−10.53	0
9	Kinase	7017	322	448.29	−10.22	0
10	Ligase	8010	230	529.41	−10.06	0
11	Decarboxylase	1293	63	345.26	−9.66	0
12	Monooxygenase	1668	73	444.87	−9.26	0
13	Metalloprotease	1100	109	553.73	−7.89	0
14	Aminopeptidase	452	39	509.17	−7.55	0
15	Dioxygenase	360	66	433.20	−7.32	0
16	Aminoacyl-tRNA synthetase	3402	37	571.83	−7.15	0
17	Protease	4423	380	549.70	−7.1	0
18	Aminotransferase	955	28	420.27	−6.02	0

Predviđeno uređene MF Ključne reči iz ovog rada (z-skor < -0.2)						
#	*name*	*n*	*avg_len*	*avg_dis*	*z*	*p*
0	Oxidoreductase	4126	472.25	0.28	-41.35	0
1	Hydrolase	7564	614.81	0.51	-26.99	0
2	Lyase	1431	481.37	0.30	-23.62	0
3	Monooxygenase	555	503.36	0.20	-20.29	0
4	Transferase	8846	631.95	0.55	-19.72	0
5	Ligase	995	693.30	0.46	-18.05	0
6	Glycosyltransferase	1134	551.26	0.40	-17.04	0
7	Glycosidase	697	570.50	0.37	-16.81	0
8	Isomerase	931	422.72	0.35	-13.60	0
9	Protease	1863	674.42	0.54	-13.20	0
10	Transducer	1703	482.28	0.41	-12.56	0
11	G-protein coupled receptor	1385	465.62	0.39	-12.45	0
12	Acyltransferase	867	531.58	0.42	-11.28	0
13	Decarboxylase	195	488.21	0.25	-10.70	0
14	Aminotransferase	202	451.05	0.24	-10.23	0
15	Aminopeptidase	130	668.72	0.37	-9.10	0
16	Serine protease	460	700.07	0.50	-8.87	0
17	Metalloprotease	507	688.25	0.56	-8.43	0
18	Methyltransferase	874	611.22	0.47	-8.33	0
19	Carboxypeptidase	116	631.16	0.37	-8.25	0
20	Threonine protease	138	246.88	0.18	-7.50	0
21	Dioxygenase	366	622.32	0.48	-7.39	0
22	Serine esterase	141	423.09	0.28	-7.21	0
23	Receptor	3424	647.92	0.59	-7.09	0
24	Nucleotidyltransferase	600	969.68	0.63	-6.18	0
25	Peroxidase	221	457.61	0.40	-5.50	0
26	Serine protease inhibitor	182	497.66	0.38	-4.92	0
27	Porin	63	318.84	0.21	-4.55	0
28	Dipeptidase	22	522.27	0.23	-4.43	0
29	Integrin	71	1038.25	0.70	-4.29	0
30	Protease inhibitor	284	446.29	0.41	-4.15	0
31	RNA-directed RNA polymerase	62	2506.11	0.87	-4.11	0
32	Neurotoxin	66	209.62	0.17	-3.80	0
33	Hemagglutinin	23	281.22	0.13	-3.37	0
34	Retinal protein	47	384.06	0.28	-3.37	0
35	Myosin	185	1275.20	0.72	-3.29	0
36	Photoreceptor protein	77	540.68	0.47	-3.02	0.01
37	Prenyltransferase	39	376.21	0.31	-2.59	0.01
38	Elongation factor	106	467.68	0.48	-2.59	0.01
39	Endonuclease	443	728.14	0.60	-2.59	0.01
40	Toxin	214	412.51	0.39	-2.56	0
41	Bacteriolytic enzyme	35	434.66	0.31	-2.53	0.01
42	Nuclease	703	668.54	0.60	-2.49	0.01
43	Ion channel impairing toxin	60	74.92	0.15	-2.47	0
44	Voltage-gated sodium channe...	22	79.55	0.05	-2.42	0.01
45	Aspartyl protease	120	942.30	0.64	-2.39	0.01
46	Hemostasis impairing toxin	55	297.55	0.35	-2.10	0.02
47	Thiol protease inhibitor	63	411.06	0.35	-2.04	0.03
48	tRNA-binding	119	575.46	0.56	-2.01	0.03

Predviđeno uređeni MF termini dobijeni iz direktnog i izvedenog mapiranja						
#	*name*	*n*	*avg_len*	*avg_dis*	*z*	*p*
0	catalytic activity	28179	569.86	0.48	-52.25	0
1	oxidoreductase activity	4986	450.49	0.30	-39.62	0
2	hydrolase activity	11097	623.94	0.53	-27.76	0
3	lyase activity	1624	480.74	0.31	-24.47	0
4	monooxygenase activity	633	511.13	0.22	-21.40	0
5	transporter activity	5388	621.16	0.53	-19.15	0
6	transferase activity	10428	619.96	0.56	-18.80	0
7	transferase activity, trans...	1343	545.86	0.41	-18.49	0
8	ligase activity	1097	675.42	0.46	-16.68	0
9	hydrolase activity, acting ...	934	548.13	0.39	-16.32	0
10	isomerase activity	1046	425.93	0.35	-14.19	0
11	transferase activity, trans...	1235	540.97	0.43	-13.31	0
12	carboxy-lyase activity	292	494.47	0.26	-13.09	0
13	peptidase activity	2211	644.11	0.53	-13.05	0
14	G-protein coupled receptor ...	1473	465.64	0.40	-12.30	0
15	carboxylic ester hydrolase ...	628	482.21	0.37	-12.01	0
16	serine-type peptidase activity	708	614.43	0.46	-10.79	0
17	aminopeptidase activity	177	633.80	0.35	-10.74	0
18	transaminase activity	230	459.43	0.26	-10.46	0
19	methyltransferase activity	947	607.08	0.48	-8.77	0
20	carboxypeptidase activity	143	596.81	0.34	-8.54	0
21	metallopeptidase activity	632	663.90	0.56	-8.44	0
22	ATP binding	7585	801.90	0.69	-8.22	0
23	dioxygenase activity	419	604.98	0.48	-7.83	0
24	peroxidase activity	282	414.57	0.34	-7.11	0
25	antioxidant activity	499	344.42	0.36	-6.89	0
26	threonine-type endopeptidas...	139	247.87	0.19	-6.79	0
27	nucleotidyltransferase acti...	892	805.20	0.59	-6.07	0
28	porin activity	76	369.24	0.22	-5.22	0
29	serine-type endopeptidase i...	218	485.65	0.40	-4.99	0
30	phospholipase A2 activity	93	377.39	0.30	-4.93	0
31	phospholipase D activity	37	778.27	0.43	-4.87	0
32	prenyltransferase activity	128	369.37	0.32	-4.70	0
33	toxin activity	285	349.98	0.32	-4.59	0
34	dipeptidase activity	41	557.73	0.34	-4.23	0
35	metalloendopeptidase activity	369	711.73	0.63	-4.22	0
36	phospholipase A2 activity (...	70	365.39	0.30	-3.90	0
37	endonuclease activity	773	658.33	0.59	-3.85	0
38	antigen binding	277	345.87	0.38	-3.68	0
39	translation elongation fact...	116	464.12	0.46	-3.35	0
40	nuclease activity	1124	659.17	0.61	-3.20	0
41	G-protein coupled photorece...	44	390.23	0.30	-2.99	0
42	chloride channel activity	195	622.49	0.62	-2.67	0.01
43	RNA-directed 5'-3' RNA poly...	75	2341.31	0.87	-2.57	0.02
44	tRNA binding	247	566.35	0.57	-2.39	0.01
45	peptidase inhibitor activity	446	493.46	0.50	-2.38	0.01
46	photoreceptor activity	91	574.22	0.54	-2.30	0.01
47	glucan endo-1,3-beta-D-gluc...	28	522.18	0.43	-2.29	0.02
48	signal transducer activity	4149	619.06	0.62	-2.17	0.02
49	aminoacyl-tRNA ligase activity	255	659.40	0.67	-1.93	0.04
50	cysteine-type endopeptidase...	262	953.41	0.64	-1.65	0.05

SLIKA 6.3: Poređenje predviđenih uređenih funkcija

a)

19 MF ključnih reči sa najznačajnijom predviđenom uređenošću (org. 2007)

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

Predviđeno uređene MF ključne reči iz ovog rada (z-skor < -0.2)

#	name	n	avg_len	avg_dis	z	p
0	Oxidoreductase	4126	472.25	0.28	-41.35	0
1	Hydrolase	7564	614.81	0.51	-26.99	0
2	Lyase	1431	481.37	0.30	-23.62	0
3	Monooxygenase	555	503.36	0.20	-20.29	0
4	Transferase	8846	631.95	0.55	-19.72	0
5	Ligase	995	693.30	0.46	-18.05	0
6	Glycosyltransferase	1134	551.26	0.40	-17.04	0
7	Glycosidase	697	570.50	0.37	-16.81	0
8	Isomerase	931	422.72	0.35	-13.60	0
9	Protease	1863	674.42	0.54	-13.20	0
10	Transducer	1703	482.28	0.41	-12.56	0
11	G-protein coupled receptor	1385	465.62	0.39	-12.45	0
12	Acyltransferase	867	531.58	0.42	-11.28	0
13	Decarboxylase	195	488.21	0.25	-10.70	0
14	Aminotransferase	202	451.05	0.24	-10.23	0
15	Aminopeptidase	130	668.72	0.37	-9.10	0
16	Serine protease	460	700.07	0.50	-8.87	0
17	Metalloprotease	507	688.25	0.56	-8.43	0
18	Methyltransferase	874	611.22	0.47	-8.33	0
19	Carboxypeptidase	116	631.16	0.37	-8.25	0
20	Threonine protease	138	246.88	0.18	-7.50	0
21	Dioxygenase	366	622.32	0.48	-7.39	0

b)

Predviđeno uređene MF ključne reči iz ovog rada (z-skor < -0.2)

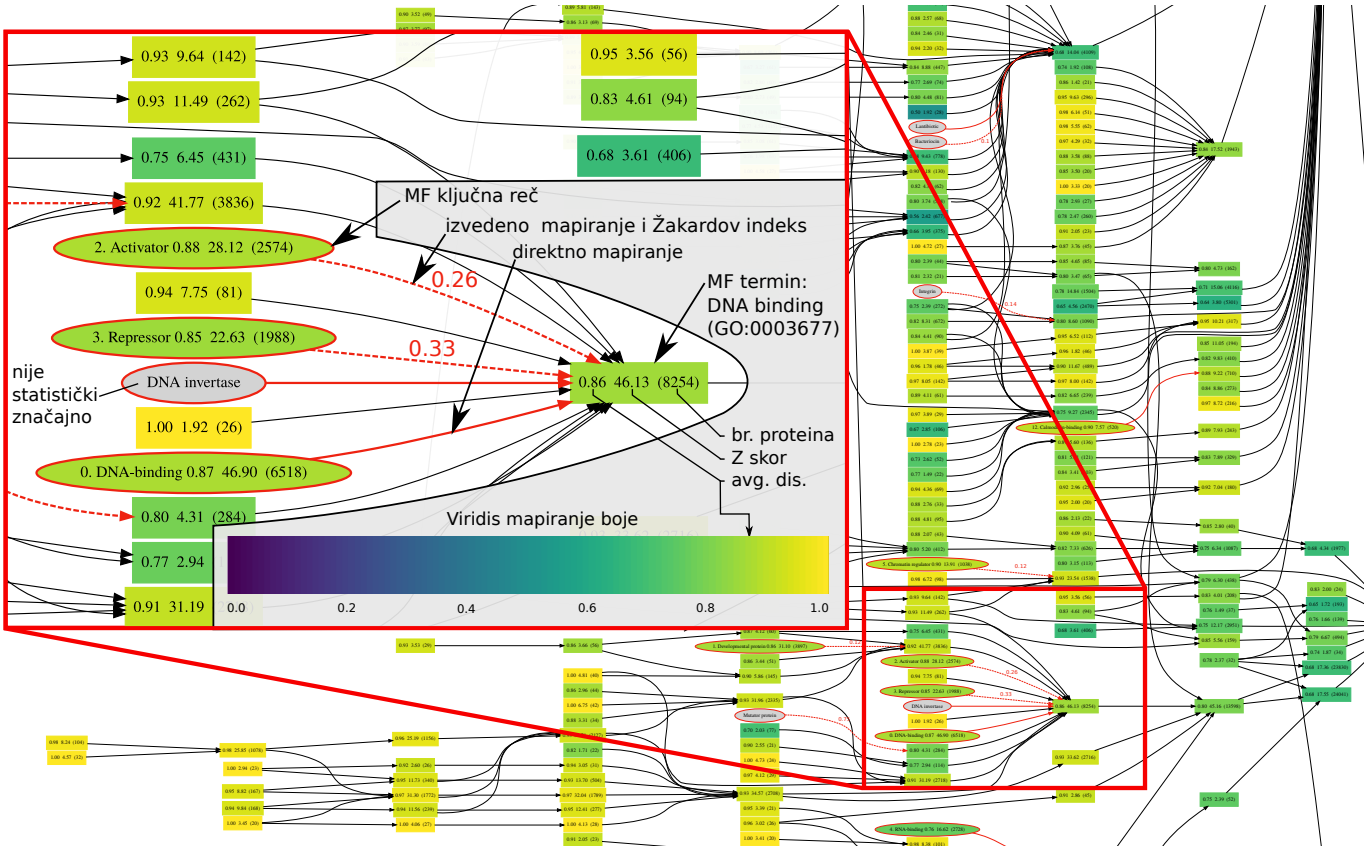
#	name	n	avg_len	avg_dis	z	p
0	Oxidoreductase	4126	472.25	0.28	-41.35	0
1	Hydrolase	7564	614.81	0.51	-26.99	0
2	Lyase	1431	481.37	0.30	-23.62	0
3	Monooxygenase	555	503.36	0.20	-20.29	0
4	Transferase	8846	631.95	0.55	-19.72	0
5	Ligase	995	693.30	0.46	-18.05	0
6	Glycosyltransferase	1134	551.26	0.40	-17.04	0
7	Glycosidase	697	570.50	0.37	-16.81	0
8	Isomerase	931	422.72	0.35	-13.60	0
9	Protease	1863	674.42	0.54	-13.20	0
10	Transducer	1703	482.28	0.41	-12.56	0
11	G-protein coupled receptor	1385	465.62	0.39	-12.45	0
12	Acyltransferase	867	531.58	0.42	-11.28	0
13	Decarboxylase	195	488.21	0.25	-10.70	0
14	Aminotransferase	202	451.05	0.24	-10.23	0
15	Aminopeptidase	130	668.72	0.37	-9.10	0
16	Serine protease	460	700.07	0.50	-8.87	0
17	Metalloprotease	507	688.25	0.56	-8.43	0
18	Methyltransferase	874	611.22	0.47	-8.33	0
19	Carboxypeptidase	116	631.16	0.37	-8.25	0
20	Threonine protease	138	246.88	0.18	-7.50	0
21	Dioxygenase	366	622.32	0.48	-7.39	0
22	Serine esterase	141	423.09	0.28	-7.21	0
23	Receptor	3424	647.92	0.59	-7.09	0
24	Nucleotidyltransferase	600	969.68	0.63	-6.18	0
25	Peroxidase	221	457.61	0.40	-5.50	0
26	Serine protease inhibitor	182	497.66	0.38	-4.92	0
27	Porin	63	318.84	0.21	-4.55	0
28	Dipeptidase	22	522.27	0.23	-4.43	0
29	Integrin	71	1038.25	0.70	-4.29	0
30	Protease inhibitor	284	446.29	0.41	-4.15	0
31	RNA-directed RNA polymerase	62	2506.11	0.87	-4.11	0
32	Neurotoxin	66	209.62	0.17	-3.80	0
33	Hemagglutinin	23	281.22	0.13	-3.37	0
34	Retinal protein	47	384.06	0.28	-3.37	0
35	Myosin	185	1275.20	0.72	-3.29	0
36	Photoreceptor protein	77	540.68	0.47	-3.02	0.01
37	Prenyltransferase	39	376.21	0.31	-2.59	0.01
38	Elongation factor	106	467.68	0.48	-2.59	0.01
39	Endonuclease	443	728.14	0.60	-2.59	0.01
40	Toxin	214	412.51	0.39	-2.56	0
41	Bacteriolytic enzyme	35	434.66	0.31	-2.53	0.01
42	Nuclease	703	668.54	0.60	-2.49	0.01
43	Ion channel impairing toxin	60	74.92	0.15	-2.47	0

Predviđeno uređeni MF termini dobijeni iz direktnog i izvedenog mapiranja

#	name	n	avg_len	avg_dis	z	p
0	catalytic activity	28179	569.86	0.48	-52.25	0
1	oxidoreductase activity	4986	450.49	0.30	-39.62	0
2	hydrolase activity	11097	623.94	0.53	-27.76	0
3	lyase activity	1624	480.74	0.31	-24.47	0
4	monooxygenase activity	633	511.13	0.22	-21.40	0
5	transporter activity	5388	621.16	0.53	-19.15	0
6	transferase activity	10428	619.96	0.56	-18.80	0
7	transferase activity, trans...	1343	545.86	0.41	-18.49	0
8	ligase activity	1097	675.42	0.46	-16.68	0
9	hydrolase activity, acting ...	934	548.13	0.39	-16.32	0
10	isomerase activity	1046	425.93	0.35	-14.19	0
11	transferase activity, trans...	1235	540.97	0.43	-13.31	0
12	carboxy-lyase activity	292	494.47	0.26	-13.09	0
13	peptidase activity	2211	644.11	0.53	-13.05	0
14	G-protein coupled receptor ...	1473	465.64	0.40	-12.30	0
15	carboxylic ester hydrolase ...	628	482.21	0.37	-12.01	0
16	serine-type peptidase activity	708	614.43	0.46	-10.79	0
17	aminopeptidase activity	177	633.80	0.35	-10.74	0
18	transaminase activity	230	459.43	0.26	-10.46	0
19	methyltransferase activity	947	607.08	0.48	-8.77	0
20	carboxypeptidase activity	143	596.81	0.34	-8.54	0
21	metallopeptidase activity	632	663.90	0.56	-8.44	0
22	ATP binding	7585	801.90	0.69	-8.22	0
23	dioxygenase activity	419	604.98	0.48	-7.83	0
24	peroxidase activity	282	414.57	0.34	-7.11	0
25	antioxidant activity	499	344.42	0.36	-6.89	0
26	threonine-type endopeptidas...	139	247.87	0.19	-6.79	0
27	nucleotidyltransferase acti...	892	805.20	0.59	-6.07	0
28	porin activity	76	369.24	0.22	-5.22	0
29	serine-type endopeptidase i...	218	485.65	0.40	-4.99	0
30	phospholipase A2 activity	93	377.39	0.30	-4.93	0
31	phospholipase D activity	37	778.27	0.43	-4.87	0
32	prenyltransferase activity	128	369.37	0.32	-4.70	0
33	toxin activity	285	349.98	0.32	-4.59	0
34	dipeptidase activity	41	557.73	0.34	-4.23	0
35	metalloendopeptidase activity	369	711.73	0.63	-4.22	0
36	phospholipase A2 activity (...	70	365.39	0.30	-3.90	0
37	endonuclease activity	773	658.33	0.59	-3.85	0
38	antigen binding	277	345.87	0.38	-3.68	0
39	translation elongation fact...	116	464.12	0.46	-3.35	0
40	nuclease activity	1124	659.17	0.61	-3.20	0
41	G-protein coupled photorece...	44	390.23	0.30	-2.99	0
42	chloride channel activity	195	622.49	0.62	-2.67	0.01
43	RNA-directed 5'-3' RNA poly...	75	2341.31	0.87	-2.57	0.02
44	tRNA binding	247	566.35	0.57	-2.39	0.01

SLIKA 6.4: Razdvojeno poređenje predviđenih uređenih funkcija. Neke tabele su skraćene.

Tabelarni pristup sa Slika 6.1 i 6.3 prikazuju samo mali podskup svih statistički značajnih MF termina. Kompletna analiza grupisanja po MF terminima predstavljena je grafovski, slikama `disorder.svg` i `order.svg` koje se nalaze na veb adresi [48]. Isečak rezultata `disorder.svg` prikazan je Slikom 6.5. Pozadinska boja funkcija kodira **neuređenost termina**, a kodirana je viridis [49] mapiranjem boja. Viridis mapiranje nulu predstavlja tamno ljubičastom a jedinicu svetlo žutom pa su tamniji, plavkasti termini uređeni dok su svetli, zelenkasti neuređeni. Rezultujuće `.svg` slike su namenjene da budu otvorene u internet pregledaču. Držanje kurzora miša iznad funkcija prikazaće dodatne informacije (ime, definiciju i sinonime), dok će levi klik miša preusmeriti korisnika na adekvatnu *AmiGO*[1] ili *UniProt* veb stranicu. Desnim klikom može da se odabere otvaranje stranice u novom tabu pregledača.



SLIKA 6.5: Isečak grafovskog prikaza rezultata (`disorder.png`).

Predloženi $P_L random$ (slučajni) model testirali smo na nivou ključnih reči izračunavši z_{rand} i p_{rand} vrednosti za sve MF ključne reči koje anotiraju bar 20 proteina. Rezultat poređenja z skor vrednosti između P_L modela (z-skor) i $P_L random$ modela (z_{rand}-skor) prikazan je na Slici 6.6. Prikazane su samo one MF ključne reči koje su statistički značajne u odnosu na p ili p_{rand} vrednosti dok su crveno obeležene one ključne reči koje su statistički značajne po jednoj ali ne i drugoj p vrednosti.

[1] *AmiGO* je skup alat za pretraživanje i prikazivanje GO termina

SLIKA 6.6: Poređenje PL i PL_{random} modela nad statistički značajnim predviđenim (ne)uređenim MF ključnim rečima

Glava 7

Diskusija

7.1 Međusobno upoređivanje MF ključnih reči

Razmotrimo prvo Sliku 6.4a koja pokazuje da se originalni rezultati i novi rezultati ne razlikuju bitno za predviđeno uređene MF ključne reči. Međutim, razlike koje postoje kao i veći izuzeci (*Kinase* i *Aminoacyl-tRNA synthetase*) mogu biti posledica drugačijih podataka, ali i modifikacije metoda originalne analize ili čak kombinacije oba faktora. Pronalaženje egzaktnog uzorka ovih razlika zahteva dodatno istraživanje koje prevazilazi obim ovog rada.

Sa druge strane, predviđeno **neuređene** MF ključne reči prikazane na Slici 6.2a imaju znatno više razlika u odnosu na originalne rezultate. Šest originalnih MF ključnih reči nije identifikovano u novim rezultatima. Ipak, *Antigen* u novoj verziji ključnih reči ne postoji dok su četiri ključne reči izbačene iz analize jer anotiraju ispod 10 CAFA3 proteina (minimum je 20). Preostala, crvena MF ključna reč *Cytokine* ima p vrednost 0.49 što je veliko odstupanje od originalnih rezultata. Značajnu razliku predstavljaju zeleno obeležene MF ključne reči koje se ne javljaju u originalnim rezultatima, a nalaze se među 20 statistički najznačajnijih novih rezultata. Među njima je preovlađujući motiv *binding*, što se ne poklapa sa originalnim rezultatima. Za razliku od GO termina, ključne reči ne sadrže datum dodavanja ili izmene pa ne možemo da proverimo da li su u pitanju nove ključne reči koje nisu postojale 2006. godine.

7.2 Upoređivanje MF ključnih reči i GO termina

Rezultati prikazni Slikama 6.2b i 6.4b sugerišu značajno poklapanje za statistički najznačajnije funkcije. Kao što je očekivano, sličnost rezultata z-skor i neuređenost tj. *avg_dis* pogotovo su izraženi za funkcije koje anotiraju slične skupove proteina. Veća nepoklapanja prisutna su prvenstveno kod ključnih reči čije je mapiranje problematično zbog razlika u nomenklaturi. Na primer, ključna reč *Bacteriolytic enzyme* nalazi se na 42. mestu dok se njen direktno mapirani MF termin *catalytic activity* nalazi na prvom mestu.

7.3 Grafovski prikaz MF termina

Grafovski prikaz na slikama `disorder.svg` i `order.svg` pruža detaljan uvid u kompleksne hijerarhijske odnose između MF termina. Ovaj prikaz otkriva strukture koje inače ne bi bile uočene. Opštije MF funkcije visoke statističke značajnosti okarakterisane su kompleksnom grafovskom strukturom potomaka. Ipak, treba naglasiti da ovu strukturu čine isključivo statistički značajni termini jer bi suprotno rezultat bio teško saglediv.

7.4 Statistička značajnost, neuređenost i broj proteina

Prosek neuređenih proteina tj. *avg_dis* (neuređenost) otkriva da funkcija ne mora da bude pretežno neuređena ili uređena da bi rezultat (F_J) bio statistički značajan. Na primer, MF termin *hydrolase activity* (Slika 6.4a) ima z-skor -27.76, međutim sadrži 53% neuređenih proteina. Ipak, prosečna dužina anotiranih proteina je 624, a verovatnoća da protein te dužine bude klasifikovan kao neuređen je 0.75. Takođe, veličina skupa anotiranih proteina (11 097 za *hydrolase activity*) povećava statističku značajnost rezultata. Verovatno je da veliki broj proteina smanjuje disperziju što vodi ka većim z-skor vrednostima. Ovo je posebno izraženo kod MF termina *catalytic activity* na Slici 6.4a. Iz ovih razloga smatramo da je neophodno uzeti sve parametre u obzir, a ne samo z-skor ili p vrednost.

7.5 $P_L random$ model

Sličnost rezultata P_L i $P_L random$ modela prikazana je na Slici 6.6. Veća odstupanja
z_{rand}-skora kod ključnih reči *Ribonucleoprotein, Ribosomal protein, Hormone* i drugih ogledaju se povećanom statističkom značajnošću. Ovo se može objasniti znatno nižom prosečnom dužinom skupa anotiranih proteina (manje od 300 AK), dok se na Slici 4.5 uočava da je $P_L random$ verovatnoća niža za proteine kraće od 300 AK. Obrnuto, uređene MF ključne reči globalno imaju niži z_{rand}-skor (veću statističku značajnost) što se takođe može objasniti kombinacijom globalno veće prosečne dužine proteina i većom $P_L random$ verovatnoćom (Slika 4.5).

MF ključna reč *Kinase* predstavlja zanimljivo odstupanje jer $P_L random$ model predviđa statistički značajnu uređenost iako je prosek neuređenih proteina 0.71 i P_L model predviđa neuređenost. Takođe, MF termin *Kinaseactivity* je statistički značajno neuređena (P_L model).

7.6 Klasifikacija neuređenog proteina

Definicija 1 neuređenosti proteina iz Potpoglavlja 4.2.1 nije uvek idealna. Na primer, ako je dužina proteina 35 AK, a neuređenost je predviđena celom dužinom, onda je jasno da nema potrebe eliminisati protein iz analize već ga treba svrstati kao neuređeni. Takođe, ako je protein dug 50 AK i sadrži predviđeni neuređeni region od 35 AK, onda treba pretpostaviti da neuređenost igra ulogu u funkciji. Ovi granični slučajevi čine suviše mali procenat proteina u CAFA3 podksupu pa je opravdano zanemariti ih. Međutim, procentualno *Swiss-Prot* sadrži značajno više kratkih proteina što nas navodi da istaknemo ovaj problem.

7.7 Nastavak istraživanja

Razvoj novih metaprediktora neuređenosti [22] svakako je razlog za nastavak istraživanja. Testiranje novih metoda analize, korišćenje drugih, većih skupova podataka i novih metaprediktora može rezultirati interesantnim rezultatima. Prikaz dobijenih rezultata bi trebalo da uključi razvoj korisničkog interfejsa koji omogućuje interaktivno istraživanje, poređenje i povezanost sa drugim resursima. Vizuelna komparacija različitih rezultata u grafovskom obliku samo je jedan od problema koje treba rešiti.

Računarsko istraživanje veze između funkcije i tipa neuređenosti je naredni pravac koji treba istražiti. Ipak, prediktori koji predviđaju tip neuređenosti prema saznanjima autora još ne postoje. Iz tog razloga buduće istraživanje treba da bude fokusirano na njihov razvoj.

Bibliografija

[1] Hongbo Xie, Slobodan Vucetic, Lilia M. Iakoucheva, Christopher J. Oldfield, A. Keith Dunker, Vladimir N. Uversky, and Zoran Obradovic. "Functional Anthology of Intrinsic Disorder. 1. Biological Processes and Functions of Proteins with Long Disordered Regions". In: *Journal of Proteome Research* 6.5 (2007), pp. 1882–1898. DOI: 10.1021/pr060392u. URL: https://www.ncbi.nlm.nih.gov/pubmed/17391014.

[2] Dejan Orčić. *Osnove biohemije (skripta)*. 2015. Chap. 3.

[3] *Slika alfa spirale i beta ploče a)*. https://www.nature.com/horizon/proteinfolding/background/figs/importance_f3.html. Pristupljeno: 24.03.2018.

[4] *Slika alfa spirale i beta ploče b)*. http://ib.bioninja.com.au/higher-level/topic-7-nucleic-acids/73-translation/protein-structure.html. Pristupljeno: 24.03.2018.

[5] *Bioinformatics*. Springer Berlin Heidelberg, 2007. Chap. 9, p. 270. DOI: 10.1007/978-3-540-69022-1. URL: https://doi.org/10.1007/978-3-540-69022-1.

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

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

[8] P. R. Romero, S. Zaidi, Y. Y. Fang, V. N. Uversky, P. Radivojac, C. J. Oldfield, M. S. Cortese, M. Sickmeier, T. LeGall, Z. Obradovic, and A. K. Dunker. "Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms". In: *Proceedings of the National Academy of Sciences* 103.22 (2006), pp. 8390–8395. DOI: 10.1073/pnas.0507916103.

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

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

[11] A.Keith Dunker, J.David Lawson, Celeste J Brown, Ryan M Williams, Pedro Romero, Jeong S Oh, Christopher J Oldfield, Andrew M Campen, Catherine M Ratliff, Kerry W Hipps, Juan Ausio, Mark S Nissen, Raymond Reeves, Chul-Hee Kang, Charles R Kissinger, Robert W Bailey, Michael D Griswold, Wah Chiu, Ethan C Garner, and Zoran Obradovic. "Intrinsically disordered protein". In: *Journal of Molecular Graphics and Modelling* 19.1 (2001), pp. 26–59. DOI:

10.1016/s1093-3263(00)00138-8. URL: https://doi.org/10.1016/s1093-3263(00)00138-8.

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

[13] Burton S. Guttman. *Biology*. William C Brown Pub, 1998, pp. 66–107. ISBN: 0697223663.

[14] Zhenling Peng, Jing Yan, Xiao Fan, Marcin J. Mizianty, Bin Xue, Kui Wang, Gang Hu, Vladimir N. Uversky, and Lukasz Kurgan. "Exceptionally abundant exceptions: comprehensive characterization of intrinsic disorder in all domains of life". In: *Cellular and Molecular Life Sciences* 72.1 (2014), pp. 137–151. DOI: 10.1007/s00018-014-1661-9. URL: https://doi.org/10.1007/s00018-014-1661-9.

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

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

[17] Rebecca B. Berlow and Peter E. Wright. "Tight complexes from disordered proteins". In: (2018). DOI: doi:10.1038/d41586-018-01694-y. URL: https://www.nature.com/articles/d41586-018-01694-y.

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

[19] Peter Tompa and Alan Fersht. *Structure and Function of Intrinsically Disordered Proteins*. Chapman and Hall/CRC, 2009. Chap. 10, 12 i 14. ISBN: 1420078925. URL: https://www.crcpress.com/Structure-and-Function-of-Intrinsically-Disordered-Proteins/Tompa-Fersht/p/book/9781420078923.

[20] Damiano Piovesan, Francesco Tabaro, Ivan Mičetić, Marco Necci, Federica Quaglia, Christopher J. Oldfield, Maria Cristina Aspromonte, Norman E. Davey, Radoslav Davidović, Zsuzsanna Dosztányi, Arne Elofsson, Alessandra Gasparini, András Hatos, Andrey V. Kajava, Lajos Kalmar, Emanuela Leonardi, Tamas Lazar, Sandra Macedo-Ribeiro, Mauricio Macossay-Castillo, Attila Meszaros, Giovanni Minervini, Nikoletta Murvai, Jordi Pujols, Daniel B. Roche, Edoardo Salladini, Eva Schad, Antoine Schramm, Beata Szabo, Agnes Tantos, Fiorella Tonello, Konstantinos D. Tsirigos, Nevena Veljković, Salvador Ventura, Wim Vranken, Per Warholm, Vladimir N. Uversky, A. Keith Dunker, Sonia Longhi, Peter Tompa, and Silvio C.E. Tosatto. "DisProt". In: 45.D1 (2016), pp. D219–D227. DOI: 10.1093/nar/gkw1056. URL: https://doi.org/10.1093/nar/gkw1056.

[21] Fanchi Meng, Vladimir N. Uversky, and Lukasz Kurgan. "Computational Prediction of Intrinsic Disorder in Proteins". In: *Current Protocols in Protein Science* (2017), pp. 2.16.1–2.16.14. DOI: 10.1002/cpps.28.

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

[23] Bo He, Kejun Wang, Yunlong Liu, Bin Xue, Vladimir N Uversky, and A Keith Dunker. "Predicting intrinsic disorder in proteins: an overview". In: *Cell Research* 19.8 (2009), pp. 929–949. DOI: 10.1038/cr.2009.87. URL: https://doi.org/10.1038/cr.2009.87.

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

[25] Chuming Chen, Hongzhan Huang, and Cathy H. Wu. "Protein Bioinformatics Databases and Resources". In: (2017), pp. 3–39. DOI: 10.1007/978-1-4939-6783-4_1.

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

[27] Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. "Gene Ontology: tool for the unification of biology". In: *Nature Genetics* 25.1 (2000), pp. 25–29. DOI: 10.1038/75556. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3037419.

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

[29] *GO veb sajt*. http://www.geneontology.org/.

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

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

[32] *UniProt veb sajt*. https://www.uniprot.org/.

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

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

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

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

[37] Damiano Piovesan, Francesco Tabaro, Lisanna Paladin, Marco Necci, Ivan Mičetić, Carlo Camilloni, Norman Davey, Zsuzsanna Dosztányi, Bálint Mészáros, Alexander M Monzon, Gustavo Parisi, Eva Schad, Pietro Sormanni, Peter Tompa, Michele Vendruscolo, Wim F Vranken, and Silvio C E Tosatto. In: 46.D1 (2017), pp. D471–D476. DOI: 10.1093/nar/gkx1071. URL: https://doi.org/10.1093/nar/gkx1071.

[38] Matt E. Oates, Pedro Romero, Takashi Ishida, Mohamed Ghalwash, Marcin J. Mizianty, Bin Xue, Zsuzsanna Dosztányi, Vladimir N. Uversky, Zoran Obradovic, Lukasz Kurgan, A. Keith Dunker, and Julian Gough. "D2P2: database of disordered protein predictions". In: *Nucleic Acids Research* 41.D1 (2012), pp. D508–D516. DOI: 10.1093/nar/gks1226. URL: https://doi.org/10.1093/nar/gks1226.

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

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

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

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

[43] D. Barrell, E. Dimmer, R. P. Huntley, D. Binns, C. ODonovan, and R. Apweiler. "The GOA database in 2009–an integrated Gene Ontology Annotation resource". In: *Nucleic Acids Research* 37.Database (2009), pp. D396–D403. DOI: 10.1093/nar/gkn803.

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

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

[46] J. L. King and T. H. Jukes. "Non-Darwinian Evolution". In: *Science* 164.3881 (1969), pp. 788–798. DOI: 10.1126/science.164.3881.788. URL: https://doi.org/10.1126/science.164.3881.788.

[47] *Adresa projekta*. https://github.com/gvinterhalter/MASTER2.

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

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