UNIVERZITET U BEOGRADU
MATEMATIČKI FAKULTET

MASTER RAD

# Računarska analiza povezanosti funkcije i neuređenosti proteina

*Autor:*
Goran VINTERHALTER

*Mentor:*
dr. Jovana KOVAČEVIĆ

ČLANOVI KOMSIJE:

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


Beograd, 2018

UNIVERZITET U BEOGRADU

# *Sažetak*

Matematički fakultet
Katedra za Računarstvo i informatiku

Master informatičar

**Računarska analiza povezanosti funkcije i neuređenosti proteina**

Goran VINTERHALTER

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

# Sadržaj

# Spisak skraćenica

| | |
|---|---|
| **LAH** | **L**ist **A**bbreviations **H**ere |
| **WSF** | **W**hat (it) **S**tands **F**or |

# Glava 1

# Uvod

## 1.1 Osnovni biološki 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 Tabli 1.1.

TABELA 1.1: Najznačajniji biološki polimeri

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

### 1.1.1 Proteini

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

[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

# 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 (Uversky, 2016). Danas neuređenost proteina je uzrokovala nastanak velikog broj hipoteza, od $D^2$ koncepta bolesti (**d2uversky2008**) pa sve do evolucije višećelijskih organizama (**romero2006**) i osobina prvih oblika života (**trifonov2000**; Uversky, 2016). Šta više, sa početkom 21. veka broj naučnih radova koji se bave ovom temom doživljava skoro eksponencijalan porast (Oldfield and Dunker, 2014),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 (pedesetih godina prošlog veka) postavila je temelje "opšteprihvaćene" **struktura-funkcija** paradigme (Dunker et al., 2001). (engl. *"The characteristic specific properties of native*[2] *proteins we attribute to their uniquely defined configurations. The denatured protein molecule we consider to be characterized by the absence of a uniquely defined configuration."*) (Mirski i Pauling) Predloženi model prilagođen je funkcionisanju enzima, čija sposobnost da katalizuju[3] zavisi od jasno definisanog geometriskog oblika koji moraju da zauzmu odnosno u koji moraju da se saviju. Substrat[4] (ključ ili funkcija) diktira oblik enzima (brave ili strukture) (Guttman, 1998). Kontrapozicijom sledi da nedostatak strukture vodi izostanku funkcije.

Prvi kontraprimer gornje teorije javio se još 1950. Protein krvne plazme, serum albumin pokazivao je veliku mogućnost vezivanja za različite partnere (Dunker et al., 2001). Ovo otkriće ukazivalo je da specifične zahteve enzima ne treba generalizovati na sve proteine. Ipak model brave i ključ i njena poboljšana varijanta **teorija indukovanog fita**[5] (engl. *induced-fit theory*) dominirale su krajem prošlog veka, zanemarujuću konstantno rastući skup funkcionalnih "ne-nativnih" proteina čije postojanje nisu mogle da objasne. Sa druge strane tehnološki napretci u razlučivanju strukture proteina jasno su demonstrirali obimno postojanje funkcionalnih proteina bez uređene 3D strukture (pri fiziološkim uslovima) od kojih su neki bili neuređeni celom dužinom (Dunker et al., 2001). Nova paradigma je bila neophodan.

Hipoteza proteinskog "trojstva" (Dunker et al., 2001) (nastala tek početkom 21 veka) predlaže da funkcija proteina može zavisiti od bilo kojeg od "tri" stanja ili tranzicije

[1] Emil Fišer bio je Nemački hemičar koji je 1894. predložio analogiju brave i ključa opisujući karakteristike enzima pivske plesni (Dunker et al., 2001).

[2] nativno stanje proteina je savijeno, operativno, funkcionalno stanje (Dunker et al., 2001). Ovaj termin bio je isprepleten sa struktura-funkcija paradigmom

[3] katalizacija podrazumeva ubrzavanje ili omogućavanje hemijske reakcije sa substratom (Guttman, 1998)

[4] substrat je molekul sa kojim enzim deluje (Guttman, 1998)

[5] Teorija indukovanog fita omekšava rigidnost brava-ključ model sugerišući da interakcija sa substratom indukuje konačni oblik enzima maksimizujući reakciju (Guttman, 1998)

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
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) (Dunker et al., 2001) model je dopunjen.
4. **Nasumično klupko** (engl. *random coil*) - gasovito stanje

Povezanost sekvence sa strukturom sugeriše da je neuređenost enkodirano inherentno[6] svojstvo (Dunker et al., 2001) stoga ove proteine nazivamo **Inherentno**[7] **Neuređeni Proteini** (engl. *Intrinsically Disorderd Proteins*) skraćeno **IDP**, a njihove neuređene ali funkcionalne regione **IDPr** (Uversky, 2016). U ovom radu pod neuređenošću proteina podrazumevaćemo inherentnu neuređenost osim ako to nije drugačije naglašeno[8].

Današnje procene zastupljenosti pronašle su da 19% aminokiselina kod eukariota, 6% kod bakterija i 4% kod arhea pripadaju IDPr (**peng2015b**). Čak 50% proteina eukariota ima bar jedan IDPr duži ili jednak od 30 uzastopnih AK (**Xue2012**) dok je za 6% do 17% predviđeno da su neurđeni celom dužinom (**tompa2002**). 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 prevazilazi obim rada zalazeći u biohemiju i biofiziku. Takođe, maloistraženi potencijal ove oblasti proizvodi veliku količinu novih saznanja. U časopisu *Nature* objavljen je rad (Berlow and Wright, 2018) koji kratko sumira najnovija saznanja koja fundamentalno menjaju poglede na mogućnost "jakog vezivanja potpuno neuređenih proteina u dinamične komplekse". Iz tih razloga navodimo samo globalne osobine IDP i IDPr kao i osobine relevantne za naš rad.

- Neuređenost je inherentno svojstvo sekvence (Dunker et al., 2001). Pokazano je da nisko očekivanje indeksa hidropatije[9] zajedno sa visokim ukupnim nabojem predstavlja bitan preduslov koji sprečava savijanje proteina u fiziološkim uslovima (Uversky, 2016). 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*). (Oldfield and Dunker, 2014; Uversky, 2016) Opisane osobine daju validnost primeni mašinskog učenja u predviđanju neuređenih regiona proteina (Oldfield and Dunker, 2014).
- PTM**??** proteina 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 (Uversky, 2016). Iako je

[6] Inherentno ili prirođeno, nasleđeno

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

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

[9] mera hidrofobnosti

PTM povezano sa neuređenošću i sugeriše beskrajne uticaje na funkciju proteina (Uversky, 2016) kompleksnost ove teme prevazilazi obime ovog istraživanja.

- IDP i IDPr su po zastupljenosti AK prostije[10] sekvence u poređenju sa domenima savijenih proteina. Ipak usled manje restrikcija (obaveznog savijanja) mogućnost interakcije sa više partnera je mnogo veća što moguće funkcije čini raznolikim (Uversky, 2016). Pomenuta interakcija kod nekih neuređenih proteina vodi do njihovog potpunog ili parcijalnog savijanja dok neki i dalje ostaju neuređeni (Uversky, 2016). Bukvalna evoluciona 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 (**oldfield20014**; Dunker et al., 2001) ali su (neki pokušaji su napravljeni (Dunker et al., 2001)). Najuopšteniji opis strukture ovih proteina dat je kao **kombinacija različitih tipova foldona**[11] (Uversky, 2016):

  - **foldon** (engl. *foldon*) je nezavisno organizujuća jedinica (region) proteina.
  - **indukativni foldon** (engl. *inducible foldon*) je IDPr koji savijanje postiže barem delom vezivajući se za partnera.
  - **ne-foldon** (engl. *non-foldon*) je IDPr koji nikad ne postiže uređenost.
  - **polu-foldon** (engl. *semi-foldon*) je IDPr koji ostaje polovično neuređen i nakon vezivanja za partnera.
  - **anti-foldon** (engl. *unfoldons*) je region proteina koji iz uređenog prelazi u neuređeno stanje u cilju vršenja funkcije.

- Gore pomenut opšti prikaz strukture nastao je iz raznih opažanja interakcije, prvestveno vezivanja proteina za partnere. Za detaljan opis i iscrpnu listu ovih i drugih pojava preporučujemo čitanje (**a2z**; Uversky, 2016) kao i poglavlja 10, 12 i 14 iz knjige (engl. *Structure and Function of Intrinsically Disordered Proteins by Peter Tompa, Alan Fersht*).

## 2.2 Funkcija proteina

Funkcija proteina može biti sagledana iz tri ugla: molekulske funkcije, biološkog procesa kome pripada i lokacije u ćeliji gde se funkcija odvija (Ashburner et al., 2000)(postoje i drugi sistemi klasifikacije**??**). Kako je cilj ovog rada molekulska funkcija Tabelom2.1 ukratko navodimo (bez poretka) ustanovljene (Xie et al., 2007) molekulske funkcije koje se pripisuju (ne)uređenosti proteina. Ovo su takođe rezultati za koje se nadamo da će naše istraživanje potvrditi.

TABELA 2.1: Odnos molekulske funkcije i uređenosti

**Ovo možda za kraj ostaviti...**
Novija istraživanja nad eksperimentalno dokazanih IDP i IDPr dovela su do kreiranja ontologija(po ugledu na GO (**GO2000**)) za opis funkcija neuređenih proteina. Ontologije su sastavni deo DisProt (**disprot7**) baze eksperimentalno dokazanih IDP i IDPr... novi prediktori postoje...

[10] prostije u smislu da sadrže manje informacija (Šenonov indeks)
[11] Zbog nove prirode termina i manjka prevedene literature autor je odlučio da usvoji naziv u originalu.

## 2.3 Eksperimentalno ispitivanje neuređenosti

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

## 2.4 Predikcija neuređenosti

Do danas napravljeno je preko 60 prediktora inherentno neuređenih proteina (**meng2017**). U radu (**meng_c2017**) hronološkim redosledom prikazane su karakteristike i dostupnost tridesetak popularnih prediktora.

Istorijski posmatrano razlikujemo tri epohe razvoja: (**meng_c2017**)

- Prva generacija (1979[12]-2001) Prvi prediktori oslanjali su se na razne fizičko-hemijske osobine proteina uključujući i svojstva**??** aminokiselina:
- Druga generacija (2002-2006)
  Ovaj period okarakterisan je korišćenjem relativno jednostavnih prediktivnih ML modela koji koriste isključivo svojstava AK ulazne sekvence[13].
- Treća generacija (2007-)
  Prediktori današnjice koriste komplikovanije ML modele. Uglavnom se podrazumeva meta-prediktor koji kombinuju rezultate nekoliko običnih ML modela. Recimo kombinacija NN, SVM i K-najblilžih suseda tehnikom glasanja.

Po arhitekturi predikotre delimo u četiri kategorije: (**meng_c2017**)

1. scorring function based
2. ML metode
3. Meta-prediktori
4. Predikcije na osnovu strukture[14].

### 2.4.1 Evaluacija ML modela

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

[12] Nakon 1979 drugi (prvi ozbiljni) prediktor nastao je tek 1997. (**meng_c2017**)

[13] Takođe javljaju se prediktori koji koriste evolutivne profile sekvence (PSSM skoring matrice) dobijene PSI-BLAST pretragom

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

### 2.4.2 PONDR familija prediktora i VSL2b

PONDR familija (engl. *Predictors of Natural Disordered Regions*) je grupa prediktora druge generacije zasnovanih na neuronskim mrežama, kraće NN . Neuronske mreže sa propagacijom unapred (engl. *feed forward NN*) sa veličinom prozora između 9 i 21 AK trenirane su na različitim trening skupovima proteinskih sekvenci. Finalni prediktor predstavlja kombinaciju nekoliko neuronskih mreža od kojih je svaka specijalizovana za regione određene dužine ili položaja. PONDR familija ima nekoliko prediktora koji se razlikuju u načinu treniranja što je postignuto kombinacijom pomenutih trening skupova. Oznaka "VSL2b" 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 regionim ($<$ 30 AK)
- L - Prediktor je treniran na skupu proteina sa **dugim** neuređenim regionima ($>$ 30 AK)

Tokom CASP7 takmičenja 2008. VSL2b je evaluiran kao prediktor sa ukupnim najtačnijim predviđanjima (**bohe2009**). Međutim, po današnjim merilima (**meng2017**) 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) naše istraživanje temelji se upravo na njemu.

VSL2b kao ulaz prima proteinsku sekvencu[15] minimalne dužine 9 AK kodiranih jednim karakterom. Podržava azbuku od 20 standardnih AK. Izlaz je niz ocena (verovatnoća) za svaku poziciju sekvence[16] koje govore da li je pozicija uređena ili neuređena. Pozicija sa vrednostima iznad 0.5 smatra se neuređenim, a suprotno uređenim.

[15] Ulaz VSL2b može biti i evolutivni profil što poboljšava rezultat, međutim zbog dodatnog koraka PSI-BLAST pretrage ovaju pristup nije korišćen. (posledice ???)

[16] Autori često koriste termin "ostatak" (engl. *residue*) kada misle na vrednost neke poziciju u sekvenci (polimeru). Kod aminokiselina "ostatak" se odnosi na R grupu po kojoj razlikujemo aminokiseline.

# Glava 3

# Baze 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 procedure i eksponencijalno povećavajući količinu dostupnih sekvenci. Da bi se razumeo uticaj NGS tehnologije razmotrimo sledeći tok događaja. Od sveže sekvencionisanih nepoznatih genoma predviđaju se potencijalni geni, od gena potencijalne proteinske sekvence. Dobijene proteinske sekvence mogu se dalje klasterovati u familije, automatski anotirati, predviđati im se struktura, osobine itd. Zatim, moguće je vršiti analize za generisanje novih bioloških znanja. Povezanost između funkcije i neuređenosti proteina je jedan primer biološkog znanja. Dakle generisanje novih informacija u jednoj oblasti (u ovom slučaju genomici) propagira se u druge oblasti bioinformatike. Ovo je samo jedan primer ali ilustruje dve bitne stvari:

1. Informacije eksponencijalno rastu uvodeći čitavu oblast **omike**[1] (engl. *omics*) u teritoriju (engl. *Big Data*) (Chen, Huang, and Wu, 2017). (U našem radu pažljivo su odabrani podaci malog obima kako bi se izbegao ovaj scenario i sve analize su urađene na klasičnom kućnom računaru.)

2. Velika povezanost između bioloških podataka.

Povezanost podataka preslikava se na baze. 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 kao što je UniProtKB, pored primarnog sadržaja održavaju i veliki broj referenci (dbxref) pokušavajući da međusobno povežu sve dostupne informacije. Konkrentno UniprotKB (feb. 2018) održava reference ka čak 164 različite baze[2]. Dakle, bioinformatika kao disciplina podrazumeva da će analize biti vršena kombinacijom informacija nekoliko različitih baza. Zbog raznovrsnosti i svrhe prikupljenih informacija postoji veliki broj kategorija[3](vrsta) baza. Na adresi `www.proteininformationresource.org/staff/chenc/MiMB/dbSummary2015.html` kategorizovane su i prikazane kvalitenije proteinski orijentisane baza (prikazana lista nipošto nije konačna) (Chen, Huang, and Wu, 2017). Za naše istraživanje bile su potrebne naredne tri kategorije:

- Baze sekvenci.
  Ove baze teoretski sadrže sve poznate sekvence i kontrolišu dodeljivanje identifikacionog broja sekvence.

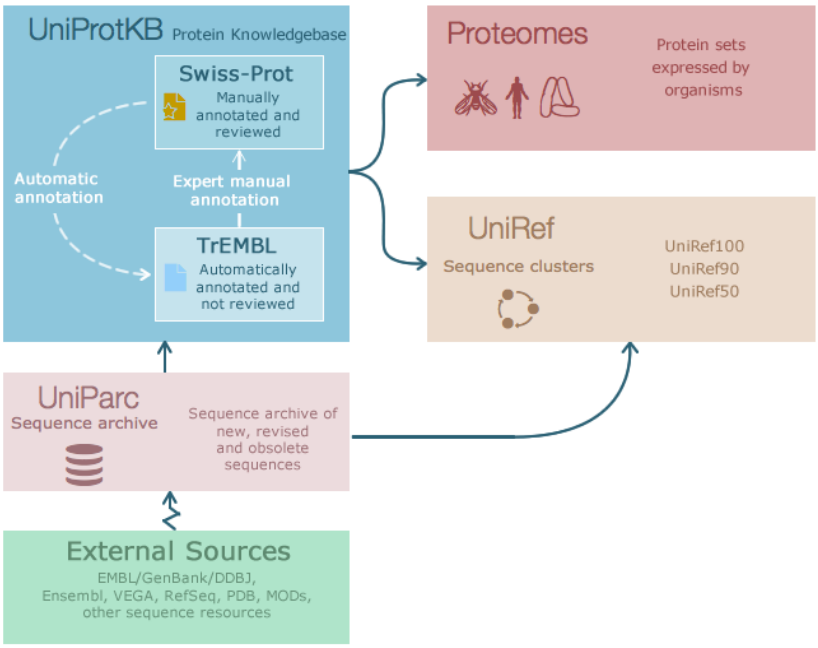
[1] termin objedinjuje gen**omiku**, proteomiku, transkriptomiku, glikomiku...
[2] `www.uniprot.org/docs/dbxref`
[3] Baze ne pripadaju ekskluzivno samo jednoj kategoriji

  - Proteinske sekvence: UniProtKB
  - DNK sekvence: (EMBL, GenBank, DDBJ)
- Baze strukture: DisProt, D2P2, MobiDB, (PDB), ...
- Baze homologija: Gene Ontology, (Protein Ontology)

## 3.1 UniProtKB/Svis-Prot

**UniProt** skraćeno od (engl. *Universal Protein Resource*) je konzorcijum nastao 2002. izmedju tri organizacije: Evropski Bioinformatički Institut (EBI), Švajcarski institut za Bioinformatiku (SIB) i Resurs Proteinskih Informacija (PIR). "Misija UniProt-a je da naučnoj zajednici obezbedi sveobuhvatan, visokokvalitetan i slobodno dostupan resurs proteinskih sekvenci i funkcionalnih informacija."[4]

UniProt obuhvata nekoliko baza i podbaza sa striktno definisanim tokom informacija Slika 3.1. Od prikazanih najbitnija je **UniprotKB** (engl. *UniProt Knowledge Base*) sačinjena od 2 podbaze.




SLIKA 3.1: Šematski prikaz povezanosti UniProt baza **??**

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

2. TrEMBL (engl. *Translated EMBL*) je nadskup Svis-Prot sekvenci čije su sekvence dobijene prevođenjem EMBL nukleinskih sekvenci, ali još nisu stigle da budu ručno proverene. Ove sekvence su redundantne i njihova obimnost posledica je masovne primene NGS tehnologija. U februaru 2018. god TrEMBLE sadržao je 107 627 435 sekvenci što je oko 200 puta više u poređenju sa 556 568 ručno

[4] www.uniprot.org

proverenih Svis-Prot sekvenci. Sve nove sekvence prvo ulaze u sastav TrEMBL da bi ručnom proverom napredovale u Svis-Prot što je donekle prikazano na Slici 3.1.

Distribucije Svis-Prot baze dostupne su u nekoliko tekstualnih formata: ravna datoteka (engl. *flat file*), XML, RDF/XML. Ravni tekstualni format zbog standardizacije prati format EMBL (**embl**) baze (Boeckmann, 2003). Unos u bazu se zove **slog** (engl. *record*) i sadrži sve informacije vezane za jedan protein. Jedan slog ilustrovaćemo uprošćenim primerom**??** u formatu ravne datoteke na kome ilustrujemo ključne osobine Svis-Prot baze:

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 (**www_svisprot**).

2. Identifikacioni broj predstavlja **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 (**www_svisprot**; Boeckmann, 2003):

   - 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 TrEML, GO mapiranje za Svis-Prot sekvence određuju se ručno (**www_svisprot**).

4. **Ključne reči** (engl. *keywords*) označene **KW** opisuju hijearhisku strukturu kontrolisanog vokabulara namenjenog opisivanju funkcije proteina. Postoji 10 kategorija ključnih reči od kojih je za naše istraživanje bitna "Molekulska funkcija"(Boeckmann, 2003). Za razliku od GO termina ključne reči prligaođene su opisivanju sadržaja isključivo Svis-Prot proteina (**www_svisprot**).

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 čuvaju različite osobine kanonske sekvence uključujući i razlike u odnosu na izoforme[5] sekvence. U našoj analizi korišćena je isključivo kanonska sekvenca. Detaljan opis pravila za biranje kanonske sekvence može se naći na (**www_svisprot**).

6. Svis-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 (*How redundant are the UniProt databases?*).

7. Postojnost proteina **PE** (engl. *Protein existance*) opisuje stepen sigurnosti da protein postoji **??**.

[5]Izoforme su alternativni oblici sekvence nastali usled: (engl. *alterntive promoter usage, alternative splicing, alternative initiation and ribosomal frameshifiting*)

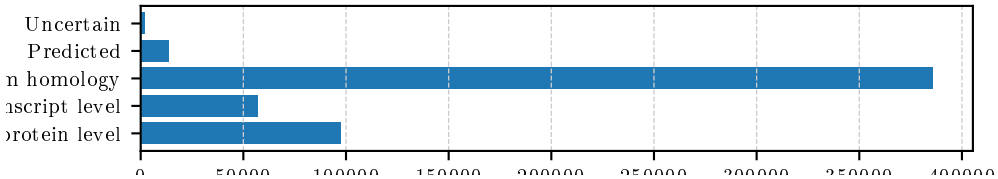
```
1  ID   ACSA_DROME              Reviewed;         670 AA.   | ime sloga, info
2  AC   Q9VP61; Q24226; Q8IH30; Q9VP60;                      | identifikacija
3  DT   19-SEP-2003, integrated into UniProtKB/Swiss-Prot. | ulazak u Svis-Prot
4  DT   01-MAY-2000, sequence version 1.                    | 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 3.2: Uprošćen primer sloga (unosa) u Svis-Prot

8. Swis-Prot takđe vrši predikcije neuređenih regiona: koristeći DISOPRED2 and CLADIST prediktor (**meng_c2017**)
   Međutim ove informacije postale su irelevantne pojavom baza MobiDB i D2P2 koje razmatramo u narednim sekcijama.

9. Zanimljiva zapažanja globalne statistike:

   - Najzastupljenije sekvence su kraće od 500 aminokiselina.
   - Postojnost oko 70% proteina potvrđeno je homologijom.
   - Zastupljeno je preko 1000 različitih organizama međutim većina Svis-Prot sekvenci pripada malom broju model organizama.




SLIKA 3.3: Histogram nivoa pouzdanosti Svis-Prot proteina

## 3.2 Disprot

Baza proteinskog neuređenja (engl. *Database of protein Disorder (DisProt)*)

## 3.3 D2P2 i MobiDB

Baze predviđenog neuređenja proteinskih sekvenci

## 3.4 Ontologije gena

**Ontologija Gena** (engl. *Gene Ontology*) ili skraćeno **GO**, predstavlja izračunato znanje o funkciji gena odnosno genskog produkta (protein, nekodirajuća RNK ili makromolekulski kompleks) (GO Consortium, 2016). 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 Svis-Prot baze[6].

Ontologija gena definiše univerzum 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 ontologija nazivaju imenski prostori (engl. *namespace*):

- **Molekulska funkcija (MF)** je biohemijska aktivnost (uključujući specifično vezivanje za ligande ili strukture) genskog produkta.

[6] Ali Svis-Prot koristi anotacije iz ontologije gena

- **Ćelijske komponente (CC)** se odnosi na mesto u ćeliji gde je genski produkt aktivan.
- **Biološki procesi (BP)** se odnose na procese kome genski produkt doprinosi.

Inspirisani sličnošću prva tri sekvencirana eukariotska organizma, GO projekat nastao je sa ciljem da unifikuje biologiju pod jedan univerzum termina za opis genskih proizvoda svih vrsta organizama (**GO2000**). Ovaj ideal je najveća razlika u odnosu na kontrolisani vokabular Svis-Prot ključnih reči koji je prilagođen za opis samo proteina sadržanih u Svis-Prot bazi.

Suštinu ontologije čine relacije između termina i pravila dedukcije koja se nad njima mogu primenjivati. Osnovnu strukturu ontologije čini direktni aciklički graf (engl. *DAG*) obrazovan roditeljskom vezom (relacijom) **is_a**. Prateći ovu relaciju termini jednog imenskog prostora recimo MF neće nikad preći u druga dva CC i BP. Ontologija stoga ima tri korena čvora MF, CC i BP (*Ontology Structure*). Primer strukture prikazan je na Slici 3.4. Pored **is_a** postoje dodatne relacije od kojih su najčešće [7]:

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

Svaka veza (relacija) ima strogo definisana pravila kompozicije koja omogućavaju automatsko rezonovanje. Recimo relacija **is_a** ima svojstvo tranzitivnosti (*Ontology Relations*):

```
A is_a    B  /\  B is_A C   =>    A is_a    C
A part_of B  /\  B is_A C   =>    A part_of C
```

Siže pravila rezonovanja prikazano je na Slici 3.5.

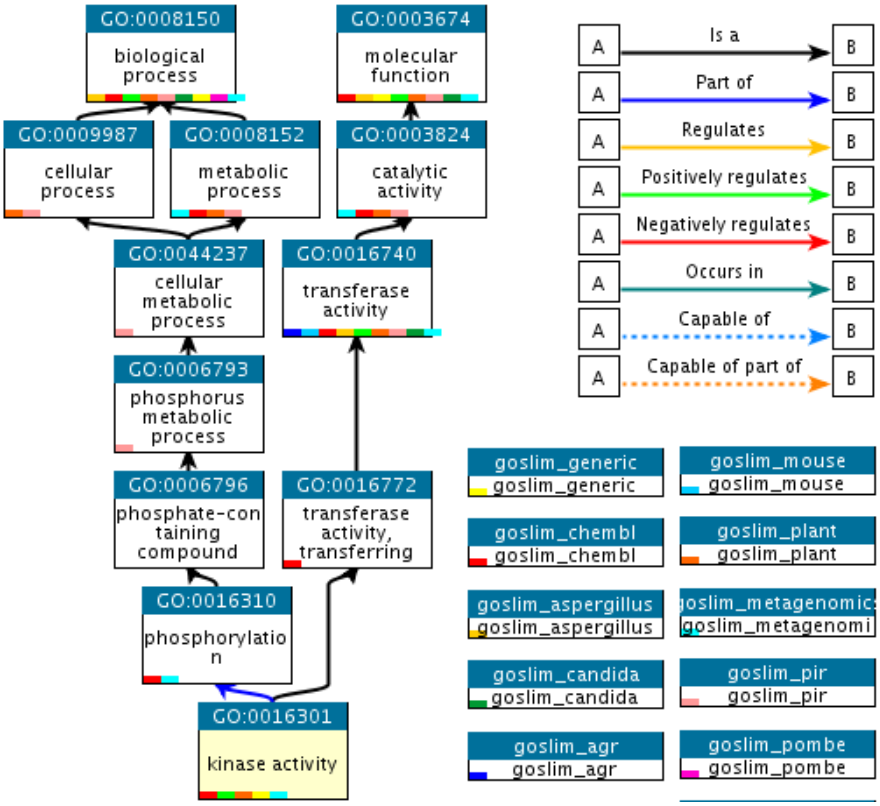
Ontologije su dostupne u nekoliko formata. U našem radu korišćen je ravni `.obo` format. Pored njega treba naglasiti postojanje `RDF/XML` i `OWL` verzija. Ove verzije namenjene su automatskom rezonovanju unutar specijalizovanih softvera i upitnih jezika [8] (protégé, SPARQL, ...).

GO termin može biti zastareo u kom slučaju se relacijom **replaced_by** pokazuje na noviji termin. Relacija **consider** ukazuju na postojanje mogućih ekvivalentnih termina. Pored glavnog univerzuma postoje i podskupovi [9] termina (GO slim) prikazani u donjem desnom delu Slike 3.4.

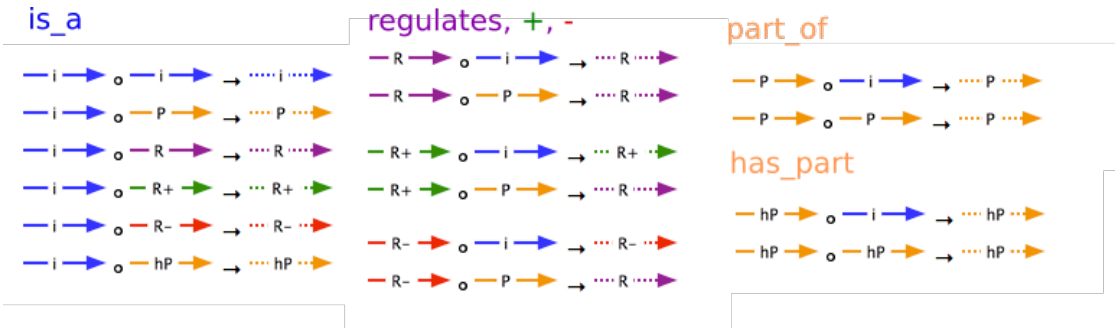
### 3.4.1 Molekulska funkcija

TODO, treba proučiti (*Molecular Function Ontology Guidelines*) možda nešto saznam o kvalitetu anotacija u Svis-Prot bazi.

[7] Vremenom se ontologija proširuje novim tipovima relacije koje su van okvira ovog rada.
[8] U našem radu korišćena je Neo4j grafovska bazu što naš postupak rezonovanja čini eksplicitnim
[9] Uglavnom ovi podskupovi predstavljaju model organizme




SLIKA 3.4: Struktura ontologije




SLIKA 3.5: Pravila rezonovanja (isprekidane relacije su rezultat)

# Glava 4

# Podaci i metode

## 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 je potrebno grupisati ih)

### 4.1.1 Podaci iz originalnog rada

U originalnom radu (Xie et al., 2007) korišćena je baza **Svis-Prot** Poglavlje 3.1, 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 značajnost posmatrane su one ključne reči kojima je bilo anotirano barem 20 proteina, tj. 710 ključnih reči.

Kao što smo pomenulli u Poglavlju 3.1 kanonske sekvence(proteini) u Svis-Prot bazi "nisu redundantne" u smislu da jedan unos u bazi predstavlja produkt jednog gena iz jedne vrste organizma. Međutim za analizu funkcija Svis-Prot **jeste statistički redundantna** (**proveriti**) jer sadrže veliku količinu **homologih** proteina (prvenstveno ortologa). Autori rada (Xie et al., 2007) su izvršili klasterovanje Svis-Prot proteina u **proteinske familije** dobivši 27 217 familija. Pri klasterovanju svaki protein ima težinu kojom doprinosi daljoj analizi. Težina svakog proteina u preseku klastera sa datom funkcijom je inverzno proporcionalna veličini preseka tako da je zbir težina svih proteina jednaka veličini preseka.

### 4.1.2 Naši podaci

U našem radu korišćen je skup proteina preuzet sa **CAFA3** takmičenja. Ovaj skup je namenjen da bude trening skup za predikciju funkcija proteina (*CAFA*).

CAFA3 trening skup je pažljivo odabran podskup Svis-Prot baze (iz 2016.) koji uključuje sve proteine iz model organizama: (engl. *Human, Mouse, Rat, S. cerevisiae, S. pombe, E. coli, A. thaliana, Dictyostelium discoideum, Zebrafish, & Bacillus cereus.* ) sa izuzetokm sekvenci (engl. *Drosophila and Candida*) koje su preuzete iz svojih respektivnih genomskih baza (Lična komunikacija sa Iddo Friedberg, PhD iz CAFA tima)

Iako ovaj pristup potencijalno proizvodi skup koji je **statisički redundantan** u našoj analizi smo pretpostavili da to nije slučaj jer je čin klasterovanja veoma računarski

zahtevan, a nismo ubeđeni da je neophodan za ovaj konkretan skup. Iz tog razloga u daljoj analizi predstavljamo uprošćenu verziju formule koja ne uračunva težinu pojedinačnog proteina.

Svis-Prot proteini su kodirani jednim karakterom koristeći **IUPAC** kodove. U podacima postoje sekvence sa nestandardnim aminokiselinam 'U' i 'O' ili višeznačnim oznakama 'B', 'J', 'X' i 'Z'. Ovakve sekvence nisu podržane od strane izabranog prediktora i za nas predstavljaju nevalidne proteinske sekvence. Pod **validnom proteinskom sekvencom** smatraćemo sekvencu koja je validan ulaz za prediktor, tj. sačinjena je od zbuke od 20 standardnih aminokiselina i ima najmanju dužinu 9[1].

CAFA3 Podaci se sastoje od dve datoteke:

1. `uniprot_sprot_exp.fasta` sadrži 66 841 protein od kojih 66 599 za našu analizu predstavljaju validnu proteinsku sekvencu. Od preostalih proteina 66.063 ima dužinu veću ili jednaku od 40 aminokiselina.

2. `uniprot_sprot_exp.txt` pridružuje funkcije u obliku **GO termina**. Zastupljeni su termini iz sva tri imenska prostora: 16.117 ćelijskih komponenti, 5 966 molekulskih funkcija i 16 117 bioloških procesa. Jednom proteinu može biti pridruženo više GO termina i obrnuto.

Naša analiza primarno je orijentisna na korišćenje GO termina za opis funkcije i razlikuje se od originalnog pristupa. Analiza sa GO terminima (grupisanje po funkciji) zahteva prvenstveno poznavanje *IS_A* roditeljske veze između termina. Takođe tokom istraživanja bile su nam potebne i ostale informacije o terminima. Pomenute informacije dobili smo iz `http://purl.obolibrary.org/obo/go.obo` dokumenta verzije 2017-12-01.

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

- `www.uniprot.org/docs/keywlist.txt` verzija 20.12.2017 sadrži detaljan opis 1188 ključnih reči od kojih 195 pripada kategoriji **Molekulsih funkcija**. Od 195 samo 145 ima mapiranje na jedan ili više GO termina.

- `ttp://geneontology.org/external2go/uniprotkb_kw2go` sadrži samo mapiranja i generiše ih **GOA projekat** (Barrell et al., 2009). Ipak ova mapiranja nisu korišćena jer ... **TODO**

Pošto je originalni rad (Xie et al., 2007) iz 2007. godine postoji razlike u vokabularu ključnih reči, razlike u samim sekvencama proteina, broj proteina i razume se anotacije ključnih reči na proteine. Iz tog razloga bilo je potrebno prvo ponoviti analizu sa vokabularom ključnih reči da bi se procenilo koliko ove razlike utiču na originalne rezultate ovog rada (Xie et al., 2007).

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

[1] Dužina 9 je minimum za VSL2b prediktor koji koristimo

## 4.2 Metod

Cilj rada je ispitivanje veze između molekulske funkcije proteina i njegove (ne)uređenosti tj. da li molekulska funkcija zavisi više od uređenosti ili neuređenosti.

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

**Realnost.**

- Broj eksperimentalno određenih neuređenih regiona je veoma mali. **Disprot baza** eksperimentalno utvrđenih neuređenih regiona ima svega 803 proteina sa opisanih 2167 neuređenih regiona (**disprot7**). Još gore pouzdanost ovih regiona je diskutabilna jer različite eksperimentalne tehnike koje su korišćene imaju različitu pouzdanost. Najveću pouzdanost nose regioni koji su eksperimentlano utvrđeni sa većim brojem eksperimentalnih tehnika[2] (**disprot7**).
- Prediktori su trenirani na malom podskupu proteina iz Disprot i PDB baze. Čak i konsenzus nekoliko različitih prediktora ne daje dovoljno pouzdane rezultate o lokaciji neuređenog regiona (**Mitic**).
- Pozitivna strana je najnoviji napredak, razvoj prediktora koji direktno pokušavaju da predvide funkciju koju IDPr obavlja (**meng_c20017**).

Jednostavna alternativa je da se pretpostavi da veći udeo neuređenih u odnosu na uređene proteine podrazumeva da funkcija zavisi više od neuređenosti. Dakle izjednačavamo uzročnost (engl. *causation*) i **korelaciju**. Međutim prvo je potrebno definisati kada protein smatramo neuređenim. Definicija mora da ima biološkog smisla, da bude prilagođena analizi, ali pored takođe ograničena je sposobnostima i preciznošću prediktora koji se korist. Više o tome u nastavku.

### 4.2.1 Predikcija neuređenosti proteina

Autori (Xie et al., 2007) koristili su **PONDR VL3E** prediktor koji postiže tačnost od 87% pri unakrsnoj validaciji nad uravnoteženim test skupom. Zbog ekonomičnosti i dostupnosti u našem radu korišćen je noviji prediktor druge generacije **PONDR VSL2b**. Relevantne karakteristike VSL2b detaljno su opisane u 2.4.2. Za potrebe analize autori (Xie et al., 2007) uvode sledeću definiciju:

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

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

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

Uslov "$\geq$ 40" u originalnom radu delom je posledica ograničenja VL3 prediktora koji je treniran na **dugim** sekvencama[3]. Mi nismo u obavezi da sledimo ovo pravilo, ali ga sledimo radi upoređivanja rezultata.

[2] Nisu ni sve eksperimentalne tehnike podjednako pouzdane

[3] L označava duge regione, $\geq$ 30 AK

### 4.2.2 Zavisnost dužine proteina i predikcije dugačkog neuređenog regiona

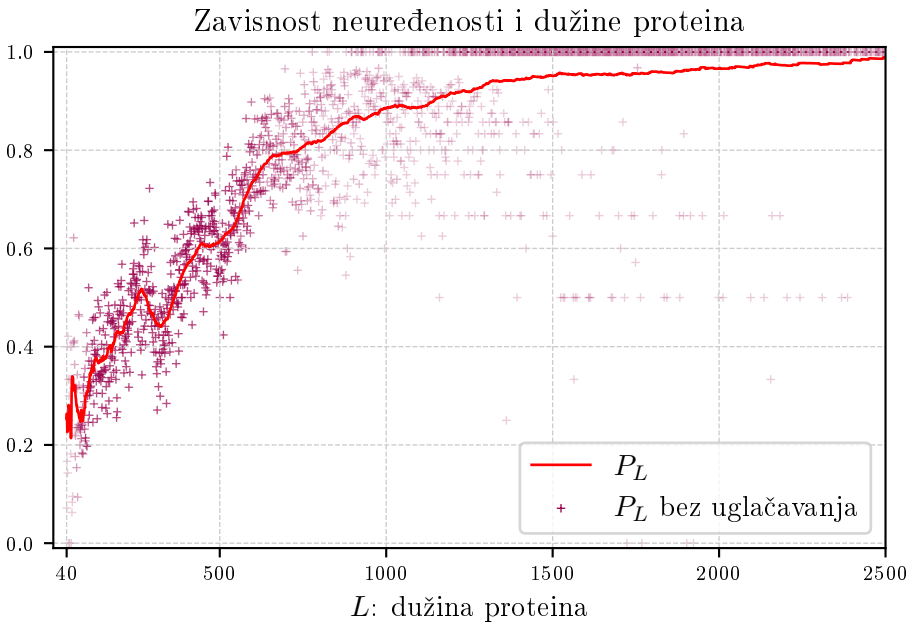
Verovatnoća da po gornjoj definiciji protein bude klasifikovan kao verovatno neuređen raste sa porastom njegove dužine. Ovo je ozbiljan problem koji utiče na statističku značajnost rezultata. Autori (Xie et al., 2007) predlažu narednu formulu da se ta verovatnoća proceni:

Neka je $S_L$ skup proteina sa dužinama između $[L-l, L+l]$ gde je $l = 0.1 * L$. Dobijamo sledeće formule:

$$S_L = \{ s_i \mid \quad |L - \|s_i\|| <= l \quad \}$$

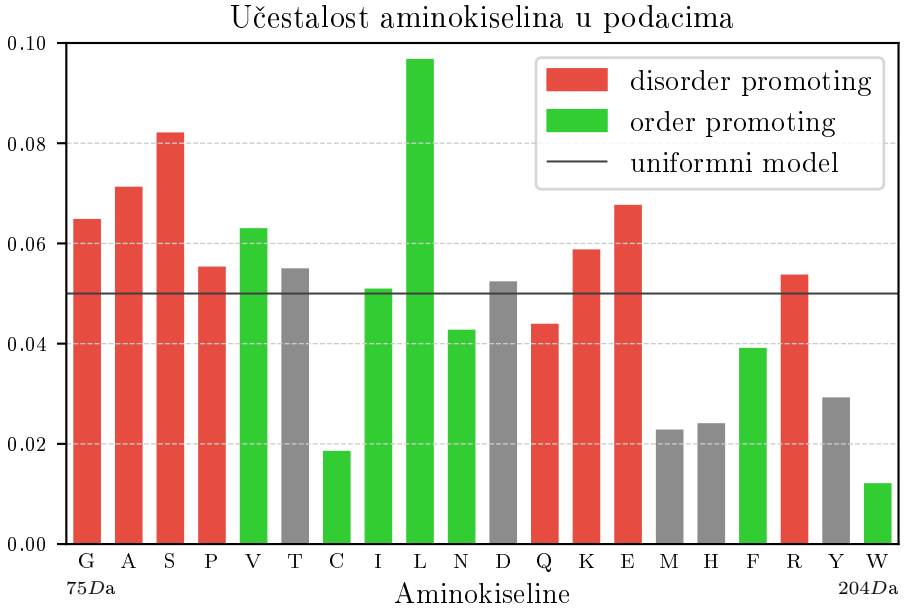
$$P_L = \frac{\sum_{s_i \in S_L} d(s_i)}{\|S_L\|}$$

Ponašanje $P_L$ predstavljeno je na Slici 4.1. Glatkoća rezulatata kontroliše se veličinom $l$ koja predstavlja prozor uprosečavanja. Kako prozor uglačavanja raste sa porastom dužine proteina ($l = 0.1 * L$) tako da prozor uprosečavanja raste sa porastom dužine proteina te $P_L$ postaje glađe sa veličinom proteina. Konstantni prozor uprosečavanja bi bila tehnika još poznata kao (engl. *rolling average*) ili (engl. *boxcar filter*) i predstavljala bi prostu vrstu konvolucije. **Trenutno ne znamo zašto se autor odlučio da veličina prozora raste sa dužinom proteina???**.




SLIKA 4.1: Punom linijom predstavljena je $P_L$ sa prozorom uprosečavanja $l = 0.1L$, a krstići predstavljaju sirove vrednosti $l = 0$

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




SLIKA 4.2: Slučajni i uniformni modeli za procenu $P_L$

Poređenje ova dva pristupa sa originalnim $P_L$ prikazano je na Slici 4.3. Originalni $P_L$ ostaje prikazan kao puna linija. Jasno se vidi da slučajni model prikazan isprekidanom linijom predstavlja vizuelno dobru aproksimaciju dok uniformni model verovatnoća prikazan tačkicama znatno odstupa i dosta sporije raste (naizgled skoro linearno). Kako VSL2b predkitor prepoznaje neuređene regione na osnovu učestalosti aminokiselina ovo ponašanje nije čudno jer je manja verovatnoća pojave aminokiselina koje promovišu neuređenost. Zbog suviše velikog odstupanja uniformni model nije korišćen u daljoj analizi.

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

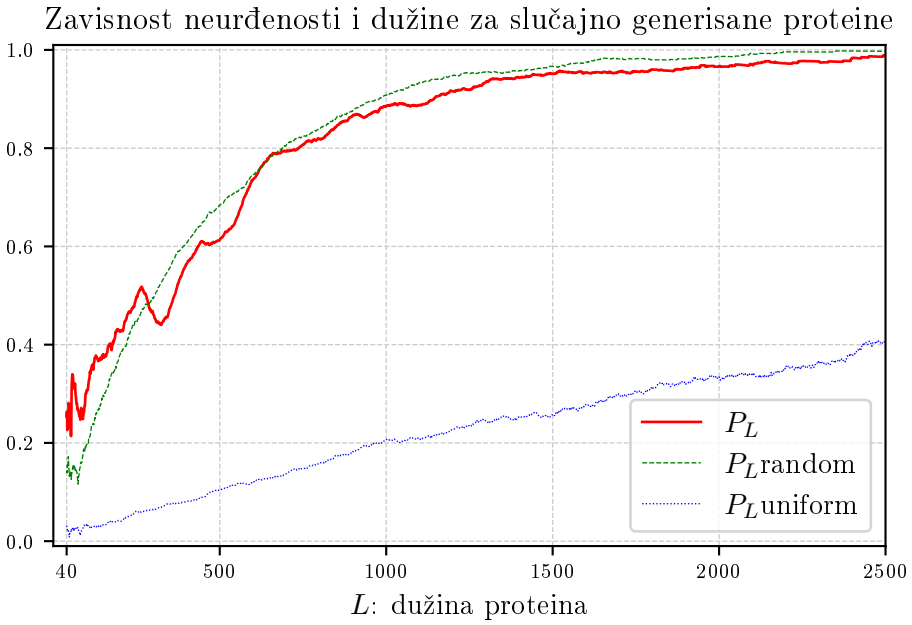
### 4.2.3 Ocenjivanje zavisnosti funkcije od neuređenosti

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

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

Nultu hiptezu koja predviđa da je rezultat $F_j$ posledica samo slučajnosti, to jest zavisi samo od $P_L$ opisana je preko slučajne veličine $Y_j$ gde je $X_L$ Bernulijeva slučajna veličina sa verovatnoćom $P(X_L = 0) = P_L$ odnosno $P(X_L = 1) = 1 - P_L$

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




SLIKA 4.3: Različiti pristupi za procenu $P_L$

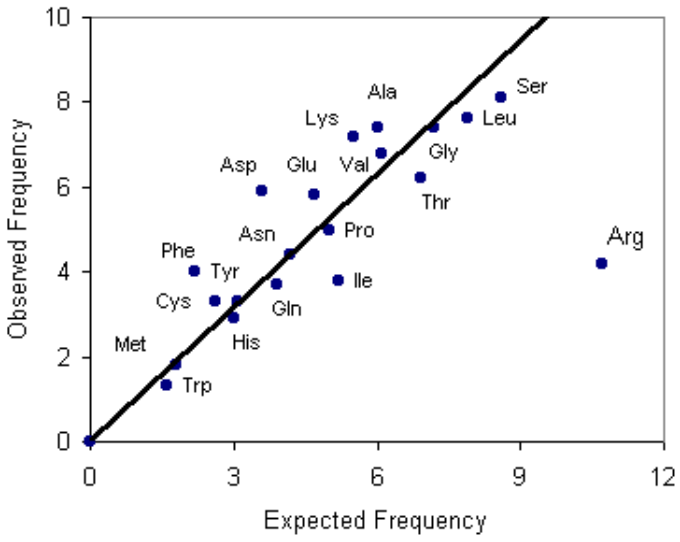
Ako $F_j$ izlazi iz intervala poverenja raspodele $Y_j$ onda funkcija $j$ sadrži značajno mnogo predviđenih neuređenih ili uređenih proteina. Preciznije ako je *p-value* $< 0.05$ funkcija $j$ je povezana sa neuređenim proteinima a ako je *p-value* $> 0.95$ funkcija $j$ je povezana sa uređenim proteinima. Suprotno ne može ništa da se tvrdi za funkciju $j$.

$Y_j$ je teško izračunati analitički te mora da se pribegne empiriskom računanju p-vrednosti. Empiriska p-vrednost određena je tako što je za 1000 realizacija $Y_j$ izračunato očekivanje da je realizacija $Y_j$ veća od $F_j$.

```
p = np.array( [yj>Fj for yj in Yj_1000] ).mean()
```

U radu (Xie et al., 2007) autori tvrde se da za veće skupove $S_j$ raspodela $Y_j$ ponaša kao normalna pa se ocena Z-skor može dobiti kao $Z_j = (F_j - \mu_j)/\delta_j$ gde je $\mu_j$ očekivanje a $\delta_j$ standardna devijacija. Dodatno p-vrednost može da se aproksimira kao $1/2(1 - erf(Z_j/2))$ [4] ako raspodela liči na normalnu. Ovo je nekad korisno jer sa 1000 realizacija $Y_j$ nema dovoljnu preciznost za p vrednost manju od $1/1000 = 0.001$. Međutim u ovom radu to nije korišćeno jer su sva sortiranja (kao i u originalnom radu) izvršena po Z-skor oceni.

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




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

# Glava 5

# Priprema podataka

## 5.1 Ontologije gena i ključne reči

## 5.2 Objedinjavanje CAFA3 i novije Svis-Prot verzije

Iz CAFA3 trening skupa izdvojeni su svi validni proteini ( dužine barem 9, i azbukom od 20 standardnih aminokiselina). Ni u jednom trenutko ne izbacujemo proteine manje od 40 aminokiseline jer je to deo analize funkcije i nismo želeli da ograničimo skup mogućih predikcija neuređenosti.

Informacije o Svis-Prot bazi dobijene su iz verzije 2017_12, iz datoteke `ftp://ftp.uniprot.org/pub/databases/uniprot/previous_releases/release-2017_12/knowledgebase/uniprot_sprot-only2017_12.tar.gz`. Pomenuta verzija ima 556 196 proteina. Zbog novijeg datuma baze postoje razlike u broju, sekvencama i anotacijama proteina u odnosu na CAFA3 verzije proteina.

Od 66 599 validnih CAFA3 proteina 66 530 ima nepromenjen **primarni identifikator** (engl. *accession number*[1]). 69 novih unosa(slogova[2]) u Svis-Prot bazu dobijena su revizijom CAFA3 proteina koji nam nedostaju. Ovo je posledica dva moguća mehanizma:

1. Unifikacija postojećih proteina u jedan novi slog. Rezultat ovog preslikavanja prikazan je na Slici 5.1. Analizom ovih promena uspešno su rekonstruisana svega 4 nova sloga. Kako je 4 suviše mali broj zbog jednostavnosti zanemarili smo sve slogove dobijene unifikacijom.

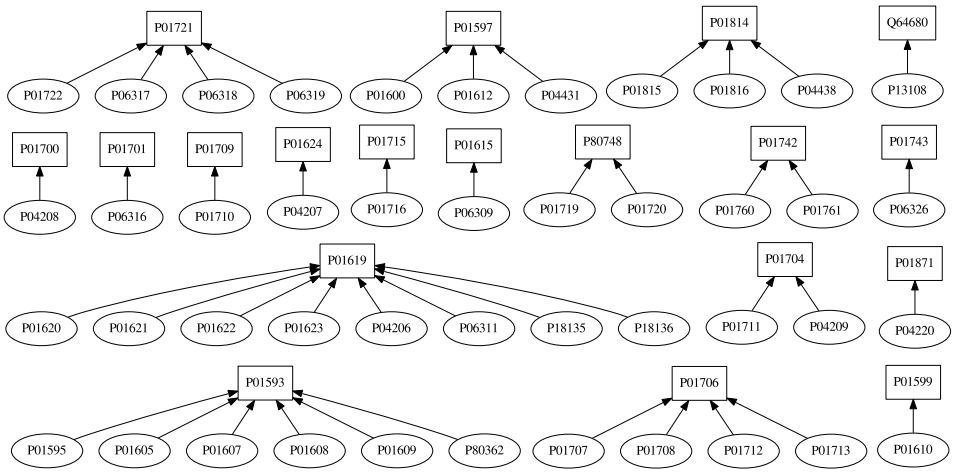
2. Specijalizacija jednog proteina u više različitih slogova. Zbog moguće statističke redundantnosti ovi slogovi su zanemareni.

Validini CAFA3 proteini anotirani su sa 5957 različitih GO termina Molekulske Funkcije (MF) od kojih je 50 zamenjeno novijim terminom i izbačeno iz najnovije `go.obo` datoteke. U Svis-Prot bazi nismo bili u mogućnosti da proverimo za M,F ali ukupno je zamenjeno 319 GO termina. CAFA3 sadrži 67 MF GO termina koji se ne javljaju u Swis-Prot anotacijama. Swis-Prot sadrži 888 MF GO termina koji se ne javljaju u CAFA3 anotacijama. Pošto Svis-Prot treba da sadrži svežije (tačnije) anotacije CAFA3 verzija anotacija je zanemarena u koristi najnovijih Svis-Prot anotacija.

Pored anotacija Svis-Prot sadrži 194 proteina čija se sekvenca razlikuje u odnosu na CAFA3 verziju proteina. Odlučeno je da ipak koristimo CAFA3 sekvence.

[1] Pod brojem se zapravo podrazumeva alfanumerička oznaka.

[2] Slog (engl. *Record*) u terminima baze podataka predstavlja zapis jednog elementa u ovom slučaju reprezentacije proteina i njegovih karakteristika. identifkovan je primarnim identifikatorm.




SLIKA 5.1: Unifikacija starih(elispe) na nove slogove u Svis-Prot bazi

# Bibliografija


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

Ashburner, Michael et al. (2000). ?Gene Ontology: tool for the unification of biology? In: *Nature Genetics* 25.1, pp. 25–29. DOI: 10.1038/75556. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3037419.

Barrell, D. et al. (2009). ?The GOA database in 2009–an integrated Gene Ontology Annotation resource? In: *Nucleic Acids Research* 37.Database, pp. D396–D403. DOI: 10.1093/nar/gkn803.

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

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

*CAFA*. http://biofunctionprediction.org/cafa/. Pristupljeno: 13.12.2017.

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

Dunker, A.Keith et al. (2001). ?Intrinsically disordered protein? In: *Journal of Molecular Graphics and Modelling* 19.1, pp. 26–59. DOI: 10.1016/s1093-3263(00)00138-8. URL: https://doi.org/10.1016/s1093-3263(00)00138-8.

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

Guttman, Burton S. (1998). ?Biology? In: pp. 66–107. URL: https://www.amazon.com/Biology-Burton-S-Guttman/dp/0697223663.

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

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

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

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

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

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

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