UNIVERZITET U BEOGRADU
MATEMATIČKI FAKULTET

MASTER RAD

Računarska analiza povezanosti funkcije i neuređenosti proteina

---

Autor:
Goran VINTERHALTER

Mentor:
dr. Jovana KOVAČEVIĆ

ČLANOVI KOMSIJE:

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

Beograd, 2017

UNIVERZITET U BEOGRADU

Sažetak

Matematički fakultet
Katedra za Računarstvo i informatiku

Master informatičar

Računarska analiza povezanosti funkcije i neuređenosti proteina

by 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 koncepti

1.2 Tehnike predvidjanja neuredjenosti

1.3 Funkcija proteina

Glava 2

Podaci i Metode

2.1 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 samo ako je potrebno grupisati ih, ili ako je potrebno mapiranje na neku drugu nomenkulaturu funkcija pa se zahteva grupisanje.

2.1.1 Podaci iz originalnog rada

U originalnom radu (Xie et al., 2007) korišćena je baza ručno proverenih proteinskih sekvenci **Svis-Prot** (engl. *Swiss-Prot*), verzija 48 iz 2005. Verzija 48 ima 201 560 proteina od kojih 196 326 imaju dužinu preko 40 aminokiselina (što je potrebno zbog definicije ref 1). 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.

Proteinske sekvence u Svis-Prot bazi "nisu redundantne" u smislu da su produkti jednog gena za jednu vrstu predstavljeni jednim entiteom (*How redundant are the UniProt databases?*). Zbog toga unosi u bazu tretiraju se kao proteini a ne pojedinačne proteinske sekvence. Međutim za analizu funkcija podaci su **statistički redundantni** jer sadrže jako mnogo **homologih** proteina. Rešenje klasterovati proteine u **proteinske familije** , čime je dobijeno 27 217 familija. Tada svaki protein ima težinu kojom doprinosi analizi funkcija tako da je težina svih proteina jedne familije uniformno raspoređena i sabira se na jedan. Za detalje konsultovati originalni rad. (Xie et al., 2007).

komentar:
Ono što autori nisu elaborirali jeste da početni uslov od minimum 20 proteina po ključnoj reči možda nije dovoljan. Ako pretpostavimo zarad ilustracije normalnu raspodelu veličina klastera proteina, očekivali bi da klaster najčešće sadrži 7 proteina. Dakle iako je 50 proteina pridruženo nekoj funkciji ona verovatno ima pridruženih svega 7 familija proteina. Kako familija sadrži proteine pod pretpostavkom istog evolutivnog porekla njihova funkcija bi trebalo da je slična pa se onda postavlja pitanje da li je 7 familija dovoljno da bi se razmatrala data ključna reč. Ovo je primarno kritika za ključne reči jer one obično predstavljaju jako opšte pojmove.

Sa druge strane za usko specijalizovane pojmove bila bi dovoljna jedna familija proteina jer bi ona predstavljala sve razne homologe (TODO Burkhard Rost, Termofili)

2.1.2 Naši podaci (CAFA3 proteini)

U ovom radu korišćen je skup proteina preuzet sa **CAFA3** takmičenja gde je korišćen kao trening skup za predikciju funkcija proteina (*CAFA*). Podaci se sastoje od dve datoteke:

1. `uniprot_sprot_exp.fasta` sadrži 41 793 proteina od kojih 41 227 ima dužinu veću od 40 aminokiselina.

2. `uniprot_sprot_exp.txt` pridružuje funkcije označene kao **GO termini** (engl. *Gene Ontology terms*). Postoje termini iz sve tri ontologije: 16 117 ćeliskih komponenti, 5 966 molekulskih funkcija i 16 117 bioloških procesa.Jednom proteinu može biti pridruženo više GO termina i obrnuto.

CAFA3 trening skup je pažljivo odabran podskp Swis-Prot proteina i smatra se da nije **statistički redundantan** (**??**). Iz tog razloga nije potrebno vršiti klasterovanje i analiza je jednostavnija te u metodama predstavljamo samo jednostavan oblik formula koje odgovaraju CAFA3 podacima.

Sekvence su kodirane jednim karakterom i koristeći **IUPAC** kodove. Među proteinima bilo je sekvenci sa nestandardnim aminokiselinam 'U' i 'O' ili višeznačnim oznakama 'B', 'J', 'X' i 'Z'. Ovi proteini se preskočeni jer ih VSL2b predkitor ne podržava. Nakon filtiranja ostalo je 41 119 proteina.

U ovom radu analiziraćemo samo termine molekulskih funkcija. Od 5 966 termina svega 358 ima pridruženo 20 ili više proteina. Zbog prirode ontologija ovaj broj se može povećati uopštavanjem termina sa malim brojem pridruženih proteina ili obrnuto specijalizacijom onih sa velikim brojem pridruživanja. Detaljan postupak uopštavanja i specijalizacije objašnjen je u metodama.

2.2 Metod

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

Idealan slučaj. Pretpostavimo da za proizvoljnu molekulsku funkciju imamo skup različitih proteina. Da bi dali korektan odgovor na ovo pitanje moramo da znamo kako neuređenost pojedinačnog proteina utiče na zadatu funkciju, da li je bitna ili nije bitna. Nije dovoljno samo posmatrati da li protein ima neuređeni region jer možda on ne utiče na funkciju. Ovaj pristup zahteva ekspertsko poznavanje svakog proteina i anotirane funkcije te stoga može da se primeni na jako malo funkcija. Takođe, trenutno svega 803 proteina ima eksperimentalno opisanu neuređenost (**disprot**)

Realnost. Možemo da damo procenu ako pretpostavimo da veći udeo neuređenih u odnosu na uređene proteine podrazumeva da funkcija zavisi više od neuređenosti. Dakle ono što želimo jeste da ispitamo **korelaciju** između funkcije i (ne)uređenosti proteina kojima je pridružena. Ali prvo potrebno je definisati šta znači da je protein neuređen. Definicija mora da ima biološkog smisla za analizu ali pored toga je ograničena je sposobnostima i preciznošću prediktora koji se korist. Više o tome u nastavku.

2.2.1 Predikcija dugih neuređenih regiona

Autori (**Xie2017**) koristili su **PONDR VL3E** prediktor koji postiže tačnost od 87% pri unakrsnoj validaciji nad uravnoteženim test skupom. Zbog ekonomičnosti i dostupnosti mi smo koristili **PONDR VSL2b** .

Oba prediktora pripadaju **PONDR** (engl. *(Predictor Of Naturally Disordered Regions)*) familiji prediktora. Ovi prediktori zasnivaju se na eksperimentalno pokazanim karakteristikama neurđenih regiona. Preciznije određene aminokiseline verovatnije su da se jave u neuređenom regionu. Neuređeni regioni imaju manje aromatičnih i hidrofobnih AK, veći ukupni naboj, veći indeks fleksibilnost i manju kompleksnost sekvence. Ove osobine izražene kao atributi koriste se za treniranje neuronske mreže (engl. *neural networks*) sa propagacijom unapred (engl. *feed forward*) koja koristi prozor veličine između 9 i 21 aminokiseline. Finalni prediktor predstavlja kombinaciju nekoliko neuronskih mreža gde je svaka od njih trenirana nad različitim podacima specijalizujući se da predviđa samo regione određene veličine ili položaja. PONDR familija ima nekoliko prediktora koji se razlikuju u načinu treniranja što se postiže kombinacijom različitih trening skupova. Oznaka "VSL2b" kodira tipove proteinskih skupova nad kojima su trenirani prediktori.

- V - Opisuje eksperimentalnu tehniku kojom je neurđenost 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)

VSL2b kao ulaz prima proteinsku sekvencu minimalne dužine 9 AK kodiranih jednim karakterom. Podržava azbuku od samo 20 standardnih AK. Izlaz je niz ocena(verovatnoća) za svaku aminokiselinu da li pripada neuređenom regionu to jest da je taj rezidual [1] neuređen. Reziduale sa vrednošću iznad 0.5 smatramo neuređenim a suprotno uređenim. Za potrebe analize autori uvode sledeću definiciju:

Definicija 1 *Protein je **verovatno/putativno 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:

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

2.2.2 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

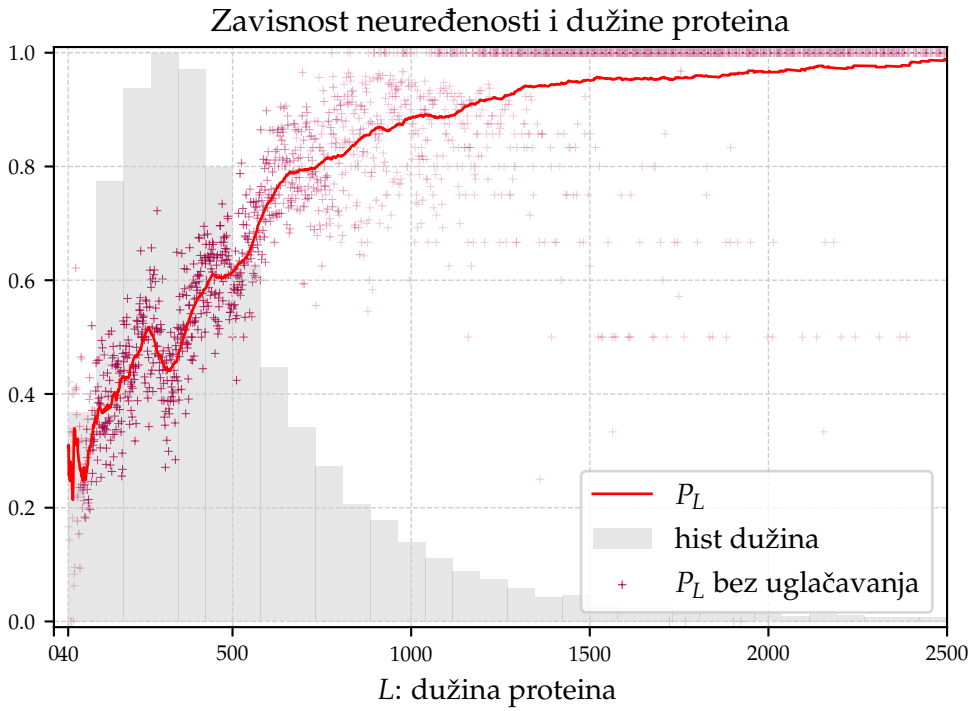
[1] Rezidual je čest naziv koji se koristi za elemente na nekoj poziciji sekvence. Koristi se za aminokiseline proteina ali i za nukleinske kiseline kod DNK i RNK. Naziv potiče od hemiskih tehnika prečišćavanja čiji su rezultati reziduali (ostaci).

statističku značajnost rezultata. Autori (Xie et al., 2007) predstavljaju način da se ta verovatnoću proceni. Označimo tu verovatnoću sa P_L gde $L \in N$ označava dužinu sekvence. Neka je V_L skup svih validnih proteina dužine L onda bismo P_L mogli da procenimo kao količnik broja verovatno neuređenih proteina iz V_L i ukupnog broja proteina u V_L. Ovaj idealan slučaj nije realan jer skup svih validnih proteina nije poznat. Pod validnošću podrazumevamo da se protein javlja prirodno kao produkt evolucije (tj. ima ili je imao funkciju u organizmu). Skup V_L aproksimiraćemo skupom svih naših proteina. Međutim, ovaj skup je suviše mali i P_L sadrži mnogo šuma. Da bi uglačali rezultat autori Xie et al., 2007 predlažu da se umesto skupa proteina egzaktne dužine L koristi skup proteina u oznaci S_L 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 grafiku 2.1. Glatkoća rezulatata kontroliše se veličinom l koja predstavlja prozor uprosečavanja. Kako je $l = 0.1 * L$ tako da prozor uprosečavanja raste sa porastom dužine proteina te P_L postaje glađe sa veličinom proteina. Za konstantni prozor uprosečavanja ova tehnika je još poznata kao (engl. *rolling average*) ili (engl. *boxcar filter*) i pripada specijalnoj vrsti konvolucije. Trenutno ne znamo zašto se autor odlučio da veličina prozora raste sa dužinom proteina.



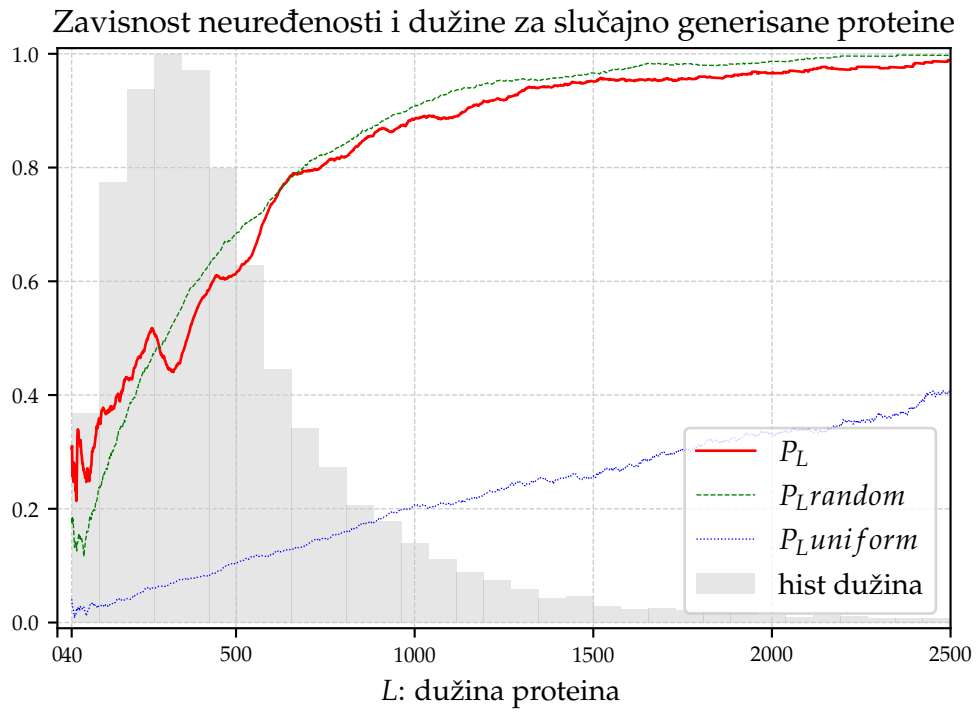
SLIKA 2.1: Punom linijom predstavljena je P_L sa prozorom uprosečavanja $l = 0.1L$, a krstići predstavljaju sirove vrednosti $l = 0$

Pored gore prikazanog 'originalnog' metoda predstavljamo još jedan pristup. **Slučajno generisani** (engl. *random generated*) proteini za procenu P_L. Razmotrićemo dva modela. Prvi je naivan 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 grafiku 2.2. Koristeći ova dva modela za svaki protein generisan je slučajan protein iste dužine koji se koristi za procenu P_L.

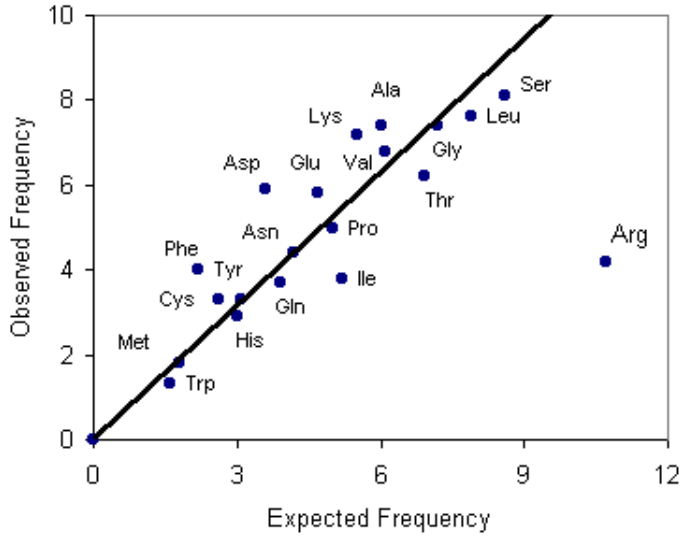
SLIKA 2.2: Učestalost aminokiselina u podacima

Poređenje ova dva pristupa sa originalnim P_L prikazano je na slici 2.3. 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 (reklo bi se 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.



SLIKA 2.3: Različiti pristupi za procenu P_L

Jedno objašnjenje 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 2.4 (*Amino acid frequency*).



SLIKA 2.4: očekivana i realna učestalost aminokiselina kod sisara

2.2.3 Ocenjivanje zavisnosti funkcije od neuređenosti

Neka je S_j skup proteina koji imaj pridruženu funkciju j. Tada procenat verovatno neuređenih proteina F_j možemo 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 opisaćemo preko slučajne veličinu Y_j.

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

Gde je X_L Bernulijeva slučajna veličina sa verovatnoćom $P(X_L = 0) = P_L$ odnosno $P(X_L = 1) = 1 - P_L$

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 možemo ništa da tvrdimo za funkciju j.

Y_j je teško izračunati analitički tako da se pribegava empiriskom računanju p-vrednosti. Empiriska p-vrednost se određuje tako što se za 1000 realizacija gleda frakcija koliko puta je realizovano Y_j bilo veće od F_j. Za veće skupove S_j raspodela Y_j liči na normalnu pa računanjem očekivanja μ_j i standardne devijacije δ_j za raspodelu Y_j možemo da izračunamo Z-skor kao $Z_j = (F_j - \mu_j)/\delta_j$. Dalje p-vrednost možemo da aproksimiramo kao $1/2(1 - erf(Z_j/2)))$ ako raspodela liči na normalnu. Ovo je nekad korisno jer sa 1000 realizacija Y_j nemamo dovoljnu preciznost da računamo p-vrednost veću ili manju od 0.01 i 0.99. Ipak koristićemo samo empiriski izračunatu p vrednost.

2.3 uopštavanje, grupisanje GO termina

2.4 Metod za upoređivanje rezultata

Glava 3

Implementacija

Glava 4

Rezultati

Glava 5

Diskusija

Bibliografija

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

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

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

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.