Maksimalne klike u analizi sličnosti proteinskih motiva

Martinić, Keti

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Keti Martinić

MAKSIMALNE KLIKE U ANALIZI SLIČNOSTI PROTEINSKIH MOTIVA

Diplomski rad

Voditelj rada:
doc. dr. sc. Pavle Goldstein

Zagreb, studeni, 2018.

Ovaj diplomski rad obranjen je dana _____ pred ispitnim povjerenstvom u sastavu:

1. _____, predsjednik
2. _____, član
3. _____, član

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Zahvaljujem svom mentoru doc. dr. sc. Pavlu Goldsteinu na uloženom trudu, strpljenju te pomoći u izradi rada. Posebna zahvala mojoj obitelji i dečku, a nadasve majci koja mi je bila velika potpora tijekom cijelog školovanja.

Sadržaj

Sadržaj **iv**

Uvod **1**

1 Pojmovi iz vjerojatnosti i statistike **2**
1.1 Vjerojatnosni prostor 2
1.2 Uvjetna vjerojatnost. Nezavisnost 3
1.3 Slučajne varijable. Funkcije distribucije 3
1.4 Primjeri slučajnih varijabli 5
1.5 Osjetljivost i specifičnost 7

2 Pojmovi iz teorije grafova **9**

3 Analiza odgovora **11**
3.1 Uvod u biološke pojmove 11
3.2 Upit 12
3.3 Odgovor 12
3.4 Težinski graf 15
3.5 0 – 1 graf 16
3.6 Maksimalna klika 16
3.7 Prag 17

4 Primjeri **24**
4.1 Osnovni pojmovi 24
4.2 Rezultati 25
4.3 Analiza rezultata 34

Bibliografija **35**

Uvod

U ovom radu bavimo se analizom strukture odgovora koji je dobiven iterativnim pretraživanjem proteoma. Iterativnim pretraživanjem se, u proteomu nekog organizma, pronalaze proteini iz određene proteinske familije zadavanjem karakterističnog niza aminokiselina za tu familiju. Točnije, pronalaze se nizovi aminokiselina dovoljno slični zadanom nizu. Ako zadamo niz aminokiselina koji je slučajno odabran iz proteoma i nije karakterističan ni za jednu familiju dobivamo odgovor koji nije biološki značajan. Naša pretpostavka je da struktura odgovora korespondira biološkoj značajnosti odgovora. Dakle, na temelju strukture, odnosno veličine najveće klike, pokušavamo doći do zaključka o biološkoj značajnosti samog odgovora. Također, promatramo najveću kliku biološki značajnog odgovora i zanima nas udio proteina iz proteinske familije koje smo sačuvali najvećom klikom. Je li veći udio proteina koji su zaista iz određene proteinske familije u najvećoj kliki ili u samom odgovoru?

Ovaj rad podijeljen je u 4 poglavlja. U prvom i drugom poglavlju definirani su pojmovi iz vjerojatnosti i statistike te iz teorije grafova koji su nužni za razumijevanje i koji se koriste u radu. U drugom poglavlju primjerima ilustriramo varijacije u sličnostima unutar odgovora, a zatim objašnjavamo prelazak na analizu grafova i opisujemo pridruživanje težinskog grafa, a potom i $0-1$ grafa, odgovoru. Nadalje, postavljamo hipotezu i opisujemo algoritam kojim tražimo posebnu strukturu unutar odgovora, tj. najveću kliku pripadnog $0-1$ grafa s obzirom na određeni prag. U zadnjem poglavlju provodimo analizu strukture odgovora koje smo dobili zadavanjem 5 upita i pretraživanjem biljnih proteomima talijinog uročnjaka, krumpira, azijske riže i rajčice te bakterije *Streptomyces avermitilis*.

Poglavlje 1

Pojmovi iz vjerojatnosti i statistike

1.1 Vjerojatnosni prostor

Definicija 1.1.1. ***Slučajni pokus*** *ili* ***slučajni eksperiment*** *je pokus čiji ishodi, tj. rezultati nisu jednoznačno određeni uvjetima u kojima izvodimo pokus.*

Definicija 1.1.2. *Osnovni polazni objekt u teoriji vjerojatnosti je neprazan skup Ω koji zovemo* ***prostor elementarnih događaja*** *i koji reprezentira skup svih ishoda slučajnog pokusa. Točke ω skupa Ω zvat ćemo* ***elementarni događaji****.*

Definicija 1.1.3. *Familija $\mathcal{F}$ podskupova od Ω ($\mathcal{F} \subset \mathcal{P}(\Omega)$) je* ***$\sigma$-algebra skupova*** *(na Ω) ako je:*

(i) $\emptyset \in \mathcal{F}$

(ii) $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$

(iii) $A_i \in \mathcal{F}, i \in \mathbb{N} \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

Definicija 1.1.4. *Neka je $\mathcal{F}$ σ-algebra na skupu Ω. Uređen par $(\Omega, \mathcal{F})$ zove se* ***izmjeriv prostor****.*

Definicija 1.1.5. *Neka je $(\Omega, \mathcal{F})$ izmjeriv prostor. Funkcija $P : \mathcal{F} \to \mathbb{R}$ je* ***vjerojatnost*** *(na $\mathcal{F}$, na Ω) ako vrijedi:*

(i) $P(A) \geqslant 0, A \in \mathcal{F}$

(ii) $P(\Omega) = 1$

(iii) $A_i \in \mathcal{F}, i \in \mathbb{N}$ i $A_i \cap A_j = \emptyset$ *za* $i \neq j \Rightarrow P\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \sum_{i=1}^{\infty} P(A_i)$.

Napomena 1.1.6. *Svojstva iz gornje definicije se nazivaju:*

(i) ***nenegativnost vjerojatnosti***

(ii) ***normiranost vjerojatnosti***

(iii) ***σ-aditivnost vjerojatnosti.***

Definicija 1.1.7. *Uređena trojka $(\Omega, \mathcal{F}, P)$, gdje je $\mathcal{F}$ σ-algebra na Ω i P vjerojatnost na $\mathcal{F}$, zove se* ***vjerojatnosni prostor****.*

Neka je $(\Omega, \mathcal{F}, P)$ vjerojatnosni prostor. Elemente σ-algebre $\mathcal{F}$ zovemo **događaji**, a broj $P(A), A \in \mathcal{F}$ zove se **vjerojatnost događaja** A.

1.2 Uvjetna vjerojatnost. Nezavisnost

Definicija 1.2.1. *Neka je $(\Omega, \mathcal{F}, P)$ proizvoljan vjerojatnosni prostor i $A \in \mathcal{F}$ takav da je $P(A) > 0$. Definirajmo funkciju $P_A : \mathcal{F} \to [0, 1]$ na sljedeći način:*

$$P_A(B) = P(B \mid A) = \frac{P(A \cap B)}{P(A)}, \quad B \in \mathcal{F}. \tag{1.1}$$

P_A je vjerojatnost na $\mathcal{F}$ i nju zovemo ***uvjetna vjerojatnost uz uvjet A****. Broj $P(B \mid A)$ zovemo* ***vjerojatnost od B uz uvjet A****.*

Definicija 1.2.2. *Neka je $(\Omega, \mathcal{F}, P)$ vjerojatnosni prostor i $A_i \in \mathcal{F}, i \in I$ proizvoljna familija događaja. Kažemo da je to* ***familija nezavisnih događaja*** *ako za svaki konačan podskup različitih indeksa $i_1, i_2, \ldots, i_k \in I$ vrijedi*

$$P\Big(\bigcap_{j=1}^{k} A_{i_j}\Big) = \prod_{j=1}^{k} P(A_{i_j}). \tag{1.2}$$

1.3 Slučajne varijable. Funkcije distribucije

Sa $\mathcal{B}$ označimo σ-algebru generiranu familijom svih otvorenih skupova na $\mathbb{R}$. $\mathcal{B}$ zovemo **σ-algebra Borelovih skupova na** $\mathbb{R}$, a elemente σ-algebre $\mathcal{B}$ zovemo **Borelovi skupovi**.

Definicija 1.3.1. *Neka je $(\Omega, \mathcal{F}, P)$ vjerojatnosni prostor. Funkcija $X : \Omega \to \mathbb{R}$ je* ***slučajna varijabla*** *(na Ω) ako je $X^{-1}(B) \in \mathcal{F}$ za proizvoljno $B \in \mathcal{B}$, tj. $X^{-1}(\mathcal{B}) \subset \mathcal{F}$.*

Definicija 1.3.2. *Funkcija $g : \mathbb{R} \to \mathbb{R}$ je* ***Borelova funkcija*** *ako je $g^{-1}(B) \in \mathcal{B}$ za svako $B \in \mathcal{B}$, tj. ako je $g^{-1}(\mathcal{B}) \subset \mathcal{B}$.*

Definicija 1.3.3. *Neka je* $(\Omega, \mathcal{F}, P)$ *vjerojatnosni prostor i* X *slučajna varijabla na* Ω*. Za* $B \in \mathcal{B}$ *stavimo*

$$P_X(B) = P(X^{-1}(B)) = P\{\omega \in \Omega : X(\omega) \in B\} = P\{X \in B\}. \tag{1.3}$$

Gornjom relacijom definirana je funkcija $P_X : \mathcal{B} \to [0,1]$ *i lako je provjeriti da je* P_X *vjerojatnost, odnosno vjerojatnosna mjera na* $\mathcal{B}$*.* P_X *zovemo* ***vjerojatnosna mjera inducirana sa*** X*, a vjerojatnosni prostor* $(\mathbb{R}, \mathcal{B}, P_X)$ *zovemo* ***vjerojatnosni prostor induciran sa*** X*.* P_X *često zovemo i* ***zakon razdiobe od*** X*.*

Definicija 1.3.4. *Neka je* X *slučajna varijabla na* Ω*.* ***Funkcija distribucije od*** X *je funkcija* $F_X : \mathbb{R} \to [0,1]$ *definirana sa*

$$\begin{aligned} F_X(x) &= P_X((-\infty, x]) = P(X^{-1}(-\infty, x]) = \\ &= P\{\omega \in \Omega : X(\omega) \leqslant x\} = P\{X \leqslant x\}, \quad x \in \mathbb{R}. \end{aligned} \tag{1.4}$$

Često ćemo stavljati $F_X = F$, ako je jasno o kojoj se slučajnoj varijabli odnosno njenoj funkciji distribucije radi.

Teorem 1.3.5. *Funkcija distribucije* F *slučajne varijable* X *je rastuća i neprekidna zdesna na* $\mathbb{R}$ *i zadovoljava*

$$\begin{aligned} F(-\infty) &= \lim_{x \to -\infty} F(x) = 0 \\ F(+\infty) &= \lim_{x \to +\infty} F(x) = 1. \end{aligned} \tag{1.5}$$

Funkciju $F : \mathbb{R} \to [0,1]$ koja ima svojstva iz prethodnog teorema zvat ćemo **vjerojatnosna funkcija distribucije** (na $\mathbb{R}$) ili, kraće, **funkcija distribucije**.

U teoriji vjerojatnosti postoje dva glavna tipa slučajnih varijabli: diskretne i neprekidne.

Definicija 1.3.6. *Slučajna varijabla* X *je* ***diskretna*** *ako postoji konačan ili prebrojiv skup* $D \subset \mathbb{R}$ *takav da je* $P\{X \in D\} = 1$*.*

Definicija 1.3.7. ***Lebesgueova mjera na*** $\mathbb{R}$ *je mjera* λ *na* $(\mathbb{R}, \mathcal{B})$ *takva da za sve* $a, b \in \mathbb{R}, a < b$ *vrijedi* $\lambda((a,b]) = \lambda([a,b]) = \lambda((a,b)) = b - a$*.*

Definicija 1.3.8. *Neka je* X *slučajna varijabla na vjerojatnosnom prostoru* $(\Omega, \mathcal{F}, P)$ *i neka je* F_X *njezina funkcija distribucije. Kažemo da je* X ***apsolutno neprekidna*** *ili, kraće,* ***neprekidna slučajna varijabla*** *ako postoji nenegativna realna Borelova funkcija* f *na* $\mathbb{R}$ *(*$f : \mathbb{R} \to \mathbb{R}_+$*) takva da je*

$$F_X(x) = \int_{-\infty}^{x} f(t) d\lambda(t), \quad x \in \mathbb{R}. \tag{1.6}$$

Ako je X neprekidna slučajna varijabla, tada se funkcija f iz (1.6) zove **funkcija gustoće vjerojatnosti od** X ili, kraće, **gustoća od** X, i ponekad je označujemo sa f_X.

1.4 Primjeri slučajnih varijabli

Eksponencijalna distribucija

Neprekidna slučajna varijabla X ima **eksponencijalnu distribuciju** ako joj je funkcija gustoće f dana sa

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x > 0 \\ 0, & x \leqslant 0, \end{cases}$$

gdje je $\lambda > 0$ fiksan. λ obično zovemo parametar eksponencijalne distribucije.

Logistička distribucija

Neka su $\mu, \beta \in \mathbb{R}, \beta > 0$. Neprekidna slučajna varijabla X ima **logističku distribuciju** s parametrima μ i β ako joj je funkcija gustoće f dana sa

$$f(x) = \frac{e^{-\frac{x-\mu}{\beta}}}{\beta\left(1 + e^{-\frac{x-\mu}{\beta}}\right)^2}, \quad x \in \mathbb{R}.$$

Neka su $p > 0, q > 0$. Slučajna varijabla X ima **generaliziranu logističku distribuciju** ako joj je funkcija gustoće f dana sa

$$f(x) = \frac{\Gamma(p+q)}{\Gamma(p)\,\Gamma(q)} \frac{e^{px}}{(1+e^x)^{p+q}}, \quad x \in \mathbb{R},$$

gdje je funkcija Γ definirana sa $\Gamma(x) = \int_0^\infty e^{-t} t^{x-1} dt, x > 0$.

Generalizirana distribucija ekstremnih vrijednosti

Generalizirana distribucija ekstremnih vrijednosti, poznata i kao Fisher-Tippettova distribucija, fleksibilan je model sa tri parametra koji kombinira Gumbelovu, Fréchetovu i Weibullovu distribuciju maksimalne ekstremne vrijednosti. Funkcija gustoće f joj je dana sa

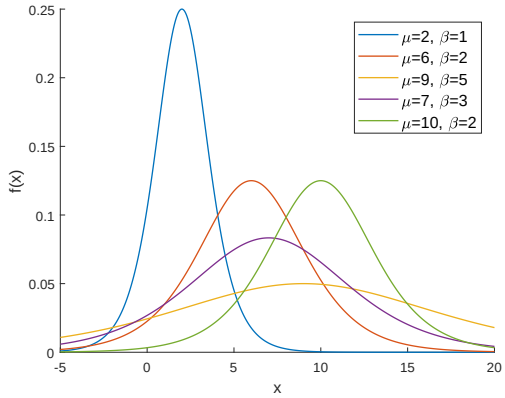
$$f(x) = \begin{cases} \frac{1}{\beta} e^{-(1+kz)^{-\frac{1}{k}}} (1+kz)^{-1-\frac{1}{k}}, & k \neq 0 \\ \frac{1}{\beta} e^{-(z+e^{-z})}, & k = 0, \end{cases}$$

gdje je $z = \frac{x-\mu}{\beta}, x \in \mathbb{R}$, te su $k, \mu, \beta \in \mathbb{R}, \beta > 0$ parametri distribucije. Također, zadan je uvjet $1 + k\frac{x-\mu}{\beta} > 0$, za $k \neq 0$.

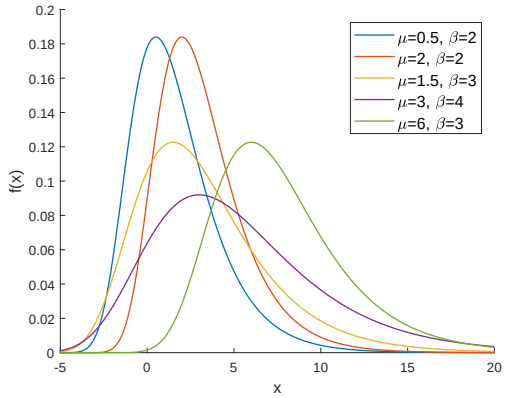
Za različite vrijednosti parametra k, $k = 0, k > 0$ i $k < 0$, dobiju se redom Gumbelova, Fréchetova i "obrnuta" Weibullova distribucija.

Neka je $p > 0$. Slučajna varijabla X ima **generaliziranu Gumbelovu distribuciju** ako joj je funkcija gustoće f dana sa

$$f(x) = \frac{1}{\Gamma(p)} e^{-px} e^{-e^{-x}}, \quad x \in \mathbb{R}.$$



Slika 1.1: Funkcije gustoće logističke distribucije za razne vrijednosti parametara



Slika 1.2: Funkcije gustoće Gumbelove distribucije za razne vrijednosti parametara

Teorem 1.4.1. *Neka su X_1 i X_2 nezavisne slučajne varijable s generaliziranom Gumbelovom distribucijom s parametrima p i q, respektivno. Tada slučajna varijabla $Y = X_1 - X_2$ ima generaliziranu logističku distribuciju s parametrima p i q.*

Teorem 1.4.2 (Fisher-Tippett (1928.), Gnedenko (1943.))**.** *Neka su $X_1, X_2, \ldots, X_n$ nezavisne, jednako distribuirane slučajne varijable i neka je $M_n = \max\{X_1, X_2, \ldots, X_n\}$. Ako postoje konstante $a_n \in \mathbb{R}, b_n > 0$ i nedegenerirana funkcija distribucije H takva da je*

$$\lim_{n\to+\infty} P\Big(\frac{M_n - a_n}{b_n} \leqslant x\Big) = H(x),$$

odnosno

$$\frac{M_n - a_n}{b_n} \xrightarrow{\mathcal{D}} H, \quad n \to +\infty,$$

tada H pripada jednoj od tri distribucije ekstremnih vrijednosti: Gumbelovoj, Fréchetovoj ili Weibullovoj distribuciji.

1.5 Osjetljivost i specifičnost

Osjetljivost i specifičnost su statističke mjere uspješnosti provedenog testa.

Osjetljivost (eng. *sensitivity*) ili **TPR** (eng. *true positive rate*) je postotak pozitivnih elemenata uzorka u odnosu na određeno stanje, odnosno CP (eng. *condition positive*) elemenata uzorka, koji su ispravno prepoznati kao pozitivni.

$$\begin{aligned}\text{TPR} &= \frac{\text{broj stvarno pozitivnih}}{\text{broj stvarno pozitivnih + broj lažno negativnih}} \\ &= \frac{\text{TP}}{\text{TP+FN}} = \frac{\text{TP}}{\text{CP}}\end{aligned}$$

Specifičnost (eng. *specificity*) ili **TNR** (eng. *true negative rate*) je postotak negativnih elemenata uzorka u odnosu na određeno stanje, odnosno CN (eng. *condition negative*) elemenata uzoraka, koji su ispravno prepoznati kao negativni.

$$\begin{aligned}\text{TNR} &= \frac{\text{broj stvarno negativnih}}{\text{broj stvarno negativnih + broj lažno pozitivnih}} \\ &= \frac{\text{TN}}{\text{TN+FP}} = \frac{\text{TN}}{\text{CN}}\end{aligned}$$

Postoje još neke mjere u ocjenjivanju testa kao što su PPV i NPV.

Pozitivna prediktivna vrijednost (PPV) (eng. *positive predictive value*) je postotak pozitivno prepoznatih elemenata uzorka koji su ispravno prepoznati kao pozitivni.

$$\begin{aligned}\text{PPV} &= \frac{\text{broj stvarno pozitivnih}}{\text{broj stvarno pozitivnih + broj lažno pozitivnih}} \\ &= \frac{\text{TP}}{\text{TP+FP}}\end{aligned}$$

Negativna prediktivna vrijednost (NPV) (eng. *negative predictive value*) je postotak negativno prepoznatih elemenata uzorka koji su ispravno prepoznati kao negativni.

$$\begin{aligned} \text{NPV} &= \frac{\text{broj stvarno negativnih}}{\text{broj stvarno negativnih + broj lažno negativnih}} \\ &= \frac{\text{TN}}{\text{TN+FN}} \end{aligned}$$

		Predviđeno stanje		
	Ukupna populacija	Ocijenjeni pozitivno	Ocijenjeni negativno	
Stvarno stanje	Pozitivno stanje (CP)	TP (stvarno pozitivni)	FN (lažno negativni)	Osjetljivost
	Negativno stanje (CN)	FP (lažno pozitivni)	TN (stvarno negativni)	Specifičnost
		PPV (pozitivna prediktivna vrijednost)	NPV (negativna prediktivna vrijednost)	

Tablica 1.1: Tablica uspješnosti testa

Pojmovi iz ovog poglavlja preuzeti su iz izvora [5], [2] i [4].

Poglavlje 2

Pojmovi iz teorije grafova

Definicija 2.0.1. ***Graf*** *G je uređeni par vrhova (V, E), gdje je V skup* ***vrhova****, a E skup $2-$podskupova od V, koje zovemo* ***bridovi****.*

Napomena 2.0.2. *Katkada gornju definiciju proširujemo tako da dopustimo* ***petlje*** *(bridove koje spajaju vrh sa samim sobom),* ***višestruke bridove*** *(više bridova između para vrhova) i* ***usmjerene bridove*** *(bridovi koji imaju orijentaciju tako da idu od jednog vrha prema drugome). Usmjereni bridovi se reprezentiraju uređenim parovima, a ne $2-$podskupovima, dok kod višestrukih bridova E postaje multiskup.*

Definicija 2.0.3. *Graf koji ima usmjerene bridove zvat ćemo* ***usmjereni graf*** *ili* ***digraf****, a graf koji ima višestruke bridove zvat ćemo* ***multigraf****.*

Napomena 2.0.4. *U ovome radu neusmjereni graf bez višestrukih bridova i petlji nazivamo $0-1$ graf.*

Definicija 2.0.5. *Kažemo da su vrhovi $u, v \in V$, u grafu $G = (V, E)$,* ***susjedni*** *ako postoji brid $e = \{u, v\} \in E$.*

Definicija 2.0.6. *Kažemo da je graf $G = (V, E)$* ***potpun*** *ukoliko je svaki par vrhova u grafu brid, te* ***nul graf*** *ukoliko nema bridova.*

Definicija 2.0.7. ***Podgraf*** *grafa $G = (V, E)$ je graf kojemu su skup vrhova i skup bridova podskupovi od V i E, respektivno.*

Definicija 2.0.8. ***Put*** *od v_0 do v_n u grafu $G = (V, E)$ je niz $(v_0, e_1, v_1, e_2, v_2, \ldots, e_n, v_n)$, gdje su vrhovi različiti (osim eventualno prvog i zadnjeg vrha), a e_i je brid $\{v_{i-1}, v_i\}$, za $i = 1, 2, \ldots, n$.*

Definicija 2.0.9. *Relacija ekvivalencije $\equiv$ na skupu vrhova V grafa $G = (V, E)$: vrhovi x i y su u relaciji, odnosno $x \equiv y$, ako postoji put u grafu od x do y.*

Definicija 2.0.10. ***Komponenta povezanosti*** *grafa $G = (V, E)$ je podgraf induciran klasom ekvivalencije gore definirane relacije ekvivalencije $\equiv$.*

Definicija 2.0.11. *Kažemo da je graf $G = (V, E)$* ***povezan*** *ako postoji samo jedna komponenta povezanosti.*

Definicija 2.0.12. ***Težinski graf*** *$G = (V, E)$ je graf s težinskom funkcijom na skupu bridova. Težinska funkcija na skupu bridova je funkcija s E u $\mathbb{R}$ ili $\mathbb{R}_0^+$.*

Pojmovi iz ovog poglavlja preuzeti su iz izvora [6].

Poglavlje 3

Analiza odgovora

3.1 Uvod u biološke pojmove

Proteini ili bjelančevine su sastavni dijelovi svake stanice, što ih čini jednom od osnovnih komponenti života na Zemlji. Imaju različite funkcije u organizmu, uključujući kataliziranje metaboličkih reakcija, repliciranje DNA i prijenos molekula unutar stanice. Izgrađene su od 20 standardnih aminokiselina koje su međusobno povezane peptidnom vezom. Aminokiseline su molekule koje sadrže amino skupinu, karboksilnu skupinu i bočni lanac po kojemu se aminokiseline međusobno razlikuju. Njihova biološka uloga je upravo izgradnja proteina. Svi proteini nekog organizma (ili stanice) koji nastaju kao posljedica ekspresije gena u određenom trenutku pod određenim uvjetima čine proteom.

Kratica	Naziv	Kratica	Naziv
A	Alanin	M	Metionin
C	Cistein	N	Asparagin
D	Asparaginska kiselina	P	Prolin
E	Glutaminska kiselina	Q	Glutamin
F	Fenilalanin	R	Arginin
G	Glicin	S	Serin
H	Histidin	T	Treonin
I	Izoleucin	V	Valin
K	Lizin	W	Triptofan
L	Leucin	Y	Tirozin

Tablica 3.1: Standardne aminokiseline

3.2 Upit

Upit je neki niz aminokiselina, najčešće duljine od 5 do 20. Iterativno pretraživanje proteoma, uz neki zadani upit, je pretraživanje s obzirom na neku funkciju sličnosti i njime dobivamo odgovor, odnosno skup nizova aminokiselina koji su dovoljno slični upitu s obzirom na tu funkciju sličnosti. Funkcija sličnosti u iterativnom pretraživanju, odnosno njezini parametri se mijenjaju u svakoj iteraciji. Naime, u svakoj novoj iteraciji pretražujemo proteom sa skupom nizova aminokiselina koji su bili dovoljno slični u prethodnoj iteraciji. Iteriranje staje kada se skup odabranih nizova ne promijeni, odnosno kada se postigne maksimalni broj iteracija.

3.3 Odgovor

Promatramo strukturu odgovora dobivenog iterativnim pretraživanjem u odnosu na neki upit. Naime, zbog same prirode iterativnog pretraživanja, ali i različitih upita, veličine proteoma, veličine i smislenosti odgovora te prirode bioloških podataka, apriori ne znamo koja je prosječna sličnost nizova aminokiselina u odgovoru. Također, sličnost između parova nizova u odgovoru može varirati od negativne pa do jako pozitivne sličnosti. U iterativnom pretraživanju tražimo nizove koji su dovoljno slični upitu s obzirom na neku funkciju sličnosti i iako se ta funkcija sličnosti mijenja, ona uzima u obzir samo prosječnu sličnost nizova koji su izabrani u određenoj iteraciji. Dakako, priroda iterativnog pretraživanja je samo jedan od gore navedenih razloga za varijaciju sličnosti u odgovoru. Upravo zbog svega navedenog analiziramo strukturu odgovora.

Definirajmo Blosum matricu i Blosum score (odgovara sličnosti nizova aminokiselina).

Definicija 3.3.1. *Blosum matrica B je 20×20 matrica, $B = (b_{ij}) \in M_{20}(\mathbb{Z})$, koja na (i, j)-tom mjestu sadrži koeficijente sličnosti i-te i j-te aminokiseline. (O Blosum matrici više u [1]). Ukratko, bazirana je na sljedećoj formuli:*

$$B(i, j) = \left\lfloor \log \frac{P(a_i \leftrightarrow b_j \,|\, M)}{P(a_i, b_j \,|\, R)} \right\rfloor, \quad a_i, b_j \in \mathcal{A}, \tag{3.1}$$

gdje su a_i i b_j aminokiseline pridružene, respektivno, i-tom i j-tom mjestu, a $\mathcal{A}$ je skup svih standardnih aminokiselina. M je model koji pretpostavlja da aminokiseline a_i i b_j imaju zajedničkog pretka, a R je random model koji pretpostavlja nezavisnost aminokiselina pa vrijedi $P(a_i, b_j \,|\, R) = P(a_i \,|\, R) \cdot P(b_j \,|\, R)$. Distribucija standardnih aminokiselina uz model R je dana sa:

A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
0.078	0.051	0.043	0.053	0.019	0.043	0.063	0.072	0.023	0.053	0.091	0.059	0.022	0.039	0.052	0.068	0.059	0.014	0.032	0.066

Definicija 3.3.2. *Blosum score s je rezultat koji odgovara sličnosti (ili povezanosti) dvaju nizova aminokiselina. Što je Blosum score veći, nizovi aminokiselina su sličniji.*

Primjer 3.3.3 (*Računanje Blosum score-a*)**.** *Neka su zadana dva niza aminokiselina jednakih duljina,* `FVFGDS` *i* `PEPLIS`*. Sličnost nizova računamo tako da zapišemo nizove jedan ispod drugog i za aminokiseline u stupcu očitavamo pripadnu vrijednost u Blosum matrici i zatim te vrijednosti zbrajamo. Aminokiseline su poredane po recima i stupcima Blosum matrice u sljedećem rasporedu:* `A R N D C Q E G H I L K M F P S T W Y V`.

`FVFGDS`
`PEPLIS`

$s = B(14,15)+B(20,7)+B(14,15)+B(8,11)+B(4,10)+B(16,16) = -4-3-4-4-4+5 = -14$

Promotrimo sljedeća dva primjera kako bi bolje razumjeli zašto analiziramo strukturu odgovora.

Primjer 3.3.4. *Pretpostavimo da imamo sljedeći upit i* 6 *nizova aminokiselina koji čine odgovor.*

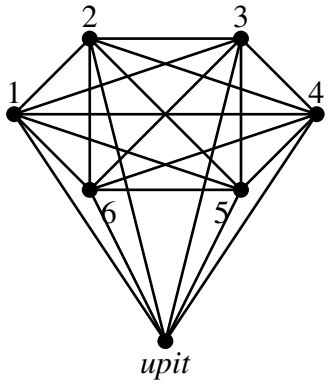
`FVFGDSLSDA` = *upit*
`FVFGDSLFDA` = 1
`FIFGDSLVDN` = 2
`FVFGDSLVDA` = 3
`FIFGDSLSDV` = 4
`FVFGDSVFDN` = 5
`FVFGDSYADT` = 6

Nizove aminokiselina iz odgovora označili smo brojevima od 1 *do* 6 *zbog lakše notacije. Sada računamo sličnosti između svih parova nizova aminokiselina, uključujući nizove iz odgovora i sam upit, pomoću prethodno definirane Blosum matrice, odnosno računanjem Blosum score-a.*

$$\begin{array}{lll} s(upit,1) = 57 & s(1,3) = 59 & s(2,6) = 48 \\ s(upit,2) = 51 & s(1,4) = 51 & s(3,4) = 52 \\ s(upit,3) = 58 & s(1,5) = 58 & s(3,5) = 49 \\ s(upit,4) = 59 & s(1,6) = 46 & s(3,6) = 49 \\ s(upit,5) = 47 & s(2,3) = 58 & s(4,5) = 44 \\ s(upit,6) = 50 & s(2,4) = 50 & s(4,6) = 49 \\ s(1,2) = 52 & s(2,5) = 56 & s(5,6) = 46 \end{array}$$

Možemo uočiti kako su svi nizovi aminokiselina iz odgovora međusobno podjednako slični. Također, ti su nizovi i podjednako slični upitu.

Kako bi ilustrirali njihovu međusobnu sličnost, upit i nizove aminokiselina iz odgovora ćemo promatrati kao vrhove, a njihovu međusobnu sličnost ćemo prikazati debljinom bridova između pripadnih vrhova.



Slika 3.1: Ilustracija za primjer 3.3.4

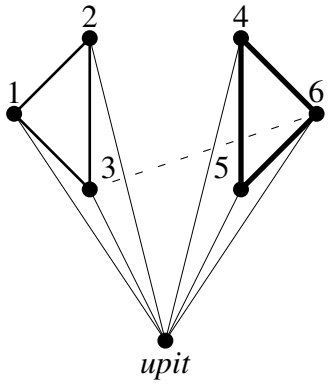
Primjer 3.3.5. *Pretpostavimo da imamo sljedeći upit i odgovor.*

$$\begin{aligned}
&\texttt{MALAFGHL} = \textit{upit}\\
&\texttt{MAVAWAWL} = 1\\
&\texttt{MALAWAKA} = 2\\
&\texttt{MVLAFAWA} = 3\\
&\texttt{KLQDFGHI} = 4\\
&\texttt{KLQEFGHL} = 5\\
&\texttt{KVQDFGHL} = 6
\end{aligned}$$

Kao i u prethodnom primjeru nizove aminokiselina iz odgovora označili smo brojevima od 1 *do* 6*. Promotrimo sada izračunate sličnosti između parova nizova aminokiselina.*

$$\begin{array}{lll}
s(\textit{upit},1) = 21 & s(1,3) = 32 & s(2,6) = -7\\
s(\textit{upit},2) = 21 & s(1,4) = -9 & s(3,4) = -1\\
s(\textit{upit},3) = 20 & s(1,5) = -5 & s(3,5) = -1\\
s(\textit{upit},4) = 20 & s(1,6) = -4 & s(3,6) = 2\\
s(\textit{upit},5) = 24 & s(2,3) = 25 & s(4,5) = 48\\
s(\textit{upit},6) = 25 & s(2,4) = -8 & s(4,6) = 50\\
s(1,2) = 33 & s(2,5) = -8 & s(5,6) = 47
\end{array}$$

Možemo uočiti dva bloka nizova aminokiselina u odgovoru. Prvi blok čine nizovi 1*,* 2 *i* 3*, a drugi blok nizovi* 4*,* 5 *i* 6*. Nizovi iz svakog bloka su podjednako slični upitu dok su blokovi, odnosno nizovi u njima, međusobno slabo povezani. Nizovi unutar svakog pojedinog bloka su pak dobro povezani, posebno drugi blok nizova. Ilustracija će najbolje pokazati međusobne sličnosti, a kako bi je pojednostavili nećemo crtati negativne sličnosti već samo pozitivne.*



Slika 3.2: Ilustracija za primjer 3.3.5

U prethodnim primjerima možemo uočiti varijaciju sličnosti unutar odgovora dok su svi nizovi u odgovoru podjednako slični upiti. Primjeri nam ilustriraju strukturu odgovora, ali su i motivacija za analizu strukture grafa pridruženog odgovoru.

3.4 Težinski graf

Odgovoru, koji smo dobili iterativnim pretraživanjem u odnosu na neki upit, pridružujemo potpun težinski graf. Nizovi aminokiselina iz odgovora postaju vrhovi grafa, a Blosum score-ovi, odnosno izračunate sličnosti između parova nizova u odgovoru, postaju težine bridova između pripadnih vrhova.

Promatramo li strukturu dobivenog težinskog grafa možemo uočiti varijacije u težinama bridova koje su posljedica varijacija u sličnosti nizova aminokiselina u odgovoru. Analizom strukture težinskog grafa pridruženog odgovoru analiziramo i strukturu samog odgovora.

Pretpostavka ili hipoteza je da bi struktura odgovora mogla biti povezana sa biološkom značajnošću odgovora, odnosno da je odgovor biološki značajan ako su svi, ili skoro svi, nizovi iz odgovora međusobno podjednako slični i ako su te sličnosti dosta dobre. Odgovor je biološki značajan, po definiciji biologa, ako su proteini, u kojima se nalaze nizovi aminokiselina iz odgovora, karakteristični za određenu proteinsku familiju. Upravo zbog navedene pretpostavke tražimo najveći podgraf težinskog grafa koji je potpun i čije težine su dovoljno velike i podjednake. Neka je prag M spomenuta dovoljno velika težina. Ako su vrhovi povezani bridom težine veće ili jednake pragu M kažemo da su vrhovi dovoljno dobro povezani.

Općenito kod analize težinskog grafa nećemo imati tako jasno podijeljene blokove ni-

zova aminokiselina kao u primjeru 3.3.5. Taj primjer je služio kao ilustracija za analizu odgovora. U većini slučajeva imat ćemo težinske grafove kod kojih težine većine bridova ne variraju previše.

3.5 $0-1$ graf

Težinskom grafu, koji smo definirali u prethodnom odjeljku, pridružujemo $0-1$ graf s obzirom na prag M. Bridove čije su težine manje od dovoljno velike težine, praga M, brišemo iz grafa, a bridove čija su težine veće ili jednake pragu M ostavljamo u grafu i više ne promatramo težine tih bridova. Sada se analiza strukture težinskog grafa, odnosno samog odgovora, svodi na analizu strukture $0-1$ grafa. Traženje najvećeg potpunog podgrafa težinskog grafa dovoljno velikih i podjednakih težina svodimo na traženje najvećeg potpunog pografa $0-1$ grafa, odnosno na traženje najveće klike $0-1$ grafa. Detaljnije o klikama u sljedećem odjeljku.

Ovisno o strukturi $0-1$ grafa, odnosno o povezanosti samog grafa, možemo dobiti različite veličine najveće klike. Ako je graf skoro potpun, tj. izbacivanjem par vrhova dobivamo potpun graf tada će najveća klika sadržavati skoro sve vrhove grafa. Ako u grafu imamo malo bridova i dosta komponenata povezanosti tada će najveća klika biti mala. Moguće je i da ne dobijemo tako jasno određene male ili velike najveće klike.

3.6 Maksimalna klika

Definicija 3.6.1. *Klika u grafu* (V, E) *je njegov podgraf s bar dva vrha, koji je potpun (tj. postoji brid između svaka dva vrha podgrafa).*

Definicija 3.6.2. *Maksimalna klika je klika koja nije sadržana u niti jednoj većoj kliki, tj. dodavanjem nekog vrha, ona prestaje biti klika.*

Definicija 3.6.3. *Najveća klika je klika koja ima najveći broj vrhova.*

U prethodnom odjeljku smo govorili o $0-1$ grafu čiju strukturu analiziramo. U $0-1$ grafu tražimo najveću kliku, tj. najveći potpun podgraf. Najveću kliku ćemo tražiti Bron-Kerbosch algoritmom koji vraća sve maksimalne klike grafa, a zatim ćemo naći najveću maksimalnu kliku, ili jednu od najvećih ako ne postoji jedinstvena. Najveća maksimalna klika ujedno je i najveća klika grafa.

Bron-Kerbosch algoritam

Bron-Kerbosch algoritam je rekurzivni algoritam koji vraća sve maksimalne klike neusmjerenog grafa. Algoritam je objavljen 1973., a osmislili su ga Nizozemski programeri

Coenraad Bron i Joep Kerbosch. U svrhu ovog rada korišten je algoritam s pivotiranjem, odnosno s biranjem pivotnog elementa. Promotrimo pseudokod algoritma s pivotiranjem.

```
bronkerbosch(R, P, X):
    if P = ∅ and X = ∅:
        return R
    odaberi pivotni vrh p ∈ P ∪ X
    for v ∈ P \ N(p):    #N(p) su susjedni vrhovi vrha p
        bronkerbosch(R ∪ {v}, P ∩ N(v), X ∩ N(v))
        P := P \ {v}
        X := X ∪ {v}
```

Listing 3.1: Bron-Kerbosch pseudokod

Imamo tri disjunktna skupa R, P i X i tražimo maksimalnu kliku koja sadrži sve vrhove iz skupa R, neke vrhove iz skupa P i nijedan vrh iz skupa X. U svakom pozivu rekurzije P i X su disjunktni skupovi čija unija sadrži vrhove koji su povezani sa svakim vrhom, odnosno elementom skupa R. Inicijalno su skupovi R i X prazni, a skup P sadrži sve vrhove grafa. Kada su skupovi P i X oboje prazni više nema vrhova koji se mogu dodati skupu R pa algoritam vraća skup R kao maksimalnu kliku. Uočimo, u slučaju da je skup P prazan, a skup X nije također više nema vrhova koji se mogu dodati skupu R, ali u tom slučaju rekurzija ne vraća maksimalnu kliku. Naime, nema maksimalne klike koja sadrži elemente skupa R, a ne sadrži elemente skupa X jer su elementi iz $P \cup X = \emptyset \cup X = X$ povezani sa svim elementima iz R.

Ako skupovi P i X nisu oboje prazni, tada biramo pivotni vrh p iz skupa $P \cup X$ tako da vrh p ima najviše susjednih vrhova ili je jedan od takvih vrhova. Za svaki vrh $v \in P$ koji je ili pivotni vrh p ili nije susjedan pivotnom vrhu p radimo sljedeće:

• Vrh v dodajemo skupu R, a iz skupova P i X izbacujemo one koji nisu susjedni sa vrhom v i radimo rekurzivni poziv. Tim rekurzivnim pozivom tražimo maksimalne klike koje sadrži vrh v.

• Vrh v dodajemo skupu X, a izbacujemo ga iz skupa P jer više ne tražimo maksimalne klike koje sadrže vrh v.

3.7 Prag

$0-1$ graf pridružen težinskom grafu je definiran s obzirom na prag M pa je njegov izgled posljedica odabira praga. Ako je prag manji od težine većine bridova tada će $0-1$ graf pridružen težinskom grafu biti skoro potpun. Ako je prag veći od težine većine bridova tada će $0-1$ graf biti nepovezan graf sa mnogo komponenata povezanosti.

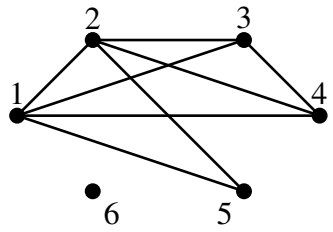
U odjeljku 3.5 smo već spomenuli kako o povezanosti $0-1$ grafa, odnosno njegovoj strukturi ovisi i veličina njegove najveće klike. Promotrimo sljedeći primjer kako bi ilustrirali važnost odabira praga.

Primjer 3.7.1. *Neka je zadan odgovor kao u primjeru 3.3.4 .*

FVFGDSLFDA = 1
FIFGDSLVDN = 2
FVFGDSLVDA = 3
FIFGDSLSDV = 4
FVFGDSVFDN = 5
FVFGDSYADT = 6

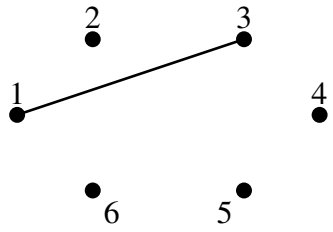
Pretpostavimo da je prag $= 44$*, tada je* $0-1$ *graf pridružen odgovoru potpun te je najveća klika* $0-1$ *grafa skup svih vrhova* $\{1,2,3,4,5,6\}$.

Ako je prag $= 50$*, tada najveću kliku* $0-1$ *grafa čine vrhovi* $\{1,2,3,4\}$. $0-1$ *graf za prag* $= 50$ *možemo vidjeti na sljedećoj slici.*



Slika 3.3: $0-1$ graf za *prag* $= 50$

Ako je pak prag $= 59$*, tada najveću kliku* $0-1$ *grafa čine samo vrhovi* $\{1,3\}$*, a pripadni* $0-1$ *graf možemo vidjeti na sljedećoj slici.*



Slika 3.4: $0-1$ graf za *prag* $= 59$

Računanje praga

1. Procjena distribucija za upit.

Za svaku poziciju u upitu procjenjujemo vjerojatnost da se određena aminokiselina pojavi na toj poziciji. Neka je upit duljine n. Dakle, i-toj poziciji u upitu pridružujemo empirijsku distribuciju $f_i = (f_{i_1}, f_{i_2}, \ldots, f_{i_{20}})$, gdje je $f_{i_j} = 1$ ako se j-ta aminokiselina pojavila na i-toj poziciji, inače vrijedi $f_{i_j} = 0, \forall i \in \{1, 2, \ldots, n\}$. Napomenimo da je redoslijed aminokiselina isti njihovom redoslijedu po recima i stupcima Blosum matrice. Budući da smo dobili n distribucija u kojima je vjerojatnost pojave 19 aminokiselina jednaka nuli primjenjujemo blagu težinsku shemu kako distribucije ne bi bile koncentrirane samo u jednoj aminokiselini. Svakoj poziciji empirijske distribucije f_i dodajemo mali pseudozbroj 0.01 te zatim normaliziramo distribuciju. Dakle, nova distribucija aminokiselina i-te pozicije je $g_i = (g_{i_1}, g_{i_2}, \ldots, g_{i_{20}})$, gdje je

$$g_{i_j} = \frac{f_{i_j} + 0.01}{1.2}, \quad i = 1, \ldots, n, j = 1, \ldots, 20. \tag{3.2}$$

Neka je $A = (a_{ij}) \in M_{20}(\mathbb{R})$ PAM matrica, tj. matrica koja na (i, j)-tom mjestu sadrži vjerojatnost da i-ta aminokiselina mutira u j-tu. Matrica A je stohastička matrica, tj. suma svakog retka je jednaka 1. Neka je $C = (c_{ij}) = A^k \in M_{20}(\mathbb{R})$ potencirana PAM matrica, gdje je k neki veliki broj. U našem slučaju uzeli smo $k = 120$. Uočimo da je matrica C također stohastička matrica, a redak c_i u matrici C predstavlja očekivani vektor mutacije i-te aminokiseline nakon faktor k vremena evolucije. Konačnu distribuciju $p_i = (p_{i_1}, p_{i_2}, \ldots, p_{i_{20}})$ za i-tu poziciju definiramo kao linearnu kombinaciju vektora redaka matrice C, $c_1, \ldots, c_{20}$, sa koeficijentima iz vektora $g_i = (g_{i_1}, g_{i_2}, \ldots, g_{i_{20}})$, za svaku poziciju $i = 1, \ldots, n$, odnosno:

$$p_i = \sum_{k=1}^{20} g_{i_k} c_k, \quad i = 1, \ldots, n. \tag{3.3}$$

Primjer 3.7.2. *Neka je zadan upit u, u =* `A`*. Duljina upita je* 1 *pa moramo procijeniti samo jednu funkciju distribucije za prvu poziciju* $f = (f_1, f_2, \ldots, f_{20})$.

Prvo pridružimo upitu u vektor f duljine 20 *koji na prvoj poziciji ima jedinicu, a na svim ostalim pozicijama nule. Naime, aminokiselina A je prva po redu aminokiselina. Slijedi da je distribucija* $f = (1, 0, 0, \ldots, 0, 0)$*. Zatim dodajemo pseudozbroj* 0.01 *i normaliziramo distribuciju f i dobivamo* $g = (0.841\dot{6}, 0.008\dot{3}, 0.008\dot{3}, \ldots, 0.008\dot{3}, 0.008\dot{3})$*. Na kraju još uzimamo u obzir očekivanu mutaciju aminokiselina nakon određenog vremena evolucije i dobivamo konačnu distribuciju p:*

$$\begin{aligned} p = (&0.603680, 0.010547, 0.016060, 0.019699, 0.010440, 0.013815, 0.025632, \\ &0.046747, 0.010245, 0.012199, 0.016167, 0.014877, 0.008567, 0.010523, \\ &0.031205, 0.054753, 0.045248, 0.008019, 0.010151, 0.031426). \end{aligned}$$

2. Simuliranje nizova aminokiselina iz procijenjenih distribucija i računanje prosječne sličnosti.

Za neki upit duljine n, za koji smo procijenili distribucije aminokiselina za svaku poziciju $i = 1, \ldots, n$ kao u prethodnom koraku, sada simuliramo nizove aminokiselina iz tih procjenjenih distribucija. U našem slučaju odabrali smo sljedeća 4 upita duljine 10:

- FVFGDSLSDA
- PEPLISEILF
- EIFECRESLT
- DHILKGQNKA.

Bitno je naglasiti da smo nizove aminokiselina odabrali proizvoljno. Također, proizvoljno smo odabrali simulirati, za sva 4 upita, po 150 nizova aminokiselina iz procjenjenih distribucija za određeni upit.

Promotrimo neke od 150 simuliranih nizova iz procijenjene distribucije za *upit* = FVFGDSLSDA.

FVFGHSIFDA	SVFGDSWSDA	AVFGDRYANS	FVFADVLSDA	FVYSDSLSDA
FVFGDSVPNA	FRKVDRLFEE	FIFMDSLSDG	FVEDQSLEDA	FVFGDSLDDA
TVFGSSLSDA	FKFGDDLGME	FTFGDTLADK	FHFENSLKAA	NVYGDNGSDA
FVFQDSLSDK	FVFGGSMEDS	FVFEDSLSDS	RDFGKSLPTA	FMFGDSLSND
YVFGDSLGDE	FVKGDTLSNT	FVGNNSLVDP	FKTGESLSDA	REYGDSLNDA
FVMGDSLSEA	FVFGESVSNT	FVFGDSLADA	FIIGVKLGDI	FTFGDSLFDE
FAFGCSLSAG	FVFQDSLDDA	FDFGDSLFDA	FIEGNGVKDT	YVFGDALGQI
FVFGATLADS	FVFGDSLSEA	FVWGDTVADF	FVFGDYCSLA	FAFGEQRADV
FVTIDNLANA	PTFADPLSDP	FVLGKGLSDA	FVFGDGLSDH	YAYGDSLSNA
FVFGDKLSDA	GVFGADLTIA	AEAGDSLSKA	TVLGDSVSDA	FYFGDSCSDA
EEFGDTFSDE	FVWADSRSDA	FRFGESLSAA	YGFSDSLSFA	FVYGDELSKH
KVFGDTLSDA	FVFGDGDSGA	FVIAATLAKP	FVFGDGLSDN	FIFKQSLSWA
FSFMDSLENR	FVLGDSLNDE	FVFGPQGSDA	FVFGDAFKDA	FVCSQGLADS
FVFANQLPDD	AVFGDSLWMA	FVFGDSLSDT	PVNGDSLFEA	FVLGLSLGDA
FVFADNLFDA	YVFGDSLNDR	YPIYGGLSDA	FVFQDSLESA	FVFGEALSDS

Sada, za svaka dva simulirana niza aminokiselina računamo njihove međusobne sličnosti, odnosno Blosum score, a potom izračunate sličnosti uprosječujemo. U gornjoj simulaciji izračunata prosječna sličnost 150 nizova iznosi 24.234. Taj postupak ponavljamo 100 puta. Znači, provodimo po 100 simulacija od 150 nizova aminokiselina te u svakoj simulaciji računamo prosječnu sličnost. Dobivamo 100 prosječnih sličnosti koje potom ponovno uprosječujemo.

Za navedena 4 upita u 100 simulacija prosječna sličnost 150 nizova iznosi:

Upit	Prosjek
FVFGDSLSDA	24.302
PEPLISEILF	25.009
EIFECRESLT	26.709
DHILKGQNKA	24.734

Tablica 3.2: Prosječne sličnosti 150 nizova aminokiselina za upite duljine 10

Uočimo da prosjeci, odnosno prosječne međusobne sličnosti 150 nizova aminokiselina, jako malo variraju za četiri prozvoljno odabrana upita duljine 10. Možemo pretpostaviti da bi i za druge upite duljine 10 taj prosjek također bio sličan i da bi iznosio približno 25.

Promotrimo kako se ponašaju prosječne sličnosti $50, 100, 150, 200, 250$ i 300 nizova, koji su simulirani iz procjenjenih distribucija za upite FVFGDSLSDA i PEPLISEILF, u 100 simulacija.

Broj simuliranih nizova	Prosjek
50	24.386
100	24.172
150	24.302
200	24.286
250	24.187
300	24.378

Tablica 3.3: *Upit* = FVFGDSLSDA

Broj simuliranih nizova	Prosjek
50	24.968
100	25.016
150	25.009
200	24.972
250	24.891
300	24.950

Tablica 3.4: *Upit* = PEPLISEILF

Vidimo da su, bez obzira na broj simuliranih nizova iz određenog upita, razlike u prosječnim sličnostima jako male. Budući da su te razlike jako male nema ih smisla promatrati.

Također, promotrimo kako se ponašaju prosječne sličnosti 150 nizova aminokiselina, u 100 simulacija, gdje su nizovi simulirani iz procjenjenih distribucija za 4 proizvoljna upita duljine 8:

- `GESGCGKS`
- `GMALAGKT`
- `FVFGDSLS`
- `PEPLISEI`.

Upit	Prosjek
`GESGCGKS`	24.312
`GMALAGKT`	18.198
`FVFGDSLS`	20.343
`PEPLISEI`	18.780

Tablica 3.5: Prosječne sličnosti 150 nizova aminokiselina za upite duljine 8

Uočimo da prosječne sličnosti imaju veću varijaciju i da su vrijednosti manje za kraće upite. Za neki drugi upit veličine 8 možemo pretpostaviti da će prosjek biti oko 20 uz veće varijacije nego kod upita duljine 10.

Prag će nam upravo biti izračunata prosječna sličnost nizova aminokiselina simuliranih iz procjenjenih distribucija za upit. Bitno je uočiti da računanje praga ovisi samo o duljini zadanog upita kojim pretražujemo proteom. Ne ovisi ni o proteomu, ni o odgovoru pa tako ni o pridruženom $0-1$ grafu. Za upite duljine 10 prosječne sličnosti jako malo variraju pa uzimamo da je $prag = 25$.

Napomena 3.7.3. *Pokažimo da je prag zaista vrlo blizu* 25 *za proizvoljne upite duljine* 10 *i da možemo postaviti* $prag = 25$.

Pretpostavimo da za upit imamo samo jednu aminokiselinu. Simuliramo li 100 *nizova duljine* 1*, iz procjenjene distribucije za tu aminokiselinu, otprilike* 66 *simuliranih aminokiselina će biti jednako onoj koja je zadana za upit. Naime, vjerojatnost da će se ista aminokiselina iz upita pojaviti simuliranjem je zadana sljedećim vjerojatnosnim vektorom,*

s obzirom na različite aminokiseline, gdje je redoslijed aminokiselina isti onom u Blosum matrici:

$$p = (0.603680, 0.675449, 0.535571, 0.592876, 0.784332, 0.613299, 0.601414, 0.713843, 0.673399, 0.606378, 0.736718, 0.695446, 0.608194, 0.729637, 0.692377, 0.563689, 0.604106, 0.790153, 0.734676, 0.657050).$$

Prosjek tih vjerojatnosti je 0.661 *pa će otprilike* 66 *simuliranih aminokiselina od njih* 100 *biti jednako aminokiselini iz upita, a za ostale* 34 *simulirane aminokiseline pretpostavimo da su, zbog jednostavnosti izračuna, uniforno izabrane među ostalim aminokiselinama.*

Ukupno imamo $uk = 4950$ *usporedbi između simuliranih aminokiselina, a* $uk_1 = 2145$ *od njih* 4950 *je broj usporedbi aminokiseline iz upita same sa sobom. Sada računamo prosjek vrijednosti na dijagonali Blosum matrice* p_d*, jer nas zanima prosječna sličnost aminokiseline same sa sobom, i taj prosjek množimo sa brojem usporedbi sam sa sobom,* uk_1*. Zatim računamo prosjek vrijednosti izvandijagonalnih elemenata Blosum matrice* p_{vd}*, jer nas zanima prosječna sličnosti različitih aminokiselina, i taj prosjek množimo sa brojem takvih usporedbi,* $uk - uk_1$*. Te vrijednosti zbrajamo i dijelimo sa ukupnim brojem usporedbi i dobivamo sljedeći rezultat:*

$$\frac{uk_1 \times p_d + (uk - uk_1) \times p_{vd}}{uk} = 2.337.$$

Kada bi dakle upit bio duljine 10, tada bi prosječna sličnost simuliranih nizova bila 23.37*. Uzmemo li pak medijan vjerojatnosti iz vjerojatnosnog vektora p i s njime računamo tada će* 67 *od* 100 *simuliranih aminokiselina biti jednako onoj iz upita te će prosječna sličnost simuliranih aminokiselina biti* 2.459*, a za upite duljine* 10 *prosječna sličnost će biti* 24.59*.*

Poglavlje 4

Primjeri

4.1 Osnovni pojmovi

GDSL lipaze su, kao i sve lipaze, enzimi koji kataliziraju hidrolizu lipida, odnosno cijepanje molekula lipida u reakciji s vodom. Karakteristična značajka GDSL lipaza je da imaju fleksibilno katalitičko mjesto (niz aminokiselina koji kataliziraju reakciju supstrata) koje mijenja svoj strukturni raspored u prisutnosti različitih supstrata. To bi moglo objasniti njihovu katalitičku multifunkcionalnost, što ih čini zanimljivim za istraživanja i primjene. GDSL lipaze su nađene u životinjama, biljkama, gljivama i bakterijama, a kopnene biljke posebno obiluju njima. Upravo bi biljke mogle biti dobar izvor vrlo obećavajućih enzima koji bi se mogli koristiti u hidrolizi i sintezi važnih spojeva koji su od biotehnološkog interesa. Stoga je traženje novih GDSL lipaza u biljakama od velike važnosti.

Promotrimo sljedeća četiri upita duljine 10: `FVFGDSLSDA`, `PEPLISEILF`, `EIFECRESLT` i `DHILKGQNKA`. Prvi upit sadrži niz aminokiselina GDSL koji je tipičan za GDSL lipaze, a ostala tri upita su slučajno odabrani nizovi aminokiselina iz nekih proteina. Provodit ćemo analizu strukture četiri odgovora koji su dobiveni iterativnim pretraživanjem sljedećih biljnih proteoma, u odnosu na spomenute upite:

- talijin uročnjak (lat. *Arabidopsis thaliana*)
- krumpir (lat. *Solanum tuberosum*)
- azijska riža (lat. *Oryza sativa*)
- rajčica (lat. *Solanum lycopersicum*).

Proteomi su dobro anotirani, odnosno imamo svu biološku informaciju o proteomima (opisana su svojstva i zapisani nazivi proteina). Odgovori su dobiveni zadavanjem upita,

proteoma i skale (parametar iterativnog pretraživanja) na IGLOSS web serveru. Više o tome u [3].

ATPaze su enzimi koji kataliziraju razgradnju ATP-a, adenozin trifosfata koji je složen organski spoj, na ADP (adenozin difosfat) i slobodan fosfatni ion. Ova reakcija oslobađa energiju koja pokreće druge kemijske reakcije u stanici koje se inače ne bi dogodile. Također, ATPaze kataliziraju i obrnutu reakciju, sintezu ATP-a. Walker-ov A motiv je niz aminokiselina u proteinima koji je povezan sa reakcijom vezivanja fosfata i pojavljuje se u ATPazama. Motiv ima niz `G-xxxx-GK-[TS]`, gdje su `G,K,T` i `S` kratice za aminokiseline glicin, lizin, treonin i serin, respektivno, a `x` označava bilo koju aminokiselinu. `[TS]` označava da zadnja aminokiselina u motivu mora biti ili treonin ili serin.

Upit `GESGCGKS` je jedan takav Walker-ov A motiv. Analizirat ćemo strukturu odgovora koji je dobiven iterativnim pretraživanjem proteoma bakterije *Streptomyces avermitilis* u odnosu na taj upit.

4.2 Rezultati

Talijin uročnjak

Talijin uročnjak (lat. *Arabidopsis thaliana*) je mala jednogodišnja cvjetnica koja pripada porodici *Brassicaceae*. Ona je popularni modelni organizam u biologiji i genetici. Naime, *Arabidopsis thaliana* je prva biljka s potpuno sekvencioniranim genomom te je stoga pogodna za razna istraživanja. Njezin proteom je vrlo dobro anotiran i za svaki protein, od njih 35176 u proteomu, znamo kojoj proteinskoj familiji pripada.

Slika 4.1: Arabidopsis thaliana

Kao što već spomenuli analizirat ćemo strukturu odgovora dobivenog pretraživanjem proteoma biljke talijinog uročnjaka u odnosu na četiri upita.

1. *Upit* = `FVFGDSLSDA`

Skala	Duljina odgovora	Prag	Veličina najveće klike
7.5	145	25	114

Tablica 4.1: Rezultat za *upit* = `FVFGDSLSDA`

Skala je parametar iterativnog pretraživanja koji utječe na broj nizova aminokiselina koji će biti odabrani dovoljno sličnima u iteraciji pretraživanja. Što je parametar veći, odabiru se sličniji nizovi, odnosno manje nizova je izabrano dovoljno sličnima u iteraciji. Preciznije, maksimalne ocjene sličnosti u iteraciji pretraživanja su logistički distribuirane. Skala je parametar te distribucije koji nam govori koliko puta ćemo se odmaknuti za parametar β od prosječne ocjene sličnosti i sve ocjene koje su veće od $\mu + skala \times \beta$ su statistički značajne. Parametar β logističke distribucije možemo izraziti pomoću standardne devijacije na sljedeći način: $\beta = \frac{\sqrt{3}}{\pi}\sigma$. U standardnoj logističkoj distribuciji sve ocjene sličnosti veće od skale su statistički značajne. Više o logističkoj distribuciji u odjeljku poglavlja Pojmovi iz vjerojatnosti i statistike 1.4.

Budući da je upit duljine 10, prag kojim definiramo $0-1$ graf postavljamo na 25. Možemo uočiti kako je najveća klika dosta velika. Naime, ona sadrži 78.621% nizova aminokiselina iz odgovora.

Za upit `FVFGDSLSDA` i proteom talijinog uročnjaka imamo biološku listu pozitivaca, odnosno imena svih proteina iz proteoma biljke u kojima bi se trebali nalaziti nizovi aminokiselina slični upitu. Proteini sa te liste su CP (eng. *condition positive*) i nazivamo ih c-pozitivci. U ovom slučaju ta lista sadrži imena 104 proteina. Biološku listu pozitivaca su sastavili biolozi koji su proučavanjem biološke strukture i upotrebom raznih metoda sastavili takav popis.

Uz odgovor, nizove aminokiselina koji su dovoljno slični upitu, imamo i popisana imena proteina u kojima se nalaze ti nizovi. Te proteine nazivamo pozitivci. Pogledajmo koliko se proteina sa biološke liste pozitivaca podudara sa proteinima u kojima se nalaze nizovi iz odgovora i iz najveće klike.

	Broj pozitivaca	TP	Različiti TP	PPV	TPR
Odgovor	145	103	91	0.710	0.875
Najveća klika	114	97	85	0.851	0.817

Tablica 4.2: Usporedba sa biološkom listom pozitivaca

Uočimo kako je veliki dio c-pozitivaca sadržan u odgovoru, ali i u najvećoj kliki. U najvećoj kliki je ostalo čak 94.175% c-pozitivaca iz odgovora.

2. *Upit* = `PEPLISEILF`

Skala	Duljina odgovora	Prag	Veličina najveće klike
7.5	138	25	59

Tablica 4.3: Rezultat za *upit* = `PEPLISEILF`

Najveća klika za upit `PEPLISEILF` je dosta manja u odnosu na upit `FVFGDSLSDA`. Naime, ona sadrži 42.754% nizova aminokiselina iz odgovora i možemo reći da se raspala, odnosno nije sačuvala većinu odgovora.

3. *Upit* = `EIFECRESLT`

Skala	Duljina odgovora	Prag	Veličina najveće klike
7.5	75	25	37

Tablica 4.4: Rezultat za *upit* = `EIFECRESLT`

Najveća klika za upit `EIFECRESLT` je također dosta manja u odnosu na upit `FVFGDSLSDA`. Ona sadrži 49.333% nizova aminokiselina iz odgovora i nije sačuvala većinu odgovora.

4. *Upit* = `DHILKGQNKA`

Skala	Duljina odgovora	Prag	Veličina najveće klike
7.5	61	25	14

Tablica 4.5: Rezultat za *upit* = `DHILKGQNKA`

Najveća klika za upit `DHILKGQNKA` je dosta mala i sadrži samo 22.951% nizova aminokiselina iz odgovora.

Krumpir

Krumpir (lat. *Solanum tuberosum*) je trajna zeljasta biljka iz porodice pomoćnica (lat. *Solanaceae*) koja je danas jedna od najvažnijih prehrambenih biljaka. Analiziramo strukturu odgovora na sljedeće upite:

1. *Upit* = `FVFGDSLSDA`

Skala	Duljina odgovora	Prag	Veličina najveće klike
7.5	130	25	95

Tablica 4.6: Rezultat za *upit* = `FVFGDSLSDA`

Najveća klika odgovora, dobivenog pretraživanjem proteoma krumpira na navedeni upit, je velika te sadrži 73.077% odgovora.

Za upit `FVFGDSLSDA` i proteom krumpira imamo biološku listu pozitivaca koja sadrži 123 proteina. Usporedimo biološku listu pozitivaca sa pozitivcima iz odgovora i najveće klike.

	Broj pozitivaca	TP	Različiti TP	PPV	TPR
Odgovor	130	91	91	0.700	0.740
Najveća klika	95	87	87	0.916	0.707

Tablica 4.7: Usporedba sa biološkom listom pozitivaca

Veliki dio c-pozitivaca je sadržan i u odgovoru i u najvećoj kliki te je u kliki ostalo 95.604% c-pozitivaca iz odgovora.

2. *Upit* = `PEPLISEILF`

Skala	Duljina odgovora	Prag	Veličina najveće klike
7.5	110	25	39

Tablica 4.8: Rezultat za *upit* = `PEPLISEILF`

Najveća klika odgovora na upit `PEPLISEILF` sadrži 35.455% nizova aminokiselina iz odgovora.

3. *Upit* = `EIFECRESLT`

Skala	Duljina odgovora	Prag	Veličina najveće klike
7.5	49	25	13

Tablica 4.9: Rezultat za *upit* = `EIFECRESLT`

Najveća klika za upit `EIFECRESLT` sadrži 26.531% nizova aminokiselina iz odgovora.

4. *Upit* = `DHILKGQNKA`

Skala	Duljina odgovora	Prag	Veličina najveće klike
7.5	50	25	9

Tablica 4.10: Rezultat za *upit* = `DHILKGQNKA`

Najveća klika za upit `DHILKGQNKA` sadrži samo 18% nizova aminokiselina iz odgovora.

Azijska riža

Azijska riža (lat. *Oryza sativa*) je najpoznatija vrsta riže. To je žitarica iz porodice trava i također jedna od najvažnijih prehrambenih biljaka koja je udomaćena širom svijeta, a potječe iz Jugoistočne Azije. Promotrimo strukturu odgovora na sljedeće upite:

1. *Upit* = `FVFGDSLSDA`

Skala	Duljina odgovora	Prag	Veličina najveće klike
7.5	183	25	122

Tablica 4.11: Rezultat za *upit* = `FVFGDSLSDA`

Najveća klika odgovora na upit `FVFGDSLSDA` je, i na proteomu azijske riže, velika i sadrži 66.667% odgovora.

Za upit `FVFGDSLSDA` i proteom azijske riže imamo biološku listu pozitivaca koja sadrži 116 proteina. Usporedimo odgovor i najveću kliku sa biološkom listom pozitivaca.

	Broj pozitivaca	TP	Različiti TP	PPV	TPR
Odgovor	183	115	96	0.628	0.828
Najveća klika	122	114	95	0.934	0.819

Tablica 4.12: Usporedba sa biološkom listom pozitivaca

Dosta c-pozitivaca je sadržano u odgovoru i u najvećoj kliki, a čak 99.130% c-pozitivaca iz odgovora je ostalo u najvećoj kliki.

2. *Upit* = `PEPLISEILF`

Skala	Duljina odgovora	Prag	Veličina najveće klike
7.5	54	25	15

Tablica 4.13: Rezultat za *upit* = `PEPLISEILF`

Najveća klika odgovora na upit `PEPLISEILF` je mala i sadrži 27.778% odgovora.

3. *Upit* = `EIFECRESLT`

Skala	Duljina odgovora	Prag	Veličina najveće klike
7.5	77	25	20

Tablica 4.14: Rezultat za *upit* = `EIFECRESLT`

Najveća klika za upit `EIFECRESLT` sadrži 25.974% nizova aminokiselina iz odgovora.

4. *Upit* = `DHILKGQNKA`

Skala	Duljina odgovora	Prag	Veličina najveće klike
7.5	219	25	101

Tablica 4.15: Rezultat za *upit* = `DHILKGQNKA`

Najveća klika za upit `DHILKGQNKA` sadrži 46.119% nizova aminokiselina iz odgovora.

Rajčica

Rajčica (lat. *Solanum lycopersicum*) je jednogodišnja biljka iz porodice pomoćnica (lat. *Solanaceae*). Uzgaja se zbog svojih plodova i važna je prehrambena biljka. Analiziramo strukturu odgovora dobivenog pretraživanjem ovog biljnog proteoma u odnosu na sljedeće upite:

1. *Upit* = `FVFGDSLSDA`

Skala	Duljina odgovora	Prag	Veličina najveće klike
7.5	128	25	88

Tablica 4.16: Rezultat za *upit* = `FVFGDSLSDA`

I na proteomu rajčice je najveća klika odgovora, na upit `FVFGDSLSDA`, velika i sadrži 68.750% odgovora. Također, za ovaj upit i proteom imamo biološku listu pozitivaca koja sadrži 108 proteina. Usporedimo naš odgovor i dobivenu najveću kliku sa biološkom listom pozitivaca.

	Broj pozitivaca	TP	Različiti TP	PPV	TPR
Odgovor	128	91	91	0.711	0.843
Najveća klika	88	86	86	0.977	0.796

Tablica 4.17: Usporedba sa biološkom listom pozitivaca

Velik dio c-pozitivaca je u odgovoru, a 94.505% ih je ostalo sadržano u najvećoj kliki. Uočimo kako čak 97.727% najveće klike čine upravo c-pozitivci.

2. *Upit* = `PEPLISEILF`

Skala	Duljina odgovora	Prag	Veličina najveće klike
7.5	49	25	8

Tablica 4.18: Rezultat za *upit* = `PEPLISEILF`

Najveća klika odgovora na upit `PEPLISEILF` sadrži samo 16.327% odgovora.

3. *Upit* = `EIFECRESLT`

Skala	Duljina odgovora	Prag	Veličina najveće klike
7.5	44	25	9

Tablica 4.19: Rezultat za *upit* = `EIFECRESLT`

Najveća klika za upit `EIFECRESLT` sadrži 20.455% nizova aminokiselina iz odgovora.

4. *Upit* = `DHILKGQNKA`

Skala	Duljina odgovora	Prag	Veličina najveće klike
7.5	62	25	14

Tablica 4.20: Rezultat za *upit* = `DHILKGQNKA`

Najveća klika za upit `DHILKGQNKA` sadrži 22.581% odgovora.

Bakterija Streptomyces avermitilis

Streptomyces avermitilis je bakterijska vrsta iz roda *Streptomyces* i porodice *Streptomycetaceae*. Analiziramo strukturu odgovora dobivenog pretraživanjem proteoma ove bakterije u odnosu na upit `GESGCGKS` koji je Walker-ov A motiv.

Skala	Duljina odgovora
4	399

Tablica 4.21: Odgovor na *upit* = `GESGCGKS`

Za upite duljine 8 prosječne sličnosti simuliranih nizova, iz procjenjenih distribucija za upit, variraju više nego za upite duljine 10. Budući da prag definiramo upravo kao prosječnu sličnost tih simuliranih nizova promatrat ćemo više pragova: $18, 19, 20, 21, 22$ i 23. S obzirom na svaki od tih pragova ćemo definirati $0-1$ graf pridružen odgovoru i tražiti njegovu najveću kliku.

Prag	Veličina najveće klike
18	307
19	302
20	295
21	287
22	279
23	273

Tablica 4.22: Rezultati za *upit* = GESGCGKS

Najveća klika odgovora na upit GESGCGKS je velika za svaki od zadanih pragova i sadrži približno 70% nizova aminokiselina iz odgovora.

Za upit koji je Walker-ov A motiv i proteom *Streptomyces avermitilis* imamo biološku listu pozitivaca koja sadrži imena 459 proteina u kojima bi se trebali nalaziti nizovi aminokiselina slični upitu. Usporedimo biološku listu pozitivaca sa proteinima u kojima se nalaze nizovi aminokiselina iz odgovora i iz najveće klike za svaki od gore spomenutih pragova.

	Broj pozitivaca	TP	Različiti TP	PPV	TPR
Odgovor	399	211	196	0.529	0.427
Najveća klika za *prag* = 18	307	205	191	0.668	0.416
Najveća klika za *prag* = 19	302	206	192	0.682	0.418
Najveća klika za *prag* = 20	295	205	191	0.695	0.416
Najveća klika za *prag* = 21	287	203	191	0.707	0.416
Najveća klika za *prag* = 22	279	201	189	0.720	0.412
Najveća klika za *prag* = 23	273	201	188	0.736	0.410

Tablica 4.23: Usporedba sa biološkom listom pozitivaca

Uočimo kako više od 50% odgovora čine c-pozitivci, a u najvećoj kliki oni čine i veći udio. Također, oko 95% pozitivaca iz odgovora je ostalo sadržano u najvećoj kliki dok je veličina klike u odnosu na odgovor manja za nekih 30% i to za svaki od pragova kojima smo definirali pripadni $0-1$ graf.

4.3 Analiza rezultata

Analizom strukture različitih odgovora dobivenih iterativnim pretraživanjem biljnih proteoma talijinog uročnjaka, krumpira, azijske riže i rajčice, u odnosu na upite `FVFGDSLSDA`, `PEPLISEILF`, `EIFECRESLT` i `DHILKGQNKA`, dobivamo vrlo slične rezultate. Naime, najveća klika odgovora na upit `FVFGDSLSDA` sadrži redom 78.621%, 73.077%, 66.667% i 68.750% nizova aminokiselina iz odgovora s obzirom na različite proteome. Nadalje, najveća klika odgovora na upit `PEPLISEILF` sadrži redom 42.754%, 35.455%, 27.778% i 16.327% nizova iz odgovora. Za upit `EIFECRESLT` najveća klika sadrži 49.333%, 26.531%, 25.974% i 20.455% nizova iz odgovora s obzirom na različite proteome, dok za upit `DHILKGQNKA` sadrži 22.951%, 18%, 45.119% i 22.581% odgovora.

Usporedili smo biološku listu pozitivaca za upit `FVFGDSLSDA` na različitim proteomima sa odgovorom na taj upit i na osnovu toga zaključili da su ti odgovori biološki značajni. Naime, proteini u kojima se nalaze nizovi aminokiselina iz odgovora su velikom većinom upravo c-pozitivci karakteristični za familiju GDSL lipaza. Uočimo također kako su najveće klike odgovora na taj upit dosta velike i sadrže većinu odgovora te da su skoro svi c-pozitivci iz odgovora ostali sačuvani u najvećim klikama. Nadalje, na svim promatranim proteomima vrijednosti TPR-a su se malo smanjile na najvećoj kliki, u odnosu na odgovor, najviše za 5.8% na proteomu *Arabidopsis thaliana* dok se na proteomu riže ta vrijednost smanjila za samo 0.9%. Vrijednosti PPV-a su se pak dosta povećale na najvećoj kliki na svim promatranim proteomima, čak i za 30.6% na proteomu riže. Odgovori na upit `PEPLISEILF`, `EIFECRESLT` i `DHILKGQNKA` nisu značajni ni na jednom proteomu budući da su ti upiti slučajno odabrani nizovi aminokiselina koji nisu karakteristični ni za jednu proteinsku familiju pa su takvi i nizovi u odgovorima. Također, pripadne najveće klike tih odgovora su se raspale. Na temelju svih navedenih rezultata uočavamo kako struktura odgovora zaista korespondira biološkoj značajnosti odgovora.

Također smo analizirali strukturu odgovora dobivenog pretraživanjem proteoma bakterije *Streptomyces avermitilis* u odnosu na Walker-ov A motiv `GESGCGKS`. Najveća klika $0-1$ grafa pridruženog odgovoru sadrži 70−ak% nizova iz odgovora, a odgovor je biološki značajan jer više od pola odgovora čine c-pozitivci. Uočimo i kako su skoro svi c-pozitivci, njih 95%, ostali sačuvani u najvećoj kliki. I na ovom primjeru uočavamo korespondenciju između strukture odgovora i njegove biološke značajnosti.

Bibliografija

[1] S. Henikoff i J. Henikoff, *Amino acid substitution matrices from protein blocks*, Proc. Natl. Acad. Sci. USA (1992), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC50453/pdf/pnas01096-0363.pdf.

[2] M. Kobovac, *Neki aspekti iterativnog pretraživanja proteoma*, Diplomski rad, Sveučilište u Zagrebu, Prirodoslovno-matematički fakultet, Matematički odsjek, 2017.

[3] B. Rabar, S. Ristov, M. Zagorščak, M. Rosenzweig i P. Goldstein, *Igloss: Iterative gapless local similarity search*, arXiv:1807.11862v1 [q-bio.QM] (2018.), https://arxiv.org/pdf/1807.11862.pdf.

[4] A. Relja, *Neki statistički aspekti prepoznavanja motiva*, Diplomski rad, Sveučilište u Zagrebu, Prirodoslovno-matematički fakultet, Matematički odsjek, 2014.

[5] N. Sarapa, *Teorija vjerojatnosti*, Školska Knjiga, Zagreb, 2002.

[6] D. Veljan, *Kombinatorna i diskretna matematika*, Algoritam, Zagreb, 2001.

Sažetak

Pretpostavka ili hipoteza u ovom radu je da struktura odgovora dobivenog iterativnim pretraživanjem proteoma nekog organizma korespondira biološkoj značajnosti odgovora. Proveli smo analizu odgovora dobivenih iterativnim pretraživanjem 4 biljna i jednog bakterijskog proteoma u odnosu na 5 upita. Na svim odgovorima tražili smo najveću kliku $0-1$ grafa pridruženog odgovoru i dobili vrlo slične rezultate. Najveće klike biološki značajnih odgovora su sačuvale oko 70% odgovora dok su se ostale najveće klike raspale. Time smo pokazali da zaista postoji korespondencija između strukture i biološke značajnosti odgovora te da bi na osnovi veličine najveće klike mogli donijeti zaključak o biološkoj značajnosti odgovora. Također, na najvećim klikama biološki značajnih odgovora se povećavaju vrijednosti PPV-a, odnosno najveće klike sadrže veći udio pravih pozitivaca nego odgovori.

Summary

This work is concerned with the analysis of the iterative scanning response. We analyze responses for various queries against 4 plant and 1 bacterial proteome. We introduce a weighted graph structure on the response and, after certain adjustments, apply a maximal clique algorithm. We show that the size of the maximal clique corresponds strongly to the biological significance of the response. Furthermore, we show that the maximal clique captures the true positive part of the response very reliably.

Životopis

Rođena sam 18.11.1994. u Supetru na otoku Braču. Osnovnu školu sam pohađala u Pučišćima, a Opću gimnaziju u Srednjoj školi Brač u Supetru. Nakon završenog srednjoškolskog obrazovanja, 2013. upisujem Preddiplomski sveučilišni studij Matematike na Prirodoslovno-matematičkom fakultetu u Zagrebu. 2016. stječem naziv sveučilišne prvostupnice matematike te iste godine upisujem Diplomski studij Matematičke statistike.