SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Antonia Berko

ITERATIVNO TRAŽENJE MOTIVA I NEKODIRAJUĆA DNA

Diplomski rad

Voditelj rada:
doc. dr. sc. Pavle Goldstein

Zagreb, veljača, 2019.

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. _____

Sadržaj

Uvod

Bioinformatika je znanost koja, kako joj ime i kaže, spaja biologiju i informatiku. Centar te znanosti je upravo matematika. U ovom radu, uz matematičko znanje i vještine, pristupamo jednom od trenutno aktualnih problema u bioinformatici. Svjetski ekonomski problem je sve manja količina plodnog tla, a sve veći broj ljudi. Glavne prehrambene namirnice diljem svijeta su riža i krumpir pa je pitanje kako omogućiti da te biljke uspiju na slanom tlu kojeg ima dovoljno, a neiskorišteno je. To je velik problem koji zahtjeva godine istraživanja i na kojem već radi brojni tim stručnjaka. Kako u tom, tako i u manjim problemima te vrste potrebno je pretraživanje proteoma organizma i otkrivanje svrhe svakog proteina u njemu. Problem kojim se mi bavimo uključuje iterativno pretraživanje proteoma, a riječ je o identificiranju biološki značajnih nizova aminokiselina u proteomu tj. onih nizova koji pripadaju proteinskoj familiji koju tražimo.

Na proteomima četiri različita organizma (među kojima i riža i krumpir) pokušat ćemo uz karakteristični motiv za proteinsku familiju koju tražimo pronaći što veći broj biološki značajnih nizova. Isto ćemo pokušati i za modificirane motive koji nisu svojim rasporedom aminokiselina sasvim vezani za tu familiju. Metodu kojom provodimo navedeni postupak nazivamo metodom centralnog motiva jer ćemo u odgovoru iterativne pretrage na naše motive - upite tražiti onaj motiv s kojim su svi drugi motivi u odgovoru najviše povezani te ćemo ga proglasiti centralnim motivom. Zatim ćemo istražiti kakav on utjecaj ima na pronalazak biološki značajnih motiva. Za sve navedeno nije dovoljno samo matematičko znanje nego i poznavanje osnovnih bioloških pojmova, programerske vještine te podrška za stručne stvari od strane biologa.

Napomenimo da "pronaći što veći broj biološki značajnih nizova" znači povećati točnost pridruživanja kodirajuće DNA. Poboljšavanjem točnosti klasifikacije proteinskih nizova ujedno poboljšavamo i analizu nekodirajuće DNA.

Ovaj rad je podijeljen u šest poglavlja. U prvom i drugom poglavlju su definirani matematički pojmovi iz vjerojatnosti, statistike i Markovljevih lanaca. U trećem poglavlju definiramo bioinformatičke pojmove i strukture kojima se koristimo. U četvrtom poglavlju objašnjavamo kroz primjer našu metodu centralnog motiva te u petom poglavlju na stvarnim podacima prikazujemo rezultate istraživanja. U zadnjem, šestom, poglavlju iznosimo zaključke rada.

Poglavlje 1

Vjerojatnost i funkcije distribucije

1.1 Vjerojatnost

Definicija 1.1.1. ***Slučajni pokus*** *ili* ***slučajni eksperiment*** *je pokus čiji ishod, tj. rezultat nije jednoznačno određen uvjetima u kojima izvodimo pokus.*

Definicija 1.1.2. *Neka je Ω neprazan skup. Familija podskupova $\mathcal{F}$ od Ω ($\mathcal{F} \subset \mathcal{P}(\Omega)$) je* ***$\sigma$-algebra skupova*** *(na Ω) ako vrijede sljedeća tri svojstva :*

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

(ii) $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$ *(zatvorenost na komplementiranje)*

(iii) $A_i \in \mathcal{F}, i \in \mathbb{N} \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$ *(zatvorenost na prebrojive unije)*

Uređen par $(\Omega, \mathcal{F})$ zove se izmjeriv prostor.

Definicija 1.1.3. *Neka je $(\Omega, \mathcal{F})$ izmjeriv prostor. Funkcija $P : \mathcal{F} \to \mathbb{R}$ je* ***vjerojatnost*** *na $\mathcal{F}$ ako vrijedi:*

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

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

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

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.

Definicija 1.1.4. *(**Uvjetna vjerojatnost**) Neka je $(\Omega, \mathcal{F}, P)$ vjerojatnosni prostor i $B \in \mathcal{F}$ takav da je $P(B) > 0$. Uvjetna vjerojatnost događaja A uz uvjet B definira se formulom:*

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

*Lako se pokaže da je P_B vjerojatnost na $\mathcal{F}$. Broj $P(B \mid A)$ zovemo **vjerojatnost od B uz uvjet A**.*

Definicija 1.1.5. *(**Nezavisni događaji**) 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*

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

1.2 Funkcije distribucije

Definicija 1.2.1. *Neka je $\mathcal{B}$ σ-algebra generirana familijom svih otvorenih skupova na $\mathbb{R}$. $\mathcal{B}$ zovemo **σ-algebra skupova** na $\mathbb{R}$, a elemente σ-algebre zovemo **Borelovi skupovi**.*

Definicija 1.2.2. *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 1.2.3. *(**Nezavisne slučajne varijable**) Neka su $X_1, X_2..., X_n$ slučajne varijable na vjerojatnosnom prostoru $(\Omega, \mathcal{F}, P)$. Kažemo da su $X_1, X_2..., X_n$ **nezavisne slučajne varijable** ako za proizvoljne $B_i \in \mathcal{B}, i = 1, 2, ..., n$ vrijedi :*

$$P\Big(\bigcap_{i=1}^{n}(X_i \in B_i)\Big) = \prod_{i=1}^{n} P(X_i \in B_i). \tag{1.3}$$

Definirajmo još vjerojatnosnu mjeru induciranu sa slučajnom varijablom X, kako bi konačno mogli definirati funkciju distribucije od X.

Definicija 1.2.4. *Neka je $(\Omega, \mathcal{F}, P)$ vjerojatnosni prostor i X slučajna varijabla na Ω. Za $B \in \mathcal{B}$ definiramo funkciju $P_X : \mathcal{B} \to [0, 1]$ sljedećom relacijom:*

$$P_X(B) = P(X^{-1}(B)) = P\{\omega \in \Omega : X(\omega) \in B\} = P\{X \in B\}. \tag{1.4}$$

*Onda je P_X vjerojatnost, odnosno vjerojatnosna mjera na $\mathcal{B}$. P_X zovemo **vjerojatnosna mjera inducirana sa X**, a vjerojatnosni prostor, koji je pridružen slučajnoj varijabli X, $(\mathbb{R}, \mathcal{B}, P_X)$ zovemo **vjerojatnosni prostor induciran sa X**.*

Nakon što smo definirali slučajnu varijablu i vjerojatnosnu mjeru induciranu njome možemo definirati funkciju distribuciju od X, a nakon toga ćemo moći nešto više reći o vrstama slučajnih varijabli - diskretnim i neprekidnim slučajnim varijablama.

Definicija 1.2.5. *Neka je X slučajna varijabla na* Ω. ***Funkcija distribucije od X*** *je funkcija* $F_X : \mathbb{R} \to [0,1]$ *definirana sa*

$$\begin{aligned} F_X(x) &= P_X((-\infty, x]) = P(X^{-1}(-\infty, x]) = \\ &= P\{\omega \in \Omega : X(\omega) \leqslant x\} = P\{X \leqslant x\}, \quad x \in \mathbb{R}. \end{aligned} \tag{1.5}$$

Teorem 1.2.6. *Funkcija distribucije F slučajne varijable X je rastuća i neprekidna zdesna na* $\mathbb{R}$ *i zadovoljava*

$$\begin{aligned} F(-\infty) &= \lim_{x \to -\infty} F(x) = 0 \\ F(+\infty) &= \lim_{x \to +\infty} F(x) = 1. \end{aligned} \tag{1.6}$$

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**.

Definicija 1.2.7. ***(Diskretna slučajna varijabla)*** *Slučajna varijabla X je* ***diskretna*** *ako postoji konačan ili prebrojiv skup* $D \subset \mathbb{R}$ *takav da je* $P\{X \in D\} = 1$.

Definicija 1.2.8. ***(Apsolutna neprekidna slučajna varijabla)*** *Slučajna varijabla X je* ***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*

$$F_X(x) = \int_{-\infty}^{x} f(t) d\lambda(t), \quad x \in \mathbb{R}. \tag{1.7}$$

Za funkciju distribucije $F_X(x)$ slučajne varijable X oblika 1.7 kažemo da je **apsolutno neprekidna funkcija distribucije** te u tom slučaju za funkciju ($f : \mathbb{R} \to \mathbb{R}_+$) kažemo da je **funkcija gustoće vjerojatnosti od** X.

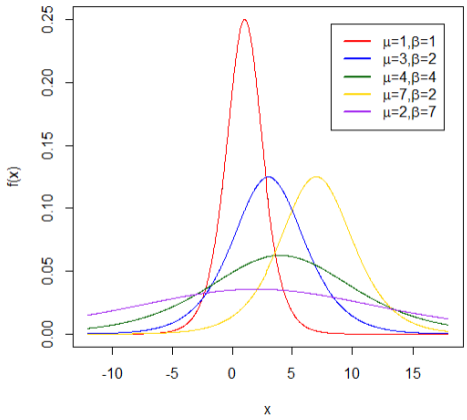
1.3 Primjeri distribucija

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

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

Navedimo još da je srednja vrijednost logističke distribucije dana sa μ, dok je varijanca jednaka $\sigma^2 = \frac{\beta^2\pi^2}{3}$. U nastavku grafički prikaz funkcije gustoće logističke distribucije.

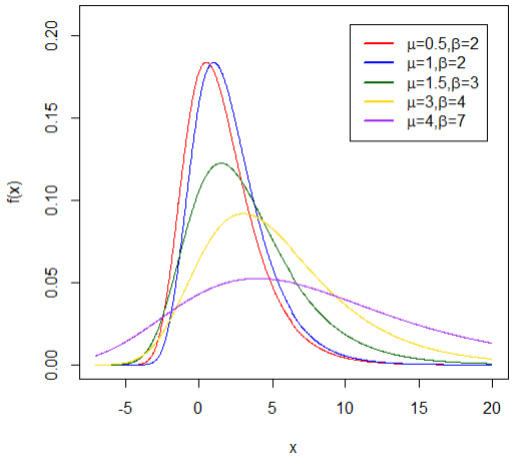


Slika 1.1: Graf funkcije gustoće logističke distribucije za razne vrijednosti parametara μ i β

Gumbel distribucija

Neka je $\mu \in R$ i $\beta > 0$. Neprekidna slučajna varijabla X ima **Gumbel distribuciju** s parametrima μ i β ako joj je funkcija gustoće dana s:

$$f(x) = \frac{1}{\beta} e^{-\frac{x-\mu}{\beta} - e^{-\frac{x-\mu}{\beta}}}, \quad x \in R \tag{1.8}$$



Slika 1.2: Graf funkcije gustoće Gumbelove distribucije za razne vrijednosti parametara μ i β

Teorem 1.3.1. *Neka su X_1 i X_2 nezavisne slučajne varijable s Gumbelovom distribucijom s parametrima p i q, respektivno. Tada slučajna varijabla $Y = X_1 - X_2$ ima logističku distribuciju s parametrima p i q.*

Teorem 1.3.2 (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.

1.4 Statistički testovi i statističke mjere

Svako statističko istraživanje temelji se na provođenju statističkog testa, koji je temeljen na hipotezama. **Statistička hipoteza** je pretpostavka o razdiobi X, gdje je X neko statističko obilježje.

Uz osnovnu, **nultu hipotezu** (H_0) postoji i **alternativna hipoteza** (H_1). Provođenjem statističkog testa dolazimo do zakjučka hoćemo li odbaciti nultu hipotezu ili nećemo. Postupak donošenja odluke o odbacivanju statističke hipoteze zove se **testiranje statističke hipoteze**.

Definicija 1.4.1. ***Test*** *(hipoteze H_0 u odnosu na alternativu H_1) je preslikavanje $\tau : \mathbb{R}^n \to \{0, 1\}$.*

Interpretacija. Ako je za realizaciju $\mathbf{x}$ uzorka $\mathbf{X}$, $\tau(x) = 1$ tada odbacujemo H_0 u korist H_1, a ako je $\tau(x) = 0$ tada ne odbacujemo H_0 u korist H_1.

Za potpuno provođenje nekog testa i zaključivanje, potrebno nam je poznavanje konkretne razine značajnosti na kojoj provodimo test. Da bi definirali pojam značajnosti prvo je potrebno reći nešto o pogreškama prve i druge vrste. Pogreška koju činimo kada odbacujemo H_0, a ona je istinita, je **pogreška prve vrste**. Pogreška koju činimo kada ne odbacujemo H_0, a istinita je H_1, je **pogreška druge vrste**. Navedeno najbolje možemo shvatiti iz sljedeće tablice:

	zaključak	
točno je	ne odbaciti H_0	odbaciti H_0
H_0	$\surd$	pogrešno! (I)
H_1	pogrešno! (II)	$\surd$

Definicija 1.4.2. *Preslikavanje $\alpha : \Theta_0 \to [0, 1]$ definirano sa :*

$$\alpha(\theta) := \gamma(\theta) = P_\theta(\boldsymbol{X} \in C) \tag{1.9}$$

je vjerojatnost pogreške prve vrste. Ovdje je C kritično područje za test τ definirano sa

$$C := \tau^{-1}(1) = \{x \in \mathbb{R}^n : \tau(x) = 1\} \tag{1.10}$$

*Kažemo da test ima **razinu značajnosti** α, ukoliko mu je značajnost manja ili jednaka α. Gdje je značajnost testa τ dana sa:*

$$\alpha_\tau := \sup_{\theta \in \Theta_0} \alpha(\theta) \tag{1.11}$$

Sada možemo definirati statističke mjere uspješnosti provedenog testa - **osjetljivost i specifičnost**. Preko tih mjera ćemo doći do definicije ključne mjere za ovaj rad, a to je **F1 score**.

Oznake (izvedene iz engleskih riječi za dane pojmove) koje ćemo koristiti:

- odgovor - rezultat pretraživanja elemenata sa određenim svojstvom na nekom skupu
- CP - elementi pozitivnog stanja u skupu
- CN - elementi negativnog stanja u skupu
- P - pozitivci, svi elementi u odgovoru
- N - negativci, svi elementi koji nisu u odgovoru

  U skupu P se mogu nalaziti:
- TP - stvarno pozitivni elementi, $TP = CP \cap P$
- FP - lažno pozitivni elementi, $FP = CN \cap P$

  U skupu N se mogu nalaziti:
- FN - lažno negativni elementi, $FN = CP \cap N$
- TN - stvarno negativni elementi $TN = CN \cap N$

Napomena. Nama će za računanje statističkih mjera trebati broj elemenata u navedenim skupovima, a to ćemo označavati sa : $|TP|$ - broj elemenata u skupu TP.

Osjetljivost (eng. *sensitivity*) ili **TPR** (eng. *true positive rate*) nam daje omjer stvarno pozitivnih elemenata odgovora u odnosu na sumu stvarno pozitivnih i lažno negativnih. Drugim riječima, u odnosu na određeno stanje - CP (eng. *condition positive*).

$$\text{TPR} = \frac{|TP|}{|CP|}$$

Specifičnost (eng. *specificity*) ili **TNR** (eng. *true negative rate*) nam daje omjer stvarno negativnih elemenata odgovora u odnosu na sumu stvarno negativnih i lažno pozitivnih. Drugim riječima, u odnosu na određeno stanje - CN (eng. *condition negative*).

$$\text{TNR} = \frac{|TN|}{|CN|}$$

Pozitivna prediktivna vrijednost (eng. *positive predictive value*) ili **PPV** je omjer broja stvarno pozitivnih elemenata i broja svih elemenata u odgovoru i time dobivamo udio onih koji su točno prepoznati kao pozitivni u kompletnom odgovoru.

$$\text{PPV} = \frac{|TP|}{|P|}$$

Napomena. Pogreška prve vrste = $1 - PPV$.

Negativna prediktivna vrijednost (eng. *negative predictive value*) ili **NPV** je omjer broja stvarno negativnih elemenata i broja svih elemenata koji nisu u odgovoru i time dobivamo udio elemenata koji su ispravno prepoznati kao negativni u skupu negativaca.

$$\text{NPV} = \frac{|TN|}{|N|}$$

		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 statističkih mjera uspješnosti testa

F1 score je jedna od mjera točnosti testa. Računa se kao harmonijska sredina TPR-a i PPV-a tj. osjetljivosti i pozitivne prediktivne vrijednosti. Formula glasi:

$$F1 = \Big(\frac{TPR^{-1} + PPV^{-1}}{2}\Big)^{-1} = 2 * \frac{TPR * PPV}{TPR + PPV} \tag{1.12}$$

F1 score može postizati vrijednosti u intervalu [0,1], gdje 1 predstavlja najbolju vrijednost, a 0 najgoru. Kada bi F1 score iznosio 1, to bi značilo da su se u odgovoru (P) našli svi elementi pozitivnog stanja (CP) i nijedni drugi. Hoćemo li našom metodom uspjeti poboljšati odgovor saznat ćemo iz usporedbe F1 score-a prije i nakon metode centralnog motiva. Također, iz usporednog grafičkog prikaza ćemo jednostavno vidjet što se događa sa F1 score-om za iste vrijednosti PPV-a, prije i nakon metode.

Pojmovi iz ovog poglavlja preuzeti su iz izvora [6], [5] i [3].

Poglavlje 2

Markovljevi lanci

Markovljevi lanci s diskretnim vremenom

Definicija 2.0.1. *Neka je S skup.* ***Slučajan proces*** *s diskretnim vremenom i prostorom stanja S je familija $X = (X_n : n \geq 0)$ slučajnih varijabli definiranih na nekom vjerojatnosnom prostoru $(\Omega, \mathcal{F}, P)$ s vrijednostima u S. Dakle, za svaki $n \geq 0$ je $X_n : \Omega \to S$ slučajna varijabla.*

Definicija 2.0.2. *Neka je S prebrojiv skup. Slučajni proces $X = (X_n : n \geq 0)$ definiran na vjerojatnosnom prostoru $(\Omega, \mathcal{F}, P)$ s vrijednostima u skupu S je* ***Markovljev lanac*** *ako vrijedi*

$$P(X_{n+1} = j \,|\, X_n = i, X_{n-1} = i_{n-1}, ..., X_0 = i_0) = P(X_{n+1} = j \,|\, X_n = i) \tag{2.1}$$

za svaki $n \geq 0$ i za sve $i_0, ..., i_{n-1}, i, j \in S$ za koje su obje uvjetne vjerojatnosti dobro definirane.

Svojstvo u relaciji (2.1) nazivamo ***Markovljevim svojstvom***

Definicija 2.0.3. *Matrica $P = (p_{ij} : i, j \in S)$ naziva se* ***stohastičkom matricom*** *ako je $p_{ij} \geq 0$, $\forall i, j \in S$ te*

$$\sum_{j \in S} p_{ij} = 1, \quad za\ sve\ i \in S. \tag{2.2}$$

Definicija 2.0.4. *Neka je $\lambda = (\lambda_i : i \in S)$ vjerojatnosna distribucija na S, te neka je $P = (p_{ij} : i, j \in S)$ stohastička matrica. Slučajni proces $X = (X_n : n \geq 0)$ definiran na vjerojatnosnom prostoru $(\Omega, \mathcal{F}, P)$ s prostorom stanja S je* ***homogen Markovljev lanac*** *s početnom distribucijom λ i prijelaznom matricom P ako vrijedi*

(i) $P(X_0 = i) = \lambda_i$ *za sve $i \in S$, te*

(ii)

$$P(X_{n+1} = j \,|\, X_n = i, X_{n-1} = i_{n-1}, ..., X_0 = i_0) = p_{ij} \tag{2.3}$$

za svaki $n \geq 0$ *i za sve* $i_0, ..., i_{n-1}, i, j \in S$.

Definicija 2.0.5. *Markovljev lanac X je* ***ireducibilan*** *ako se prostor stanja S sastoji samo od jedne klase komuniciranja tj. za sve* $i, j \in S$ *vrijedi* $i \longleftrightarrow j$.

Definicija 2.0.6. *Slučajan proces* $X = (X_n : n \geq 0)$ *definiran na vjerojatnosnom prostoru* $(\Omega, \mathcal{F}, P)$ *zove se* ***stacionaran*** *ako za sve* $k \geq 0$ *i sve* $n \geq 0$ *, slučajni vektori* $(X_0, X_1, ..., X_k)$ *i* $(X_n, X_{n+11}, ..., X_{n+k})$ *imaju istu distribuciju.*

Definicija 2.0.7. *Neka je* $X = (X_n : n \geq 0)$ *Markovljev lanac s prebrojivim skupom stanja S i prijelaznom matricom P . Vjerojatnosna distribucija* $\pi = (\pi_i : i \in S)$ *na S je* ***stacionarna distribucija*** *(ili invarijantna) Markovljevog lanca X ako vrijedi*

$$\pi = \pi P \tag{2.4}$$

odnosno po komponentama

$$\pi_j = \sum_{k \in S} \pi_k p_{kj}, \ \ za \ \ sve \ \ j \in S \tag{2.5}$$

Teorem 2.0.8. *Neka je* $X = (X_n : n \geq 0)$ (π, P)*- Markovljev lanac gdje je* π *stacionarna distribucija za P. Tada je X* ***stacionaran proces****. Preciznije, X je stacionaran uz vjerojatnost* $P_\pi = \sum_{i \in S} \pi_i P_i$ *za sve* $j \in S$.

Specijalno, n-ta potencija matrice P dana je s $P^n = (p_{ij}^{(n)} : i, j \in S)$, gdje je

$$p_{ij}^{(n)} = \sum_{i_1 \in S} ... \sum_{i_{n-1} \in S} p_{ii_1} p_{i_1 i_2} ... p_{i_{n-2} i_{n-1}} p_{i_{n-1} j} \tag{2.6}$$

Propozicija 2.0.9. *Neka je S konačan skup stanja te pretpostavimo da za neki* $i \in S$*,*

$$\lim_{n \to \infty} p_{ij}^{(n)} = \pi_j \quad , \ za \ \ sve \ \ j \in S \tag{2.7}$$

tada je $\pi = (\pi_j : j \in S)$ *stacionarna distribucija.*

Definicija 2.0.10. *Niz* $\lambda = (\lambda_i : i \in S)$ *naziva se mjera ako je* $\lambda_i \in [0, \infty)$ *za sve* $i \in S$. *Mjera* λ *je netrivijalna ako postoji* $i \in S$ *takav da* $\lambda_i > 0$. *Neka je* $X = (X_n : n \geq 0)$ *Markovljev lanac s prijelaznom matricom P. Netrivijalna mjera* λ *na S je* ***invarijantna mjera*** *Markovljevog laca X (odnosno prijelazne matrice P) ako vrijedi*

$$\lambda = \lambda P \tag{2.8}$$

odnosno po komponentama

$$\lambda_j = \sum_{k \in S} \lambda_k p_{kj}, \ \ za \ \ sve \ \ j \in S \tag{2.9}$$

Definicija 2.0.11. *Neka je $X = (X_n : n \geq 0)$ Markovljev lanac na skupu stanja S s prijelaznom matricom P. Vjerojatnosna distribucija $\pi = (\pi_i : i \in S)$ naziva se* ***graničnom distirbucijom*** *Markovljevog lanca X ako za sve $i, j \in S$ vrijedi*

$$\lim_{n\to\infty} p_{ij}^{(n)} = \pi_j \tag{2.10}$$

Propozicija 2.0.12. *Neka je π granična distribucija Markovljevog lanca X. Tada je π i stacionarna distribucija.*

Lema 2.0.13. *Pretpostavimo da je Markovljev lanac X ireducibilan i aperiodičan. Tada za sve $i, j \in S$ postoji $n_0 = n_0(i, j) \in N$ takav da je $p_{ij}^{(n)} > 0$, $\forall n \geq n_0$.*

Teorem 2.0.14. *Neka je λ proizvoljna vjerojatnosna distribucija na skupu stanja S. Pretpostavimo da je $X = (X_n : n \geq 0)$ (λ, P) - Markovljev lanac koji je ireducibilan i aperiodičan te ima stacionarnu distribuciju π . Tada je*

$$\lim_{n\to\infty} P(X_n = j) = \pi_j \text{ za sve } j \in S \tag{2.11}$$

$$\lim_{n\to\infty} p_{ij}^{(n)} = \pi_j \ \forall i, j \in S \tag{2.12}$$

tj. stacionarna distribucija je ujedno i granična.

Markovljevi lanci s neprekidnim vremenom

Definicija 2.0.15. *Vjerojatnosna distribucija $\lambda = (\lambda_i, i \in S)$ se zove granična distribucija Markovljevog lanca $X = (X_t : t \geq 0)$ ako vrijedi*

$$\lim_{t\to\infty} P_{ij} = \lambda_j \ \forall i, j \in S \tag{2.13}$$

Teorem 2.0.16. ***(Konvergencija prema graničnoj distribuciji)***
Neka je $X = (X_t : t \geq 0)$ ireducibilan i regularan Markovljev lanac s generatorskom matricom Q i polugrupom $(P(t) : t \geq 0)$. Pretpostavimo da X ima invarijantnu distribuciju λ. Tada je λ ujedno i granična distribucija.

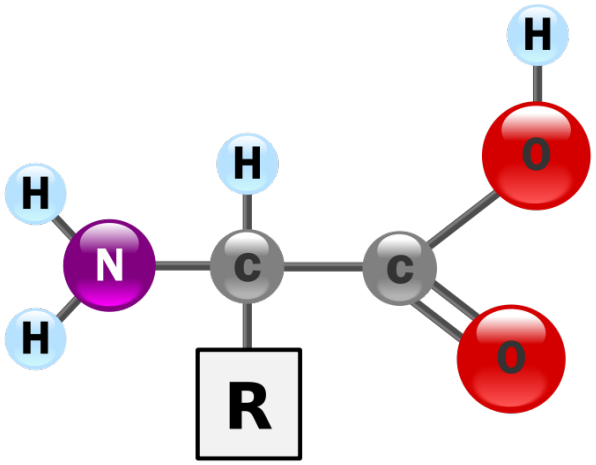
Dokaz ovog teorema može se naći u [8] . Ostali pojmovi iz ovog poglavlja preuzeti su iz izvora [7] .

Poglavlje 3

Bioinformatički pojmovi

3.1 Biološki pojmovi

Uz matematičke pojmove, za razumijevanje rada potrebno je upoznati se i sa osnovim biološkim pojmovima koje ćemo koristiti. Sva iterativna pretraživanja i testiranja koja provodimo su na proteomima biljaka, stoga definiranje pojmova možemo započeti od osnovnih strukturnih jedinica za nastanak proteoma, a to su aminokiseline. **Aminokiseline** su molekule koje sadrže amino skupinu, karboksilnu skupinu i bočni lanac. Ono što razlikuje jednu aminokiselinu od druge je bočni lanac koji može biti vrlo jednostavan kao kod glicina ili složen kao kod triptofana.



Slika 3.1: Struktura aminokiseline

Njihova glavna biološka uloga je, upravo, izgradnja proteina. Aminokiseline se peptidnom vezom (vezom između karboksilne skupine jedne aminokiseline i amino skupine druge aminokiseline) međusobno vežu u lanac i tako tvore proteine, a sve to se događa u "tvornici bjelančevina" tj. na ribosomima. Proteini su kemijske tvari koje upravljaju svim životnim procesima stanice. Izgrađeni su od 20 različitih aminokiselina koje navodimo u tablici:

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

Proteom je skup svih proteina i proteinskih oblika koje organizam proizvodi tijekom života te koji nastaju kao posljedica ekspresije gena u određenom trenutku u vremenu, pod određenim uvjetima. Napomenimo još da ćemo kroz rad spominjati riječ "motiv" sa značenjem "niz aminokiselina neke konačne duljine".

GDSL lipaze

Lipaze su skupine enzima koji kataliziraju hidrolizu lipida. Postoji velik broj neotkrivenih lipaza i onih novo otkrivenih, jedne od takvih su GDSL lipaze. Nama će u ovom radu, za analiziranje i testiranje, biti zanimljivi motivi koji su biološki karakteristični za GDSL lipaze. Želimo vidjeti možemo li identificirati takve motive u nekom proteomu. Razlog zbog kojeg su GDSL lipaze zanimljive u svijetu genetičkih istraživanja je taj što imaju multifunkcionalna svojstva. Jedno od glavnih je široka specifičnost supstrata tj. ove lipaze imaju fleksibilno katalitičko (aktivno) mjesto koje mijenja svoju strukturu u prisutnosti različitih supstrata. GDSL lipaze su podskupina lipotičkih enzima, a razlikuju se od ostalih podskupina po tome što ne sadrži uobičajeni GxSxS motiv. Razlog njihove velike važnosti u prehrambenoj, farmaceutskoj i biomedicinskoj industriji je taj što imaju potencijal za primjenu u hidrolizi i sintezi važnih spojeva u biološkom svijetu. Te lipaze su već

pronađene u raznim živim organizmima. Otkriveno je da bi biljke mogle biti poseban izvor GDSL lipaza i zato je od velikog biološkog interesa pronalazak tj. identificiranje što većeg broja GDSL lipaza u proteomima biljaka.

Bitno je za napomenuti da proteinska familija GDSL lipaza ne sadrži nužno nizove sa GDSL motivom. Štoviše, takvih je vrlo malo. Iz tog razloga, traženje novih GDSL lipaza i predstavlja problem jer se ne zna po kojem točno zapisu ih se može pronaći. Pokažimo to na primjeru motiva biološki karakterističnih za GDSL lipaze kod biljke talijin uročnjak. Navedimo neke motive iz te skupine : `LVFGDSTIDT`, `FNFGDSNSDT`, `IVFGDSIMDT`, `FVLGDSLVDA`, `FVFGDSLVDS`, `IVFGDSTVDS`, `FAFGDSLFEA`, `LIFGDSTVDT`, `FVFGDSMSDN`, `FVFGDSVFDN`, `YAFGDSFTDT`. U toj skupini se nalazi 118 motiva, mi smo ih naveli 11 od kojih u samo 3 možemo uočiti niz aminokiselina GDSL. Što potvrđuje prethodno rečeno.

Biološka lista pozitivaca

Za svaki proteom biljke, koju ćemo koristiti u istraživanju, imamo biološku listu pozitivaca. Listu određenu od strane biologa, koji su različitim metodama i analizama utvrdili koji točno proteini će se nalaziti na toj listi za koji proteom. Pozitivci na listi su imena proteina u kojima su smješteni nizovi aminokiselina koji su, zbog bioloških razloga, karakteristični za proteinsku familiju GDSL lipaza. To su upravo oni proteini pozitivnog stanja, koje smo prethodno označili kao CP. Svi oni koji nisu na toj listi su CN.

3.2 Blosum matrica i score

Ostajemo kod pojmova aminokiselina, ali ćemo vidjeti kako ćemo ih matematički, matrično "spremiti" i koristiti u kasnijem iterativnom pretraživanju. Definirajmo zato pojmove **Blosum matrica** i **Blosum score**. Skraćenica BLOSUM dolazi od engleskog naziva za tu matricu - *BLOcks SUbstitution Matrix*.

Definicija 3.2.1. *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 [4]). 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 $P(a_i, b_j \mid R) = P(a_i \mid R) \cdot P(b_j \mid R)$ *. Distribucija standardnih aminokiselina uz model R je dana sa:*

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

Definicija 3.2.2. *Blosum score s je rezultat koji odgovara sličnosti (ili povezanosti) dvaju nizova aminokiselina. Što je Blosum score veći, nizovi aminokiselina su sličniji.*

3.3 Iterativno pretraživanje. IGLOSS i BLAST.

Svako testiranje u ovom radu započinjemo iterativnim pretraživanjem pa je potrebno objasniti što je to uopće. Pretraga će imati svoj upit i odgovor na taj upit. Definirajmo prvo ta dva pojma, a zatim ćemo opisati kako se izvršavaju iteracije te pomoću kojih iterativnih pretraživača ćemo dobivati odgovor. Skupovi na kojem ćemo raditi pretrage bit će proteomi biljaka.

Upit i odgovor

Upit je "naš" pojam kojim nazivamo niz aminokiselina, obično duljine od 5 do 20, kojim pokrećemo pretraživanje. U našim testovima koristit ćemo upite duljine 10. Upit može sadržavati jedan ili više nizova istih duljina. Na zadani upit dobivamo preko iterativnih pretraživača nizove aminokiselina koje nazivamo odgovorom na dani upit. Za svaki je niz, koji se našao u odgovoru, temeljem funkcije sličnosti određeno da je dovoljno sličan upitu. Nizovi u odgovoru međusobno nisu povezani tj. prilikom dobivanja odgovora samo znamo da je svaki niz sličan upitu, no o njihovoj međusobnoj povezanosti ne znamo ništa i smatramo ih međusobno nezavisnima. Funkcija sličnosti je za svaki pretraživač drugačije definirana, zbog toga za isti upit, na istom skupu podataka, ali za dva različita pretraživača nećemo dobiti isti odgovor.

Iterativno pretraživanje

Riječ je o pretraživanju skupa podataka, na temelju nekog kriterija, koje se iterira, odnosno ponavlja sve do trenutka kad ponavljanje više nema smisla jer se rezultati pretrage ponavljaju. Naš skup podataka su proteomi biljaka koji sadrže veliku količinu podataka za ručnu pretragu i zato koristimo servere koji to čine u svega nekoliko minuta. Glavni pretraživač koji koristimo je **IGLOSS** (eng. *iterative gapless local similarity search*). Njegov zadatak je pronaći u proteomu sve nizove koji su slični danom upitu, a sličnost odredi preko svoje **funkcije sličnosti** koja se temelji na logističkoj distribuciji. Razinu sličnosti po kojoj želimo dobiti odgovor biramo sami tako da postavimo skalu na željenu razinu.

Skala je parametar logističke distribucije koji nam govori koliko puta ćemo se odmaknuti za paramerat β od prosječne ocjene sličnosti. Ona utječe na rezultat odgovora na način da se sve ocjene veće od vrijednosti ($\mu + skala * \beta$) proglase značajnima i s njima se kreće u drugi krug iteracija. Jasno je da što je skala manja više ocjena će biti veće od ($\mu + skala * \beta$) dakle više nizova će upasti u odgovor. Ukratko, što veću sličnost želimo između upita i odgovora to veću skalu moramo postaviti i obratno. Pokažimo to na primjeru proteoma biljke talijin uročnjak, za upit= FVFGDSLSDA.

Skala	Duljina odgovora
6	223
10	76

Tablica 3.2: Duljina odgovora ovisno o skali

Kada postavimo upit i odredimo skalu server pokreće pretragu. U prvom krugu iterativnog pretraživanja računat će se sličnost upita sa svakim nizom te duljine u proteomu. Svi oni nizovi čije ocjene se proglase značajnima idu u drugi krug iterativnog pretraživanja i proglašavamo ih pozitivnima. U svakom krugu se parametri funkcije sličnosti mijenjaju jer svaki idući krug ovisi o prethodnom. Iteriranje staje kada više nema promjena u popisu pozitivaca ili kada je postignut unaprijed određen broj ponavljanja. Kad iteriranje stane mi dobivamo nizove aminokiselina koji predstavljaju odgovor na naš upit.

Još jedan od iterativnih pretraživača koji koristimo je **BLAST** (eng. *basic local alignment search tool*). Proces pretrage na njemu je sličan kao na IGLOSS-u, osim što on ima svoju funkciju sličnosti. Još jedna razlika između BLAST-a i IGLOSSS-a je ta što su na BLAST-u maksimalne ocjene sličnosti Gumbel distribuirane, dok su na IGLOSS-u logistički distribuirane.

Podaci iz ovog poglavlja preuzeti su iz izvora [2] i [1].

Poglavlje 4

Metoda centralnog motiva

U prethodna tri poglavlja smo postavili temelje za razumijevanje rada. Sada možemo objasniti metodu koju koristimo u radu, na primjeru pokazati kako se metoda provodi i na kraju objasniti u koje svrhe tj. s kojim ciljem ju provodimo. Podaci na kojima provodimo istraživanje su veliki i teško bi bilo koristeći njih objasniti metodu zato smo uzeli jedan primjer kojim ćemo ilustrirati svaki korak, a sve što navedemo će vrijediti i na našim, stvarnim, podacima.

Objašnjenje metode. Primjer.

Neka je $X = (X_n : n \geq 0)$ Markovljev lanac, odnosno niz nezavisnih, jednakodistribuiranih slučajnih varijabli s vrijednostima u skupu $A = \{1, 2, 3, 4, 5, 6, 7\}$. Skup stanja čini 7 motiva od kojih je svaki niz duljine 8. Označimo ih:

- MALAFGHL = 1
- MAVAWAWL = 2
- MALAWAKA = 3
- MVLAFAWA = 4
- KLQDFGHI = 5
- KLQEFGHL = 6
- KVQDFGHL = 7

Vjerojatnosna distribucija Markovljevog lanca dana je sa $\pi = (\frac{1}{7}, \frac{1}{7}, \frac{1}{7}, \frac{1}{7}, \frac{1}{7}, \frac{1}{7}, \frac{1}{7})$. Za početak, formiramo score matricu S ovog Markovljevog lanca. Retke i stupce matrice čini tih 7 elemenata skupa A pa ćemo dobiti matricu dimenzije 7x7. Neka je $S_7 \in M_7$, tada će element

matrice $s_{ij} \in R$, gdje su i,j=1,2,...,7, sadržavati iznos score-a između niza u i-tom retku te niza u j-tom stupcu, odnosno taj element će nam reći koliko su i-ti i j-ti motiv međusobno povezani, što je veći broj to su motivi povezaniji. Za računanje score-a koristimo Blosum matricu u kojoj su u stupcima i recima poredane redom aminokiseline, njih 20. Kako se računa vrijednost score-a pokažimo na dva motiva našeg primjera. Zapišemo motive jedan ispod drugoga:

MAVAWAWL
MALAWAKA

Motivi su jednakih duljina te možemo računati score. Za svake dvije aminokiseline koje su jedna ispod druge očitamo iz Blosum matrice vrijednost povezanosti te dvije aminokiseline i to učinimo za svake dvije aminokiseline u motivima te na kraju zbrojimo dobivene vrijednosti. To izgleda ovako:

$score = B(M,M) + B(A,A) + B(V,L) + B(A,A) + B(W,W) + B(A,A) + B(W,K) + B(L,A) = B(12,12) + B(0,0) + B(19,10) + B(0,0) + B(17,17) + B(0,0) + B(17,11) + B(10,0) = 7 + 5 + 1 + 5 + 15 + 5 - 3 - 2 = 33$

Sada znamo da će se u score matrici na pozicijama S_{23} i S_{32} nalaziti vrijednost 33. Da je matrica simetrična vrlo je jasno. Na isti način izračunamo ostale vrijednosti i formiramo matricu :

$$S = \begin{bmatrix} 53 & 21 & 21 & 20 & 20 & 24 & 25 \\ 21 & 62 & 33 & 32 & -9 & -5 & -4 \\ 21 & 33 & 53 & 25 & -8 & -8 & -7 \\ 20 & 32 & 25 & 55 & -1 & -1 & 2 \\ 20 & -9 & -8 & -1 & 57 & 48 & 50 \\ 24 & -5 & -8 & -1 & 48 & 55 & 47 \\ 25 & -4 & -7 & 2 & 50 & 47 & 57 \end{bmatrix}$$

Uočavamo da u matrici postoje i negativne vrijednosti što nam govori da ti motivi međusobno nisu povezani. Isto tako možemo uočiti da su prva 4 motiva međusobno dobro povezana te prvi i zadnja 3 da su isto tako međusobno dobro povezani. Oni koji su nepovezani jer imaju negativne vrijednosti će ostati nepovezani i kada te vrijednosti postavimo na 0. Negativnih vrijednosti smo se još mogli riješiti na način eksponenciranja svih elemenata, no odlučili smo se za ovaj način. Bitno nam je zbog idućeg koraka, normalizacije matrice, izbjeći negativne elemente. Stoga, sljedeće što radimo na putu za pronalazak centralnog motiva je normalizacija matrice S. Matricu normaliziramo tako da za svaki redak sumiramo elemente u tom retku i svaki element tog retka podijelimo sa sumom koju smo dobili. Na taj način dobivamo matricu kojoj je suma svakog retka 1 i kojoj su svi elementi $s_{ij} \geq 0$,

drugim riječima, upravo smo formirali stohastičku matricu, što ne bi bilo moguće da nismo eliminirali negativne elemente matrice.
Napomena. S obzirom da smo provodili normalizaciju po recima mi smo formirali matricu koja nakon transponiranja postaje stohatička.

Normalizirana matrica :

$$P = \begin{bmatrix} 0.288043 & 0.11413 & 0.11413 & 0.108696 & 0.108696 & 0.130435 & 0.13587 \\ 0.141892 & 0.418919 & 0.222973 & 0.216216 & 0. & 0. & 0. \\ 0.159091 & 0.25 & 0.401515 & 0.189394 & 0. & 0. & 0. \\ 0.149254 & 0.238806 & 0.186567 & 0.410448 & 0. & 0. & 0.014925 \\ 0.114286 & 0. & 0. & 0. & 0.325714 & 0.274286 & 0.285714 \\ 0.137931 & 0. & 0. & 0. & 0.275862 & 0.316092 & 0.270115 \\ 0.138122 & 0. & 0. & 0.01105 & 0.276243 & 0.259669 & 0.314917 \end{bmatrix}$$

Neka je to prijelazna matrica P našeg Markovljevog lanca, gdje element p_{ij} predstavlja vjerojatnost da iz mjesta i dođemo u mjesto j. Lako je za uočiti da je u svakom retku najveća vrijednost na mjestu p_{ii}, a elementi na pozicijama gdje i-ti i j-ti motiv nisu povezani su 0 tj. nikako ne možemo od mjesta i doći u mjesto j u jednom koraku. To uočavamo iz prijelazne matrice, a isto možemo uočiti i iz samog pogleda na motive, vidimo da se prva 4 motiva po svojim aminokiselinama znatno razlikuju od preostala tri te isto tako da su sva 4 međusobno slična. Drugim riječima, lanac nije ireducibilan jer postoji više klasa komuniciranja, a mi želimo imati samo jednu klasu. Još uvijek ne možemo ništa reći o centralnom motivu. Zato preostaje potenciranje prijelazne matrice i uočavanje centralnog motiva. Prema teoremu o konvergenciji prema graničnoj distribuciji znamo da moramo uzeti dovoljno veliki n za potenciju kako bi zaključili o kojoj graničnoj distribuciji je riječ. Dovoljno veliki n tražimo vrlo "opipljivim" putem. Nakon što smo matricu digli na 10. potenciju i na 100. potenciju uočili smo da rezultati nisu baš isti, no kada smo digli na 200. potenciju dobili smo matricu gotovo istu kao za potenciju 100 te iz tog razloga naš n=100 predstavlja dovoljno veliki n, odnosno kada iz teorema iskoristimo $n \to \infty$ uzimamo podatke za n=100. Pogledajmo koji su to elementi matrice:

$$P^{(100)} = \begin{bmatrix} 0.163121 & 0.131206 & 0.117021 & 0.118794 & 0.155142 & 0.154255 & 0.160461 \\ 0.163121 & 0.131206 & 0.117021 & 0.118794 & 0.155142 & 0.154255 & 0.160461 \\ 0.163121 & 0.131206 & 0.117021 & 0.118794 & 0.155142 & 0.154255 & 0.160461 \\ 0.163121 & 0.131206 & 0.117021 & 0.118794 & 0.155142 & 0.154255 & 0.160461 \\ 0.163121 & 0.131206 & 0.117021 & 0.118794 & 0.155142 & 0.154255 & 0.160461 \\ 0.163121 & 0.131206 & 0.117021 & 0.118794 & 0.155142 & 0.154255 & 0.160461 \\ 0.163121 & 0.131206 & 0.117021 & 0.118794 & 0.155142 & 0.154255 & 0.160461 \end{bmatrix}$$

Lako se vidi da je

$$\lambda = (0.163121, 0.131206, 0.117021, 0.118794, 0.155142, 0.154255, 0.160461) \tag{4.1}$$

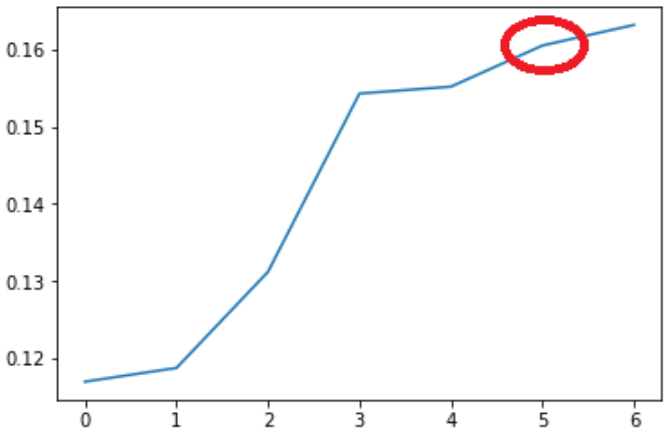
stacionarna distribucija. Skup A je konačan i vrijedi

$$\lim_{n\to\infty}(p_{12})^n = \lambda_2 = 0.131206 \tag{4.2}$$

$$\lim_{n\to\infty}(p_{42})^n = \lambda_2 = 0.131206 \tag{4.3}$$

Navedeno vrijedi za svaki $j = 1, 2, ..., 7$.
Da je stacionarna distribucija ujedno i granična slijedi iz Teorema 2.0.14 jer su sve pretpostavke teorema zadovoljene - postojanje Markovljevog lanca koji je ireducibilan i aperiodičan te postojanje njegove stacionarne distribucije.

Konačno, kada smo dobili graničnu distribuciju možemo odrediti centralni motiv, onaj motiv za koji granična distribucija postiže najveću vrijednost. Vidimo da je to prvi motiv, no isto tako vidimo da vrlo blizu njemu je i sedmi motiv. S obzirom da ne možemo odrediti značenje izraza "vrlo blizu" teško je u ovoj situaciji odrediti hoće li centralni biti samo prvi motiv ili ćemo proglasiti 1. i 7. motiv centralnima. Jedini način za odluku je provođenje testa za oba slučaja i utvrđivanje koji daje bolje rezultate. Pogledajmo grafički prikaz graničnih vrijednosti za 7 motiva:



Slika 4.1: Grafički prikaz graničnih vrijednosti

Grafički prikaz nam može pomoći pri odabiru centralnog motiva tako da za centralni/e uzmemo one koji se nalaze u rastu nakon zadnjeg "skoka".

Cilj metode

Na proteomima različitih biljaka, sa različitim upitima provodit ćemo testove. Svaki upit je zadan sa određenim razlogom. Nakon iterativne pretrage, objašnjene u [3], za svaki upit dobivamo odgovor (P) te za njega izračunamo sve statističke mjere. Dakle, uz biološku listu pozitivaca koju imamo za svaki proteom i odgovor svakog upita izračunamo TP, PPV, TPR te iz tih mjera i F1 score. Ako F1 score nije jednak 1, a naravno da neće biti, prostora za poboljšanje njegove vrijednosti ima i zato na dobivenom odgovoru krećemo sa metodom centralnog motiva. Sad je jasno da je glavni cilj metode poboljšati odgovor, drugim riječima povećati F1 score. "Poboljšati odgovor" možemo još protumačiti kao dobiti novi odgovor u kojem će se naći više nizova sa CP liste, a manje onih koji nisu na toj listi. Da bi dobili novi odgovor prvo moramo pronaći novi upit, a novi upit će biti naš centralni motiv ili više njih. S tim upitom ponovo provedemo iterativno pretraživanje. Na dobivenom odgovoru ponovimo postupak traženja statističkih mjera. Uz tablice sa rezultatima F1 score-a i grafičkih prikaza vrlo jednostavno ćemo zaključiti je li se metoda pokazala uspješnom ili ne.

Radi provjere stabilnosti metode na par upita smo izvršili nekoliko iteracija pretrage centralnog motiva kako bi vidjeli hoće li se oni ponavljati ili će svaki puta otići u drugom smjeru. Naša intuitivna pretpostavka je da će centralni motiv u odgovoru centralnog motiva dati iste ili većinom iste nizove kao i početni centralni motiv te da ćemo nakon već jedne ili dvije iteracije biti na početnom odgovoru. Kroz provjeru to se pokazalo dobrom pretpostavkom. U nekoliko slučajeva se čak dogodilo da su centralni motivi u odgovoru centralnog motiva bili upravo oni sami. Metoda ima dobru stabilnost.

Poglavlje 5

Rezultati

Testiranje provodimo na proteomima četiri kopnene biljke :

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

za tri različita upita :

- FVFGDSLSDA
- FVFNSDLSDA
- VFFGDSLSDN

uz pomoć dva iterativna pretraživača:

- IGLOSS
- BLAST

5.1 Proteomi biljaka

Navedene četiri biljke su jedne od modelnih organizama za razna genomska istraživanja koja su vezana za otkrivanje novih GDSL lipaza. Neke zbog biološke strukture, a neke zbog ekonomske važnosti. Za svaku od biljaka imamo, već spomenutu, listu bioloških pozitivaca (CP). Od navednih, najpogodnija za biološka istraživanja je ***Arabidopsis thaliana***, biljka

male veličine, sa malim genomom koji je potpuno sekvencioniran te se za svaki protein u njenom proteomu skoro pa zna kojoj proteinskoj familiji pripada. Lista CP proteoma ove biljke sadrži 104 različita proteina, a zajedno sa izoformima ima ih 118. Mi ćemo uvijek gledati liste sa izoformima. Izoformi su različite forme istog proteina.

Ostale tri biljke su također često korištene u genomskim istraživanjima. Zbog prehrambene važnosti tih biljaka širom svijeta, od velikog ekonomskog interesa bi bilo otkriti što više bioloških funkcija njihovih gena. Isto tako i otkriti nove GDSL lipaze zbog njihove funkcije. Primjerice, za ***Oryza sativu*** poznata su svega dva gena GDSL lipaza. Lista CP ove biljke sadrži 155 proteina dok liste CP za ***Solanum tuberosum*** i ***Solanum lycopersicum*** sadrže redom 123 i 108 proteina.

5.2 Upiti

Motivom konsenzusom smatramo upit **FVFGDSLSDA**. Za taj motiv je otkriveno da je on najviše sličan onima koji se nalaze na listi bioloških pozitivaca. Trenutno se njega smatra karakterističnim za GDSL lipaze. Uočimo kako taj upit u sebi sadrži niz aminokiselina GDSL. Pretpostavka nam je da će F1 score za ovaj upit biti veći nego za druga dva, malo "kriva", upita. Drugi upit **FVFNDSLSDA** se razlikuje od motiva konsenzusa po središnjem djelu jer ne sadrži niz GDSL, ali okolne aminokiseline mu se podudaraju sa motivom konsenzusom. Treći upit **VFFGDSLSDN** sadrži niz GDSL, ali su mu okolne aminokiseline drugačije od onih kod motiva konsenzusa. Zbog toga, odgovor će povući dosta FP motiva čime će se F1 score smanjiti u odnosu na motiv konsenzus.

Testiranje možemo podijeliti u dvije skupine. Prva skupina će biti iterativno pretraživanje svakog upita uz pomoć BLAST-a, a druga skupina uz pomoć IGLOSS-a. No, pretraživanje sa centralnim upitima u obje skupine radimo preko IGLOSS-a.

5.3 Hipoteze testa

Svako statističko testiranje treba imati postavljene hipoteze, kako smo i definirali u poglavlju 1. Hipoteze našeg testa su :

$$H_0 : \text{Proteini iz odgovora ne pripadaju proteinskoj familiji}$$
$$H_1 : \text{Proteini iz odgovora pripadaju proteinskoj familiji}$$

Za svaki test koji provedemo računat ćemo F1 score te ako je njegova vrijednost dovoljno velika odbacit ćemo nultu hipotezu, u suprotnom, nećemo odbaciti. Problem nastaje kod pojma "dovoljno velika" jer nemamo i ne možemo odrediti točnu razinu značajnosti primjerenu za ovakvu vrstu testa. Bez razine značajnosti ne možemo ni donositi odluke o odabiru hipoteze.

5.4 Kratka analiza rezultata - tablično i grafički

BLAST. Upit = FVFGDSLSDA.

1. **Talijin uročnjak**

CP	Duljina odgovora - P	TP	PPV	TPR	F1-score
118	106	75	0.71	0.64	0.67

Tablica 5.1: Rezultat za *upit = FVFGDSLSDA*

Uzimamo 106 pozitivaca, provodimo postupak traženja centralnih motiva te se u ovom slučaju odlučujemo za jedan centralni. Pomoću IGLOSS-a tražimo odgovor na upit centralnog motiva i dobivamo sljedeće rezultate:

CP	Duljina odgovora - P	TP	PPV	TPR	F1-score
118	101	84	0.83	0.71	0.77

Tablica 5.2: Rezultat za *centralni motiv*

Uspoređujući rezultate F1 score-a za upit i centralni motiv, dolazmo do zaključka da je metoda centralnog motiva uspješna. Iterativnom pretragom sa centralnim motivom poboljšali smo odgovor tj. u tom odgovoru smo dobili više onih koji se nalaze u CP, a manje onih koji se ne nalaze u CP. Pogrešku prve vrste smo smanjili za 31% što je jako dobro poboljšanje.

2. **Azijska riža**

CP	Duljina odgovora - P	TP	PPV	TPR	F1-score
155	117	87	0.74	0.56	0.64

Tablica 5.3: Rezultat za *upit = FVFGDSLSDA*

Uzimamo 117 pozitivaca, provodimo postupak traženja centralnih motiva te se odlučujemo za tri centralna motiva jer njihove granične vrijednosti odudaraju od ostalih. Pomoću IGLOSS-a tražimo odgovor na upit sastavljen od sva tri centralna motiva i dobivamo sljedeće rezultate:

CP	Duljina odgovora - P	TP	PPV	TPR	F1-score
155	131	119	0.91	0.77	0.83

Tablica 5.4: Rezultat za *centralne motive*

Uspoređujući rezultate F1 score-a za upit i centralne motive, dolazmo do zaključka da je metoda centralnog motiva uspješna. Iterativnom pretragom sa centralnim motivima poboljšali smo odgovor. U odgovoru centralnih motiva smo dobili 14 pozitivaca više pa je bila veća vjerojatnost da će se nalaziti i više onih koji su u CP, ali mi smo dobili čak 32 više onih koji su u CP i na taj način povećali i PPV i TPR, a time i F1 score. Pogrešku prve vrste smo smanjili za 65% što je izvrsno poboljšanje.

3. **Rajčica**

CP	Duljina odgovora - P	TP	PPV	TPR	F1-score
108	109	77	0.71	0.71	0.71

Tablica 5.5: Rezultat za *upit = FVFGDSLSDA*

Uzimamo 108 pozitivaca, provodimo postupak traženja centralnih motiva. Određujemo centralni motiv i pretragom dobivamo sljedeće:

CP	Duljina odgovora - P	TP	PPV	TPR	F1-score
108	100	91	0.91	0.84	0.87

Tablica 5.6: Rezultat za *centralni motiv*

Metoda centralnog motiva i u ovom slučaju daje veliko poboljšanje odgovora. Ovdje je još zanimljivo uočiti da iako se pretragom sa centralnim motivom dobila manja lista pozitivaca, na njoj se našlo više TP elemenata. Iako bi za očekivati bilo da TPR padne u slučaju manje pozitivaca, TPR je znatno povećan i F1 score je veći.

4. **Krumpir**

Pokažimo još rezultate za krumpir i uočimo kako je i na tom organizmu došlo do velikog skoka F1 score-a .

CP	Duljina odgovora - P	TP	PPV	TPR	F1-score
123	123	76	0.62	0.62	0.62

Tablica 5.7: Rezultat za *upit = FVFGDSLSDA*

CP	Duljina odgovora - P	TP	PPV	TPR	F1-score
123	106	91	0.86	0.74	0.8

Tablica 5.8: Rezultat za *centralni motiv*

IGLOSS. Upit = FVFGDSLSDA.

1. **Talijin uročnjak**

CP	Duljina odgovora - P	TP	PPV	TPR	F1-score
118	105	87	0.83	0.74	0.78

Tablica 5.9: Rezultat za *upit = FVFGDSLSDA*

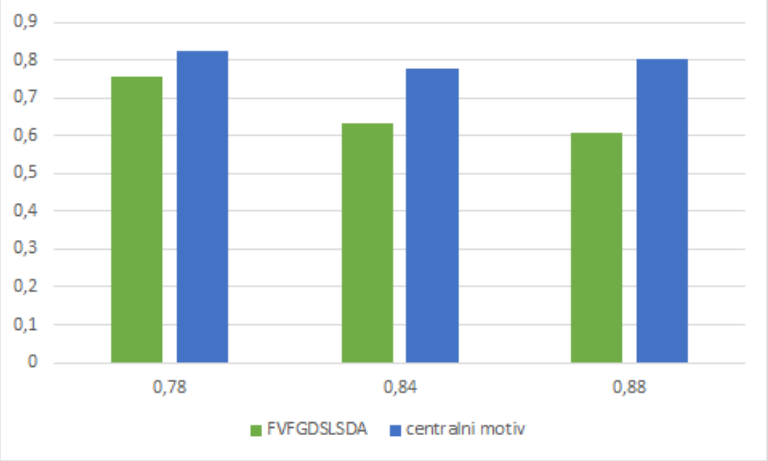
Uzimamo 105 pozitivaca, provodimo postupak traženja centralnih motiva te uzimamo jedan centralni motiv koji zbog vrijednosti granične distribucije malo odskače od preostalih. Pomoću IGLOSS-a tražimo odgovor na upit centralnog motiva i dobivamo sljedeće rezultate:

CP	Duljina odgovora - P	TP	PPV	TPR	F1-score
118	130	105	0.81	0.89	0.85

Tablica 5.10: Rezultat za *centralni motiv*

Uspoređujući rezultate F1 score-a za upit i centralni motiv, dolazmo do zaključka da je metoda centralnog motiva uspješna. Uočavamo kako se sa centralnim motivom povećao broj pozitivaca u odgovoru te se PPV vrlo malo smanjio, ali smo pokupili gotovo 90% elemenata sa CP liste i time se F1 score jako povećao. Pogrešku prve vrste smo smanjili za 31% što je jako dobro poboljšanje .

Za ovaj primjer pokazujemo i usporedni graf F1-scora. Graf na x - osi ima tri različite PPV vrijednosti, a na y-osi pripadne F1 score-ove upita i centralnog motiva.



Slika 5.1: Usporedni graf F1 score-a - Talijin uročnjak

Uočavamo da za isti PPV u sva tri slučaja dobivamo veći F1 score kod centralnog motiva. Moramo primijetiti i kako je F1 score dosta velik i za početni upit pa se i minimalno poboljšanje smatra uspjehom.

2. **Krumpir**

CP	Duljina odgovora - P	TP	PPV	TPR	F1-score
123	119	91	0.76	0.74	0.75

Tablica 5.11: Rezultat za *upit = FVFGDSLSDA*

Uzimamo 119 pozitivaca, provodimo postupak traženja centralnih motiva te se u ovom slučaju odlučujemo se za više centralnih motiva zbog vrijednosti njihovih graničnih distribucija koje su približno jednake. Pomoću IGLOSS-a tražimo odgovor na upit koji se sastoji od svih centralnih motiva i dobivamo sljedeće rezultate:

CP	Duljina odgovora - P	TP	PPV	TPR	F1-score
123	100	87	0.87	0.71	0.78

Tablica 5.12: Rezultat za *centralne motive*

Pretragom proteoma sa centralnim motivima dobili smo broj pozitivaca za 19 manji od pozitivaca upita, ali se broj TP smanjio samo za 4 i zbog toga je PPV porastao jer je čak 87% motiva iz P-centralnog i u CP. Zbog toga se TPR malo smanjio. Zbog većeg rasta PPV-a od smanjenja TPR-a imamo nešto veći F1 score pa je i ovdje metoda uspješna.

3. **Rajčica**

Rezultati kod rajčice su slični rezultatima kod krumpira i također se odlučujemo za više centralnih motiva jer time dobivamo nešto bolji F1 score nego kod jednog centralnog motiva, a u svakom slučaju veći nego kod početnog upita.

CP	Duljina odgovora - P	TP	PPV	TPR	F1-score
108	111	91	0.82	0.84	0.83

Tablica 5.13: Rezultat za *upit = FVFGDSLSDA*

CP	Duljina odgovora - P	TP	PPV	TPR	F1-score
108	97	90	0.93	0.83	0.88

Tablica 5.14: Rezultat za *centralne motive*

Pogrešku prve vrste smo smanjili za oko 30%

4. **Azijska riža**

CP	Duljina odgovora - P	TP	PPV	TPR	F1-score
155	118	112	0.95	0.72	0.82

Tablica 5.15: Rezultat za *upit = FVFGDSLSDA*

CP	Duljina odgovora - P	TP	PPV	TPR	F1-score
155	124	112	0.90	0.72	0.8

Tablica 5.16: Rezultat za *centralne motive*

Rezultati pokazuju da smo nakon pretrage sa upitom dobili odgovor koji ima PPV vrijednost čak 0.95 i to nismo uspjeli poboljšati sa centralnim motivima i zbog toga je došlo do vrlo malog pada F1 score-a.

BLAST. Upit = FVFNDSLSDA.

Za ovaj malo krivi upit uočimo kako su vrijednosti i PPV-a i TPR-a znatno niže od onih kod karakterističnog upita, samim tim pronalaskom centralnog motiva lakše će biti poboljšati odgovor tj. povećati rezultat F1 score-a.

1. **Talijin uročnjak**

Iterativnom pretragom preko BLAST-a za ovaj upit dobili smo odgovor duljine 53, od čega je samo 34 niza unutar 118 CP-a što objašnjava vrlo mali F1 score.

CP	Duljina odgovora - P	TP	PPV	TPR	F1-score
118	53	34	0.64	0.29	0.40

Tablica 5.17: Rezultat za *upit = FVFNDSLSDA*

U 53 pozitivca našli smo više centralnih motiva te nakon pokretanja pretrage s njima, rezultati su sljedeći:

CP	Duljina odgovora - P	TP	PPV	TPR	F1-score
118	52	46	0.88	0.39	0.54

Tablica 5.18: Rezultat za *centralne motive*

Za gotovo jednaku veličinu odogovora dobili smo znatno veći PPV, a time i F1 score.

2. **Krumpir**

CP	Duljina odgovora - P	TP	PPV	TPR	F1-score
123	81	48	0.59	0.39	0.47

Tablica 5.19: Rezultat za *upit = FVFNDSLSDA*

BLAST nam za krumpir daje 81 pozitivca, ali i F1 score kao i kod talijinog uročnjaka manji od 0.5 pa zadatak poboljšavanja toga i nije toliko težak. Nakon pretrage sa centralnim motivima rezultati su puno bolji :

CP	Duljina odgovora - P	TP	PPV	TPR	F1-score
123	90	80	0.89	0.65	0.75

Tablica 5.20: Rezultat za *centralni motivi*

Uz veliki skok F1 score-a uspjeli smo i pogrešku prve vrste smanjiti za čak 73%.

3. **Rajčica**

Kod ove biljke, nakon pretrage preko BLAST-a situacija je slična kao i kod krumpira.

CP	Duljina odgovora - P	TP	PPV	TPR	F1-score
108	64	41	0.64	0.38	0.48

Tablica 5.21: Rezultat za *upit = FVFNDSLSDA*

Ono što je ovdje zanimljivo za pokazati je kako skala i odluka hoćemo li uzeti jedan ili više centralnih motiva utječu na odgovor. Prvo imamo primjer gdje smo uzeli jedan centralni motiv te veću skalu. U tom slučaju nam je PPV gotovo 1, ali TPR nije previsok, no u svakom slučaju F1 score je jako skočio.

CP	Duljina odgovora - P	TP	PPV	TPR	F1-score
108	64	63	0.98	0.58	0.73

Tablica 5.22: Rezultat za *centralni motiv*

Druga situacija je kada smo uzeli više centralnih motiva te manju skalu. U tom slučaju duljina odgovora se jako povećala, PPV je ostao dosta visok, no nije približno 1, ali je TPR vrlo visok, za više od 50% veći nego kod upita i time je F1 score znatno bolji nego kod upita, ali i bolji nego kod jednog centralnog motiva sa većom skalom.

CP	Duljina odgovora - P	TP	PPV	TPR	F1-score
108	106	92	0.87	0.85	0.86

Tablica 5.23: Rezultat za *centralne motive*

4. **Azijska riža**

CP	Duljina odgovora - P	TP	PPV	TPR	F1-score
155	70	47	0.7	0.3	0.41

Tablica 5.24: Rezultat za *upit = FVFNDSLSDA*

Kako vidimo, lista pozitivaca je mala obzirom na CP listu. Iz grafičkog prikaza uočili smo da su dva motiva centralna. Nakon što smo s njima pokrenuli pretragu PPV je skočio na 0.95 što je u dosadašnjim rezultatima drugi najbolji PPV . Duljina odgovora nam nije niti blizu jednaka duljini CP liste i zato TPR ne može biti visok, ali smo F1 score jako poboljšali što se vidi iz sljedeće tablice :

CP	Duljina odgovora - P	TP	PPV	TPR	F1-score
155	111	105	0.95	0.68	0.79

Tablica 5.25: Rezultat za *centralne motive*

IGLOSS. Upit = FVFNDSLSDA.

Sve kao i na BLAST-u za ovaj upit napravili smo i na IGLOSS-u. Kod talijinog uročnjaka, rajčice i azijske riže dolazi do dobrog poboljšanja dok kod krumpira je to poboljšanje vrlo malo, ali ipak postoji.

Pogledajmo rezultate:

1. **Talijin uročnjak**

CP	Duljina odgovora - P	TP	PPV	TPR	F1-score
118	87	66	0.76	0.56	0.64

Tablica 5.26: Rezultat za *upit = FVFNDSLSDA*

CP	Duljina odgovora - P	TP	PPV	TPR	F1-score
118	98	87	0.89	0.74	0.81

Tablica 5.27: Rezultat za *centralne motive*

2. **Krumpir**

CP	Duljina odgovora - P	TP	PPV	TPR	F1-score
123	108	84	0.78	0.68	0.73

Tablica 5.28: Rezultat za *upit = FVFNDSLSDA*

CP	Duljina odgovora - P	TP	PPV	TPR	F1-score
123	101	83	0.82	0.67	0.74

Tablica 5.29: Rezultat za *centralne motive*

3. **Rajčica**

CP	Duljina odgovora - P	TP	PPV	TPR	F1-score
108	107	86	0.8	0.8	0.8

Tablica 5.30: Rezultat za *upit = FVFNDSLSDA*

CP	Duljina odgovora - P	TP	PPV	TPR	F1-score
108	105	95	0.9	0.88	0.89

Tablica 5.31: Rezultat za *centralne motive*

4. **Azijska riža**

CP	Duljina odgovora - P	TP	PPV	TPR	F1-score
155	122	102	0.84	0.66	0.74

Tablica 5.32: Rezultat za *upit = FVFNDSLSDA*

CP	Duljina odgovora - P	TP	PPV	TPR	F1-score
155	132	117	0.89	0.75	0.81

Tablica 5.33: Rezultat za *centralne motive*

BLAST. Upit = VFFGDSLSDN.

Treći upit za koji provodimo istraživanje je onaj koji se za najviše aminokiselina razlikuje od konsenzusa. Kako smo i pretpostavili, vidjet ćemo da je F1 score uvjerljivo najmanji od svega dosad te da je razlika tog i F1 score-a centralnog motiva najveća. Pogledajmo rezultate:

1. **Talijin uročnjak**

CP	Duljina odgovora - P	TP	PPV	TPR	F1-score
118	107	32	0.30	0.27	0.28

Tablica 5.34: Rezultat za *upit = VFFGDSLSDN*

U odogovoru od 107 nizova su se našla samo 32 TP zato što je pretraživač povukao puno onih nizova koji su slični upitu po okolnim aminokiselinama, a takvi nizovi se znatno razlikuju od onih na listi CP. Uočimo, centralni motiv je neki od ta 32 koji je sličan upitu po centralnom djelu GDSL. S njim kad provedemo pretragu dobivamo:

CP	Duljina odgovora - P	TP	PPV	TPR	F1-score
118	104	88	0.85	0.75	0.80

Tablica 5.35: Rezultat za *centralni motiv*

F1 score se skoro tri puta povećao, a pogreška prve vrste se smanjila za gotovo 80% što je uvjerljivo najbolje poboljšanje dosad.

2. **Krumpir**

Kod krumpira je slična situacija.

CP	Duljina odgovora - P	TP	PPV	TPR	F1-score
123	147	34	0.23	0.28	0.25

Tablica 5.36: Rezultat za *upit = VFFGDSLSDN*

CP	Duljina odgovora - P	TP	PPV	TPR	F1-score
123	191	77	0.85	0.63	0.72

Tablica 5.37: Rezultat za *centralni motiv*

U odgovoru upita sa centralnim motivom pogreška prve vrste se smanjila više nego kod talijinog uročnjaka, ali je ovdje duljina odgovora puno veća kod centralnog nego kod početnog upita pa poboljšanje kod uročnjaka smatramo značajnijim.

3. **Rajčica**

CP	Duljina odgovora - P	TP	PPV	TPR	F1-score
108	114	33	0.29	0.31	0.30

Tablica 5.38: Rezultat za *upit = VFFGDSLSDN*

CP	Duljina odgovora - P	TP	PPV	TPR	F1-score
108	93	81	0.87	0.75	0.81

Tablica 5.39: Rezultat za *centralni motiv*

Kod rajčice pretragom sa centralnim motivom za dani upit se događa poboljšanje bolje nego kod talijinog uročnjaka. Uočimo da smo duljinu odgovora smanjili za 21 niz, a broj TP se povećao za 48 nizova . F1 score je znatno veći, a ono najzanimljivije, pogreška prve vrste se smanjila za 82%.

4. **Azijska riža**

Testiranje upita na azijskoj riži daje isto odlične rezultate. Uočimo veliko poboljšanje score-a, ali i skoka PPV vrijednosti sa 0.44 na 0.93, što je od velikog značaja.

CP	Duljina odgovora - P	TP	PPV	TPR	F1-score
155	107	47	0.44	0.30	0.36

Tablica 5.40: Rezultat za *upit = VFFGDSLSDN*

CP	Duljina odgovora - P	TP	PPV	TPR	F1-score
155	118	110	0.93	0.71	0.81

Tablica 5.41: Rezultat za *centralni motiv*

IGLOSS. Upit = VFFGDSLSDN.

Sa istim upitom provodimo testiranje sa IGLOSS-om. Iako za upit, za svaku biljku IGLOSS daje bolji F1 score od BLAST-a, rezultati sa centralnim motivom će opet pokazati veliko poboljšanje.

1. **Talijin uročnjak**

CP	Duljina odgovora - P	TP	PPV	TPR	F1-score
118	94	41	0.45	0.36	0.49

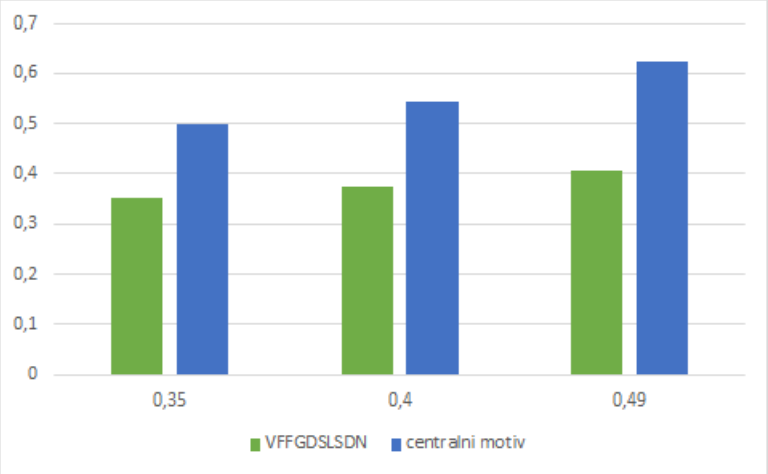
Tablica 5.42: Rezultat za *upit = VFFGDSLSDN*

CP	Duljina odgovora - P	TP	PPV	TPR	F1-score
118	98	87	0.89	0.74	0.81

Tablica 5.43: Rezultat za *centralne motive*

U početnom odgovoru uzeli smo više centralnih motiva te nakon pretrage s njima F1 score je sa 0.49 skočio na 0.81, ali se i PPV znatno povećao. Pogledajmo sada iz usporednog grafa što se događa sa F1 score-om za iste PPV vrijednosti pretrage s upitom i sa

centralnim motivima.



Slika 5.2: Usporedni graf F1 score-a - Talijin uročnjak

Za isti PPV razlike u F1 score-u nisu, kao u tablici, u vrijednosti od 0.30, ali su sve veće od 0.12 što ukazuje na jako dobro poboljšanje odgovora. Možemo još uočiti da što je veća vrijednost PPV-a to je i veće poboljšanje F1 score-a pa tako za PPV=0.49 imamo porast F1 score-a za 0.22.

2. **Krumpir**

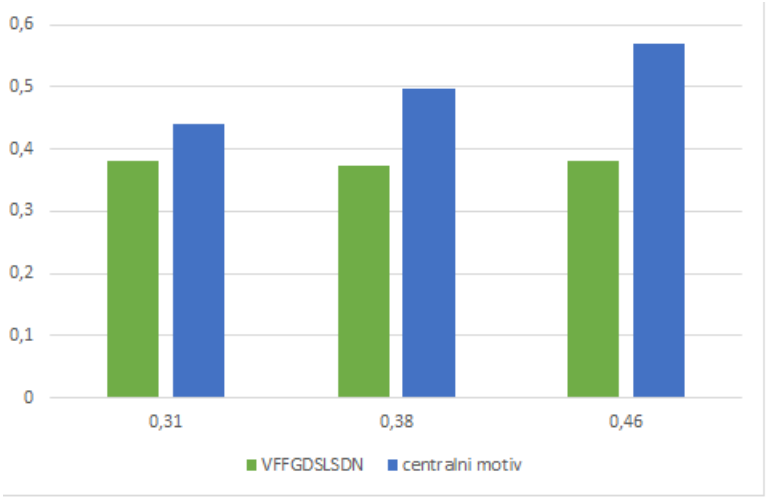
CP	Duljina odgovora - P	TP	PPV	TPR	F1-score
123	96	41	0.43	0.33	0.37

Tablica 5.44: Rezultat za *upit = VFFGDSLSDN*

CP	Duljina odgovora - P	TP	PPV	TPR	F1-score
123	84	76	0.9	0.62	0.73

Tablica 5.45: Rezultat za *centralni motiv*

Za krumpir isto imamo usporedni graf F1 score-ova kako bi vidjeli da metoda ima neku vrstu stabilnosti i po pitanju različitih biljaka. Poboljšanja su čak približno jednaka kao i kod talijinog uročnjaka. Pogledajmo iz grafa.



Slika 5.3: Usporedni graf F1 score-a - Krumpir

Pogledajmo sada rezultate prije i nakon metode za ostala dva organizma. Rezultati idu u istom smjeru kao i za prethodna dva organizma pa nema potrebe za dodatnim komentiranjem.

3. **Azijska riža**

CP	Duljina odgovora - P	TP	PPV	TPR	F1-score
155	137	68	0.50	0.44	0.47

Tablica 5.46: Rezultat za *upit = VFFGDSLSDN*

CP	Duljina odgovora - P	TP	PPV	TPR	F1-score
155	118	110	0.93	0.71	0.81

Tablica 5.47: Rezultat za *centralni motiv*

4. **Rajčica**

CP	Duljina odgovora - P	TP	PPV	TPR	F1-score
108	106	53	0.5	0.49	0.49

Tablica 5.48: Rezultat za *upit = VFFGDSLSDN*

CP	Duljina odgovora - P	TP	PPV	TPR	F1-score
108	109	91	0.83	0.84	0.83

Tablica 5.49: Rezultat za *centralne motive*

Poglavlje 6

Analiza rezultata i uspješnost metode

Iako smo za svaku vrstu testa imali kratki osvrt na rezultate, u poglavlju [5], sada ćemo grupno sagledati testiranja i donijeti zaključak o uspješnosti metode. Ovo malo istraživanje ima par nedostataka na kojima se treba poraditi u slučaju daljnjeg razvoja metode. No, pitanje je može li se naći rješenje za sve nedostatke. Na kraju ćemo kratko reći nešto i o njima.

Testiranja smo provodili na dva iterativna pretraživača. Za svaki početni upit i svaku biljku je odmah vidljivo da IGLOSS daje bolje rezultate. Kod rajčice, za upit FVFNDSLSDA, F1 score za BLAST iznosi 0.48 dok za IGLOSS iznosi 0.80. Dakle, bez ikakve metode poboljšavanja uočavamo da je IGLOSS za sve kombinacije testa poboljšanje BLAST-a. Napomenimo još jednom da smo nakon pronalaska centralnog motiva u odgovoru iz BLAST-a pretragu s njim vršili na IGLOSS-u. Stoga, nije teško za zaključiti da je naša metoda centralnog motiva dala daleko veće poboljšanje u kombinaciji pretraživača BLAST/IGLOSS, u odnosu na one vezane samo za IGLOSS. To je vidljivo već na prvom upitu, karakterističnom motivu GDSL lipaza - FVFGDSLSDA. Kod sve 4 biljke nakon pretrage ovog upita na BLAST-u i provođenja metode centralnog motiva, rezultati F1 score-a su skočili između 0.1 i 0.19, što je dobro. Najbolje poboljšanje je uočeno kod azijske riže gdje prije metode F1 iznosi 0.64, a nakon metode 0.83. Nakon pretrage istog upita na IGLOSS-u i provođenja metode do poboljšanja F1 score-a je došlo kod talijinog uročnjaka, rajčice, krumpira, ali kod azijske riže se F1 score smanjio za 0.02. Taj jedini slučaj na kojem nije metoda bila uspješna ne možemo smatrati relevantim za donošenje odluke o uspješnosti metode jer je u tom slučaju PPV=0.95 što je gotovo savršeno i to poboljšati nije lako. Ono što je još važno za uočiti je da su sve PPV vrijednosti visoke za ovaj upit baš zato što je karakterističan za tu proteinsku familiju te svaki uspjeh poboljšanja kod tog upita je od značaja. Za modificirani upit FVFNDSLSDA BLAST, za svaku biljku, daje odgovor sa niskim F1 score-om, svaki je manji od 0.5. Nakon provedene metode poboljšanja najveći skok F1 score-a se dogodio kod rajčice gdje je od 0.48 skočio na 0.86. Kod svih biljaka

je uočeno veliko pobljšanje. Za isti upit i pretragu na IGLOSS-u također su poboljšanja vidljiva kod sve četiri biljke unatoč tome što se ovdje F1 score-ovi i prije metode kreću oko 0.7. Uvjerljivo najmanji F1 score se postiže kod pretrage na BLASTU za upit VFFGDSLSDN te tako kod krumpira on iznosi 0.25. Isto tako, uvjerljivo najbolje poboljšanje je u tom slučaju, skoro tri puta se povećao F1 score te nakon metode kod krumpira iznosi 0.72. U tim okvirima je poboljšanje i kod ostale tri biljke. IGLOSS za taj upit daje bolje rezultate F1 score-a od BLASTA, ali metodom ih mi još više poboljšavamo. Primjerice, kod krumpira smo ga poboljšali sa vrijednosti 0.37 na 0.73.

Zaključno, možemo reći da je metoda centralnog motiva uspješna kod sve 4 biljke za sva tri upita. Najveća poboljšanja su uočena kod upita VFFGDSLSDN, a najmanja kod upita FVFGDSLSDA što je bilo i za očekivati. Ne možemo reći kod koje biljke se najviše poboljšanja dogodilo jer se od testa do testa rezultati razlikuju.

Nedostaci metode

Neki od nedostataka u ovom istraživanju su varijabilnost rezultata u ovisnosti o odabiru skale te odabiru centralnih motiva. Rekli smo već da u pretraživaču IGLOSS skalu postavljamo sami i da o njoj ovisi duljina odgovora, isto tako ovisi jako i F1 score. Za svaki test u poglavlju [5] proizvoljno smo postavljali skalu. U jedinom slučaju kada metoda nije bila uspješna, kod azijske riže, da smo postavili skalu drugačije ni duljina odgovora ni PPV ne bi bili isti te bi možda u tom slučaju uspjeli i tamo poboljšati rezultat. Dakle, nemamo konkretno rješenje s kojom skalom treba pretraživati upit s centralnim motivom. Za odabir centralnog motiva vidjeli smo da nema pravila. Nekad smo uzeli jedan centralni motiv, a nekad više njih. Kad iz grafičkog prikaza nije jasno vidljivo gdje je zadnji "skok" tada intuitivno biramo koje ćemo motive uzeti za centralne. Na jednom testu smo, u poglavlju [5], pokazali rezultate sa odabirom jednog centralnog motiva te sa odabirom više njih. U oba slučaja smo uočili uspješnost metode, ali je na jednom slučaju ipak bila više uspješna. Nije još sasvim jasno kako prepoznati i odrediti što će biti centralno.

Glavni nedostatak metode je taj što rezultati ovise o odgovoru na početni upit. Kada bi za upit uzeli niz aminokiselina čiji odgovor ne sadrži biološki značajne proteine tada ne bi imali mogućnost poboljšati ga. Isto tako, kada bi za neki upit odgovor sadržavao vrlo malo biološki značajnih proteina među kojima se ne bi našao centralni motiv tada također ne bi mogli poboljšati odgovor. U našim testovima, to nije bio slučaj. Primjerice, kod upita VFFGDSLSDN, na nekim testovima udio biološki značajnih proteina u odgovoru bio je svega 23%, ali centralni motiv se našao upravo u tih 23% i mi smo njime značajno poboljšali odgovor.

Bibliografija

[1] Warren; Miller Webb; Myers Eugene; Lipman David Altschul, Stephen; Gish, *Basic local alignment search tool*, Journal of Molecular Biology (1990), https://blastalgorithm.com/.

[2] M. Zagorsčak M. Rosenzweig i P. Goldstein B. Rabar, S. Ristov, *Igloss: Iterative gapless local similarity search*, arXiv:1807.11862v1 [q-bio.QM] (2008), https://arxiv.org/pdf/1807.11862.pdf.

[3] M. Huzak, *Statistika (prezentacije)*, https://web.math.pmf.unizg.hr/nastava/stat/index.php?sadrzaj=predavanja.php.

[4] 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.

[5] K. Martinić, *Maksimalne klike u analizi sličnosti proteinskih motiva*, Diplomski rad, Sveučilište u Zagrebu, Prirodoslovno-matematički fakultet, Matematički odsjek, 2018.

[6] N. Sarapa, *Teorija vjerojatnosti*, Školska knjiga, 2002.

[7] Z. Vondraček, *Markovljevi lanci (skripta)*, 2008.

[8] ———, *Slučajni procesi (skripta)*, https://web.math.pmf.unizg.hr/~vondra/sp17-predavanja.html.

Sažetak

Iterativnom pretragom, uz pomoć upita, na proteomu tražimo proteine iz određene klase, nazivamo ih biološki značajnim proteinima. Odgovor na dani upit može se dobiti raznim iterativnim pretraživačima te on može i ne mora sadržavati biološki značajne proteine. U ovom radu analiziramo tehniku poboljšanja točnosti iterativnog pretraživanja. Točnost mjerimo statističkom mjerom F1 score, a tehnika se naziva metoda centralnog elementa. Odgovor na dani upit promatramo kao Markovljev proces na konačnom skupu. Koristeći rezultate iz Markovljevih lanaca pronalazimo težište, centar, odgovora. Postupak iterativne pretrage ponovimo sa centralnim elementom te usporedimo statističke mjere točnosti testa prije i nakon korištenja metode.

Tehnika se provodi za različite upite, na proteomima četiri različite biljke - talijin uročnjak, rajčica, krumpir i azijska riža. Nakon analize rezultata za sve kombinacije upita, biljaka i iterativnih pretraživača, dolazimo do zaključka da našom tehnikom poboljšavamo točnost pretrage, a negdje i bitno poboljšavamo.

Summary

Motif scanning is a very important part of bioinformatics. In this work, we are concerned with improving accuracy of certain iterative scanning tools. The method that we propose - a Central Element Method - is based on interpreting the response to a query as a finite Markov process, and results from Markov chain theory are used to obtain the center. The iterative search process is then repeated with the central element, and the statistical measures of test accuracy before and after are compared. The accuracy is measured by F1 score.

The method is applied for various queries, on proteomes of four different plants – thale cress, tomato, potato and Asian rice. It turns out that this technique improves search accuracy and, in some instances, improves it significantly.

Životopis

Rođena sam 7. srpnja 1994. u Spaichingenu, Republika Njemačka. Nakon završene osnovne škole upisala sam XV. gimnaziju, informatički smjer, u Zagrebu. Po završetku srednjoškolskog obrazovanja, upisujem Preddiplomski studij matematike na Prirodoslovno-matematičkom fakultetu u Zagrebu. 2016. godine upisujem Diplomski studij Matematičke statistike na istom fakultetu.