Clustering i klasifikacija proteinskih nizova

Višek, Ivana

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Ivana Višek

CLUSTERING I KLASIFIKACIJA PROTEINSKIH NIZOVA

Diplomski rad

Voditelj rada:
doc. dr. sc. Pavle Goldstein

Zagreb, srpanj, 2022.

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. _____

Majci, Ocu, Sestri i Bratu za svaki uspon i pad podijeljen s njima.
Prijateljima koji su uljepšali ovo razdoblje.
Mentoru za strpljenje i velikoj pomoći u izradi ovog rada.

Sadržaj

Uvod

Proteom je skup svih proteina nekog organizma. Proteini su složene molekule, sastavljene od aminokiselina, koje su sastavni dio stanice svih živih bića. U proteinskoj familiji nalaze se proteini sa zajedničkim evolucijskim podrijetlom koje je zaslužno za njihova ista svojstva. U ovom radu analizirat ćemo skup motiva, tj. dijelova proteina. Pretražit ćemo motive te pronaći one koji su međusobno najsličniji, a pritom odbaciti one koji nisu. Važno je napomenuti da se bavimo nenadziranom klasifikacijom, odnosno iz motiva nekih proteina, za koje ne znamo jesu li slični upitu, dolazimo do zaključka da se oni koji jesu, grupiraju. Kako je svaka kugla zadana središtem i radijusom, bavit ćemo se pronalaskom upravo tih elemenata. Metoda razvijena u ovom radu vraća kuglu koja sadrži najviše motiva iz danog proteina te testiranjem pokazujemo da upravo ta kugla sadrži motive koji su najsličniji upitu.

Rad se sastoji od tri poglavlja. U prva dva poglavlja navodimo važne pojmove, definicije i teoreme iz biologije, bioinformatike te linearne algebre i statistike koji su važni za razumijevanje rada. Za kraj detaljno opisujemo algoritam i analiziramo podatke. Analizom uočavamo da rezultati potvrđuju pretpostavku grupiranja međusobno sličnih proteina.

Poglavlje 1

Bioinformatika

1.1 Biološki pojmovi

Proteini ili bjelančevine su sastavni dijelovi svake stanice, što ih čini jednom od osnovnih komponenta života na zemlji. Imaju različite funkcije u organizmu poput kataliziranja metaboličkih reakcija, repliciranje DNA, ali su i glavni čimbenici u rastu i razvoju svih tjelesnih tkiva. Izgrađeni su od 20 standardnih aminokiselina koje su međusobno povezane peptidnom vezom. Svakoj od aminokiselina pridruženo je jedno slovo engleske abecede kao u tablici:

Oznaka	Naziv	Oznaka	Naziv
A	Alanin	M	Metionin
C	Cistenin	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 1.1: Standardne aminokiseline

Neka je R slučajna varijabla koja predstavlja aminokiselinu s distribucijom

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

Vjerojatnosti navedene u distribuciji predstavljaju vjerojatnosti pojavljivanja pripadne aminokiseline u prostoru proteina.

Skup svih proteina nekog organizma naziva se proteom. Sastoji se od različitih proteinskih familija koje su zaslužne za različita svojstva organizma. Promatrat ćemo proteinsku familiju GDSL lipaza. **GDSL lipaze** jedan su od primjera lipaza, enzima koji sudjeluju kao katalizatori u razgradnji lipida odnosno masti. Iako su elementi GDSL lipaza važni u raznim staničnim procesima poput razvoja biljaka i zaštite organizma, još uvijek nisu u potpunosti istražene te je razumijevanje istih vrlo ograničeno. Kako se nalaze u biljkama, životinjama i u bakterijama mogle bi biti bogat izvor obećavajućih enzima stoga su od velikog interesa bioinformatike. U ovom radu promatrat ćemo četiri biljke u kojima se nalaze GDSL lipaze. Proteini se sastoje od velikog broja aminokiselina pa ćemo raditi s motivima. Motiv je niz od 5 do 20 aminokiselina. Ako su 2 motiva dovoljno slična, smatrat ćemo da se nalaze u proteinima koji pripadaju istoj proteinskoj familiji.

Poglavlje 2

Matematički pojmovi

U ovom poglavlju navode se teoremi, definicije, propozicije i napomene iz linearne algebre, statistike te uspješnosti modela. Pojmovi su preuzeti iz izvora [2], [3], [4], [5].

2.1 Linearna algebra

Definicija 2.1.1. *Neka je $\mathbb{F}$ skup na kojem su definirane binarne operacije zbrajanja $+ : \mathbb{F} \times \mathbb{F} \to \mathbb{F}$ i množenja $\cdot : \mathbb{F} \times \mathbb{F} \to \mathbb{F}$ koje imaju sljedeća svojstva:*

1) $\alpha + (\beta + \gamma) = (\alpha + \beta) + \gamma, \ \forall \alpha, \beta, \gamma \in \mathbb{F}$*;*

2) postoji $0 \in \mathbb{F}$ *sa svojstvom* $\alpha + 0 = 0 + \alpha = \alpha, \ \forall \alpha \in \mathbb{F}$*;*

3) za svaki $\alpha \in \mathbb{F}$*, postoji* $-\alpha \in \mathbb{F}$ *tako da je* $\alpha + (-\alpha) = (-\alpha) + \alpha = 0$*;*

4) $\alpha + \beta = \beta + \alpha, \ \forall \alpha, \beta \in \mathbb{F}$*;*

5) $(\alpha\beta)\gamma = \alpha(\beta\gamma), \ \forall \alpha, \beta, \gamma \in \mathbb{F}$*;*

6) postoji $1 \in \mathbb{F} \setminus \{0\}$ *sa svojstvom* $1 \cdot \alpha = \alpha \cdot 1 = \alpha, \ \forall \alpha \in \mathbb{F}$*;*

7) za svaki $\alpha \in \mathbb{F}, \ \alpha \neq 0,$ *postoji* $\alpha^{-1} \in \mathbb{F}$ *tako da je* $\alpha\alpha^{-1} = \alpha^{-1}\alpha = 1$*;*

8) $\alpha\beta = \beta\alpha, \ \forall \alpha, \beta \in \mathbb{F}$*;*

9) $\alpha(\beta + \gamma) = \alpha\beta + \alpha\gamma, \ \forall \alpha, \beta, \gamma \in \mathbb{F}$*.*

Tada kažemo da je uređena trojka $(\mathbb{F}, +, \cdot)$ *polje. Elemente polja nazivamo skalarima.*

Napomena 2.1.2. *Skup realnih brojeva $\mathbb{R}$ s uobičajenim operacijama zbrajanja i množenja je polje.*

Definicija 2.1.3. *Neka je V neprazan skup na kojem su zadane binarne operacije zbrajanja $+ : V \times V \to V$ i operacija množenja skalarima iz polja $\mathbb{F}$, $\cdot : \mathbb{F} \times V \to V$. Kažemo da je uređena trojka $(V, +, \cdot)$ vektorski prostor nad poljem $\mathbb{F}$ ako vrijedi:*

1) $a + (b + c) = (a + b) + c, \ \forall a, b, c \in V;$

2) postoji $0 \in V$ *sa svojstvom* $a + 0 = 0 + a = a, \ \forall a \in V;$

3) za svaki $a \in V$, *postoji* $-a \in V$ *tako da je* $a + (-a) = (-a) + a = 0;$

4) $a + b = b + a, \ \forall a, b \in V;$

5) $\alpha(\beta a) = (\alpha\beta)a, \ \forall \alpha, \beta \in \mathbb{F}, \forall a \in V;$

6) $(\alpha + \beta)a = \alpha a + \beta a, \ \forall \alpha, \beta \in \mathbb{F}, \forall a \in V;$

7) $\alpha(a + b) = \alpha a + \alpha b, \ \forall \alpha \in \mathbb{F}, \forall a, b \in V;$

8) $1 \cdot a = a \cdot 1, \ \forall a \in V.$

Definicija 2.1.4. *Za prirodne brojeve m i n, preslikavanje*

$$A : \{1, 2, \ldots, m\} \times \{1, 2, \ldots, n\} \to \mathbb{F}$$

naziva se matrica tipa (m, n) s koeficijentima iz polja $\mathbb{F}$.

Definicija 2.1.5. *Neka je V vektorski prostor nad poljem $\mathbb{F}$. Skalarni produkt na V je preslikavanje $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{F}$ koje ima sljedeća svojstva:*

1) $\langle x, x \rangle \geq 0, \ \forall x \in V;$

2) $\langle x, x \rangle = 0 \Leftrightarrow x = 0;$

3) $\langle x_1 + x_2, y \rangle = \langle x_1, y \rangle + \langle x_2, y \rangle, \ \forall x_1, x_2, y \in V;$

4) $\langle \alpha x, y \rangle = \alpha \langle x, y \rangle, \ \forall \alpha \in \mathbb{F}, \forall x, y \in V;$

5) $\langle x, y \rangle = \overline{\langle y, x \rangle}, \ \forall x, y \in V.$

Napomena 2.1.6. *U $\mathbb{R}^n$ kanonski skalarni produkt definiran je s*

$$\langle (x_1, \ldots, x_n), (y_1, \ldots, y_n) \rangle = \sum_{i=1}^{n} x_i y_i.$$

Definicija 2.1.7. *Vektorski prostor na kojem je definiran skalarni produkt zove se unitaran prostor.*

Definicija 2.1.8. *Neka je V unitaran prostor. Norma na V je funkcija $\|\cdot\| : V \to \mathbb{R}$ definirana s*

$$\|x\| = \sqrt{\langle x, x\rangle}.$$

Propozicija 2.1.9. *Norma na unitarnom prostoru V ima sljedeća svojstva:*

1) $\|x\| \geq 0, \ \forall x \in V;$

2) $\|x\| = 0 \Leftrightarrow x = 0;$

3) $\|\alpha x\| = |\alpha| \, \|x\|, \ \forall \alpha \in \mathbb{F}, \forall x \in V;$

4) $\|x + y\| \leq \|x\| + \|y\|, \ \forall x, y \in V.$

Definicija 2.1.10. *Svako preslikavanje $\|\cdot\| : V \to \mathbb{R}$ na vektorskom prostoru V sa svojstvima iz propozicije 2.1.9 naziva se norma. Tada $(V, \|\cdot\|)$ zovemo normirani prostor.*

Definicija 2.1.11. *Norma koja potječe od kanonskog skalarnog produkta na $\mathbb{F}^n$, definiranog u napomeni 2.1.6, dana je formulom*

$$\|(x_1, \ldots, x_n)\| = \sqrt{\sum_{i=1}^{n} |x_i|^2}.$$

Ova se norma zove euklidska norma.

Definicija 2.1.12. *Neka je V normiran prostor. Metrika ili udaljenost vektora x i y je funkcija $d : V \times V \to \mathbb{R}$ definirana s*

$$d(x, y) = \|x - y\|.$$

Propozicija 2.1.13. *Metrika na normiranom prostoru ima sljedeća svojstva:*

1) $d(x, y) \geq 0, \ \forall x, y \in V;$

2) $d(x, y) = 0 \Leftrightarrow x = y, \ \forall x, y \in V;$

3) $d(x, y) = d(y, x), \ \forall x, y \in V;$

4) $d(x, y) \leq d(x, z) + d(z, y), \ \forall x, y, z \in V.$

Definicija 2.1.14. *Neka je $X \neq \emptyset$. Svaka funkcija $d : X \times X \to \mathbb{R}$ sa svojstvima iz propozicije 2.1.13 naziva se metrika ili udaljenost. Tada (X, d) zovemo metrički prostor.*

Definicija 2.1.15. *Neka su* $x = (x_1, \ldots, x_n)$ *i* $y = (y_1, \ldots, y_n)$ *proizvoljni vektori u* $\mathbb{R}^n$. *Metrika na* $\mathbb{R}^n$, *inducirana euklidskom normom iz definicije 2.1.11, dana je s*

$$d((x_1, \ldots, x_n), (y_1, \ldots, y_n)) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}.$$

Ova metrika naziva se euklidska metrika, a prostor $\mathbb{R}^n$ *s tom metrikom nazivamo euklidski prostor.*

Definicija 2.1.16. *Neka je* (X, d) *metrički prostor. Za proizvoljno* $a \in \mathbb{R}$ *i proizvoljan* $r > 0 \in \mathbb{R}$ *skup*

$$K(a, r) = \{x \in X \mid d(a, x) < r\},$$

nazivamo otvorena kugla u X, sa centrom a i radijusom r.

Definicija 2.1.17. *U euklidskom prostoru* $\mathbb{R}^n$ *otvorena kugla sa centrom* $a \in \mathbb{R}^n$ *i radijusom* $r > 0 \in \mathbb{R}$ *dana je s*

$$K(a, r) = \left\{ x \in \mathbb{R}^n \mid \sqrt{\sum_{i=1}^{n} (a_i - x_i)^2} < r \right\}.$$

2.2 Statistika

Matematičko očekivanje i varijanca

Definicija matematičkog očekivanja provodi se u tri koraka. Prvo se definira matematičko očekivanje jednostavne slučajne varijable, zatim nenegativne slučajne varijable i na kraju opće slučajne varijable.

Neka je $(\Omega, \mathcal{F}, \mathbb{P})$ vjerojatnosni prostor. Označimo sa $\mathcal{K}$ skup svih jednostavnih slučajnih varijabli definiranih na Ω, a sa $\mathcal{K}_+$ skup svih nenegativnih funkcija iz $\mathcal{K}$.

Neka je $X \in \mathcal{K},\ X = \sum_{k=1}^{n} x_k \mathcal{K}_{A_k}$, gdje su $A_1, A_2, \ldots, A_n \in \mathcal{F}$ međusobno disjunktni.

Definicija 2.2.1. *Matematičko očekivanje od X ili, kraće, očekivanje od X koje označavamo sa* $\mathbb{E}[X]$ *definira se sa:*

$$\mathbb{E}[X] = \sum_{k=1}^{n} x_k \mathbb{P}(A_k).$$

Propozicija 2.2.2. *1. Neka je $c \in \mathbb{R}$ i $X \in \mathcal{K}$. Tada je $\mathbb{X} = c\mathbb{E}X$.*

2. Za $X, Y \in \mathcal{K}$ vrijedi $\mathbb{E}(X + Y) = \mathbb{E}X + \mathbb{E}Y$.

3. Neka su $X, Y \in \mathcal{K}$ i $X \leq Y$. Tada je $\mathbb{E}X \leq \mathbb{E}Y$.

Neka je X nenegativna slučajna varijabla definirana na Ω. Prema teoremu 3.0.1 postoji rastući niz $(X_n)_{n\in\mathbb{N}}$ nenegativnih jednostavnih slučajnih varijabli takav da je $X = \lim_{n\to\infty} X_n$. Iz propozicije 2.2.2 slijedi da je niz $(\mathbb{E}[X_n])_{n\in\mathbb{N}}$ rastući niz u $\mathbb{R}_+$, dakle postoji $\lim_{n\to\infty} \mathbb{E}[X_n]$ koji može biti jednak i $+\infty$.

Definicija 2.2.3. *Matematičko očekivanje od X ili, kraće, očekivanje od X definira se sa*

$$\mathbb{E}[X] = \lim_{n\to\infty} \mathbb{E}[X_n].$$

Neka je sada X proizvoljna slučajna varijabla na Ω. Vrijedi $X = X^+ - X^-$, X^+, X^- su slučajne varijable i $X^+, X^- \geq 0$.

Definicija 2.2.4. *Kažemo da matematičko očekivanje od X ili kraće, očekivanje od X postoji ili da je definirano ako je barem jedna od veličina $\mathbb{E}[X^+], \mathbb{E}[X^-]$ konačna, tj. vrijedi $min\{\mathbb{E}[X^+], \mathbb{E}[X^-]\} < +\infty$. Tada po definiciji stavljamo*

$$\mathbb{E}[X] = \mathbb{E}[X^+] - \mathbb{E}[X^-].$$

Neka je X slučajna varijabla na vjerojatnosnom prostoru $(\Omega, \mathcal{F}, \mathbb{P})$ i $r > 0$.

Definicija 2.2.5. *$\mathbb{E}(X^r)$ zovemo r-ti moment od X, a $\mathbb{E}(|X|^r)$ zovemo r-ti apsolutni moment od X*

Definicija 2.2.6. *Neka $\mathbb{E}X$ postoji (tj. konačno je). Tada $\mathbb{E}[(X-\mathbb{E}X)^r]$ zovemo r-ti apsolutni centralni moment od X.*

Definicija 2.2.7. *Varijanca od X koju označavamo sa $Var(X)$ ili σ_X^2 jest drugi centralni moment od X, dakle*

$$Var(X) = \mathbb{E}[(X - \mathbb{E}[X])^2].$$

Napomena 2.2.8. *Pozitivan drugi korijen iz varijance nazivamo standardna devijacija i označavamo sa σ_X.*

Opisna analiza podataka

Za razumijevanje ovog rada još će biti potrebno znanje o radu s podacima. Navodimo pojmove deskriptivne statistike tj. pojmove aritmetičke sredine, standardne devijacije i varijance uzorka te standardizaciju podataka.

Neka su

$$x_1,\ x_2, \ldots,\ x_n \tag{2.1}$$

vrijednosti (opažanja) varijable X koje čine skup podataka. Ako je X numerička varijabla, tada je to niz brojeva. Neka je u nastavku X numerička varijabla.
Aritmetička sredina uzorka (2.1) je mjera centralne tendencije i definirana je kao:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Varijanca uzorka ili podataka (2.1) je mjera raspršenja podataka i predstavlja prosječno kvadratno odstupanje podataka od njihove aritmetičke sredine i dana je formulom:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

Iz prethodnih definicija slijedi da je **standardna devijacija uzorka** drugi korijen varijance i zadana je formulom:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

Standardizacija podataka je česta procedura u statistici prije obrade podataka i izgradnje modela ili algoritma. Podaci se transformiraju oduzimanjem očekivanja i dijeljenjem sa standardnom devijacijom uzorka:

$$x_i' = \frac{x_i - \overline{x}}{s}. \tag{2.2}$$

2.3 Klasifikacija i uspješnost modela

Klasifikacija

Klasifikacija podataka je problem određivanja pripadnosti opservacije nekoj skupini odnosno klasi. Razlikujemo nadziranu i nenadziranu klasifikaciju. U nadziranom radu modeli i algoritmi rade s poznatim skupom ulaznih i poznatim skupom izlaznih podataka stoga unaprijed znaju klase tih podataka. Nenadzirana klasifikacija koristi podatke koji nemaju unaprijed određene klase, već se na podacima pokušava odrediti neku sličnost i separirati ih u klase po toj sličnosti.

Mjere uspješnosti

Da bi se ocijenila uspješnost nekog modela, definirane su mjere uspješnosti koje se temelje na pojmovima iz matrice uspješnosti (eng. *confusion matrix*) prikazanoj sljedećom tablicom.

		Predviđeno stanje		
	Ukupna populacija	Ocijenjeni pozitivno (P)	Ocijenjeni negativno (N)	
Stvarno stanje	Pozitivno stanje (CP)	TP (stvarno pozitivni)	FN (lažno negativni)	Osjetljivost (TPR)
	Negativno stanje (CN)	FP (lažno pozitivni)	TN (stvarno negativni)	Specifičnost (TNR)
		Preciznost (PPV)	Negativna prediktivna vrijednost (NPV)	

Tablica 2.1: Tablica uspješnosti

Napomena 2.3.1. *U ovom radu će se provjera broja TP (eng. True Positives) i ostalih brojeva iz matrice uspješnosti (FP, FN, TN) vršiti na temelju liste CP (eng. Condition Positive). Lista CP sadrži sve proteine za koje je pripadnost određenoj familiji već utvrđena, biološki poznata. Dakle, u savršenom testu bi svi proteini s liste CP bili pozitivno ocijenjeni, a svi proteini koji nisu na listi CP bi bili negativno ocijenjeni.*

Slijede definicije nekih od mjera uspješnosti modela za binarnu klasifikaciju:
Osjetljivost ili TPR (eng. *True Positive Rate*) je postotak pozitivnih elemenata uzorka u odnosu na određeno stanje, odnosno CP elemenata uzorka, koji su ispravno prepoznati kao pozitivni.

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

Specifičnost ili TNR (eng. *True Negative Rate*) je postotak negativnih elemenata uzorka u odnosu na određeno stanje, odnosno CN (eng. *Condition Negative*) elemenata uzorka, koji su ispravno prepoznati kao negativni.

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

Preciznost ili PPV (eng. *Positive Predictive Value*) je omjer broja stvarno pozitivnih elemenata uzorka i broja elemenata uzorka koji su modelom prepoznati kao pozitivni.

$$\text{PPV} = \frac{\text{broj stvarno pozitivnih}}{\text{broj stvarno pozitivnih } + \text{ broj lažno pozitivnih}} = \frac{\text{TP}}{\text{P}}$$

Negativna prediktivna vrijednost ili NPV (eng. *Negative Predictive Value*) je omjer broja stvarno negativnih elemenata uzorka i broja elemenata uzorka koji su modelom prepoznati kao negativni.

$$\text{NPV} = \frac{\text{broj stvarno negativnih}}{\text{broj stvarno negativnih} \ + \ \text{broj lažno negativnih}} = \frac{\text{TN}}{\text{N}}$$

F_β-score je mjera uspješnosti modela koja povezuje osjetljivost i preciznost. Dobiva se kao harmonijska sredina osjetljivosti i preciznosti modela, uz težinski faktor β.

$$F_\beta = \frac{\left(\beta^2 + 1\right) \cdot PPV \cdot TPR}{\beta^2 \cdot PPV + TPR}$$

U ovom radu, kao mjera uspješnosti modela koristit će se F_1-score ($\beta = 1$):

$$F_1 = \frac{2 \cdot PPV \cdot TPR}{PPV + TPR} \tag{2.3}$$

Napomena 2.3.2. *Sve navedene mjere postižu vrijednosti isključivo na intervalu [0, 1]. Model je uspješniji po nekoj od navedenih mjera, što je ta mjera bliže broju 1.*
β faktor u F_β-score određuje kojoj mjeri dajemo veću težinu. Za $\beta < 1$ daje se više važnosti minimiziranju lažno pozitivnih. Za $\beta > 1$ daje se više važnosti minimiziranju lažno negativnih.

Poglavlje 3

Analiza problema i algoritam

Problem ovog rada je pronalazak nekih dovoljno sličnih motiva. Rješenje tog problema ponudit ćemo pronalaskom optimalne kugle koja sadrži što više motiva. Upit i skalu pretraživanja smo unijeli u IGLOSS server [8] koji nam je ponudio niz motiva za koje je on procijenio da su dovoljno slični upitu. Skala pretraživanja je parametar koji postavlja granicu dovoljne sličnosti. Što je skala veća, više se kažnjava odstupanje od upita, pa odgovor ima sličnije motive. Nakon što nam IGLOSS da svoje kandidate za motive, među njima se nalaze oni koji su iz proteina koji zaista pripadaju toj familiji (eng. *true positives*) i onih koji nisu u njoj (eng. *false positives*). Želimo ga poboljšati izbacivanjem što više lažno pozitivnih elemenata, ali tom eliminacijom ne želimo eliminirati i prave pozitivne elemente. Uspješnost metode ćemo mjeriti F_1 − *scoreom*.

Pretpostavka je da se pravi pozitivni elementi grupiraju pa prelazimo u vektorski prostor kako bismo mogli iskoristiti sva njegova svojstva. Problem smo time sveli na traženje središta i radijusa kugle sa što više pravih pozitivaca u sebi.

Prelazak u vektorski prostor

Upit kao i odgovor će biti desetorke aminokiselina koje moramo opisati numeričkim vrijednostima kako bismo na njima mogli provesti razna testiranja. Ta problematika je opisana i riješena u članku [1]. Definira se preslikavanje u $\mathbb{R}^5$ koje svakoj aminokiselini pridružuje 5-dimenzionalni vektor. Svaka koordinata tog vektora označava jedno ili kombinaciju više svojstava te aminokiseline. Koordinate vektora ćemo nazivati *faktorima*. *Faktor I* se odnosi na polaritet aminokiseline, *Faktor II* je vezan za sekundarnu strukturu, *Faktor III* se odnosi na molekularni volumen, *Faktor IV* odražava raznolikost kodona (relativnu kompoziciju aminokiselina u različitim proteinima) te *Faktor V* odgovara elektrostatičkom naboju aminokiseline.

AMINOKISELINA	Faktor I	Faktor II	Faktor III	Faktor IV	Faktor V
A	-0.591	-1.302	-0.733	1.570	-0.146
C	-1.343	0.465	-0.862	-1.020	-0.255
D	1.050	0.302	-3.656	-0.259	-3.242
E	1.357	-1.453	1.477	0.113	-0.837
F	-1.006	-0.590	1.891	-0.397	0.412
G	-0.384	1.652	1.330	1.045	2.064
H	0.336	-0.417	-1.673	-1.474	-0.078
I	-1.239	-0.547	2.131	0.393	0.816
K	1.831	-0.561	0.533	-0.277	1.648
L	-1.019	-0.987	-1.505	1.266	-0.912
M	-0.663	-1.524	2.219	-1.005	1.212
N	0.945	0.828	1.299	-0.169	0.933
P	0.189	2.081	-1.628	0.421	-1.392
Q	0.931	-0.179	-3.005	-0.503	-1.853
R	1.538	-0.055	1.502	0.440	2.897
S	-0.228	1.399	-4.760	0.670	-2.647
T	-0.032	0.326	2.213	0.908	1.313
V	-1.337	-0.279	-0.544	1.242	-1.262
W	-0.595	0.009	0.672	-2.128	-0.184
Y	0.260	0.830	3.097	-0.838	1.512

Tablica 3.1: Faktori

Niz od n aminokiselina sada odgovara $5n$-dimenzionalnom vektoru. Nakon ovakvog pridruživanja možemo koristiti alate za koje znamo da vrijede u svakom vektorskom prostoru.

Priprema podataka

U ovom radu upit će biti FVFGDSLSDA. Upit sadrži niz aminokiselina GDSL, koji je karakterističan za promatranu proteinsku familiju. Korištenjem takvog upita žele se, uz pomoć IGLOSS servera, dobiti najbolji kandidati za familiju GDSL lipaza. Odgovor kojeg ćemo promatrati je skup motiva, a svaki element tog skupa će, kao i upit, biti niz aminokiselina duljine 10. Pridruživanjem koje smo opisali u prethodnom potpoglavlju dobijemo 50-dimenzionalne vektore tj. nalazimo se u $\mathbb{R}^{50}$.

Kako bismo riješili eventualne probleme s neravnomjerno raspršenim podacima, moramo standardizirati podatke. Naime, naši podaci nemaju definirane mjerne jedinice pa je moguće da su nam varijanca i raspon podataka po jednoj koordinati veći od onih po ostalim koordinatama, čime gubimo formu kugle u kojoj bi sve koordinate trebale imati

podjednak utjecaj. Provedbom standardizacije podataka izbjegavamo taj problem. Standardizaciju smo prilagodili kako bismo izbjegli dijeljenje brojem blizu nule pa smo dijelili sa standardnom devijacijom uvećanom za 0.1. Neka su $x_1, x_2, ..., x_n$ vrijednosti koje čine skup podataka tada je za $i = 1, 2, ..., n$:

$$x_i' = \frac{x_i - \overline{x}}{s + 0.1}. \tag{3.1}$$

Radijus i središte optimalne kugle

Za razumijevanje pronalaska radijusa potrebna su nam sljedeća 2 teorema:

Teorem 3.0.1. *Površina kvadrata nad hipotenuzom pravokutnog trokuta jednaka je zbroju površina kvadrata nad njegovim katetama.*

Teorem 3.0.2. *Očekivana udaljenost dvije točke koje su uniformno distribuirane u kugli u n-dimenzionalnom prostoru teži u $r\sqrt{2}$ kada $n \to \infty$, gdje je r radijus te kugle.*

Prvi teorem je detaljno opisan i obrađen u izvoru [6, str. 55], dok je drugi svima poznat Pitagorin teorem iz kojeg slijedi formula za udaljenost 2 točke. Podaci koje koristimo su motivi duljine 10 pa računamo očekivanu udaljenost dviju desetorki. Neka su $X = (x_1, x_2, \ldots, x_{10})$, $Y = (y_1, y_2, \ldots, y_{10})$, dvije desetorke aminokiselina. Očekivanje kvadrata euklidske udaljenosti X i Y je:

$$\mathbb{E}\left[d^2(X,Y)\right] = \mathbb{E}\left[\left((x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_{10} - y_{10})^2\right)\right]$$

Iz svojstva očekivanja slijedi:

$$\mathbb{E}\left[d^2(X,Y)\right] = \sum_{k=1}^{10} \mathbb{E}\left[(\bar{x}_i - \bar{y}_i)^2\right]$$

S obzirom na to da ne razlikujemo pozicije aminokiselina u tim desetorkama, jer nemamo informaciju o istim, možemo označiti s $\bar{a}_i$ i $\bar{a}_j$ aminokiseline koje pripadaju prosječnoj distribuciji aminokiselina te dobijemo:

$$\mathbb{E}\left[d^2(X,Y)\right] = 10\mathbb{E}\left[\left(\bar{a}_i - \bar{a}_j\right)^2\right]$$

Za sljedeći korak trebamo definirati distribuciju aminokiselina.

Pretpostavimo da se aminokiseline pojavljuju s vjerojatnostima r_k, 1.1, $k \in \{1, 2, \ldots, 20\}$,

te neka su A_i $i \in \{1, 2, \ldots, 20\}$ distribucije zadane nekom aminokiselinom za koju ćemo reći da je očuvana koeficijentom očuvanosti α. Tada je:

$$A_i \sim \begin{pmatrix} a_1^i & a_2^i & \ldots & a_{20}^i \\ p_1^i & p_2^i & \ldots & p_{20}^i \end{pmatrix}, \qquad i, j \in \{1, 2, \ldots, 20\}$$

gdje broj u sufiksu pokraj oznake slučajne varijable označava redni broj aminokiseline iz niza prostora aminokiselina.

Sada možemo izračunati izraz s desne strane zadnje jednakosti. Neka su a_i^k i a_j^k neke dvije aminokiseline iz distribucije A_k. Tada vrijedi:

$$\mathbb{E}\left[\left(a_i^k - a_j^k\right)^2\right] = \sum_{i,j=1}^{20} \left(a_i^k - a_j^k\right)^2 p_j^k p_j^k$$

Distribuciju A_k određuje aminokiselina koja je odabrana s vjerojatnošću pojavljivanja te aminokiseline u prostoru proteina kojeg promatramo pa slijedi da je očekivanje za prosječnu distribuciju jednako:

$$\mathbb{E}\left[\left(\bar{a}_i - \bar{a}_j\right)^2\right] = \sum_{k=1}^{20} p_k \sum_{i,j=1}^{20} \left(a_i^k - a_j^k\right)^2 p_j^k p_j^k = 10.8724$$

Sada zaključujemo da je očekivani kvadrat udaljenosti neke dvije desetorke aminokiselina jednak $10 \cdot 10.8724$, korjenovanjem dobijemo da je očekivana udaljenost jednaka $\sqrt{10} \cdot 3.2973$. Kako aminokiseline mogu biti i bliže, gornji rezultat možemo interpretirati kao maksimalnu udaljenost za dvije točke koje prikazuju desetorku aminokiselina. Sada iz teorema 3.0.2 slijedi da je $r = \frac{\sqrt{10} \cdot 3.2973}{\sqrt{2}} = 3.2973 \cdot \sqrt{5}$

Uočimo da je radijus proporcionalan standardnoj devijaciji što povlači da je radijus nakon transformacije uzorka proporcionalan standardnoj devijaciji prije i nakon transformacije pa dobijemo sljedeću jednakost:

$$r_{new} = r_{old} \frac{std_{new}}{std_{old}}$$

pri čemu indeksi *new* i *old* označavaju podatke za uzorak nakon odnosno prije standardizacije, pri čemu je r_{old} procijenjeni radijus $3.2973 \cdot \sqrt{5}$.

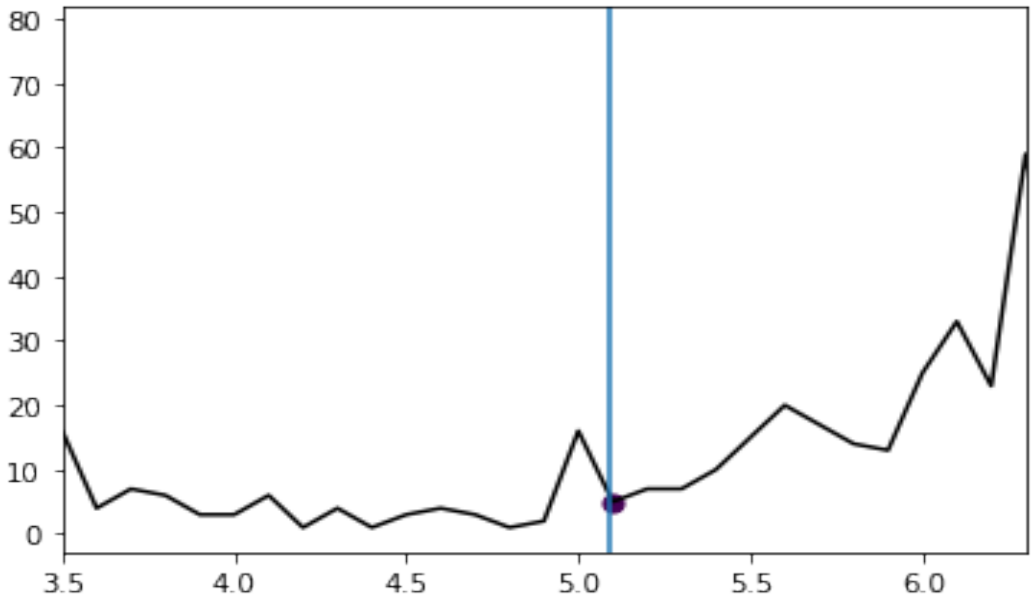
Nakon pronalaska radijusa tražene kugle, preostalo nam je pronaći i njeno središte. Metoda pronalaska središta opisana je u sljedećem poglavlju.

3.1 Metoda

Metoda je u cijelosti implementirana u programskom jeziku *Python*

Težište svih pravih pozitivaca

Odgovor IGLOSS algoritma ima određeni broj lažnih pozitivaca koje želimo isključiti iz kugle. Uočili smo kako su, za standardizirane podatke, biološki pozitivci distribuirani u blizini upita, a da su dalje od središta raspršeni oni koje je IGLOSS pogrešno ocijenio. Dakle, za uspješnost algoritma bi bilo idealno pronaći težište pravih pozitivaca te tu točku uzeti za središte kugle koju tražimo, a zatim testirati procjenu radijusa. Naši podaci nam to omogućavaju jer znamo koji su pravi a koji lažni pozitivci stoga ćemo pogledati kako bi bilo u idealnom slučaju. Kugla koji smo dobili sadrži preko 95% pravih pozitivaca. Na grafu niže vidimo crnu točku koja prikazuje radijus maksimalnog $F_1 - scorea$ za kuglu sa središtem u težištu pravih pozitivaca, te plavu vertikalnu liniju koja siječe graf u točki $F_1 - scorea$ za kuglu procijenjenog radijusa i središta u težištu pravih pozitivaca. Uočavamo da plava vertikalna linija prelazi preko točke što znači da smo našom kuglom postigli maksimalni $F_1 - score$.



Slika 3.1: Prikaz optimalnog radijusa sa središtem u težištu svih pravih pozitivaca

Problem se javlja kada u podacima ne znamo koji su pravi pozitivci a koji lažni pa smo razvili metodu u kojoj nam ta informacija nije potrebna.

Metoda prolaska po svim točkama

Rješavanju ranije spomenutog problema, za podatke u kojima ne znamo koji su pravo a koji lažno pozitivni moramo pristupiti na drugačiji način. Pretpostavka je da se u kugli s procijenjenim radijusom, koja sadrži najveći broj točaka, nalazi najveći broj pravih pozitivaca a malen broj lažnih pozitivaca. Vođeni tom pretpostavkom, metoda prolazi kroz sve točke te oko svake radi kuglu procijenjenog radijusa. Kao rezultat daje najgušću kuglu, to jest onu s najviše točaka u njoj.

Na početku metode važno je definirati radijus koji smo procijenili kao što je opisano ranije. Zatim, prolazimo kroz sve točke i računamo udaljenost od trenutne do svake druge u podacima, te one koje su udaljene manje od radijusa stavljamo u kuglu. Važan korak je zapamtiti koja od svih tih kugla ima najveći broj točaka jer će se po pretpostavci u njoj nalaziti najveći broj pravih pozitivaca. Tim dijelom algoritma smo došli do kugle koja sadrži maksimalan broj točaka ali središte nam je upitno.

Može se dogoditi da točke nisu ravnomjerno raspršene po kugli već se grupiraju na nekom dijelu kugle pa to želimo riješiti pomicanjem središta. Središte ćemo pomaknuti tako da uzmemo aritmetičku sredinu točaka iz kugle i tu točku postavimo kao središte nove, optimalnije kugle. Taj korak ponavljamo do trenutka kada algoritam ne počne vračati istu vrijednost. Drugi mogući ishod je da algoritam počne prebacivati središte s jedne točke na drugu, time smo ušli u beskonačnu petlju, pa ga moramo zaustaviti i jednu od te dvije točke postaviti kao središte. Napravili smo 5 pomaka središta te uočili da tu metoda staje, odnosno tada počne vraćati isto središte. Na kraju metode smo provjerili udaljenost upita od središta kugle te dobili rezultate da se upit nalazi unutar kugle.

3.2 Primjeri i rezultati

Uspješnost modela ispitana je na četiri različita proteoma:

- Talijin uročnjak (lat. *Arabidopsis thaliana*)
- Rajčica (lat. *Solanum lycopersicum*)
- Krumpir (lat. *Solanum tuberosum*)
- Azijska riža (lat. *Oriza sativa*)

U svim primjerima korišten je upit FVFGDSLSDA. IGLOSS-ov odgovor su desetorke koje su *Condition Positives* (CP), za njih smatramo da pripadaju familiji GDSL lipaza. Za svaki od proteoma korištene su skale pretraživanja od 3 do 7. Svaki od motiva u nizu koji

smo dobili ima oznaku je li TP (*True Positive*) ili FP (*False Positive*) pa smo mjeru uspješnosti računali u odnosu na to. Rezultate metode opisane ranije smo usporedili s rezultatima algoritma pronalaska kugle procijenjenog radijusa te središta, opisanog u diplomskom radu Marka Ivekovića [4]. Za svaki proteom i svaku skalu pretraživanja navest ćemo broj nizova aminokiselina koje je model vratio kao odgovor (n), broj bioloških pozitivaca koje je model dao kao rezultat (TP), osjetljivost modela (TPR), preciznost (PPV), radijus te mjeru uspješnosti $F_1 - score$ koji smo izračunali s podacima kugle u odnosu na podatke odgovora.

Važno je napomenuti da se metoda kugle ne nadovezuje na algoritam traženja središta u radu [3] pa ne znamo kakve rezultate očekujemo.

Talijin uročnjak

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

Slika 3.2: Arabidopsis thaliana

1. Skala pretraživanja je 3.

Model	TPR	PPV	$F_1 - score$	n	TP	radijus
Kugla sa središtem u težištu svih TP	0.98	0.88	0.93	115	102	4.86
Metoda prolaska po svim točkama	0.89	0.86	0.88	107	93	4.86
Algoritam iz [4]	0.52	0.94	0.67	93	88	4.86

2. Skala pretraživanja je 4.

Model	TPR	PPV	$F_1 - score$	n	TP	radijus
Kugla sa središtem u težištu svih TP	0.98	0.89	0.93	114	102	5.02
Metoda prolaska po svim točkama	0.88	0.92	0.90	100	92	5.02
Algoritam iz [4]	0.52	0.92	0.66	95	88	5.02

3. Skala pretraživanja je 5.

Model	TPR	PPV	$F_1 - score$	n	TP	radijus
Kugla sa središtem u težištu svih TP	0.96	0.91	0.93	109	100	5.29
Metoda prolaska po svim točkama	0.94	0.90	0.92	108	98	5.29
Algoritam iz [4]	0.55	0.92	0.68	101	93	5.29

4. Skala pretraživanja je 6.

Model	TPR	PPV	$F_1 - score$	n	TP	radijus
Kugla sa središtem u težištu svih TP	0.75	0.85	0.79	91	78	5.7
Metoda prolaska po svim točkama	0.73	0.85	0.78	89	76	5.7
Algoritam iz [4]	0.42	0.88	0.57	80	71	5.7

5. Skala pretraživanja je 7.

Model	TPR	PPV	$F_1 - score$	n	TP	radijus
Kugla sa središtem u težištu svih TP	0.79	0.87	0.79	87	76	6.0
Metoda prolaska po svim točkama	0.73	0.87	0.79	87	76	6.0
Algoritam iz [4]	0.39	0.88	0.54	76	67	6.0

Rajčica

Rajčica (lat. *Solanum lycopersicum*) je jednogodišnje povrće razgranate, zeljaste stabljike. Uzgaja se na svim kontinentima te je važna prehrambena namirnica.

1. Skala pretraživanja je 3.

Model	TPR	PPV	$F_1 - score$	n	TP	radijus
Kugla sa središtem u težištu svih TP	0.80	0.90	0.85	96	87	4.9
Metoda prolaska po svim točkama	0.70	0.89	0.78	85	76	4.9
Algoritam iz [4]	0.40	0.91	0.55	74	68	4.9

2. Skala pretraživanja je 4.

Model	TPR	PPV	$F_1 - score$	n	TP	radijus
Kugla sa središtem u težištu svih TP	0.78	0.93	0.85	91	85	5.05
Metoda prolaska po svim točkama	0.68	0.92	0.78	80	74	5.05
Algoritam iz [4]	0.37	0.94	0.53	67	63	5.05

3. Skala pretraživanja je 5.

Model	TPR	PPV	$F_1 - score$	n	TP	radijus
Kugla sa središtem u težištu svih TP	0.79	0.93	0.85	82	86	5.31
Metoda prolaska po svim točkama	0.68	0.93	0.79	79	74	5.31
Algoritam iz [4]	0.36	0.93	0.52	66	62	5.31

4. Skala pretraživanja je 6.

Model	TPR	PPV	$F_1 - score$	n	TP	radijus
Kugla sa središtem u težištu svih TP	0.81	0.95	0.88	92	88	5.64
Metoda prolaska po svim točkama	0.80	0.96	0.87	90	87	5.64
Algoritam iz [4]	0.45	0.97	0.62	79	77	5.64

5. Skala pretraživanja je 7.

Model	TPR	PPV	$F_1 - score$	n	TP	radijus
Kugla sa središtem u težištu svih TP	0.80	0.97	0.88	89	87	6.08
Metoda prolaska po svim točkama	0.80	0.98	0.88	88	87	6.08
Algoritam iz [4]	0.48	0.98	0.65	83	82	6.08

Krumpir

Krumpir (lat. *Solanum tuberosum*) je trajna biljka iz porodice pomoćnica. Uzgaja se na svim kontinentima te je s približno 18 milijuna hektara površine, četvrta kultura u svijetu.

Slika 3.3: Solanum tuberosum

1. Skala pretraživanja je 3.

Model	TPR	PPV	$F_1 - score$	n	TP	radijus
Kugla sa središtem u težištu svih TP	0.71	0.87	0.78	101	88	4.89
Metoda prolaska po svim točkama	0.65	0.86	0.74	93	80	4.89
Algoritam iz [4]	0.39	0.88	0.54	75	66	4.89

2. Skala pretraživanja je 4.

Model	TPR	PPV	$F_1 - score$	n	TP	radijus
Kugla sa središtem u težištu svih TP	0.69	0.90	0.78	95	86	4.95
Metoda prolaska po svim točkama	0.65	0.89	0.75	89	80	4.95
Algoritam iz [4]	0.40	0.89	0.55	76	68	4.95

3. Skala pretraživanja je 5.

Model	TPR	PPV	$F_1 - score$	n	TP	radijus
Kugla sa središtem u težištu svih TP	0.70	0.90	0.79	96	87	5.3
Metoda prolaska po svim točkama	0.64	0.89	0.74	88	79	5.3
Algoritam iz [4]	0.37	0.88	0.52	71	63	5.3

4. Skala pretraživanja je 6.

Model	TPR	PPV	$F_1 - score$	n	TP	radijus
Kugla sa središtem u težištu svih TP	0.74	0.92	0.82	100	92	5.79
Metoda prolaska po svim točkama	0.73	0.91	0.81	99	91	5.79
Algoritam iz [4]	0.45	0.90	0.60	85	77	5.79

5. Skala pretraživanja je 7.

Model	TPR	PPV	$F_1 - score$	n	TP	radijus
Kugla sa središtem u težištu svih TP	0.72	0.91	0.80	97	89	6.25
Metoda prolaska po svim točkama	0.72	0.91	0.80	97	89	6.25
Algoritam iz [4]	0.46	0.90	0.61	86	78	6.25

Azijska riža

Azijska riža (lat. *Oryza sativa*) je najpoznatija od 19 vrsta riže. To je žitarica iz porodice trava raširena pretežno po tropskim i suptropskim predjelima Azije i Afrike.

1. Skala pretraživanja je 3.

Model	TPR	PPV	$F_1 - score$	n	TP	radijus
Kugla sa središtem u težištu svih TP	0.88	0.90	0.89	114	103	5.09
Metoda prolaska po svim točkama	0	0	0	718	0	5.09

2. Skala pretraživanja je 4.

Model	TPR	PPV	$F_1 - score$	n	TP	radijus
Kugla sa središtem u težištu svih TP	0.81	0.93	0.87	102	95	5.40
Metoda prolaska po svim točkama	0	0	0	719	0	5.40

3. Skala pretraživanja je 5.

Model	TPR	PPV	$F_1 - score$	n	TP	radijus
Kugla sa središtem u težištu svih TP	0.77	0.94	0.85	95	90	5.9
Metoda prolaska po svim točkama	0	0	0	714	0	5.9

4. Skala pretraživanja je 6.

Model	TPR	PPV	$F_1 - score$	n	TP	radijus
Kugla sa središtem u težištu svih TP	0.93	0.91	0.92	118	108	5.73
Metoda prolaska po svim točkama	0.92	0.91	0.91	117	107	5.73

5. Skala pretraživanja je 7.

Model	TPR	PPV	$F_1 - score$	n	TP	radijus
Kugla sa središtem u težištu svih TP	0.92	0.93	0.92	115	107	6.05
Metoda prolaska po svim točkama	0.90	0.91	0.90	115	105	6.05

Analiza rezultata

U slučajevima talijinog uročnjaka, rajčice i krumpira, metoda prolaska po svim točkama i pronalaska najgušće kugle u podacima daje kuglu s vrlo visokim $F_1 - scoreom$ koji je manji za od 0 do 5 posto u odnosu na kuglu određenu sa središtem u težištu svih TP. IGLOSS odgovor prepoznaje više TP od algoritma kugle, no to je i očekivano s obzirom na to da algoritam kugle samo smanjuje broj podataka koje je IGLOSS dao kao odgovor. Primjećujemo da je uspješnost algoritma ograničena brojem pozitivaca koje IGLOSS uspije prepoznati. U većini slučajeva, povećavanjem skale, poveća se i očekivani radijus no nema pravila za povećanje odnosno smanjivanje $F_1 - scorea$. Osim očito velike vrijednosti $F_1 - scorea$ koja pokazuje da je metoda vrlo uspješna, pogledamo li PPV (vrijednost preciznosti) također uočavamo visoke vrijednosti što pokazuje da je uspješna u odbacivanju lažnih pozitivaca, bez isključivanja značajnog broja pravih pozitivaca. Provjerili smo i udaljenost upita od središta kugle te vidjeli da se upit nalazi unutar kugle što također govori o uspješnosti pronalaska motiva sličnih upitu. Još jedna važna stavka je robusnost. Uočavamo da se u metodi kugle $F_1 - score$ značajno ne mijenja u odnosu na skale, dok u IGLOSS algoritmu povećanjem broja podataka dolazi do značajnog pada $F_1 - scorea$.

U odnosu na algoritam koji je obrađen u radu [4], došli smo do poboljšanja algoritma za 20-30 posto što nam govori da središte određeno prolaskom po svim točkama i pomaknuto na težište najgušće kugle daje bolje rezultate. Osim drugačijeg pronalaska središta, transformacija podataka je napravljena na drugačiji način što također poboljšava uspješnost algoritma.

Za razliku od talijinog uročnjaka, rajčice i krumpira, podaci dobiveni iz proteoma riže ne daju dobre rezultate. Podaci se grupiraju u kuglu prosječnog radijusa između 5 i 6, kao i u drugim podacima, ali u najgušćoj kugli procijenjenog radijusa se ne nalazi niti jedan TP. Uočavamo da povećavanjem skale tj. smanjenjem broja podataka koje nam IGLOSS da kao odgovor, dolazi do poboljšanja uspješnosti metode. Na nižim skalama, točnije na skalama 3, 4 i 5, najgušća kugla sadrži prosječno 715 proteina ali niti jedan nije TP dok veće skale daju rezultate jednako dobre kao i kod drugih biljaka.

Važno je napomenuti da metoda kugle daje podatke koji u prosjeku sadrže 95% pravih pozitivaca što je veliko poboljšanje u odnosu na IGLOSS koji prosječno daje 15%-20%. Također, kugla ne samo da se većinski sastoji od TP, već obuhvaća minimalno 90% pravih pozitivaca koje je dao IGLOSS, stoga možemo zaključiti da smo bez pretpostavke i znanja

o tome koji su proteini TP a koji FP došli do izvrsnih rezultata. Metoda je nenadziranom klasifikacijom uspješno pronašla kuglu s velikim brojem TP i visokim $F_1 - scoreom$.

Bibliografija

[1] W. R. Atchley, J.Zhao, A.D. Fernandes, T. Drüke, *Solving the protein sequence metric problem*. Proc. Natlc., Acad. Sci. USA 2005., 102 (18) 6395-6400.

[2] D. Bakić, *Linearna algebra*, Školska knjiga, Zagreb, 2008.

[3] V. Bokšić, *Proteinski motivi i klasifikacija*, Diplomski rad, Sveučilište u Zagrebu, Prirodoslovno-matematički fakultet (Matematički odsjek), 2021.

[4] M. Iveković, *Traženje proteinskih motiva i klasifikacija*, Diplomski rad, Sveučilište u Zagrebu, Prirodoslovno-matematički fakultet (Matematički odsjek), 2022.

[5] M. Huzak, *Vjerojatnost i matematička statistika*, predavanja, 2006., dostupno na `http://aktuari.math.pmf.unizg.hr/docs/vms.pdf`.

[6] Maurice George Kendall, Patrick Alfred Pierce Moran, *Geometrical probability*, Hafner Publishing Company, 1963, London

[7] Braslav Rabar, Keti Nižetić, Maja Zagorščak, Kristina Gruden, Pavle Goldstein, *A Clique-Based Method for Improving Motif Scanning Accuracy*, University of Zagreb, Faculty of Science, Mathematics Department and National Institute of Biology, Department of Biotechnology and Systems Biology

[8] B. Rabar, M. Zagorščak, S. Ristov, M. Rosenzweig i P. Goldestein, *IGLOSS: iterative gapeless local similarity search*, Bioinformatics **35** (2019), br. 18, 3491-3492, ISNN 1367-4803, `https://academic.oup.com/bioinformatics/article/35/18/3491/5306940`.

[9] https://www.enciklopedija.hr/

Sažetak

Problematika ovog rada je nenadzirana klasifikacija niza motiva nekog proteina. Prelaskom u vektorski prostor želi se motive klasificirati ovisno o tome jesu li dovoljno slični upitu ili nisu.

Razvijena metoda traži najgušću kuglu sa zadanim radijusom. Testiranje točnosti radijusa je provedeno nad podacima u kojima se zna je li motiv TP ili nije. Određeno je težište svih TP, ta točka je uzeta kao središte kugle zadanog radijusa te rezultati pokazuju da je radijus zaista optimalan. Zatim, prolaskom po svim točkama te traženjem najgušće kugle u podacima, dobije se kugla s neravnomjerno raspoređenim podacima unutar nje. Kako bi podaci bili ravnomjerno raspršeni oko središta, središte se pomakne na težište točaka koje se nalaze unutar kugle. Taj korak se ponavlja 5 puta te metoda staje pronalaskom težišta koje se sljedećom iteracijom ne pomiče.

Summary

This thesis is concerned with unsupervised classification of protein motifs. Using standard methods, the problem is reduced to some geometrical considerations - searching for the optimal n-dimensional ball - in a certain Euclidean space.

The radius of the n-ball has been derived previously, and we test that it is, indeed, the optimal solution. We develop a method for optimizing the center of the ball and test it on four standard plant proteomes, with very good results.

Životopis

Rođena sam 2. svibnja 1996. godine u Zagrebu. Osnovnu i srednju školu završila sam u Zagrebu te 2015. godine upisala Prirodoslovno-matematički fakultet Sveučilišta u Zagrebu. 2018. godine završavam preddiplomski sveučilišni studij matematike, nastavnički smjer te iste godine, na istom fakultetu, upisujem diplomski studij Matematička statistika kojeg završavam obranom ovog rada.