Klasifikacija proteinskih fragmenata

Radnić, Josipa

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Josipa Radnić

KLASIFIKACIJA PROTEINSKIH FRAGMENATA

Diplomski rad

Voditelj rada:
doc. dr. sc. Pavle Goldstein

Zagreb, ožujak 2023.

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

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

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Zahvaljujem mentoru doc. dr. sc. Pavlu Goldsteinu na velikom strpljenju i suradnji pri pisanju ovog rada, na svim satima provedenim u raspravama i na uloženom vremenu i trudu. Hvala na velikoj podršci da rad bude što kvalitetniji.

Neizmjerno hvala mojim roditeljima i sestrama koji su bili bezuvjetna podrška za svaki uspon i pad kroz cijeli život. Hvala za svaku riječ, svaki zagrljaj i svaki dolazak u Zagreb koji nisam bila svjesna koliko mi je potreban.

Također, veliko hvala rodicama i prijateljima koji su uvijek bili uz mene i dijelili sve lijepe i tmurne dane sa mnom i putovali ovim stazama uz mene.

Hvala i svima onima koji se nisu pronašli u ovim zahvalama, a sigurno zaslužuju čuti riječ hvala.

Život je sa svima vama bio ljepši i lakši, i hvala vam na tome!

Sadržaj

Uvod

Proteom je skup svih proteina nekog organizma. Proteini su složene molekule, sastavljene od aminokiselina, koje su sastavni dio stanice svih živih bića. Obavljaju širok spektar funkcija unutar organizama, uključujući kataliziranje metaboličkih reakcija, replikaciju DNA, reagiranje na podražaje, osiguravanje strukture stanicama i organizmima i prijenos molekula s jednog mjesta na drugo. Njihova povezanost s osnovnim životnim svojstvima jedinki razlog je proučavanja i određivanja pripadnosti proteinskim familijama.

U proteinskoj familiji nalaze se proteini koji imaju zajedničko evolucijsko podrijetlo, što se odražava u njihovim srodnim funkcijama i sličnostima u strukturi. Problem traženja proteina koji pripadaju istoj proteinskoj familiji jedno je od glavnih istraživanja u bioinformatici. Zbog velike količine podataka o proteinima dobivenih sekvenciranjem genoma, postoji potreba za pouzdanim automatskim metodama za analizu i klasifikaciju sekvenci proteina. Iterativno pretraživanje proteoma je standardna metoda klasifikacije koja se bazira na konceptu sličnosti. Zadavanjem upita, koji je karakterističan niz aminokiselina za proteinsku familiju od interesa, pronalazi se skup proteina koji pripadaju istoj proteinskoj familiji, ali uspješnost je ograničena.

U ovom radu istražujemo način na koji bi se povećala točnost standardnog modela pretraživanja. Nakon što iterativnom metodom dobijemo kandidate za traženu proteinsku familiju, primjenjujemo metodu koja traži kuglu s najviše proteina najsličnijih upitu na temelju bliskosti proteina.

Rad se sastoji od tri poglavlja. U prvom poglavlju navedeni su matematički pojmovi iz linearne algebre, vjerojatnosti i statistike važni za razumijevanje rada. Ujedno su definirane mjere uspješnosti. U drugom poglavlju navedeni su pojmovi iz bioinformatike te je objašnjeno iterativno pretraživanje proteoma i prelazak u vektorski prostor. Konačno, u posljednjem poglavlju opisuje se problem i ideja te algoritam koji pospješuje iterativni model. Također, prikazani su numerički i grafički rezultati istraživanja.

Poglavlje 1

Matematički pojmovi

U ovom poglavlju navode se teoremi, definicije, propozicije i napomene iz linearne algebre, vjerojatnosti i statistike te uspješnosti modela. Pojmovi su preuzeti iz izvora [2], [3], [4], [5], [8], [9] i [10].

1.1 Linearna algebra

Definicija 1.1.1. *Neka je* $\mathbb{F}$ *neki skup na kojem su definirane 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****, a elemente polja nazivamo skalarima.*

Napomena 1.1.2. *Skup realnih brojeva $\mathbb{R}$ s uobičajenim operacijama zbrajanja i množenja je polje.*

Definicija 1.1.3. *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 1.1.4. *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 1.1.5. *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 1.1.6. *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 1.1.7. *Vektorski prostor na kojem je definiran skalarni produkt zove se* ***unitarni prostor****.*

Definicija 1.1.8. *Neka je V unitaran prostor.* ***Norma*** *na V je funkcija $\|\cdot\| : V \to \mathbb{R}$ definirana s*

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

Propozicija 1.1.9. *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 1.1.10. *Svako preslikavanje $\|\cdot\| : V \to \mathbb{R}$ na vektorskom prostoru V sa svojstvima iz propozicije 1.1.9 naziva se norma. Tada $(V, \|\cdot\|)$ zovemo* ***normirani prostor****.*

Definicija 1.1.11. *Norma koja potječe od kanonskog skalarnog produkta na $\mathbb{R}^n$, definiranog u napomeni 1.1.6, dana je formulom*

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

Ova norma se zove ***euklidska norma****.*

Definicija 1.1.12. *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 1.1.13. *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 1.1.14. *Neka je $X \neq \emptyset$. Svaka funkcija $d : X \times X \to \mathbb{R}$ sa svojstvima iz propozicije 1.1.13 naziva se metrika ili udaljenost. Tada (X, d) zovemo* ***metrički prostor****.*

Definicija 1.1.15. *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 1.1.11, 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* $\mathbb{R}^n$ *zajedno s tom metrikom nazivamo* ***euklidski prostor****.*

Definicija 1.1.16. *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, s centrom a i radijusom r.*

Definicija 1.1.17. *U euklidskom prostoru* $\mathbb{R}^n$ *otvorena kugla s centrom* $a \in \mathbb{R}^n$ *i radijusom* $r > 0 \in \mathbb{R}$ *dana je s*

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

1.2 Vjerojatnost i statistika

Vjerojatnosni prostor

Definicija 1.2.1. ***Slučajni pokus*** *ili* ***slučajni eksperiment*** *je pokus čiji ishodi nisu jednoznačno određeni.*

Definicija 1.2.2. ***Prostor elementarnih događaja*** Ω *je neprazan skup koji reprezentira skup svih ishoda slučajnog pokusa. Elemente* ω *skupa* Ω *nazivamo* ***elementarni događaji****.*

Definicija 1.2.3. *Familija* $\mathcal{F}$ *podskupova od* Ω ($\mathcal{F} \subset \mathcal{P}(\Omega)$) *je* ***$\sigma$-algebra skupova*** *na* Ω *ako je:*

1) $\emptyset \in \mathcal{F}$*;*

2) $A \in \mathcal{F} \implies A^c \in \mathcal{F}$*;*

3) $A_i \in \mathcal{F}, i \in \mathbb{N} \implies \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

Definicija 1.2.4. *Neka je* $\mathcal{F}$ *σ-algebra na skupu* Ω*. Uređen par* $(\Omega, \mathcal{F})$ *zove se* ***izmjeriv prostor****.*

Definicija 1.2.5. *Neka je* $(\Omega, \mathcal{F})$ *izmjeriv prostor. Funkcija* $\mathbb{P} : \mathcal{F} \to \mathbb{R}$ *je* ***vjerojatnost*** *(na* $\mathcal{F}$*, na* Ω*) ako vrijedi:*

1) $\mathbb{P}(A) \geq 0, \forall A \in \mathcal{F}$*;*

2) $\mathbb{P}(\Omega) = 1$*;*

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

Definicija 1.2.6. *Uređena trojka* $(\Omega, \mathcal{F}, \mathbb{P})$*, gdje je* $\mathcal{F}$ *σ-algebra na* Ω*, a* $\mathbb{P}$ *je vjerojatnost na* $\mathcal{F}$*, zove se* ***vjerojatnosni prostor****.*

Slučajna varijabla

Definicija 1.2.7. *Neka je* S *proizvoljan neprazan skup i* $\mathcal{A}$ *familija podskupova od* S ($\mathcal{A} \subset \mathcal{P}(S)$)*. Sa* $\sigma(\mathcal{A})$ *označimo najmanju* σ*-algebru podskupova od* S *koja sadrži* $\mathcal{A}$*. Nju nazivamo* ***σ-algebra generirana sa*** $\mathcal{A}$*.*

Definicija 1.2.8. *Neka je s $\mathcal{B}$ označena σ-algebra generirana familijom svih otvorenih skupova na $\mathbb{R}$. $\mathcal{B}$ zovemo **σ-algebra Borelovih skupova na** $\mathbb{R}$, a elemente σ-algebre $\mathcal{B}$ zovemo **Borelovi skupovi**.*

Definicija 1.2.9. *Neka je $(\Omega, \mathcal{F}, \mathbb{P})$ 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 1.2.10. *Neka je $(\Omega, \mathcal{F}, P)$ vjerojatnosni prostor i $X : \Omega \to \mathbb{R}^n$. Kažemo da je X **n-dimenzionalan slučajan vektor** (ili, kraće, **slučajan vektor**) (na Ω) ako je $X^{-1}(B) \in \mathcal{F}$ za svako $B \in \mathcal{B}^n$, tj. $X^{-1}(\mathcal{B}^n) \subset \mathcal{F}$.*

Definicija 1.2.11. *Neka je X slučajna varijabla na $(\Omega, \mathcal{F}, P)$. X je **jednostavna slučajna varijabla** ako je njezino područje vrijednosti konačan skup.*

X je jednostavna slučajna varijabla ako i samo ako je

$$X = \sum_{k=1}^{n} x_k \mathcal{K}_{A_k},$$

gdje su $x_1, x_2, \ldots, x_n$ realni brojevi, a $A_1, A_2, \ldots, A_n$ međusobno disjunktni događaji, $\bigcup_{k=1}^{n} A_k = \Omega$. $\mathcal{K}_{A_k}$ označava **karakterističnu funkciju** skupa A_k.

Neka su $X_1, X_2 : \Omega \to \mathbb{R}$. Tada definiramo funkcije $X_1 \vee X_2$ i $X_1 \wedge X_2$ na Ω, relacijama:

$$(X_1 \vee X_2)(\omega) = max\{X_1(\omega), X_2(\omega)\}, \ \omega \in \Omega, \tag{1.1}$$

i

$$(X_1 \wedge X_2)(\omega) = min\{X_1(\omega), X_2(\omega)\}, \ \omega \in \Omega.$$

Pomoću funkcije (1.1) definiramo pozitivan i negativan dio realne funkcije X na Ω:

$$X^+ = X \vee 0, \ X^- = (-X) \vee 0.$$

X^+ i X^- su nenegativne realne funkcije i vrijedi:

$$X = X^+ - X^-$$

$$|X| = X^+ + X^-.$$

Korolar 1.2.12. *X je slučajna varijabla ako i samo ako su X^+ i X^- slučajne varijable.*

Teorem 1.2.13. *Neka je X nenegativna slučajna varijabla na Ω. Tada postoji rastući niz $(X_n, n \in \mathbb{N})$ nenegativnih jednostavnih slučajnih varijabli takav da je $X = \lim_{n\to\infty} X_n$ (na Ω).*

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 općenite slučajne varijable.
Neka je $(\Omega, \mathcal{F}, \mathbb{P})$ vjerojatnosni prostor. Označimo s $\mathcal{K}$ skup svih jednostavnih slučajnih varijabli definiranih na Ω, a s $\mathcal{K}_+$ 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 1.2.14. ***Matematičko očekivanje*** *od X ili, kraće,* ***očekivanje*** *od X označavamo s* $\mathbb{E}[X]$ *i definira se s:*

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

Propozicija 1.2.15. *1. Neka je* $c \in \mathbb{R}$ *i* $X \in \mathcal{K}$. *Tada je* $\mathbb{E}(cX) = c\mathbb{E}X$.

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 sada X **nenegativna slučajna varijabla** definirana na Ω. Prema teoremu 1.2.13 postoji rastući niz $(X_n)_{n\in\mathbb{N}}$ nenegativnih jednostavnih slučajnih varijabli takav da je $X = \lim_{n\to\infty} X_n$. Iz prethodne propozicije 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 1.2.16. ***Matematičko očekivanje*** *od X ili, kraće,* ***očekivanje*** *od X definira se s*

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

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

Definicija 1.2.17. *Kažemo da* ***matematičko očekivanje*** *od X ili, kraće,* ***očekivanje*** *od X postoji (ili 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*

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

Definicija 1.2.18. *Neka je X slučajna varijabla na* $(\Omega, \mathcal{F}, \mathbb{P})$ *i neka je* $\mathbb{E}[X]$ *konačno. Tada definiramo* ***varijancu*** *od X koju označavamo s* $Var(X)$ *ili* σ_X^2 *na sljedeći način:*

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

Napomena 1.2.19. *Pozitivan drugi korijen iz varijance nazivamo* ***standardna devijacija*** *i označavamo sa* σ_X.

Funkcija distribucije

Definicija 1.2.20. *Neka je X slučajna varijabla na* Ω. ***Funkcija distribucije od X*** *je funkcija* $F_X : \mathbb{R} \to [0,1]$ *definirana s:*

$$F_X(x) = \mathbb{P}(X^{-1}(\langle -\infty, x])) = \mathbb{P}\{\omega \in \Omega : X(\omega) \leq x\} = \mathbb{P}\{X \leq x\},\ x \in \mathbb{R}.$$

Napomena 1.2.21. *Ako je jasno o kojoj se slučajnoj varijabli, odnosno njenoj funkciji distribucije, radi piše se F umjesto* F_X.

Teorem 1.2.22. *Funkcija distribucije F slučajne varijable X je rastuća i neprekidna zdesna na* $\mathbb{R}$ *te zadovoljava:*

$$F(-\infty) = \lim_{x \to -\infty} F(x) = 0$$
$$F(+\infty) = \lim_{x \to +\infty} F(x) = 1.$$

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

Definicija 1.2.23. *Funkcija* $g : \mathbb{R} \to \mathbb{R}$ *je* ***Borelova funkcija*** *ako je* $g^{-1}(B) \in \mathcal{B}$ *za svako* $B \in \mathcal{B}$, *tj. ako je* $g^{-1}(\mathcal{B}) \subset \mathcal{B}$.

Definicija 1.2.24. *Neka je* $(\Omega, \mathcal{F}, \mathbb{P})$ *vjerojatnosni prostor i X slučajna varijabla na* Ω. *Slučajna varijabla X je* ***diskretna*** *ako postoji konačan ili prebrojiv skup* $D \subset \mathbb{R}$ *takav da je* $\mathbb{P}\{X \in D\} = 1$.

Definicija 1.2.25. *Neka je X slučajna varijabla na vjerojatnosnom prostoru* $(\Omega, \mathcal{F}, \mathbb{P})$ *i neka je* F_X *njezina funkcija distribucije. Kažemo da je X* ***apsolutno neprekidna*** *ili, kraće,* ***neprekidna slučajna varijabla*** *ako postoji nenegativna realna Borelova funkcija f na* $\mathbb{R}$ *(*$f : \mathbb{R} \to \mathbb{R}_+$*) takva da je*

$$F_X(x) = \int_{-\infty}^{x} f(t) d\lambda(t),\ x \in \mathbb{R}. \tag{1.2}$$

Ako je X neprekidna slučajna varijabla, tada se funkcija f iz (1.2) zove **funkcija gustoće vjerojatnosti od X**, tj. od njezine funkcije distribucije F_X ili, kraće, **gustoća od X** i ponekad je označavamo s f_X.

Definicija 1.2.26. *Neka su* $\mu, \sigma \in \mathbb{R},\ \sigma > 0$. *Neprekidna slučajna varijabla X ima* ***normalnu distribuciju s parametrima*** μ *i* σ^2 *ako joj je gustoća f dana s*

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}},\ x \in \mathbb{R}.$$

To ćemo označavati s $X \sim N(\mu, \sigma^2)$.

Napomena 1.2.27. *Slučajna varijabla X ima* ***jediničnu normalnu distribuciju*** *ako je* $X \sim N(0,1)$*, dakle*

$$f(x) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{x^2}{2}},\ x \in \mathbb{R}.$$

Opisna analiza podataka

U ovom dijelu ćemo se podsjetiti definicija iz deskriptivne statistike koje će nam biti potrebne u daljnjem razumijevanju rada. Navodimo pojmove kao što su aritmetička sredina, standardna devijacija uzorka te varijanca uzorka i standardizacija podataka.

Neka su

$$x_1,\ x_2, \ldots,\ x_n \tag{1.3}$$

n 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 podataka ili uzorka (1.3) je mjera centralne tendencije i definirana je kao:

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

Varijanca uzorka ili podataka (1.3) 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:

$$x_i' = \frac{x_i - \overline{x}}{s}. \tag{1.4}$$

1.3 Klasifikacija i uspješnost modela

Klasifikacija

U statistici, **klasifikacija** je problem identificiranja pripadnosti opservacije nekoj od klasa ili kategorija. Razlikujemo nadziranu i nenadziranu klasifikaciju. U nadziranoj klasifikaciji su unaprijed određene klase na temelju skupa poznatih podataka i pridružujemo nove opservacije određenoj klasi pomoću zadane funkcije sličnosti. U nenadziranoj klasifikaciji nemamo unaprijed određenu pripadnost opservacije klasama. Model pokušava bez znanja o podacima i klasama uočiti strukturiranost među opservacijama i razdvojiti ih u kategorije po sličnosti. Osnovna razlika između nadzirane i nenadzirane klasifikacije jest što nadzirana zahtjeva unaprijed poznate označene podatke.

Mjere uspješnosti

Kako bi se ocijenila uspješnost nekog modela, definirane su mjere uspješnosti modela. One 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 1.1: Tablica uspješnosti

Napomena 1.3.1. *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 modelu bi svi proteini s liste CP imali oznaku 1, a svi proteini koji nisu na listi CP bi imali oznaku 0.*

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.

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

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.

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

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** ($\beta = 1$):

$$F_1 = \frac{2 \cdot PPV \cdot TPR}{PPV + TPR} \tag{1.5}$$

Napomena 1.3.2. *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 2

Bioinformatika

2.1 Biološki pojmovi

Proteini ili bjelančevine su prisutni u svim živim bićima i najvažnija su tvar u tijelu, što ih čini osnovom života na Zemlji. Proteini su glavni izvor tvari za rast i razvoj svih tjelesnih tkiva. Izgrađeni su od aminokiselina koje su međusobno povezane peptidnom vezom. Aminokiseline su organski spojevi koji sadrže amino i karboksilnu skupinu, te bočni lanac po kojem se međusobno razlikuju. Proteini su izgrađeni od 20 standardnih aminokiselina, svaka označena velikim slovom engleske abecede, prikazanih u tablici 2.1. Svojstva proteina određuju duljina lanca i raspored u njemu, a promjenom samo jedne karike u lancu nastat će nova bjelančevina.

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 2.1: Standardne aminokiseline

Skup svih proteina nekog organizma je proteom. Proteom se sastoji od različitih proteinskih familija od kojih je svaka zaslužna za određeno funkcijsko svojstvo organizma. Važnost određivanja pripadnosti proteina proteinskoj familiji je bitno za razumijevanje uloge proteina. Kada bismo uspjeli odrediti pripadnost proteina proteinskoj familiji znali bismo više o svojstvima jedinki, a time se otvara mogućnost unaprjeđenja određene vrste putem genetske modifikacije.

U ovom radu promatrat će se familija transkripcijskih faktora **MADS-box**. MADS-box je sačuvani motiv sekvence, a geni koji sadrže ovaj motiv nazivaju se obitelj gena MADS-box. Proteinska familija MADS-box dobila je ime kao akronim od četiri glavna člana familije koji su: MCM1 (iz *Saccharomyces cerevisiae*, pupajući kvasac), AGAMOUS (iz *Arabidopsis thaliana*, talijin uročnjak), DEFICIENS (iz *Antirrhinum majus*, "snapdragon") i SRF (iz *Homo Sapiens*).

MADS-box geni imaju različite funkcije. Kod životinja MADS-box geni uključeni su u razvoj mišića i staničnu proliferaciju i diferencijaciju, a u biljkama u kontroli svih glavnih aspekata razvoja, uključujući razvoj muškog i ženskog gametofita, razvoj embrija i sjemena, kao i razvoj korijena, cvijeta i ploda. Transkripcijski faktori ključne su komponente regulatornih mreža koje integriraju okolišne znakove i usklađene odgovore na staničnoj razini, uključujući one koji impliciraju stresno stanje. Razvojna i evolucijska važnost biljnih transkripcijskih faktora MADS domene posljedica je njihovog svestranog načina vezanja DNA i kombinatorne multimerizacije. Zbog toga je otkrivanje novih biljnih članova familije MADS-box od velikog interesa.

2.2 Iterativno pretraživanje proteoma

Iterativno pretraživanje proteoma je standardna metoda pronalaska proteina iz iste proteinske familije. Cilj iterativnog pretraživanja je da za određeni upit dobijemo nizove aminokiselina dovoljno slični zadanom upitu (s obzirom na određenu funkciju sličnosti). Time metoda kao ulazni parametar prima karakteristični motiv za proteinsku familiju od interesa. **Motiv** (ili upit) je niz aminokiselina duljine od 5 do 20 koji mutira na specifičan način. Ako protein sadrži dovoljno sličan niz tada ga algoritam svrstava u pripadnu proteinsku familiju. Pri svakoj iteraciji parametri funkcije sličnosti se mijenjaju i promatramo proteom sa skupom proteina koji su bili dovoljno slični u prethodnoj iteraciji. Iteriranje staje kada skup proteina - **odgovor**, ostaje nepromijenjen ili kad se dosegne maksimalan broj iteracija. Proteini su u familije svrstani s određenom uspješnosti.

U ovom radu za iterativno pretraživanje proteoma koristi se **IGLOSS server** opisan u izvoru [7]. Za funkciju sličnosti IGLOSS server koristi *log likelihood ratio* (LLR) koja je ocijenjena pomoću logističke distribucije. **Skala pretraživanja** je parametar koji postavlja granicu "dovoljne sličnosti". Odgovor čini skup proteina čija je sličnost veća ili jednaka od zadane skale pretraživanja. Što je skala veća više se kažnjava odstupanje od motiva pa su odabrani sličniji nizovi, i obrnuto. Slijedi da je broj podataka u odgovoru obrnuto proporcionalan skali pretraživanja. Za kraj ćemo definirati BLOSUM matricu (2.1) i BLOSUM score koji se koriste za ocjenu sličnosti dvaju nizova aminokiselina.

Definicija 2.2.1. ***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. Bazirana je na sljedećoj formuli:*

$$B(i,j) = \left\lfloor \log \frac{\mathbb{P}\left(a_i \leftrightarrow b_j \mid M\right)}{\mathbb{P}\left(a_i, b_j \mid R\right)} \right\rfloor, \quad a_i, b_j \in \mathcal{A}, \tag{2.1}$$

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 $\mathbb{P}(a_i, b_j \mid R) = \mathbb{P}(a_i \mid R) \cdot \mathbb{P}(b_j \mid R)$. Distribucija standardnih aminokiselina uz model R dana je s:

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

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	5	−2	−1	−2	−1	−1	−1	0	−2	−1	−2	−1	−1	−3	−1	1	0	−3	−2	0
R	−2	7	−1	−2	−4	1	0	−3	0	−4	−3	3	−2	−3	−3	−1	−1	−3	−1	−3
N	−1	−1	7	2	−2	0	0	0	1	−3	−4	0	−2	−4	−2	1	0	−4	−2	−3
D	−2	−2	2	8	−4	0	2	−1	−1	−4	−4	−1	−4	−5	−1	0	−1	−5	−3	−4
C	−1	−4	−2	−4	13	−3	−3	−3	−3	−2	−2	−3	−2	−2	−4	−1	−1	−5	−3	−1
Q	−1	1	0	0	−3	7	2	−2	1	−3	−2	2	0	−4	−1	0	−1	−1	−1	−3
E	−1	0	0	2	−3	2	6	−3	0	−4	−3	1	−2	−3	−1	−1	−1	−3	−2	−3
G	0	−3	0	−1	−3	−2	−3	8	−2	−4	−4	−2	−3	−4	−2	0	−2	−3	−3	−4
H	−2	0	1	−1	−3	1	0	−2	10	−4	−3	0	−1	−1	−2	−1	−2	−3	2	−4
I	−1	−4	−3	−4	−2	−3	−4	−4	−4	5	2	−3	2	0	−3	−3	−1	−3	−1	4
L	−2	−3	−4	−4	−2	−2	−3	−4	−3	2	5	−3	3	1	−4	−3	−1	−2	−1	1
K	−1	3	0	−1	−3	2	1	−2	0	−3	−3	6	−2	−4	−1	0	−1	−3	−2	−3
M	−1	−2	−2	−4	−2	0	−2	−3	−1	2	3	−2	7	0	−3	−2	−1	−1	0	1
F	−3	−3	−4	−5	−2	−4	−3	−4	−1	0	1	−4	0	8	−4	−3	−2	1	4	−1
P	−1	−3	−2	−1	−4	−1	−1	−2	−2	−3	−4	−1	−3	−4	10	−1	−1	−4	−3	−3
S	1	−1	1	0	−1	0	−1	0	−1	−3	−3	0	−2	−3	−1	5	2	−4	−2	−2
T	0	−1	0	−1	−1	−1	−1	−2	−2	−1	−1	−1	−1	−2	−1	2	5	−3	−2	0
W	−3	−3	−4	−5	−5	−1	−3	−3	−3	−3	−2	−3	−1	1	−4	−4	−3	15	2	−3
Y	−2	−1	−2	−3	−3	−1	−2	−3	2	−1	−1	−2	0	4	−3	−2	−2	2	8	−1
V	0	−3	−3	−4	−1	−3	−3	−4	−4	4	1	−3	1	−1	−3	−2	0	−3	−1	5

Slika 2.1: BLOSUM matrica

Definicija 2.2.2. ***BLOSUM score*** *s je rezultat koji odgovara sličnosti (ili povezanosti) dvaju nizova aminokiselina. Što je BLOSUM score veći, nizovi aminokiselina su sličniji. BLOSUM score dvaju nizova standardnih aminokiselina dobiva se zbrajanjem sličnosti pojedinačnih aminokiselina po poziciji, pri čemu su te sličnosti prethodno definirane BLOSUM matricom.*

2.3 Prelazak u vektorski prostor

Nedostatak prirodne metrike za usporedbu nizova slova spriječava obradu nad takvim podacima. Zbog toga se javlja potreba za opisom aminokiselina numeričkim vrijednostima. Navedena problematika je opisana i riješena u članku [1]. Definirano je preslikavanje u $\mathbb{R}^5$ koje svakoj aminokiselini pridružuje 5-dimenzionalni vektor. Preslikavanje "čuva" sve važne fizikalno-kemijske informacije o aminokiselini. Svaka koordinata vektora (*faktor*) opisuje jedno ili više svojstava odgovarajuće aminokiseline. *Faktor I* opisuje polaritet aminokiseline, *Faktor II* ima veze sa sekundardnom strukturom, *Faktor III* se odnosi na molekularni volumen, *Faktor IV* odražava raznolikost kodona (relativnu kompoziciju aminokiselina u različitim proteinima) te *Faktor V* opisuje elektrostatički naboj 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 2.2: Faktori

Nizu od n aminokiselina pridružujemo $5n$-dimenzionalni vektor. Nakon opisa aminokiselina numeričkim vrijednostima spremni smo za primjenu matematičkog alata.

Pojmovi iz ovog poglavlja preuzeti su iz izvora [1], [7], [8], [9] i [10].

Poglavlje 3

Analiza problema i rezultati

3.1 Opis problema i ideja

U ovom radu cilj je poboljšati uspješnost iterativnog pretraživanja proteoma. Cilj nije promijeniti metodu direktno nego u kombinaciji s iterativnim pretraživanjem riješiti problem pronalaska proteina koji pripadaju proteinskoj familiji od interesa. Nakon što iterativna metoda izbaci svoje kandidate za proteinsku familiju, među njima se nalaze proteini koji uistinu pripadaju familiji (eng. *true positives*) i oni koji ne pripadaju (eng. *false positives*). Iz odgovora, dobivenog putem IGLOSS servera, želimo eliminirati što više lažnih pozitivaca, a ujedno zadržati prave pozitivce. Kako bi to postigli koristit ćemo mogućnost prelaska u vektorski prostor gdje možemo promatrati distribuciju nizova u prostoru, njihove udaljenosti i iskoristiti matematičko znanje.

Reprezentacijom motiva točkama u višedimenzionalnom prostoru omogućava nam da postavimo i provjerimo dvije pretpostavke:

- pravi pozitivci se grupiraju u blizini upita (motiva) dok lažni pozitivci su razbacani i udaljeniji od upita
- pravi pozitivci se mogu smjestiti unutar kugle u $\mathbb{R}^{5n}$ koja sadrži upit (motiv)

U nastavku će se uspostaviti da su pretpostavke točne. Time smo problem sveli na traženje središta i radijusa kugle kojoj je mjera uspješnosti modela F_1 najveća.

Priprema podataka

Upit koji je korišten u ovom radu je RQVTFSKRRNGLLKKA. Upit je niz aminokiselina karakterističnih za MADS-box familiju. Navedeni upit se koristi kako bi uz pomoć IGLOSS servera dobili najbolje kandidate za MADS-box familiju. Odgovori će, kao i upit, biti nizovi aminokiselina duljine 16. Prelaskom u vektorski prostor dobiveni podaci su transformirani u 80-dimenzionalne vektore, odnosno nalazimo se u prostoru $\mathbb{R}^{80}$. Budući da promatramo euklidsku udaljenost između središta kugle i proteina, koja mjeri udaljenosti po pojedinim koordinatama i zbraja ih, želimo izbjeći da su varijanca i raspon podataka po jednoj koordinati veći od ostalih koordinata. Ako dođe do takve situacije, tada će euklidska udaljenost biti dominirana tom koordinatom i zbog toga gubimo formu kugle u kojoj bi sve koordinate trebale imati jednak utjecaj. Standardizacijom podataka se riješi potencijalni problem i dobijemo podatke gdje nam je utjecaj svih koordinata jednak. Standardizaciju smo prilagodili na način da dijelimo sa standardnom devijacijom uvećanom za 0.1 kako bismo izbjegli dijeljenje s brojem blizu nule.

Neka su $x_1, x_2, ..., x_n$ vrijednosti koje čine skup podataka tada je za $i = 1, 2, ..., n$:

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

Sada možemo tražiti "idealnu" kuglu za koju nema straha da jedna od koordinata izdominira udaljenost.

Kugla oko težišta pravih pozitivaca

Nakon što smo pripremili podatke za obradu možemo početi tražiti središte i radijus kugle koja će zadržati prave pozitivce, a eliminirati lažne pozitivce. Nakon što IGLOSS server izbaci popis pozitivaca, među kojima ima pravih i lažnih, želimo naći kuglu koja će dati što veći F_1 score. Kada bismo koristili činjenicu da znamo koji su pravi pozitivci ima smisla uzeti za središte kugle težište svih pravih pozitivaca. Nakon fiksiranja središta želimo pronaći optimalni radijus za koji dobivamo najveći F_1 score, odnosno uspješnost modela je najveća. Pronalazak optimalnog radijusa će se raditi iteracijom po vrijednostima od 2 do 9 s pomakom veličine 0.01 gdje u svakoj iteraciji izračuna F_1 score. Nakon prolaska po svim vrijednostima dobijemo za koji radijus je kugla sa središtem u težištu pravih pozitivaca ima najveći F_1 score. Kugla oko težišta pravih pozitivaca s optimalnim radijusom će se gledati kao "idealan" slučaj, odnosno bit će gornja granica za daljnje dobivene rezultate. Ako rezultati budu jako blizu "idealnom" slučaju znamo da je vrlo dobar.

Procijenjeni radijus kugle

U radu [10] je optimalni radijus nevjerojatno bio blizu vrijednosti procijenjenog radijusa te motivirano tim rezultatima također će se u ovom radu proučiti rezultati dobiveni s procijenjenim radijusom. Za razumijevanje sljedećih rezultata potrebni su sljedeći teoremi:

Teorem 3.1.1. *Površina kvadrata nad hipotenuzom pravokutnog trokuta jednaka je zbroju površina kvadrata nad njegovim katetama.*

Teorem 3.1.2. *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.*

Teorem 3.1.2 je detaljno opisan i obrađen u izvoru [12, str. 55], a teorem 3.1.1 je poznati Pitagorin teorem iz kojeg slijedi formula za udaljenost dviju točaka.

Pretpostavimo da se aminokiseline pojavljuju s vjerojatnostima $p_k,\ k \in \{1, 2, \ldots, 20\}$ zadanima u distribuciji aminokiselina navedenoj u poglavlju 2.2 te neka su $A_i, i \in \{1, 2, \ldots, 20\}$ distribucije zadane nekom aminokiselinom za koju ćemo pretpostaviti da je očuvana koeficijentom očuvanosti $\alpha = 0.68$. 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, a vjerojatnosti p_j^i su jednake

$$p_j^i = \alpha \cdot \mathbb{1}_{i=j} + (1 - \alpha) \cdot p_j$$

gdje broj u sufiksu pokraj vjerojatnosti p_j označava redni broj aminokiseline iz niza prostora aminokiselina.

Podaci su motivi duljine 16 pa računamo očekivanu udaljenost dvaju 16-dimenzionalnih vektora. Neka su $X = (x_1, x_2, \ldots, x_{16})$ i $Y = (y_1, y_2, \ldots, y_{16})$. Očekivanje kvadrata euklidske udaljenosti X i Y je:

$$\mathbb{E}\left[d^2(X, Y)\right] = \mathbb{E}\left[\left((x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_{16} - y_{16})^2\right)\right]$$

Iz svojstva očekivanja slijedi:

$$\mathbb{E}\left[d^2(X, Y)\right] = \sum_{k=1}^{16} \mathbb{E}\left[(x_k - y_k)^2\right]$$

S obzirom da nemamo nikakve pretpostavke o položaju aminokiselina po pozicijama u vektorima, možemo označiti s $\bar{a}_i$ i $\bar{a}_j$ aminokiseline koje pripadaju prosječnoj distribuciji aminokiselina te dobijemo:

$$\mathbb{E}\left[d^2(X,Y)\right] = \sum_{k=1}^{16} \mathbb{E}\left[\left(\bar{a}_i - \bar{a}_j\right)^2\right] = 16 \cdot \mathbb{E}\left[\left(\bar{a}_i - \bar{a}_j\right)^2\right]$$

Izračunat ćemo izraz s desne strane prethodne jednakosti. Neka su a_i^k i a_j^k neke dvije aminokiseline iz distribucije A_k. Tada vrijedi:

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

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:

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

Sada slijedi da je:

$$\mathbb{E}\left[d^2(X,Y)\right] = 16 \cdot \mathbb{E}\left[\left(\bar{a}_i - \bar{a}_j\right)^2\right] = 16 \cdot 10.8724$$

Zaključujemo da je očekivani kvadrat udaljenosti dva 16-dimenzionalna vektora aminokiselina jednak $16 \cdot 10.8724$ pa korjenovanjem dobijemo da je očekivana udaljenost jednaka $\sqrt{16} \cdot 3.2973$. Kako aminokiseline mogu biti i bliže, gornji rezultat možemo interpretirati kao maksimalnu udaljenost za dvije točke koje prikazuju 16-dimenzionalne vektore aminokiselina. Time udaljenost dvaju nizova aminokiselina duljine 16 ne bi smjela biti veća od dobivenog rezultata.
Sada iz teorema 3.1.2 slijedi da je $r = \frac{\sqrt{16} \cdot 3.2973}{\sqrt{2}} = \sqrt{8} \cdot 3.2973$

Uočimo da je radijus proporcionalan standardnoj devijaciji što povlači da je radijus prije i nakon standardizacije podataka proporcionalan standardnoj devijaciji prije i nakon standardizacije. Zbog toga imamo sljedeću jednakost:

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

gdje r_{new} označava traženi radijus, r_{old} radijus prije standardizacije podataka (izračunat gore), te s_{new} i s_{old} redom označavaju nakon, odnosno prije, standardizacije podatke.

Nakon izračunate procjene radijusa tražene kugle preostalo nam je pronaći središte bez poznavanja pravih pozitivaca. Time ne možemo kao središte staviti težište pravih pozitivaca pa ćemo pokušati pronaći središte kugle s procijenjenim radijusom metodom prolaska po svim točkama.

Metoda prolaska po svim točkama

Do sad je spomenuto način na koji ćemo naći "idealan" slučaj koji ćemo gledati kao gornju granicu svih sljedećih rezultata, a to je uzimanjem za središte kugle težište pravih pozitivaca i tražimo optimalni radijus. Također, obrađeno je kako izračunati procijenjeni radijus te s tim procijenjenim radijusom možemo napraviti kuglu oko težišta pravih pozitivaca. Međutim, u oba slučaja moramo poznavati podatke i znati koji su pravi pozitivci. U interesu je pronaći središte kugle nenadziranom klasifikacijom, odnosno napraviti kuglu s procijenjenim radijusom bez ikakvog znanja o podacima.

Nakon što je IGLOSS server dao svoje kandidate i podatke smo prebacili u vektorski prostor i standardizirali sada su nam podaci 80-dimenzionalni vektori koje će se u nastavku zvati točkama. Ideja je da iterativno prođemo kroz sve točke kao kandidate za središte kugle s procijenjenim radijusom te nađemo "najgušću" (u smislu s najviše točaka) kuglu. Pretpostavka iza ideje je da će u "najgušćoj" kugli biti zadržani pravi pozitivci, a što moguće više eliminirani lažni pozitivci. U većini slučajeva pretpostavka je bila istinita, no u nekima je stvarala problem i nalazila kugle gdje nema nijednog pravog pozitivca. Kako bismo riješili navedeni problem iskoristit će se pretpostavka o grupiranju pravih pozitivaca oko upita (motiva) na način da uzimamo kao kandidata za središte kugle sve točke u blizini upita, odnosno sve točke udaljene od upita manje od procijenjenog radijusa. Time je metoda vrlo brza i u svim slučajevima dobra. Ovom metodom smo na brz i uspješan način uspjeli naći kuglu s procijenjenim radijusom bez poznavanja podataka.

3.2 Primjeri i rezultati

U ovom radu uspješnost modela ispitana je na pet različitih proteoma:

- Talijin uročnjak (lat. *Arabidopsis thaliana*)
- Sirak (lat. *Sorghum*)
- Kukuruz (lat. *Zea mays*)
- Krumpir (lat. *Solanum tuberosum*)
- Soja (lat. *Glycine max*)

Kod svih proteoma korišten je upit RQVTFSKRRNGLLKKA za iterativno pretraživanje proteoma. Iako u poglavlju 1.3 su već spomenuti sljedeći pojmovi, u svrhu boljeg razumijevanja sljedećih rezultata dodat ćemo dodatna objašnjenja pojmova iz navedenog poglavlja.

Za proteome talijin uročnjak i krumpir dana je lista *Condition Positives* (CP), odnosno lista proteina koji su biološki utvrđeni da pripadaju MADS-box familiji. Za preostala tri proteoma CP lista je dobivena pretraživanjem po anotaciji proteoma s ključnim riječima za MADS-box familiju. Mjere uspješnosti izračunate su usporedbom rezultata modela s tim listama. Svi proteini koji se ne nalaze na CP listi smatraju se *Condition Negatives* (CN), odnosno biološki negativnima. Svi proteini koje je iterativni model vratio kao rezultat označeni su s P (**pozitivni**, eng. *Positives*), dok su svi ostali proteini iz danog proteoma koji nisu u rezultatu označeni s N (**negativni**, eng *Negatives*). Za ilustraciju slijede odnosi između definiranih pojmova i pojmova iz tablice uspješnosti:

$$TP = P \cap CP, \quad FP = P \cap CN,$$
$$TN = N \cap CN, \quad FN = N \cap CP.$$

Za svaki od proteoma korištene su tri skale pretraživanja u rasponu od 4 do 9 kako bi se pokazala učinkovitost metoda s obzirom na različit broj podataka. Skale pretraživanja su birane na način da je red veličine uzorka koju izbaci IGLOSS server približno jednak kod svih proteoma. Za svaku od skala pretraživanja prikazani su rezultati za kuglu sa središtem u težištu pravih pozitivaca s optimalnim radijusom, za kuglu sa središtem u težištu pravih pozitivaca s procijenjenim radijusom i za metodu prolaska po svim točkama u blizini upita. U tablici su navedeni osjetljivost modela (TPR), preciznost (PPV), mjera uspješnosti F_1-score, broj bioloških pozitivaca unutar kugle (TP), broj nizova aminokiselina koji se nalaze unutar kugle (n) i radijus kugle (r).

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. Kod uročnjaka motiv MADS-boxa je uključen u cvjetni razvoj. Njezin proteom je vrlo dobro anotiran i za svaki protein, od njih 35176 u proteomu, znamo kojoj proteinskoj familiji pripada. Duljina liste CP je 124.

Slika 3.1: Talijin uročnjak

1. Skala pretraživanja je 6, a veličina odgovora je 331.

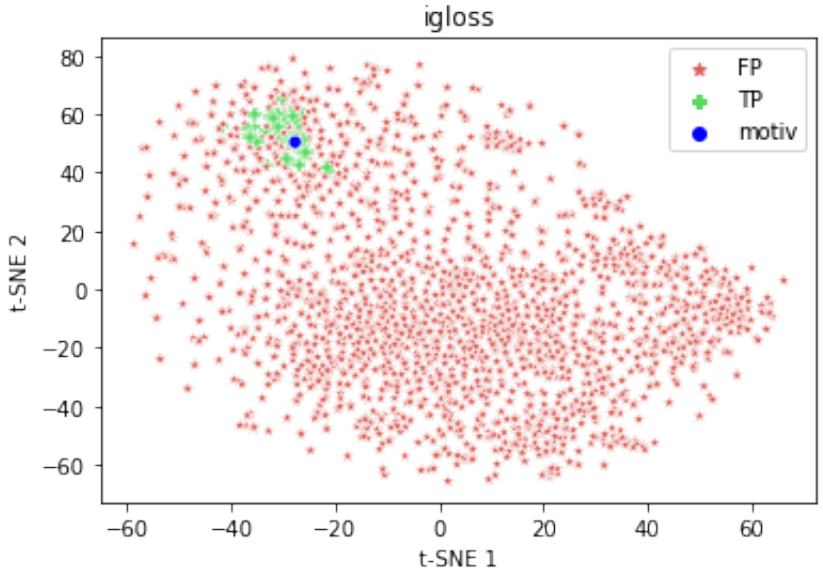
Model	TPR	PPV	F_1-score	TP	n	r
kugla oko težišta TP s optimalnim r	0.814	0.98	0.889	101	103	7.549
kugla oko težišta TP s procijenjenim r	0.774	1.0	0.872	96	96	6.957
metoda prolaska po svim točkama	0.742	1.0	0.852	92	92	6.972

2. Skala pretraživanja je 5, a veličina odgovora je 681.

Model	TPR	PPV	F_1-score	TP	n	r
kugla oko težišta TP s optimalnim r	0.806	1.0	0.893	100	100	6.969
kugla oko težišta TP s procijenjenim r	0.766	1.0	0.868	95	95	6.533
metoda prolaska po svim točkama	0.758	1.0	0.862	94	94	6.832

3. Skala pretraživanja je 4, a veličina odgovora je 2290.

Model	TPR	PPV	F_1-score	TP	n	r
kugla oko težišta TP s optimalnim r	0.823	0.99	0.898	102	103	7.339
kugla oko težišta TP s procijenjenim r	0.75	1.0	0.857	93	93	6.489
metoda prolaska po svim točkama	0.75	1.0	0.857	93	93	6.788



Slika 3.2: Talijin uročnjak t-SNE prikaz

Iz gornjih tablica možemo primijetiti za svaku skalu pretraživanja, odnosno bez obzira na veličinu odgovora, kod kugle oko težišta pravih pozitivaca s optimalnim radijusom F_1-score je vrlo visok, tj. blizu 0.9. Također, preciznost je često jednaka 1 što znači da kugla uspješno uzima bitne podatke. Kod ostale dvije kugle F_1-scoreovi su vrlo blizu "idealnom" rezultatu. Za nenadziranu metodu prolaska po svim točkama oko upita F_1-score je manji za manje od 4% od "idealnog" rezultata. Time vidimo da metoda jako dobro radi iako ne poznaje podatke i uspijeva pronaći kuglu koja će zadržati prave pozitivce, a eliminirati što više lažnih. Pomoću Slike 3.2 vidimo da je nevjerojatno kako u hrpi crvenih točaka je pronađeno i pokupljeno veliki broj zelenih točaka i potvrđujemo da je pretpostavka o grupiranju pravih pozitivaca oko upita ispravna. Ista stvar će se ustvrditi i kod preostalih proteoma. Također, možemo primijetiti da je procijenjeni radijus uvijek manji od optimalnog radijusa te možemo pretpostaviti da bi postojalo poboljšanje u rezultatima s povećanjem radijusa.

Sirak

Sirak (lat. *Sorghum*) je rod brojnih biljnih vrsta iz porodice trava. Neke od njih se uzgajaju kao žitarice, neke kao stočna hrana ili za proizvodnju alkoholnih pića i sirupa. Pogodna područja za rast su tropski i suptropski dijelovi svijeta te su mnoge vrste otporne na sušu i visoku temperaturu. Njezin proteom ima 41380 proteina, a duljina liste CP je 96.

Slika 3.3: Sirak

1. Skala pretraživanja je 8, a veličina odgovora je 129.

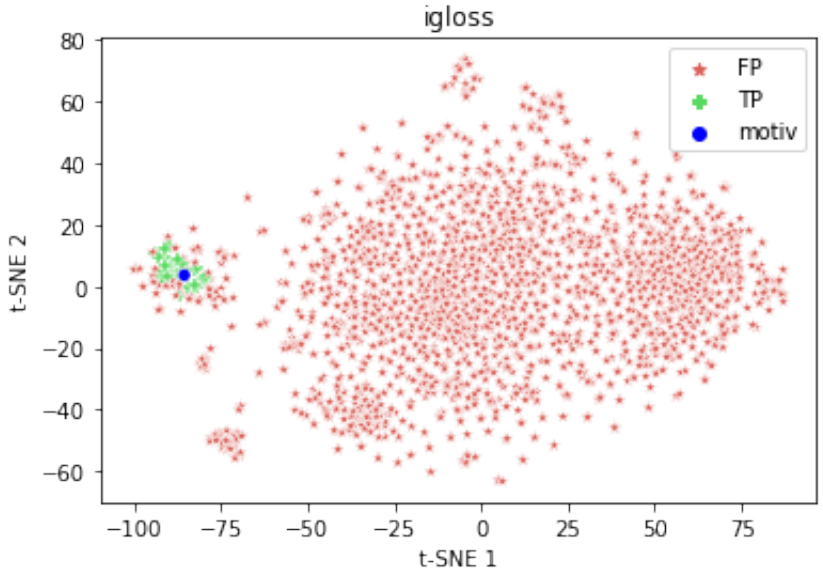
Model	TPR	PPV	F_1-score	TP	n	r
kugla oko težišta TP s optimalnim r	0.865	0.965	0.912	83	86	8.829
kugla oko težišta TP s procijenjenim r	0.802	0.963	0.875	77	80	8.054
metoda prolaska po svim točkama	0.822	0.963	0.887	79	82	8.554

2. Skala pretraživanja je 7, a veličina odgovora je 1535.

Model	TPR	PPV	F_1-score	TP	n	r
kugla oko težišta TP s optimalnim r	0.865	0.965	0.912	83	86	8.539
kugla oko težišta TP s procijenjenim r	0.739	0.959	0.835	71	74	6.903
metoda prolaska po svim točkama	0.781	0.961	0.862	75	78	7.903

3. Skala pretraživanja je 6, a veličina odgovora je 2429.

Model	TPR	PPV	F_1-score	TP	n	r
kugla oko težišta TP s optimalnim r	0.864	0.965	0.912	83	86	8.459
kugla oko težišta TP s procijenjenim r	0.76	0.96	0.848	73	76	6.734
metoda prolaska po svim točkama	0.781	0.961	0.862	75	78	7.734



Slika 3.4: Sirak t-SNE prikaz

Rezultati su vrlo slični kao kod talijinog uročnjaka. Iz gornjih tablica možemo primijetiti, bez obzira na veličinu odgovora, kod kugle oko težišta pravih pozitivaca s optimalnim radijusom F_1-score je 0.912. Također, preciznost je vrlo blizu 1 što znači da kugla uspješno uzima bitne podatke. Kod ostale dvije kugle F_1-scoreovi su jako blizu "idealnom" rezultatu. Za nenadziranu metodu prolaska po svim točkama oko upita F_1-score je manji za manje od 5% od "idealnog" rezultata. Time vidimo da metoda jako dobro radi iako ne poznaje podatke i uspijeva pronaći kuglu koja će zadržati prave pozitivce, a eliminirati što više lažnih. Iz Slike 3.4 vidimo isto kako uspijeva pokupiti taj mali udio zelenih točaka među hrpom crvenih točaka i kako se zelene točke grupiraju oko upita. Također, možemo primijetiti da je procijenjeni radijus uvijek manji od optimalnog radijusa, a svejedno daje visoke rezultate. Kod svih skala pretraživanja nenadzirana metoda daje viši F_1-score nego kod nadzirane s procijenjenim radijusom, ali treba uzeti u obzir da se ipak radi o većem radijusu kod nenadzirane što dijelom pridonosi većem rezultatu.

Kukuruz

Kukuruz (lat. *Zea mays*) je žitarica koju su prvi udomaćili domorodački narodi u južnom Meksiku prije otprilike 10 000 godina. Postao je osnovna hrana u mnogim dijelovima svijeta, a ukupna proizvodnja kukuruza nadmašuje proizvodnju pšenice ili riže. Osim što ga ljudi izravno konzumiraju koristi se također za kukuruzni etanol, stočnu hranu i druge proizvode. Kod kukuruza motiv MADS-boxa je uključen u cvjetni razvoj. Njezin proteom ima 63236 proteina, a duljina liste CP je 115.

Slika 3.5: Kukuruz

1. Skala pretraživanja je 9, a veličina odgovora je 173.

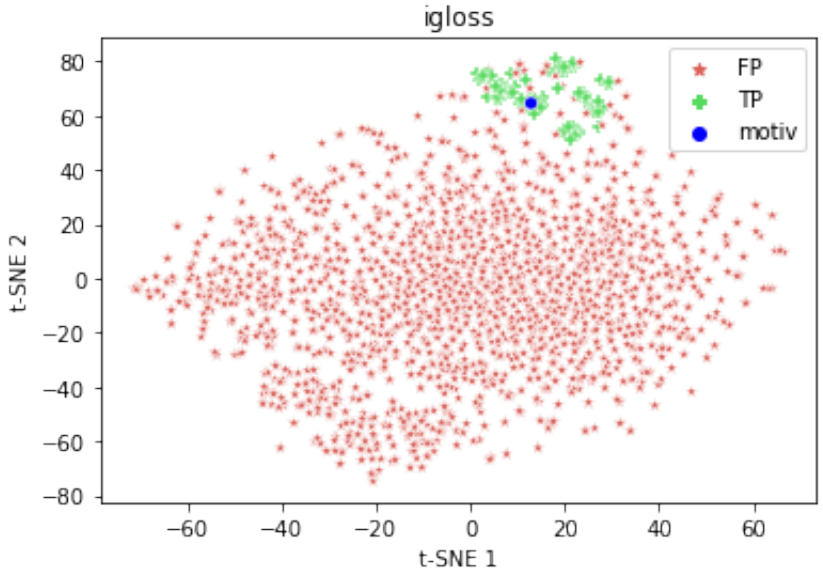
Model	TPR	PPV	F_1-score	TP	n	r
kugla oko težišta TP s optimalnim r	0.869	0.714	0.784	100	140	8.789
kugla oko težišta TP s procijenjenim r	0.913	0.719	0.805	105	146	9.835
metoda prolaska po svim točkama	0.913	0.719	0.805	105	146	9.835

2. Skala pretraživanja je 8, a veličina odgovora je 1751.

Model	TPR	PPV	F_1-score	TP	n	r
kugla oko težišta TP s optimalnim r	0.965	0.716	0.822	111	155	8.609
kugla oko težišta TP s procijenjenim r	0.895	0.715	0.795	103	144	6.837
metoda prolaska po svim točkama	0.886	0.713	0.791	102	143	6.837

3. Skala pretraživanja je 6, a veličina odgovora je 2271.

Model	TPR	PPV	F_1-score	TP	n	r
kugla oko težišta TP s optimalnim r	0.947	0.717	0.816	109	152	7.189
kugla oko težišta TP s procijenjenim r	0.930	0.718	0.811	107	149	6.890
metoda prolaska po svim točkama	0.930	0.718	0.811	107	149	6.890



Slika 3.6: Kukuruz t-SNE prikaz

Iz gornjih tablica možemo primijetiti za svaku skalu pretraživanja, odnosno bez obzira na veličinu odgovora, kod kugle oko težišta pravih pozitivaca s optimalnim radijusom F_1-score je visok, tj. vrijednost mu je blizu 0.8. Možemo primijetiti da je kod skale 9, za razliku od dosadašnjih rezultata, F_1-score malo veći kod ostale dvije kugle nego kod "idealnog" slučaja. Pretpostavljam da je zbog jako malog broja odgovora procijenjeni radijus veći nego optimalni radijus time je i rezultat malo viši. Zanemarujući ovu iznimku vidimo da za nenadziranu metodu prolaska po svim točkama oko upita F_1-score je manji za manje od 3% od "idealnog" rezultata. Možemo reći da metoda jako dobro radi iako ne poznaje podatke. Iz Slike 3.6 vidimo da je jako veliki broj crvenih točaka i da su zelene točke raspršenije nego očekivano. Bez obzira na tu raspršenost kugla s nenadziranom metodom skroz solidno uspije pronaći prave kandidate. Iako je preciznost manja nego kod uročnjaka i sirka svejedno se smatra solidnim rezultatom budući da je osjetljivost modela dosta visoka na svim skalama pretraživanja.

Krumpir

Krumpir (lat. *Solanum tuberosum*) je trajna zeljasta biljka iz porodice pomoćnica. U početku je uzgajan kao ukrasna biljka, a danas je jedna od najvažnijih prehrambenih biljaka. Za krumpir su uzeta dva proteoma, stari i novi, te za oba su napravljeni rezultati.

Slika 3.7: Krumpir

Proteom iz starih podataka ima 35004 proteina i duljina liste CP je 113. U sljedećim tablicama nalaze se rezultati za krumpir sa starim proteomom.

1. Skala pretraživanja je 7, a veličina odgovora je 246.

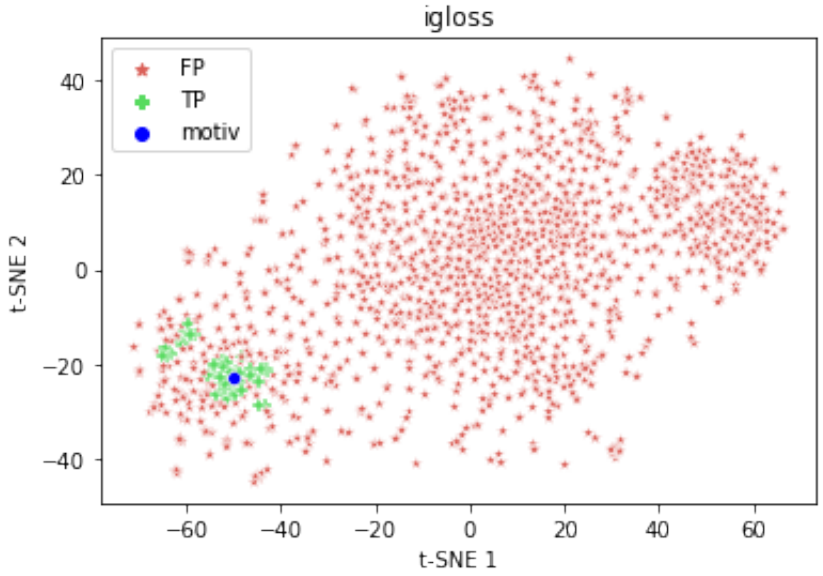
Model	TPR	PPV	F_1-score	TP	n	r
kugla oko težišta TP s optimalnim r	0.734	0.747	0.741	83	111	7.799
kugla oko težišta TP s procijenjenim r	0.699	0.766	0.731	79	103	7.366
metoda prolaska po svim točkama	0.699	0.752	0.725	79	105	7.366

2. Skala pretraživanja je 6, a veličina odgovora je 1332.

Model	TPR	PPV	F_1-score	TP	n	r
kugla oko težišta TP s optimalnim r	0.752	0.758	0.755	85	112	7.729
kugla oko težišta TP s procijenjenim r	0.895	0.715	0.795	76	92	6.750
metoda prolaska po svim točkama	0.681	0.802	0.737	77	96	6.765

3. Skala pretraživanja je 5, a veličina odgovora je 1740.

Model	TPR	PPV	F_1-score	TP	n	r
kugla oko težišta TP s optimalnim r	0.699	0.822	0.755	79	96	6.659
kugla oko težišta TP s procijenjenim r	0.690	0.821	0.749	78	95	6.646
metoda prolaska po svim točkama	0.672	0.8	0.730	76	95	6.646



Slika 3.8: Krumpir t-SNE prikaz

Iz gornjih tablica možemo primijetiti za svaku skalu pretraživanja, odnosno bez obzira na veličinu odgovora, kod kugle oko težišta pravih pozitivaca s optimalnim radijusom F_1-score nije veći od 0.755. Uz score koji nije "obećavajući", preciznost i osjetljivost također nisu visoki, stoga dolazimo do sumnje uzima li metoda bitne podatke. Postavlja se pitanje radi li se o fenomenu poput evolucijskog *splita* ili o nečemu drugome poput loše anotacije, upita koji nije dovoljno karakterističan za familiju i slično. No, uzevši u obzir rezultat "idealnog" slučaja, F_1-score dobiven nenadziranom metodom prolaska po svim točkama je jako blizu "idealnom" rezultatu. Vidimo da metoda s obzirom na "idealan" slučaj radi dobro iako ne poznaje podatke. Iz Slike 3.8 vidimo da se zelene točke razdvajaju u dvije skupine gdje je jedna grupirana oko upita. Iz slike možemo pretpostaviti da postoji mogućnost evolucijskog *splita* ili metoda pronalazi neki drugi transkripcijski faktor koji je vrlo sličan MADS-boxu. Kako bismo provjerili radi li se o stvarnom fenomenu ili lažnom signalu provjerit ćemo imamo li isto ponašanje i na novom proteomu krumpira.

Proteom iz novih podataka ima 56103 proteina i duljina liste CP je 160. U sljedećim tablicama nalaze se rezultati za krumpir s novim proteomom.

1. Skala pretraživanja je 9, a veličina odgovora je 687.

Model	TPR	PPV	F_1-score	TP	n	r
kugla oko težišta TP s optimalnim r	0.631	0.721	0.673	101	140	8.719
kugla oko težišta TP s procijenjenim r	0.699	0.766	0.731	75	111	6.969
metoda prolaska po svim točkama	0.5375	0.699	0.607	86	123	8.469

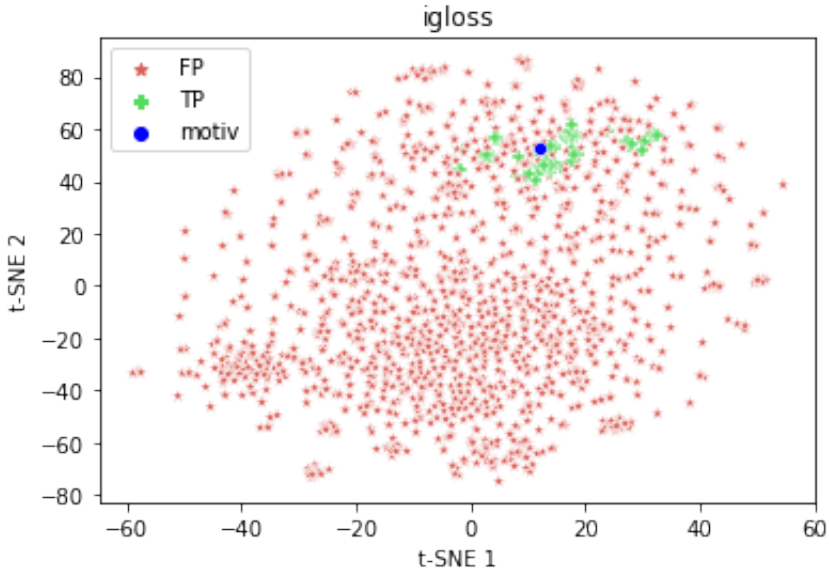
2. Skala pretraživanja je 6, a veličina odgovora je 1448.

Model	TPR	PPV	F_1-score	TP	n	r
kugla oko težišta TP s optimalnim r	0.656	0.709	0.681	105	148	8.269
kugla oko težišta TP s procijenjenim r	0.506	0.686	0.582	81	118	8.177
metoda prolaska po svim točkama	0.568	0.705	0.629	91	129	6.765

3. Skala pretraživanja je 5, a veličina odgovora je 1829.

Model	TPR	PPV	F_1-score	TP	n	r
kugla oko težišta TP s optimalnim r	0.643	0.705	0.673	103	146	7.719
kugla oko težišta TP s procijenjenim r	0.525	0.694	0.598	84	121	6.601
metoda prolaska po svim točkama	0.531	0.696	0.602	85	122	7.100

Iako je postojala nada u novim podacima i boljoj anotaciji unutar novog proteoma, ponašanje je slično kao kod krumpira iz starih podataka. Iz gornjih tablica možemo primijetiti za svaku skalu pretraživanja, odnosno bez obzira na veličinu odgovora, kod kugle oko težišta pravih pozitivaca s optimalnim radijusom F_1-score nije veći od 0.681. Uz score koji je nizak također je i preciznost i osjetljivost niska, stoga ne možemo zaključiti da metoda ispravno uzima bitne podatke. No, uzevši u obzir rezultat "idealnog" slučaja F_1-score dobiven nenadziranom metodom prolaska po svim točkama oko upita je jako blizu tom rezultatu. Time vidimo da metoda s obzirom na "idealan" slučaj radi dobro iako ne poznaje podatke.

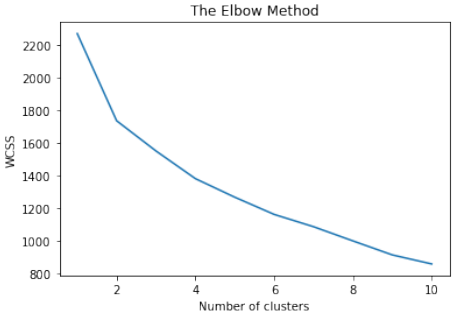


Slika 3.9: Krumpir t-SNE prikaz

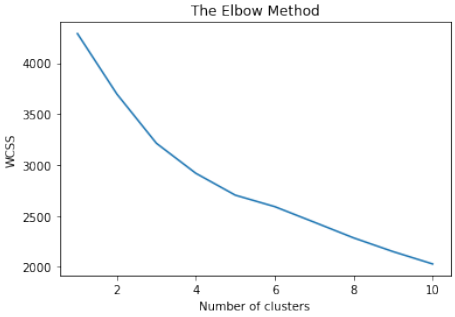
Iz Slike 3.9 vidimo da se zelene točke ne razdvajaju samo u dvije skupine kao što je to bili na Slici 3.8, nego postoji i raspršenost pravih pozitivaca.

Motivirana sumnjom o razdvajanju pravih pozitivaca u dva klastera htjela sam vidjeti je li moguće napraviti novu kuglu koja će pokupiti ostatak pravih pozitivaca. Kako bih to postigla morala sam koristiti znanje o podacima, odnosno nadziranom klasifikacijom pronaći drugu kuglu. Kako su unutar prve kugle već pokupljeni neki pravi pozitivci, a želimo s novom kuglom pokupiti ostatak pravih pozitivaca, za početak ćemo iz skupine pravih pozitivaca izbaciti sve koji su pokupljeni s prvom kuglom. Pošto koristimo znanje o podacima uzeti ćemo za središte kugle težište preostalih pravih pozitivaca i napraviti kuglu oko tog težišta. Na taj način se kod starih i novih podataka krumpira uspije pokupiti barem 50% preostalih pravih pozitivaca.

Također, htjela sam provjeriti metodom lakta bi li postojao kakav zaključak o razdvajanju u klastere. Metoda lakta je objašnjena u izvoru [25]. Iz Slike 3.10a možemo vidjeti kod starog proteoma da radi "blagi lakat" za broj klastera jednak dva. No, kod Slike 3.10b možemo vidjeti kako se za novi proteom ne može konkretno zaključiti da se radi o dva klastera. Iako slika kod starog proteoma daje sumnju o dva klastera, metoda lakta kod novog proteoma više daje sumnju na to da su podaci "bučni" i ne mora biti u pitanju evolucijski *split*.



(a) Metoda lakta stari proteom



(b) Metoda lakta novi proteom

Slika 3.10: Metoda lakta za krumpir

Soja

Soja (lat. *Glycine max*) je mahunarka visoke hranjive vrijednosti. Postoje razne sorte soje razlikovane po obliku zrna, boji, okusu i kemijskim svojstvima. Soja je jedna od biljaka kojima se genetički manipulira te se genetički modificirana soja koristi u sve više proizvoda. Njezin proteom ima 74683 proteina, a duljina liste CP je 205.

Slika 3.11: Soja

1. Skala pretraživanja je 7, a veličina odgovora je 590.

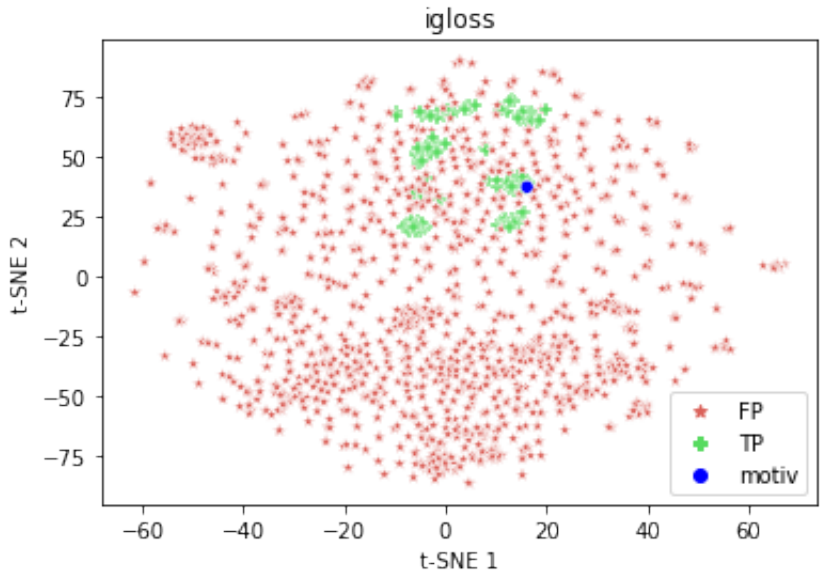
Model	TPR	PPV	F_1-score	TP	n	r
kugla oko težišta TP s optimalnim r	0.848	0.763	0.803	174	228	8.519
kugla oko težišta TP s procijenjenim r	0.785	0.777	0.781	161	207	7.502
metoda prolaska po svim točkama	0.771	0.774	0.772	158	204	7.502

2. Skala pretraživanja je 6, a veličina odgovora je 1570.

Model	TPR	PPV	F_1-score	TP	n	r
kugla oko težišta TP s optimalnim r	0.819	0.774	0.796	168	217	8.029
kugla oko težišta TP s procijenjenim r	0.765	0.773	0.769	157	203	6.912
metoda prolaska po svim točkama	0.760	0.772	0.766	156	202	6.912

3. Skala pretraživanja je 5, a veličina odgovora je 2024.

Model	TPR	PPV	F_1-score	TP	n	r
kugla oko težišta TP s optimalnim r	0.834	0.76	0.795	171	225	7.689
kugla oko težišta TP s procijenjenim r	0.770	0.774	0.772	158	204	6.725
metoda prolaska po svim točkama	0.765	0.773	0.769	157	203	6.890

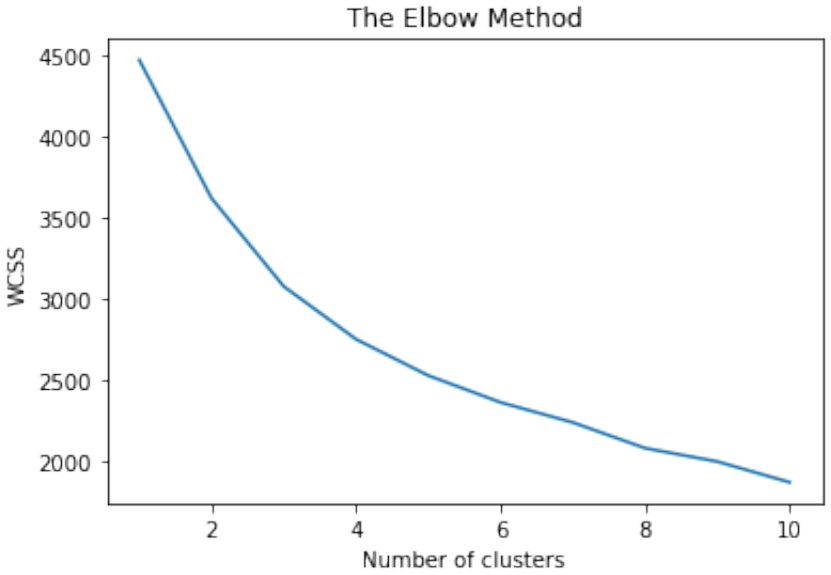


Slika 3.12: Soja t-SNE prikaz

Iz gornjih tablica možemo primijetiti za svaku skalu pretraživanja, odnosno bez obzira na veličinu odgovora, kod kugle oko težišta pravih pozitivaca s optimalnim radijusom F_1-score je visok, tj. vrijednost mu je blizu 0.8. Iako je F_1-score blizu rezultatima kao kod kukuruza ipak su vrijednosti osjetljivosti znatno manje nego kod kukuruza. Uzevši u obzir rezultat "idealnog" slučaja, F_1-score nenadzirane metode prolaska po svim točkama oko upita je jako blizu "idealnom" rezultatu. Time vidimo da metoda s obzirom na "idealan" slučaj radi dobro iako ne poznaje podatke. No, iako za nenadziranu metodu F_1-score je manji za manje od 3% od "idealnog" rezultata, pitamo se koliko metoda dobro uzima bitne podatke. Iz Slike 3.12 vidimo da je jako veliki broj crvenih točaka i da su zelene točke jako raspršene i grupiraju se na više dijelova.

Kako smo kod krumpira dobili sumnju o razdvajanju pravih pozitivaca u klastere provjerit ćemo i za soju je li moguće napraviti novu kuglu koja će pokupiti ostatak pravih pozitivaca. Na isti princip kao kod krumpira, s novom kuglom uspijevam pokupiti preko

60% preostalih pravih pozitivaca. Također, provjerila sam što bi metoda lakta mogla sugerirati te su isti zaključci kao kod metode lakta za novi proteom krumpira. Iz slike 3.13 možemo pretpostaviti da su podaci vrlo moguće "bučni" i ne može se zaključiti da se radi o dva ili više klastera.



Slika 3.13: Metoda lakta za soju

Za kraj sam još pogledala postoji li mogućnost da je "problem" u odgovoru koji daje IGLOSS, odnosno bili se dobio veći F_1-score ako se uzme neki drugi server za iterativno pretraživanje proteoma. Motivirano time uzimam BLAST server koji radi na malo drugačiji način od IGLOSS-a. BLAST server je opisan u izvoru [26]. Rezultate sam napravila za različite *e*-value, no pokazani su samo za one kod kojih sam dobila najveći odgovor od BLAST-a. Također, u svrhu usporedbe napravila sam rezultate i na kukuruzu. Za svaki proteom prikazani su rezultati za iste modele kao kod IGLOSS servera i navedeni su osjetljivost modela (TPR), preciznost (PPV), mjera uspješnosti F_1-score i veličina odgovora dobivena BLAST serverom.

Krumpir (stari proteom)				
Model	TPR	PPV	F_1-score	odgovor
kugla oko težišta TP s optimalnim r	0.433	0.816	0.566	402 (96)
kugla oko težišta TP s procijenjenim r	0.433	0.765	0.553	402 (96)
metoda prolaska po svim točkama	0.433	0.753	0.551	402 (96)

Soja				
Model	TPR	PPV	F_1-score	odgovor
kugla oko težišta TP s optimalnim r	0.556	0.721	0.628	408 (180)
kugla oko težišta TP s procijenjenim r	0.580	0.721	0.643	408 (180)
metoda prolaska po svim točkama	0.580	0.721	0.643	408 (180)

Kukuruz				
Model	TPR	PPV	F_1-score	odgovor
kugla oko težišta TP s optimalnim r	0.869	0.699	0.775	425 (170)
kugla oko težišta TP s procijenjenim r	0.886	0.698	0.782	425 (170)
metoda prolaska po svim točkama	0.886	0.698	0.782	425 (170)

Uspoređujući gornje tablice s tablicama dobivenim za svaki proteom uz server IGLOSS možemo primijetiti da su F_1-scoreovi još manji. Kod krumpira su rezultati vrlo loši i osjetljivost je jako niska, no treba uzeti u obzir da se ipak radi o samo 96 valjanih podataka na kojima se provodio postupak. Kod soje je odgovor ipak veći, no osjetljivost je također niska i F_1-score je za barem 20% manji. No, kod kukuruza vidimo da je F_1-score vrlo blizu onome što smo dobili s IGLOSS-om.

Analiza rezultata

U ovom radu istražuje se način na koji bi se povećala točnost iterativnog modela pretraživanja. Pomoću IGLOSS servera dobije se lista proteina iz koje želimo izvući prave pozitivce, a izbaciti lažne pozitivce. Kako bismo to postigli za početak prebacujemo podatke u vektorski prostor i standardiziramo, te koristimo matematičko znanje kako bi našli prave pozitivce. Pretpostavka je da se pravi pozitivci grupiraju oko upita što je potvrđeno pomoću rezultata kugle oko težišta pravih pozitivaca i t-SNE prikaza. Dodatno, pretpostavka je da se može nenadziranim načinom doći do "najgušće" kugle i da će dobivena kugla zadržati prave pozitivce i upit, a eliminirati što više lažnih pozitivaca.

Navedene pretpostavke su provjerene na pet različitih proteoma i za svaki je uspješno potvrđena pretpostavka o grupiranju pravih pozitivaca oko upita. Naime, može se primijetiti da su rezultati varirali po proteomima. Kod talijinog uročnjaka i sirka rezultati su pokazali da su obje pretpostavke potvrđene i uspješnost je vrlo visoka. Međutim, kod krumpira i soje nastao je problem poput grupiranja pravih pozitivaca u više klastera i rapršenosti pravih pozitivaca. Mogući uzrok tog fenomena je što se radi o biljkama koje su kultivari te su industrijski više orijentirani, dok kod talijinog uročnjaka i sirka to nije slučaj. Kukuruz je također kultivar i industrijski orijentiran pa je opravdano da uspješnost nije vrlo visoka, no ipak je osjetljivost visoka i bolji su rezultati nego kod soje i krumpira.

Kako bismo ispitali detaljnije rezultate dobivene za krumpir i soju dodatno je napravljena analiza i na BLAST serveru, s čime je pokazano da je uspješnost još manja. Za usporedbu je napravljena analiza i na kukuruzu kako bi potvrdili da su rezultati s BLAST-om korektni. Također, napravljena je metoda lakta kako bi se provjerilo postoji li mogućnost evolucijskog *splita*, odnosno radi li se o podfamilijama ili nekim drugim transkripcijskim faktorima. Tom metodom se ne može potvrditi pretpostavka o postojanju više klastera.

Za kraj možemo primijetiti da su kod svih rezultata procijenjeni radijusi bili manji od optimalnih radijusa, stoga možemo pretpostaviti da bi s većim radijusom uspješnost bila veća. Prisjetimo se da su rezultati dobiveni uz pretpostavku da je koeficijent očuvanosti jednak 0.68. Kada bi koeficijent očuvanosti bio manji procijenjeni radijus bi bio veći, pa možemo pretpostaviti da bi i uspješnost bila veća smanjivanjem koeficijenta očuvanosti.

Pojmovi iz ovog poglavlja preuzeti su iz izvora [6], [7], [8], [9] i [10], te iz izvora od [13] do [26].

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

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

[5] Š. Ungar, *Metrički prostori*, predavanja, 2016., dostupno na https://www.mathos.unios.hr/metricki/metricki.pdf.

[6] M. Pathak, *Introduction to t-SNE*, dostupno na https://www.datacamp.com/community/tutorials/introduction-t-sne, (2018.).

[7] 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.

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

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

[10] I. Višek, *Clustering i klasifikacija proteinskih nizova*, Diplomski rad, Sveučilište u Zagrebu, Prirodoslovno-matematički fakultet (Matematički odsjek), 2022.

[11] 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

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

[13] https://en.wikipedia.org/wiki/Protein

[14] https://www.ebi.ac.uk/training/online/courses/protein-classification-intro-ebi-resources/protein-classification/what-are-protein-families/

[15] https://genomebiology.biomedcentral.com/articles/10.1186/gb-2001-3-1-reviews2001

[16] https://hr.wikipedia.org/wiki/Aminokiselina

[17] https://en.wikipedia.org/wiki/MADS-box

[18] https://www.frontiersin.org/articles/10.3389/fpls.2019.00853/full

[19] https://www.sciencedirect.com/science/article/pii/B9780128008546000087

[20] https://hr.wikipedia.org/wiki/Sirak

[21] https://en.wikipedia.org/wiki/Maize

[22] https://www.plantea.com.hr/krumpir/

[23] https://hr.wikipedia.org/wiki/Soja$_$(biljna$_$vrsta)

[24] Nekolicina slika priloženih u zadnjem potpoglavlju dostupne su na: https://en.wikipedia.org/wiki/Zea$_$(plant) https://news.yale.edu/2021/10/14/weed-winter-how-plants-detect-seasonal-changes

[25] https://www.analyticsvidhya.com/blog/2021/01/in-depth-intuition-of-k-means-clustering-algorithm-in-machine-learning/

[26] Altschul S.F. et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, 25, 3389–3402.

Sažetak

U ovom diplomskom radu promatra se problem klasifikacije proteina određenog proteoma u proteinske familije. Uz pomoć opisivanja aminokiselina numeričkim vektorima želi se povećati točnost standardnog modela pretraživanja.

Nakon uvedenih pojmova, matematičkih i bioloških, nužnih za razumijevanje ovog rada, navode se metode pomoću kojih se pronalazi kugla u vektorskom prostoru. Nadziranim i nenadziranim načinom nalazi se kugla koja sadržava motive koji su biološki dokazani da pripadaju određenoj proteinskoj familiji. Istraživanje se provodi na pet različitih proteoma i promatra se uspješnost na različitim veličinima skupova podataka. Glavna mjera za uspješnost modela je F_1-score. U radu su potvrđene pretpostavke o grupiranju pravih pozitivaca oko upita i postojanju kugle koja sadrži te prave pozitivce. Model se pokazao robustan na promjene broja podataka i nenadzirana klasifikacija se pokazala vrlo brzim rješenjem.

Summary

This thesis is concerned with classification of proteins into protein families. In particular, the aim is to increase the accuracy of a motif scanning procedure by using a description of amino acids as vectors.

After describing the mathematical and some biological background we present our Euclidean ball searching techniques. With supervised and unsupervised classification we found a sphere that contains motifs that have been shown to belong to the designated protein family. This study was carried out on five different proteomes and for different sizes of data sets. F_1-score was the primary metric used for measuring success of a model. Finally, we confirmed the hypothesis regarding clustering of true positive hits and the existence of sphere that contains them.

Životopis

Rođena sam u Splitu, 18. ožujka 1998. godine. Školovanje sam započela u OŠ kneza Trpimira u Kaštel Kambelovcu, nakon koje sam upisala III. gimnaziju u Splitu. Nakon završenog srednjoškolskog obrazovanja 2016. godine upisala sam preddiplomski sveučilišni studij Matematika na Prirodoslovno-matematičkom fakultetu u Zagrebu. Zvanje sveučilišnog prvostupnika matematike sam stekla 2020. godine kada sam upisala i diplomski sveučilišni studij Matematička statistika na istom fakultetu.

Osnovno i srednje glazbeno obrazovanje završila sam u Glazbenoj školi Josipa Hatzea u Splitu s glavnim predmetom klavir. U slobodno vrijeme volim šetati u prirodi, trenirati, čitati knjige i družiti se s bližnjima.