Proteinski motivi i klasifikacija proteinskih familija

Kokan, Kristin

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Kristin Kokan

PROTEINSKI MOTIVI I KLASIFIKACIJA PROTEINSKIH FAMILIJA

Diplomski rad

Voditelj rada:
doc. dr. sc. Pavle Goldstein

Zagreb, studeni 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. _____

Mojoj obitelji.

Zahvaljujem mentoru doc. dr. sc. Pavlu Goldsteinu na entuzijazmu, susretljivosti i velikoj pomoći tijekom pisanja ovog rada.
Najveće hvala mojoj mami na beskonačnoj podršci i tati na svakom savjetu.
Hvala svim prijateljima koji su obogatili moje studentsko putovanje.
Ivane, sve je bilo lakše i zabavnije uz tebe, hvala!
I za kraj, Maria, budi uporna i samo nebo ti je granica!

Sadržaj

Uvod

Proteini su građevni blokovi života. Uvrštava ih se u jedne od najvažnijih tvari u tijelu. Proteini su nizovi aminokiselina koji se međusobno razlikuju u sastavu i slijedu aminokiselina. Svaka aminokiselina ima svoju ulogu i funkciju te zajedno čine raznolik skup proteina koji je zaslužan za mnoge ključne procese u tijelu, poput rasta, popravka i održavanja tijela u optimalnom stanju.

Proteinska familija je skupina proteina koji su međusobno evolucijski povezani. Ti proteini najčešće imaju slična svojstva, strukturu i funkciju. Jedno je od bitnijih istraživanja u bioinformatici je problem traženja proteina koji pripadaju istoj proteinskoj familiji.

Budući da proteini predstavljaju veoma duge sekvence aminokiselina, umjesto cijelih nizova često se promatraju samo kratki bolje očuvani podnizovi koje zovemo motivi. Identifikacijom motiva, proteine ćemo klasificirati u promatranu proteinsku familiju. Zadavanjem upita koji je karakterističan niz aminokiselina za proteinsku familiju od interesa, pronalazi se skup proteina koji pripadaju istoj proteinskoj familiji.

U ovom radu fokusirat ćemo se na specifičnu proteinsku familiju i analizirati motive koji je karakteriziraju. Razvijamo metodu za pronalaženje središta i radijusa kugle koja obuhvaća relevantne motive.

Ovaj rad sastoji se od tri poglavlja. Prvo poglavlje obuhvaća osnovne pojmove iz matematike, tj. linearne algebre, vjerojatnosti i statistike koji su neophodni za razumijevanje. U tom dijelu također definiramo mjere uspješnosti. Drugo poglavlje se fokusira na biološki kontekst problema, pretraživanje motiva pomoću iterativnih metoda pretraživanja i njihovo prevođenje u vektorski prostor. Naposljetku, treće poglavlje pruža detaljan opis problema i algoritma koji je korišten te grafički i numerički prikaz rezultata dobivenih u radu.

Poglavlje 1

Matematička pozadina

Teoremi, definicije, propozicije i napomene iz linearne algebre, vjerojatnosti i statistike te uspješnosti modela navedeni su u ovom poglavlju. Pojmovi su preuzeti iz izvora [1], [2], [3], [4], [5], [8] i [13].

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 Teorija vjerojatnosti

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 Evaluacija modela

Klasifikacija

Klasifikacija podataka je proces određivanja pripadnosti opažanja određenoj grupi ili klasi. Postoje dvije osnovne vrste klasifikacije: nenadzirana i nadzirana.

Nenadzirana klasifikacija radi bez prethodnog poznavanja klasa u podacima. Model unaprijed ne zna koje klase postoje, već pokušava pronaći sličnosti između ulaznih podataka i temeljem tih sličnosti definira klase ili grupe. Ovdje model traži obrasce u podacima samostalno, bez unaprijed postavljenih ciljeva za klasifikaciju.

U nadziranoj klasifikaciji, model koristi podatke za koje već unaprijed znamo klasifikaciju ili pripadnost određenim klasama. Na temelju ovih poznatih podataka, model se „uči" kako klasificirati nove ulazne podatke. Model „uči" iz iskustva i podataka te ga koristimo za predviđanje klasa za nove, nepoznate podatke.

Nenadzirana klasifikacija često se koristi za istraživanje nepoznatih struktura u podacima i za otkrivanje potencijalnih klasa ili grupa, dok se nadzirana klasifikacija često koristi za predviđanje i razvrstavanje podataka temeljem prethodno poznatih klasa.

Mjere uspješnosti

Da bismo procijenili koliko je dobar naš model, koristimo mjere koje se temelje na informacijama iz matrice uspješnosti (eng. *confusion matrix*). Nakon što utvrdimo koje mjere koristimo za procjenu modela, možemo ih usporediti kako bismo odabrali najbolji model.

		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 zadano 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 dano 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 Struktura proteina

Proteini su lanci aminokiselina. Sastav i redoslijed aminokiselina u proteinu određuje njegovu strukturu i funkciju. Postoji dvadeset standardnih različitih aminokiselina koje se mogu naći u prirodi. Aminokiseline imaju jednaku strukturu, a razlikuju se po bočnom lancu koji je dio strukture. Aminokiseline označavamo velikim slovom engleske abecede, kao što je prikazano u tablici 2.1.

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

2.2 Proteinska familija

Proteom organizma obuhvaća sve proteine prisutne u tom organizmu. Unutar proteoma nalazimo raznolike skupine proteinskih familija, pri čemu svaka familija obavlja specifičnu ulogu u organizmu. Određivanje kojoj proteinskoj familiji pripada pojedini protein ima iznimno velik značaj jer nam omogućuje bolje razumijevanje funkcije tog proteina.

Kada uspješno utvrdimo kojoj proteinskoj familiji pripada određeni protein, dobivamo dublji uvid u karakteristike organizma. Ovo znanje otvara vrata za potencijalno poboljšanje karakteristika organizma putem genetske modifikacije, što može biti korisno u različitim područjima, uključujući poljoprivredu i biomedicinska istraživanja.

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 familija gena MADS-box. Proteinska familija MADS-box dobila je ime kao akronim od inicijala četiri transkripcijska faktora koji su prvi otkriveni u ovoj familiji: mini kromosomska održivost 1 MCM1, (iz *Saccharomyces cerevisiae*, pupajući kvasac), agamozan AG (iz *Arabidopsis thaliana*, talijin uročnjak), deficijentan DEF (iz *Antirrhinum majus*, „snapdragon") i faktor odgovora seruma SRF (iz *Homo Sapiens*).

MADS-box geni obavljaju raznovrsne uloge u različitim organizmima. Ključni su regulatori svakog aspekta razvoja reprodukcije biljaka. „Igraju" posebno istaknute uloge u kontroli vremena cvjetanja, arhitekturi cvatova, određivanju identiteta cvjetnih organa i razvoju sjemena. U kontekstu biljnih organizama, transkripcijski faktori MADS domene imaju poseban značaj zbog svoje sposobnosti preciznog vezanja za DNA i mogućnosti da se združe u različite konfiguracije. Iz tog razloga, identifikacija novih članova MADS-box familije u biljnim organizmima predstavlja područje istraživanja s iznimnim interesom i potencijalom za otkrivanje novih regulacijskih mehanizama. U životinjama, sudjeluju u razvoju mišića te u procesima proliferacije i razlikovanja stanica, uglavnom vezanih za različite aspekte razvoja i funkcionalnosti mišića i stanica. U gljivama, MADS-box geni obavljaju različite funkcije, uključujući odgovor na feromone i regulaciju metabolizma arginina. Ovi geni imaju raznolike uloge u biokemijskim procesima unutar gljiva.
Razvojna i evolucijska važnost MADS-box gena široko je priznata.

2.3 Motiv i pretraživanje motiva

Motiv (ili upit) je niz aminokiselina duljine najmanje 5 koji mutira na specifičan način. Metode za pretragu motiva često se primjenjuju u analizi nizova u području bioinformatike. Iterativno pretraživanje proteina predstavlja uobičajenu tehniku za identificiranje proteina unutar iste proteinske familije. Glavna svrha ovog postupka jest pronaći nizove aminokiselina koji su dovoljno slični zadanom upitu, posebno s obzirom na određenu razinu funkcionalne sličnosti. Ova metoda započinje s konsenzus nizom koji odgovara motivu specifičnim za proteinsku familiju koja nas zanima i koristi ga kao početnu točku za pretragu i identifikaciju drugih srodnih proteina. Odgovor na pretragu generira se u dva koraka. Prvo, svi rezultati lokalnog poravnanja rangiraju se prema sličnosti s motivom. Zatim se odabiru samo oni rezultati čija je sličnost iznad određenog praga. Važno je napomenuti da oba koraka mogu biti izvor grešaka. Primjerice, ako rezultate netočno rangiramo s preniskim pragom, možemo dobiti velik broj identificiranih motiva, uključujući i one koji nisu biološki relevantni. S druge strane, postavljanje previsokog praga može propustiti neke stvarne biološki važne motive. Različite metode pretraživanja motiva pokušavaju rješavati ove izazove na različite načine.

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 2.3.1. ***BLOSUM matrica*** *B je 20×20 matrica, $B = (b_{ij}) \in M_{20}(\mathbb{Z})$, koja na (i, j)-tom mjestu sadrži koeficijente sličnosti i-te i j-te aminokiseline. 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 2.3.2. ***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.*

2.4 Prelazak u vektorski prostor

Nedostatak prirodne metrike za usporedbu nizova slova sprečava obradu nad takvim podacima. Stoga je potrebno opisati aminokiseline pomoću numeričkih vrijednosti. Navedena problematika je opisana i riješena u članku [11]. 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 aminokiselina duljine n odgovara $5n$-dimenzionalni vektor. Ovako opisane aminokiseline koriste se u istraživanjima posvećenim razumijevanju evolucije, te strukturnih i funkcionalnih svojstava proteina.

Pojmovi iz ovog poglavlja preuzeti su iz izvora [5], [7], [8], [11] i [13].

Poglavlje 3

Analiza problema i rezultati

3.1 Cilj rada

Cilj ovog diplomskog rada je utvrditi kako se biljke ponašaju (u smislu MADS-boxa) i ima li razlike između divljih i uzgajanih biljaka te pokušati shvatiti razlog. Nakon što iterativna metoda izbaci svoje kandidate za proteinsku familiju, među njima se nalaze motivi iz proteina koji stvarno pripadaju familiji (eng. *true positives*) i oni koji ne pripadaju (eng. *false positives*). Želimo eliminirati što više lažnih pozitivaca, a zadržati prave pozitivce iz odgovora kojeg dobijemo koristeći IGLOSS server. Da bismo to postigli, iskoristit ćemo mogućnost prelaska u vektorski prostor. Na taj način možemo promatrati distribuciju nizova u prostoru i analizirati njihove udaljenosti.

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 djelomično točne. Time smo problem sveli na traženje središta i radijusa kugle za koju je mjera uspješnosti modela F_1 najveća. Od interesa će se pokazati i veličina radijusa kugle te ćemo proučavati i što bi se dogodilo kad bismo taj radijus malo povećali. Očekujemo da ćemo za divlje biljke dobiti slične rezultate, a da će se uzgajane biljke međusobno ponašati slično, ali različito u odnosu na divlje.

Priprema podataka

U ovom radu korištena su dva upita RQVTFSKRRNGLLKKA i KTNRQVTFSKRRNGLLKKAYEL, s tim da je naglasak stavljen na duljem upitu. RQVTFSKRRNGLLKKA je standardni upit. Upiti su nizovi aminokiselina karakterističnih za MADS-box familiju. Navedeni upiti koriste se kako bi uz pomoć IGLOSS servera dobili najbolje kandidate za MADS-box familiju. Služimo se s dva upita kako bismo bolje shvatili što se dogodilo s razvojom biljke, preciznije rečeno ne gledamo cijele gene nego samo MADS-box. Također, koristimo više skala i koeficijenata sličnosti kako bi se s obzirom na različit broj podataka pokazala učinkovitost metode. Odgovori će, kao i upiti, biti nizovi aminokiselina jednakih duljina. Ako je upit 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. Naime, ako se to dogodi, euklidska udaljenost bila bi dominirana tom koordinatom. Time bi se izgubila forma kugle u kojoj bi sve koordinate trebale imati jednak utjecaj. Standardizacijom podataka riješi se potencijalni problem. 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, a $\overline{x}$ i s aritmetička sredina i standardna devijacija podataka, redom. Tada je za $i = 1, 2, ..., n$:

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

Kugla oko težišta pravih pozitivaca

Nakon pripreme podataka za analizu, krenut ćemo u potragu za središtem i radijusom kugle koja će zadržati prave pozitivce, istovremeno eliminirajući lažne. Kada IGLOSS server generira popis pozitivaca, koji uključuju kako prave, tako i lažne, naš cilj je pronaći kuglu koja će maksimizirati F_1 score. Kada bismo koristili činjenicu da znamo koji su pravi pozitivci, smisleno je za središte kugle uzeti težište svih pravih pozitivaca. Nakon fiksiranja središta želimo pronaći optimalni radijus za koji je uspješnost modela najveća, odnosno najveći F_1 score. Pronalazak optimalnog radijusa će se raditi iteracijom po vrijednostima od 2 do 9 s pomakom veličine 0.01 gdje se u svakoj iteraciji izračuna F_1 score. Nakon prolaska po svim vrijednostima dobijemo za koji radijus 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. Ako rezultati budu jako blizu „idealnom" slučaju znamo da su vrlo dobri.

Procijenjeni radijus kugle

U radu [5] je optimalni radijus bio blizu vrijednosti procijenjenog radijusa pa se na temelju tih rezultata u ovom radu proučavaju 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 [6, str. 55], a teorem 3.1.1 je poznati Pitagorin teorem iz kojeg slijedi formula za udaljenost dviju točaka u višedimenzionalnom prostoru.

Pretpostavimo da se aminokiseline pojavljuju s vjerojatnostima $p_k,\ k \in \{1, 2, \ldots, 20\}$ zadanima u distribuciji aminokiselina navedenoj u poglavlju 2.3 te neka su $A_i, i \in \{1, 2, \ldots, 20\}$ distribucije zadane nekom aminokiselinom za koju ćemo pretpostaviti da je očuvana koeficijentom očuvanosti $\alpha = 0.55$. 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.

Račun provodimo u slučaju gdje koristimo kraći upit (RQVTFSKRRNGLLKKA koji je duljine 16). Podaci su motivi duljine 16 pa računamo očekivanu udaljenost dvaju 16-dimenzionalnih vektora. Neka su $X = (x_1, x_2, \ldots, x_{16})$ i $Y = (y_1, y_2, \ldots, y_{16})$ dva promatrana niza. 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 nikakvih saznanja o aminokiselinama, pretpostavimo da se radi o nekim prosječnim aminokiselinama te označimo s $\bar{a}_i$ i $\bar{a}_j$ aminokiseline koje pripadaju prosječnoj distribuciji aminokiselina pa 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:

$$\mathbb{E}\left[\left(\bar{a}_i - \bar{a}_j\right)^2\right] = \sum_{k=1}^{20} p_k \sum_{i,j=1}^{20} \left(a_i^k - a_j^k\right)^2 p_i^k p_j^k = 14.1062$$

Sada slijedi da je:

$$\mathbb{E}\left[d^2(X,Y)\right] = 16 \cdot \mathbb{E}\left[\left(\bar{a}_i - \bar{a}_j\right)^2\right] = 16 \cdot 14.1062$$

Zaključujemo da je očekivani kvadrat udaljenosti dva 16-dimenzionalna vektora aminokiselina jednak $16 \cdot 14.1062$ pa korjenovanjem dobijemo da je očekivana udaljenost jednaka $\sqrt{16} \cdot 3.7558$.
Sada iz teorema 3.1.2 slijedi da je $r = \frac{\sqrt{16} \cdot 3.7558}{\sqrt{2}} = \sqrt{8} \cdot 3.7558$

Analogan račun provodimo za upit duljine 22 (KTNRQVTFSKRRNGLLKKAYEL). Tada dobijemo da je $r = \frac{\sqrt{22} \cdot 3.7558}{\sqrt{2}} = \sqrt{11} \cdot 3.7558$

Uočimo da je radijus proporcionalan standardnoj devijaciji što povlači da je radijus prije i nakon standardizacije podataka proporcionalan standardnoj devijaciji prije i nakon standardizacije. Neka su r_{old} i r_{new} radijusi podataka prije i poslije standardizacije, a std_{old} i std_{new} standardne devijacije podataka prije i poslije standardizacije, redom. Tada je

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

Detaljniji raspis vezan za procijenjeni radijus kugle može se proučiti u [8].

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 sada je spomenut način na koji ćemo naći „idealan" slučaj koji ćemo gledati kao gornju granicu svih sljedećih rezultata. To postižemo postavljanjem težišta pravih pozitivaca za središte kugle i traženjem optimalnog radijusa. Također smo razmotrili kako izračunati procijenjeni radijus, koji nam omogućuje stvaranje kugle oko težišta pravih pozitivaca. Ipak, u oba slučaja moramo poznavati podatke i znati koji su pravi pozitivci. U cilju nam je pronaći središte kugle nenadziranom klasifikacijom, odnosno napraviti kuglu s procijenjenim radijusom ne oslanjajući se na prethodno znanje o podacima.

Nakon što je IGLOSS server dao svoje kandidate, podatke smo prebacili u vektorski prostor i standardizirali. Podaci postaju 80-dimenzionalni vektori koji 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 ideje je da će u „najgušćoj" kugli biti zadržani pravi pozitivci, a što je više moguće 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. Ova metoda je izuzetno efikasna, pružajući brze i pouzdane rezultate. Na ovaj način smo uspješno identificirali kuglu s procijenjenim radijusom, čak i bez prethodnog poznavanja podataka, što dodatno potvrđuje njezinu brzu i učinkovitu primjenu.

Grafički prikaz podataka

Za bolje shvaćanje i predodžbu strukture podataka koristimo alat za vizualizaciju visokodimenzionalnih podataka *t-Distributed Stochastic Neighbor Embedding* skraćeno t-SNE. Radi se o statističkoj metodi koja smanjuje dimenziju podataka tako da očuva lokalnu strukturu. To znači da točke koje su blizu jedna drugoj u smanjenoj dimenziji tumačimo kao bliske i u visoko-dimenzionalnom prostoru. Na taj način, pri prijelazu na dvije dimenzije, možemo ilustrirati mogućnost razdvajanja pravih pozitivaca od lažnih.

Ovo potpoglavlje blisko slijedi izvore [5] i [8].

3.2 Rezultati

U ovom radu promatramo rezultate dobivene na sedam različitih proteoma:

- Talijin uročnjak (lat. *Arabidopsis thaliana*)
- Divlji kupus (lat. *Brassica oleracea var. oleracea*)
- Crvena škripavica (lat. *Capsella rubella*)
- Krumpir (lat. *Solanum tuberosum*)
- Soja (lat. *Glycine max*)
- Crvotočina (lat. *Selaginella moellendorffii*)
- Amborela (lat. *Amborella trichopoda*)

Kod svih proteoma korišteni su upiti RQVTFSKRRNGLLKKA i KTNRQVTFSKRRNGLLKKAYEL za iterativno pretraživanje proteoma. Promatramo i nekoliko skala pretraživanja i koeficijenata očuvanosti, a posebno smo izdvojili što se dogoditi kad su koeficijenti očuvanosti jednaki 0.55 i 0.58. Iako su u poglavlju 1.3 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 preostale proteome 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 negativnim stanjem. 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 skale pretraživanja u rasponu od 2 do 8 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 i radijus kugle (r).

Arabidopsis thaliana

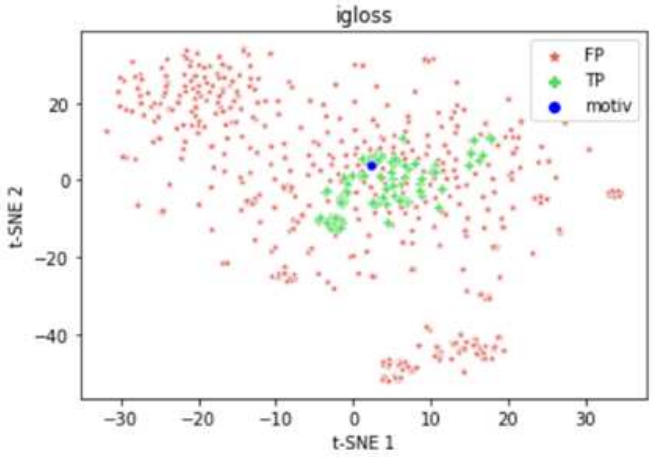
Talijin uročnjak (lat. *Arabidopsis thaliana*) je mala jednogodišnja cvjetnica iz porodice krstašica. Poznata je kao modelna biljka u znanosti, ima svoje prednosti zbog svojih malih dimenzija i brzog razvojnog ciklusa. To omogućava znanstvenicima brže provođenje eksperimenata i istraživanje genetskih aspekata biljnog razvoja. Posebno je važna u istraživanjima gena koji kontroliraju razvoj biljaka, uključujući gene MADS-box. 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. Dulji upit, skala pretraživanja je 5, a koeficijent očuvanosti je 0.55.

Model	TPR	PPV	F_1-score	r
kugla oko težišta TP s optimalnim r	0.887	0.991	0.936	8.567
kugla oko težišta TP s procijenjenim r	0.879	1.0	0.936	8.567
metoda prolaska po svim točkama	0.798	1.0	0.888	8.567
metoda prolaska po svim točkama + povećan radijus za 0.8	0.887	0.991	0.936	9.367

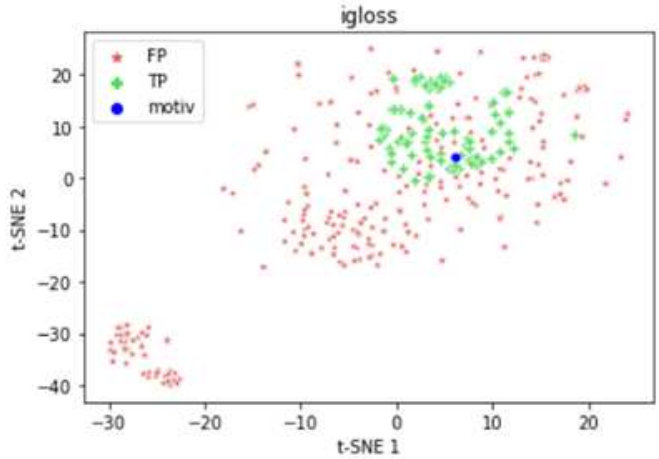
Slični rezultati dobiju se kada je koeficijent sličnosti 0.58.



Slika 3.2: Talijin uročnjak t-SNE prikaz

2. Dulji upit, skala pretraživanja je 6, a koeficijent očuvanosti 0.55.

Model	TPR	PPV	F_1-score	r
kugla oko težišta TP s optimalnim r	0.887	0.991	0.936	8.84
kugla oko težišta TP s procijenjenim r	0.879	1.0	0.936	8.751
metoda prolaska po svim točkama	0.798	1.0	0.888	8.751
metoda prolaska po svim točkama + povećan radijus za 0.8	0.887	0.991	0.936	9.551



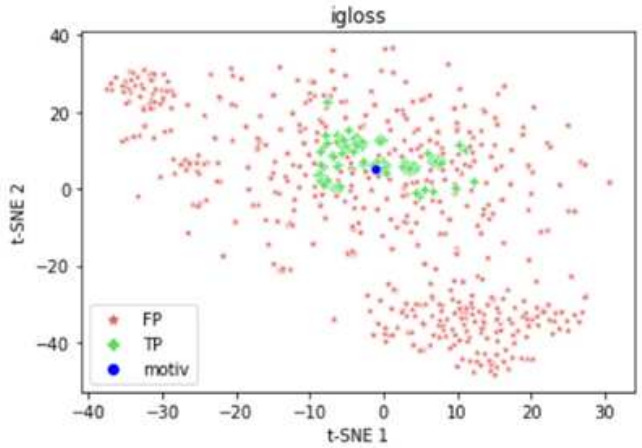
Slika 3.3: Talijin uročnjak t-SNE prikaz

Vidimo da je jedina razlika u odnosu na prijašnje rezultate u radijusu. Slike t-SNE prikazi poprilično slično izgledaju neovisno o skali pretraživanja.

Pogledajmo što se dogodi kad je u pitanju kraći upit.

3. Kraći upit, skala pretraživanja je 5, a koeficijent očuvanosti 0.58.

Model	TPR	PPV	F_1-score	r
kugla oko težišta TP s optimalnim r	0.806	1.0	0.893	6.97
kugla oko težišta TP s procijenjenim r	0.815	0.962	0.882	7.258
metoda prolaska po svim točkama	0.782	0.96	0.862	7.258
metoda prolaska po svim točkama + povećan radijus za 0.8	0.831	0.88	0.855	8.058



Slika 3.4: Talijin uročnjak t-SNE prikaz

Iz gornjih tablica možemo primijetiti za obje skale pretraživanja F_1-score je vrlo visok, tj. oko 0.9. Također, preciznost je često jednaka 1 što znači da kugla uspješno uzima bitne podatke. Time vidimo da i nadzirana i nenadzirana metoda jako dobro rade, uspijevaju pronaći kuglu koja će zadržati prave pozitivce, a eliminirati što više lažnih. Pomoću Slike 3.4 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.

Brassica oleracea var. oleracea

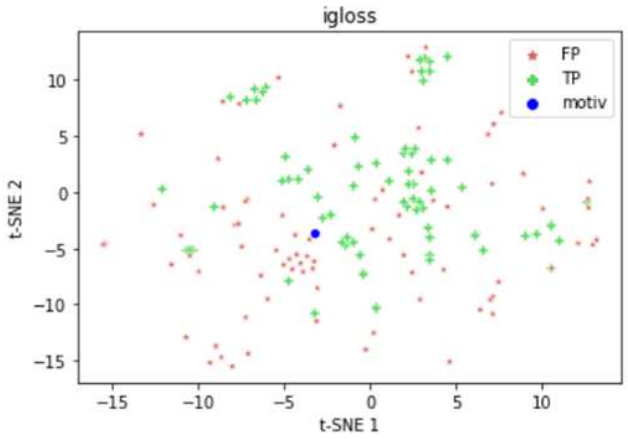
Divlji kupus (lat. *Brassica oleracea var. oleracea*) podvrsta je *Brassica oleracea*. To je divlji predak mnogih poznatih i široko uzgajanih sorti povrća, uključujući kupus, kelj, brokulu i cvjetaču. Kultivacija ove podvrste divljeg kupusa kroz generacije dovela je do razvoja raznolikog niza uzgojenih oblika koje danas poznajemo.

Slika 3.5: Divlji kupus

1. Dulji upit, skala pretraživanja je 5, a koeficijent sličnosti 0.55.

Model	TPR	PPV	F_1-score	r
kugla oko težišta TP s optimalnim r	0.467	0.612	0.53	8.94
kugla oko težišta TP s procijenjenim r	0.467	0.612	0.53	9.118
metoda prolaska po svim točkama	0.447	0.602	0.513	9.118
metoda prolaska po svim točkama + povećan radijus za 0.8	0.48	0.619	0.541	9.318

Ovdje uvodimo i novu mjeru nTPR što označava broj TP unutar naših pravih podataka, tj. gledamo je li problem u IGLOSS serveru, da dosta TP nije prepoznao, da su neki FP zapravo pritajeni MADS-box. U ovom slučaju je nTPR = 0.612, što je veće nego TPR pa vidimo da se potencijalni problem nalazi i u IGLOSS-u.

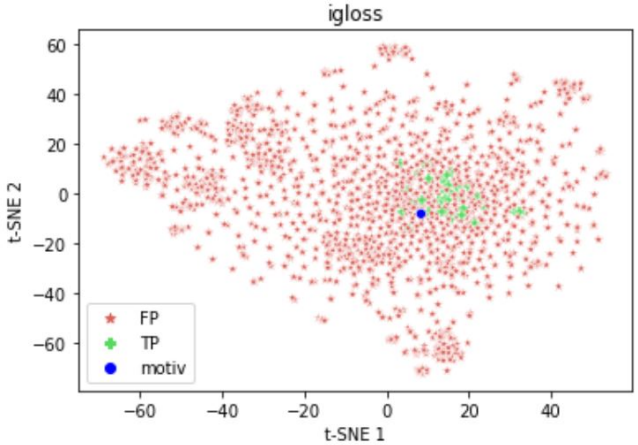


Slika 3.6: Divlji kupus t-SNE prikaz

2. Dulji upit, skala pretraživanje je 3, a koeficijent sličnosti 0.55.

Model	TPR	PPV	F_1-score	r
kugla oko težišta TP s optimalnim r	0.678	0.669	0.673	8.82
kugla oko težišta TP s procijenjenim r	0.461	0.609	0.524	7.242
metoda prolaska po svim točkama	0.414	0.583	0.485	7.242
metoda prolaska po svim točkama + povećan radijus za 0.2	0.474	0.615	0.535	7.442

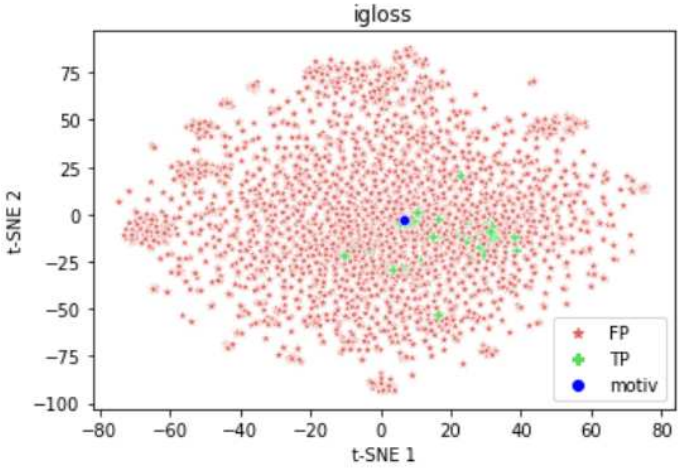
Rezultati skale pretraživanja 3 čine se relativno sličnima prijašnjim rezultatima kada je u pitanju bila skala 5.



Slika 3.7: Divlji kupus t-SNE prikaz

3. Kraći upit, skala pretraživanje je 3, a koeficijent sličnosti 0.55.

Model	TPR	PPV	F_1-score	r
kugla oko težišta TP s optimalnim r	0.658	0.662	0.66	6.88
kugla oko težišta TP s procijenjenim r	0.658	0.613	0.635	7.078
metoda prolaska po svim točkama	0.632	0.596	0.613	7.078
metoda prolaska po svim točkama + povećan radijus za 0.2	0.691	0.303	0.421	7.278



Slika 3.8: Divlji kupus t-SNE prikaz

Rezultati ne izgledaju slično rezultatima dobivenim na temelju proučavanja talijinog uročnjaka. Iz grafičkog prikaza je vidljiva velika razlika, a i brojke (npr. F_1-score) su mnogo niže.

Capsella rubella

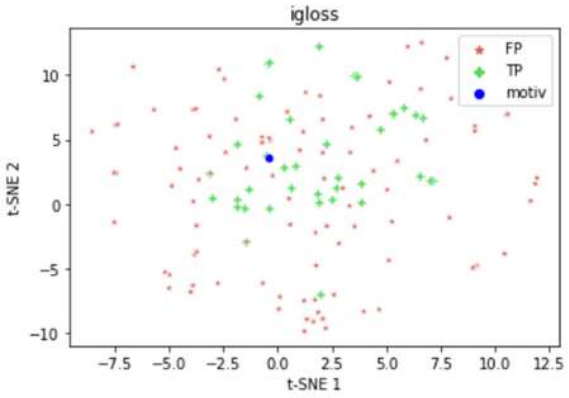
Crvena škripavica (lat. *Capsella rubella*) bliski je rođak biljke talijin uročnjak. S obzirom na to, očekujemo i relativno slične rezultate. Ovu vrstu vrlo često možemo pronaći u mediteranskom području. Također se koristi kao model u proučavanju evolucije kompatibilnosti biljaka.

Slika 3.9: Crvena škripavica

1. Dulji upit, skala pretraživanja je 3, a koeficijent očuvanosti 0.55.

Model	TPR	PPV	F_1-score	r
kugla oko težišta TP s optimalnim r	0.602	0.723	0.657	8.96
kugla oko težišta TP s procijenjenim r	0.513	0.69	0.589	8.268
metoda prolaska po svim točkama	0.398	0.634	0.489	8.268
metoda prolaska po svim točkama + povećan radijus za 0.2	0.522	0.694	0.596	8.468

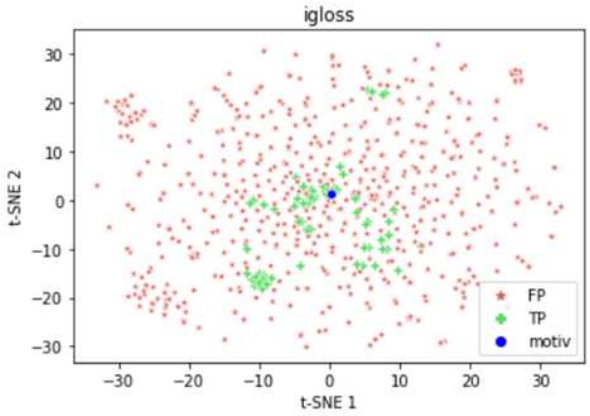
nTPR = 0.752 što nam kao i u divljeg kupusa ukazuje na nepouzdanost IGLOSS-a.



Slika 3.10: Crvena škripavica t-SNE prikaz

2. Kraći upit, skala pretraživanja je 3, koeficijent očuvanosti 0.55, a nTPR = 0.805.

Model	TPR	PPV	F_1-score	r
kugla oko težišta TP s optimalnim r	0.717	0.659	0.686	7.43
kugla oko težišta TP s procijenjenim r	0.717	0.653	0.684	7.504
metoda prolaska po svim točkama	0.655	0.627	0.641	7.504
metoda prolaska po svim točkama + povećan radijus za 0.2	0.743	0.4	0.52	7.704



Slika 3.11: Crvena škripavica t-SNE prikaz

Primjećujemo sličnost u rezultatima između divljeg kupusa i crvene škripavice. S druge strane, rezultati na talijinom uročnjaku mnogo drukčiji izgledaju. Već ovdje pada hipoteza da se divlje biljke ponašaju slično talijinom uročnjaku.

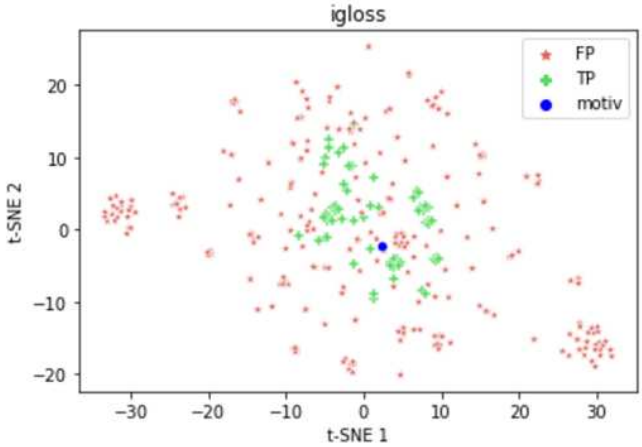
Solanum tuberosum

Krumpir (lat. *Solanum tuberosum*) je trajna zeljasta biljka iz porodice pomoćnica. Krumpir je jedna od najvažnijih i najrasprostranjenijih biljnih kultura na svijetu, te igra ključnu ulogu u prehrani mnogih ljudi. Proteom ima 56103 proteina i duljina liste CP je 160. U sljedećim tablicama nalaze se rezultati za krumpir s proteomom.

Slika 3.12: Krumpir

1. Dulji upit, skala pretraživanja je 7, koeficijent očuvanosti 0.55.

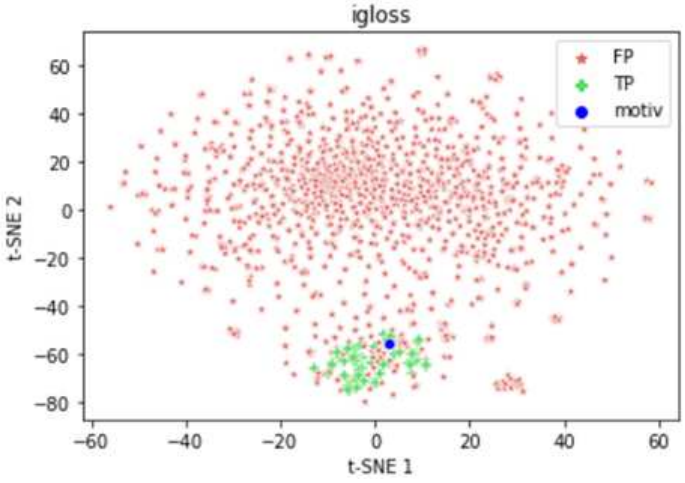
Model	TPR	PPV	F_1-score	r
kugla oko težišta TP s optimalnim r	0.588	0.696	0.637	8.92
kugla oko težišta TP s procijenjenim r	0.588	0.696	0.637	9.126
metoda prolaska po svim točkama	0.575	0.697	0.63	9.126
metoda prolaska po svim točkama + povećan radijus za 0.8	0.613	0.641	0.626	9.926



Slika 3.13: Krumpir t-SNE prikaz

2. Dulji upit, skala pretraživanja je 6, a koeficijent očuvanosti 0.58.

Model	TPR	PPV	F_1-score	r
kugla oko težišta TP s optimalnim r	0.544	0.702	0.613	8.84
kugla oko težišta TP s procijenjenim r	0.588	0.718	0.647	9.245
metoda prolaska po svim točkama	0.544	0.702	0.613	9.245
metoda prolaska po svim točkama + povećan radijus za 0.8	0.606	0.678	0.64	10.045



Slika 3.14: Krumpir t-SNE prikaz

Iz gornjih tablica možemo primijetiti za obje skale pretraživanja, odnosno bez obzira na veličinu odgovora, kod kugle oko težišta pravih pozitivaca s optimalnim radijusom F_1-score nije veći od 0.637. 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. Početne dvije pretpostavke nisu ispunjene u slučaju krumpira, to je vidljivo iz 3.13.

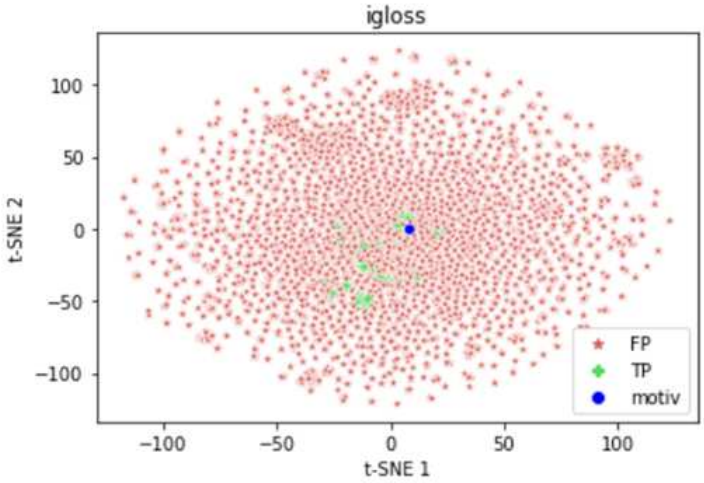
Glycine max

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 3.15: Soja

1. Dulji upit, skala pretraživanja je 2, koeficijent očuvanosti 0.55.

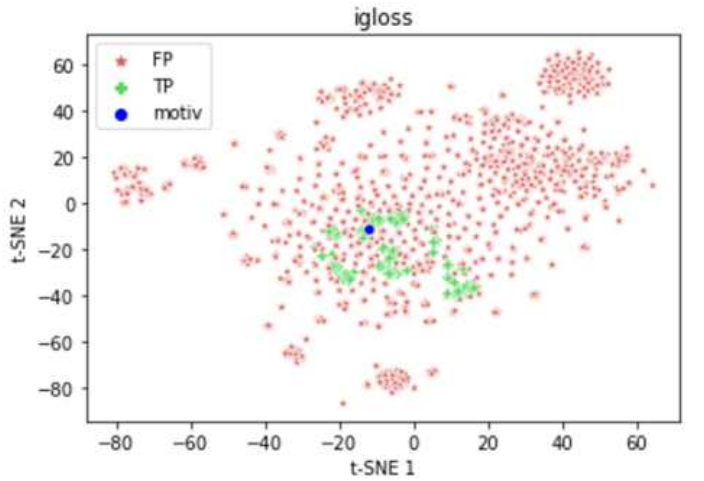
Model	TPR	PPV	F_1-score	r
kugla oko težišta TP s optimalnim r	0.82	0.768	0.792	8.04
kugla oko težišta TP s procijenjenim r	0.824	0.748	0.784	8.373
metoda prolaska po svim točkama	0.78	0.766	0.773	8.373
metoda prolaska po svim točkama + povećan radijus za 0.8	0.863	0.639	0.734	9.173



Slika 3.16: Soja t-SNE prikaz

2. Dulji upit, skala pretraživanja je 5, a koeficijent očuvanosti 0.55.

Model	TPR	PPV	F_1-score	r
kugla oko težišta TP s optimalnim r	0.824	0.764	0.793	8.78
kugla oko težišta TP s procijenjenim r	0.824	0.765	0.793	8.784
metoda prolaska po svim točkama	0.776	0.768	0.772	8.784
metoda prolaska po svim točkama + povećan radijus za 0.2	0.834	0.7	0.762	8.984



Slika 3.17: 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. 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 otprilike 3% od „idealnog" rezultata, pitamo se koliko metoda dobro uzima bitne podatke. Iz Slike 3.17 vidimo da je jako veliki broj crvenih točaka i da su zelene točke jako raspršene i grupiraju se na više dijelova. Vidimo da se soja i krumpir slično ponašaju.

Selaginella moellendorffii

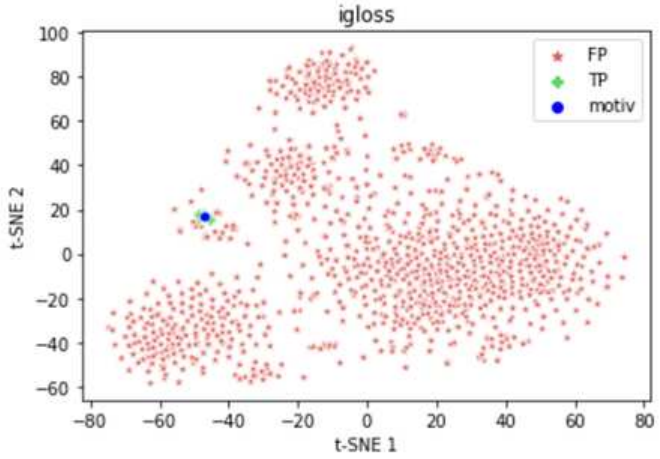
Crvotočina (lat. *Selaginella*) je uobičajeni naziv za skupinu biljaka koje pripadaju porodici Selaginellaceae, a lat. *Selaginella moellendorffii* je vrsta iz te skupine. Rod (lat. *Selaginella*) su primitivne vaskularne biljke. Nisu prave mahovine jer imaju vaskularno tkivo i razmnožavaju se putem spora, što ih čini srodnima papratima i drugim vaskularnim biljkama.

Slika 3.18: Crvotočina

1. Dulji upit, skala pretraživanja je 7, a koeficijent očuvanosti 0.55.

Model	TPR	PPV	F_1-score	r
kugla oko težišta TP s optimalnim r	0.425	1.0	0.596	8.96
kugla oko težišta TP s procijenjenim r	0.5	1.0	0.667	9.026
metoda prolaska po svim točkama	0.25	1.0	0.4	9.026
metoda prolaska po svim točkama + povećan radijus za 0.8	0.4	1.0	0.57	9.826

nTPR = 0.95. Opet se čini da je problem u IGLOSS-u te da nije prepoznao sve TP. Iz Slike 3.19 čak se čini da bi nTPR trebao biti jednak 1, ali rezultat je ipak sasvim zadovoljavajući s obzirom da je PPV = 1.

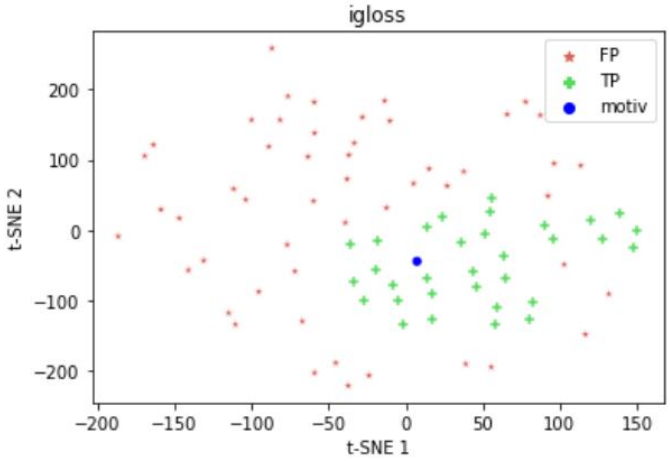


Slika 3.19: Crvotočina t-SNE prikaz

2. Dulji upit, skala pretraživanja je 8, a koeficijent očuvanosti 0.55.

Model	TPR	PPV	F_1-score	r
kugla oko težišta TP s optimalnim r	0.9	1.0	0.947	8.23
kugla oko težišta TP s procijenjenim r	0.9	1.0	0.947	9.026
metoda prolaska po svim točkama	0.8	1.0	0.889	9.026
metoda prolaska po svim točkama + povećan radijus za 0.8	0.9	1.0	0.947	9.826

Povećanje radijusa daje bolji F_1-score.



Slika 3.20: Crvotočina t-SNE prikaz

Amborella trichopoda

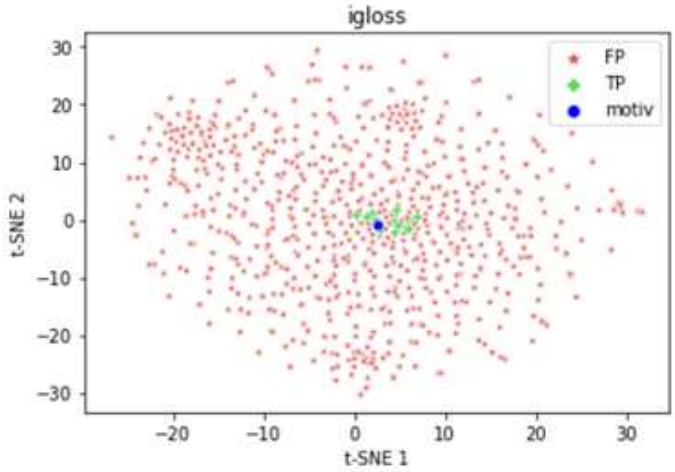
Amborela (lat. *Amborella trichopoda*) je rijetka i prapovijesna biljka koja se smatra najbližim živim srodnicima svim drugim cvjetnicama (angiospermama). Ova biljka predstavlja značajan fokus u botaničkim istraživanjima zbog svoje ključne uloge u razumijevanju evolucije i porijekla cvjetnica.

Slika 3.21: Amborela

1. Dulji upit, skala pretraživanja je 3, a koeficijent očuvanosti 0.55.

Model	TPR	PPV	F_1-score	r
kugla oko težišta TP s optimalnim r	0.655	0.594	0.623	8.37
kugla oko težišta TP s procijenjenim r	0.655	0.594	0.623	8.474
metoda prolaska po svim točkama	0.517	0.5	0.508	8.474
metoda prolaska po svim točkama + povećan radijus za 0.8	0.621	0.353	0.45	9.274

Povećanje radijusa ovdje ne izgleda poželjno. nTPR = 0.828.

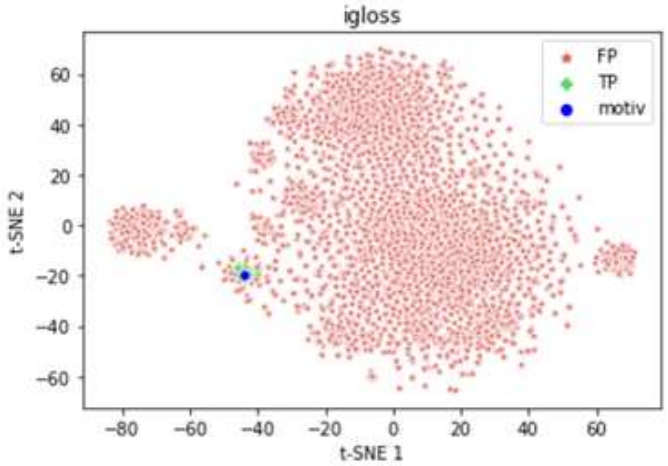


Slika 3.22: Amborela t-SNE prikaz

2. Dulji upit, skala pretraživanja je 5, a koeficijent očuvanosti 0.55.

Model	TPR	PPV	F_1-score	r
kugla oko težišta TP s optimalnim r	0.62	0.6	0.61	8.68
kugla oko težišta TP s procijenjenim r	0.621	0.6	0.61	9.066
metoda prolaska po svim točkama	0.517	0.556	0.536	9.066
metoda prolaska po svim točkama + povećan radijus za 0.8	0.517	0.417	0.462	9.866

nTPR = 0.724, što je malo niži broj od nTPR-a od amborele i skale pretraživanja 3.



Slika 3.23: Amborela t-SNE prikaz

Amborela se ne ponaša točno onako kako smo očekivali. Očekivan je puno viši nTPR, posebice ako promatramo sliku 3.23 koja sugerira da će nTPR biti jako blizu 1. Pretpostavljamo da među FP postoje slični transkripcijski faktori koje je IGLOSS krivo označio. Njih je IGLOSS krivo povukao, oni su na slikama označeni kao FP jer nisu anotirani kao MADS-box. Mogućnost je da postoji slična familija transkripcijskih faktora koja nije istražena, a druga mogućnost je da ih je trebalo označiti kao MADS-box, ali ih nitko nije dovoljno istražio. Problem je ili u anotaciji ili interpretaciji tih podataka.

Analiza rezultata

U ovom radu istražuje se ponašaju li se u smislu MADS-boxa divlje biljke na jedan način, a uzgajane na drugi. Prilikom izrade ovog rada krenuli smo s pretpostavkom da će distribucija mutacija u MADS-box familiji transkripcijskih faktora kod divljih biljaka izgledati kao talijin uročnjak, a u kultiviranih neće (zbog rezultata dobivenih u [5]).

Proučavajući grafičke rezultate, uočavamo neka grupiranja (pravi pozitivci koji su označeni kao zelene točkice mogu se staviti u jednu kuglu). Grupiranje je moguće ako su točkice dovoljno blizu. Pretpostavka je da bliskost u euklidskom prostoru direktno korespondira s bliskosti motiva u evoluciji. Ako postoji više takvih grupiranja, to zovemo klasterima. Ako su klasteri vidljivi, tim rezultatima puno bolje odgovara model s više kugala nego jednom.

Distribucija MADS-box gena u divljeg kupusa i crvene škripavice izgleda kao da je došla od više različitih predaka: možda su biljke bile pod stresom ili su rasle u blizini drugog srodnog bilja pa su postale drukčije u smislu MADS-box familije. Njihova distribucija mutacija u MADS-box motivu je prešla iz očekivanog divljeg modela evolucije u križani oblik. Postoji mogućnost da je divlji kupus oprašen npr. industrijskim zeljem, a za crvenu škripavicu se objašnjenje potencijalno krije u križanju s nekom uzgajanom biljkom. Mogući dio objašnjenja zašto se divlji kupus ponaša drukčije od talijinog uročnjaka saznajemo iz [9]. Navodi se da je gen za cvjetanje (FLC) MADS-box gen koji djeluje kao glavni regulator vremena cvatnje kod talijinog uročnjaka. FLC inhibira cvjetanje izravnim potiskivanjem aktivnosti središnjih promotora cvjetanja, SOC1, cvatnog lokusa D (FD), i cvatnog lokusa T (FT) u talijinog uročnjaka. S druge strane, usjevi divljeg kupusa uzgajaju se radi sjemena, stoga potpuno blokiranje cvjetanja nije prikladno. Umjesto toga, cilj je proizvesti zimske i proljetne sorte koje omogućuju rast i proizvodnju sjemena pod različitim uvjetima okoline i koji omogućuju prilagodbu i ubrzanje ciklusa rasta.

Soja je na oba upita slična, vide se krpice, srodne nakupine MADS-box gena, očekujemo da jedna nakupina dolazi iz jednog izvora, a soja je industrijska biljka više puta križana pa je rezultat očekivan za nju. Krumpir (pogotovo skala 7) sličan je soji. Vidljivo je grupiranje pravih pozitivaca u više klastera i njihova raspršenost. Mogući uzrok tog fenomena je što se radi o biljkama koje su puno puta križane s raznim sortama.

Dvije divlje biljke koje se ponašaju kao talijin uročnjak su amborela i crvotočina - izgledaju skroz stisnuto.

Uz sitne modifikacije, nenadzirana metoda radi gotovo jednako dobro kao i nadzirana. S druge strane, pretpostavke za primjenu metode nisu sasvim ispunjene i po svemu sudeći tehnika clusteringa bi bila dobar dodatak u primjeni ove metode na uzgajane biljke. Evolucija MADS-box familije transkripcijskih faktora pokazuje se kao relativno kompliciran fenomen: inicijalna pretpostavka da će divlje biljke u smislu MADS-box familije slijediti predložak iz talijinog uročnjaka se pokazuje kao većinom nepotpuna ili netočna.

Dulji upit pokazao se relativno sličan kraćem upitu. Kod talijinog uročnjaka, dulji upit

se pokazao nešto boljim u odnosu na kraći pa se tu krije motivacija zašto smo se većinom bavili duljim upitom.

Pojmovi iz ovog poglavlja preuzeti su iz izvora [5], [7], [8], te iz izvora [12]-[27].

Bibliografija

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

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

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

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

[5] J. Radnić, *Klasifikacija proteinskih fragmenata*, Diplomski rad, Sveučilište u Zagrebu, Prirodoslovno-matematički fakultet (Matematički odsjek), 2023.

[6] UniProt, https://www.uniprot.org/.

[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] D. Horvat, *Proteinski motivi i klasifikacija u proteinske familije*, Diplomski rad, Sveučilište u Zagrebu, Prirodoslovno-matematički fakultet (Matematički odsjek), 2023.

[9] Susanne Schilling, Sirui Pan, Alice Kennedy, Rainer Melzer, *MADS-box genes and crop domestication: the jack of all traits*, Journal of Experimental Botany, Volume 69, Issue 7, 16 March 2018, Pages 1447–1469, https://doi.org/10.1093/jxb/erx479.

[10] Gramzow, L., Theissen, G. *A hitchhiker's guide to the MADS world of plants*, Genome Biol 11, 214 (2010), https://doi.org/10.1186/gb-2010-11-6-214.

[11] 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.

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

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

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

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

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

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

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

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

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

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

[22] https://en.wikipedia.org/wiki/Brassica_oleracea

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

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

[25] https://en.wikipedia.org/wiki/Capsella_rubella

[26] https://en.wikipedia.org/wiki/Selaginella

[27] https://en.wikipedia.org/wiki/Amborella

Sažetak

Ovaj diplomski rad bavi se promatranjem MADS-box familije transkripcijskih faktora. Cilj rada je povećati točnost metode pronalaženja MADS-box motiva u biljnim proteomima i, ujedno, proučiti filogenetsku strukturu MADS-box familije.

Promatrano je sedam biljnih proteoma, gdje se MADS-box familiju tražilo s dva karakteristična motiva različitih duljina. Ispostavlja se da je filogenetska struktura MADS-box familije nezavisna od izbora motiva za pretragu. Isto tako, djelomično je potvrđeno poznato opažanje o različitoj strukturi MADS-box familije u divljih i uzgajanih biljaka, uz nekoliko značajnih izuzetaka.

Summary

This thesis examines the MADS-box family of transcription factors. The aim of the thesis is to improve the accuracy of the search method of MADS-box motifs in plant proteomes and, simultaneously, to study the phylogenetic structure of the MADS-box family.

We examined seven plant proteomes, where we searched for the MADS-box family using two characteristic motifs of different lengths. It turns out that the phylogenetic structure of the MADS-box family is independent of the motif selection for the search. Likewise, the known observation of the different structure of the MADS-box family in wild and cultivated plants has been partially confirmed, with a few significant exceptions.

Životopis

Rođena sam 10. 1. 2000. u Splitu. Osnovnoškolsko obrazovanje završila sam 2014. godine u Osnovnoj školi Pujanki u Splitu. Pohađala sam Treću gimnaziju u Splitu (MIOC), koju sam završila 2018. godine. Iste godine preselila sam se u Zagreb te upisala preddiplomski studij Matematika na Prirodoslovno-matematičkom fakultetu. Završetkom preddiplomskog studija stekla sam titulu sveučilišnog prvostupnika matematike 2021. godine kada sam upisala i diplomski sveučilišni studij Matematička statistika na istom fakultetu.

Tijekom studija bila sam demonstratorica iz kolegija Statistika i Teorija brojeva. U slobodno vrijeme bila sam dio ženske futsal sekcije za fakultet, trenirala trčanje te trčala kros u sklopu PMF sekcije. Za izniman uspjeh u izvannastavnim aktivnostima uručena mi je prodekanova nagrada za najbolje studente Matematičkog odsjeka za akademsku godinu 2020./2021., a istu sam nagradu za izniman uspjeh, za najuspješnije studente završnih godina svih preddiplomskih, diplomskih i integriranog studija, dobila za akademsku godinu 2021./2022. Od prosinca 2022. godine zaposlena sam u automobilskoj diviziji u IT tvrtci *Visage Technologies* na poziciji student developer.