SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Igor Keserin

ANALIZA PROTEINSKIH MOTIVA

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. _____

Sadržaj

Sadržaj **iii**

Uvod **1**

1 Matematička pozadina **2**
1.1 Linearna algebra 2
1.2 Teorija vjerojatnosti 5

2 Biološka pozadina **11**
2.1 Aminokiseline i struktura proteina 11
2.2 Proteinske domene i motivi 12
2.3 WRKY transkripcijski faktori 12

3 Metoda klasifikacije proteina **14**
3.1 Prikupljanje podataka 14
3.2 Metoda pretraživanja 14
3.3 Klasifikacija motiva 15
3.4 Ideja metode 15
3.5 Prelazak u euklidski prostor $\mathbb{R}^{5n}$ 16
3.6 Pronalaženje najgušće kugle 17
3.7 Uspješnost metode 19

4 Rezultati razvijene metode **21**
4.1 Subjekti testiranja 21
4.2 Rezultati analize WRKY motiva 22
4.3 Diskusija o WRKY motivima 27
4.4 Rezultati analize cinkovih prstiju 28
4.5 Diskusija o cinkovim prstima 31
4.6 Zaključak 31

Bibliografija **32**

Uvod

Proteini su velike, kompleksne molekule s važnom ulogom u našem tijelu. Ključni su za mnoštvo funkcija stanica. Sudjeluju u prijenosu raznih molekula, replikaciji DNA lanaca, komunikaciji stanica te pružaju potporu stanicama u obliku citoskeleta. Proteini su građeni od jednog ili više presavijanih lanaca aminokiselina (polipeptida) te mogu biti raznih dužina. Informacija o sintezi i redoslijedu aminokiselina u proteinima sadržana je unutar DNA organizma. Promjenom jedne aminokiseline u proteinu dobivamo sasvim novi protein sa moguće drugim svojstvima.

Skup svih proteina nekog organizma nazivamo njegovim proteomom. Evolucijski srodne grupe proteina svrstavamo u proteinske familije. Takvi proteini potiču od zajedničkog pretka te često imaju slične funkcije te slične nizove aminokiselina. Proteinske familije su često okarakterizirane kraćim nizovima aminokiselina koji su zajednički svim proteinima u familiji, te nizove zovemo proteinskim motivima. Generalizacija i modifikacija već poznatih motiva sve češća je tema u današnjoj literaturi.

U ovom radu promatrati ćemo transkripcijske faktore (TF), vrstu proteina odgovornih za brzinu transkripcije (prepisivanje genske informacije sadržane u DNA u RNA). Obavljaju svoju ulogu vezanjem na specifičan dio DNA niza, koji se zove karakteristični motiv. Uloga im je da omogućuju ili blokiraju aktivnost gena kako bi ti geni djelovali u potrebnim stanicama te u potrebnoj mjeri i određenom vremenu. Različite grupe transkripcijskih faktora odgovorne su za smrt stanica te njihov rast i diobu kao i za migraciju i organizaciju stanica tijekom embrionalnog razvoja.

Cilj rada testiranje je i primjena već razvijene metode pronalaženja i filtriranja motiva pripadnih proteinskih familija. U tu svrhu promatrati ćemo motive iz WRKY familije transkripcijskih faktora. Zbog kratke duljine i velike varijabilnosti, pripadnost motiva teško je za predvidjeti uporabom računala. Unatoč tome pokazati ćemo kako je razvijena metoda relativno uspješnija od tipičnog pretraživača lokalnog poravnavanja. Navesti ćemo matematičke pojmove te objasniti biološku pozadinu potrebnu za razvijanje metode. Nadalje, objasniti ćemo metodu korištenu za dobivanje rezultata te navesti i prokomentirati te rezultate.

Poglavlje 1

Matematička pozadina

Svi pojmovi iz ovog poglavlja preuzeti su iz izvora [4], [10], [5].

1.1 Linearna algebra

Definicija 1.1.1. *Neka je $\mathbb{F}$ neki skup na kojem su zadane binarne operacije zbrajanja $+ : \mathbb{F} \times \mathbb{F} \to \mathbb{F}$ i množenja $\cdot : \mathbb{F} \times \mathbb{F} \to \mathbb{F}$ koje imaju sljedeća svojstva:*

1. $\alpha + (\beta + \gamma) = (\alpha + \beta) + \gamma$;
2. $\exists 0 \in \mathbb{F}$ *sa svojstvom* $\alpha + 0 = 0 + \alpha = \alpha$, $\forall \alpha \in \mathbb{F}$;
3. $\forall \alpha \in \mathbb{F}$, $\exists - \alpha \in \mathbb{F}$ *tako da je* $-\alpha + \alpha = \alpha + (-\alpha) = 0$;
4. $\alpha + \beta = \beta + \alpha$, $\forall \alpha, \beta \in \mathbb{F}$;
5. $(\alpha \cdot \beta) \cdot \gamma = \alpha \cdot (\beta \cdot \gamma)$, $\forall \alpha, \beta, \gamma \in \mathbb{F}$;
6. $\exists 1 \in \mathbb{F} \setminus \{0\}$ *sa svojstvom* $1 \cdot \alpha = \alpha \cdot 1 = \alpha$, $\forall \alpha \in \mathbb{F}$;
7. $\forall \alpha \in \mathbb{F}$, $\alpha \neq 0$, $\exists \alpha^{-1} \in \mathbb{F}$ *tako da je* $\alpha \cdot \alpha^{-1} = \alpha^{-1} \cdot \alpha = 1$;
8. $\alpha \cdot \beta = \beta \cdot \alpha$, $\forall \alpha, \beta \in \mathbb{F}$;
9. $\alpha \cdot (\beta + \gamma) = \alpha \cdot \beta + \alpha \cdot \gamma$, $\forall \alpha, \beta, \gamma \in \mathbb{F}$.

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 $+ : V \times V \to \mathbb{F}$ 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. $\exists 0 \in V$ *sa svojstvom* $a + 0 = 0 + a = a$, $\forall a \in V$;
3. $\forall a \in V$, $\exists -a \in V$ *tako da je* $-a + a = a + (-a) = 0$;
4. $a + b = b + a$, $\forall a, b \in V$;
5. $(\alpha \cdot \beta) \cdot a = \alpha \cdot (\beta \cdot a)$, $\forall \alpha, \beta \in \mathbb{F}$, $\forall a \in V$;
6. $(\alpha + \beta) \cdot a = \alpha \cdot a + \beta \cdot a$, $\forall \alpha, \beta \in \mathbb{F}$, $\forall a \in V$;
7. $\alpha \cdot (a + b) = \alpha \cdot a + \alpha \cdot b$, $\forall \alpha \in \mathbb{F}$, $\forall a, b \in V$;
8. $1 \cdot a = a$, $\forall a \in V$.

Napomena 1.1.4. *Skup $\mathbb{R}^n$ s uobičajenim operacijama zbrajanja i množenja je vektorski prostor nad $\mathbb{R}$. Kažemo da je $(\mathbb{R}^n, +, \cdot)$* ***realan*** *vektorski prostor.*

Definicija 1.1.5. *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 1.1.6. *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. $\langle x, x \rangle = 0 \iff x = 0$;
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. $\langle \alpha x, y \rangle = \alpha \langle \alpha x, y \rangle$, $\forall x, y \in V$, $\forall \alpha in \mathbb{F}$;
5. $\langle x, y \rangle = \overline{\langle y, x \rangle}$, $\forall x, y \in V$.

Napomena 1.1.7. *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 1.1.8. *Vektorski prostor na kojem je definiran skalarni produkt zovemo* ***unitaran*** *prostor.*

Definicija 1.1.9. *Neka je V vektorski prostor.* ***Norma*** *na V je preslikavanje $\|\cdot\|: V \to \mathbb{R}$ definirano s*

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

Propozicija 1.1.10. *Norma na unitarnom prostoru V ima sljedeća svojstva:*

1. $\|x\| \geq 0$, $\forall x \in V$;
2. $\|x\| = 0 \iff x = 0$;
3. $\|\alpha x\| = |\alpha|\|x\|$, $\forall \alpha \in \mathbb{F}$, $\forall x \in V$;
4. $\|x + y\| \leq \|x\| + \|y\|$, $\forall x, y \in V$.

Napomena 1.1.11. *Svaka funkcija $\|\cdot\|: V \to \mathbb{R}$ na vektorskom prostoru V sa svojstvima iz propozicije 1.1.10 naziva se norma. Tada $(V, \|\cdot\|)$ zovemo* ***normirani prostor****.*

Napomena 1.1.12. *Norma koja potječe od kanonskog skalarnog produkta na $\mathbb{R}^n$, definiranog u napomeni 1.1.7, dana je formulom:*

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

Ova norma naziva se ***euklidska norma****.*

Definicija 1.1.13. *Neka je V normirani prostor.* ***Metrika*** *ili* ***udaljenost*** *na prostoru V je preslikavanje $d : V \times V \to \mathbb{R}$ definirano s:*

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

Propozicija 1.1.14. *Metrika na normiranom prostoru ima sljedeća svojstva:*

1. $d(x, y) \geq 0$, $\forall x, y \in V$;
2. $d(x, y) = 0 \iff x = y$, $\forall x, y \in V$;
3. $d(x, y) = d(y, x)$, $\forall x, y \in V$;

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

Napomena 1.1.15. *Svaka funkcija $d : X \times X \to \mathbb{R}$ na skupu X sa svojstvima iz propozicije 1.1.14 naziva se metrika ili udaljenost. Tada (X, d) zovemo* ***metrički prostor****.*

Napomena 1.1.16. *Metrika koja potječe od euklidske norme na $\mathbb{R}^n$ definirane u napomeni 1.1.12, dana je formulom*

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

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

Definicija 1.1.17. *Neka je (X, d) metrički prostor te neka je $a \in X$ i $r \in \mathbb{R}$, $r > 0$. Skup*

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

nazivamo ***otvorena kugla*** *u X sa središtem u a i radijusom r.*

Napomena 1.1.18. *U n-dimenzionalnom euklidskom prostoru $\mathbb{R}^n$ otvorena kugla sa središtem u a i radijusom r dana je sa:*

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

1.2 Teorija vjerojatnosti

Vjerojatnosni prostor

Definicija 1.2.1. ***Slučajni pokus*** *ili* ***slučajni eksperiment*** *je pokus čiji ishodi, tj. rezultati nisu jednoznačno određeni uvjetima u kojima izvodimo pokus.*

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 je σ-algebra skupova na Ω ako je:*

1. $\emptyset \in \mathcal{F}$;

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

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

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

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

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

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

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

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

Slučajna varijabla

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

Definicija 1.2.8. *Neka je sa $\mathcal{B}$ 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. Funkciju $X : \Omega \to \mathbb{R}$ zovemo **slučajna varijabla** (na Ω) ako je $X^{-1}(B) \in \mathcal{F}$ za proizvoljan $B \in \mathcal{B}$, tj. $X^{-1}(\mathcal{B}) \subset \mathcal{F}$.*

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

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

Napomena 1.2.12. *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.

Napomena 1.2.13. *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 1.2.14. *X je slučajna varijabla ako i samo ako su* X^+ *i* X^- *slučajne varijable.*

Teorem 1.2.15. *Neka je X nenegativna slučajna varijable 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 sa $\mathcal{K}$ skup svih jednostavnih slučajnih varijabli definiranih na Ω, a sa $\mathcal{K}_+$ skup svih nenegativnih funkcija iz $\mathcal{K}$.

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

Definicija 1.2.16. ***Matematičko očekivanje*** *od X ili, kraće,* **očekivanje** *od X označavamo sa:*

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

Neka je X ***nenegativna slučajna varijabla*** *definirana na Ω. Prema teoremu 1.2.15 postoji rastući niz $(X_n)_{n\in\mathbb{N}}$ nenegativnih slučajnih varijabli takav da je $X = \lim_{n\to\infty} \mathbb{E}[X_n]$ koji može biti jednak i $+\infty$.*

Definicija 1.2.17. ***Matematičko očekivanje*** *od X ili, kraće,* ***očekivanje*** *od X definira se sa:*

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

Neka je X proizvoljna slučajna varijabla na Ω. Tada postoje slučajne varijable $X^+,\ X^- \geq 0$ tako da $X = X^+ - X^-$.

Definicija 1.2.18. *Kažemo da* ***Matematičko očekivanje*** *od X postoji ako je barem jedna od veličina $\mathbb{E}[X^+],\ \mathbb{E}[X^-]$ konačna. Tada definiramo*

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

Teorem 1.2.19. *Osnovna svojstva matematičkog očekivanja*

1. *Ako $\mathbb{E}[X]$ postoji i neka $c \in \mathbb{R}$, tada postoji $\mathbb{E}[cX]$ i vrijedi*

$$\mathbb{E}[cX] = c\mathbb{E}[X].$$

2. *Ako je $X \leq Y$, tada je*

$$\mathbb{E}[X] \leq \mathbb{E}[Y].$$

u smislu da

$$\text{ako je } -\infty < \mathbb{E}[X], \text{ tada je } -\infty < \mathbb{E}[Y] \text{ i } \mathbb{E}[X] \leq \mathbb{E}[Y]$$

ili

$$\text{ako je } \mathbb{E}[Y] < \infty, \text{ tada je } \mathbb{E}[X] < \infty \text{ i } \mathbb{E}[X] \leq \mathbb{E}[Y].$$

3. *Ako $\mathbb{E}[X]$ postoji, tada je*

$$|\mathbb{E}[X]| \leq \mathbb{E}[|X|].$$

4. *Ako $\mathbb{E}[X]$ postoji, tada postoji $\mathbb{E}[X\mathcal{K}_A]$ za svako $A \in \mathcal{F}$. Ako je $\mathbb{E}[X]$ konačno, tada je $\mathbb{E}[X\mathcal{K}_A]$ konačno za svako $A \in \mathcal{F}$.*

5. Neka su X i Y nenegativne slučajne varijable. Tada vrijedi

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$

Definicija 1.2.20. *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 sa* $Var(X)$ *ili* σ_X^2 *na sljedeći način:*

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

Napomena 1.2.21. *Pozitivan drugi korijen iz varijance nazivamo* ***standardna devijacija*** *od X i označavamo s* σ_X.

Funkcija distribucije

Definicija 1.2.22. *Neka je X slučajna varijabla na* Ω. ***Funkcija distribucije od X*** *je funkcija* $F_X : \mathbb{R} \to [0, 1]$ *definirana sa:*

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

Napomena 1.2.23. *Ako je jasno o kojoj se slučajnoj varijabli radi, piše se F umjesto* F_X.

Teorem 1.2.24. *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 navedena svojstva zovemo **vjerojatnosna funkcija distribucije** ili kraće, **funkcija distribucije**.

Klasifikacija slučajnih varijabli

Definicija 1.2.25. *Slučajna varijabla X je* ***diskretna*** *ako postoji konačan ili prebrojiv skup* $D \subset \mathbb{R}$ *takav da je* $\mathbb{P}\{X \in D\} = 1$.

Diskretne slučajne varijable obično zadajemo tako da zadamo skup $D = \{x_1, x_2, \dots\}$ i brojeve $p_n = \mathbb{P}\{X \in x_n\}$, što zapisujemo u obliku:

$$X \sim \begin{pmatrix} x_1 & x_2 & \dots & x_n & \dots \\ p_1 & p_2 & \dots & p_n & \dots \end{pmatrix} \tag{1.2}$$

Tablicu (1.2) zovemo **distribucija** ili **zakon distribucije** slučajne varijable X. U (1.2) je $x_n \in \mathbb{R}$, $x_i \neq x_j$ za $i \neq j$, $p_n > 0$ i $\sum_n p_n = 1$.

Definicija 1.2.26. *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 1.2.27. *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*

$$F_X(x) = \int_{-\infty}^{x} f(t)d\lambda(t), \ x \in \mathbb{R}. \tag{1.3}$$

Ako je X neprekidna slučajna varijabla, tada se funkcija f iz (1.3) zove **funkcija gustoće vjerojatnosti** od X, tj. od njezine funkcije distribucije F_x, ili kraće, **gustoća** od X i označavamo je sa f_X.

Definicija 1.2.28. *Neka su μ, $\sigma \in \mathbb{R}$, σ¿0. Neprekidna slučajna varijabla X ima **normalnu** distribuciju s parametrima μ i σ^2 ako joj je gustoća f_X dana s:*

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

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

Napomena 1.2.29. *X ima **jediničnu normalnu distribuciju** ako je $X \sim N(0, 1)$, dakle*

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

Poglavlje 2

Biološka pozadina

2.1 Aminokiseline i struktura proteina

Aminokiseline su skupina organskih molekula, sadržana u svim živim bićima. Građene su od amino skupine ($-NH_2$) i karboksilne skupine ($-COOH$). Život na zemlji je kompleksan i različit, ali proteini svih živih organizama građeni su od 20 osnovnih aminokiselina. Navedene aminokiseline prikazane su u tablici 2.1.1.

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

Tablica 2.1.1: Aminokiseline

Aminokiseline se vežu (peptidnim vezama) u trodimenzionalne lance, proteine, čije duljine variraju, u prosjeku od 50 do 2000, dok se neki (Titin) sastoje od više od 30 000 aminokiselina. Niz aminokiselina u proteinu određen je nizom pripadnog gena koji je sadržan u gentičkom kodu organizma. Proteini međudjeluju s raznim molekulama, uključujući, s drugim proteinima, s lipidima, ugljikohidratima i s dijelovima DNA. Ključno pitanje

u istraživanju proteina je njihova evolucija, svrstavanjem proteina u proteinske familije cilj je kategorizirati njihovo zajedničko porijeklo. Klasifikacija proteina temelji se na uspoređivanju njihovih nizova aminokiselina, tako se svojstva novo otkrivenih proteina već mogu previdjeti na temelju njihove pripadnosti nekoj proteinskoj familiji.

2.2 Proteinske domene i motivi

Proteinski polipeptidni lanac podijeljen je na proteinske domene, u prosjeku duge od 50 do 250 aminokiselina. Svaka domena tvori trodimenzionalnu strukturu koja je presavijena neovisno od ostatka proteina. Proteini imaju više domena i iste domene mogu biti dio raznih proteina. DNA-vezujuća domena (DVD) nezavisna je i često dio većeg proteina sa drugim domenama, koje često reguliraju funkciju DNA-vezujuće domene. Jedna od uloga DNA-vezujuće domene regulacija je transkripcije putem transkripcijskih faktora.

Kraće podnizove aminokiselina (većinom do 20 aminokiselina) sa specifičnom zajedničkom ulogom u organizmu zovemo proteinski motivi, ili kraće motivi. Motivi su često evolucijski bolje očuvani dijelovi proteina te funkcijski djeluju kao zasebna jedinica te su time korisni za lakšu klasifikaciju pripadnih proteina. Tu činjenicu i mi koristimo u svrhu razvijanja nove metode.

2.3 WRKY transkripcijski faktori

Jedna od najvećih familija transkripcijskih faktora (kod biljaka) upravo su WRKY transkripcijski faktori (TF), dužine oko 60 aminokiselina. Igraju ključnu ulogu u ekspresiji gena (prepisivanje i prevođenje informacije iz gena u funkcionalni genski produkt) u embrionalnom razvoju, starenju, otpornosti na patogene i reakcijama na abiotski stres. Unatoč tome, njihova klasifikacija ostaje nejasna.

Ime im je dodijeljeno zbog WRKY motiva, koji su sastavni dio njihove DNA-vezajuće domene (DVD). Slično, dio njihove DVD čine i cinkovi-prsti (eng. *zinc-fingers*), manji motivi okarakterizirani s jednim ili više cinkovih iona koji stabiliziraju njihovu strukturu. Često različiti transkripcijski faktori dijele iste DVD.

Prijašnje klasifikacije razvrstavaju WRKY TF u tri grupe ovisno o broju WRKY domena i vrsti cinkovog prsta kojeg sadrže:

1. Grupa: sadrže dvije WRKY domene i svaka sadrži C2H2 cinkov prst,
2. Grupa: sadrže jednu WRKY domenu i C2H2 cinkov prst,
3. Grupa: sadrže jednu WRKY domenu i C2HC cinkov prst.

Međutim, u članku [6] otkriveno je da su skoro svi WRKY TF okarakterizirani motivima xWRKYGQK ili xWRKYGEK sa cinkovim prstima CxCxHTC ili CxCxHxH (gdje je x oznaka za bilo koju aminokiselinu). Također, kako bi ušli u trag njihovom evolucijskom porijeklu, WRKY transkripcijski faktori podijeljeni su u pet grana (eng. *clade*). Opet, slično, ovisno o broju WRKY domena i vrsti cinkovog prsta kojeg sadrže.

Poglavlje 3

Metoda klasifikacije proteina

Metoda je izvedena po uzoru na rad [5].

3.1 Prikupljanje podataka

U svrhu testiranja metode potrebno je prikupiti informacije o nizovima raznih proteina. Uz to nam trebaju i njihove anotacije, kako bi utvrdili pripadnost proteina proteinskim familijama. To smo napravili koristeći UniProt-ovu [2] (*Universal Protein*) bazu UniProtKB. Za nekoliko organizama smo preuzeli sve proteine s označenom pripadnosti tom organizmu i istim taksonomskim nazivom (eng. *taxonomy ID*). To nam je povoljno zbog većih uzoraka pa time i mnogobrojnijih rezultata na upite.

3.2 Metoda pretraživanja

Jedan od potrebnih alata za postizanje ciljeva rada je metoda pretraživanja motiva. Metoda se temelji na algoritmu poravnavanja (eng. *sequence alignment*). Ulazna komponenta je upit, motiv, kojim se pronalaze dovoljno slični podnizovi (iste duljine) iz duljeg niza aminokiselina (npr. proteina). Sličnost je definirana funkcijom sličnosti koja se može razlikovati u raznim varijantama metode. Razlikujemo globalne i lokalne metode poravnavanja. Za pretraživanje potrebnih motiva koristimo pretraživač IGLOSS [9] (kr. *iterative gapless local similarity search*), koji koristi lokalnu varijantu metode. IGLOSS dopušta biranje skale, time određujemo dozvoljenu sličnost upita i rezultata, što je veća skala to je veća sličnost. Također, možemo postaviti ključna mjesta u motivu, te će aminokiseline biti jače očuvane s manjom varijacijom.

Pretpostavljamo da motivi ne sadrže rupe, (eng. *gaps*) tada duljinu upita možemo označiti s n. U slučaju da pretraživač pronađe više motiva iz istog proteina, možemo pro-

matrati samo najsličniji. Međutim, kako u našem slučaju isti protein može sadržavati više motiva iste vrste, nećemo u tom smislu filtrirati rezultate upita.

3.3 Klasifikacija motiva

Potrebno je uvesti pojmove putem kojih ćemo mjeriti uspješnost metode. Proteine preuzete iz baze UniProt-a, za koje je anotirano da pripadaju promatranoj proteinskoj familiji označimo s CP (eng. *condition positive*). Slično, proteine iz iste baze za koje nije anotirano da pripadaju promatranoj proteinskoj familiji označimo s CN (eng. *condition negative*). Nadalje, proteine koje smo dobili putem pretraživača označimo sa P (eng. *positive*), a preostale sa N (eng. *negative*). Također, definiramo i presjeke navedenih skupova:

- TP (eng. *true positive*)=CP∩P,
- FP (eng. *false positive*)=CN∩P,
- TN (eng. *true negative*)=CN∩N,
- FN (eng. *false negative*)=CP∩N.

3.4 Ideja metode

U radu [11] razvijen je novi pristup problemu pretraživanja motiva. Rezultat razvijene metode podskup je skupa motiva dobivenih nekim pretraživačem. Dakle, metoda nije osmišljena kao samostalan alat klasifikacije, već kao filter postojeće skupine motiva. Motive biramo s ciljem da izaberemo što veći udio onih s anotacijom koja indicira na pripadnost željenoj familiji (TP) te pri tome želimo minimizirati udio onih koje smo dobili kao rezultat pretraživača, ali ne pripadaju promatranoj familiji po anotaciji (FP).

Željenu filtraciju postići ćemo promatranjem međusobne sličnosti motiva. Korištenjem rezultata dobivenih u članku [3], motive ćemo prevesti u $5n$-dimenzionalni euklidski prostor. Tako će se sličniji motivi preslikati u bliže točke u $\mathbb{R}^{5n}$. Takvim preslikavanjem omogućiti ćemo uporabe raznih matematičkih alata, ali i uvesti novi izvor greške. Zatim ćemo tražiti najgušće kugle, određivanjem optimalnog radijusa te optimalnog središta.

Napomenimo da u opisu metode naizmjenično koristimo izraze motivi ili točke, što je opravdano zbog njihovog bijekcijskog odnosa.

U opisanom postupku pretpostavka je da će motivi iz skupa TP biti po parovima gušće raspoređeni od drugih te da se okupljaju u kugle pa su kao takvi, lakši za identificirati. Valjanost te pretpostavke lako ćemo testirati mjerom kvalitete rezultata metode.

3.5 Prelazak u euklidski prostor $\mathbb{R}^{5n}$

Velik problem za rigorozne statističke analize nizova aminokiselina je tzv. problem metrike nizova (eng. *sequence metric problem*) tj. pouzdano preslikavanje nizova kako bi se istaknuli njihovi odnosi. Standardna notacija aminokiselina ne zadovoljava tu potrebu, npr. aminokiselina leucin (L) svojstvima je sličnija valinu (V) nego alaninu (A), iako je "abecedna udaljenost" leucina i valina veća.

Navedeni problem riješen je u članku [3], gdje je dano preslikavanje aminokiselina u pet faktora od kojih je svaki realan broj, pa je time definirano preslikavanje sa skupa aminokiselina u $\mathbb{R}^5$. Pravilo pridruživanja opisanog preslikavanja dano je u tablici 3.5.1.

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

Tablica 3.5.1: Preslikavanje aminokiselina u $\mathbb{R}^5$

Slično, ulančavanjem vektora iz $\mathbb{R}^5$, definirano je preslikavanje motiva duljine n u euklidski prostor $\mathbb{R}^{5n}$.

Kako bi uspješno primijenili metodu, podatke je također potrebno standardizirati. Naime, ako je varijanca po jednoj koordinati prevelika, udaljenost točaka biti će dominirana

tom koordinatom, time ćemo izgubiti željeni oblik kugle. Označimo s $\overline{x}$ aritmetičku sredinu te sa s standardnu devijaciju podataka. Također, kako bi izbjegli dijeljenje s jako malim brojem, standardnoj devijaciji dodajem 0.1. Radimo sljedeće transformacije podataka:

$$\frac{x_i - \overline{x}}{s + 0.1}.$$

3.6 Pronalaženje najgušće kugle

U ovom odjeljku opisati ćemo postupak korišten za određivanje "optimalnog" skupa motiva. Kao što smo već objasnili, to ćemo postići pronalaskom kugle te promatranjem motiva unutar te kugle. Po definiciji 1.1.17, kugla je jednoznačno određena njenim radijusom i središtem pa će se metoda svesti na prvo procjenu radijusa i zatim pronalazak središta.

Procjena radijusa

Najprije ćemo izračunati očekivanu udaljenost dva niza aminokiselina uzorkovanih iz α-koveksne kombinacije distribucija, gdje parametar $\alpha \in \langle 0, 1\rangle$ zadajemo *a priori*. Potrebno je definirati prosječnu distribuciju aminokiselina:

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

Pripadne vjerojatnosti označimo sa $r_i,\ i \in \{1, 2, \ldots, 20\}$.

Parametar α zovemo *koeficijent očuvanosti*, a predstavlja relativnu frekvenciju dominantne aminokiseline po svakom stupcu u hipotetskom profilu motiva, uprosječenu po svim stupcima. Označimo s A_i prosječne distribucije aminokiselina s pretpostavkom očuvanosti i-te aminokiseline $\alpha \cdot 100\%$, dakle:

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

gdje su

$$p_j^i = \alpha \cdot \mathcal{K}_{\{i=j\}} + (1 - \alpha) \cdot r_j,\ j \in \{1, 2, \ldots, 20\}.$$

Odredimo sada očekivanu udaljenost dvaju nizova aminokiselina duljine n. Pretpostavimo dodatno da svaki niz ima točno $n - k < n$ fiksnih pozicija. Bez smanjenja općenitosti pretpostavljamo da se radi o zadnjih $n - k$ pozicija. Označimo navedene nizove s:

- $X = (X_1, X_2, \ldots, X_k, c_{k+1}, \ldots, c_n)$,
- $Y = (Y_1, Y_2, \ldots, Y_k, c_{k+1}, \ldots, c_n)$.

Izračunajmo sada njihovu očekivanu udaljenost

$$\mathbb{E}[d(X,Y)^2] = \mathbb{E}[\sum_{i=1}^{k}(X_i - Y_i)^2 + \sum_{i=k+1}^{n}(c_i - c_i)^2] = \mathbb{E}[\sum_{i=1}^{k}(X_i - Y_i)^2].$$

Kako nemamo dodatnih informacija o prvih k aminokiselina, pretpostavimo da se radi o nekim prosječnim aminokiselinama X_0 i Y_0, tada dobivamo:

$$\mathbb{E}[d(X,Y)^2] = \mathbb{E}[\sum_{i=1}^{k}(X_0 - Y_0)^2] = k \cdot \mathbb{E}[(X_0 - Y_0)^2].$$

Očekivani kvadrat udaljenosti dvije aminokiseline uzorkovane iz A_i je

$$\sum_{j,k=1}^{20}(a_j^i - a_k^i)^2 p_j^i p_k^i.$$

Koeficijent očuvanosti α u uzorcima varira od 0.88 do 0.91 pa se odlučujemo za $\alpha = 0.88$, uvrštavanjem dobijemo:

$$\mathbb{E}[(X_0 - Y_0)^2] = \sum_{i=1}^{20} r_i \sum_{j,k=1}^{20}(a_j^i - a_k^i)^2 p_j^i p_k^i = 4.56.$$

Odnosno

$$\mathbb{E}[d(X,Y)^2] = k \cdot 4.56.$$

Pa je

$$\mathbb{E}[d(X,Y)] = \sqrt{k} \cdot 2.14.$$

Sljedeći teorem nam je potreban za procjenu radijusa, njegov se dokaz može naći u [7].

Teorem 3.6.1. *Očekivana udaljenost dvije točke koje su uniformno distribuirane u kugli radijusa r u n-dimenzionalnom prostoru teži u* $r\sqrt{2}$ *kada* $n \to \infty$.

Koristeći procjenu udaljenosti dvije aminokiseline te pomoću teorema 3.6.1 dobijemo:

$$r_{old} = \frac{\mathbb{E}[d(X,Y)]}{\sqrt{2}} = \frac{\sqrt{k} \cdot 2.14}{\sqrt{2}}.$$

Još nam ostaje dobivenu procjenu radijusa prilagoditi standardiziranim podacima. Označimo s σ_{old} i σ_{new} standardne devijacije podataka prije i poslije standardizacije. Kako su radijus i standardna devijacija proporcionalne veličine, slijedi konačna procjena radijusa:

$$r_{new} = r_{old} \cdot \frac{\sigma_{new}}{\sigma_{old}} = \frac{\sqrt{k} \cdot 2.14}{\sqrt{2}} \cdot \frac{\sigma_{new}}{\sigma_{old}}.$$

Procjena središta

Sada konačno tražimo najgušću kuglu u $\mathbb{R}^{5n}$ s procijenjenim radijusom r_{new}, u smislu da od cjelokupne populacije motiva iz skupa P, kugla sadrži što više motiva iz skupa TP te što manje iz skupa FP.

Počinjemo biranjem jednog pozitivca, koji će biti središte u prvoj iteraciji. Biramo ga tako da dobivena kugla sadrži najveći broj drugih pozitivaca. Zatim, kako bi centrirali točke unutar kugle oko središta, iterativno postavljamo novo središte na težište točaka unutar trenutačne kugle. Postupak ponavljamo sve dok ne dobijemo istu točku kao težište.

Također, kako bi algoritam završio u konačnom broju koraka, postavljamo maksimalan broj iteracija na 10. Napomenimo da je u svakoj analizi metoda konvergirala u manje od 6 koraka.

3.7 Uspješnost metode

Kako bi testirali uspješnost metode, usporediti ćemo ju s referentnom metodom, čiji je rezultat jednak primjeni modela kugle u idealnom slučaju. Uvedimo prvo nekoliko omjera pomoću kojih provodimo usporedbu.

Mjere uspješnosti

Omjere koje ćemo definirati lakše je predočiti pomoću tablice 3.7.1.

	P	N
CP	TP	FN
CN	FP	TN

Tablica 3.7.1: Matrica uspješnosti

Opisom cilja metode, jasno je da njenu uspješnost možemo mjeriti omjerima veličina skupova TP, TN, FP ili FN i sume veličina u tom stupcu, to su skupovi CP, CN, P ili N.

Ukupno možemo dobiti osam omjera, koje možemo grupirati u četiri para. Upareni omjeri su međusobno komplementarni tj. u sumi daju jedan (npr. $\frac{\mathrm{TP}}{\mathrm{P}} + \frac{\mathrm{FP}}{\mathrm{P}} = 1$). Posebno su nam korisni omjeri osjetljivost (eng. *sensitivity* ili *true positive rate*) i preciznost (eng. *specificity* ili *positive predictive value*) koji su redom definirani s:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{CP}},$$

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{P}}.$$

Također, kao mjeru točnosti modela, definiramo i F_1-*score* (ili kraće, F_1), uzimanjem harmonijske sredine omjera TPR i PPV

$$F_1\text{-}score = 2 \cdot \frac{PPV \cdot TPR}{PPV + TPR}.$$

Referentna metoda

Dobivene omjere i F_1-*score* ćemo usporediti s referentnom metodom (eng. *benchmark*). Za razliku od već opisane metode, u prvoj iteraciji, središte postavljamo na težište svih TP motiva. Zatim, tražimo radijus tako da maksimiziramo F_1-*score* te, slično kao prije, iterativno postavljamo novo središte na težište kugle sve dok ne dobijemo istu točku ili dok ne napravimo deset iteracija.

Ovom metodom također testiramo valjanost pretpostavke da se motivi iz TP gušće okupljaju u kuglama. Ako pretpostavka nije točna, jasno je da će se to odraziti na uspješnost metode.

Poglavlje 4

Rezultati razvijene metode

Zbog visoke dimenzije, dobivene rezultate metode nije moguće vizualizirati, iz tog razloga koristimo alat *t-Distributed Stochastic Neighbor Embedding* [1] ili kraće t-SNE koji rješava upravo taj problem. Riječ je o statističkoj metodi koja reducira dimenziju podataka te pri tome čuva njihovu lokalnu strukturu. Korištenjem navedenog alata smanjujemo dimenziju podataka na 2, čime omogućujemo njihovu vizualnu interpretaciju.

4.1 Subjekti testiranja

Metodu testiramo preuzimanjem šest skupova proteina pripitomljenih i divljih biljaka. Među njima je divlja biljka talijin uročnjak, dobro anotirana biljka često korištena u modelima u botanici i genetici. Također, kao primjer potencijalno lošije anotirane biljke izabrali smo divlju bananu. Do sada navedene te soju izabrali smo zbog moguće usporedbe s rezultatima rada [5]. Ostale biljke uključuju rajčicu, rižu i suncokret, koji smo izabrali zbog njihove uporabe u članku [6]. Napomenimo da su riža, rajčica, soja i suncokret pitome biljke te su, kako imaju veliku ekonomsku važnost, potencijalno dobro anotirane. Kao pretraživač motiva koristimo iterativni pretraživač IGLOSS.

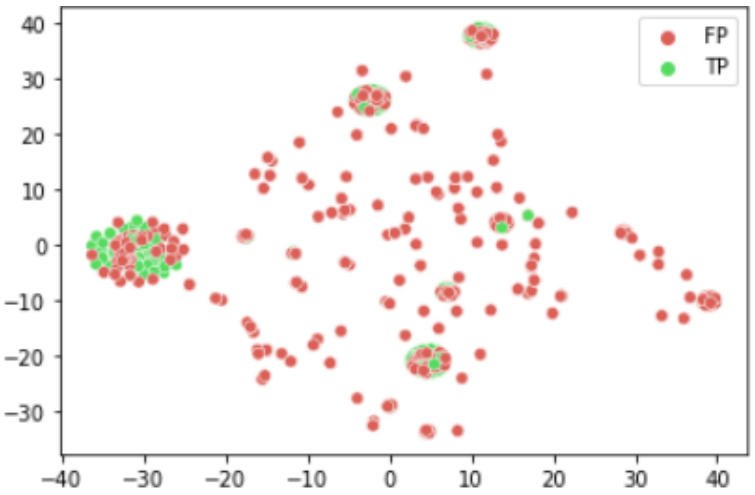
Cilj je pronalaženje motiva iz WRKY familije transkripcijskih faktora. Kao glavne građevne jedinice navedene familije, promatramo WRKYGQK motive te kao upit koristimo xWRKYGxK (gdje je x oznaka za bilo koju aminokiselinu). Navedeni upit izabran je po uzoru na rezultate članka [6]. IGLOSS zahtjeva unošenje skale, koju ćemo postaviti na 5, 6 i 7 kako bi mogli usporediti rezultate metode ovisno sličnosti o motiva i upita.

Kako je WRKY familija određena i cinkovim prstima, analiziramo i njihovu prisutnost u proteomu. To ćemo raditi na biljci talijin uročnjak zbog dobre kvalitete anotacije.

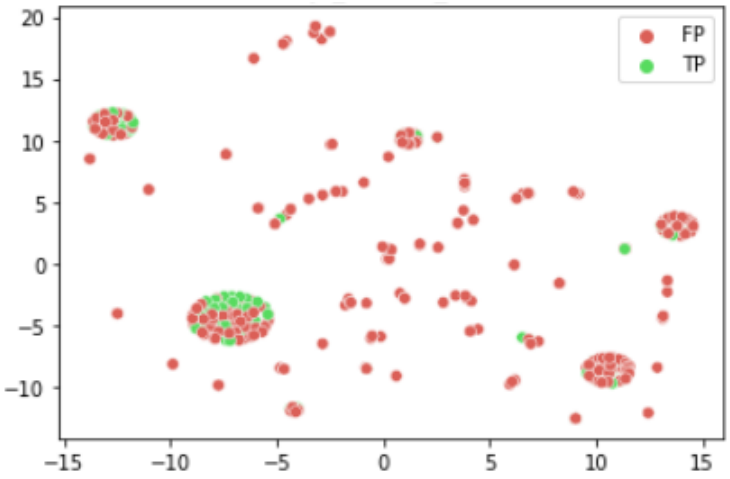
4.2 Rezultati analize WRKY motiva

Kao što smo najavili, rezultate prvo prikazujemo grafički, koristeći alat t-SNE. U natavku slijede rezultati primjene metode na WRKY motivima, za svih 6 navedenih biljaka. Pri tome prikazujemo samo one koji su dobiveni IGLOSS-ovom skalom 6.

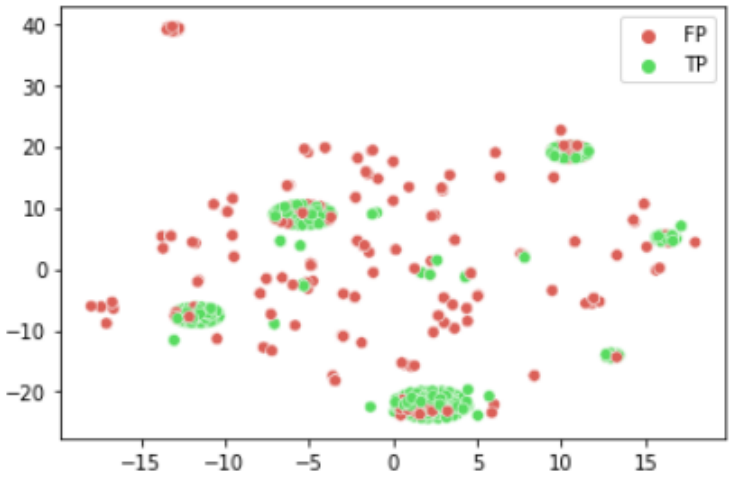
Ilustrirana je separacija pravih pozitivaca od lažnih. Zelenom bojom označena je pripadnost točaka skupu pravih pozitivaca, a crvenom njihova pripadnost skupu lažnih.



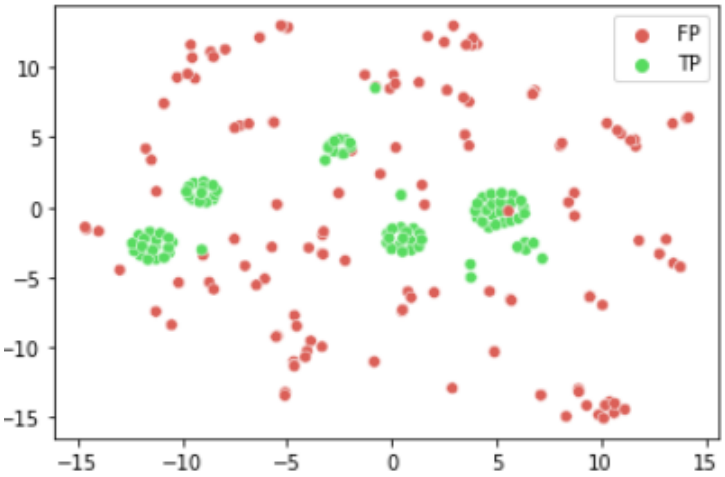
Slika 4.2.1: Talijin uročnjak, skala 6



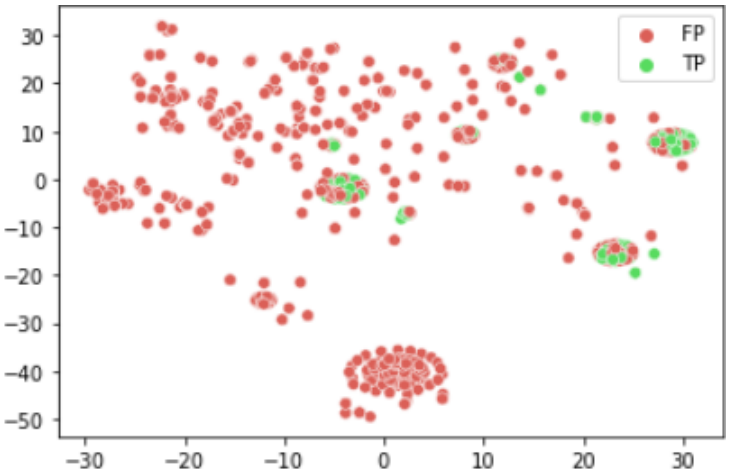
Slika 4.2.2: Divlja banana, skala 6



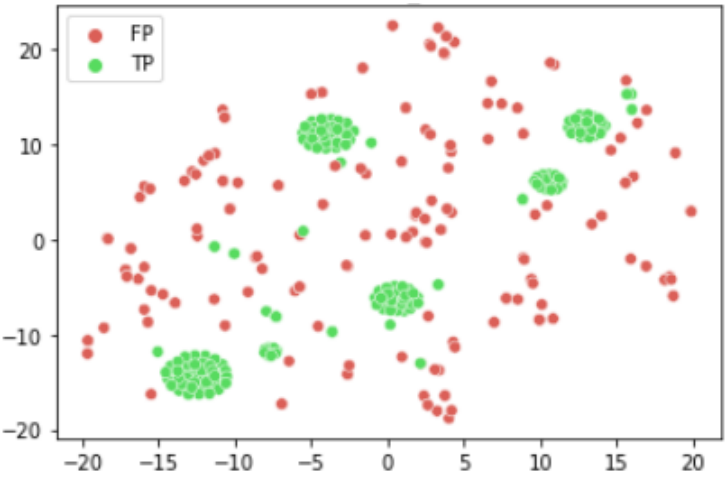
Slika 4.2.3: Soja, skala 6



Slika 4.2.4: Rajčica, skala 6



Slika 4.2.5: Riža, skala 6



Slika 4.2.6: Suncokret, skala 6

Nadalje, prikazujemo rezultate metode zajedno s rezultatima referentne metode, kako bi ih mogli usporediti. Dajemo omjere, kao što su TPR, PPV te F_1-*score*, zajedno s procijenjenim i optimalnim radijusom.

Budući da uspješnost metode ovisi o kvaliteti rezultata pretraživača, prikazujemo i prilagođene omjere kako bi testirali uspješnost isključivo naše metode. Naime, omjer TPR i F_1-*score* ovise o veličini skupa CP, uzimamo u obzir sve motive iz tog skupa, ali u zagradama navodimo i omjere za koje u obzir uzimamo samo one motive iz CP koje je pronašao i pretraživač.

Skala	Metoda	TP	P	PPV	TPR	F_1-*score*	Radijus
5	Referentna	213	417	0.51	0.93(1.00)	0.66(0.68)	3.19
	IGLOSS+Kugla	213	417	0.51	0.93(1.00)	0.66(0.68)	3.50
6	Referentna	213	417	0.51	0.93(1.00)	0.66(0.68)	3.23
	IGLOSS+Kugla	213	417	0.51	0.93(1.00)	0.66(0.68)	3.83
7	Referentna	207	404	0.51	0.90(0.97)	0.65(0.67)	3.76
	IGLOSS+Kugla	213	417	0.51	0.93(1.00)	0.66(0.68)	4.34

Tablica 4.2.1: Talijin uročnjak (lat. *Arabidopsis thaliana*)

Skala	Metoda	TP	P	PPV	TPR	F_1-*score*	Radijus
5	Referentna	79	232	0.34	0.86(0.86)	0.49(0.49)	2.94
	IGLOSS+Kugla	90	279	0.32	0.98(0.98)	0.49(0.49)	3.72
6	Referentna	91	283	0.32	0.99(0.99)	0.49(0.49)	5.3
	IGLOSS+Kugla	87	277	0.31	0.95(0.95)	0.47(0.47)	4.34
7	Referentna	86	266	0.32	0.93(0.93)	0.48(0.48)	4.28
	IGLOSS+Kugla	90	279	0.32	0.98(0.98)	0.49(0.49)	4.63

Tablica 4.2.2: Divlja banana (lat. *Musa acuminata*)

Skala	Metoda	TP	P	PPV	TPR	F_1-*score*	Radijus
5	Referentna	319	389	0.82	0.93(0.97)	0.87(0.89)	3.90
	IGLOSS+Kugla	319	389	0.82	0.93(0.97)	0.87(0.89)	3.63
6	Referentna	319	389	0.82	0.93(0.97)	0.88(0.89)	4.57
	IGLOSS+Kugla	315	385	0.82	0.92(0.95)	0.87(0.88)	4.10
7	Referentna	320	396	0.81	0.95(0.99)	0.87(0.89)	7.08
	IGLOSS+Kugla	313	383	0.82	0.93(0.97)	0.87(0.89)	4.46

Tablica 4.2.3: Soja (lat. *Glycine max*)

Skala	Metoda	TP	P	PPV	TPR	F_1-*score*	Radijus
5	Referentna	96	99	0.97	0.84(0.97)	0.90(0.97)	3.93
	IGLOSS+Kugla	95	97	0.98	0.83(0.96)	0.90(0.97)	3.49
6	Referentna	97	99	0.98	0.86(0.99)	0.92(0.98)	4.01
	IGLOSS+Kugla	95	97	0.98	0.84(0.97)	0.90(0.97)	3.71
7	Referentna	97	99	0.98	0.86(0.96)	0.92(0.98)	4.55
	IGLOSS+Kugla	95	97	0.98	0.84(0.97)	0.90(0.97)	4.03

Tablica 4.2.4: Rajčica (lat. *Solanum lycopersicum*)

Skala	Metoda	TP	P	PPV	TPR	F_1-*score*	Radijus
5	Referentna	146	340	0.43	0.92(0.97)	0.59(0.59)	3.78
	IGLOSS+Kugla	146	334	0.44	0.92(0.97)	0.59(0.60)	3.50
6	Referentna	146	336	0.43	0.93(0.98)	0.59(0.60)	4.30
	IGLOSS+Kugla	146	336	0.43	0.93(0.98)	0.59(0.60)	3.75
7	Referentna	141	326	0.43	0.89(0.93)	0.58(0.59)	3.96
	IGLOSS+Kugla	146	336	0.43	0.92(0.97)	0.59(0.60)	4.29

Tablica 4.2.5: Riža (lat. *Oryza sativa*)

Skala	Metoda	TP	P	PPV	TPR	F_1-*score*	Radijus
5	Referentna	171	175	0.98	0.91(0.98)	0.94(0.98)	4.14
	IGLOSS+Kugla	158	159	0.99	0.84(0.90)	0.91(0.95)	3.47
6	Referentna	172	173	0.99	0.92(0.99)	0.96(0.99)	4.45
	IGLOSS+Kugla	161	162	0.99	0.87(0.93)	0.93(0.96)	3.88
7	Referentna	167	167	1.00	0.90(0.96)	0.95(0.98)	4.80
	IGLOSS+Kugla	161	161	1.00	0.87(0.93)	0.93(0.96)	4.11

Tablica 4.2.6: Suncokret (lat. *Helianthus annuus*)

4.3 Diskusija o WRKY motivima

Prvo ćemo komentirati grafičke prikaze. Primjećujemo da se kod većine biljaka jasno razaznaje grupacija točaka u 5 ili 6 "kugli". Dakle, vidimo da se motivi nisu grupirali u jednu kuglu kao što smo pretpostavili. Navedeno opažanje zahtjeva prilagodbu modela, kako velik broj zaključaka dolazi od navedene pretpostavke.

Unatoč tome, iz slika primjećujemo da ipak možemo naslutiti sličnost motiva u pojedinim kuglama. Zaključujemo da bi grupe trebalo promatrati pojedinačno, što zahtjeva dodatnu analizu. Nadalje, navedene kugle su u slučaju soje (slika 4.2.3), rajčice (slika 4.2.4) i suncokreta (slika 4.2.6) pretežno zelene tj. većinom sadrže prave pozitivce.

Ovo bi mogla biti posljedica evolucije WRKY TF, tj. njihovog razvoja od zajedničkog pretka u nekoliko različitih grana. Takva se tvrdnja podudara s njihovom grupacijom u članku [6] u 5 grupa.

Napomenimo također, da postoje grupacije koje pretežno sadrže lažne pozitivce, što je možda posljedica njihove pripadnosti nekim drugim familijama.

Usporedimo PPV, TPR te F_1-*score* dobiven primjenom modela kugle sa njihovim početnim vrijednostima, dobivenim isključivo korištenjem pretraživača. U tu svrhu prikazujemo navedene omjere u skali 6. Također, u zagradama (u postotcima) prikazujemo koliko su omjeri porasli nakon primjene metode, odnosno pali u slučaju TPR-a.

Vrsta	PPV	TPR	F_1-*score*
Talijin uročnjak	0.23(122%)	0.93(−0%)	0.36(83%)
Divlja banana	0.21(48%)	1.00(−5%)	0.35(34%)
Soja	0.48(71%)	0.96(−4%)	0.64(36%)
Rajčica	0.38(158%)	0.87(−3%)	0.53(70%)
Riža	0.14(207%)	0.95(−2%)	0.25(136%)
Suncokret	0.44(125%)	0.94(−7%)	0.60(55%)

Tablica 4.3.1: Uspješnost pretraživača

Dakle, iz tablice 4.3.1 utvrđujemo da je PPV u prosjeku veći za 122%. Cilj je bio izbaciti što manje pravih pozitivaca što smo uspjeli, kako je TPR u prosjeku manji za 4%. Cjelokupna uspješnost modela je također veća, na što ukazuje F_1 koji je u prosjeku veći za 69%.

Napomenimo također, kako smo korištenjem raznih skala dobili konzistentne rezultate. Najveće razlike PPV i TPR omjera primjećujemo kod divlje banane. međutim ta razlika je otprilike 3%, dok je najveća razlika omjera F_1 4%.

Provjerimo nadalje, koliko se rezultati naše metode razlikuju od rezultata dobivenih u idealnom slučaju, tj. od referentne metode. Najveća razlika u F_1 je kod divlje banane (za skalu 6) te iznosi oko 4% te kod suncokreta (za skalu 6) oko 3%. Kod ostalih uzoraka je navedena razlika manja od 2%.

Prokomentirati ćemo još i radijuse te ih usporediti s idealnim slučajem. Za skale 5, 6 i 7, u prosijeku procijenjeni radijusi iznose redom 3.55, 3.94 i 4.31, a u idealnom slučaju 3.65, 4.31 i 4.74. Primijetimo prvo da povećavanje skale rezultira povećanju radijusa. Nadalje, idealna metoda daje nešto veće radijuse, iako kod talijinog uročnjaka (na skalama 5, 6 i 7), divlje banane (na skalama 5 i 7) te kod riže (na skali 7) opažamo suprotnu pojavu.

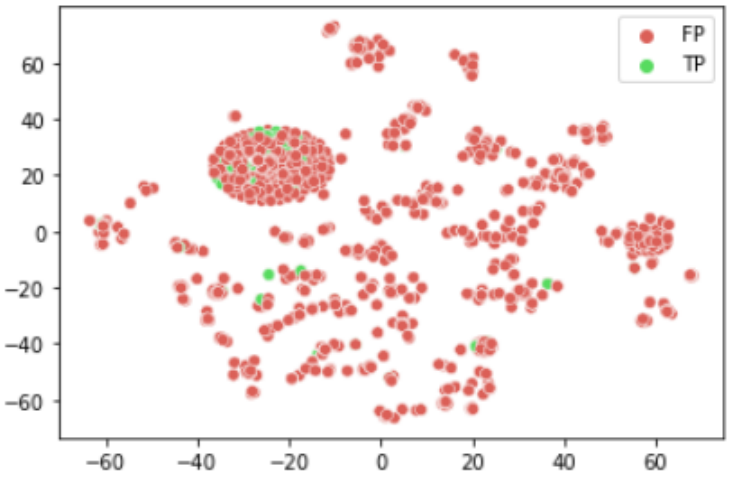
4.4 Rezultati analize cinkovih prstiju

Naziv cinkov prst (eng. *zinc finger*) dodijeljen je raznim strukturama, neke od njih su C2H2 tip, koji su i najbolje opisani, cinkov zglob (eng. *zinc knuckle*), cinkova vrpca (eng. *zinc ribbon*), C2HC i C2HC5. Njihova raznovrsnost otežava posao pronalzka svih anotacija i time njihove klasifikacije.

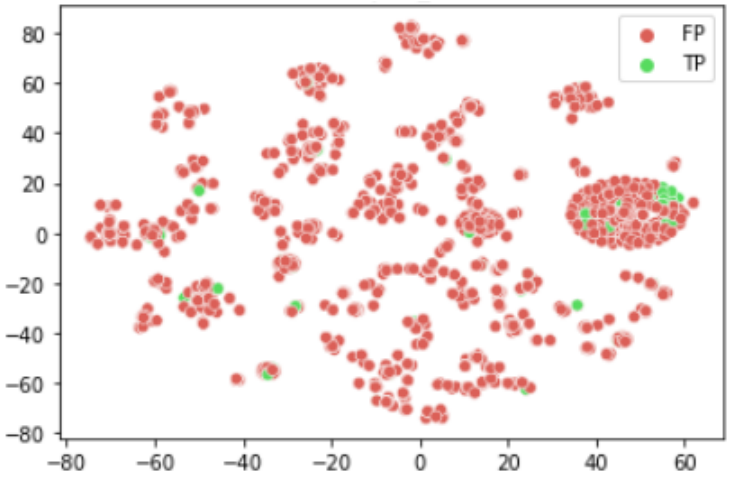
Napomenimo da smo u analizi cinkovih prsti u rezultatu pretraživača češće dobili više motiva u istom proteinu, što nije neobično jer znamo da isti protein može imati više cinkovih prsti.

Kao što smo napomenuli, testiramo metodu na biljci talijin uročnjak. Kao upite koristimo CxCxHxH sa skalama 6 i 7 te FCxCxHxHR i EICxCxHxHYE sa skalom 8. Prikažimo

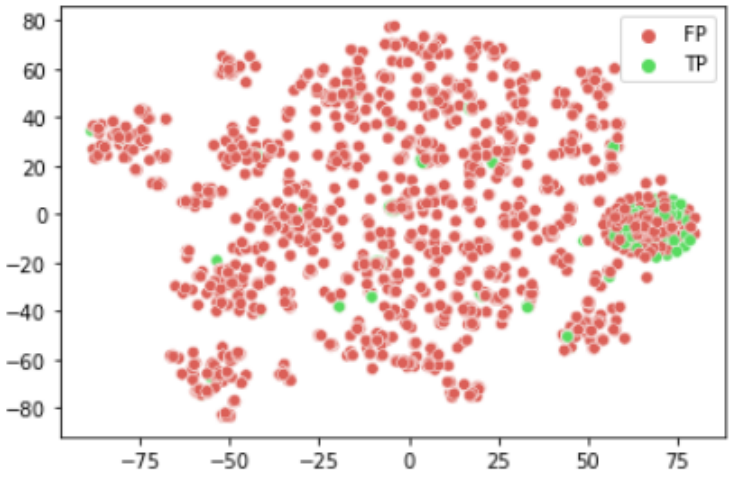
prvo dobivene rezultate grafički, prikazujemo samo za upite CxCxHxH i FCxCxHxHR. Kako se radi o drugoj familiji motiva, odvojeno procjenjujemo koeficijent očuvanosti. U prosijeku dobivamo $\alpha = 0.82$.



Slika 4.4.1: Upit CxCxHxH, skala 6



Slika 4.4.2: Upit CxCxHxH, skala 7



Slika 4.4.3: Upit FCxCxHxHR, skala 8

Analogno analizi WRKY motiva prikazujemo mjere uspješnosti metode, zajedno s referentnom metodom u sljedećim tablicama:

Skala	Metoda	TP	P	PPV	TPR	F_1-*score*	Radijus
6	Referentna	79	595	0.13	0.05(0.45)	0.07(0.21)	2.88
	IGLOSS+Kugla	128	1870	0.07	0.08(0.74)	0.08(0.13)	5.38
7	Referentna	64	470	0.14	0.04(0.31)	0.06(0.19)	1.06
	IGLOSS+Kugla	171	2316	0.07	0.11(0.74)	0.09(0.14)	5.43

Tablica 4.4.1: Upit CxCxHxH

Skala	Metoda	TP	P	PPV	TPR	F_1-*score*	Radijus
8	Referentna	62	319	0.19	0.04(0.40)	0.07(0.26)	3.03
	IGLOSS+Kugla	130	1451	0.09	0.09(0.83)	0.09(0.16)	5.60

Tablica 4.4.2: Upit FCxCxHxHR

Skala	Metoda	TP	P	PPV	TPR	F_1-*score*	Radijus
8	Referentna	55	199	0.28	0.04(0.27)	0.06(0.27)	3.11
	IGLOSS+Kugla	113	984	0.11	0.07(0.55)	0.09(0.19)	5.33

Tablica 4.4.3: Upit EICxCxHxHYE

4.5 Diskusija o cinkovim prstima

Na grafičkim prikazima ne vidi se jasna grupacija cinkovih prsti (motiva iz TP). Međutim, kod svakog upita primjećujemo jednu veću kuglu. Za razliku od kugli dobivenih u slučaju uzoraka WRKY motiva, one su sada pretežno crvene tj. sadrže više motiva iz FP. Navedeno opažanje ćemo također potkrijepiti analizom omjera uspješnosti, koji potvrđuju lošiju kvalitetu rezultata. U svakom slučaju, cinkovi prsti se ne grupiraju, što je moguća posljedica lošije anotacije ili činjenice da su teži za pronaći.

Kako nas slike vode do zaključka da ima više, razmaknutih grupa motiva iz TP, zaključujemo (slično kao kod WRKY motiva) da bi grupe trebalo promatrati pojedinačno.

Pogledajmo sada uspješnost modela. Prvo radimo usporedbu s rezultatima bez modela (samo pretraživač). F_1 je redom: 0.08, 0.08, 0.07, 0.07, uspoređivanjem sa rezultatima iz tablice zaključujemo da je naša metoda ista kod prvog te nešto bolja kod drugih upita (za 13% do 29%). PPV je redom 0.06, 0.06, 0.05, 0.05, vidimo da je malo bolji nakon korištenja metode kod prva 2 upita, za 17%, a puno bolji kod zadnja 2, za 80% i 220%. Pogledajmo još TPR, on je redom 0.11, 0.13, 0.1, 0.13, lošiji je za 27%, 15%, 10% i 56%. Zaključujemo da je F_1 malo bolji, a kod upita gdje je PPV puno bolji - TPR je puno lošiji.

Iz tablice također primjećujemo da referentna metoda "prioritizira" PPV - ima puno bolji PPV, ali je TPR dosta lošiji, također ima bolji F_1 u zagradi (uzimamo u obzir samo motive koje je izbacio pretraživač) nego izvan, što je očekivano jer je cilj algoritma da se taj F_1 maksimizira. Vidimo također da je procijenjeni radijus dosta veći nego optimalni, međutim to nije problem s obzirom na potencijalno lošu anotaciju.

4.6 Zaključak

Na kraju zaključujemo da opisana metoda može biti vrlo uspješna u klasifikaciji proteinskih motiva. Iz grafičkih prikaza ne možemo zaključiti da je ispunjena pretpostavka o grupaciji motiva u jednu kuglu. Međutim, u slučaju WRKY motiva rezultirala je znatnim povećanjem F_1-*score*-a, a da je pri tome održala TPR što većim. Unatoč tome, vidimo da je potreban oprez. Slabija uspješnost metode primijenjene na cinkove prste indicira na njenu osjetljivost na kvalitetu rezultata pretraživača i anotacije proteina.

Bibliografija

[1] *t-SNE*, (Studeni, 2023), `https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html`.

[2] *UniProt*, (Studeni, 2023), `https://www.uniprot.org/`.

[3] W. R. Atchley, J. Zhao, A. D. Fernandes i T. Drüke, *Solving the protein sequence metric problem*, Proceedings of the National Academy of Sciences of the United States of America (2005).

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

[5] Dolores Horvat, *Proteinski motivi i klasifikacija u proteinske familije*, Sveučilište u Zagrebu, Prirodoslovno-matematički fakultet (2023).

[6] K. T. Hsin, M. C Hsieh, Y. H Lee, K. C. Lin i Y. S. Cheng, *Insight into the Phylogeny and Binding Ability of WRKY Transcription Factors*, International Journal of Molecular Sciences (2022).

[7] M. G. Kendall i P. A. P. Moran, *Geometrical probability*, Hafner Publishing Company, London, 1963.

[8] B. Rabar, K. Nižetić, M. Zagorščak, K. Gruden i P.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.

[9] B. Rabar, M. Zagorščak, S. Ristov, M. Rosenzweig i P. Goldestein, *IGLOSS: iterative gapless local similarity search*, Bioinformatics (2019).

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

[11] I. Višek, *Clustering i klasifikacija proteinskih nizova*, Sveučilište u Zagrebu, Prirodoslovno-matematički fakultet (2022).

Sažetak

U ovom diplomskom radu proučavamo problem klasifikacije proteina u proteinske familije. Motivi su kraći nizovi aminokiselina sa zajedničkim ulogom u organizmu. Često su evolucijski bolje očuvani dijelovi proteina te su time korisni za njihovu klasifikaciju.

Testiramo i objašnjavamo metodu koja je razvijena za problem klasifikacije motiva. Metoda je implementirana kao nadogradnja standardnih tehnika pretraživanja motiva, gdje se, nakon ulaganja objekata u odgovarajući Euklidski prostor, koriste geometrijske metode optimizacije i strojnog učenja.

Uspješnost postupka testiramo na 6 biljnih proteoma na kojima promatramo familiju WRKY transkripcijskih faktora. Metoda rezultira relativno uspješnom klasifikacijom, iako pretpostavka o međusobnoj bliskosti nije u potpunosti ispunjena.

Summary

In this thesis, we study the classification problem for protein families. Motifs are shorter sequences of amino acids with a common role in organisms. They are often better preserved parts of proteins and are useful for their classification.

We test and explain a method that was developed for the motif classification problem. The method is implemented as an addition to standard motif searching techniques, where, after the transfer of objects to a corresponding Euclidean space, we use geometric optimization and machine learning methods.

We test the success of the procedure on 6 plant proteomes in which we analyze the family of WRKY transcription factors. The method results in a relatively successful classification, although the assumption of mutual closeness is not fully fulfilled.

Životopis

Rođen sam 26. kolovoza 1998. godine u Bjelovaru. Školovanje sam započeo u III. osnovnoj školi Bjelovar, nakon koje sam upisao Prirodoslovno-matematičku gimnaziju u Bjelovaru. 2017. godine upisujem preddiplomski studij matematike na Prirodoslovno-matematičkom fakultetu u Zagrebu, nakon čijeg završetka upisujem diplomski studij Matematička statistika na istom fakultetu.