Proteinski motivi i klasifikacija u proteinske familije

Horvat, Dolores

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Dolores Horvat

PROTEINSKI MOTIVI I KLASIFIKACIJA U PROTEINSKE FAMILIJE

Diplomski rad

Voditelj rada:
doc. dr. sc. Pavle Goldstein

Zagreb, veljača 2023.

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

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

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Zahvaljujem mentoru doc. dr. sc. Pavlu Goldsteinu na uloženom vremenu, pomoći i savjetima pri izradi ovog diplomskog rada.

Sadržaj

Uvod

Proteini su najraznovrsnije makromolekule u živim bićima. Skup svih proteina koje organizam proizvede tijekom života zovemo njegovim proteomom. Proteini imaju mnoštvo bitnih uloga u organizmu. Sudjeluju u izgradnji stanica, tkiva i organa (kolagen), imunološkoj obrani tijela (antitijela), ubrzavanju kemijskih reakcija (enzimi), prijenosu tvari (hemoglobin). Također, neki hormoni po kemijskom su sastavu proteini (inzulin).

Protein je jedinstveno zadan nizom aminokiselina od kojih je sastavljen. Ovisno o nizu, poprima prostornu stukturu te kao takav postaje funkcionalan. U posljednjih 20 godina, završetkom projekta određivanja ljudskog genoma (eng. *Human genome project*), dolazi do sve veće dostupnosti tehnologije i naglog pada cijena sekvenciranja. To je rezultiralo golemim brojem dobivenih proteinskih nizova (sekvenci) što je motiviralo razvoj novih računalnih metoda koje bi omogućile njihovu pohranu, obradu i analizu.

Karakteriziranje proteina na temelju pronađene sličnosti između njegovog niza i familije nizova drugih proteina uobičajen je pristup u bioinformatici. Umjesto cijelih nizova, često se u njima promatraju samo kratki očuvani podnizovi koje zovemo motivi, a upravo time ćemo se baviti i u ovom radu. Identifikacijom motiva, proteine ćemo klasificirati u promatranu proteinsku familiju. Za takve proteine smatra se da dijele evolucijsko podrijetlo što se odražava na njihovu sličnost u strukturi i funkciji.

Cilj ovog rada testirati je novorazvijenu metodu za identifikaciju motiva tražeći karakteristične motive VQ proteinske familije. Ideja metode filtrirati je odgovor neke druge metode koja pretražuje motive te joj tako ponuditi svoj pristup u njihovom prepoznavanju. Za razliku od tipičnih pretraživača motiva, koji traže odgovore "dovoljno" slične nekom fiksnom motivu, nova metoda traži odgovor u kojem su svi međusobno "dovoljno" slični. To čini prelaskom u euklidski prostor gdje umjesto sličnosti promatra udaljenosti između potencijalnih motiva.

Ovaj rad sastoji se od četiri poglavlja. U prvom poglavlju navedeni su osnovni matematički pojmovi i rezultati na kojima se metoda temelji. U drugom poglavlju objašnjena je biološka pozadina klasifikacije proteina te je rečeno na kojem pristupu rade tipični pretraživači motiva i s kojim izvorima greške se susreću. Nadalje, metoda koju ćemo testirati detaljno je opisana u trećem poglavlju. Konačno, u četvrtom poglavlju iznosimo rezultate klasifikacije.

Poglavlje 1

Matematička podloga

Pojmovi iz ovog poglavlja preuzeti su iz izvora [3], [9] i [12].

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

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

Napomena 1.1.4. *Skup $\mathbb{R}^n$ s uobičajenim operacijama zbrajanja i množenja je vektorski prostor nad poljem $\mathbb{R}$. Kažemo još 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) $\langle x, x \rangle = 0 \Leftrightarrow x = 0;$

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 1.1.7. *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 zove se* ***unitaran prostor****.*

Definicija 1.1.9. *Neka je V unitaran prostor.* ***Norma*** *na V je funkcija $\|\cdot\| : V \to \mathbb{R}$ definirana s*

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

Propozicija 1.1.10. *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) $\|x + y\| \leq \|x\| + \|y\|, \ \forall x, y \in V.$

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*

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

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

Definicija 1.1.13. *Neka je V normirani prostor.* ***Metrika*** *ili* ***udaljenost*** *na skupu 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) $d(x, y) = 0 \Leftrightarrow x = y, \ \forall x, y \in V;$

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

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

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 $a \in X$ 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 ($\mathcal{F} \subset \mathcal{P}(\Omega)$) 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 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. 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 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 1.2.14. ***Matematičko očekivanje*** *od X ili kraće, **očekivanje** od X označavamo sa $\mathbb{E}[X]$ i definira se sa:*

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

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 $X = \lim_{n\to\infty} X_n$. Niz $(\mathbb{E}[X_n])_{n\in\mathbb{N}}$ je rastući niz u $\mathbb{R}_+$, dakle postoji $\lim_{n\to\infty} \mathbb{E}[X_n]$ koji može biti jednak i $+\infty$.

Definicija 1.2.15. ***Matematičko očekivanje*** *od X ili, kraće, **očekivanje** od X definira se sa*

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

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

Definicija 1.2.16. *Kažemo da* ***matematičko očekivanje*** *od X, ili kraće,* ***očekivanje*** *od X* ***postoji*** *(ili da je definirano) ako je barem jedna od veličina $\mathbb{E}[X^+], \mathbb{E}[X^-]$ konačna, tj. vrijedi $min\{\mathbb{E}[X^+], \mathbb{E}[X^-]\} < +\infty$. Tada po definiciji stavljamo*

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

Teorem 1.2.17. *(osnovna svojstva matematičkog očekivanja)*

1) Ako $\mathbb{E}[X]$ postoji i $c \in \mathbb{R}$, tada $\mathbb{E}[cX]$ postoji 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

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

ili

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

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 1.2.18. *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 sa $Var(X)$ 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*** *od X 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 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 1.2.21. *Ako je jasno o kojoj se slučajnoj varijabli 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**.

Klasifikacija slučajnih varijabli

Definicija 1.2.23. *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 $p_n = \mathbb{P}\{X = x_n\}$, što zapisujemo u obliku tablice

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

Tablicu (1.2) zovemo **distribucija** ili **zakon razdiobe 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 1.2.24. *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 1.2.25. *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*

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

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 ponekad je označavamo sa 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. *X je* ***jedinična normalna distribucija*** *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

Problem klasifikacije proteina

2.1 Struktura, funkcija i evolucija proteina

Proteini su građeni od jednog ili više lanaca aminokiselina. Postoji 20 tzv. standardnih aminokiselina koje nalazimo u proteinima živih bića. Ovisno o redoslijedu aminokiselina, pripadni lanci savijaju se u različite trodimenzionalne strukture. Kao takvi, proteini mogu djelovati s drugim molekulama te izvršavati svoju funkciju. Struktura proteina ima važnu ulogu u njegovoj funkciji — ako protein izgubi svoj prirodni oblik, može postati nefunkcionalan.

Oznaka	Aminokiselina	Oznaka	Aminokiselina
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

Niz aminokiselina u proteinu određen je genima. Iz tog razloga, izmjene gena uslijed mutacija dovode i do promjene aminokiselina u proteinu. Dakle, zaključke o evolucijskim odnosima proteina možemo dobiti usporedbom pripadnih nizova aminokiselina — što su

nizovi sličniji, mogli bi biti srodniji. U puno slučajeva, promjena svega nekoliko aminokiselina perturbira, ali ne uništi strukturu proteina čime i njegova funkcija ostaje očuvana.

2.2 Proteinske familije i motivi

Za proteine, koji imaju zajedničko evolucijsko podrijetlo, reći ćemo da se nalaze u istoj proteinskoj familiji. Proteinska klasifikacija ima važnu ulogu u predviđanju strukture i funkcije novootkrivenih proteina. Naime, proteinska familija često sadrži već dobro okarakterizirane proteine. Pronalaskom novog proteina, njegova svojstva moguće je predložiti na temelju familije kojoj je predviđeno da on pripada.

Kao što se već može naslutiti, klasifikacija proteina temeljit će se na promatranju sličnosti između pripadnih nizova aminokiselina. Nakon što je pronađen novi niz, uobičajen je pristup promatrati njegovu sličnost sa svim okarakteriziranim nizovima u bazi, i to po cijeloj duljini. Često značajne sličnosti nema, međutim moguće je da se novi niz može razumno dobro karakterizirati s očuvanim podnizom — motivom. Motiv je kratak niz, obično duljine od 10 do 20, aminokiselina. Često se može povezati s prepoznatljivim dijelom u strukturi koji izvršava određenu funkciju. U evoluciji, motiv je ostao očuvaniji od ostalih dijelova te je imao običaj ponašati se kao zasebna jedinica u proteinu. Sve to čini identifikaciju motiva u proteinima važnim aspektom u klasifikaciji proteina i upravo time ćemo se baviti u ostatku ovog rada.

2.3 Pretraživanje motiva

Metode pretraživanja motiva (eng. *motif scanning methods*) su često korištene metode za analizu nizova u bioinformatici. Za ulaz primaju motiv — to je upit — te pronalaze njemu "dovoljno" slične podnizove u danom skupu nizova — to su motivi koje su pretraživači identificirali. Tipično, metode se temelje na algoritmu lokalnog poravnanja i funkciji sličnosti. Odgovor se generira u dva koraka. Najprije se svi rezultati lokalnog poravnanja rangiraju po sličnosti s upitom, a zatim se odabiru samo oni čija je sličnost iznad određenog praga (eng. *treshold*). Oba koraka mogu biti izvor greške. Tako će npr. netočno rangiranje s preniskim pragom rezultirati velikim odgovorom koji će sadržavati puno biološki pogrešnih motiva dok će previsok prag propustiti neke biološki prave motive. Različiti pretraživači se na različite načine nose s tim problemom.

U radu [4] je razvijena metoda s alternativnim pristupom problemu. Metodu ćemo detaljno opisati u idućem poglavlju.

Poglavlje 3

Metoda klasifikacije proteina

3.1 Ideja metode

Razvijena metoda nije osmišljena da djeluje samostalno kao tipičan pretraživač motiva. Za ulaz ne prima upit niti cijele proteinske nizove, već skup potencijalnih motiva među kojima odabire one za koje smatra da zaista jesu motivi. Dakle, na metodu možemo gledati kao na filter koji djeluje na odgovor nekog pretraživača, pružajući pretraživaču svoj pristup u odabiru odgovora.

Iz tog razloga, u terminima klasifikacije, recimo da je dobiven odgovor nekog pretraživača (pozitivci) takav da sadrži motive skoro svih proteina u uzorku koji biološki pripadaju promatranoj familiji (pravi pozitivci) te velik broj motiva pronađenih u proteinima koji navedenoj familiji ne pripadaju (lažni pozitivci). Primijetimo da je uzimanjem dovoljno niskog praga opisani slučaj upravo očekivani odgovor pretraživača. Cilj metode pronaći je optimalan podskup pravih pozitivaca u dobivenom skupu pozitivaca. Drugim riječima, želi se odbaciti što više lažnih pozitivaca, a pritom zadržati prave pozitivce.

Za razliku od pretraživača, koji promatraju sličnost s upitom, razvijena metoda se temelji na promatranju međusobne sličnosti. To će se realizirati na nešto drugačiji način što će omogućiti korištenje novih matematičkih alata i pokazati se iznimno brzim u izvršavanju. Uz numeričku reprezentaciju aminokiselina, prelazi se u euklidski prostor gdje se umjesto sličnosti promatra međusobna udaljenost. Pretpostavka je da su pravi pozitivci gušće raspoređeni od lažnih pozitivaca, te dodatno, da su smješteni u nekoj kugli u prostoru. Metoda pokušava pronaći radijus i središte optimalne kugle. Uočimo da smo pronalaskom kugle dobili točke takve da se svake dvije nalaze "dovoljno" blizu.

U nastavku dajemo detaljan opis razvijene metode. Metoda je u cijelosti implementirana u programskom jeziku *Python*.

3.2 Priprema podataka

Analizu započinjemo traženjem pozitivaca pomoću nekog pretraživača motiva. Neka je upit duljine n. Pretpostavljamo da motivi nemaju praznina (eng. *gaps*) pa su dobiveni nizovi aminokiselina od kojih je svaki duljine n.

Ako pretraživač pronađe više podnizova u istom proteinu, promatramo samo podniz koji je najsličniji upitu. Kao mjeru sličnosti koristimo *e-value* — što je ta vrijednost manja, upit i niz su sličniji. Sada je protein–motiv korespondencija 1–1 čime su riješene potencijalne nedoumice oko pitanja klasifikacije.

Prelazak u euklidski prostor

Nizovi aminokiselina sastavljeni su od slova. Nedostatak prirodne metrike za usporedbu tako opisanih podataka stvara prepreku u provođenju statističke analize. Naime, "udaljenost" slova aminokiselina u alfabetu se ne odražava na njihovu sličnost u fizikalno-kemijskim svojstvima. Rješenje tog problema ponuđeno je u članku [13]. Definirano je preslikavanje sa skupa aminokiselina u 5-dimenzionalni euklidski prostor $\mathbb{R}^5$ s pravilom pridruživanja danim u tablici 3.1. Ovako opisane aminokiseline pogodne su za korištenje u različitim analizama koje su usmjerene prema razumijevanju evolucije te strukturnih i funkcionalnih svojstava proteina.

Koristeći ovakav pristup, naši nizovi aminokiselina duljine n odgovaraju $5n$-dimenzionalnim vektorima.

Standardizacija podataka

Sada, kada se nalazimo u $\mathbb{R}^{5n}$, promatramo udaljenost među podacima. Kako bi traženje kugle koja bi obuhvaćala dio podataka bila razumna strategija, podatke treba najprije standardizirati. Naime, ako je varijanca podataka po jednoj koordinati znatno veća od one po ostalim koordinatama, euklidska udaljenost bila bi dominirana tom koordinatom. Time bi se izgubila forma kugle u kojoj bi sve koordinate trebale imati jednak utjecaj. Dodatno, standardizaciju ćemo prilagoditi kako bismo izbjegli dijeljenje jako malim brojem čime ćemo smanjiti mogućnost pojave *outliera*. Dakle, neka su $\bar{x}$ i s aritmetička sredina i standardna devijacija podataka, redom. Svaki podatak x_i transformiramo na sljedeći način

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

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 3.1: Opis aminokiselina numeričkim vektorima

3.3 Grafički prikaz podataka

Kako bismo dobili uvid u strukturu podataka koristimo alat za vizualizaciju visoko-dimenzionalnih podataka *t-Distributed Stochastic Neighbor Embedding* ili skraćeno t-SNE. Riječ je o statističkoj metodi koja reducira dimenziju podataka čuvajući lokalnu strukturu — točke koje su susjedne u prostoru reducirane dimenzije interpretiramo kao susjedne i u visoko-dimenzionalnom prostoru. Tako prelaskom u dvije dimenzije ilustriramo mogućnost separacije pravih pozitivaca od lažnih. Više o metodi i njenoj primjeni u *Python*-u može se naći u [10].

3.4 Mjere uspješnosti

Kako bismo mjerili uspješnost metode, potrebno je definirati mjere koje ćemo koristiti. Prije toga uvedimo oznake. Proteine u uzorku za koje je anotirano da pripadaju proma-

tranoj familiji označavamo sa CP (eng. *condition positive*), dok one ostale označavamo sa CN (eng. *condition negative*). S druge strane, proteine koje je metoda vratila kao odgovor označavamo sa P (eng. *positive*), a preostale sa N (eng. *negative*). Sada, svakom podatku, ovisno o njegovom stvarnom i predviđenom stanju, pridružujemo jedan od četiri moguća ishoda: TP (eng. *true positive*), FN (eng. *false negative*), FP (eng. *false positive*) ili TN (eng. *true negative*). Odnose navedenih grupa ilustriramo matricom uspješnosti (eng. *confusion matrix*) u koju se unose njihove veličine te se na taj način dobiva cjelokupan uvid u rezultat klasifikacije.

		Predviđeno stanje	
		P	**N**
Stvarno stanje	**CP**	**TP**	**FN**
	CN	**FP**	**TN**

Slika 3.1: Matrica uspješnosti

Postoji osam omjera koje možemo izračunati iz tablice 3.1. Oni dolaze u četiri komplementarna para tj. suma svakog od ta četiri para jednaka je 1. Omjere dobivamo dijeljenjem veličine svake od četiri grupa TP, FN, FP, TN s njihovom sumom u tom retku ili stupcu — što su veličine preostale četiri grupe CP, CN, P, N. One koje koristimo kao mjere uspješnosti naše metode su osjetljivost (eng. *true positive rate*)

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{CP}} \tag{3.1}$$

i preciznost (eng. *positive predictive value*)

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{P}}. \tag{3.2}$$

Napomenimo da korištenje omjera (eng. *false positive rate*)

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{CN}}$$

kao mjere uspješnosti u kontekstu našeg problema ne bi imalo smisla. Naime, broj CN-ova je nekoliko redova veličine veći od broja CP-ova pa bi za svaki razuman ishod metode FPR bio blizu nule. Konačno, uzimanjem harmonijske sredine od (3.1) i (3.2) dobivamo

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

3.5 *Benchmark*

Model kugle polazi od pretpostavke da su pravi pozitivci smješteni u nekoj kugli u prostoru. Ako pretpostavka nije posve ispunjena, jasno je da će se to odraziti na uspješnost metode. U tu svrhu napravljena je referentna metoda (eng. *benchmark*) čija je uspješnost jednaka uspješnosti metode kugle u idealnom slučaju. Referentna metoda najprije računa težište pravih pozitivaca te ga postavlja za središte kugle. Zatim za dano središte pronalazi optimalan radijus tj. radijus za koji se postiže maksimalan F_1-*score*.

3.6 Model kugle

U ovom odjeljku konačno ćemo dati procjenu kugle kojom ćemo pokušati odvojiti prave pozitivce od lažnih. Napomenimo da ćemo koristiti nešto drugačiji pristup pri procjeni radijusa, nego što je to napravljeno u radu [4] — pretpostavit ćemo da su neke pozicije u motivu fiksne. Kažimo i nešto o oznakama koje ćemo koristiti. U nastavku računamo s numeričkim vektorima duljine $5n$, a ne s nizovima aminokiselina duljine n. Međutim, radi intuitivnijeg shvaćanja procjene radijusa ponašat ćemo se kao da radimo s nizovima, a ne njihovim numeričkim reprezentacijama. Tako ćemo koristiti oznake koje nisu matematički korektne, ali su dovoljno sugestivne čime je opravdano njihovo korištenje.

Procjena radijusa

Procjena radijusa ovisit će o parametru α kojeg zadajemo *a priori*. Najprije ćemo izračunati očekivanu udaljenost dva niza aminokiselina koji su uzorkovani iz α-konveksne kombinacije distribucija, uz pretpostavku nekih fiksnih pozicija te prosječne distribucije aminokiselina duž ostalih pozicija. Koristeći taj rezultat, procjenu radijusa ćemo dobiti pozivanjem na teorem iz područja vjerojatnosne geometrije. Dodatno, procjenu će biti potrebno prilagoditi standardiziranim podacima.

Definirajmo najprije potrebne distribucije. Neka je R prosječna distribucija aminokiselina dana sa

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

Pripadne vjerojatnosti označimo sa r_i, $i \in \{1, 2, \ldots, 20\}$. Parametar $\alpha \in (0, 1)$ zovemo koeficijentom očuvanosti, a predstavlja relativnu frekvenciju dominantne aminokiseline po svakom stupcu u hipotetskom profilu motiva, uprosječenu po svim stupcima. Neka su A_i prosječne distribucije aminokiselina s pretpostavkom očuvanosti i-te aminokiseline u postotku $\alpha \cdot 100$ dane sa

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

gdje su

$$p^i_j = \alpha \cdot \mathbb{1}_{\{i=j\}} + (1-\alpha) \cdot r_j, \;\; j \in \{1, 2, \ldots, 20\}.$$

Promatramo dva niza aminokiselina duljine n. Pretpostavimo da je $n - k < n$ pozicija u nizu fiksno. Bez smanjenja općenitosti možemo pretpostaviti da se radi o zadnjih $n - k$ pozicija. Aminokiseline na preostalih k pozicija uzorkovane su iz nekih distribucija. Tako neka su $X = (X_1, \ldots, X_k, c_{k+1}, \ldots, c_n)$ i $Y = (Y_1, \ldots, Y_k, c_{k+1}, \ldots, c_n)$ dva promatrana niza. Želimo izračunati njihovu očekivanu udaljenost. Koristeći definiciju euklidske udaljenosti i linearnost očekivanja dobivamo

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

S obzirom da nemamo nikakvih saznanja o aminokiselinama na prvih k pozicija, pretpostavimo da se radi o nekim "prosječnim" aminokiselinama X_0 i Y_0 čime dobivamo

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

Očekivani kvadrat udaljenosti dvije aminokiseline uzorkovane iz A_i je

$$\sum_{j,k=1}^{20}(a^i_j - a^i_k)^2 p^i_j p^i_k.$$

Uprosječivanjem po distribuciji R i uvrštavanjem $\alpha = 0.53$ (u 12 proteoma koje smo koristili α je varirao od 43% do 57%) dobivamo

$$\mathbb{E}\Big[(X_0 - Y_0)^2\Big] = \sum_{i=1}^{20} r_i \sum_{j,k=1}^{20}(a^i_j - a^i_k)^2 p^i_j p^i_k = 14.54.$$

Slijedi

$$\mathbb{E}[d^2(X,Y)] = k \cdot 14.54.$$

Konačno dobivamo

$$\mathbb{E}[d(X,Y)] = \sqrt{k} \cdot 3.81.$$

Sada kada smo odredili očekivanu udaljenost, navodimo teorem koji ćemo iskoristiti za dobivanje procjene radijusa. Dokaz teorema se može naći u [6].

Teorem 3.6.1. *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.*

Pozivanjem na teorem 3.6.1 dobivamo procjenu radijusa

$$r_{old} = \frac{\sqrt{k} \cdot 3.81}{\sqrt{2}}.$$

Dobivenu procjenu preostaje prilagoditi standardiziranim podacima. Neka su std_{old} i std_{new} standardne devijacije podataka prije i poslije standardizacije, redom. Radijus i standardna devijacija su proporcionalne veličine pa slijedi konačna procjena radijusa

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

Uvrštavanjem r_{old} dobivamo

$$r_{new} = \frac{\sqrt{k} \cdot 3.81}{\sqrt{2}} \cdot \frac{std_{new}}{std_{old}}.$$

Procjena središta

Nakon što smo procijenili radijus kugle, preostaje procijeniti njeno središte. U nastavku dajemo iterativni algoritam pomoću kojega to činimo.

Prije toga potrebno je pronaći najgušću kuglu čije središte postavljamo kao inicijalnu iteraciju. Za fiksnu točku iz skupa podataka promatramo kuglu sa središtem u toj točki i procijenjenim radijusom. Prolaskom po svim točkama, za najgušću kuglu odabiremo onu unutar koje se nalazi najveći broj točaka.

Zatim, kako bi točke unutar kugle bile centrirane oko središta, algoritam u svakoj iteraciji za središte postavlja težište točaka unutar kugle. Zaustavlja se kada vrati isto težište. Postoji mogućnost da metoda ne konvergira pa se dodatno algoritmu kao parametar zadaje maksimalan broj iteracija. Time osiguravamo izvođenje u konačnom broju koraka.

U svim analizama koje smo proveli, primijetili smo da je metoda konvergirala u najviše sedam iteracija, a najčešće unutar tri.

Poglavlje 4

Rezultati klasifikacije proteina

4.1 Testovi

U radu [4] metoda je testirana na četiri biljna proteoma u kojima se tražilo članove familije GDSL lipaza. Metoda se pokazala vrlo uspješnom čime je ujedno i potvrđena pretpostavka o grupiranju pravih pozitivaca. U ovom radu metodu ćemo testirati traženjem proteina koji pripadaju VQ proteinskoj familiji.

VQ (valin-glutamin) familiju čine proteini koji se mogu opisati prisustvom motiva Fxx-hVQxhTG (h označava hidrofobnu aminokiselinu A, C, F, G, I, L, M, P, V ili W, a x označava bilo koju aminokiselinu). Po imenu familije zovemo ih VQ proteinima. VQ proteini specifični su za biljke, međusobno djeluju s WRKY transkripcijskim faktorima, a imaju važnu ulogu u rastu i razvoju biljaka te njihovu odgovoru na različite vrste abiotskog i biotskog stresa. Do danas su identificirani i okarakterizirani u brojnim biljnim vrstama, međutim, znanje o njima je i dalje vrlo ograničeno.

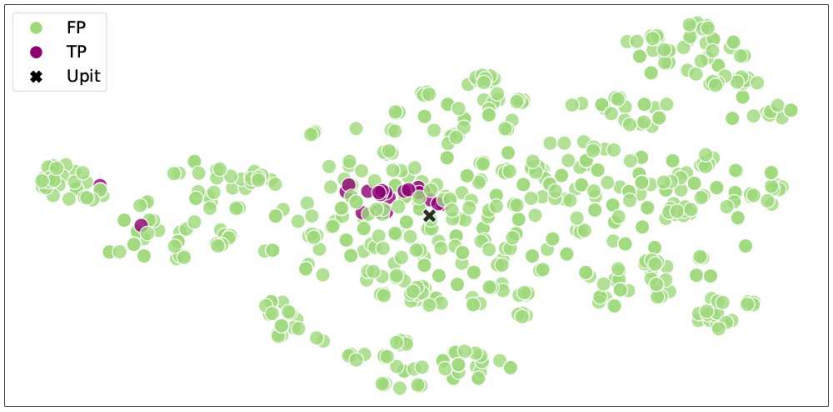
Iz baze UniProt [11] preuzeli smo 11 biljnih proteoma. Također, preuzeli smo i jednu stariju verziju proteoma ITAG2.3 [5] kako bismo vidjeli postoji li značajna razlika u anotaciji s nedavno preuzetim proteomom iste vrste s UniProt-a. Dakle, riječ je o 11 biljnih vrsta. Među njima, nalazi se talijin uročnjak što je popularna modelna biljka u botanici i genetici s vrlo dobro anotiranim proteomom. Odabrali smo i divlju bananu koja nije ekonomski važna biljka pa je potencijalno lošije anotirana. S druge strane, sve ostale odabrane vrste imaju veliku ekonomsku važnost. Vrsta za koju su odabrana dva proteoma je rajčica.

Kako bismo mogli primijeniti našu metodu, potrebni su nam kandidati za motive. U tu svrhu odlučili smo koristiti iterativni pretraživač IGLOSS [2]. IGLOSS za ulaz prima tzv. skalu koja mjeri sličnost između upita i odgovora. Što je skala veća, veća je i sličnost. To nam omogućava jednostavnu kontrolu nad dobivenim odgovorom pa smo za svaki proteom odabrali po tri različite skale tako da dobiveni PPV bude u prosjeku 4%, 7% i 14% (izuzev proteoma ITAG2.3 rajčice gdje PPV iznosi oko 1% i 3%). IGLOSS omogućava unos bilo

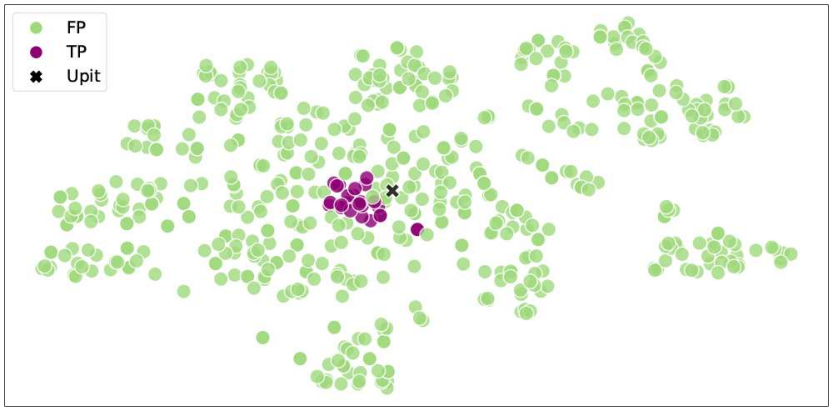
koje aminokiseline uz oznaku x pa smo za upit mogli koristiti FxxxVQxxTG.

4.2 Rezultati

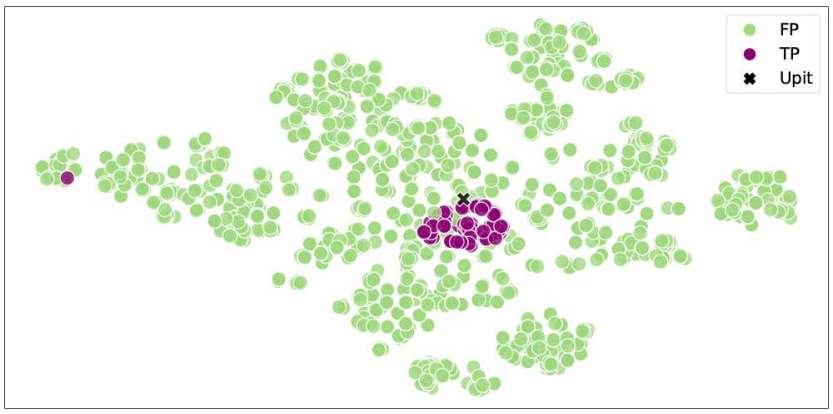
Kao što smo i najavili, svaki uzorak prikazali smo grafički koristeći t-SNE. U nastavku za svaki proteom dajemo prikaze samo po najmanjim skalama.



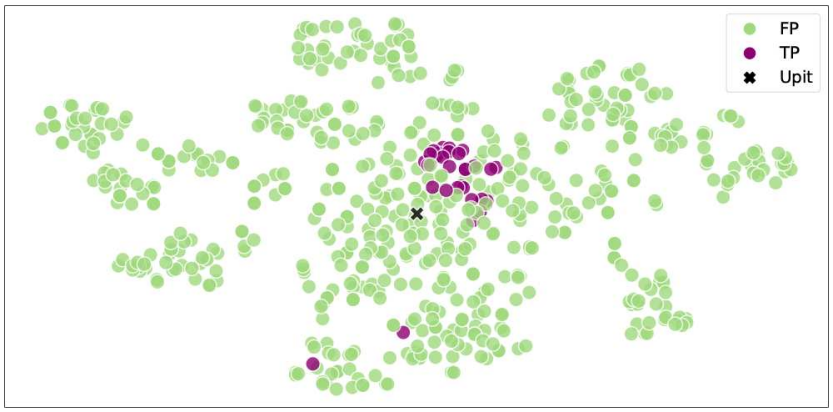
Slika 4.1: Talijin uročnjak, skala 4.0



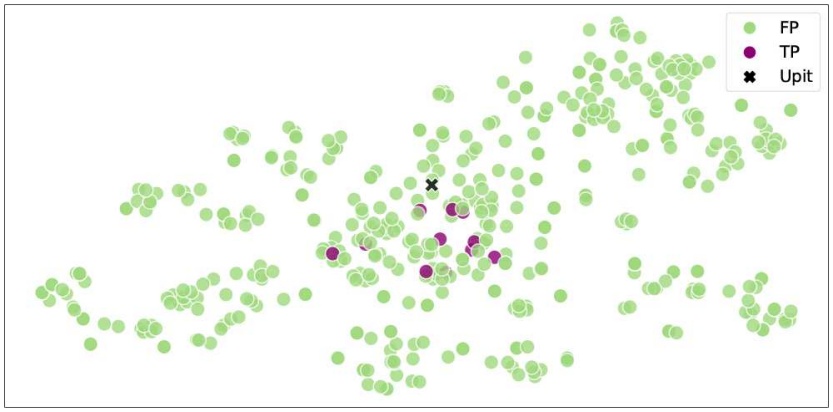
Slika 4.2: Čaj, skala 4.0



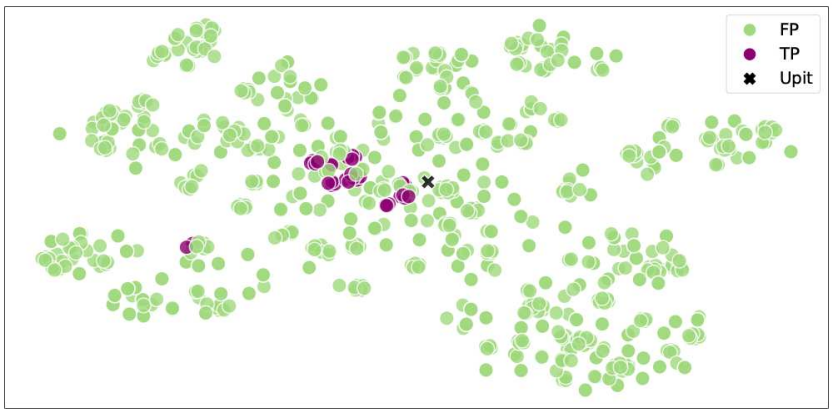
Slika 4.3: Soja, skala 4.0



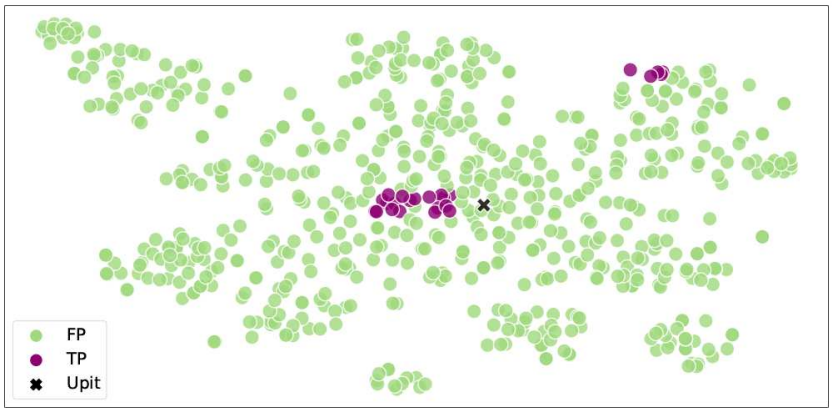
Slika 4.4: Obični ječam, skala 4.0



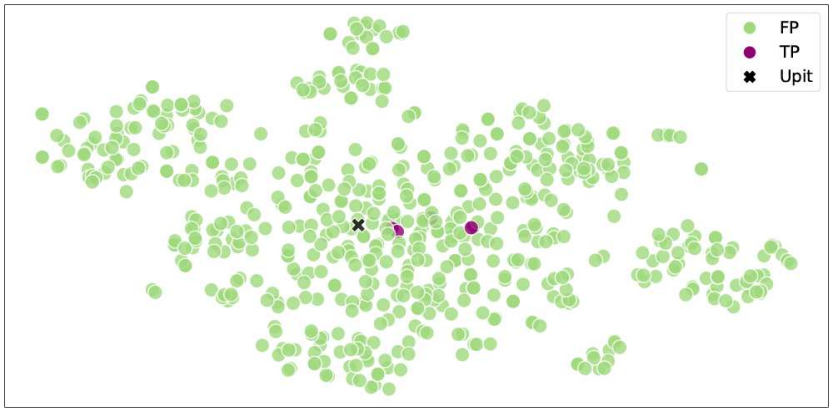
Slika 4.5: Divlja banana, skala 5.0



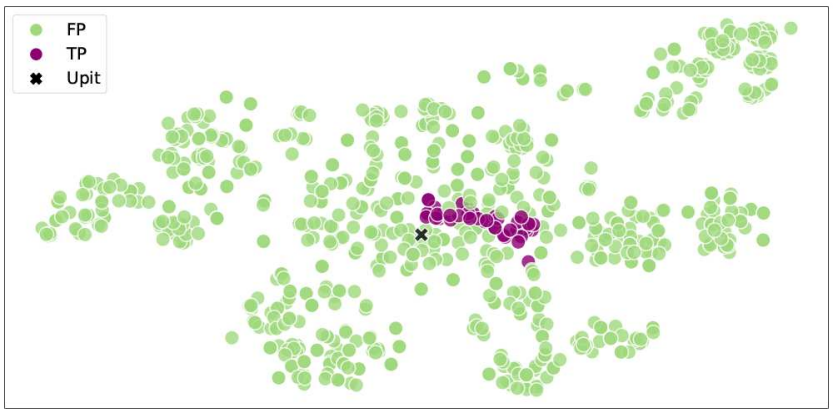
Slika 4.6: Pravi duhan, skala 5.0



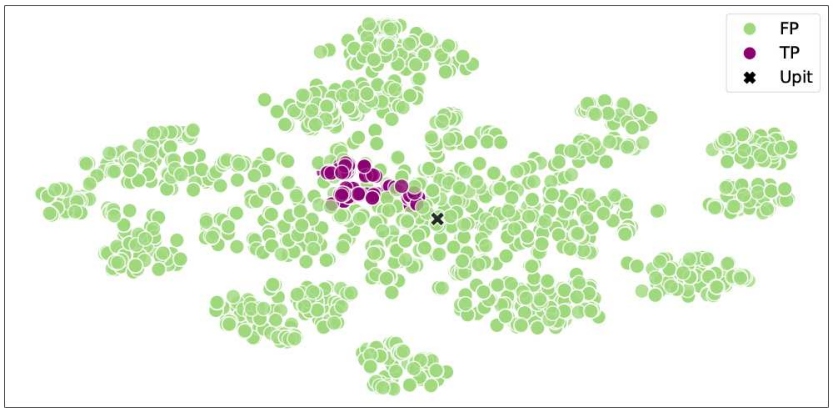
Slika 4.7: Rajčica UniProt, skala 4.0



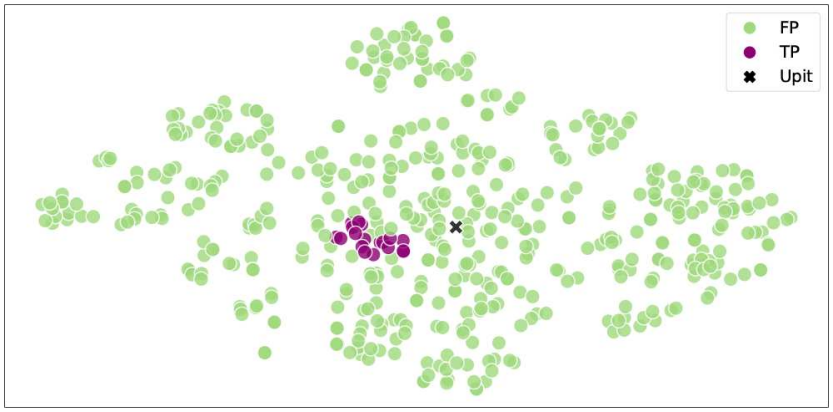
Slika 4.8: Rajčica ITAG2.3, skala 4.0



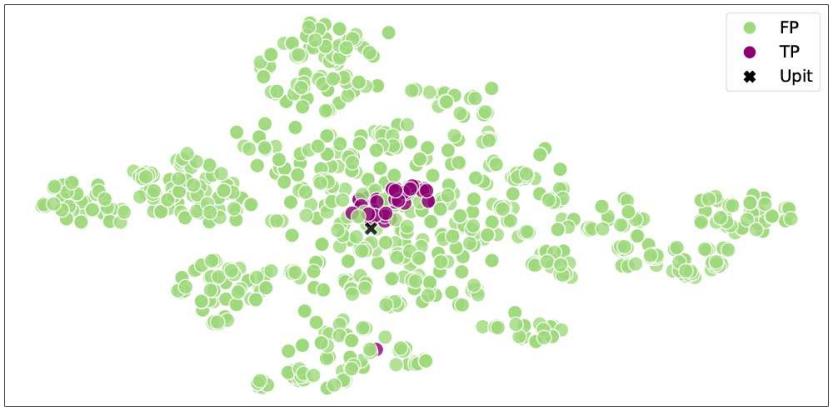
Slika 4.9: Sirak, skala 4.0



Slika 4.10: Obična pšenica, skala 4.0



Slika 4.11: Vinova loza, skala 4.0



Slika 4.12: Kukuruz, skala 4.0

Nakon toga, konačno dajemo i rezultate metode. Budući da naša metoda djeluje u kombinaciji s IGLOSS-om, uspješnost metode smo računali uzimajući u obzir i uspješnost IGLOSS-a. Dakle, računajući TPR metode, a time i F_1-*score* (u daljnjem tekstu: F1), u obzir smo uzeli cijeli skup CP-ova, neovisno o tome jesu li se svi nalazili u skupu pozitivaca. Međutim, kako bismo dobili bolji uvid u uspješnost isključivo naše metode, u zagradama navodimo TPR koji u obzir uzima samo prave pozitivce koje je IGLOSS pronašao. Dakle, uspješnost u zagradama nema utjecaj IGLOSS-a.

Skala	Metoda	TP	P	PPV	TPR	F1	Radijus
4.0	*Benchmark*	13	19	0.68	0.59 (0.62)	0.63 (0.65)	3.48
	IGLOSS + Kugla	17	47	0.36	0.77 (0.81)	0.49 (0.50)	4.58
5.0	*Benchmark*	13	17	0.76	0.59 (0.62)	0.67 (0.68)	3.47
	IGLOSS + Kugla	17	44	0.39	0.77 (0.81)	0.52 (0.53)	4.63
6.0	*Benchmark*	16	27	0.59	0.73 (0.76)	0.65 (0.66)	3.70
	IGLOSS + Kugla	17	39	0.44	0.77 (0.81)	0.56 (0.57)	4.58

Tablica 4.1: Talijin uročnjak (lat. *Arabidopsis thaliana*)

Skala	Metoda	TP	P	PPV	TPR	F1	Radijus
4.0	*Benchmark*	24	27	0.89	0.62 (0.96)	0.73 (0.92)	4.23
	IGLOSS + Kugla	23	32	0.72	0.59 (0.92)	0.65 (0.81)	4.62
5.0	*Benchmark*	25	30	0.83	0.64 (1.00)	0.72 (0.91)	4.36
	IGLOSS + Kugla	23	31	0.74	0.59 (0.92)	0.66 (0.82)	4.67
6.0	*Benchmark*	25	29	0.86	0.64 (1.00)	0.74 (0.92)	4.40
	IGLOSS + Kugla	25	30	0.83	0.64 (1.00)	0.72 (0.91)	4.70

Tablica 4.2: Čaj (lat. *Camellia sinensis*)

Skala	Metoda	TP	P	PPV	TPR	F1	Radijus
4.0	*Benchmark*	79	82	0.96	0.94 (0.98)	0.95 (0.97)	4.25
	IGLOSS + Kugla	79	92	0.86	0.94 (0.98)	0.90 (0.92)	4.66
5.0	*Benchmark*	79	84	0.94	0.94 (0.98)	0.94 (0.96)	4.31
	IGLOSS + Kugla	79	87	0.91	0.94 (0.98)	0.92 (0.94)	4.69
6.0	*Benchmark*	79	83	0.95	0.94 (0.98)	0.95 (0.96)	4.41
	IGLOSS + Kugla	79	83	0.95	0.94 (0.98)	0.95 (0.96)	4.70

Tablica 4.3: Soja (lat. *Glycine max*)

Skala	Metoda	TP	P	PPV	TPR	F1	Radijus
4.0	*Benchmark*	27	30	0.90	0.87 (0.87)	0.89 (0.88)	3.81
	IGLOSS + Kugla	29	41	0.71	0.94 (0.94)	0.81 (0.81)	4.62
5.0	*Benchmark*	27	31	0.87	0.87 (0.90)	0.87 (0.88)	3.77
	IGLOSS + Kugla	29	41	0.71	0.94 (0.97)	0.81 (0.82)	4.60
6.0	*Benchmark*	27	31	0.87	0.87 (0.93)	0.87 (0.90)	3.81
	IGLOSS + Kugla	28	38	0.74	0.90 (0.97)	0.81 (0.84)	4.49

Tablica 4.4: Obični ječam (lat. *Hordeum vulgare subsp. vulgare*)

Skala	Metoda	TP	P	PPV	TPR	F1	Radijus
5.0	*Benchmark*	7	34	0.21	0.58 (0.64)	0.30 (0.32)	3.23
	IGLOSS + Kugla	11	91	0.12	0.92 (1.00)	0.21 (0.21)	4.73
6.0	*Benchmark*	7	30	0.23	0.58 (0.64)	0.33 (0.34)	3.21
	IGLOSS + Kugla	9	70	0.13	0.75 (0.82)	0.22 (0.22)	4.78
7.0	*Benchmark*	5	20	0.25	0.42 (0.45)	0.31 (0.32)	3.34
	IGLOSS + Kugla	11	87	0.13	0.92 (1.00)	0.22 (0.23)	5.22

Tablica 4.5: Divlja banana (lat. *Musa acuminata subsp. malaccensis*)

Skala	Metoda	TP	P	PPV	TPR	F1	Radijus
5.0	*Benchmark*	20	24	0.83	0.63 (0.65)	0.71 (0.73)	3.30
	IGLOSS + Kugla	29	65	0.45	0.91 (0.94)	0.60 (0.61)	4.69
6.0	*Benchmark*	20	24	0.83	0.63 (0.65)	0.71 (0.73)	3.33
	IGLOSS + Kugla	29	62	0.47	0.91 (0.94)	0.62 (0.63)	4.71
7.0	*Benchmark*	20	24	0.83	0.63 (0.65)	0.71 (0.73)	3.59
	IGLOSS + Kugla	26	58	0.45	0.81 (0.84)	0.58 (0.59)	5.13

Tablica 4.6: Pravi duhan (lat. *Nicotiana tabacum*)

Skala	Metoda	TP	P	PPV	TPR	F1	Radijus
4.0	*Benchmark*	20	21	0.95	0.63 (0.83)	0.75 (0.89)	4.32
	IGLOSS + Kugla	19	23	0.83	0.59 (0.79)	0.69 (0.81)	4.50
5.0	*Benchmark*	19	19	1.00	0.59 (0.79)	0.75 (0.88)	4.43
	IGLOSS + Kugla	19	23	0.83	0.59 (0.79)	0.69 (0.81)	4.57
6.0	*Benchmark*	24	29	0.83	0.75 (1.00)	0.79 (0.91)	5.09
	IGLOSS + Kugla	19	19	1.00	0.59 (0.79)	0.75 (0.88)	4.53

Tablica 4.7: Rajčica (UniProt, lat. *Solanum lycopersicum*)

Skala	Metoda	TP	P	PPV	TPR	F1	Radijus
4.0	*Benchmark*	5	12	0.42	1.00 (1.00)	0.59 (0.59)	3.61
	IGLOSS + Kugla	5	26	0.19	1.00 (1.00)	0.32 (0.32)	4.51
5.0	*Benchmark*	5	12	0.42	1.00 (1.00)	0.59 (0.59)	3.65
	IGLOSS + Kugla	5	26	0.19	1.00 (1.00)	0.32 (0.32)	4.59
6.0	*Benchmark*	5	12	0.42	1.00 (1.00)	0.59 (0.59)	3.63
	IGLOSS + Kugla	5	23	0.22	1.00 (1.00)	0.36 (0.36)	4.51

Tablica 4.8: Rajčica (ITAG2.3, lat. *Solanum lycopersicum*)

Skala	Metoda	TP	P	PPV	TPR	F1	Radijus
4.0	*Benchmark*	40	44	0.91	0.89 (0.93)	0.90 (0.92)	4.45
	IGLOSS + Kugla	40	46	0.87	0.89 (0.93)	0.88 (0.90)	4.67
5.0	*Benchmark*	40	43	0.93	0.89 (0.93)	0.91 (0.93)	4.48
	IGLOSS + Kugla	40	45	0.89	0.89 (0.93)	0.89 (0.91)	4.72
6.0	*Benchmark*	40	44	0.91	0.89 (0.93)	0.90 (0.92)	4.42
	IGLOSS + Kugla	40	47	0.85	0.89 (0.93)	0.87 (0.89)	4.72

Tablica 4.9: Sirak (lat. *Sorghum bicolor*)

Skala	Metoda	TP	P	PPV	TPR	F1	Radijus
4.0	*Benchmark*	109	122	0.89	0.88 (0.95)	0.89 (0.92)	4.27
	IGLOSS + Kugla	112	138	0.81	0.90 (0.97)	0.85 (0.88)	4.61
5.0	*Benchmark*	109	122	0.89	0.88 (0.95)	0.89 (0.92)	4.27
	IGLOSS + Kugla	112	140	0.80	0.90 (0.97)	0.85 (0.88)	4.65
6.0	*Benchmark*	108	118	0.92	0.87 (0.94)	0.89 (0.93)	4.32
	IGLOSS + Kugla	108	126	0.86	0.87 (0.94)	0.86 (0.90)	4.62

Tablica 4.10: Obična pšenica (lat. *Triticum aestivum*)

Skala	Metoda	TP	P	PPV	TPR	F1	Radijus
4.0	*Benchmark*	18	19	0.95	0.90 (1.00)	0.92 (0.97)	3.88
	IGLOSS + Kugla	18	22	0.82	0.90 (1.00)	0.86 (0.90)	4.58
5.0	*Benchmark*	18	19	0.95	0.90 (1.00)	0.92 (0.97)	3.91
	IGLOSS + Kugla	18	22	0.82	0.90 (1.00)	0.86 (0.90)	4.64
6.0	*Benchmark*	18	18	1.00	0.90 (1.00)	0.95 (1.00)	3.97
	IGLOSS + Kugla	18	20	0.90	0.90 (1.00)	0.90 (0.95)	4.71

Tablica 4.11: Vinova loza (lat. *Vitis vinifera*)

Skala	Metoda	TP	P	PPV	TPR	F1	Radijus
4.0	*Benchmark*	42	54	0.78	0.86 (0.88)	0.82 (0.83)	3.93
	IGLOSS + Kugla	45	65	0.69	0.92 (0.94)	0.79 (0.80)	4.61
5.0	*Benchmark*	45	59	0.76	0.92 (0.94)	0.83 (0.84)	4.51
	IGLOSS + Kugla	45	63	0.71	0.92 (0.94)	0.80 (0.81)	4.69
6.0	*Benchmark*	42	50	0.84	0.86 (0.88)	0.85 (0.86)	3.96
	IGLOSS + Kugla	45	62	0.73	0.92 (0.94)	0.81 (0.82)	4.66

Tablica 4.12: Kukuruz (lat. *Zea mays*)

4.3 Diskusija

Najprije ćemo komentirati grafičke prikaze. Vidimo da se pravi pozitivci u većini uzoraka grupiraju. Soju i vinovu lozu možemo izdvojiti kao primjere gdje je grafički pretpostavka o grupiranju odlično zadovoljena. S druge strane, ako pogledamo prave pozitivce iz divlje banane ili ITAG2.3 rajčice, vidimo da su potpuno raspršeni te da se među njima nalazi velik broj lažnih pozitivaca. Slično, ako pogledamo talijin uročnjak i pravi duhan možemo vidjeti da je popriličan broj lažno pozitivnih odgovora susjedan pravim odgovorima. Kasnije, kada

budemo komentirali rezultate *Benchmark*-a, i to one bez utjecaja IGLOSS-a, vidjet ćemo da nas je t-SNE naveo na dobre zaključke.

Nadalje, zanimljivo je primijetiti susjedstvo upita i pravih pozitivaca. Za veliku većinu uzoraka upit se ne nalazi u blizini. Izuzeci su soja i kukuruz gdje izgleda da se upit grupira zajedno s pravim odgovorima. Ovo bi moglo sugerirati da upit možda nije reprezentativan, a izgleda i da bi mogao biti različit za različite biljke. Također, zanimljivo je uočiti da postoje i grupiranja lažnih pozitivaca (potencijalno je riječ o nekim drugim familijama) koja bi mogla biti pronađena kao najgušća kugla. Kasnije ćemo komentirati da se upravo to dogodilo s uzorcima iz sirka. Primijetimo još da se u nekim uzorcima mogu uočiti i *outlieri*. Moguće je da je riječ o biološkim izuzecima ili se pojavljuju kao posljedica naše standardizacije ili pak krive anotacije.

U nastavku ćemo komentirati rezultate metode. Napomenimo da ćemo ITAG2.3 rajčicu trenutno izostaviti iz analize te ćemo je zasebno komentirati. Promotrimo koliko je metoda bila uspješna u postizanju svog cilja. Nastojala je zadržati što više pravih pozitivaca što pokazuje TPR koji je u prosjeku pao 7%. S druge strane, metoda je izbacila značajan broj lažnih pozitivaca — PPV je u prosjeku porastao za 60%. Ovime smo u prosjeku postigli rast od F1 za 57%.

Metoda se pokazala robusnom na promjene u veličini uzorka. Naime, odabirom različitih skala dobivali smo uzorke znatno različitih veličina (npr. broj pozitivaca koje je IGLOSS vratio iz proteoma soje na skali 4 je 1434, a na skali 5 je 920). Međutim, za uglavnom sve proteome razlika u F1 po različitim skalama je manja od 5%. Iznimke su proteomi talijinog uročnjaka, čaja i rajčice s razlikama manjima od 7%.

Kako bismo vidjeli koliko je metoda mogla biti uspješnija, napravimo usporedbu s uspješnosti koju je postigao *Benchmark*. Razlika u F1 je za većinu uzoraka manja od 10%. Preciznije, razlika za 15 uzoraka je manja od 5%. Za njih 12 je između 5% i 10%. Ostalih 6 uzoraka dobivenih iz proteoma talijinog uročnjaka, čaja, divlje banane i duhana imaju nešto veće razlike, a one iznose od 10% do 15%.

Ako usporedimo optimalni i procijenjeni radijus, vidimo da i tu veće razlike postoje. Optimalni radijus se nalazi u rasponu od 3.21 do 5.09, s prosjekom 4.01. S druge strane, procijenjeni radijus se kreće od 4.49 do 5.22, a u prosjeku iznosi 4.68. Opaženo je da se smanjivanjem procijenjenog radijusa prema optimalnom radijusu F1 povećava, no kad ta razlika postane dovoljno mala F1 značajno padne jer procjena središta također odstupa od težišta pravih pozitivaca.

U gotovo svim provedenim testovima, pronađena najgušća kugla je zaista ona koja sadrži značajan broj pravih pozitivaca. Izuzeci su uzorci iz proteoma sirka na skalama 4.0 i 5.0 u kojima isprva pronađena najgušća kugla u sebi nije sadržavala niti jedan pravi odgovor, niti upit. Iz tog razloga, na ta dva uzorka smo naknadno uvjetovali pronalazak najgušće kugle koja sadrži upit te tako pronašli prave odgovore. Što se tiče ostalih uzoraka, provjerili smo da dobivene kugle zaista sadrže upit.

Kao što je već rečeno, broj pravih pozitivaca koje je kugla pronašla je odozgo ograničen s brojem pravih pozitivaca koje je IGLOSS pronašao. Ovdje to još jednom naglašavamo jer je upravo to značajno utjecalo i na neke naše rezultate. Naime, IGLOSS-ov TPR za sve tri skale na proteomu čaja iznosi svega 64%. Nešto veći, ali i dalje znatno manji od svih ostalih, je TPR za sve tri skale na proteomu UniProt rajčice u iznosu od 75%.

Što se tiče ispunjavanja pretpostavke o grupiranju pravih pozitivaca, vidimo da je ona najlošije ispunjena za prave pozitivce iz proteoma divlje banane, talijinog uročnjaka i duhana, redom. Ovdje je bitno napomenuti da pronađene kugle na uzorku iz divlje banane sadrže oko 83% motiva pronađenih u hipotetskim proteinima. Riječ je o proteinima za koje je predviđeno da postoje, no za to imamo manjak eksperimentalnih dokaza. Za ostale proteome možemo reći da je pretpostavka uglavnom dobro ispunjena.

Preostaje komentirati ITAG2.3 rajčicu. U njenom proteomu pronašli smo svega 5 proteina za koje je anotirano da sadrže VQ motiv. U UniProt rajčici pronašli smo 32 takva. Na slici 4.13 nalazi se ispis motiva koje je naša kugla pronašla u ITAG2.3 rajčici na skali 6.0, njihova udaljenost od središta kugle, je li motiv promatran kao TP ili FP te anotacija. Zanimljivo je primijetiti da su svi bojom označeni motivi u ITAG2.3 rajčici anotirani kao VQ proteini u UniProt rajčici. Dakle, vidimo da zaista postoji značajna razlika u anotaciji između ta dva proteoma. Međutim, na slici 4.13 bitno je uočiti da određene sugestije za VQ proteine u anotaciji postoje.

```
2.24 FP FRALVQKLTG >Solyc03g119410.1.1 Unknown Protein IPR008889 VQ
2.29 FP FKALVQKLTG >Solyc06g069100.1.1 NAC domain protein IPR003441 protein
2.72 TP FRAVVQRLTG >Solyc03g117500.1.1 VQ motif-containing protein IPR008889 VQ
2.73 FP FRALVQQLTG >Solyc10g078440.1.1 Unknown Protein IPR008889 VQ
2.74 FP FRDVVQKLTG >Solyc04g074520.1.1 Unknown Protein IPR008889 VQ
2.84 FP FRSLVQELTG >Solyc01g096510.2.1 Sigma factor binding protein 1 IPR008889 VQ
2.97 FP FKSIVQKLTG >Solyc07g056600.1.1 Tobacco rattle virus-induced protein variant 2 IPR008889 VQ
2.99 FP FKSVVQKFTG >Solyc10g007580.1.1 Tobacco rattle virus-induced protein variant 2 IPR008889 VQ
3.02 FP FRALVQQFTG >Solyc02g068460.1.1 Unknown Protein IPR008889 VQ
3.07 FP FRSVVQQLTG >Solyc06g061190.1.1 Hydroxyproline-rich glycoprotein family protein IPR008889 VQ
3.25 FP FKQVVQMLTG >Solyc02g078030.1.1 DNA-binding WRKY VQ IPR008889 VQ
3.25 FP FKQVVQMLTG >Solyc07g063070.1.1 DNA-binding WRKY VQ IPR008889 VQ
3.29 FP FREVVQRLTG >Solyc12g088490.1.1 DNA-binding WRKY VQ
3.64 TP FRAMVQEFTG >Solyc04g055050.1.1 VQ motif family protein expressed IPR008889 VQ
3.64 FP FRALVQQHTG >Solyc02g068470.1.1 Unknown Protein
3.78 TP FRQMVQEFTG >Solyc04g073950.1.1 VQ motif family protein expressed IPR008889 VQ
3.83 FP FRAMVQQFTG >Solyc07g043250.1.1 Unknown Protein IPR008889 VQ
3.94 TP FMSVVQRLTG >Solyc11g005720.1.1 VQ motif family protein IPR008889 VQ
3.95 TP FMSLVQRLTG >Solyc06g060470.1.1 VQ motif family protein IPR008889 VQ
4.03 FP FRAVVQQYTG >Solyc02g031990.1.1 Unknown Protein
4.06 FP FREMVQQVTG >Solyc10g077130.1.1 Atcambp25-binding protein OF IPR008889 VQ
4.10 FP FQLLVDHITG >Solyc03g063340.2.1 DNA polymerase subunit delta-2 IPR007185 DNA polymerase...
4.30 FP FYAIVHRVTG >Solyc10g044670.1.1 Phytochrome A
```

Slika 4.13: Rezultati metode za ITAG2.3 rajčicu, skala 6.0

Na kraju, možemo zaključiti da rezultati i ovog rada potvrđuju da je razvijena uspješna metoda. U većini slučajeva metoda je odbacila značajan broj lažno pozitivnih odgovora, a da je pritom zadržala što više pravih odgovora. Također, za većinu proteoma smo vidjeli i da je pretpostavka o grupiranju zadovoljena. Bitno je naglasiti da je metoda izuzetno brza, a uz to se pokazala i robusnom na promjene u broju podataka.

Bibliografija

[1] B. Rabar, K. Nižetić, M. Zagorščak, K. Gruden, 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.

[2] B. Rabar, M. Zagorščak, S. Ristov, M. Rosenzweig, P. Goldestein, *IGLOSS: iterative gapless local similarity search*, Bioinformatics, 35 (2019), 3491-3492.

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

[4] I. Višek, *Clustering i klasifikacija proteinskih nizova*, Diplomski rad, Sveučilište u Zagrebu, Prirodoslovno-matematički fakultet, 2022.

[5] International Tomato Genome Sequencing Project, `https://solgenomics.net/organism/Solanum_lycopersicum/genome` (veljača 2022.).

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

[7] M. Iveković, *Traženje proteinskih motiva i klasifikacija*, Diplomski rad, Sveučilište u Zagrebu, Prirodoslovno-matematički fakultet, 2022.

[8] M. Šarić, *Proteini*, Završni rad, Sveučilište u Zagrebu, Prirodoslovno-matematički fakultet, Zagreb, 2022.

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

[10] t-SNE, `https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html` (veljača 2022.).

[11] UniProt, `https://www.uniprot.org/` (veljača 2022.).

[12] V. Bokšić, *Proteinski motivi i klasifikacija*, Diplomski rad, Sveučilište u Zagrebu, Prirodoslovno-matematički fakultet, 2021.

[13] W. R. Atchley, J. Zhao, A. D. Fernandes, T. Drüke, *Solving the protein sequence metric problem*, Proceedings of the National Academy of Sciences of the United States of America, 102 (2005), 6395-400.

[14] H. Ding, G. Yuan, S. Mo, Y. Qian, Y. Wu, Q. Chen, X. Xu, X. Wu, C. Ge, *Identification, characterization and expression analysis of the VQ motif-containing gene family in tea plant (Camellia sinensis)*, BMC Genomics 19, 710 (2018), 29-39.

[15] J. Xiong , *Essential Bioinformatics*, Cambridge University Press, 2006.

Sažetak

Ovaj diplomski rad proučava problem klasifikacije proteina u proteinske familije. Proteini se klasificiraju identifikacijom kratkog niza aminokiselina kojeg sadrže, a zovemo ga motivom. Motiv je ostao djelomično očuvan u evoluciji te se može povezati s prepoznatljivim dijelom u strukturi proteina koji izvršava određenu funkciju.

Za razliku od tipičnih pretraživača motiva koji promatraju sličnost potencijalnih motiva sa zadanim motivom, u ovom radu se opisuje i testira metoda koja promatra njihovu međusobnu sličnost. Metoda je zamišljena da djeluje u kombinaciji s nekim pretraživačem motiva pružajući mu tako svoj pristup u traženju odgovora. Uz pomoć opisa aminokiselina numeričkim vektorima, problem se smješta u euklidski prostor gdje se umjesto sličnosti promatra udaljenost. Pretpostavka je da se motivi, pronađeni u proteinima koji biološki pripadaju promatranoj familiji, grupiraju. Cilj metode pronaći je optimalnu kuglu koja će obuhvatiti što veći broj pravih odgovora uz što manje pogrešnih.

Metoda je testirana na dvanaest biljnih proteoma u čijim proteinima se tražilo karakteristične motive VQ proteinske familije. U većini slučajeva, metoda se pokazala uspješnom što je ujedno i potvrdilo pretpostavku o grupiranju.

Summary

This thesis studies the protein family classification problem. Proteins are classified by identifying a short amino acids sequence contained within them. This short sequence is called a motif. During evolution, the motif has remained partly conserved and can be associated with the recognizable part of the protein structure that performs a distinct function.

Unlike typical motif scanners that observe the similarity of potential motifs with a given motif, this thesis describes and tests the method that considers pairwise similarity. The method works in combination with another motif scanner by providing additional criteria for motif identification. By describing amino acids using numerical vectors the problem is placed in Euclidean space where distance is considered instead of similarity. The assumption is that the motifs, found in proteins that belong biologically to the observed family, will group. The goal of this method is to find the optimal ball that will contain as many correct hits and as few wrong ones as possible.

The method was tested on twelve plant proteomes whose proteins were searched for containing characteristic motifs of the VQ protein family. In most of the cases, the method proved to be successful which in turn confirmed true positives clustering assumption.

Životopis

Rođena sam 1997. godine u Zagrebu. Nakon završene Osnovne škole Novska u Novskoj, srednjoškolsko obrazovanje nastavljam u XI. Gimnaziji u Zagrebu. Godine 2015. upisujem preddiplomski sveučilišni studij Matematika na Matematičkom odsjeku Prirodoslovno-matematičkog fakulteta u Zagrebu. Titulu prvostupnice matematike stječem 2020. godine, a iste godine i upisujem diplomski sveučilišni studij Matematička statistika na istom fakultetu.