Analiza proteinskih poravnanja

Belcar, Eva

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Eva Belcar

ANALIZA PROTEINSKIH PORAVNANJA

Diplomski rad

Voditelj rada:
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. _____

Sadržaj

Uvod

Proteini su jedne od najvažnijih tvari organizma. Utječu na rast i razvoj tkiva, pomažu u izgradnji mišića i organa, a odgovorni su i za mnoge druge procese. Svaki protein građen je od aminokiselina i ima svoju funkciju koja se određuje na temelju tog niza. Skup proteina nekog organizma naziva se proteom. U ovom radu želimo odrediti koji proteini imaju istu funkciju u proteomu nekog organizma. Takvi proteini čine proteinsku familiju i za njih kažemo da su slični. Kako su proteini dugi nizovi aminokiselina, u svakom od njih tražit ćemo karakterističan podniz od deset aminokiselina za odabranu proteinsku familiju. Taj podniz naziva se motiv.

Bavit ćemo se određenom proteinskom familijom, odnosno motivom koji ju opisuje. Prevodimo ih u vektorski prostor pomoću faktora koji opisuju aminokiseline. Razvijamo metodu kojom tražimo središte i radijus kugle koja sadrži najviše motiva. Pokazuje se da se, uz neke ispunjene uvjete, u pet promatranih proteoma traženi motivi grupiraju u toj kugli.

Prvo ćemo definirati matematičke pojmove i teoreme koji će nam biti potrebni za razvoj metode. Za provjeru točnosti modela definiramo pojmove i mjere iz područja strojnog učenja, a za razumijevanje tematike rada definirat ćemo pojmove iz biologije i bioinformatike. Nakon toga opisujemo razvijanje metode i navodimo rezultate.

Poglavlje 1

Matematički pojmovi

Navodimo definicije, teoreme, propozicije i napomene iz područja linearne algebre, metričkih prostora, statistike i mjera uspješnosti iz izvora [2], [6], [8], [10], [4] i [3]

1.1 Linearna algebra i metrički prostori

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

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

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

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

Definicija 1.1.3. *Neka je V neprazan skup na kojem su zadane binarne operacije zbrajanja $+ : V \times V \to V$ i operacija množenja skalarima iz polja $\mathbb{F}$, $\cdot : \mathbb{F} \times V \to V$. Kažemo da je uređena trojka $(V, +, \cdot)$ vektorski prostor nad poljem $\mathbb{F}$ ako vrijedi:*

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

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

Definicija 1.1.7. *Neka je V unitaran prostor. Norma na V je funkcija $\| \cdot \| : V \to \mathbb{R}$ definirana s*

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

Teorem 1.1.8 (Cauchy-Schwarzova nejednakost)**.** *Neka je V unitaran prostor. Tada je*

$$|\langle x, y\rangle|^2 \leq \langle x, x\rangle\langle y, y\rangle, \quad \forall x, y \in V$$

Jednakost vrijedi ako i samo ako vrijedi $y = \alpha x$ za neki $\alpha \in \mathbb{F}$.

Dokaz. Neka je V unitaran prostor i neka su $x, y \in V$. Primijetimo da je skalarni produkt antilinearan na drugom argumentu, odnosno da vrijedi

$$\langle x, \alpha y\rangle = \overline{\alpha}\langle x, y\rangle, \quad \forall x, y \in V, \forall \alpha \in \mathbb{F}.$$

Tvrdnja je očita ako $x = 0$ ili $y = 0$, pa pretpostavimo da su x i y netrivijalni vektori i neka je $\alpha \in \mathbb{F}$ proizvoljan skalar. Tada je

$$0 \leq \langle x - \alpha y, x - \alpha y\rangle = \langle x, x\rangle - \alpha\langle y, x\rangle - \overline{\alpha}\langle x, y\rangle + \alpha\overline{\alpha}\langle y, y\rangle$$

U gornju nejednakost uvrstimo $\alpha = \frac{\langle x,y\rangle}{\langle y,y\rangle}$. Uočimo da je to dozvoljeno jer znamo $y \neq 0$, što povlači $\langle y, y\rangle \neq 0$. Tada vrijedi

$$0 \leq \langle x, x\rangle - \frac{\langle x, y\rangle}{\langle y, y\rangle}\langle y, x\rangle - \frac{\langle y, x\rangle}{\langle y, y\rangle}\langle x, y\rangle + \frac{\langle x, y\rangle\langle y, x\rangle}{\langle y, y\rangle}\langle y, y\rangle$$

Uočimo da zadnja dva izraza u zbroju daju 0. Kada jednakost pomnožimo s $\langle y, y\rangle$ dobivamo

$$0 \leq \langle x, x\rangle\langle y, y\rangle - \langle x, y\rangle\langle y, x\rangle$$

manipulacijama dolazimo do željene nejednakosti

$$|\langle x, y\rangle|^2 = \langle x, y\rangle\langle y, x\rangle \leq \langle x, x\rangle\langle y, y\rangle$$

Očito je da se jednakost postiže za $y = \alpha x$. S druge strane, ako vrijedi jednakost, onda provedeni račun pokazuje da vrijedi $y = \alpha x$. □

Propozicija 1.1.9. *Norma na unitarnom prostoru V ima sljedeća svojstva:*

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

Dokaz. Jedino svojstvo koje nije trivijalno dokazati je zadnje koje se naziva nejednakost trokuta. Za dokaz svojstva koristimo Cauchy–Schwarzov teorem 1.1.8 iz kojeg slijedi

$$\begin{aligned}\|x+y\|^2 &= \langle x+y, x+y\rangle = \langle x,x\rangle + \langle x,y\rangle + \langle y,x\rangle + \langle y,y\rangle \\ &= \langle x,x\rangle + \langle y,y\rangle + \mathrm{Re}\langle x,y\rangle \\ &\leq \langle x,x\rangle + \langle y,y\rangle + 2|\langle x,y\rangle| \\ &\leq \langle x,x\rangle + \langle y,y\rangle + 2\|x\|\,\|y\| = (\|x\| + \|y\|)^2 .\end{aligned}$$

□

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

Definicija 1.1.11. *Norma koja potječe od kanonskog skalarnog produkta na $\mathbb{F}^n$, definiranog u napomeni 1.1.5, dana je formulom*

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

Ova se norma zove euklidska norma.

Definicija 1.1.12. *Neka je V normiran prostor. Metrika ili udaljenost vektora x i y je funkcija $d : V \times V \to \mathbb{R}$ definirana s*

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

Propozicija 1.1.13. *Metrika na normiranom prostoru ima sljedeća svojstva:*

1. $d(x,y) \geq 0,\ \forall x,y \in V$;
2. $d(x,y) = 0 \Leftrightarrow x = y,\ \forall x,y \in V$;
3. $d(x,y) = d(y,x),\ \forall x,y \in V$;
4. $d(x,y) \leq d(x,z) + d(z,y),\ \forall x,y,z \in V$.

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

Definicija 1.1.15. *Neka su $x = (x_1,\ldots,x_n)$ i $y = (y_1,\ldots,y_n)$ proizvoljni vektori u $\mathbb{R}^n$. Metrika na $\mathbb{R}^n$, inducirana euklidskom normom iz definicije 1.1.11, dana je s*

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

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

Definicija 1.1.16. *Neka je (X, d) metrički prostor. Za proizvoljan $a \in X$ i proizvoljan $r > 0 \in \mathbb{R}$ skup*

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

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

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

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

1.2 Teorija vjerojatnosti

Matematičko očekivanje i varijanca

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

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

$$X = \sum_{k=1}^{n} x_k \mathbb{1}_{A_k}$$

gdje su $x_1, x_2, \ldots, x_n$ realni brojevi, a $A_1, A_2, \ldots, A_n \in \mathcal{F}$ međusobno disjunktni događaji.

Definicija 1.2.1. *Matematičko očekivanje od X ili, kraće, očekivanje od X koje označavamo sa* $\mathbb{E}[X]$ *definira se sa:*

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

Propozicija 1.2.2. *1. Neka je* $c \in \mathbb{R}$ *i* $X \in \mathcal{K}$. *Tada je* $\mathbb{E}[cX] = c\mathbb{E}X$.

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

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

Teorem 1.2.3. *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.

Neka je X nenegativna slučajna varijabla definirana na Ω. Prema teoremu 1.2.3 postoji rastući niz $(X_n)_{n\in\mathbb{N}}$ nenegativnih jednostavnih slučajnih varijabli takav da je $X = \lim_{n\to\infty} X_n$. Iz propozicije 1.2.2 slijedi da je niz $(\mathbb{E}[X_n])_{n\in\mathbb{N}}$ rastući niz u $\mathbb{R}_+$, dakle postoji $\lim_{n\to\infty} \mathbb{E}[X_n]$ koji može biti jednak i $+\infty$. Definiramo matematičko očekivanje nenegativne slučajne varijable.

Definicija 1.2.4. *Matematičko očekivanje od X ili, kraće, očekivanje od X definira se sa*

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

Konačno, neka je sada X proizvoljna slučajna varijabla na Ω, tada vrijedi

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

gdje su X^+, X^- slučajne varijable i $X^+, X^- \geq 0$.

Definicija 1.2.5. *Kažemo da matematičko očekivanje od X ili kraće, očekivanje od X postoji ili da je definirano ako je barem jedna od veličina* $\mathbb{E}[X^+], \mathbb{E}[X^-]$ *konačna, tj. vrijedi* $min\{\mathbb{E}[X^+], \mathbb{E}[X^-]\} < +\infty$. *Tada po definiciji stavljamo*

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

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

Definicija 1.2.6. $\mathbb{E}(X^r)$ *zovemo r-ti moment od X, a* $\mathbb{E}(|X|^r)$ *zovemo r-ti apsolutni moment od X*

Definicija 1.2.7. *Neka* $\mathbb{E}X$ *postoji (tj. konačno je). Tada* $\mathbb{E}[(X-\mathbb{E}X)^r]$ *zovemo r-ti apsolutni centralni moment od X.*

Definicija 1.2.8. *Varijanca od X koju označavamo sa* $Var(X)$ *ili* σ_X^2 *jest drugi centralni moment od X, dakle*

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

Napomena 1.2.9. *Pozitivan drugi korijen iz varijance nazivamo standardna devijacija i označavamo sa* σ_X.

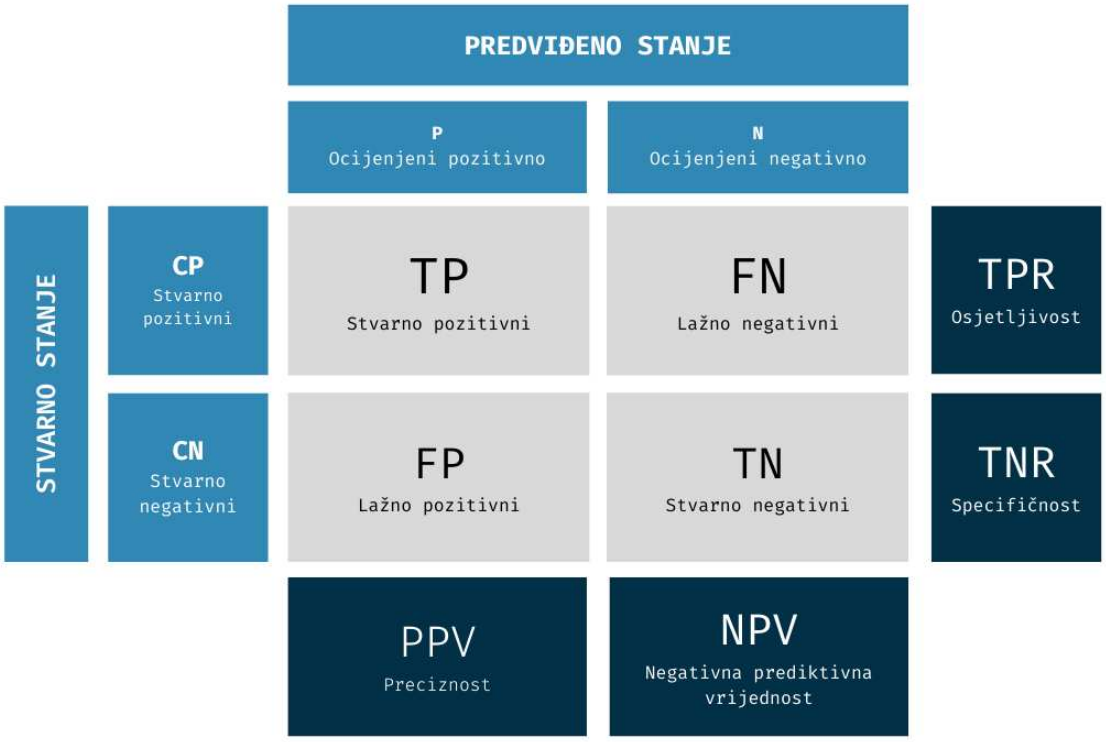
1.3 Evaluacija modela

Klasifikacija

Klasifikacija podataka je problem određivanja pripadnosti opservacije nekoj skupini odnosno klasi. Razlikujemo dvije vrste klasifikacije, nadzirana i nenadzirana. U nadziranom slučaju, model koristi podatke kojima su poznati izlazni podaci, odnosno zna se kojoj klasi pripadaju. Na temelju tih podataka uči kako klasificirati nove ulazne podatke. U nenadziranom slučaju, model unaprijed ne zna koje klase postoje, nego pokušava pronaći sličnosti između ulaznih podatka i na temelju tih sličnosti definira klase.

Mjere uspješnosti

Za ocjenjivanje uspješnosti modela definiramo mjere koje se temelje na pojmovima iz matrice uspješnosti (eng. *confusion matrix*). Jednom kada je poznata mjera kojom evaluiramo model, možemo ih uspoređivati i odabrati najbolji.



Tablica 1.1: Tablica uspješnosti

Napomena 1.3.1. *U ovom radu će od najveće važnosti biti lista CP-a (eng. Condition Positive). Ona sadrži sve proteine za koje je pripadnost odabranoj familiji već utvrđena. Na temelju te liste ćemo izračunati TP (eng. True Positives) i ostale brojeve iz matrice uspješnosti (FP, FN, TN). Dakle, u savršenom testu bi svi proteini s liste CP bili pozitivno ocijenjeni, a svi proteini koji nisu na listi CP bi bili negativno ocijenjeni.*

Navodimo definicije nekih od mjera uspješnosti modela za binarnu klasifikaciju, odnosno klasifikaciju u kojoj su moguća samo dva ishoda.

Osjetljivost ili TPR (eng. *True Positive Rate*) je postotak pozitivnih elemenata uzorka u odnosu na određeno stanje, odnosno CP elemenata uzorka, koji su ispravno prepoznati kao pozitivni.

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

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

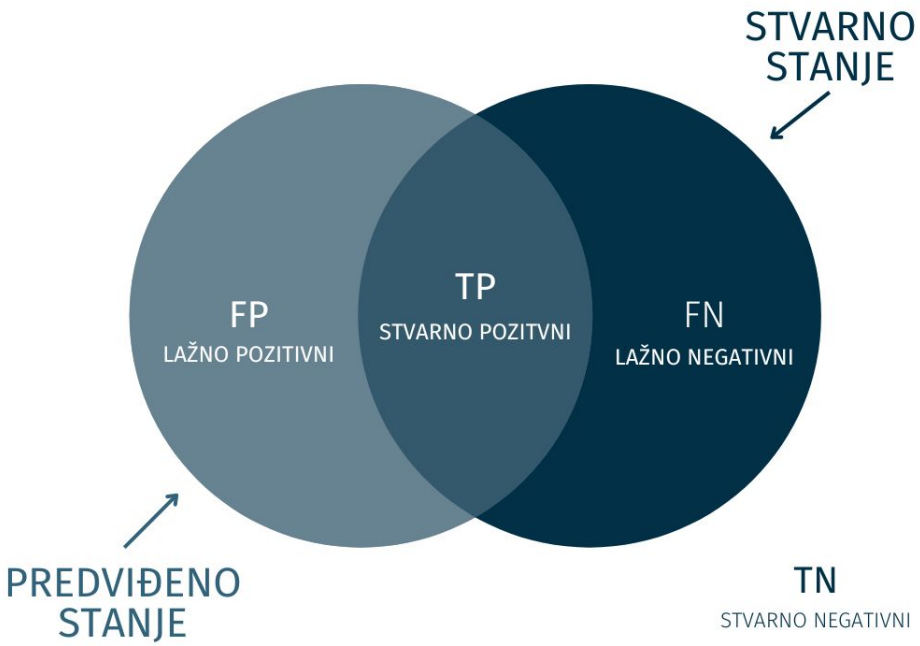
$$\text{TNR} = \frac{\text{broj stvarno negativnih}}{\text{broj stvarno negativnih} \ + \ \text{broj lažno pozitivnih}} = \frac{\text{TN}}{\text{TN} + \text{FP}} = \frac{\text{TN}}{\text{CN}}$$

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

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

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

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



Slika 1.1: Odnos preciznosti i osjetljivosti

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

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

Napomena 1.3.2. *Sve navedene mjere postižu vrijednosti isključivo na intervalu* $[0, 1]$. *Model je uspješniji po nekoj od navedenih mjera, što je ta mjera bliže broju* 1.
Faktor β *u* F_β*-score određuje kojoj mjeri dajemo veću težinu. Za* $\beta < 1$ *daje se više važnosti minimiziranju lažno pozitivnih. Za* $\beta > 1$ *daje se više važnosti minimiziranju lažno negativnih.*

Osjetljivost je omjer točno predviđenih pozitivnih elemenata i ukupnog broja stvarno pozitivnih, dok je preciznost omjer točno predviđenih pozitivnih elemenata i ukupnog broja predviđenih pozitivnih. Te mjere su najčešće suprotstavljene; ako povećamo jednu mjeru, druga se smanji i obrnuto. Također, ni jedna od njih ne uzima u obzir broj stvarnih negativaca.

U ovom radu će biti bitno izbaciti što više lažno pozitivnih i zadržati što više stvarno pozitivnih podataka, zato ćemo kao mjeru uspješnosti modela koristiti F_1-score s parametrom $\beta = 1$ jer ne želimo dati više važnosti ni preciznosti ni osjetljivosti:

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

Poglavlje 2

Bioinformatika

Aminokiseline su molekule koje rijetko dolaze u slobodnom stanju. Međusobno su povezane peptidnom vezom koja omogućava stvaranje lanca aminokiselina koji formira protein. Svi poznati proteini živog svijeta građeni su od samo dvadeset aminokiselina. Kombinacije i redoslijed aminokiselina koje ih izgrađuju igraju ulogu u određivanju neke određene funkcije u organizmu, no sam redoslijed nije dovoljan za definiranje funkcije proteina. Za utvrđivanje uloge proteina potrebno je znati i njegovu prostornu strukturu. Kompleksan proces formiranja 3D strukture proteina se zove *protein folding*.

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

Tablica 2.1: Standardne aminokiseline

Skup svih proteina nekog organizma zove se proteom, a skup svih proteina s istom funkcijom čine proteinsku familiju. U ovom radu ćemo promatrati proteinsku familiju katepsina čije uloge uključuju razgradnju proteina, metaboličku regulaciju, imunološki odgovor i odgovornost za mnoge druge funkcije. Također, izvanstanični katepsini su povezani

s mnogim bolestima kao što su rak, metastaziranje raka i kardiovaskularne bolesti. Više o samim katepsinima, nalazi se u članku [11].

Promatrat ćemo katepsine u pet različitih biljnih proteoma. Nećemo gledati cijele proteine unutar proteoma, već ćemo raditi s podnizovima od deset aminokiselina, odnosno motivima. Smatramo da su proteini iz proteinske familije katepsina ako su im pripadni motivi slični motivu katepsina.

Poglavlje 3

Algoritam i rezultati

3.1 Opis problema

Poznati pretraživači motiva za određeni motiv, odnosno niz aminokiselina koji predstavlja neku proteinsku familiju, i skalu pretraživanja daju potencijalno slične motive upitu. To rade tako da uspoređuju poravnate proteine iz proteoma s upitom i rangiraju ih. Skala pretraživanja određuje nakon kojeg mjesta se odbacuju svi koji su rangirani ispod njega. Što je ona veća to su kriteriji za sličnost stroži i odgovor pretraživača je manji broj nizova.

Cilj ovog diplomskog rada je utvrditi prave sličnosti među potencijalnim motivima međusobnim uspoređivanjem i tako doći do klasifikacije proteinske familije. Za određen upit, proteom i skalu pretraživanja, dobivamo niz motiva koje pretraživač smatra dovoljno sličnima upitu u danom proteomu. Motive koji su zaista slični upitu ćemo nazivati pravim pozitivcima, a ostatak lažnim pozitivcima. Želimo izbaciti većinu lažnih i pri tome zadržati većinu pravih pozitivaca na temelju određenih sličnosti bez informacije koji kandidati su zapravo oni koje tražimo.

Određivanje pravih pozitivaca je kompleksan i dugotrajan proces. Zahtijeva puno manualnog rada, no pouzdan je način za klasifikaciju. Postupkom opisanom u nastavku bismo na brzi način mogli odrediti slične motive međusobnim uspoređivanjem bez obzira na skalu pretraživanja.

3.2 Ideja i koncept

Motiv je niz aminokiselina od kojih je svaka opisana s pet faktora prema [1]. Faktori su dani u tablici.

i	AMINOKISELINA	Faktor I	Faktor II	Faktor III	Faktor IV	Faktor V	p_i
1	A	-0.591	-1.302	-0.733	1.570	-0.146	0.078
2	C	-1.343	0.465	-0.862	-1.020	-0.255	0.019
3	D	1.050	0.302	-3.656	-0.259	-3.242	0.053
4	E	1.357	-1.453	1.477	0.113	-0.837	0.063
5	F	-1.006	-0.590	1.891	-0.397	0.412	0.039
6	G	-0.384	1.652	1.330	1.045	2.064	0.072
7	H	0.336	-0.417	-1.673	-1.474	-0.078	0.023
8	I	-1.239	-0.547	2.131	0.393	0.816	0.053
9	K	1.831	-0.561	0.533	-0.277	1.648	0.059
10	L	-1.019	-0.987	-1.505	1.266	-0.912	0.091
11	M	-0.663	-1.524	2.219	-1.005	1.212	0.022
12	N	0.945	0.828	1.299	-0.169	0.933	0.043
13	P	0.189	2.081	-1.628	0.421	-1.392	0.052
14	Q	0.931	-0.179	-3.005	-0.503	-1.853	0.043
15	R	1.538	-0.055	1.502	0.440	2.897	0.051
16	S	-0.228	1.399	-4.760	0.670	-2.647	0.068
17	T	-0.032	0.326	2.213	0.908	1.313	0.059
18	V	-1.337	-0.279	-0.544	1.242	-1.262	0.066
19	W	-0.595	0.009	0.672	-2.128	-0.184	0.014
20	Y	0.260	0.830	3.097	-0.838	1.512	0.032

Tablica 3.1: Tablica aminokiselina, faktori koji ih opisuju i vjerojatnost pojavljivanja u prostoru proteina

Svaki od faktora opisuje jednu karakteristiku aminokiseline. *Faktor I* opisuje polaritet, *Faktor II* je faktor sekundarne strukture, *Faktor III* se odnosi na veličinu ili volumen molekule, *Faktor IV* opisuje relativnu kompoziciju određene aminokiseline u različitim proteinima, a *Faktor V* upućuje na elektrostatički naboj te aminokiseline.

Ideja rada je opisati niz od n aminokiselina pomoću faktora, čime bismo prešli u $5n-$dimenzionalni vektorski prostor. Nakon prelaska u taj prostor, upit i potencijalne kandidate možemo smatrati vektorima.

Međusobno uspoređivanje motiva i proučavanje sličnosti među njima smo prelaskom u vektorski prostor sveli na određivanje udaljenosti između vektora. Vođeni time, pretpostavljamo da se slični motivi grupiraju u kugli, a postupak se svodi na traženje središta i radijusa te kugle.

Opis i priprema podataka

Kao pretraživač motiva u ovom radu koristimo IGLOSS server čiji detaljniji opis se može naći u radu [7]. Promatramo proteinsku familiju katepsina čija funkcija je razgradnja proteina, a motiv koji ih opisuje je niz aminokiselina duljine deset, QGQCGSCWAF.

Upit koji dajemo IGLOSS serveru je gore spomenuti niz aminokiselina, skalu pretraživanja i proteom. Output koji dobivamo su poravnati nizovi aminokiselina duljine deset iz proteina u zadanom proteomu koje pretraživač smatra dovoljno sličnima za danu skalu. IGLOSS izbacuje rangirane motive i dodjeljuje im scorove koji su logistički distribuirani, a iz istog proteina može naći više poravnatih motiva. To rješavamo tako da odaberemo onaj odgovor koji ima najmanji e-value. To je p-vrijednost pomnožena s brojem napravljenih usporedbi, odnosno s brojem proteina u proteomu, a p-vrijednost je vjerojatnost pogreške prve vrste u odnosu na kritično područje kojemu je opažena vrijednost testne statistike granična vrijednost. Time dobivamo skup podataka jedinstvenih proteina čiji podnizovi čine kandidate za slične motive.

Pomoću tablice 3.2, točnije pomoću pet faktora koji opisuju svaku aminokiselinu, prevodimo nizove od deset aminokiselina u vektore dimenzije 5×10 i tako prelazimo u vektorski prostor $\mathbb{R}^{50}$. Kako svaki faktor opisuje neko obilježje pojedine aminokiseline, potrebno je standardizirati koordinate da bi bile usporedive. U suprotnome bismo mogli doći do izobličenja kugle ako jedna koordinata ima veći utjecaj od druge. Da izbjegnemo dijeljenje s jako malim brojem, prilagodit ćemo standardizaciju tako da dijelimo sa standardnom devijacijom povećanom za 0.1

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

pri čemu su $x_1, x_2 \ldots, x_n$ vrijednosti koje čine skup podataka. U našem slučaju je to $i-$ta koordinata svih n nizova koje dobijemo kao output pretraživača, $\overline{x}$ je njihova aritmetička sredina i s pripadna standardna devijacija.

3.3 *Benchmark*

Provjeru samog koncepta ovog rada radimo tako da pronalazimo željenu kuglu sa svim informacijama koje su nam dostupne. Ključnu ulogu igra lista pravih pozitivaca, točnije činjenica da znamo točno koje vektore želimo imati u kugli. Tim postupkom ćemo izračunati gornju ocjenu za uspješnost modela koji razvijamo kasnije kada ćemo htjeti pronaći slične motive bez da unaprijed znamo koji su.

Očekujemo i pojavljivanje lažnih pozitivaca u istoj kugli, pa ćemo tražiti kuglu maksimiziranjem F_1-scora. Pošto u svakom koraku znamo točno što se nalazi u nekoj kugli, odnosno koliko se pravih i koliko lažnih nalazi u njoj, tako u svakom koraku možemo uspoređivati F_1-score.

Nakon što pripremimo podatke, kao središte početne kugle stavljamo težište pravih pozitivaca. Određivanje optimalnog središta željene kugle je sam po sebi težak problem, no smatramo da je u ovom trenutku to dovoljno dobar izbor. Nadalje, iteriramo radijus kugle s gornjim središtem tako da dobijemo maksimalan F_1-score.

Na gore opisan način dobivamo gornju ocjenu, odnosno *Benchmark*, za uspješnost modela kojim određujemo slične motive.

3.4 Modeliranje i validacija

Za razvijanje modela kojim dolazimo do kugle u kojoj se grupiraju slični motivi, popis pravih pozitivaca koristimo samo za validaciju tog modela. Na samom početku dolazimo do dva bitna pitanja; što uzeti kao središte kugle i kako procijeniti radijus kojim ćemo opisati kuglu? Rješenja ta dva problema opisana su u nastavku.

Problem radijusa

U poglavlju 3.3, optimalni radijus smo dobili iteracijom radijusa jer smo mogli razlikovati prave od lažnih pozitivaca i u svakom koraku evaluirati F_1-score. Kada ne znamo tu razliku, potrebno je procijeniti radijus s kojim ćemo opisati kuglu, a to ćemo napraviti pomoću računanja očekivane udaljenosti dvije aminokiseline iz neke distribucije i sljedećeg teorema:

Teorem 3.4.1. *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.*

Dokaz teorema 3.4.1 izveden je u [5].

Empirijska distribucija aminokiselina zadana je rednim brojem i vjerojatnostima p_i navedenima u zadnjem stupcu tablice 3.2. Neka je A_i distribucija aminokiseline i očuvane koeficijentom očuvanosti α[1] dana s:

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

[1]Koeficijent očuvanosti α je koeficijent konveksne kombinacije koji opisuje u kolikom postotku očekujemo pojavljivanje aminokiseline i na j-toj poziciji.

gdje broj u sufiksu označava redni broj aminokiseline u tablici 3.2, a vjerojatnosti p^i_j se računaju kao:

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

Kao procjenu radijusa uzimamo očekivanu euklidsku udaljenost između dva niza aminokiselina duljine deset. Označimo dva takva niza s $X = (x_1, x_2, \ldots, x_{10})$ i $Y = (y_1, y_2, \ldots, y_{10})$. Očekivanje kvadrata euklidske udaljenosti možemo zapisati kao:

$$\mathbb{E}\left[d^2(X,Y)\right] = \mathbb{E}\left[(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_{10} - y_{10})^2\right] \tag{3.1}$$

Za svaku od 10 pozicija u nizu nemamo pretpostavku o aminokiselinama koje se pojavljuju na njima, a ako s $\overline{a}_i$ i $\overline{a}_j$ označimo aminokiseline koje pripadaju prosječnoj distribuciji, tada jednakost 3.1 postaje:

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

Primjenom svojstava očekivanja, desnu stranu možemo zapisati kao sumu:

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

Nadalje, želimo izračunati očekivanje s desne strane. Neka su a^k_i i a^k_j neke dvije aminokiseline iz distribucije A_k, tada vrijedi:

$$\mathbb{E}\left[(a^k_i - a^k_j)^2\right] = \sum_{i,j}^{20} \left(a^k_i - a^k_j\right)^2 p^k_i p^k_j$$

Primijetimo da distribuciju A_k odabiremo s vjerojatnošću pojavljivanja aminokiseline koja je opisuje, p_k, iz toga slijedi da je očekivanje prosječne distribucije:

$$\mathbb{E}\left[(\overline{a}_i - \overline{a}_j)^2\right] = \sum_{k}^{20} p_k \sum_{i,j}^{20} \left(a^k_i - a^k_j\right)^2 p^k_i p^k_j$$

Nakon supstitucije vjerojatnosti i faktora u gornju jednadžbu, dobijemo da je očekivani kvadrat euklidske udaljenosti dvije desetorke aminokiselina jednak:

$$\mathbb{E}\left[d^2(X,Y)\right] = 10 \cdot 9.7399$$

Prema tome, očekivana udaljenost dvije aminokiseline je $3.1209\sqrt{10}$. Primjenom teorema 3.4.1, slijedi da je očekivani radijus te kugle:

$$r = \frac{3.1209\sqrt{10}}{\sqrt{2}} = 3.1209\sqrt{5} \tag{3.2}$$

Uočimo kako je gornji radijus izračunat za nestandardizirane podatke. Također, radijus je proporcionalan standardnoj devijaciji, što znači da se poveznica između radijusa i standardne devijacije podataka, prije i nakon standardizacije, može zapisati kao:

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

pri čemu indeksi *new* i *old* označavaju podatke nakon, odnosno prije, standardizacije. Iz gornje jednakosti i procjene radijusa na nestandardiziranim podacima 3.2, dolazimo do izraza za željenu procjenu radijusa:

$$r_{new} = 3.1209\sqrt{5} \cdot \frac{std_{new}}{std_{old}}$$

Problem središta

Drugi problem s kojim se susrećemo kada ne znamo razliku između pravih i lažnih pozitivaca je pronalazak središta kugle. Prema pretpostavci, slični motivi se grupiraju u kugli procijenjenog radijusa, zato želimo naći najgušću kuglu istog radijusa, a za moguća središta ćemo uzeti sve motive, odnosno točke, iz outputa.

Metodu započinjemo procjenom radijusa kako je opisano u prošlom potpoglavlju, zatim prolazimo po svim točkama i njih tretiramo kao potencijalna središte kugle s gore spomenutim radijusom i provjeravamo koliko se točaka nalazi u toj kugli. Pamtimo onu točku koja daje kuglu s najviše točaka. Još jedan problem s kojim smo se susreli u ovom koraku je pronalazak najgušće kugle koja ne sadrži ni jednog pravog pozitivca. To je moguće jer se u outputu mogu nalaziti slični motivi koji nisu slični našem upitu. Zato od algoritma za traženje najgušće kugle zahtijevamo da se upit nalazi u samoj kugli.

Odabir jedne točke outputa kao središte najgušće kugle daje kuglu s potencijalno neravnomjerno raspršenim točkama. Taj problem ćemo riješiti pomicanjem središta bez promijene radijusa. Kao novo središte ćemo postaviti težište svih točaka u kugli, zatim ponovo izbrojati točke u kugli jer je moguće da će u novu kuglu upasti neke nove ili će biti izbačene stare točke. Ponavljamo postupak do kada algoritam ne počinje izbacivati isto težište.

3.5 Rezultati i analiza

U pet različita proteoma smo pronalazili slične motive upitu i ispitali uspješnost razvijenog modela, a to su:

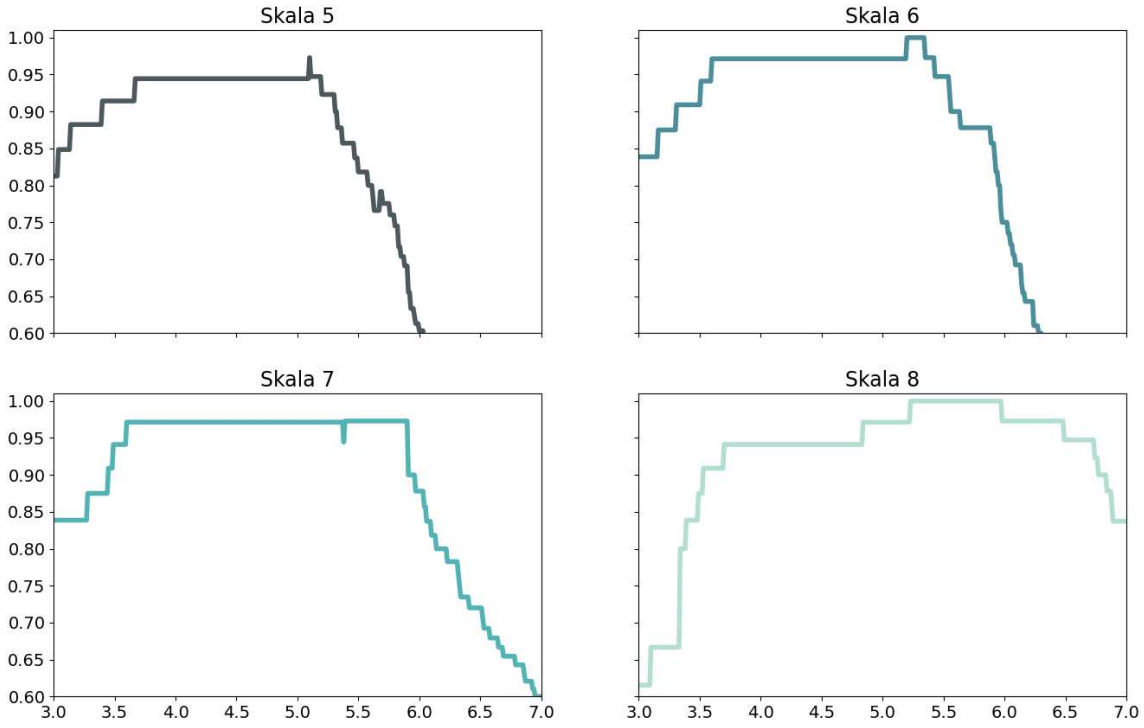
- Obična blitva (lat. *Beta Vulgaris*)
- Azijska riža (lat. *Oryza sativa*)
- Krumpir (lat. *Solanum tuberosum*)
- Rajčica (lat. *Solanum lycopersicum*)
- Talijin uročnjak (lat. *Arabidopsis thaliana*)

Bilježimo koliko dobar rezultat smo dobili pomoću modela i uspoređujemo rezultate s *benchmarkom* izračunatim na način opisan u poglavlju 3.3.

U nastavku su prikazane tablice s dobivenim rezultatima za svih pet proteoma. Ispisujemo skalu, radijus, broj potencijalno sličnih motiva, n, broj pravih pozitivaca nađenih u IGLOSS outputu, TP, broj motiva unutar najgušće kugle, n_{ball}, broj pravih pozitivaca unutar kugle, n_{TP}, i F_1-score.

	Skala	Radijus	n	TP	n_{ball}	n_{TP}	F_1-score
Benchmark	5	5.10	443	19	18	18	0.9730
Model	5	5.05	443	19	19	17	0.8947
Benchmark	6	5.20	222	18	18	18	1.000
Model	6	5.20	222	18	19	17	0.9189
Benchmark	7	5.39	138	18	19	18	0.9730
Model	7	5.27	138	18	18	17	0.9444
Benchmark	8	5.23	47	18	18	18	1.000
Model	8	5.74	47	18	18	18	1.000

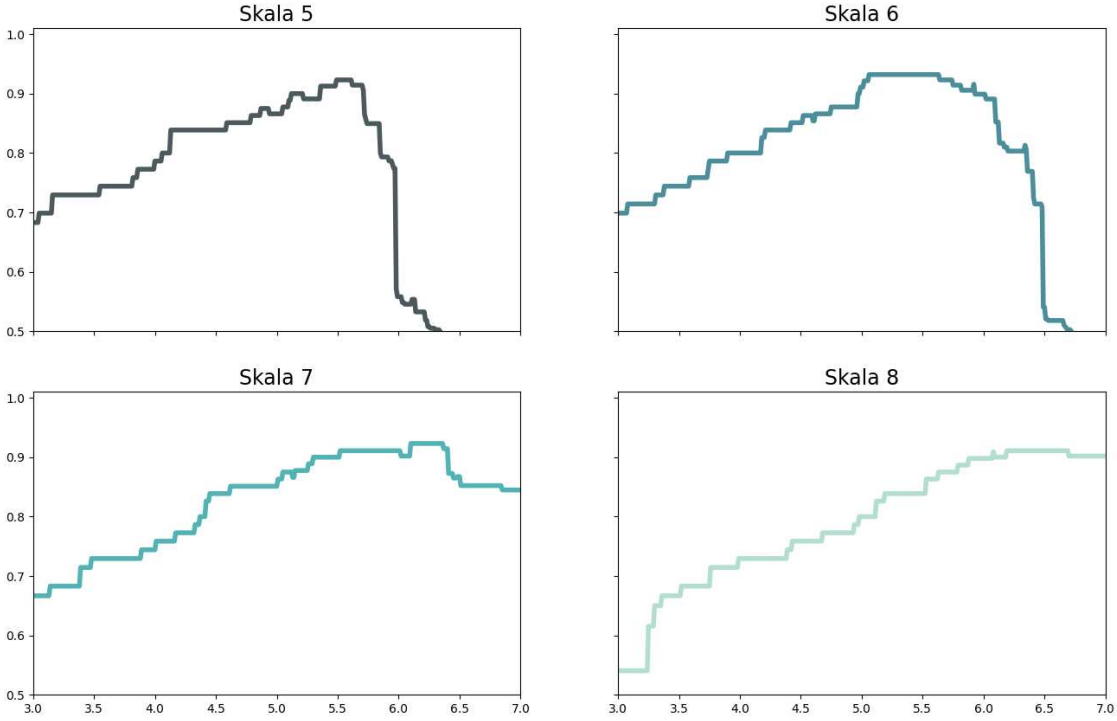
Tablica 3.2: Obična blitva



Slika 3.1: Ponašanje F_1-scora s obzirom na radijus u proteomu obične blitve sa različitim skalama

	Skala	Radijus	n	TP	n_{ball}	n_{TP}	F_1-score
Benchmark	5	5.49	1356	54	50	48	0.9231
Model	5	5.18	1356	54	43	42	0.8660
Benchmark	6	5.06	594	54	49	48	0.9320
Model	6	5.11	594	54	44	43	0.8776
Benchmark	7	6.10	374	54	50	48	0.9231
Model	7	5.27	374	54	42	42	0.8750
Benchmark	8	6.19	308	54	47	46	0.9109
Model	8	5.60	308	54	39	39	0.8387

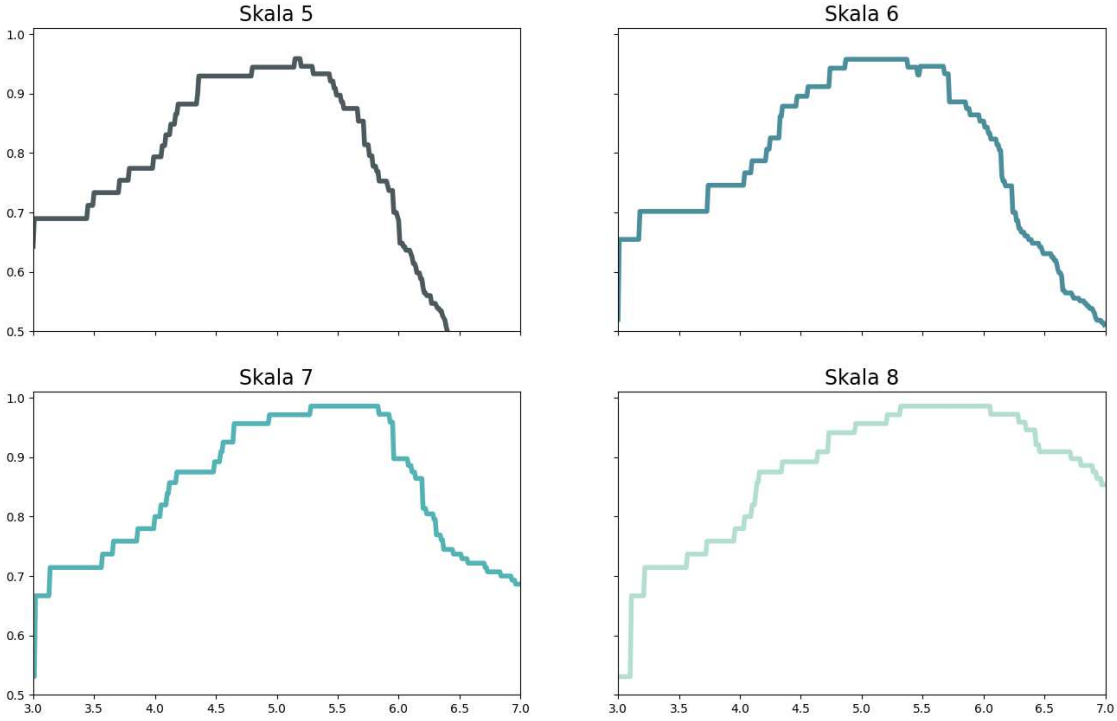
Tablica 3.3: Azijska riža



Slika 3.2: Ponašanje F_1-scora s obzirom na radijus u proteomu azijske riže sa različitim skalama

	Skala	Radijus	n	TP	n_{ball}	n_{TP}	F_1-score
Benchmark	5	5.15	466	37	36	35	0.9589
Model	5	5.11	466	37	35	34	0.9444
Benchmark	6	4.87	242	36	35	34	0.9577
Model	6	5.33	242	36	35	34	0.9577
Benchmark	7	5.28	143	35	36	35	0.9859
Model	7	5.54	143	35	36	35	0.9859
Benchmark	8	5.32	91	35	36	35	0.9859
Model	8	5.70	91	35	36	35	0.9859

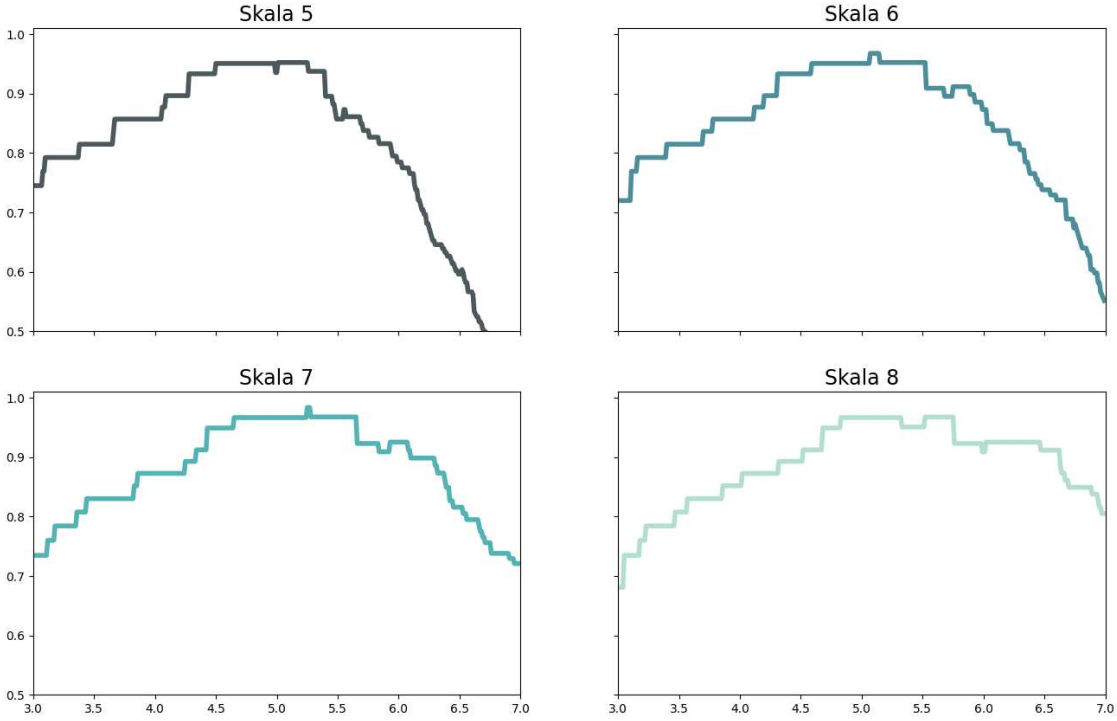
Tablica 3.4: Krumpir



Slika 3.3: Ponašanje F_1-scora s obzirom na radijus u proteomu krumpira sa različitim skalama

	Skala	Radijus	n	TP	n_{ball}	n_{TP}	F_1-score
Benchmark	5	5.01	517	32	31	30	0.9524
Model	5	5.17	517	32	31	30	0.9524
Benchmark	6	5.07	278	32	30	30	0.9677
Model	6	5.32	278	32	31	30	0.9524
Benchmark	7	5.25	162	31	30	30	0.9836
Model	7	5.50	162	31	31	30	0.9677
Benchmark	8	5.52	101	31	31	30	0.9677
Model	8	5.76	101	31	34	29	0.8923

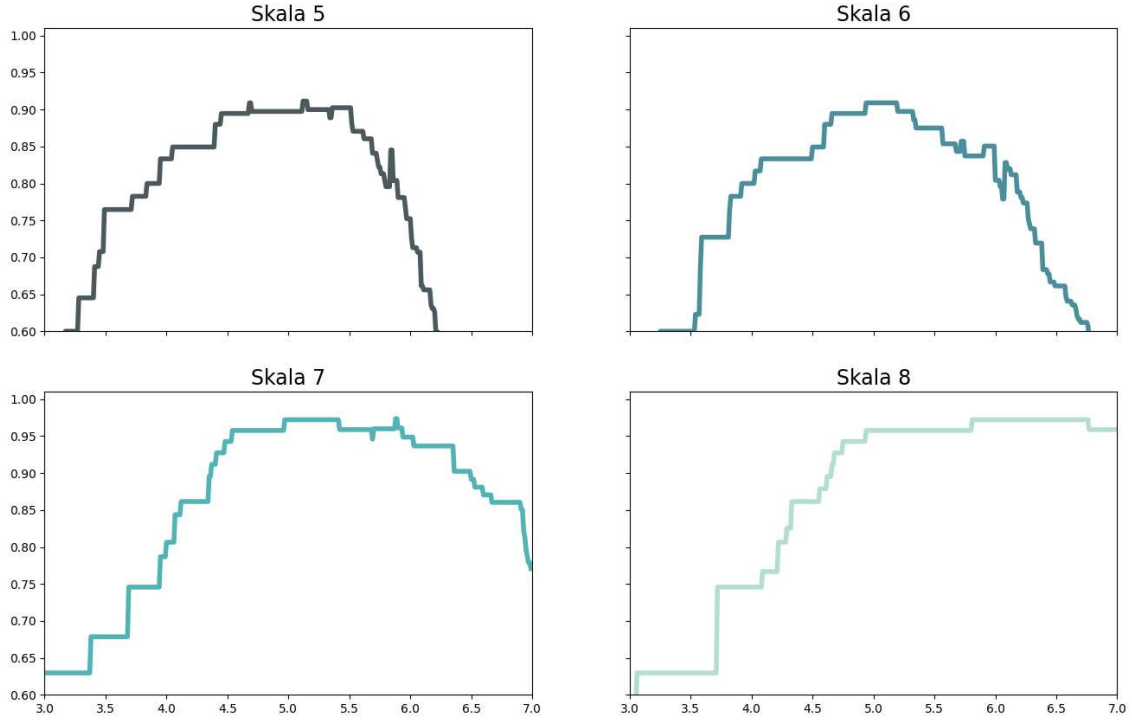
Tablica 3.5: Rajčica



Slika 3.4: Ponašanje F_1-scora s obzirom na radijus u proteomu rajčice sa različitim skalama

	Skala	Radijus	n	TP	n_{ball}	n_{TP}	F_1-score
Benchmark	5	5.12	509	42	37	36	0.9114
Model	5	5.00	509	42	37	35	0.8861
Benchmark	6	4.94	236	42	35	35	0.9091
Model	6	5.22	236	42	35	35	0.9091
Benchmark	7	5.88	129	37	39	37	0.9737
Model	7	5.50	129	37	36	35	0.9589
Benchmark	8	5.81	73	37	35	35	0.9722
Model	8	5.89	73	37	34	34	0.9577

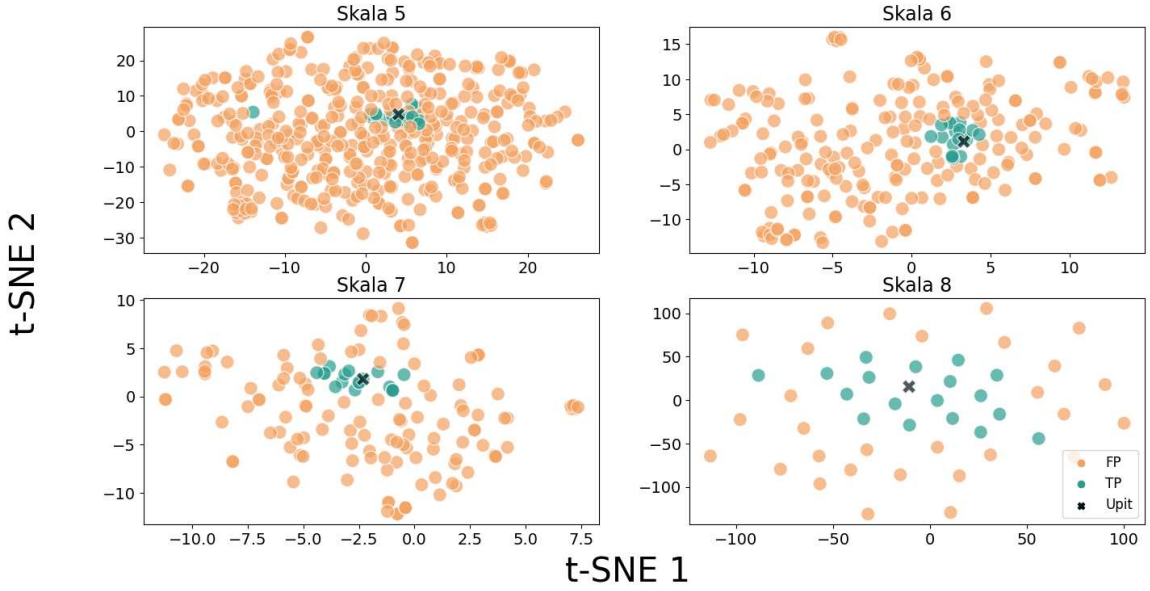
Tablica 3.6: Talijin uročnjak



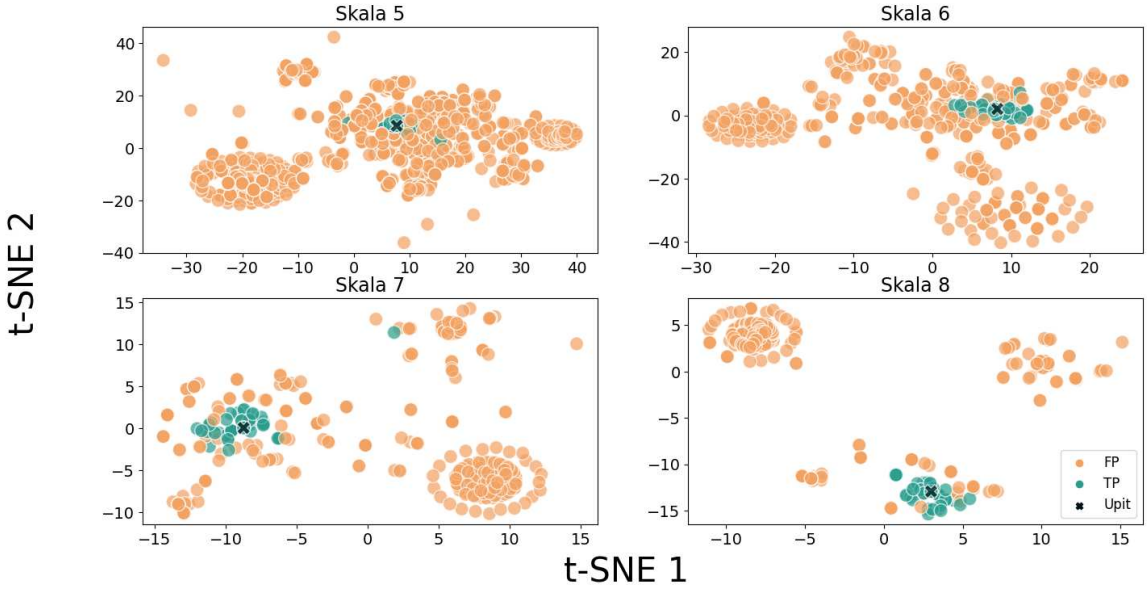
Slika 3.5: Ponašanje F_1-scora s obzirom na radijus u proteomu Talijinog uročnjak sa različitim skalama

t–SNE

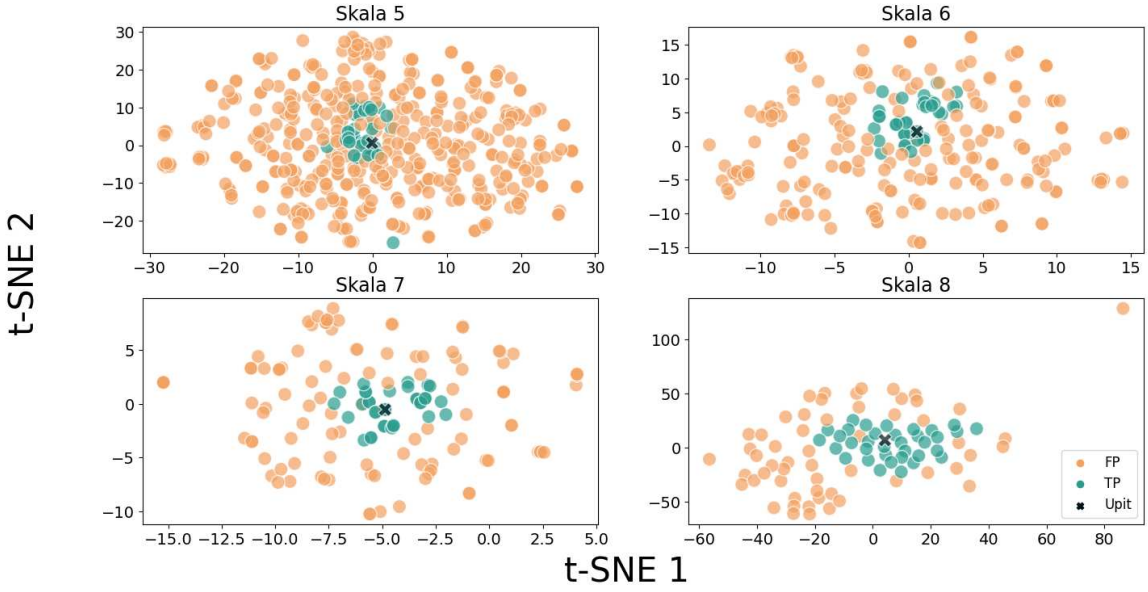
Vektorski prostor u kojem razvijamo model je $\mathbb{R}^{50}$, no zbog dimenzije ne možemo vizualizirati podatke i tako provjeriti ima li naša pretpostavka smisla. U tu svrhu, kao pomoć u vizualizaciji visoko - dimenzionalnih podataka koristili smo Pythonov alat *sklearn.manifold.TSNE* koji omogućava vizualizaciju u dvije ili tri dimenzije. Način rada paketa je izvan okvira ovog rada, no opisan je u članku [9].



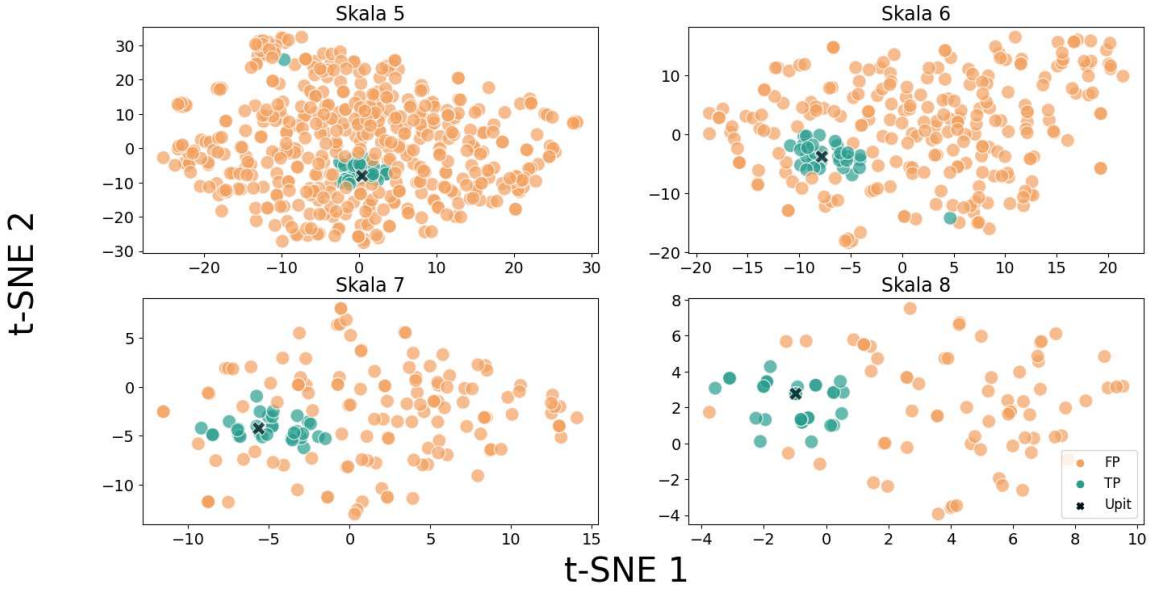
Slika 3.6: Prikaz pravih i lažnih pozitivaca obične blitve i upita u 2D pomoću *t–SNE* paketa



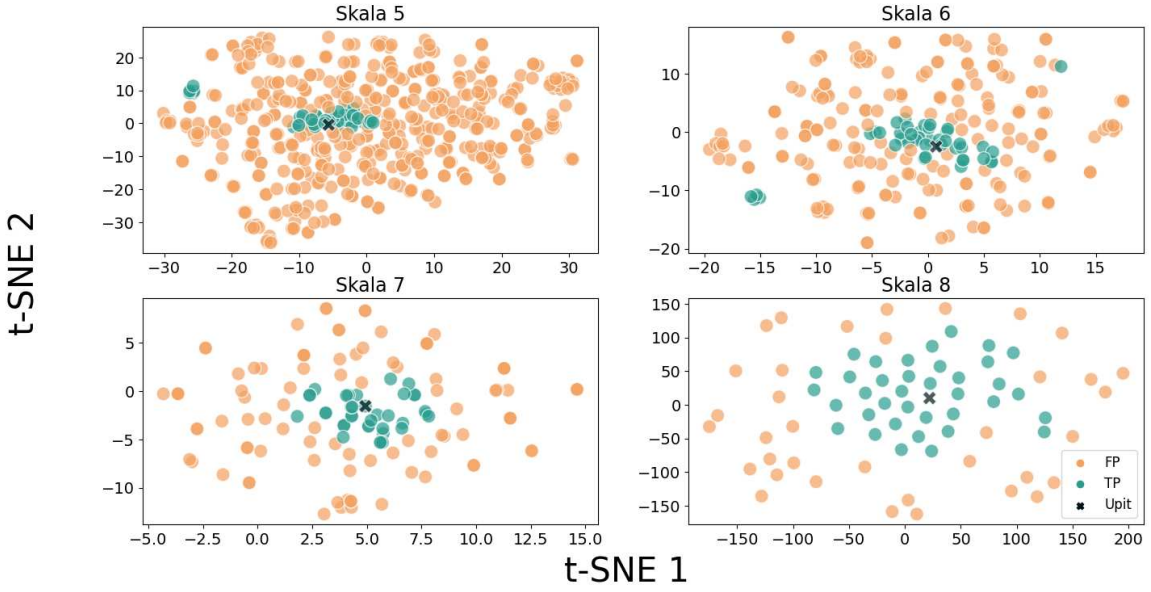
Slika 3.7: Prikaz pravih i lažnih pozitivaca riže i upita u 2D pomoću *t–SNE* paketa



Slika 3.8: Prikaz pravih i lažnih pozitivaca krumpira i upita u 2D pomoću *t–SNE* paketa



Slika 3.9: Prikaz pravih i lažnih pozitivaca rajčice i upita u 2D pomoću *t–SNE* paketa



Slika 3.10: Prikaz pravih i lažnih pozitivaca uročnjaka i upita u 2D pomoću *t–SNE* paketa

Analiza

Optimalni F_1-score je u svim proteomima i na svim skalama veći od 0.9. Vrijedi primijetiti da kod obične blitve taj optimalni score dosegne i 1.0 za skalu 6 i 8, što znači da smo u tim slučajevima uspješno pronašli najgušću kuglu koja sadrži isključivo prave pozitivce koje je IGLOSS dao u outputu.

Kada uspoređujemo pripadne F_1-scorove koje smo dobili modeliranjem, vidimo da u većini slučajeva oni ne odstupaju više od 0.06 ili 6% od benchmarka. U nekoliko slučajeva dolazi do većeg odstupanja, na primjer, kod obične blitve sa skalom 5, 3.2, vidimo razliku od skoro 9%, no skup pravih pozitivaca je mali, tako da mala promjena ima veliki utjecaj na rezultat. U tom slučaju, na prvom grafu 3.1 postoji vrh. To upućuje na osjetljivost F_1-scora s obzirom na radijus, drugim riječima, čak i da *pogodimo* središte naše kugle, malo odstupanje od optimalnog radijusa pridonosi naglom padu scora. Ta osjetljivost je izražena i kod rajčice na skali 8 na grafu 3.4. Procijenjeni radijus je u tom slučaju veći od optimalnog, a s grafa se vidi nagli pad F_1-scora oko te vrijednosti.

Smanjenjem skale dolazi do naglog pada točnosti IGLOSS outputa jer dolazi do velikog povećanja ukupnog broja potencijalnih sličnih motiva dok broj pravih sličnih motiva ostaje gotovo isti. S druge strane, F_1-score modela razvijenog u ovom radu ne pada nužno sa smanjenjem skale, što ukazuje na to da je metoda robusna. Isto tako, povećanjem skale u svih pet slučajeva dolazi do sporijeg pada F_1-scora s obzirom na iterirani radijus. To je očekivano zato što se broj kandidata smanjuje.
Rezultati za Azijsku rižu su nešto lošiji od ostalih proteoma. Najgušća kugla koju smo dobili modeliranjem sadrži oko 77% IGLOSS-ovih pozitivaca, no sadrži maksimalno jedan lažni pozitivac. Procijenjeni radijus je na svim skalama, osim na skali 6, manji od optimalnog i modelirane kugle sadrže uočljivo manje točaka što povlači to da velik dio pravih pozitivaca nije uključen.

Pomoću *t-SNE* paketa smo vizualizirali podatke u dvodimenzionalnom prostoru. Bitno je napomenuti da je ovo samo pomoćni alat što znači da samo na temelju njega ne možemo donositi zaključke, no može biti od velike koristi u razumijevanju rezultata. Na tim prikazima imamo nekoliko bitnih stvari za uočiti. Jedna od njih je kod obične blitve na skali 5. Čak ni u optimalnom slučaju nismo uspjeli uključiti jedan motiv u kuglu, a na prvom grafu 3.6 vidimo jednog pravog pozitivca koji je izrazito udaljen od ostatka. Isto to vidimo na nižim skalama ostalih proteoma. Jedna zanimljiva pojava koju vidimo iz ovih vizuala je kod riže na svim skalama na slici 3.7. Postoji očita pravilna grupacija (na nižim skalama vidimo čak i dvije) koja ne sadrži ni jednog pravog pozitivca. Upravo iz ovog razloga smo zahtijevali da kod pronalaženja najgušće kugle upit bude sadržan u samoj kugli. U suprotnome smo u nekim slučajevima mogli dobiti da najgušća kugla bude kugla koja sadrži motive koji nisu slični upitu, no to ne znači da oni nisu međusobno slični.

Ono što smo dobili razvojem ovog modela je klasifikacija sličnih motiva na temelju međusobne usporedbe neovisno o odabranoj skali. Za sve spomenute skale smo značajno smanjili broj lažnih pozitivaca bez da smo izbacili puno pravih pozitivaca. Metoda koja je implementirana je brza, robusna i optimizacijom smo smanjili količinu ručne provjere rezultata.

Bibliografija

[1] W. R. Atchley, J. Zhao, A. D. Fernandes i T. Druke, *Solving the protein sequence metric problem.*, (2005), https://www.pnas.org/doi/full/10.1073/pnas.0408677102.

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

[3] T. Duričić i A. Merćep, *Strojno učenje - bilješke s predavanja*, 2015./2026.

[4] M. Iveković, *Traženje proteinskih motiva i klasifikacija*, (2022).

[5] Maurice George Kendall i Patrick Alfred Pierce Moran, *Geometrical probability*, Hafner Publishing Company, 1963.

[6] S. Mardešić, *Matematička analiza u n-dimenzionalnom realnom prostoru*, Školska knjiga, 1974.

[7] B. Rabar, M. Zagorščak, S. Ristov, M. Rosenzweig i P. Goldestein, *Convex functions*, Bioinformatics **35**, 3491–3492, https://academic.oup.com/bioinformatics/article/35/18/3491/5306940.

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

[9] Laurens van der Maaten i Geoffrey Hinton, *Visualizing Data using t-SNE*, Journal of Machine Learning Research 9 (2008), https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf.

[10] I. Višek, *Clustering i klasifikacija proteinskih nizova*, (2022).

[11] T. Yadati, T. Houben, A. Bitorina i R. Shiri-Sverdlov, *The Ins and Outs of Cathepsins: Physiological Function and Role in Disease Management*, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7407943/.

Sažetak

Tematika ovog rada je klasifikacija proteina u proteinsku familiju katepsina na temelju karakterističnih podnizova aminokiselina (motiva) i međusobnog uspoređivnja. Cilj je izbaciti što više lažno pozitivnih i pri tome zadržati stvarno pozitivne motive iz outputa pretraživača motiva.

Problematika rada svodi se na prelazak u višedimenzionalni vektorski prostor prevođenjem aminokiselina u numeričke vektore koji predstavljaju određena obilježja tih aminokiselina. Pronalazi se kugla koja sadrži najviše motiva, uključujući i motiv katepsina. Prvo se računa benchmark za uspješnost modela, zatim se razvija model procjenjivanjem radijusa kugle i prolaskom po svim točkama za određivanje njezinog središta. Središte se pomiče u težište točaka u kugli i taj postupak se ponavlja do kada se središte ne stabilizira. Na kraju se model validira i uspoređuje s benchmarkom.

U pet proteoma napravljena je klasifikacija katepsina. Rezultati pokazuju da model uspješno i brzo izbacuje lažne pozitivce, a da pritom zadrži većinu pravih pozitivaca neovisno o veličini početnog uzorka.

Summary

The subject of this thesis is classification of proteins into the Cathepsin protein family based on characteristic sequences of amino acids (motifs) and pairwise comparison. The goal is to reduce the number of false positives and keep true positive motifs from the output of a motif scanner.

The problem comes down to transitioning into multidimensional vector space by interpreting amino acids as numeric vectors which represent certain characteristics of those amino acids. A ball containing most motifs is found, including the Cathepsin motif. Firstly, a benchmark for the performance of the model is calculated, then the model is being developed by estimating the expected radius and by optimizing the center of the ball. At the end, the model is being validated and the performance is compared to the benchmark.

The classification of Cathepsins is made in five proteoms. The results show that the fast model successfully decreases the number of false positive motifs but manages to keep true positives regardless of the size of the initial sample.

Životopis

Rođena sam 3. lipnja 1997. godine u Varaždinu. Obrazovanje započinjem u Osnovnoj školi Ivana Kukuljevića Sakcinskog u Ivancu te 2012. upisujem XV. gimnaziju u Zagrebu. Nakon završene prirodoslovno–matematičke gimnazije, 2016. godine nastavljam školovanje na preddiplomskom studiju Matematike na Prirodoslovno–matematičkom fakultetu u Zagrebu, a nakon toga upisujem diplomski studij matematičke statistike na istom fakultetu.