Proteinski motivi i klasifikacija

Bokšić, Vinko

Proteinski motivi i klasifikacija

Bokšić, Vinko

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Vinko Bokšić

PROTEINSKI MOTIVI I KLASIFIKACIJA

Diplomski rad

Voditelj rada:
doc. dr. sc. Pavle Goldstein

Zagreb, srpanj 2021.

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 velikoj pomoći pri pisanju ovog rada, na uloženom vremenu, strpljenju i trudu.
Veliko hvala mojoj majci Deani na bezuvjetnoj podršci tijekom cijelog školovanja, koja je u meni probudila zanimanje za matematikom.
Veliko hvala ocu Zoranu koji mi je omogućio sve što mi je potrebno u životu i puno više.
Hvala bratu Miru, sestri Marieti, rođaku Anti i svim prijateljima koji su bili uz mene.

Sadržaj

Uvod

Proteom je skup svih proteina nekog organizma. Proteini su velike, složene molekule, sastavljene od aminokiselina, koje su sastavni dio stanice svih živih bića. Izvršavaju širok spektar funkcija unutar organizma, među kojima su kataliziranje metaboličkih reakcija, replikacija DNA i reagiranje na podražaje. Upravo njihova povezanost s osnovnim životnim svojstvima jedinki (sposobnost apsorpcije kisika, otpornost na sušu, ...) razlog je proučavanja i određivanja pripadnosti proteinskim familijama. Mogućnost uzgoja žitarica u bočatoj vodi riješila bi veliki svjetski problem gladi.

U proteinskoj familiji nalaze se proteini koji imaju zajedničko evolucijsko podrijetlo i koji su zaslužni za isto svojstvo. Problem traženja proteina koji pripadaju istoj proteinskoj familiji jedno je od glavnih zanimanja Bioinformatike. Zbog sve detaljnijih i opširnijih podataka o proteinima dobivenih sekvenciranjem genoma postoji potreba za pouzdanim, računarskim metodama za klasifikaciju proteina. Iterativna metoda pretraživanja proteoma, kojom se zadavanjem karakterističnog niza aminokiselina pronalazi skup proteina koji pripadaju istoj proteinskoj familiji, standardna je metoda klasifikacije. Metoda se bazira na konceptu sličnosti, pa je uspješnost ograničena.

U ovom radu istražuje se algoritam koji bi unaprijedio model iterativnog pretraživanja, povećavajući njegovu uspješnost. Nakon što iterativno pretraživanje da svoje kandidate za traženu proteinsku familiju, primjenjuje se dodatni filter koji reducira broj podataka na temelju geometrijske strukturiranosti i bliskosti proteina koji sadržavaju zajedničko svojstvo od interesa.

Ovaj rad sastoji se od tri poglavlja. U prvom poglavlju navedeni su pojmovi iz linearne algebre te vjerojatnosti i statistike koji su nužni za daljnje razumijevanje rada. Također su definirane i mjere uspješnosti. Drugo poglavlje bavi se biološkom pozadinom problema, iterativnim pretraživanje proteoma te prelaskom u vektorski prostor. Konačno, u trećem poglavlju opisuje se ideja i algoritam koji pospješuje iterativni model te su prikazani grafički i numerički rezultati istraživanja.

Poglavlje 1

Matematički pojmovi

U ovom poglavlju navode se teoremi, definicije, propozicije i napomene iz linearne algebre, vjerojatnosti i statistike te uspješnosti modela. Pojmovi su preuzeti iz izvora [2], [3], [4], [7] i [9].

1.1 Linearna algebra

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

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

2) postoji $0 \in \mathbb{F}$ *sa svojstvom* $\alpha + 0 = 0 + \alpha = \alpha, \ \forall \alpha \in \mathbb{F}$*;*

3) za svaki $\alpha \in \mathbb{F}$*, postoji* $-\alpha \in \mathbb{F}$ *tako da je* $\alpha + (-\alpha) = (-\alpha) + \alpha = 0$*;*

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

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

6) postoji $1 \in \mathbb{F} \setminus \{0\}$ *sa svojstvom* $1 \cdot \alpha = \alpha \cdot 1 = \alpha, \ \forall \alpha \in \mathbb{F}$*;*

7) za svaki $\alpha \in \mathbb{F}, \ \alpha \neq 0, \ postoji \ \alpha^{-1} \in \mathbb{F}$ *tako da je* $\alpha\alpha^{-1} = \alpha^{-1}\alpha = 1$*;*

8) $\alpha\beta = \beta\alpha, \ \forall \alpha, \beta \in \mathbb{F}$*;*

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

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

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

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

1) $a + (b + c) = (a + b) + c, \ \forall a, b, c \in V$;

2) postoji $0 \in V$ *sa svojstvom* $a + 0 = 0 + a = a, \ \forall a \in V$;

3) za svaki $a \in V$, *postoji* $-a \in V$ *tako da je* $a + (-a) = (-a) + a = 0$;

4) $a + b = b + a, \ \forall a, b \in V$;

5) $\alpha(\beta a) = (\alpha\beta)a, \ \forall \alpha, \beta \in \mathbb{F}, \forall a \in V$;

6) $(\alpha + \beta)a = \alpha a + \beta a, \ \forall \alpha, \beta \in \mathbb{F}, \forall a \in V$;

7) $\alpha(a + b) = \alpha a + \alpha b, \ \forall \alpha \in \mathbb{F}, \forall a, b \in V$;

8) $1 \cdot a = a \cdot 1, \ \forall a \in V$.

Definicija 1.1.4. *Za prirodne brojeve m i n, preslikavanje*

$$A : \{1, 2, \ldots, m\} \times \{1, 2, \ldots, n\} \to \mathbb{F}$$

naziva se ***matrica*** *tipa (m, n) s koeficijentima iz polja $\mathbb{F}$.*

Definicija 1.1.5. *Neka je V vektorski prostor nad poljem $\mathbb{F}$.* ***Skalarni produkt*** *na V je preslikavanje $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{F}$ koje ima sljedeća svojstva:*

1) $\langle x, x \rangle \geq 0, \ \forall x \in V$;

2) $\langle x, x \rangle = 0 \Leftrightarrow x = 0$;

3) $\langle x_1 + x_2, y \rangle = \langle x_1, y \rangle + \langle x_2, y \rangle, \ \forall x_1, x_2, y \in V$;

4) $\langle \alpha x, y \rangle = \alpha \langle x, y \rangle, \ \forall \alpha \in \mathbb{F}, \forall x, y \in V$;

5) $\langle x, y \rangle = \overline{\langle y, x \rangle}, \ \forall x, y \in V$.

Napomena 1.1.6. *U $\mathbb{R}^n$ kanonski skalarni produkt definiran je s*

$$\langle (x_1, \ldots, x_n), (y_1, \ldots, y_n) \rangle = \sum_{i=1}^{n} x_i y_i.$$

Definicija 1.1.7. *Vektorski prostor na kojem je definiran skalarni produkt zove se* ***unitarni prostor****.*

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

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

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

1) $\|x\| \geq 0, \ \forall x \in V;$

2) $\|x\| = 0 \Leftrightarrow x = 0;$

3) $\|\alpha x\| = |\alpha| \, \|x\|, \ \forall \alpha \in \mathbb{F}, \forall x \in V;$

4) $\|x + y\| \leq \|x\| + \|y\|, \ \forall x, y \in V.$

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

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

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

Ova norma zove se ***Euklidska norma****.*

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

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

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

1) $d(x, y) \geq 0, \ \forall x, y \in V;$

2) $d(x, y) = 0 \Leftrightarrow x = y, \ \forall x, y \in V;$

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

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

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

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

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

Ova metrika naziva se ***Euklidska metrika****, a prostor* $\mathbb{R}^n$ *zajedno s tom metrikom nazivamo* ***Euklidski prostor****.*

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

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

nazivamo ***otvorena kugla*** *u X, 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 Vjerojatnost i statistika

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 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 napokon 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*

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

Definicija 1.2.17. *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 1.2.18. *Pozitivan drugi korijen iz varijance nazivamo **standardna devijacija** i označavamo sa σ_X.*

Funkcija distribucije

Definicija 1.2.19. *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 1.2.20. *Ako je jasno o kojoj se slučajnoj varijabli radi, piše se F umjesto F_X.*

Teorem 1.2.21. *Funkcija distribucije F slučajne varijable X je rastuća i neprekidna zdesna na $\mathbb{R}$, te zadovoljava:*

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

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

Definicija 1.2.22. *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 1.2.23. *Neka je X slučajna varijabla na vjerojatnosnom prostoru $(\Omega, \mathcal{F}, \mathbb{P})$ i neka je F_X njezina funkcija distribucije. Kažemo da je X **apsolutno neprekidna** ili kraće, **neprekidna slučajna varijabla** ako postoji nenegativna realna Borelova funkcija f na $\mathbb{R}$ ($f : \mathbb{R} \to \mathbb{R}_+$) takva da je*

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

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

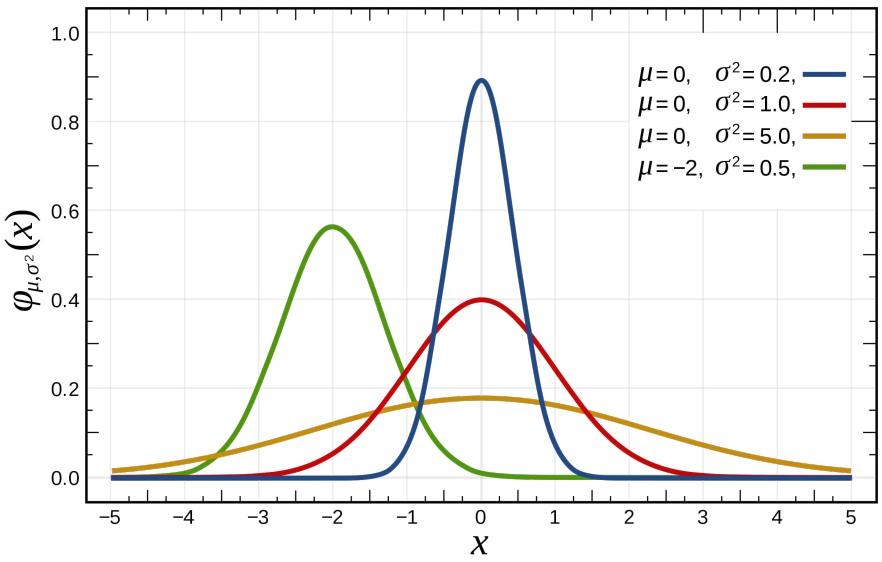
Definicija 1.2.24. *Neka su $\mu, \sigma \in \mathbb{R}$, $\sigma > 0$. Neprekidna slučajna varijabla X ima **normalnu distribuciju s parametrima** μ i σ^2 ako joj je gustoća f dana s*

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

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

Napomena 1.2.25. *X je **jedinična normalna distribucija** ako je $X \sim N(0, 1)$, dakle*

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



Slika 1.1: Funkcija gustoće normalne distribucije za različite parametre

Opisna statistika

Za razumijevanje ovog rada još će bit potrebno znanje o radu s podacima (mjerenjima). Navodimo pojmove kao što su aritmetička sredina, standardna devijacija uzorka i varijanca uzorka i standardizacija podataka.

Neka su

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

n vrijednosti (opažanja) varijable X koje čine skup podataka. Ako je X numerička varijabla, tada je to niz brojeva. Neka je u nastavku X numerička varijabla.

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

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

Varijanca uzorka ili podataka (1.3) je mjera raspršenja podataka i predstavlja prosječno kvadratno odstupanje podataka od njihove aritmetičke sredine i dana je formulom:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

Iz prethodnih definicija slijedi da je **standardna devijacija uzorka** drugi korijen varijance i zadana je formulom:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

Standardizacija podataka je česta procedura u statistici prije obrade podataka i izgradnje modela ili algoritma. Podaci se transformiraju oduzimanjem očekivanja i dijeljenjem sa standardnom devijacijom uzorka:

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

Rezultat nam govori koliko je standardnih devijacija pojedini podatak pomaknut od aritmetičke sredine uzorka. Procedura se provodi kako bi se izbjegao nejednolik raspon i raspršenje među podacima, što može dovesti do toga da model daje više značajnosti varijablama koje imaju veći raspon. To dovodi do potpuno krivih zaključaka kod algoritama koji koriste udaljenost među podacima. Poslije transformacije, svi novi podaci su normalno distribuirani s očekivanjem 0 i varijancom 1.

1.3 Klasifikacija i uspješnost modela

Klasifikacija

U statistici, **klasifikacija** je problem određivanja pripadnosti opservacije nekoj od skupa kategorija (klasa). Postoje nadzirana i nenadzirana klasifikacija. U nadziranoj klasifikaciji pridruživanje opservacije određenoj klasi temelji se na skupu poznatih podataka koji sadrže opservacije kojima je klasa već određena ili unaprijed poznata. Uz pomoć zadane funkcije sličnosti, opservacija se svrstava u klasu čiji su joj elementi najsličniji. U nenadziranoj klasifikaciji model pokušava, bez prijašnjeg znanja o podacima i klasama, uočiti strukturiranost među opservacijama i separirati ih u kategorije. Osnovna razlika između ova dva tipa jest što nadzirana klasifikacija zahtijeva poznate, unaprijed označene podatke.

Mjere uspješnosti

Da bi se ocijenila uspješnost nekog modela, definirane su mjere uspješnosti modela. One se temelje na pojmovima iz matrice uspješnosti (eng. *confusion matrix*) prikazanoj sljedećom tablicom.

		Predviđeno stanje		
	Ukupna populacija	Ocijenjeni pozitivno (P)	Ocijenjeni negativno (N)	
Stvarno stanje	Pozitivno stanje (CP)	TP (stvarno pozitivni)	FN (lažno negativni)	Osjetljivost (TPR)
	Negativno stanje (CN)	FP (lažno pozitivni)	TN (stvarno negativni)	Specifičnost (TNR)
		Preciznost (PPV)	Negativna prediktivna vrijednost (NPV)	

Tablica 1.1: Tablica uspješnosti

Napomena 1.3.1. *U ovom radu će se provjera broja TP (eng. True Positives) i ostalih brojeva iz matrice uspješnosti (FP, FN, TN) vršiti na temelju liste CP (eng. Condition Positive). Lista CP sadrži sve proteine za koje je pripadnost određenoj familiji već utvrđena, biološki poznata. Dakle, u savršenom modelu bi svi proteini sa liste CP imali oznaku 1, a svi proteini koji nisu na listi CP bi imali oznaku 0.*

Slijede definicije nekih od mjera uspješnosti modela za binarnu klasifikaciju:
Osjetljivost ili **TPR** (eng. *True Positive Rate*) je postotak pozitivnih elemenata uzorka u odnosu na određeno stanje, odnosno CP elemenata uzorka, koji su ispravno prepoznati kao pozitivni.

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

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

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

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

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

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

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

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

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

U ovom radu, kao mjera uspješnosti modela koristit će se F_1**-score** ($\beta = 1$):

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

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

Poglavlje 2

Bioinformatika

2.1 Biološki pojmovi

Proteini ili bjelančevine su, uz vodu, najvažnije tvari u tijelu, stoga čine osnovu života na zemlji. Izgrađeni su od aminokiselina, nanizanih u lance, koje su međusobno povezane peptidnom vezom. Aminokiseline su molekule koje sadrže amino skupinu, karboksilnu skupinu i bočni lanac po kojem se međusobno razlikuju. Postoji 20 standardnih aminokiselina koje izgrađuju proteine, svaka označena velikim slovom Engleske abecede, prikazane u tablici 2.1. Duljina lanca i raspored određuju svojstva proteina, a promjenom samo jedne karike u lancu nastat će nova bjelančevina, potpuno novih osobina.

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

Proteom je skup svih proteina nekog organizma. U njemu se nalaze različite proteinske familije, od kojih je svaka zaslužna za određeno funkcijsko svojstvo organizma. Važnost određivanja pripadnosti proteina proteinskoj familiji jest razumijevanje uloge proteina. Kada bi to uspjeli, znali bi više o svojstvima jedinki, a što je bitnije, otvara se mogućnost unaprjeđenja određene vrste putem genetske modifikacije. Saznanje koji proteini doprinose rastu biljaka u bočatoj vodi, može rezultirati novom vrstom žitarica kojoj je pogodno i bočato tlo.

U ovom radu promatrat će se GDSL lipaze. **GDSL lipaze** jedan su od primjera lipaza, enzima koji sudjeluju kao katalizatori u razgradnji lipida (masti). Njihova posebnost je što imaju fleksibilno katalitičko mjesto koje mijenja svoj raspored u prisutnosti različitih supstrata. GDSL lipaze nađene su u biljkama, životinjama i bakterijama. Razumijevanje biljnih GDSL enzima je, zasad, vrlo ograničeno, a upravo bi biljke mogle biti bogat izvor obećavajućih enzima. Njihova katalitička multifunkcionalnost mogla bi se koristiti u hidrolizi i sintezi spojeva koji su od velikog interesa u biotehnologiji (prehrambena industrija, tekstilna, kozmetička). Stoga je otkrivanje novih biljnih GDSL lipaza iznimno važno.

2.2 Iterativno pretraživanje proteoma

Iterativno pretraživanje proteoma standardna je metoda pronalaska proteina iz iste proteinske familije. Metoda kao ulazni parametar prima karakterističnu motiv proteinske familije. **Motiv** (ili upit) je kratak niz aminokiselina (slova), duljine od 5 do 20, koji je ostao djelomično sačuvan selekcijskim pročišćavanjem ili evolucijom. Ako protein sadrži dovoljno sličan niz, s obzirom na neku funkciju sličnosti, tada ga algoritam svrstava u pripadnu proteinsku familiju. U iterativnom pretraživanju, parametri funkcije sličnosti se mijenjaju pri svakoj iteraciji. U svakoj iteraciji promatramo proteom sa skupom proteina koji su bili dovoljno slični u prethodnoj iteraciji. Iteriranje staje kada skup proteina - odgovor, ostaje nepromijenjen ili kada se dosegne maksimalan broj iteracija. Proteini su u familije svrstani s određenom uspješnosti.

U ovom radu za iterativno pretraživanje proteoma koristi se IGLOSS server, opisan u izvoru [6]. IGLOSS server za funkciju sličnosti koristi *log likelihood ratio* (LLR) koja je ocijenjena pomoću logističke regresije. **Skala pretraživanja** je parametar koji postavlja granicu "dovoljne sličnosti". Odgovor čini skup proteina čija je sličnost veća ili jednaka od zadane skale pretraživanja. Ako je skala veća, više se kažnjava odstupanje od motiva, pa su sličniji nizovi odabrani. Kao posljedica slijedi da je broj podataka u odgovoru obrnuto proporcionalan skali pretraživanja. Za kraj nam preostaje definirati BLOSUM matricu i BLOSUM score, koji se koriste za ocjenu sličnosti dvaju nizova aminokiselina.

Definicija 2.2.1. *BLOSUM matrica B je* 20×20 *matrica,* $B = (b_{ij}) \in M_{20}(\mathbb{Z})$*, koja na* (i, j)*-tom mjestu sadrži koeficijente sličnosti i-te i j-te aminokiseline. Bazirana je na sljedećoj formuli:*

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

gdje su a_i *i* b_j *aminokiseline pridružene, respektivno, i-tom i j-tom mjestu, a* $\mathcal{A}$ *je skup svih standardnih aminokiselina. M je model koji pretpostavlja da aminokiseline* a_i *i* b_j *imaju zajedničkog pretka, a R je random model koji pretpostavlja nezavisnost aminokiselina, pa vrijedi* $\mathbb{P}(a_i, b_j \mid R) = \mathbb{P}(a_i \mid R) \cdot \mathbb{P}(b_j \mid R)$*. Distribucija standardnih aminokiselina uz model R dana je sa:*

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

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

Slika 2.1: BLOSUM matrica

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

2.3 Prelazak u vektorski prostor

Nedostatak prirodne metrike za usporedbu nizova sastavljenih od slova sprječava statističku analizu i obradu nad takvim podacima. Zbog toga se javlja potreba za opisom aminokiselina numeričkim vrijednostima. Ta problematika je opisana i riješena u članku [1]. Definirano je preslikavanje u $\mathbb{R}^5$ koje svakoj aminokiselini pridružuje 5-dimenzionalni numerički vektor. Preslikavanje "čuva" sve važne fizikalno-kemijske informacije o aminokiselini. Svaka koordinata vektora (*faktor*) odgovara jednom ili kombinaciji više svojstava. *Faktor I* se odnosi na polaritet aminokiseline, *Faktor II* je faktor sekundarnog naboja, *Faktor III* je molekularni volumen, *Faktor IV* odražava raznolikost kodona (relativnu kompoziciju aminokiselina u različitim proteinima) te *Faktor V* odgovara elektrostatičkom naboju aminokiseline.

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

Tablica 2.2: Faktori

Niz od n aminokiselina sada odgovara $5n$-dimenzionalnom vektoru. Naoružani numeričkim opisom aminokiselina sada smo spremni za primjenu matematičkog alata.

Pojmovi iz ovog poglavlja preuzeti su iz izvora [1], [4], [6] i [8].

Poglavlje 3

Analiza problema i algoritam

3.1 Opis problema i ideja

Cilj ovog rada je poboljšati uspješnost iterativnog pretraživanja proteoma. To nećemo učiniti tako da mijenjamo metodu direktno, već ćemo kreirati algoritam koji će u kombinaciji s iterativnim pretraživanjem dati bolje rezultate. Nakon što nam iterativna metoda da svoje kandidate za proteinsku familiju, među njima se nalaze proteini koji zaista pripadaju toj familiji (eng. *true positives*) i oni koji po prirodi nisu u njoj (eng. *false positives*). Želimo primijeniti filter koji će iz odgovora, dobivenog putem IGLOSS servera, eliminirati što više lažnih pozitivaca, a pritom sačuvati prave pozitivce. Za to ćemo koristiti mogućnost prelaska u vektorski prostor, gdje možemo promatrati raspored nizova u prostoru, njihove udaljenosti i iskoristiti matematičko znanje. Dakle, sa stohastičkog modeliranja IGLOSS servera, koje se temelji na vjerojatnostima dobivenim pomoću biološke pozadine samih podataka, prelazimo na proučavanje geometrije među njima - sagledavamo stvari iz potpuno drugog ugla i time pokušavamo unaprijediti model.

Ako sada proteine zamislimo kao točke u višedimenzionalnom prostoru, glavna ideja jest da se pravi pozitivci nalaze zgusnuti u blizini upita (motiva), dok su lažni pozitivci razbacani i udaljeniji od upita. Pretpostavimo još da su ti pravi pozitivci smješteni u nekoj kugli u $\mathbb{R}^n$ koja sadrži upit (kasnije ćemo pokazati da je to valjana pretpostavka). Ono što nama preostaje je odrediti središte i radijus kugle u kojoj je mjera uspješnosti modela F_1, definirana u (1.5), najveća.

Priprema podataka

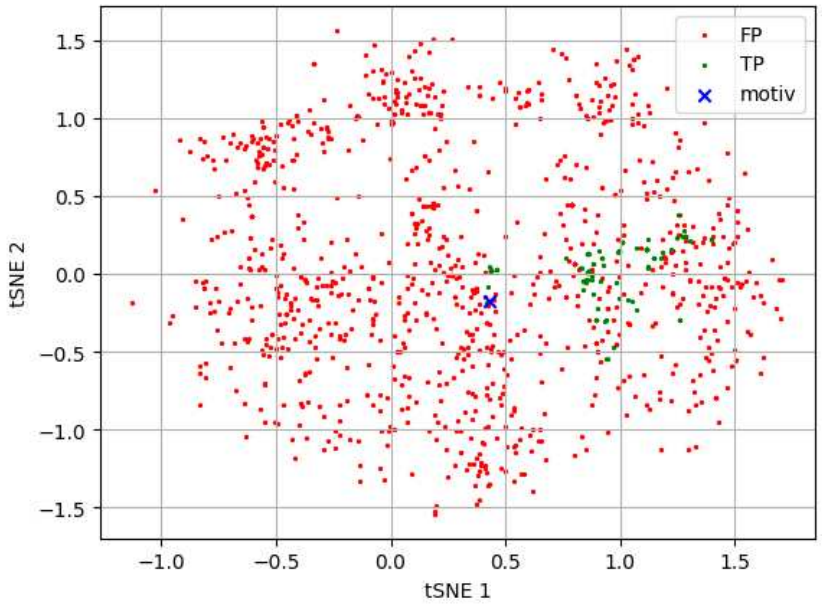
U ovom radu radit će se s upitom FVFGDSLSDA. Upit sadrži niz aminokiselina GDSL, koji je karakterističan za promatranu proteinsku familiju. Korištenjem takvog upita žele se, uz pomoć IGLOSS servera, dobiti najbolji kandidati za familiju GDSL lipaza. Kao i upit,

svi odgovori će biti nizovi aminokiselina duljine 10. Slijedi da prelaskom u vektorski prostor, naši podaci su 50-dimenzionalni vektori. S obzirom na to da promatramo euklidsku udaljenost između centra kugle i proteina, koja mjeri udaljenosti po pojedinim koordinatama i zbraja ih, želimo izbjeći da su nam varijanca i raspon podataka po jednoj koordinati veći od ostalih koordinata. Ako se to desi, euklidska udaljenost bit će dominirana tom koordinatom, jer će razlika između vrijednosti centra i proteina po toj koordinati biti veća od ostalih. Time gubimo formu kugle u kojoj bi sve koordinate trebale imati jednak utjecaj. Taj problem riješen je standardizacijom podataka, objašnjena u (1.4). Sada su podaci po svakoj koordinati normalno distribuirani s očekivanjem 0 i standardnom devijacijom 1. Za svaki protein, vrijednost po pojedinoj koordinati nam govori koliko je standardnih devijacija taj vektor udaljen od očekivanja za tu koordinatu. Konačno, dobili smo podatke gdje nam je utjecaj svih koordinata jednak i možemo tražiti idealnu kuglu bez straha da će nam jedna od koordinata izdominirati udaljenost.

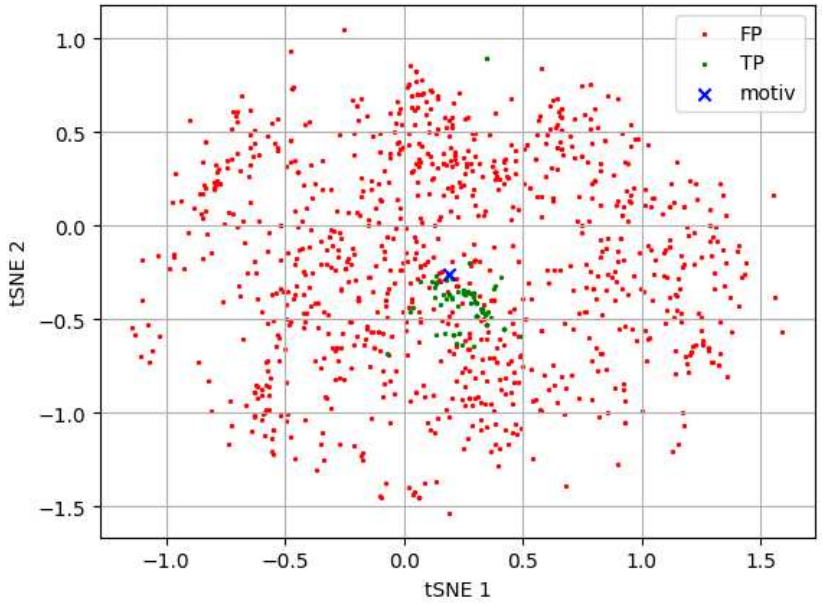
Prikaz u 2D

Prije samog definiranja algoritma želimo se uvjeriti da zaista ima smisla promatrati kuglu u $\mathbb{R}^n$. Jasno je da ne možemo crtati podatke u 50-dimezionalnom koordinatnom sustavu, koji je za nas nezamisliv, i pokušati uočiti formu. Stoga, moramo reducirati dimenziju podataka. U paketu *Scikit-Learn*, programskog jezika *Python*, postoji funkcija **t-SNE** koja provodi statističku proceduru *t-distributed Stohastic Neighbor Embedding*. t-SNE je ne-linearna metoda smanjivanja dimenzije podataka, koja čuva lokalnu strukturu. Matematička pozadina metode je kompliciranа, ali ideja je jednostavna - preslikava podatke u manje-dimenzionalan prostor čuvajući okolinu i susjedstvo točke. To je upravo ono što mi želimo postići u svrhu vizualizacije ponašanja podataka u većim dimenzijama. Ako su dva proteina blizu jedan drugom u 50-dimenzionalnom prostoru, želimo da se ta struktura translatira i u 2D. To nam omogućava uočavanje uzoraka među njima - nalaze li se zaista pravi pozitivci u kugli grupirani oko upita, te možemo li na taj način odvojiti "dobre" proteine od onih "loših". Više o pozadini same metode i njene primjene u *Python*-u možete pronaći u [5].

Sada imamo alat pomoću kojeg ćemo naše podatke preslikati u standardni dvodimenzionalni koordinatni sustav, čuvajući pritom informacije od interesa. Slijedi rezultat metode. Prikazat ćemo za normalne (početne) i standardizirane podatke kako bi mogli i grafički obrazložiti korištenje standardiziranih podataka. Analiza je provedena nad proteomom biljke Talijin uročnjak, sa skalom sličnosti 3.5. Dobiveno je 1339 nizova aminokiselina ocijenjenih da pripadaju familiji GDSL lipaza, ali samo njih 98 su stvarni, biološki pozitivci. Pravi pozitivci su označeni zelenom bojom, lažni s crvenom, a upit s plavom bojom.



Slika 3.1: t-SNE, normalni podaci



Slika 3.2: t-SNE, standardizirani podaci

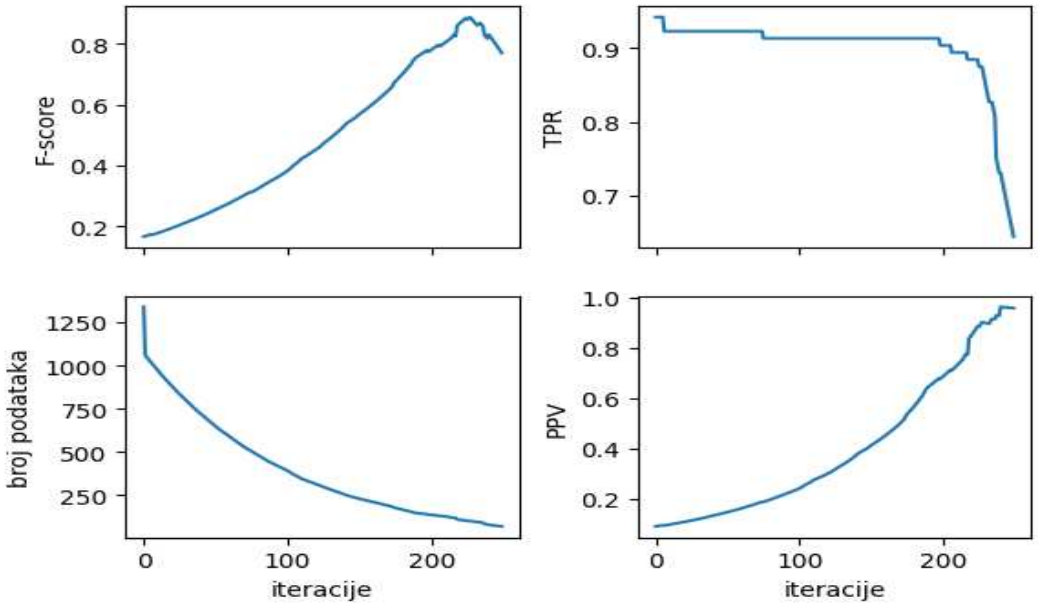
Prvo uočavamo razliku između normalnih i standardiziranih podataka. Kod standardiziranih podataka vidimo jednoličnu raspršenost svih podataka oko ishodišta, što je posljedica toga da sve koordinate imaju jednak utjecaj. Također, pravi pozitivci su gušće raspoređeni oko upita i zauzimaju manje područje. Kod normalnih podataka prisutna je veća koncentracija lažnih pozitivaca u okolini pravih te bi uspješnost algoritma koji traži kuglu koja sadrži upit bila manja jer bi se više pogrešnih proteina nalazilo u njoj. Analogni zaključci proizlaze za različite skale pretraživanja i broj podataka pa iz svega navedenog zaključujemo da ćemo se u daljnjem istraživanju koristiti isključivo standardiziranim podacima koji daju bolju mogućnost odvajanja stvarnih pozitivaca.

Drugo, iz grafa standardiziranih podataka uočavamo da zaista ima smisla promatrati kuglu koja sadrži upit. Zelene točke mogli bi vrlo lijepo opisati krugom u kojem se nalazi i motiv. Kako je slika dobivena t-SNE algoritmom koji čuva strukturu podataka slijedi da analogne zakonitosti vrijede i većim dimenzijama - biološke pozitivce možemo opisati kuglom i u izvornom, 50-dimenzionalnom prostoru. Time je ideja traženja kugle nužna.

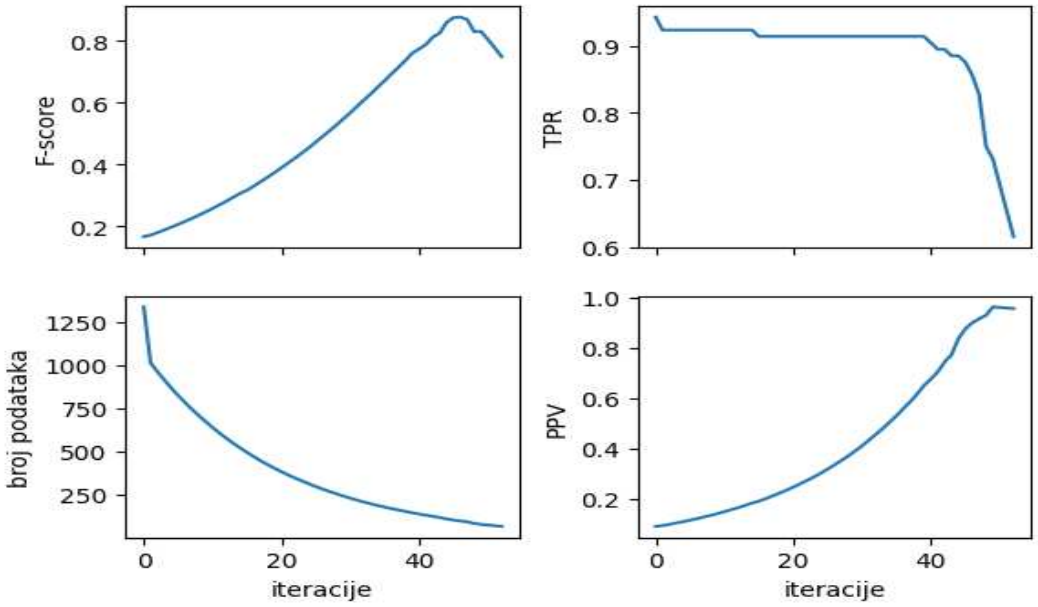
Algoritam

Slijedi i definicija algoritma. Algoritam je u cijelosti implementiran u programskom jeziku *Python*. Uočili smo kako su, za standardizirane podatke, biološki pozitivci distribuirani gusto u središtu svih podataka i u blizini upita, a da su dalje od središta raspršeni oni koje je IGLOSS pogrešno ocijenio. Dakle, za uspješnost algoritma ključno je da se eliminiraju podaci udaljeniji od centra. Vođeni tom idejom, razvijen je iterativni algoritam koji će u svakoj iteraciji odbacivati rubne elemente. Na početku jedne iteracije algoritam računa središte svih točaka kao aritmetičku sredinu svih podataka, po svakoj koordinati. Zatim se odbacuje određeni postotak najudaljenijih točaka od dobivenog središta. U sljedećoj iteraciji računa se aritmetička sredina svih preostalih točaka koja postaje novo središte, nakon čega se ponovno odbacuje postotak najudaljenijih točaka. Time se svakom iteracijom eliminiraju rubni podaci i kugla se stišće prema pravim pozitivcima. Algoritam se zaustavlja kada kugla dosegne jednak ili manji radijus od zadanog. Postotak točaka koje se izbacuju u svakoj iteraciji zadaje se kao parametar algoritma i iznosi između 1% i 20%. Veći postoci rezultiraju bržem izvođenju algoritma, ali pod cijenom da se preskoči najbolja kugla. Pokazat će se da se algoritam brzo izvodi i kada je 1% izbačenih te se taj postotak preporučuje koristiti jer je najprecizniji.

Sada ćemo grafički prikazati kako algoritam radi na jednom primjeru. Kao i prije, nad proteomom biljke Talijin uročnjak, dobiveno je 1339 nizova aminokiselina ocijenjenih da pripadaju familiji GDSL lipaza, od kojih je 98 pravih pozitivaca. Prikazujemo i uspoređujemo rezultate za dva različita postotka izbacivanja - 1% i 5%. Za svaki od postotaka prikazana su četiri grafa - kako se kroz iteracije mijenjaju mjera uspješnosti (F-score), osjetljivost (TPR), preciznost (PPV) i broj podataka, definirani u poglavlju (1.3).



Slika 3.3: prikaz rada algoritma, 1% izbačenih podataka u svakoj iteraciji



Slika 3.4: prikaz rada algoritma, 5% izbačenih podataka u svakoj iteraciji

Krivulje na grafovima s 1% izbačenih podataka su nešto oštrije i detaljnije, s više "skalovitih" prijelaza, a to je posljedica većeg broja iteracija (249 u usporedbi s 52). U ovom radu, kao mjera uspješnosti modela korišten je F-score. U oba slučaja jasno je vidljivo da se kroz iteracije model sve više i više pospješuje, dok ne dođe do maksimuma, nakon čega počne opadati. Ako promotrimo graf osjetljivosti (TPR), koji nam govori koliki je udio ukupnih bioloških pozitivaca (eng. *condition positives*) model prepoznao kao "dobri", uočavamo da je pad F-score povezan s naglim padom vrijednosti osjetljivosti modela. Krivulja preciznosti (PPV) prikazuje koliki postotak od svih elemenata koje je model ocijenio kao pozitivni predstavljaju biološki pozitivci (koliko je algoritam precizan). Iz tog grafa se vidi kako s vremenom u kugli ostaju samo stvarni pozitivci, a oni loši se izbacuju, što je i bila ideja algoritma.

Sve ovo interpretiramo kao sljedeće - kroz iteracije kugla se zaista stišće prema području pravih pozitivaca i izbacuju se oni pogrešni. Sve do trenutka kada radijus kugle ne postane dovoljno mali da se u sljedećim iteracijama krenu izbacivati i oni biološki pozitivci - to je taj nagli pad u TPR krivulji, s kojim krene i smanjivanje F-score. To se sprječava tako što se algoritam zaustavlja kada se postigne radijus zaustavljanja. Određivanje radijusa zaustavljanja prelazi domenu ovog diplomskog rada i on će se utvrditi eksperimentalno, višestrukim provođenjem i ponavljanjem mjerenja dok se ne uoče zakonitosti koje će dovesti do egzaktne brojke. Pošto je moguće dobiti vrijednost radijusa za kojeg se postiže maksimalan F-score ideja je da će dobiveni radijus biti približno jednak njemu, uz minimalna odstupanja.

Maksimalna vrijednost F-score algoritma s 1% izbačenih podataka je 0.888 uz brzinu izvođenja 1.1 sekunda, a za algoritam s 5% izbačenih podataka je 0.876 uz vrijeme izvođenja 0.28 sekundi. Općenito će se, u daljnjoj prezentaciji rezultata, pokazati da model s 1% izbačenih podataka daje bolje rezultate, između 1 i 3 posto, uz zanemarivo dulje vremensko izvođenje algoritma. Zanemarivo zato što se svi primjeri, čak i oni s više od 2000 podataka, izvode unutar 3 sekunde, dok je ostalim algoritmima u svijetu bioinformatike koji se bave istom problematikom, poput metode najveće klike, za lošije rezultate potrebno i do nekoliko minuta (a za još veće baze podataka i par sati). Stoga, za postizanje najboljih rezultata cijena koje sekunde ne predstavlja nikakav problem.

Za kraj ćemo samo napomenuti da princip rada i rezultat iterativnog pretraživanja proteoma garantira da tijekom izvođenja algoritma kugla neće "odlutati" od okoline stvarnih pozitivaca, što je ključno za efikasnost modela. U odgovoru IGLOSS servera, najviše je proteina distribuirano oko upita, upravo gdje se nalaze i pravi pozitivci, a to osigurava da će se s iteracijama centar kugle nalaziti u tom području. U konačnici, kugla će se sužavati zadržavajući biološke pozitivce.

3.2 Primjeri i rezultati

U ovom radu, uspješnost modela ispitana je na tri različita proteoma:

- Talijin uročnjak (lat. *Arabidopsis thaliana*)
- Krumpir (lat. *Solanum tuberosum*)
- Rajčica (lat. *Solanum lycopersicum*)

Slika 3.5: Talijin uročnjak

Kod svih proteoma korišten je upit FVFGDSLSDA. Za svaki od proteoma zasebno, dana je lista *Condition Positives* (CP) - proteini koji su biološki utvrđeni da pripadaju familiji GDSL lipaza. Mjere uspješnosti izračunate su usporedbom rezultata modela s tim listama. Svi proteini koji se ne nalaze na toj listi smatraju se *Condition Negatives* (CN) - biološki negativni. Svi proteini koje je model vratio kao rezultat označeni su s P (**pozitivni**, eng. *Positives*), dok su svi ostali proteini iz danog proteoma koji nisu u rezultatu označeni s N (**negativni**, eng *Negatives*). Za ilustraciju, slijede odnosi između definiranih pojmova i pojmova iz tablice uspješnosti:

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

Za svaki od proteoma korištene su četiri skale pretraživanja - 5, 4, 3.5 i 3 kako bi se pokazala učinkovitost algoritma s obzirom na različit broj podataka. Za svaku od skala pretraživanja prikazana je tablica rezultata za IGLOSS, algoritam s 5% izbačenih podataka u svakoj iteraciji i algoritam s 1% izbačenih podataka u svakoj iteraciji. U tablici su navedeni osjetljivost modela (TPR), preciznost (PPV), mjera uspješnosti F-score, vrijeme izvođenja, broj bioloških pozitivaca koje je model vratio kao rezultat (TP), broj nizova aminokiselina koje je model vratio kao odgovor (n) i radijus kugle. Za algoritam kugle sve navedene mjere uzete su za kuglu u kojoj se postiže maksimalan F-score, pa tako i radijus kao pokazatelj koliki radijus možemo očekivati u eksperimentalnom računu. Vrijeme izvođenja i radijus kugle nisu dostupni za IGLOSS, a to je ionako bitno samo za usporedbu različitih postotaka izbacivanja.

Napominjem još jednom da se algoritam kugle nadovezuje na iterativno pretraživanje proteoma - uzima odgovor IGLOSS servera i primjenjuje dodatni filter, stoga je za očekivati da će njegova uspješnost biti veća. Navodimo rezultate IGLOSS servera samo kako bi mogli uočiti koliko sam algoritam pospješuje iterativno pretraživanje.

Talijin uročnjak

Talijin uročnjak (lat. *Arabidopsis thaliana*) je mala jednogodišnja cvjetnica. Ona je popularni modelni organizam u biologiji i genetici jer je prva biljka s potpuno sekvenciranim genomom te je stoga pogodna za istraživanja. Njezin proteom je vrlo dobro anotiran i za svaki protein, od njih 35176 u proteomu, znamo kojoj proteinskoj familiji pripada. Duljina liste CP iznosi 104.

1. Skala pretraživanja je 5.

Model	TPR	PPV	F-score	Vrijeme	TP	radijus	n
Algoritam kugla - 1%	0.865	0.909	0.887	0.38s	90	5.84	99
Algoritam kugla - 5%	0.846	0.907	0.876	0.08s	88	5.79	97
IGLOSS	0.894	0.263	0.407	-	93	-	421

2. Skala pretraživanja je 4.

Model	TPR	PPV	F-score	Vrijeme	TP	radijus	n
Algoritam kugla - 1%	0.885	0.902	0.893	0.80s	92	5.74	102
Algoritam kugla - 5%	0.885	0.893	0.889	0.16s	92	5.72	103
IGLOSS	0.942	0.139	0.243	-	98	-	882

3. Skala pretraživanja je 3.5.

Model	TPR	PPV	F-score	Vrijeme	TP	radijus	n
Algoritam kugla - 1%	0.875	0.901	0.888	1.11s	91	5.50	101
Algoritam kugla - 5%	0.856	0.899	0.876	0.28s	89	5.40	99
IGLOSS	0.942	0.092	0.167	-	98	-	1339

4. Skala pretraživanja je 3.

Model	TPR	PPV	F-score	Vrijeme	TP	radijus	n
Algoritam kugla - 1%	0.865	0.882	0.874	1.98s	90	5.38	102
Algoritam kugla - 5%	0.865	0.874	0.870	0.39s	90	5.41	103
IGLOSS	0.923	0.061	0.115	-	96	-	1993

Krumpir

Krumpir (lat. *Solanum tuberosum*) je trajna zeljasta biljka, jedna od najvažnijih prehrambenih namirnica. Duljina liste CP je 123.

1. Skala pretraživanja je 5.

Model	TPR	PPV	F-score	Vrijeme	TP	radijus	n
Algoritam kugla - 1%	0.748	0.902	0.818	0.39s	92	6.09	102
Algoritam kugla - 5%	0.731	0.900	0.807	0.08s	90	6.08	100
IGLOSS	0.772	0.245	0.372	-	95	-	389

2. Skala pretraživanja je 4.

Model	TPR	PPV	F-score	Vrijeme	TP	radijus	n
Algoritam kugla - 1%	0.740	0.778	0.758	0.92s	91	5.82	117
Algoritam kugla - 5%	0.707	0.784	0.744	0.22s	87	5.80	111
IGLOSS	0.772	0.109	0.192	-	95	-	893

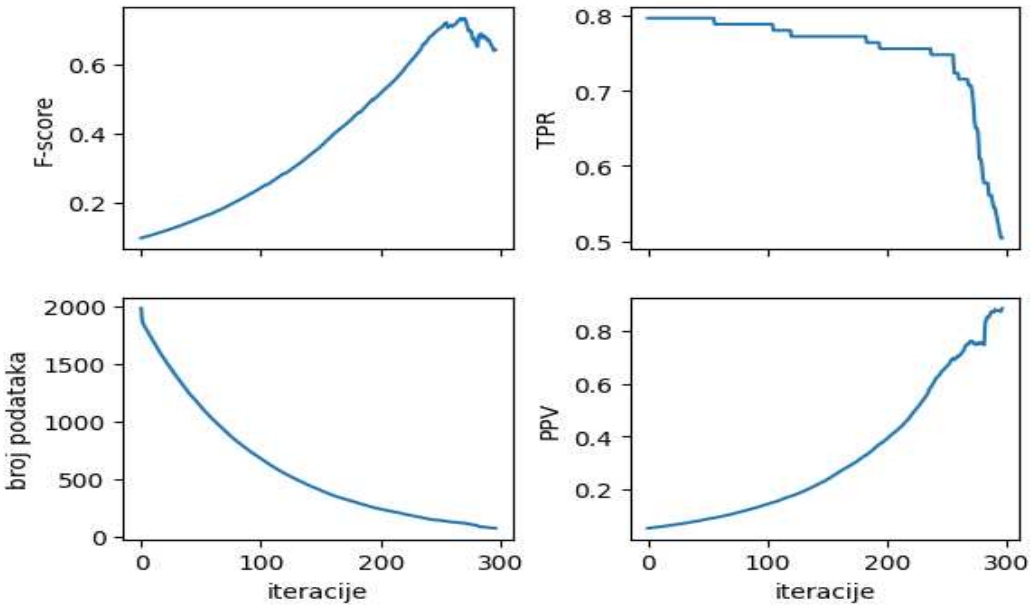
3. Skala pretraživanja je 3.5.

Model	TPR	PPV	F-score	Vrijeme	TP	radijus	n
Algoritam kugla - 1%	0.707	0.791	0.747	1.61s	87	5.68	110
Algoritam kugla - 5%	0.707	0.777	0.740	0.33s	87	5.66	112
IGLOSS	0.780	0.075	0.137	-	96	-	1334

4. Skala pretraživanja je 3.

Model	TPR	PPV	F-score	Vrijeme	TP	radijus	n
Algoritam kugla - 1%	0.707	0.763	0.734	2.24s	87	5.55	114
Algoritam kugla - 5%	0.707	0.750	0.728	0.49s	87	5.55	116
IGLOSS	0.797	0.052	0.098	-	98	-	1983

Za skalu pretraživanja 3 prikazat ćemo i grafički rad algoritma kako bi se uočile iste zakonitosti koje vrijede i za Talijin uročnjak.



Slika 3.6: prikaz rada algoritma, krumpir - 1% izbačenih podataka u svakoj iteraciji

Rajčica

Rajčica (lat. *Solanum lycopersicum*) je jednogodišnja biljka, uzgaja se zbog svojih plodova i važna je prehrambena namirnica. Duljina liste CP je 108.

1. Skala pretraživanja je 5.

Model	TPR	PPV	F-score	Vrijeme	TP	radijus	n
Algoritam kugla - 1%	0.815	0.936	0.871	0.39s	88	6.01	94
Algoritam kugla - 5%	0.815	0.926	0.867	0.08s	88	6.06	95
IGLOSS	0.880	0.249	0.389	-	95	-	387

2. Skala pretraživanja je 4.

Model	TPR	PPV	F-score	Vrijeme	TP	radijus	n
Algoritam kugla - 1%	0.796	0.935	0.860	0.87s	86	5.63	92
Algoritam kugla - 5%	0.796	0.935	0.860	0.19s	86	5.63	92
IGLOSS	0.870	0.110	0.195	-	94	-	882

3. Skala pretraživanja je 3.5.

Model	TPR	PPV	F-score	Vrijeme	TP	radijus	n
Algoritam kugla - 1%	0.824	0.918	0.869	1.48s	89	5.62	97
Algoritam kugla - 5%	0.824	0.918	0.869	0.28s	89	5.65	97
IGLOSS	0.889	0.078	0.144	-	96	-	1292

4. Skala pretraživanja je 3.

Model	TPR	PPV	F-score	Vrijeme	TP	radijus	n
Algoritam kugla - 1%	0.824	0.918	0.869	2.44s	89	5.45	97
Algoritam kugla - 5%	0.824	0.890	0.856	0.50s	89	5.52	100
IGLOSS	0.898	0.051	0.097	-	97	-	2014

Analiza rezultata

U svim slučajevima algoritam kugle s 1% izbačenih podataka daje bolje rezultate od algoritma s 5% izbačenih podataka, te uspijeva zadržati više bioloških pozitivaca. IGLOSS odgovor prepoznaje više TP od algoritma kugle, no to je i očekivano s obzirom na to da algoritam kugle "reže" podatke dobivene od IGLOSS servera pa pritom i odbaci neke stvarne pozitivce. Iz toga slijedi da je uspješnost algoritma kugle ograničena odozgo s brojem bioloških pozitivaca koje IGLOSS uspije prepoznati. Pogledom na vrijednosti preciznosti (PPV) i F-score modela uočavamo koliko je model kugle uspješan u odbacivanju lažnih pozitivaca, bez isključivanja značajnog broja TP. Još jedna bitna stavka je robusnost algoritma s obzirom na broj podataka. Kako se mijenja količina podataka vrijednost F-score algoritma ostaje slična - vrlo visoka uz minimalne promjene. Kod IGLOSS modela, porastom broja podataka značajno opada i uspješnost modela - u prosjeku sa 0.4 na 0.1.

Rezultati pokazuju da odgovor algoritma kugle u prosjeku 90% posto čine biološki pozitivci što su izvrsni rezultati u usporedbi s IGLOSS-ovih 15%, a i općenito. Uz to se uspijeva prepoznati oko 85% od svih proteina u proteomu biljke koji zaista pripadaju traženoj proteinskoj familiji. Dakle, razvijen je iznimno uspješan i brz algoritam koji je uz to i robustan na promjene broja podataka. Postižu se bolji rezultati uz višestruko manje vrijeme izvršavanja od svih algoritama u bioinformatici koji se bave istom problematikom! Možda i najzanimljivija stavka je da je ovo primjer nenadzirane klasifikacije - algoritam iz "ničega" uspijeva prepoznati "nešto" i to vrlo dobro. Bez ikakvih prethodnih informacija o proteinima i njihovoj pripadnosti familiji model pronalazi ciljanu skupinu uz veliku uspješnost.

Za kraj ćemo se dotaknuti i radijusa najbolje kugle. Iz samo nekoliko primjera mogu se uočiti neke zakonitosti koje vrijede. Iznos radijusa je vrlo stabilan i kreće se između 5.4 i 6. S porastom broja podataka smanjuje se radijus i s promjenom proteoma mijenja se raspon u kojem se radijus zaustavljanja nalazi stoga će i to biti faktor u određivanju njegovog iznosa. Sve ovo nam daje razlog za optimizam kako će se uz eksperimentalna istraživanja moći egzaktno odrediti najbolji radijus zaustavljanja.

Pojmovi iz ovog poglavlja preuzeti su iz izvora [4], [5] i [6].

Bibliografija

[1] W. R. Atchley, J.Zhao, A.D. Fernandes, T. Drüke, *Solving the protein sequence metric problem*. Proc. Natlc., Acad. Sci. USA 2005., 102 (18) 6395-6400.

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

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

[4] I. Kapec, *Točnost pretraživanja, clustering i klasifikacija*, Diplomski rad, Sveučilište u Zagrebu, Prirodoslovno-matematički fakultet (Matematički odsjek), 2021.

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

[6] B. Rabar, M. Zagorščak, S. Ristov, M. Rosenzweig i P. Goldestein, *IGLOSS: iterative gapeless local similarity search*, Bioinformatics **35** (2019), br. 18, 3491-3492, ISNN 1367-4803, https://academic.oup.com/bioinformatics/article/35/18/3491/5306940.

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

[8] H.Tušek, *Analiza proteinskih nizova iz Covid-a 19*, Diplomski rad, Sveučilište u Zagrebu, Prirodoslovno-matematički fakultet (Matematički odsjek), 2021.

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

Sažetak

Ovaj diplomski rad proučava problematiku klasifikacije proteina u proteinske familije. Uz pomoć opisa aminokiselina numeričkim vektorima želi se nadopuniti i pospješiti metoda iterativnog pretraživanja proteoma. Proučava se raspored proteina u vektorskom prostoru te se pokušava uočiti geometrijska struktura među njima.

Nakon navođenja matematičkih i bioloških pojmova nužnih za razumijevanje ovog rada i uvođenja strukture podataka na kojima se provodi analiza, prezentira se algoritam koji pronalazi kuglu u vektorskom prostoru koja najbolje grupira proteine koji su biološki dokazani da pripadaju određenoj proteinskoj familiji. Postupak je proveden na proteomima talijinog uročnjaka, krumpira i rajčice. Promatrao se učinak algoritma na manjim i većim skupovima podataka. Glavna mjera za uspješnost modela je F_1-score. Rezultati pokazuju da je razvijen algoritam koji pospješuje iterativno pretraživanje - značajno povećava postotak biološki ispravnih proteina u odgovoru, odbacujući one pogrešne. Model se pokazao robustan na promjene broja podataka, uz iznimno brzo vremensko izvođenje algoritma.

Summary

This thesis covers the issue of classification of proteins into protein families. With the help of the description of amino acids by numerical vectors, the aim is to improve the method of iterative proteome search. By studying the distribution of proteins in the vector space, we try to observe the geometric structure between them.

After stating the mathematical and biological concepts necessary to understand this paper and introducing the data structure on which the analysis is performed, an algorithm that finds a sphere in the vector space that groups proteins, which are biologically proved to belong to a particular protein family, is presented. The procedure was performed on proteomes of thale cress, potato and tomato. Algorithm performance was measured on both small and large data sets. F_1-score was the primary metric used for model success. Results obtained show that the algorithm improves iterative search - it significantly increases the percentage of biologically correct proteins in the response, by removing the wrong ones. The model proved to be robust to changes in the data size, with extremely fast execution time.

Životopis

Rođen sam u Splitu, 25. ožujka 1997. godine. Školovanje započinjem u Osnovnoj školi Manuš u Splitu, nakon koje upisujem III. gimnaziju, također u Splitu. Po završetku srednjoškolskog obrazovanja 2015. godine, odlazim u Zagreb gdje upisujem preddiplomski studij matematike na Prirodoslovno-matematičkom fakultetu. Zvanje sveučilišnog prvostupnika matematike stječem 2018. godine, kada upisujem i diplomski studij Matematičke statistike na istom fakultetu.

Slobodno vrijeme volim provoditi uz obitelj i prijatelje. Hobiji su mi svi oblici sporta - najviše nogomet i košarka, te ribolov.