Traženje proteinskih motiva i klasifikacija

Iveković, Marko

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Marko Iveković

TRAŽENJE PROTEINSKIH MOTIVA I KLASIFIKACIJA

Diplomski rad

Voditelj rada:
doc. dr. sc. Pavle Goldstein

Zagreb, svibanj, 2022.

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

Problem s kojim se susrećemo u ovom radu je problem nenadzirane klasifikacije, gdje promatramo skup proteina nekog organizma i te proteine želimo klasificirati u određene proteinske familije. Skup proteina nekog organizma naziva se proteom, a svaki se protein sastoji od nizova aminokiselina. U svakom proteinu tražimo karakterističan podniz aminokiselina promatrane proteinske familije (takav niz aminokiselina se zove motiv) te se tako problem svodi na klasifikaciju kratkih nizova aminokiselina.

Koristeći opis aminokiselina numeričkim faktorima taj problem smještamo u vektorski prostor te se vodimo idejom da bi nizovi aminokiselina. tj. potencijalni motivi, trebali biti gušće raspoređeni od ostalih nizova aminokiselina u tom prostoru ako su zapravo motivi. Tehnika opisana u ovom radu pronalazi središte i radijus kugle koja bi klasificirala promatrane nizove aminokiselina na motive i na one koji to nisu.

Prvo ćemo definirati matematičke pojmove koji su nam potrebni za razvoj algoritma koji pronalazi kuglu koja klasificira nizove aminokiselina. Nakon toga ćemo obraditi pojmove iz bioinformatike koji su nam potrebni za razumijevanje rada te onda smjestiti problem u vektorski prostor i svesti ga na problem klasifikacije u višedimenzionalnom vektorskom prostoru.

Poglavlje 1

Matematički pojmovi

1.1 Linearna algebra

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. *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 unitaran prostor.*

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

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

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

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

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

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

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

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

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

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. *Neprazan skup Ω koji reprezentira skup svih ishoda slučajnog pokusa zovemo prostor elementarnih događaja. Elemente ω skupa Ω zovemo elementarni događaji.*

Definicija 1.2.3. *Familija $\mathcal{F}$ podskupova od Ω ($\mathcal{F} \subset \mathcal{P}(\Omega)$) jest σ-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 $\mathbb{P} : \mathcal{F} \to \mathbb{R}$ jest vjerojatnost (na $\mathcal{F}$, na Ω) ako vrijedi*

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

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 *i* $\mathbb{P}$ *vjerojatnost na* $\mathcal{F}$*, zove se vjerojatnosni prostor.*

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 zovemo σ-algebra generirana sa* $\mathcal{A}$*.*

Definicija 1.2.8. *Neka je* $\mathbb{R}$ *skup realnih brojeva. Sa* $\mathcal{B}$ *označimo σ-algebru generiranu 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* $X : \Omega \to \mathbb{R}$ *jest 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

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

gdje su $x_1,\ x_2, \ldots,\ x_n$ realni brojevi, a $A_1,\ A_2, \ldots,\ A_n$ međusobno disjunktni događaji, $\bigcup_{k=1}^{n} A_k = \Omega$.

Neka su $X_1,\ X_2 : \Omega \to \mathbb{R}$. Tada definiramo funkcije $X_1 \vee X_2$ i $X_1 \wedge X_2$ na Ω, relacijama:

$$(X_1 \vee X_2)(\omega) = max\{X_1(\omega), X_2(\omega)\},\ \omega \in \Omega, \tag{1.1}$$

i

$$(X_1 \wedge X_2)(\omega) = min\{X_1(\omega), X_2(\omega)\},\ \omega \in \Omega.$$

Pomoću funkcije (1.1) definiramo pozitivan i negativan dio realne funkcije X na Ω:

$$X^+ = X \vee 0,\ X^- = (-X) \vee 0.$$

X^+ i X^- su nenegativne realne funkcije i vrijedi:

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

$$|X| = X^+ + X^-.$$

Korolar 1.2.12. *X je slučajna varijabla ako i samo ako su X^+ i X^- slučajne varijable.*

Teorem 1.2.13. *Neka je X nenegativna slučajna varijabla na Ω. Tada postoji rastući niz $(X_n, n \in \mathbb{N})$ nenegativnih jednostavnih slučajnih varijabli takav da je $X = \lim_{n\to\infty} X_n$ (na Ω).*

Definicija 1.2.14. *Neka je X slučajna varijabla na Ω. Funkcija distribucije od X jest 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 1.2.15. *Ako je jasno o kojoj se slučajnoj varijabli, odnosno njenoj funkciji distribucije radi, pišemo $F_X = F$.*

Teorem 1.2.16. *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 svojstva iz prethodnog teorema zovemo vjerojatnosna funkcija distribucije (na $\mathbb{R}$) ili, kraće, funkcija distribucije.

Definicija 1.2.17. *Funkcija $g : \mathbb{R} \to \mathbb{R}$ jest 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 1.2.18. *Neka je $(\Omega, \mathcal{F}, \mathbb{P})$ vjerojatnosni prostor i X slučajna varijabla na Ω. Slučajna varijabla X je diskretna ako postoji konačan ili prebrojiv skup $D \subset \mathbb{R}$ takav da je $\mathbb{P}(X \in D) = 1$*

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

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

Prethodnu tablicu zovemo distribucija ili zakon razdiobe slučajne varijable X.

Definicija 1.2.19. *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*

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

Napomena 1.2.20. *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 1.2.21. *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 $X \sim N(\mu, \sigma^2)$.

Napomena 1.2.22. *X je jedinična normalna distribucija ako je $X \sim N(0,1)$, dakle je*

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

Matematičko očekivanje

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. Sa $\mathcal{K}$ označimo 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 1.2.23. *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 1.2.24. *1. Neka je $c \in \mathbb{R}$ i $X \in \mathcal{K}$. Tada je $\mathbb{X} = c\mathbb{E}X$.*

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

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

Neka je X nenegativna slučajna varijabla definirana na Ω. Prema teoremu 1.2.13 postoji rastući niz $(X_n)_{n\in\mathbb{N}}$ nenegativnih jednostavnih slučajnih varijabli takav da je $X = \lim_{n\to\infty} X_n$. Iz propozicije 1.2.24 slijedi da je niz $(\mathbb{E}[X_n])_{n\in\mathbb{N}}$ rastući niz u $\mathbb{R}_+$, dakle postoji $\lim_{n\to\infty} \mathbb{E}[X_n]$ koji može biti jednak i $+\infty$.

Definicija 1.2.25. *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 X proizvoljna slučajna varijabla na Ω. Vrijedi $X = X^+ - X^-$, X^+, X^- su slučajne varijable i $X^+, X^- \geq 0$.

Definicija 1.2.26. *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 1.2.27. *$\mathbb{E}(X^r)$ zovemo r-ti moment od X, a $\mathbb{E}(|X|^r)$ zovemo r-ti apsolutni moment od X*

Definicija 1.2.28. *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 1.2.29. *Varijanca od X koju označavamo sa $Var(X)$ ili σ_X^2 jest drugi cetralni moment od X, dakle*

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

Napomena 1.2.30. *Pozitivan drugi korijen iz varijance nazivamo standardna devijacija i označavamo sa σ_X.*

Opisna analiza podataka

U ovom ćemo se dijelu podsjetiti definicija iz opisne ili deskriptivne statistike koje će nam biti potrebne u daljnjem radu, a to su aritmetička sredina, standardna devijacija te varijanca uzorka.

Neka su

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

n vrijednosti varijable X koje čine skup podataka. Ako je X numerička varijabla, tada je to niz brojeva.

Neka je u X numerička varijabla. Aritmetička sredina brojeva 1.3 je broj:

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

Najčešće korištena mjera raspršenja skupa numeričkih podataka je standardna devijacija. Stadardna devijacija je srednje kvadratno odstupanje podataka od njihove aritmetičke sredine. Definirana je formulom:

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

Standardna devijacija je, kako smo već definirali, drugi korijen varijance. Varijanca skupa podataka 1.3 je mjera raspršenja podataka i predstavlja prosječno kvadratno odstupanje podataka od njihove aritmetičke sredine i dana je formulom:

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

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:

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

1.3 Klasifikacija i uspješnost modela

Klasifikacija

Klasifikacija je problem identificiranja pripadnosti objekta nekoj od skupa klasa na način da su u istim klasama sličniji objekti nego u dvije različite klase. Navedimo ovdje dva pristupa klasifikacije objakata. Prvi je nadzirana klasifikacija koji koristi unaprijed određene klase te se tada objekt pridružuje u klasu koja joj je najsličnija. Sljedeći pristup je nenadzirana klasifikacija, gdje nemamo unaprijed određene klase, već se na podacima pokušavaju odrediti neki obrasci sličnosti među objektima i separirati među te klase.

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 ($\beta = 1$):

$$F_1 = \frac{2 \cdot PPV \cdot TPR}{PPV + TPR} \tag{1.4}$$

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

Bjelančevine ili proteini su, uz vodu, najvažnije tvari u tijelu. Najvažniji su čimbenik u rastu i razvoju svih tjelesnih tkiva. Glavni su izvor tvari za izgradnju mišića, krvi, kože, kose, noktiju i unutarnjih organa, uključujući srce i mozak. Sastavni su dijelovi svake stanice što ih čini osnovom života na Zemlji. Izgrađene su od aminokiselina koje su međusobno povezane peptidnom vezom. Postoji 20 standardnih aminokiselina koje izgrađuju proteine te se svakoj aminokiselini pridružuje jedno slovo engleske abecede na način prikazan u tablici:

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

Neka je R slučajna varijabla koja predstavlja aminokiselinu s distribucijom

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

Vjerojatnosti navedene u ovoj distribuciji predstavljaju vjerojatnosti pojavljivanja pripadne aminokiseline u nekom prostoru proteina.

Skup svih proteina nekog organizma naziva se proteom. Sastoji se od različitih proteinskih familija koje su zaslužne za različita svojstva organizma. Jedna od tih proteinskih familija je familija GDSL lipaza. To je familija koja sadrži enzime čija je karakteristika fleksibilno katalitičko mjesto koje mijenja svoj položaj ovisno o prisustvu pojedinih supstrata. To bi svojstvo moglo objasniti njihovu katalitičku multifunkcionalnost, što ih čini privlačnom temom za istraživanje.

GDSL hidrolaze se mogu pronaći u mnogim organizmima te su nedavne analize otkrile široku rasprostranjenost u nekim kopnenim biljkama. Iako elementi familije GDSL lipaza sudjeluju u velikom broju staničnih procesa, kao što su razvoj biljaka, zaštita organizma od patogena i stresa te ih se veže uz svojstvo hidrofilnosti i hidrofobnosti, još su slabo istražene. Njihova do sad uočena multifunkcionalnost daje nam naslutiti da bi biljke mogle biti izvor vrlo korisnih enzima za primjenu u hidrolizi i sintezi ostalih zanimljivih i korisnih spojeva u biotehnologiji. Iz tog je razloga otkrivanje novih biljnih GDSL lipaza iznimno važno.

2.2 Prelazak u vektorski prostor

U ovom radu promatramo k-torke aminokiselina koje su predstavljene slovima abecede, pa kako bi na njima mogli provesti bilo kakve statističke analize, potreban nam je opis aminokiselina nekim numeričkim vrijednostima. Taj problem je riješen ovdje [1]. Definira se preslikavanje u $\mathbb{R}^5$ koje svakoj aminokiselini pridružuje 5-dimenzionalni vektor, gdje svaka koordinata vektora opisuje neko svojstvo odgovarajuće aminokiseline. Prva koordinata opisuje svojstvo polariteta aminokiseline, druga koordinata opisuje svojstvo sekundarnog naboja, treća koordinata opisuje molekularni volumen, četvrta koordinata opisuje raznolikost kodona (relativnu kompoziciju aminokiselina u različitim proteinima), a peta koordinata opisuje elektrostatički naboj aminokiseline. Tako definiranom nizu od n aminokiselina je pridružen $5n$-dimenzionalan vektor, pa kako promatramo nizove od 10 aminokiselina, takvom je nizu pridružen 50-dimenzionalni vektor

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

Poglavlje 3

Analiza problema i algoritam

Problem s kojim se u ovom radu susrećemo je da iz niza proteina određenog proteoma pokušamo pronaći proteine koji pripadaju određenoj proteinskoj familiji. Kako bi ponudili rješenje tog problema prvo ćemo definirati što je motiv. Motiv je niz aminokiselina, duljine od 5 do 20, koji je ostao djelomično sačuvan tijekom evolucije. Taj niz je karakterističan za neku proteinsku familiju. Jedan od modela pronalaženja proteina koji pripadaju određenoj proteinskoj familji je taj da se kao ulazni podatak navodi neki upit (motiv) te se unutar proteoma pronalaze podnizovi nizova proteina koji su dovoljno slični upitu, pa ako protein sadrži takav podniz, model ga klasificira kao da pripada određenoj proteinskoj familiji. Model koji ćemo mi u ovom radu koristiti za spomenuto pretraživanje proteoma je IGLOSS server [8]. IGLOSS server kao upit prihvaća neki niz aminokiselina tj. motiv, proteom koji se pretražuje te neki broj koji označava skalu pretraživanja, gdje je skala pretraživanja parametar koji postavlja granicu dovoljne sličnosti. Model pretražuje zadani proteom te daje odgovor koji sadrži skup nizova aminokiselina koji su dovoljno slični nizu aminokiselina kojeg smo zadali.

Kada smo pomoću IGLOSS servera dobili skup potencijalnih motiva, među njima se nalazi popriličan broj lažno pozitivnih nizova aminokiselina, odnosno nizova aminokiselina koji ne pripadaju traženoj proteinskoj familiji. Mi bismo taj skup htjeli smanjiti tako da pronađemo način koji bi eliminirao lažno pozitivne elemente, a da pri tome zadržimo elemente koji bi se stvarno trebali nalaziti u tom skupu. Ideja je da dobivene nizove aminokiselina (potencijalne motive) prikažemo kao točke te tada očekujemo da će točke koje predstavljaju nizove aminokiselina koji zaista pripadaju promatranoj proteinskoj familiji (motive) biti gušće raspoređene od ostalih točaka. Iz tog ćemo razloga preći u vektorski prostor gdje možemo iskoristiti njegova svojstva. Sada smo taj problem sveli na traženje središta i radijusa optimalne kugle, a gdje nam je mjera uspješnosti F_1 score.

Pronalazak središta kugle je obrađen u diplomskom radu [3], a do njega se dolazi na sljedeći način. Kao upit koji dajemo IGLOSS serveru koristimo upit koji sadrži niz aminokiselina GDSL iz razloga što je karakterističan za promatranu proteinsku familiju. Korištenjem takvog upita žele se dobiti najbolji kandidati za familiju GDSL lipaza. Kao i upit, svi odgovori će biti nizovi aminokiselina duljine 10. Slijedi da su prelaskom u vektorski prostor naši podaci 50-dimenzionalni vektori. Sada na podacima dobivenima kao rezultat upita IGLOSS-u provodimo postupak standardizacije podataka. Standardizacija podataka se izvodi na način da se podaci transformiraju oduzimanjem očekivanja i dijeljenjem sa standardnom devijacijom uzorka. Ako su $x_1, x_2, \ldots, x_n$ n vrijednosti varijable koje čine skup podataka, tada je:

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

Provedbom standardizacije podataka izbjegavamo problem nedefiniranih mjernih jedinica uz vrijednosti podataka te, iz tog razloga, raspršenosti i nejednolikog raspona podataka. Uz to osiguravamo da naš algoritam koji pokušava pronaći optimalne kugle zaista može pronaći tražene kugle, a ne da radimo s elipsoidima.

Sada kada smo standardizirali podatke, vodimo se idejom da su biološki pozitivci, tj. motivi, 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 tocaka 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 stišče prema pravim pozitivcima. Algoritam se zaustavlja kada kugla dosegne jednak ili manji radijus od zadanog.

Nakon pronalaska središta tražene kugle, preostalo nam je pronaći i njen radijus tako da ćemo se sad usredotočiti na pronalazak spomenutog radijusa.

3.1 Algoritam

Pretpostavimo da se aminokiseline pojavljuju s vjerojatnostima p_k, $k \in \{1, 2, \ldots, 20\}$ zadanima u distribuciji aminikiselina navedenoj u prethodnom poglavlju te neka su A_i $i \in \{1, 2, \ldots, 20\}$ distribucije zadane nekom aminokiselinom za koju ćemo reći da je očuvana

koeficijentom očuvanosti α:

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

gdje broj u sufiksu pokraj oznake slučajne varijable označava redni broj aminokiseline iz niza prostora aminokiselina, a vjerojatnosti p_j^i su jednake

$$p_j^i = \alpha \cdot \mathbb{1}_{\{i=j\}} + (1 - \alpha) p_j$$

gdje broj u sufiksu pokraj vjerojatnosti p_j također označava redni broj aminokiseline iz niza prostora aminokiselina. Nadalje koristimo teorem koji je opisan u izvoru [6, str. 55]:

Teorem 3.1.1. *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.*

Računamo očekivanu udaljenost dviju 10-orki aminokiselina. Neka su $X = (x_1, x_2, \dots, x_{10})$, $Y = (y_1, y_2, \dots, y_{10})$, dvije 10-orke aminokiselina. Očekivanje kvadrata euklidske udaljenosti dviju 10-orki jednako je:

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

Kako nemamo nikakve pretpostavke o položaju aminokiselina po pozicijama u tim 10-orkama, odnosno ne razlikujemo pozicije, označimo s $\bar{a}_i$ i $\bar{a}_j$ aminokiseline koje pripadaju prosječnoj distribuciji aminokiselina:

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

Iz svojstva očekivanja slijedi:

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

Izračunajmo sada izraz sa desne strane jednakosti. Neka su a_i^k i a_j^k neke dvije aminokiseline iz distribucije A_k. Tada vrijedi:

$$\mathbb{E}\left[\left(a_i^k - a_j^k\right)^2\right] = \sum_{i,j=1}^{20} \left(a_i^k - a_j^k\right)^2 p_j^k p_j^k$$

Kako distribuciju A_k određuje aminokiselina koja je odabrana s vjerojatnosću pojavljivanja te aminokiseline u prostoru proteina kojeg promatramo, slijedi da je očekivanje za prosječnu distribuciju jednako:

$$\mathbb{E}\left[\left(\bar{a}_i - \bar{a}_j\right)^2\right] = \sum_{k=1}^{20} p_k \sum_{i,j=1}^{20} \left(a_i^k - a_j^k\right)^2 p_j^k p_j^k = 10.8724$$

pa je očekivani kvadrat udaljenosti za neke dvije 10-orke aminokiselina jednak $10 \cdot 10.8724$, te je tada očekivana udaljenost jednaka $\sqrt{10} \cdot 3.2973$, što možemo interpretirati kao očekivanu maksimalnu udaljenost za dvije točke koje prikazuju 10-orku aminokiselina, tj. udaljenost dvije instance 10-orke aminokiselina ne bi smjela biti veća od ovog rezultata uz pretpostavku očuvanosti α.

Sada kada imamo očekivanu maksimalnu udaljenost, iz teorema 3.1.1 slijedi da je $r = \frac{\sqrt{10} \cdot 3.2973}{\sqrt{2}} = 3.2973 \cdot \sqrt{5}$

Računamo standardne devijacije podataka prije standardizacije i nakon standardizacije, te budući da je izračunati radijus proporcionalan sa standardnom devijacijom, radijus nakon standardizacije proporcionalan je sa standardom devijacijom nakon standardizacije. Vrijedi da je traženi radijus jednak:

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

gdje r_{new} označava traženi radijus, std_{old} označava standardnu devijaciju podataka prije standardizacije, std_{new} standardnu devijaciju podataka nakon standardizacije, a r_{old} radijus kojeg smo izračunali s vrijednošću $3.2973 \cdot \sqrt{5}$ Na taj smo način dobili procjenu radijusa tražene optimalne kugle:

$$r_{new} = 3.2973 \cdot \sqrt{5} \cdot \frac{std_{new}}{std_{old}}$$

3.2 Primjeri i rezultati

Procjenu radijusa testirat ćemo na tri proteoma, a to su rajčica, krumpir i talijin uročnjak. Njihovi opisi preuzeti su iz izvora [13].Provedeni račun i sljedeće primjere implementirali smo i proveli u programskom jeziku *Python*.

U sva tri primjera kao upit koji smo uputili serveru IGLOSS bio je FVFGDSLSDA, a on nam je dao odgovor za koje 10-orke u pojedinom proteomu on smatra da pripadaju familiji GDSL lipaza. Također, bila nam je dostupna i lista bioloških pozitivaca, to jest lista 10-orki aminokiselina za koje je stvarno određeno da pripadaju familiji GDSL lipaza, te smo tako mogli testirati procjenu radijusa tražene kugle. Usporedbu smo proveli tako da smo usporedili procijenjeni radijus s radijusom kugle koja ima središte koje je pronađeno na način opisan ovdje [3] i radijusom u kojem mjera F_1 poprima maksimum. Postupak smo proveli na spomenutim proteomima na skalama pretraživanja od 3 do 7.

Za svaki proteom i za svaku skalu navest ćemo radijus u kojem se postiže maksimalan F_1-score, procjenu radijusa, njihovu razliku, maksimalan F_1-score i kategorizaciju F_1-scorea procijenjenog radijusa. Kategorije F_1-scorea procijenjenog radijusa označit ćemo na sljedeći način: < 1, < 2, < 3, < 4, < 5 te > 5 gdje svaka kategorija označava unutar koliko se posto F_1-score procijenjenog radijusa razlikuje od F_1-scorea radijusa s maksimalnim F-scoreom, odnosno kod kategorije s oznakom > 5 vrijedi da je razlika između F_1-scorea procijenjenog radijusa i F_1-scorea radijusa s maksimalnim F_1-scoreom veća od 5 posto.

Rajčica

Rajčica (*Lycopersicon esculentum*) je jednogodišnje povrće razgranate, zeljaste stabljike iz porodice pomoćnica (*Solanaceae*), podrijetlom iz Perua. Fiziološki zreli plodovi sočne su i mesnate bobe, različita oblika i veličine, najčešće crvene boje, skupljeni u grozdove; rabe se za jelo svježi ili prerađeni (sok, koncentrat, pelat, kečap i dr.). Suha tvar ploda (5 do 7%) sadrži ugljikohidrate (fruktoza, glukoza), organske kiseline, bjelančevine, minerale, vlakna i vitamine.

Slika 3.1: Rajčica

Skala	Optimalni radijus	Procijenjeni radijus	Razlika	Maksimalan F_1-score	Kategorija
3	5.4	5.38900	0.0110	0.6666	<2
4	5.5	5.59758	0.0976	0.6666	<3
5	6	5.79480	0.2052	0.6666	<5
6	6.3	6.34550	0.0455	0.6769	<1
7	6.6	6.93300	0.3330	0.6770	<1

Krumpir

Krumpir (*Solanum tuberosum*), trajna je zeljasta biljka iz porodice pomoćnica (*Solanaceae*), s razgranjenom, do jednog metra visokom nadzemnom stabljikom, tzv. cimom, koja nosi listove, i s mnogobrojnim podzemnim stabljikama (stolonima), koje su na krajevima odebljale u kuglaste, jajaste ili valjkaste gomolje – krumpire. Peteročlani cvjetovi skupljeni su u paštitaste cvatove bijelih, svijetloplavih, ljubičastih ili ružičastih cvjetova, a iz plodnice se, ovisno o odlici i uvjetima uzgoja, razvijaju plodovi, zelene bobe s mnogo sjemenki. Gomolj je glavni rezervni dio biljke, koji služi za njezino prezimljenje i razmnožavanje. Krumpir se u kulturi ne razmnožava sjemenom već podzemnim gomoljima, na kojima se iz njihovih pupova (oka) zameću novi izdanci. Sjeme se može upotrebljavati u oplemenjivanju novih odlika. Odabiranjem i križanjem uzgojene su mnoge odlike, koje se dijele prema uporabi krumpira (za sjeme, ljudsku i stočnu hranu, kao industrijska sirovina), duljini vegetacije (rane, srednje rane, srednje kasne i kasne), boji kožice (žute, smeđe do crvene boje), boji ”mesa” (od bijele do žute boje). Danas u Hrvatskoj ima oko 50 odlika krumpira.

Slika 3.2: Krumpir

Skala	Optimalni radijus	Procijenjeni radijus	Razlika	Maksimalan F_1-score	Kategorija
3	5.5	5.32000	0.1800	0.6175	>5
4	5.8	5.56300	0.2370	0.6323	<4
5	6.1	5.91400	0.1860	0.6764	<4
6	6.6	6.65560	0.0556	0.6815	<1
7	7.1	7.23768	0.1377	0.6692	<1

Talijin uročnjak

Talijin uročnjak (*Arabidopsis thaliana*) jedna je od desetak vrsta biljaka iz porodice krstašica. Popularna je u istraživanjima u biologiji i genetici jer ima potpuno sekvenciran genom. Njen proteom je vrlo dobro anotiran i za skoro svaki protein, od njih 35176 u proteomu, znamo kojoj proteinskoj familiji pripada.

Slika 3.3: Talijin uročnjak

Skala	Optimalni radijus	Procijenjeni radijus	Razlika	Maksimalan F_1-score	Kategorija
3	5.4	5.50870	0.1087	0.7153	<2
4	5.6	5.59790	0.0021	0.7324	<2
5	6	6.11500	0.1150	0.7208	<2
6	6.5	6.66200	0.1620	0.6031	<1
7	7.1	7.04100	0.0590	0.6035	<1

Analiza rezultata

Na promatranim primjerima rajčice, krumpira i talijinog uročnjaka uočavamo da su radijusi u kojem se postiže maksimalan F_1-score i radijus kojeg smo procijenili blizu te, što je još i bitnije, da su male razlike u F_1 scoreovima među tim radijusima. U gotovo svim promatranim primjerima i skalama, razlika između maksimalnog F_1-scorea i F_1-scorea procijenjenog radijusa bila je unutar 5 posto, najčešće barem unutar 2 posto, osim na skali 3 kod primjera rajčice gdje je taj F_1-score bio malo veći od toga. Stoga zaključujemo da je ovo dobra procjena optimalnog radijusa, tj. radijusa u kojem se postiže maksimalan F_1-score.

Pronalaskom središta i radijusa dobivamo kuglu koju u prosjeku 90% čine biološki pozitivci u odnosu na 15% kada gledamo podatke koje daje IGLOSS. Uz to se uspijeva prepoznati oko 85% od svih proteina u proteomu biljke koji zaista pripadaju traženoj proteinskoj familiji. Kao rezultat, pronašli smo algoritam koji je brz, uspješan i robustan, a da smo pri tome koristili nenadziranu klasifikaciju, odnosno, bez unaprijed određenih zakonitosti smo uspjeli pronaći skup točaka koji te točke uspješno klasificira u promatranu proteinsku familiju.

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

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

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

[6] Maurice George Kendall, Patrick Alfred Pierce Moran, *Geometrical probability*, Hafner Publishing Company, 1963, London

[7] Braslav Rabar, Keti Nižetić, Maja Zagorščak, Kristina Gruden, Pavle Goldstein, *A Clique-Based Method for Improving Motif Scanning Accuracy*, University of Zagreb, Faculty of Science, Mathematics Department and National Institute of Biology, Department of Biotechnology and Systems Biology

[8] 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`.

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

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

[11] Ivan Vujaklija, Ana Bielen, Tina Paradžik, Siniša Biđin, Pavle Goldstein, Dušica Vujaklija, *An effective approach for annotation of protein families with low sequence similarity and conserved motifs : identifying GDSL hydrolases across the plant kingdom*, BMC bioinformatics, 17 (2016), 91-1 doi:10.1186/s12859-016-0919-7, dostupno na: https://bmcbioinformatics.biomedcentral.com/track/pdf/10.1186/s12859-016-0919-7.pdf

[12] Slike priložene u radu dostupne su na: https://commons.wikimedia.org/wiki/File:Solanum_tuberosum_Beate.jpg
https://commons.wikimedia.org/wiki/File:Arabidopsis_thaliana_JdP_2013-04-28.jpg
https://commons.wikimedia.org/wiki/File:Flower_of_Solanum_lycopersicum.jpg

[13] https://www.enciklopedija.hr/

Sažetak

U ovom diplomskom radu promatra se problem klasifikacije niza proteina određenog proteoma u neku proteinsku familiju. U svakom proteinu tražimo karakterističan podniz aminokiselina promatrane proteinske familije (motiv) te se tako problem svodi na klasifikaciju nizova aminokiselina na to jesu li oni motivi ili nisu. Koristeći postojeći iterativni algoritam dobiva se rezultat koji je skup potencijalnih motiva u kojem se nalazi popriličan broj lažno pozitivnih elemenata te je cilj eliminirati lažno pozitivne, a da se pritom sačuvaju oni koji su stvarno pozitivni, odnosno motivi.

Rješavanju tog problema pristupa se tako da se promatra opis aminokiselina numeričkim faktorima te se dane nizove aminokiselina smješta u vektorski prostor, gdje se među njima pokušava uočiti neka geometrijska struktura. Prethodno je razvijen algoritam koji u tom prostoru pronalazi središte kugle koja klasificira promatrane nizove aminokiselina, dok je pitanje radijusa ostalo otvoreno. U ovom radu se uz prethodno spomenuto smještanje nizova aminokiselina u vektorski prostor pronalazi procjena radijusa spomenute kugle.

Izračunata procjena radijusa i središta kugle testira se na primjerima proteoma rajčice, krumpira i talijinog uročnjaka uz F_1-score kao mjeru uspješnosti algoritma. Rezultati pokazuju da algoritam kojim se pronalazi središte i radijus određuje kuglu koja je bila i cilj rada, a to je rješenje problema eliminiranja lažno pozitivnih nizova, a da se pritom sačuvaju oni koji su stvarno pozitivni. Uz to je taj algoritam brz te robustan na promjene veličine uzorka.

Summary

This thesis covers the classification problem of protein sequences of specified proteome into some protein family. In every protein we are searching for a characteristic subsequence of aminoacids (a motif) so the problem is reduced to classification of aminoacids. Using the existing iterative algorithm we get a set of potential motifs in which there is a certain number of false positive elements. Our goal is to eliminate the false positives while the true positive elements (motifs) remain.

We approach the problem by examining a description of aminoacids in terms of numerical factors and by embedding the aminoacid sequences into a vector space and studying geometry of this embedding. An existing algorithm finds a center of a sphere which classifies the sequences, while the question of the radius remained open. In this thesis we estimate the radius of this sphere.

The classification procedure is tested on three plant proteomes, using F_1-score as a measure of accuracy. The results show that the algorithm gives a good solution to the problem. Furthermore, it is fast and robust with respect to the changes in data size.

Životopis

Rođen sam 5. srpnja 1993. godine u Zagrebu. Školovanje sam započeo u Osnovnoj školi Josipa Broza u Kumrovcu, nakon koje upisujem gimnaziju Antuna Gustava Matoša u Zaboku. Nakon završetka srednjoškolskog obrazovanja upisujem preddiplomski studij matematike, nastavnički smjer, na Prirodoslovno-matematičkom fakultetu u Zagrebu te nakon toga i diplomski studij Matematička statistika na istom fakultetu.