Clustering i klasifikacija proteinskih nizova

Bevc, Laura

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Laura Bevc

CLUSTERING I KLASIFIKACIJA PROTEINSKIH NIZOVA

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. _____

Zahvaljujem mentoru doc.dr.sc. Pavlu Goldsteinu na velikoj pomoći pri pisanju ovog rada, na uloženom vremenu, strpljenju i trudu.

Sadržaj

Uvod

Teški akutni respiratorni sindrom koronavirus 2 (skraćeno SARS-CoV-2) je zarazni virus koji uzrokuje bolest dišnih puteva COVID-19. Spada u RNA viruse i sadrži četiri strukturna proteina. To su proteini E (eng. *envelope*), M (eng. *membrane*), N (eng. *nucleocapsid*) i S (eng. *spike*). Uzrok je pandemije koronavirusa koja je počela krajem 2019. godine i proširila se po cijelom svijetu. Brzini širenja pandemije pridonijele su i brojne mutacije na virusu.

U ovom radu će se analizirati protein S poznat i kao protein šiljka. To je glikoprotein koji tvori strukturu koja se nalazi na površini virusa te pomaže virusu da se veže za stanicu u tijelu. Sastoji se od linearnog lanca koji sadrži 1273 aminokiseline. Sekvencioniran je mnogo puta te je tako identificirano na tisuće različitih varijanti.

Cilj ovog rada je u promatranom uzorku pokušati pronaći pouzdan način grupiranja nizova različitih varijanti S proteina. Dodatna pitanja koja se nameću ukoliko se pronađu takve pouzdane grupe su odgovaraju li one raznim varijantama virusa te koje su mutacije karakteristične za njih.

Ovaj rad se sastoji od četiri poglavlja. U prvom poglavlju su navedeni pojmovi iz linearne algebre te statistike nužni za razumijevanje ostatka rada. Uz definirane pojmove objašnjen je i algoritam k-means koji je korišten u analizama. U drugom poglavlju definirana je biološka pozadina problema, objašnjeni su osnovni biološki pojmovi te je definiran način prikazivanja aminokiselina u vektorskom prostoru. U trećem poglavlju su analizirani i pripremljeni podaci na kojima su rađeni izračuni i na kojima je proveden k-means algoritam. Konačno, u četvrtom poglavlju vizualni prikaz podataka indicira moguće značajne klastere. Ideja je testirati postoje li klasteri koji odgovaraju nekim varijantama te koje su mutacije karakteristične za njih.

Poglavlje 1

Matematički pojmovi

U ovom poglavlju navode se teoremi, definicije, propozicije i napomene iz linearne algebre i statistike. 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* $\cdot : \mathbb{F} \times \mathbb{F} \to \mathbb{F}$ *sa sljedećim svojstvima:*

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 $\mathbb{F}$.*

Napomena 1.1.5. *Djelovanje svake takve funkcije A piše se tablično, u m redaka i n stupaca gdje se u i-ti i j-ti stupac piše funkcijsku vrijednost $A(i, j)$. U tom smislu kažemo da je A matrica s m redaka i n stupaca. Uobičajeno se ta funkcijska vrijednost $A(i, j)$ označava kao a_{ij}.*

$$A_{m,n} = \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,1} & a_{m,2} & \cdots & a_{m,n} \end{pmatrix}$$

Definicija 1.1.6. *Neka je V vektorski prostor nad poljem $\mathbb{F}$.* ***Skalarni produkt*** *na V je preslikavanje $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{F}$ sa sljedećim svojstvima:*

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

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

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

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

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

Napomena 1.1.7. *U* $\mathbb{R}^n$ *kanonski skalarni produkt definiran je s*

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

Definicija 1.1.8. *Vektorski prostor na kojem je definiran skalarni produkt zove se* ***unitarni prostor****.*

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

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

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

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

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

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

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

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

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

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

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

Definicija 1.1.13. *Neka je V 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 1.1.14. *Metrika na normiranom prostoru ima sljedeća svojstva:*

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

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

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

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

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

Definicija 1.1.16. *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 1.1.12, 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 1.1.17. *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 1.1.18. *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 Statistika

Slučajna varijabla

Definicija 1.2.1. *Neka je* $(\Omega, \mathcal{F}, P)$ *vjerojatnosni prostor. Funkcija* $X : \Omega \to \mathbf{R}$ *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 1.2.2. *Neka je* $(\Omega, \mathcal{F}, P)$ *vjerojatnosni prostor. 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 1.2.3. *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.

Propozicija 1.2.4. *Neka je* $X : \Omega \to \mathbf{R}^n$*,* $X = (X_1, X_2, ... X_n)$*. Tada je X slučajan vektor ako i samo ako je* X_k *slučajna varijabla za svaki* $k = 1, 2, ..., n$.

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

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 1.2.6. ***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. Tada postoji rastući niz $(X_n)_{n\in\mathbb{N}}$ nenegativnih jednostavnih slučajnih varijabli takav da je $X = \lim_{n\to\infty} X_n$.

Definicija 1.2.7. ***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 1.2.8. *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 je po definiciji*

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

Definicija 1.2.9. *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 1.2.10. *Pozitivan drugi korijen iz varijance nazivamo* ***standardna devijacija*** *i označavamo sa* σ_X.

Funkcija distribucije

Definicija 1.2.11. *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 1.2.12. *Ako je jasno o kojoj se slučajnoj varijabli radi, piše se F umjesto* F_X.

Teorem 1.2.13. *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 1.2.14. *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 1.2.15. *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 1.2.16. *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 $X \sim N(\mu, \sigma^2)$.

Opisna statistika

S obzirom na problem kojim se bavi ovaj rad potrebno je navesti i pojmove kao što su aritmetička sredina, standardna devijacija uzorka, varijanca uzorka, medijan, mod i standardizacija podataka.

Neka je

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

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

Aritmetička sredina podataka ili uzorka (1.3) definirana je kao:

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

Vrijednosti (1.3) možemo urediti:

$$x_{(1)} \leq\ x_{(2)} \leq \ldots \leq\ x_{(n)}$$

Medijan skupa podataka (1.3) je vrijednost od X za koju vrijedi da je 50% svih podataka u skupu manje ili jednako toj vrijednosti, a 50% svih podataka je veće ili jednako.

Mod je vrijednost obilježja X koja se u skupu (1.3) pojavljuje najviše puta, odnosno ima najveću frekvenciju.

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 K-means algoritam

K-means clustering je proces grupiranja n točaka u k klastera u kojem svaka točka pripada klasteru s najbližim centrom. Neka su podaci reprezentirani skupom vektora

$$X = \{x_1,\ x_2, \ldots,\ x_n\} \subseteq \mathbb{R}^n \tag{1.5}$$

Za proizvoljan $k \in \mathbb{N}$ neka je $\{C_1, \ldots, C_k\}$ k klastera, a $c_1, \ldots, c_k$ su pripadni centri. Cilj algoritma je pronaći optimalnu k-particiju skupa X. To se postiže minimizacijom funkcije cilja f.

Definicija 1.3.1. *Neka je c_i centar klastera C_i. Funkcija cilja je definirana sa*

$$f(C_1, \ldots, C_k, c_1, \ldots, c_k) = \sum_{i=1}^{k} \sum_{x \in C_i} d^2(x, c_i), \tag{1.6}$$

gdje je $d(\cdot, \cdot)$ Euklidska udaljenost.

Poglavlje 2

Bioinformatika

2.1 Biološki pojmovi

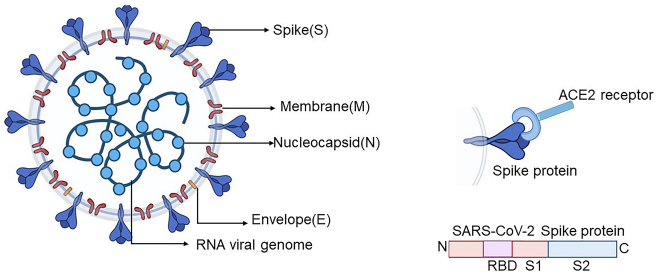
Proteini ili bjelančevine su, uz vodu, najvažnije tvari u tijelu. Sastavni su dijelovi svake stanice, a to ih čini osnovom života. Izgrađene su od aminokiselina koje su međusobno povezane peptidnom vezom. U kemijskom smislu, aminokiseline se definiraju kao molekule koje sadrže amino skupinu, karboksilnu skupinu i bočni lanac prema kojem se međusobno razlikuju. Međusobno su povezane u makromolekule proteina. U proteinima se može naći dvadeset različitih vrsta aminokiselina, svaka označena velikim slovom engleske abecede. Prikazane su u tablici 2.1

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

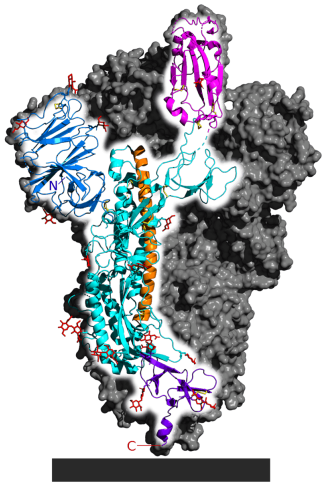
Tablica 2.1: Standardne aminokiseline

Protein šiljka

S protein ili protein šiljka (eng. *spike protein*) je glikoprotein koji tvori strukturu koja se nalazi na površini virusa, što pomaže virusu da se veže za stanicu u tijelu i uđe u nju. Najveći je od četiri strukturna proteina pronađena u koronavirusima. Preostala tri proteina su protein E, protein N te protein M. Proteini šiljka formiraju trimere te imaju prepoznatljiv oblik krune. S protein se sastoji od linearnog lanca koji sadrži 1273 aminokiseline. Tijekom pandemije COVID-19, genom SARS-CoV-2 virusa sekvencioniran je mnogo puta, što je rezultiralo identifikacijom tisuća različitih varijanti.



Slika 2.1: Shematski prikaz koronavirusa SARS-CoV-2



Slika 2.2: Mikroskopski prikaz trimera S proteina

Poravnanje proteina

Mutacija je trajna promjena genskog materijala. Evolucijski procesi su mutacijski događaji na slučajnom mjestu u proteinskom nizu. Ti događaji obuhvaćaju dodavanje jedne ili više aminokiselina proteinu, zamjenu jedne aminokiseline drugom te izostavljanje jedne ili više aminokiselina u proteinu.

Poravnanje nizova je jedan od najtočnijih prikaza evolucijskih procesa. Najjednostavniji način kojim se može odrediti jesu li dva proteina povezana je poravnavajući njihove nizove aminokiselinskih ostataka. Poravnate sekvence se tipski prikazuju kao redovi u matrici. Praznine se umeću između ostataka tako da se identična ili slična slova poravnavaju.

2.2 Prikaz aminokiselina u vektorskom prostoru

Nedostatak prirodne metrike za usporedbu nizova sastavljenih od slova sprečava statističke analize takvih podataka. Iz toga razloga je potrebno opisati aminokiseline numeričkim vrijednostima. Taj problem je opisan i riješen u članku [1]. Definirano je preslikavanje koje svakoj aminokiselini pridruži 5-dimenzionalni vektor. Ovakvo preslikavanje "čuva" sve bitne fizikalno-kemijske informacije o aminokiselini. Svaka koordinata vektora odgovara jednom svojstvu ili kombinaciji više njih. Te koordinate nazivamo faktorima. *Faktor I* predstavlja polaritet, *Faktor II* je faktor sekundarne strukture, *Faktor III* predstavlja molekularni volumen ili veličinu aminokiseline, *Faktor IV* odražava raznolikost kodona (relativnu kompoziciju aminokiselina u različitim proteinima) te *Faktor V* označava elektrostatski naboj aminokiseline.

Pojmovi i slike iz ovog poglavlja preuzeti su iz izvora [1], [4], [6], [8] [10], [11] i [12].

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

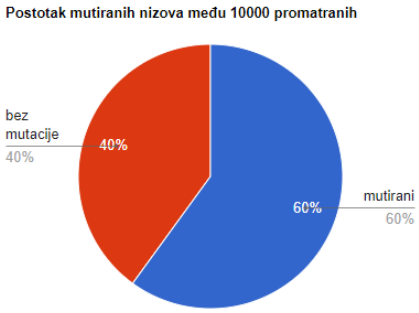
Analiza podataka i k-means algoritam

3.1 Analiza problema i opis podataka

U ovom diplomskom radu je promatrana datoteka sa 10000 zapisa proteina S koji su odabrani iz veće datoteke koja sadrži oko 20000 nizova proteina. Prilikom odabira se pazilo da svaki zapis bude duljine 1273 te da ne sadrži niti jedno slovo koje nije oznaka jedne od 20 aminokiselina. Dakle, odbačeni su svi zapisi koji su imali delecije ili insercije. Na taj je način dobivena datoteka sa višestruko poravnatim nizovima. Potom je na tako odabranim podacima napravljena transformacija svake od aminokiselina u 5-dimenzionalni vektor na način kako je objašnjeno u 2.2. Tako su promatrani podaci prebačeni u vektorski prostor. S obzirom da se radi o nizovima duljine 1273, promatrani podaci su prikazani kao 6365-dimenzionalni vektori. Radi lakše analize podataka, nizovi su spremljeni u matrice od 1273 retka i 5 stupaca.

Na svakoj od 1273 pozicije je promatrano koja aminokiselina se pojavljuje najveći broj puta te je tako za svaku poziciju dobivena dominantna aminokiselina. Među 10000 nizova pronađen je niz koji se sastoji od isključivo dominantnih aminokiselina, odnosno za uzorak je pronađen mogući niz predak. Usporedbom pronađenog niza s preostalim proteinima dobiven je broj pozicija na kojima su se dogodile promjene. Izračunato je da se na 496 pozicija dogodila mutacija unutar 6001 niza. Pozicije s promjenama su se mogle izračunati i preko matrice varijance, tako da se prvo napravi matrica srednjih vrijednosti po svakoj poziciji pojedinačno. Na mjestima gdje je varijanca različita od 0 dogodila se neka promjena.

Podaci su preuzeti iz izvora [13].

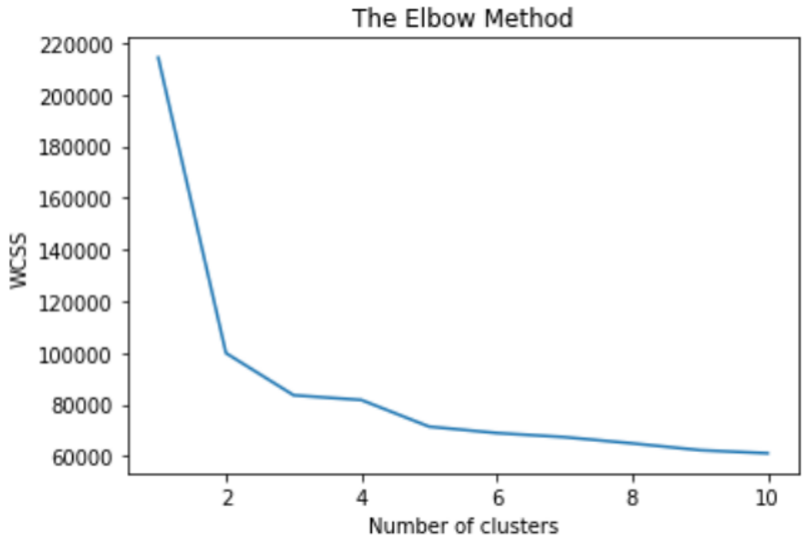


Slika 3.1: Postotak mutiranih nizova

3.2 K-means algoritam

Prethodnom analizom proteina dobiveni su podaci koji su spremni za primjenu k-means algoritma. Cilj je dobiti odgovor na pitanje postoji li pouzdan način grupiranja podataka u promatranom uzorku.

Procjena broja mogućih klastera je napravljena pomoću metode lakta, objašnjene u [8]. Rezultati te metode prikazani su na slici u nastavku.



Slika 3.2: Metoda lakta

Sa slike se vidi da se ne može sa sigurnošću zaključiti je li podjela u 2 klastera prirodna. U nastavku će se analizirati što se događa prilikom grupiranja u 2 klastera te ima li smisla gledati takvu podjelu.

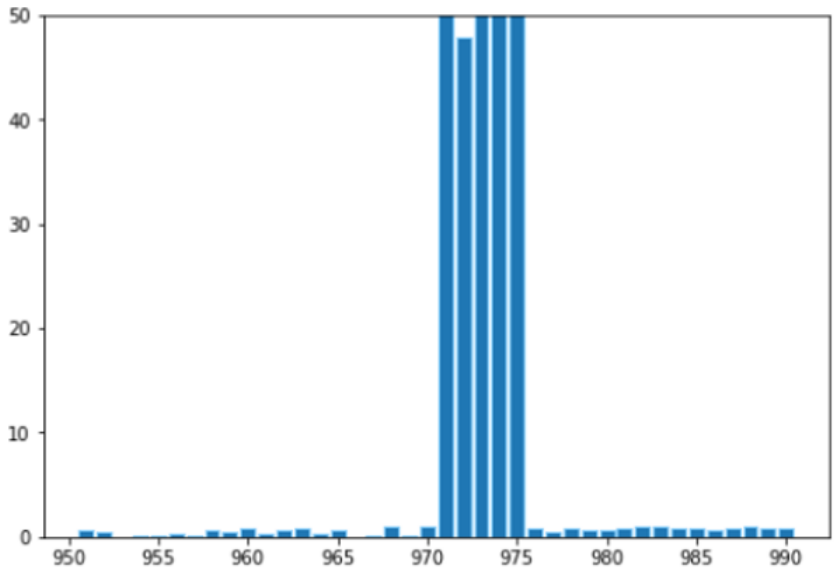
Određivanje značajnih pozicija

Radi jednostavnosti računa, promatrati će se samo proteini u kojima su se dogodile mutacije, njih 6001. Napomena, takav način promatranja može utjecati na rezultate daljnje analize.

Koristeći k-means++ algoritam uz predodređen broj klastera, što je u ovom slučaju 2, nizovi su podijeljeni u klastere. Prvi sadrži 3241 proteina, a drugi 2760. Rangiranje koje pokazuje koja je najznačajnija aminokiselina za promatranu podjelu napravljeno je preko omjera. Omjer predstavlja statistiku kojom se mjeri je li koordinata bolje opisana s jednim ili dva centra, a postupak se ponavlja dva puta. Prvo se radi po svakoj koordinati posebno, a zatim po nizu od 5 koordinata.

$$O(j) = \frac{\sum_{i=1}^{br}(x_{i,j} - \overline{x_j})^2}{\sum_{k_1 \in K_1}(x_{k_1,j} - \overline{x_{j,k_1}})^2 + \sum_{k_2 \in K_2}(x_{k_2,j} - \overline{x_{j,k_2}})^2}, j = 1, 2, ..., l \qquad (3.1)$$

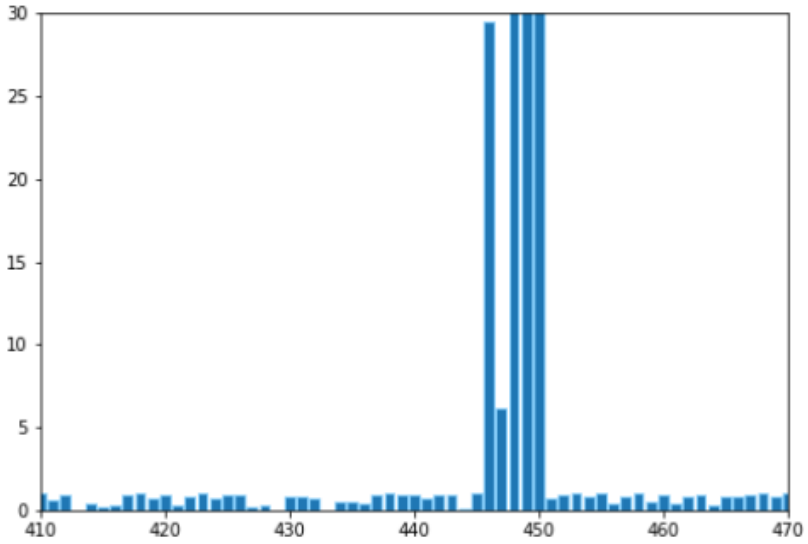
U formuli (3.1) l predstavlja duljinu vektora, br broj vektora koji su sudjelovali, K_1 prvi klaster, a K_2 drugi klaster. Tako je dobiven omjer za sve koordinate vektora S proteina. Slijedi prikaz rangiranja koordinata na kojem se vide najznačajnije koordinate.



Slika 3.3: Najznačajnije koordinate u uzorku

Na prethodnoj slici vidljivo je da je omjer najveći u 971., 972., 973., 974. i 975. koordinati. Dakle, tih 5 koordinata je najviše utjecalo kod podjele u dva klastera. S obzirom da je svaka aminokiselina prikazana kao 5-dimenzionalni vektor, nakon što se rezultat podijeli s 5 dobije se pozicija aminokiseline. Radi se o 195. po redu od mutiranih pozicija, odnosno gledajući sve pozicije to je 477. aminokiselina. Oni nizovi koji nisu mutirali na toj poziciji imaju aminokiselinu S.

Ponavljanjem prethodnog postupka na podskupu koji sadrži 3000 proteina od ukupnog skupa od 10000 proteina dobiveno je da su se mutacije dogodile u 1256 nizova na 358 različitih pozicija. K-means algoritmom su podaci podijeljeni u klastere tako da su u prvom 82 proteina, a u drugom 1174. Slijedi prikaz rangiranja koordinata za ovaj poduzorak.



Slika 3.4: Najznačajnije koordinate u poduzorku

Na prethodnoj slici vidljivo je da je omjer najveći u 446., 447., 448., 449. i 450. koordinati. To znači da se radi o 90. poziciji unutar niza mutiranih pozicija, odnosno gledajući sve pozicije radi se o 253. aminokiselini. Oni nizovi koji nisu mutirali na toj poziciji imaju aminokiselinu D.

Uzimanjem poduzorka zaključeno je ono što je i pretpostavljeno analizom slike (3.2). K-means algoritam će uvijek podijeliti zadani skup podataka u traženi broj klastera (u ovom slučaju 2), međutim nije pronađena pozicija za koju bi se moglo zaključiti da je značajna prilikom podjele. Dakle, kao što je već napisano, metodom lakta ne može se zaključiti postoje li aminokiseline koje stabilno dijele uzorak na određeni broj klastera.

Poglavlje 4

Analiza problema i algoritam

4.1 Opis problema pretrage značajnih grupa

Iako metodom k-means nije dobivena dominantna podjela podataka, svejedno će se pokušati pronaći postoje li klasteri koji su u korespondenciji s nekim varijantama virusa. Prikazat će se podatke u 2-dimenzionalnom prostoru i iz vizualnog prikaza će se pokušati pronaći mogući značajni klasteri te ukoliko se pronađu pogledat će se koje su karakteristične mutacije za njih. Nadalje, za odgovarajuće mutacije će se pogledati je li neka od njih nova bitna mutacija koja nije do sada pronađena.

Standardizacija podataka

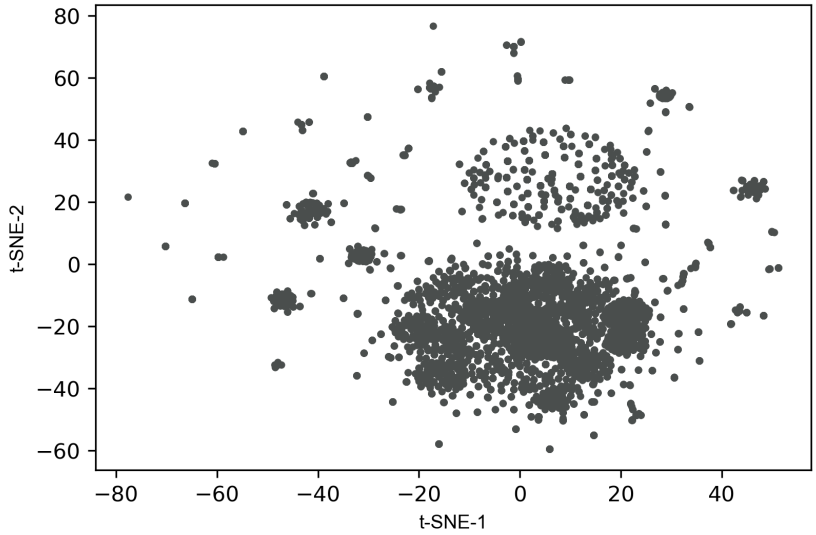
Kako je svaka aminokiselina prikazana kao 5-dimenzionalni vektor, promatraju se podaci u 6365-dimenzionalnom vektorskom prostoru. Ideja je promatrati euklidske udaljenosti među proteinima, odnosno mjeriti udaljenosti po pojedinim koordinatama i zbrajati ih. Iz tog razloga je potrebno izbjeći da varijanca i raspon podataka po jednoj koordinati budu veći od ostalih koordinata. Točnije, treba izbjeći dominiranost euklidske udaljenosti takvom koordinatom jer se na taj način gubi forma kugle u kojoj bi sve koordinate trebale imati jednaki utjecaj. Taj problem je riješen standardizacijom podataka, objašnjenom u (1.4). Standardizirani podaci su po svakoj koordinati normalno distribuirani s očekivanjem 0 i standardnom devijacijom 1. Na ovaj način su dobiveni podaci u kojima je utjecaj svih koordinata jednak te se može nastaviti sa daljnjom analizom.

Prikaz podataka u 2D

Kako nije moguće nacrtati podatke u 6365-dimenzionalnom koordinatnom sustavu, niti ih zamisliti kao takve, koristit će se redukcija dimenzije podataka. U paketu *Scikit-Learn*,

programskog jezika *Python*, postoji funkcija *t-SNE* koja provodi statističku proceduru *t-distributed Stohastic Neighbor Embedding*. To je nelinearna metoda smanjivanja dimenzije podataka koja čuva lokalnu strukturu. Ona preslikava podatke u manje dimenzionalan prostor na način da čuva okolinu i susjedstvo svake točke. To je točno ono što zadovoljava početnu ideju, gledanje susjedstva, grupiranja i udaljenosti. Odnosno, ako su dva proteina blizu jedan drugome u 6365-dimenzionalnom prostoru, onda će biti i u 2-dimenzionalnom. Više o metodi i njenim primjenama je moguće pronaći u [5].

Koristeći prethodno definiranu funkciju dobivena je slika podataka u 2-dimenzionalnom koordinatnom sustavu prikazana u nastavku.



Slika 4.1: t-SNE prikaz, standardizirani podaci

Sa prethodne slike uočeno je kako ima smisla promatrati odgovaraju li male grupice podataka, koje su se formirale oko dviju većih grupa, traženim značajnim klasterima. Odnosno, ima smisla pogledati koji su to podaci unutar grupica i zbog koje promjene su se oni grupirali na takav način. Vidljivo je i kako ima smisla zamisliti dvije velike kugle s velikom većinom podataka i puno manjih kuglica koje okružuju te dvije i koje su bliže rubovima slike. Kako je slika dobivena *t-SNE* algoritmom koji čuva strukturu podataka, slijedi da bi jednake zakonitosti vjerojatno vrijedile i u većim dimenzijama.

4.2 Pronalazak značajnih grupa

S obzirom na prethodni vizualni prikaz ima smisla tražiti kuglice koje će u konačnici biti u korespondenciji s nekim mutacijama. Proizvoljno je odabrano nekoliko mutacija te je promatrano koliko još ima proteina s takvom mutacijom, koliko njih su međusobno jednaki u potpunosti, a koliko njih ima još neku drugačiju mutaciju. Na grafičkim prikazima proteini s odabranom mutacijom su označeni drugom bojom. Izračunat je radijus kugle u kojoj se nalaze nizovi s odabranom mutacijom. Provjereno je i upada li neki od proteina koji nema takvu mutaciju u tu kuglu. Taj postupak je napravljen na način da su se promatrale euklidske udaljenosti među svakim od proteina koji su bili u blizini promatrane kugle s proteinima koji se nalaze u kugli. Svi rezultati i promatranja su popraćeni zaključcima. Dodatno su mutirani proteini promatrani kao jedan klaster, a ostali kao drugi te je napravljen logaritamski omjer opisan u 3.2 kako bi se pokazalo da se oni stvarno dijele na temelju odabrane mutacije i kako bi se pogledalo postoji li još neka mutacija koja utječe na takvu podjelu.

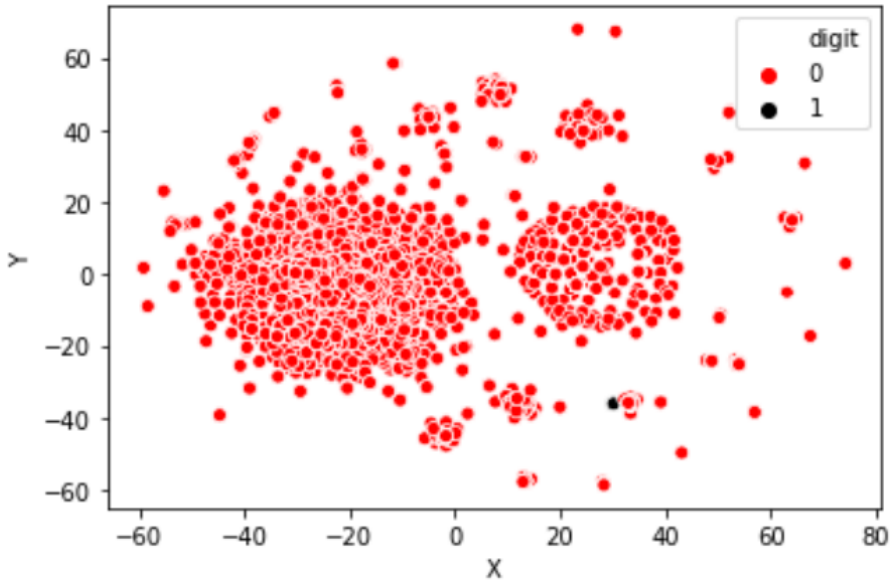
Pojmovi iz ovog poglavlja preuzeti su iz izvora [14].

Rezultati

Za analizu su odabrane sljedeće mutacije:

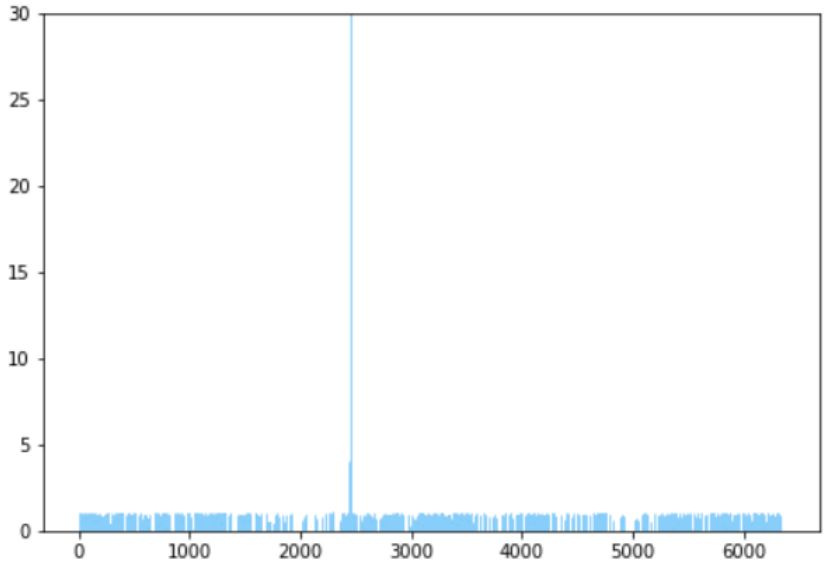
- H146Y
- F490S

Prvo je promatrana mutacija na poziciji 490 i to ona u kojoj se na toj poziciji umjesto aminokiseline F našla aminokiselina S. Uočeno je da postoji 7 različitih proteina u promatranom zapisu od 10000 koji imaju takvu mutaciju. Od njih sedam tri su jedinstvena i imaju i neke druge mutacije, dok su preostali jednaki jednom od ta tri. Na slici 4.2, dobivenoj *t-SNE* algoritmom, svi proteini s promatranom mutacijom su prikazani crnom bojom, a preostali proteini su prikazani crvenom bojom.



Slika 4.2: t-SNE prikaz proteina s mutacijom F490S

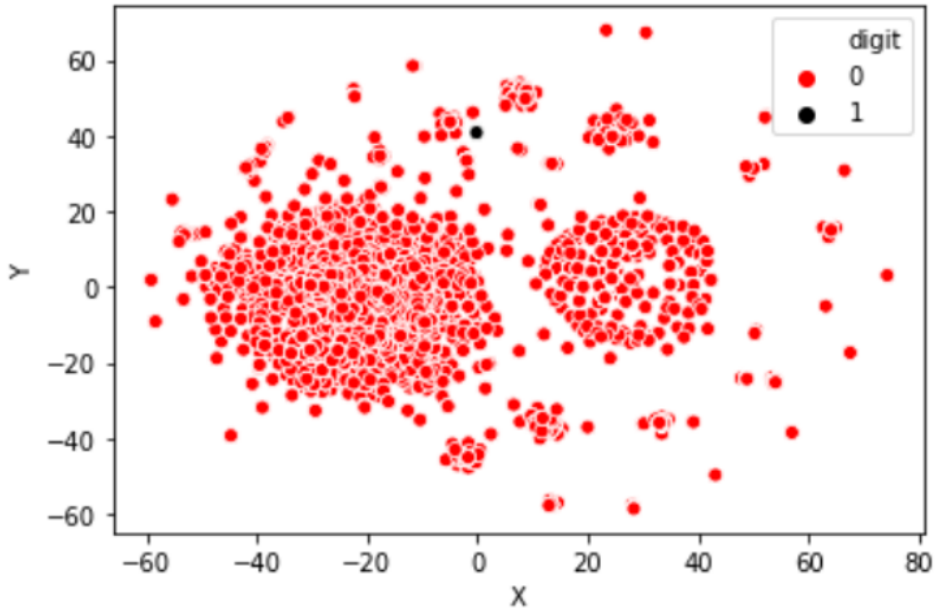
Na slici je vidljivo da su se proteini s promatranom mutacijom grupirali i da čine jednu "kuglicu". Računajući euklidske udaljenosti među njima i uzimajući maksimalnu, dobiveno je da je radijus kugle koju oni formiraju jednak 8.06. Provjereno je i upada li neki od proteina koji nema tu mutaciju unutar te kugle i dobiveno je da ne upada niti jedan drugi, što i odgovara slici. Zatim su podaci podijeljeni u klastere na način da su ovih 7 s mutacijom unutar jednog klastera, a preostali unutar drugog. Napravljen je i logaritamski omjer kojim se pokušalo pronaći postoji li još neka aminokiselina koja je zaslužna za takvu podjelu u 2 klastera.



Slika 4.3: Najznačajnije koordinate za promatrane klastere

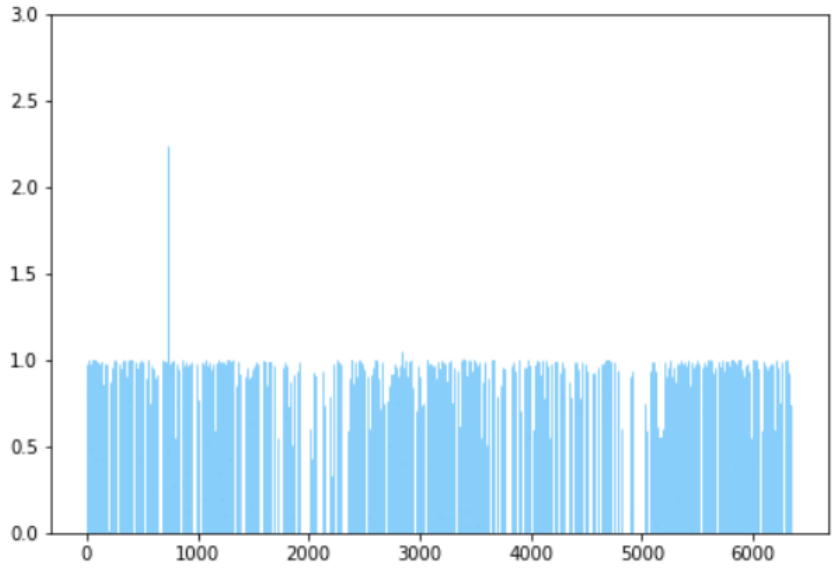
Sa slike je vidljivo da postoji samo jedna mutacija odgovorna za ovakvu podjelu i to je upravo ona na poziciji 490.

Ista analiza je ponovljena za mutaciju na poziciji 146 na kojoj se umjesto aminokiseline H našla aminokiselina Y. Uočeno je da postoji 11 različitih proteina koji imaju takvu mutaciju, a 6 ih je jedinstveno. Na slici 4.4 dobivenoj *t-SNE* algoritmom, crnom bojom su prikazani svi ti proteini s promatranom mutacijom, a preostali proteini su crvene boje.



Slika 4.4: t-SNE prikaz proteina s mutacijom H146Y

Na slici je vidljivo da su se proteini s promatranom mutacijom grupirali. Računajući je dobiveno da je radijus kugle koju oni formiraju jednak 4.25. Provjereno je i upada li neki od proteina koji nema tu mutaciju unutar te kugle i dobiveno je da ne upada niti jedan drugi. Zatim su proteini podijeljeni u klastere na način da je njih 11 sa promatranom mutacijom unutar jednog klastera, a preostali unutar drugog. Napravljen je logaritamski omjer kojim se pokušalo pronaći postoji li još neka aminokiselina koja je zaslužna za takvu podjelu u 2 klastera.



Slika 4.5: Najznačajnije koordinate za promatrane klastere

Sa slike je vidljivo da postoji samo jedna mutacija odgovorna za ovakvu podjelu i to je upravo ona na poziciji 146.

S obzirom da su pronađene kugle koje su u korespondenciji s promatranim mutacijama te da su one udaljene od najveće grupe podataka ima smisla zapitati se jesu li te mutacije značajne. Provjerom na stranicama s objavljenim mutacijama na spike proteinu, [15], te mutacije su pronađene kao česte mutacije.

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` (veljača 2022.).

[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` (veljača 2022.).

[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, dostupno na `https://academic.oup.com/bioinformatics/article/35/18/3491/5306940` (veljača 2022.).

[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` (veljača 2022.).

[10] L. M., Q. S., *Antibodies and Vaccines Target RBD of SARS-CoV-2*, dostupno na `https://www.frontiersin.org/articles/10.3389/fmolb.2021.671633/full` (veljača 2022.).

[11] *Coronavirus spike protein*, dostupno na `https://en.wikipedia.org/wiki/Coronavirus_spike_protein` (veljača 2022.).

[12] L. J. Catania, *Foundations of Artificial Intelligence in Healthcare and Bioscience*, 2021., dostupno na https://www.sciencedirect.com/topics/engineering/spike-protein (veljača 2022.).

[13] *GISAID database* dostupno na https://www.gisaid.org/ (veljača 2022.).

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

[15] *Spike mutations* dostupno na https://github.com/cov-lineages/constellations/blob/main/constellations/misc/spike_mutations.csv (veljača 2022.).

Sažetak

Ovaj diplomski rad bavi se analizom nizova proteina S iz SARS-CoV-2 koronavirusa, a za cilj ima pogledati postoji li pouzdan način grupiranja promatranog uzorka te pronaći koje su mutacije karakteristične za tako dobivene grupe. Sve analize potrebne za zaključke su se provodile nad aminokiselinama prikazanim kao vektorima.

Na početku su navedeni matematički i biološki pojmovi potrebni za razumijevanje ovog rada te je uvedena struktura podataka na kojima se provodila analiza. Zatim je proveden k-means algoritam nad pripremljenim podacima. Rezultat nije dao pouzdan način grupiranja podataka te se dolazi do zaključka da uzorak nije reprezentativan za ciljeve filogenetske analize. Iako nije pronađena dominantna podjela svejedno je nastavljena analiza kako bi se saznalo postoje li neki značajni klasteri te koje su mutacije karakteristične za njih. Prikazom u 2-dimenzionalnom prostoru pronađene su grupe podataka koje su geometrijski promatrane kao kuglice koje su bile udaljene od većine podataka, odnosno od dviju većih kugli. Odabrane su dvije od tih kuglica, jedna koja je sadržavala sve proteine s mutacijom H146Y i druga koja je sadržavala proteine s mutacijom F490S. S obzirom da su se ti podaci grupirali i izdvojili od ostalih, provjereno je i potvrđeno da se to dogodilo isključivo zbog spomenutih mutacija. U konačnici je provjereno jesu li spomenute varijante virusa već postojeće i česte mutacije te se pokazalo da jesu.

Summary

This thesis deals with the analysis of S-proteins sequences from SARS-CoV-2 coronavirus. The aim of the analysis is to detect reliable clustering of the sample, and determine sequence positions that are relevant or important for this division into clusters. All analyses required for the conclusions were performed on the amino acid sequences represented as vectors in a suitable real vector space.

At the beginning, the mathematical and biological concepts needed to understand this paper are listed and the structure of the data on which the analysis was performed is introduced. After that, the k-means algorithm was applied. Results did not provide a reliable way of grouping the data and it is concluded that the sample is not representative. Since no dominant clustering was detected, visual analysis was carried out to detect any significant small clusters. Two such clusters were selected, one containing proteins with the H146Y mutation and the other containing proteins with the F490S mutation. Since these sequences were grouped and separated from the others, it was verified that this occurred solely due to the mentioned mutations. At the end, we successfully checked for these variants on the list of already described, frequent mutations.

Životopis

Rođena sam u Splitu, 05. prosinca 1995. godine. Školovanje započinjem u Osnovnoj školi Ostrog u Kaštel Lukšiću, nakon koje upisujem I. gimnaziju u Splitu. Po završetku srednjoškolskog obrazovanja upisujem preddiplomski studij matematike i informatike na Prirodoslovno-matematičkom fakultetu u Splitu. Nakon završenog prediplomskog studija, 2019. godine upisujem diplomski sveučilišni studij Matematičke statistike na Prirodoslovno-matematičkom fakeltute u Zagrebu.