# Analiza proteinskih nizova iz CoViD-a 19

**Tušek, Helena**

# Analiza proteinskih nizova iz CoViD-a 19

**Tušek, Helena**

**SVEUČILIŠTE U ZAGREBU**

**PRIRODOSLOVNO–MATEMATIČKI FAKULTET**

**MATEMATIČKI ODSJEK**

Helena Tušek

# ANALIZA PROTEINSKIH NIZOVA IZ COVID-A 19

Diplomski rad

Voditelj rada:
doc.dr.sc. Pavle Goldstein

Zagreb, veljača, 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. _____

*Na početku, željela bih izraziti zahvalnost mentoru doc.dr.sc. Pavlu Goldsteinu na razumijevanju, strpljenju i trudu koji je uložio kako bi mi pomogao prilikom pisanja ovog diplomskog rada. Hvala mojoj obitelji bez čijih: "Neno, ti to možeš" nikada ne bih položila sve što je trebalo položiti. Hvala za svaku riječ podrške, svaku vožnju do fakulteta i svaki zagrljaj za koji nisam ni znala koliko mi je potreban. I šećer na kraju, jer oni su to bili svih ovih godina, željela bih zahvaliti i prijateljima. Karlo, hvala na svemu! Svaki pad bio je lakši s tobom, a svaki prolaz još ljepši. Hvala onima koji su kao "moji" krenuli u ovo putovanje sa mnom, a i onima koji su "moji" postali tijekom studiranja. Veliko hvala svima, razvedrili ste i najtmurnije dane.*

*Deda, ovo će biti naša diploma!*

# Sadržaj

# Uvod

Krajem 2019. godine, u Kini, gradu Wuhanu, pojavio se virus SARS- COV- 2 (Sars-coronavirus-2) koji uzrokuje bolest COVID-19. Nedugo zatim, virus se počeo širiti velikom brzinom i par mjeseci kasnije, rijetko koja država nije bilježila svoje prve slučajeve. Nije bilo dovoljno što je bolest bila velika nepoznanica za cijelo čovječanstvo, ubrzo je i sam virus počeo mutirati i bilježiti još veću brzinu širenja. Koronavirus je RNA virus kuglastog oblika. Sadrži četiri do šest strukturnih polipeptida, a među njima su značajna četiri: E, N, M i S. Polipeptidi su jednostruki linearni polimerni lanci aminokiselina koje su povezane peptidnim vezama. Ovaj će se rad baviti detaljnom analizom četiri navedena proteina. Cilj rada je primijeniti tehnike strojnog učenja na analizu višestrukih poravnanja proteina koronavirusa. Odnosno, primjenom tehnika klasteriranja i analize pozicija pokušavaju se otkriti najznačajnije, to jest, dominantne mutacije u aminokiselinskom nizu. Za očekivati je da će značajne pozicije (otkrivene na taj način) imati utjecaj na samu virulentnost i smrtnost koronavirusa. Rad se sastoji od triju poglavlja. U prvom će poglavlju biti navedeni pojmovi iz linearne algebre, statistike i strojnog učenja koji će biti potrebne za razumijevanje ostatka rada. U drugom će se poglavlju opisati struktura podataka i bit će objašnjeno kako podatke prilagoditi algoritmima strojnog učenja, a u ovom slučaju metodi klasteriranja (grupiranja). Konačno, treće se poglavlje bavi analizom rezultata. Odnosno, bit će objašnjeno je li evolucija drugačije djelovala na četiri spomenuta proteina, koje mutacije su dobivene i što nam to znači za sam virus. Program analize proteinskih nizova rađen je u programskom jeziku Python, dok su neke procjene rađene u programu R.

# Poglavlje 1

# Matematički pojmovi

U ovom se poglavlju navode neke definicije, propozicije i napomene iz linearne algebre, statistike i strojnog učenja. Pojmovi iz linearne algebre prate [2], iz statistike [5] i [6], a iz područja strojnog učenja [4] i [8].

## 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 iduća svojstva:*

1. $\alpha + (\beta + \gamma) = (\alpha + \beta) + \gamma$;
2. $\exists 0 \in \mathbb{F}$ *sa svojstvom* $\alpha + 0 = 0 + \alpha = \alpha, \forall \alpha \in \mathbb{F}$;
3. $\forall \alpha \in \mathbb{F}, \exists - \alpha \in \mathbb{F}$ *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. $\exists 1 \in \mathbb{F} \setminus \{0\}$ *sa svojstvom* $1 \cdot \alpha = \alpha \cdot 1 = \alpha$, $\forall \alpha \in \mathbb{F}$;

*7.* $\forall \alpha \in \mathbb{F},\ \alpha \neq 0, \exists \alpha^{-1} \in \mathbb{F}$ *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* $\mathbb{F}$ *polje, a elemente polja nazivamo skalarima.*

**Napomena 1.1.2.** *Skup realnih brojeva 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.* $\exists 0 \in V$ *sa svojstvom* $a + 0 = 0 + a = a, \forall a \in V$*;*

*3.* $\forall a \in V, \exists -a \in V$ *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*

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

*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. Običaj je da 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:*

$$\langle \cdot, \cdot \rangle : V \times V \to \mathbb{F}$$

*koje ima sljedeća svojstva:*

1. $\langle x, x \rangle \geq 0, \forall x \in V;$
2. $\langle x, x \rangle = 0 \Leftrightarrow x = 0, \forall x \in V;$
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. $\langle x, y \rangle = \overline{\langle y, x \rangle}, \forall x, y \in V.$

**Definicija 1.1.7.** *Neka je V vektorski prostor nad poljem* $\mathbb{F}$ *s definiranim skalarnim produktom. Tada V nazivamo unitarnim prostorom.*

**Definicija 1.1.8.** *Euklidski prostor je unitaran realni prostor.*

**Napomena 1.1.9.** *U* $\mathbb{R}^n$ *skalarni produkt je obično definiran s*

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

**Definicija 1.1.10.** *Neka je $V$ vektorski prostor nad $\mathbb{F}$ i $M \subseteq V, M \neq \emptyset$. Ako je i $(M, +, \cdot)$ vektorski prostor nad $\mathbb{F}$ uz iste operacije iz $V$, kažemo da je $M$ potprostor od $V$.*

**Definicija 1.1.11.** *Neka je $V$ unitaran prostor. Norma na $V$ je funkcija*

$$\|\cdot\| : V \to \mathbb{R}$$

*definirana s*

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

**Propozicija 1.1.12.** *Norma na unitarnom prostoru $V$ ima sljedeća svojstva:*

1. $\|x\| \geq 0, \forall x \in V$;
2. $\|x\| = 0 \Leftrightarrow x = 0, \forall x \in V$;
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 1.1.13.** *Svaka funkcija $\|\cdot\| : V \to \mathbb{R}$ na vektorskom prostoru $V$ sa svojstvima iz 1.1.12 naziva se norma. Tada $(V, \|\cdot\|)$ zovemo normirani prostor.*

**Definicija 1.1.14.** *Neka je $V$ unitaran 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 1.1.15.** *Metrika na unitarnom 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 1.1.16.** *Neka je $X \neq \emptyset$. Svaka funkcija $d : X \times X \to \mathbb{R}$ sa svojstvima iz 1.1.15 naziva se metrika ili udaljenost. Tada $(X, d)$ zovemo metrički prostor.*

## 1.2 Statistika

**Definicija 1.2.1.** *Neka je* $(\Omega, \mathcal{F}, P)$ *vjerojatnosni prostor. Funkcija* $X : \Omega \to \mathbf{R}$ *je slučajna varijabla (na* $\Omega$*) 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* $\Omega$*) 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 $x_1,\ x_2, ...,\ x_n$ realni brojevi, a $A_1,\ A_2, ...,\ A_n$ međusobno disjunktni događaji, $\bigcup_{k=1}^{n} A_k = \Omega$.

**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 $\Omega$, relacijama:

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

i

$$(X_1 \wedge X_2)(\omega) = min\{X_1(\omega),\ X_2(\omega)\},\ \omega \in \Omega \tag{1.2}$$

Pomoću funkcije (1.1) definiramo pozitivan i negativan dio realne funkcije $X$ na $\Omega$:

$$X^+ = X \vee 0,\ X^- = (-X) \vee 0. \tag{1.3}$$

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

$$X = X^+ - X^- \tag{1.4}$$

$$|X| = X^+ + X^-. \tag{1.5}$$

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

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 $(\Omega, \mathcal{F}, P)$ vjerojatnosni prostor. Sa $\mathcal{K}$ označimo skup svih jednostavnih slučajnih varijabli $\Omega$, 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 $x_1, x_2, ..., x_n$ realni brojevi, a $A_1, A_2, ..., A_n$ međusobno disjunktni događaji.

**Definicija 1.2.6.** *Matematičko očekivanje od X ili, kraće, očekivanje od X koje označavamo sa* $\mathbb{E}(X)$ *definira se s*

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

Neka je $X$ sada nenegativna slučajna varijabla definirana na $\Omega$. 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 s*

$$\mathbb{E}(X) = \lim_{n\to\infty} \mathbb{E}(X_n).$$

Konačno, možemo navesti definiciju matematičkog očekivanja za slučaj opće slučajne varijable.

**Definicija 1.2.8.** *Kažemo da matematičko očekivanje od X ili, kraće, očekivanje od X, koje označavamo sa* $\mathbb{E}X$ *postoji ili da je definirano ako je barem jedna od veličina* $\mathbb{E}\, X^+$ *ili* $\mathbb{E}\, X^-$ *konačna, tj. ako 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 vjerojatnosnom prostoru* $(\Omega, \mathcal{F}, P)$ *i* $r > 0$. $\mathbb{E}(X^r)$ *zovemo r-ti moment od X, a* $\mathbb{E}(|X^r|)$ *zovemo r-ti apsolutni moment od X.*

**Definicija 1.2.10.** *Neka* $\mathbb{E}X$ *postoji (tj. konačno je). Tada* $\mathbb{E}[(X - \mathbb{E}X)^r]$ *zovemo r-ti centralni moment od X, a* $\mathbb{E}[|X - \mathbb{E}X|^r]$ *zovemo r-ti apsolutni centralni moment od X.*

**Definicija 1.2.11.** *Varijanca od X koju označavamo sa VarX ili* $\sigma_x^2$ *je drugi centralni moment od X, dakle je*

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

**Definicija 1.2.12.** *Standardna devijacija od X koju označavamo sa $\sigma_X$ je pozitivan drugi korijen iz varijance.*

Ovo sve je bilo potrebno navesti jer će se ovaj diplomski rad baviti raspršenjem podataka te će biti izračunate aritmetička sredina i varijanca, za što su potrebne mjere lokacije, mjere raspršenja i navedeni momenti. Postoji mnogo različitih mjera centralnih tendencija skupova podataka. Navedene su tri najvažnije: medijan, mod i aritmetička sredina. Neka su

$$x_1,\ x_2, ...,\ x_n \tag{1.6}$$

n vrijednosti 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 brojeva (1.6) je broj:

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

Vrijednosti (1.6) možemo urediti:

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

Medijan skupa podataka (1.6) je vrijednost od $X$ za koju vrijedi da je 50% svih podataka u skupu manje ili jednako toj vrijednosti i 50% svih podataka je veće ili jednako joj.

Mod je vrijednost obilježja $X$ koja se u skupu (1.6) pojavljuje najviše puta, odnosno ima najveću frekvenciju.

Najčešća korištena mjera raspršenja skupa numeričkih podataka je standardna devijacija. Standardna devijacija je zapravo srednje kvadratno odstupanje podataka od njihove aritmetičke sredine, tj. zadana je formulom

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

Odnosno, iz prethodno navedenih definicija za varijancu i standardnu devijaciju slijedi da je varijanca zadana formulom

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

## 1.3 Strojno učenje

Danas je strojno učenje jedno od najaktivnijih područja računarske znanosti te postoji mnogo uspješnih primjena strojnog učenja. To su na primjer: otkrivanje znanja u velikim skupovima podataka, programske implementacije koje nije moguće riješiti klasičnim programiranjem, bioinformatika i mnogi drugi.

Strojno učenje može se definirati kao proces programiranja računala kako bi se moglo optimizirati izvođenje kriterija koristeći podatke ili stečeno iskustvo.

Glavna razlika između primijenjene statistike i samog strojnog učenja je u veličini skupa podataka na kojima se radi. Primijenjena statistika se obično primjenjuje na manjim skupovima dok je u strojnom učenju naglasak na velikoj količini podataka.

Ime nam dolazi iz činjenice da strojno učenje omogućava strojevima da uče, to jest, da postaju sve bolji u rješavanju problema u odnosu na "iskustvo". Kako stroj (računalo) uči na podacima to znači da što je veća količina podataka dostupna, to će računalo točnije riješiti problem. Glavna podjela strojnog učenja je na:

1. Nadzirano učenje (*Supervised Learning*)
2. Nenadzirano učenje (*Unsupervised Learning*)
3. Učenje s podrškom

Slijedi kratko objašnjenje razlike između tri navedena oblika učenja. Nadzirano učenje uzima eksplicitnu informaciju o primjerima i vrijednostima njihove ciljne varijable i cilj mu je napraviti model koji će raditi predikcije na još neviđenim (novim) primjerima. Pod nadzirano učenje spada klasifikacija, regresija te predikcija (*forecasting*).

Nenadzirano učenje uzima primjere bez ikakve anotacije ili povratne informacije o njihovoj kategorizaciji. Cilj mu je grupirati primjere, odnosno otkriti pravilnost među podacima i projekcija podataka u niže-dimenzionalne prostore. Pod nenadzirano učenje spada grupiranje (*clustering*), otkrivanje detekcija-iznimaka, te kompresija podataka.

Za razliku od nadziranog i nenadziranog učenja, u podržanom učenju nema označenih primjera dobrog i lošeg djelovanja, te čovjek sudjeluje u ovom načinu učenja time što dodjeljuje nagrade. Odnosno, čovjek određuje u kojim uvjetima te za što će sustav biti "nagrađen" i ono se najviše koristi uz učenje sekvenci uspjeha (roboti).

U ovom će diplomskom biti govora o nenadziranom strojnom učenju, to jest, k-means clusteringu koji dijeli $n$ točaka u $k$ klastera i u kojem svaka točka pripada klasteru s najbližim centrom.

## K-means algoritam

K-means algoritam je jednostavan, intuitivan proces koji služi za grupiranje podataka. Neka su podaci reprezentirani skupom vektora $X = \{ x_1,\ x_2,\ x_3, ...,\ x_n\} \subseteq \mathbb{R}^n$, a cilj postupka je pronaći optimalnu $k$-članu particiju $\{ c_1,\ c_2, ...,\ c_k\}$ skupa X. To se postiže minimizacijom funkcije cilja $f$, definirane u terminima klasterima $C_1,\ C_2, ...,\ C_k$ i središtima klastera $c_1,\ c_2, ...,\ c_k$.

Funkcija $f$ zadana je s:

$$f(\ C_1,\ C_2, ...,\ C_k,\ c_1,\ c_2, ...,\ c_k) = \sum_{i=1}^{k} \sum_{x \in\ C_i} d\ ^2(x,\ c_i),$$

gdje je $d(\cdot, \cdot)$ Euklidska udaljenost definirana s:

$$d(x, y) = (\sum_{i=1}^{m} (\ x_i -\ y_i)\ ^2)^{\frac{1}{2}},$$

a $x = (\ x_1, ...,\ x_m)$, $y = (\ y_1, ....,\ y_m) \in \mathbb{R}^m$.

Proces optimizacije k-means algoritma - također poznatom kao Loydov algoritam započinje tako da se za fiksni broj klastera $k$ i $X$ inicijalno odabere $k$ središta (centara) $c_1,\ c_2, ...,\ c_k$ i dalje se izvršava iteriranjem sljedeća dva koraka:

1. korak: pridruživanje svake točke x klasteru s najbližim središtem

$$C_i^{(t+1)} = \{x : d(x,\ c_i^{(t)}) \leq d(x,\ c_j^{(t)}), \forall j\}$$

2. korak: određivanje središta za novi raspored klastera

$$c_i^{(t+1)} = \frac{1}{|\ C_i^{(t)}|} \sum_{x \in\ C_i^{(t)}} x.$$

Ovdje $C_i^{(t)} i\ c_i^{(t)}$ označavaju $i$-ti klaster i $i$-to središte klastera u $t$-toj iteraciji, redom, dok $|\ C_i^{(t)}|$ označava veličinu skupa $C_i^{(t)}$. U drugom koraku postavljamo središte $c_i$ kao središte i-tog klastera $C_i$.

Početna konfiguracija "bilo particija bilo središte" može biti nasumično odabrana ili određena na neki drugi način. Algoritam se obično zaustavlja stabilizacijom particija ili nakon unaprijed određenog broja koraka. Kroz 1. i 2. korak se smanjuje vrijednost funkcije cilja $f$. Ako označimo s $f^{(i)}$ vrijednost funkcije $f$ u i-toj iteraciji, dobivamo padajući niz:

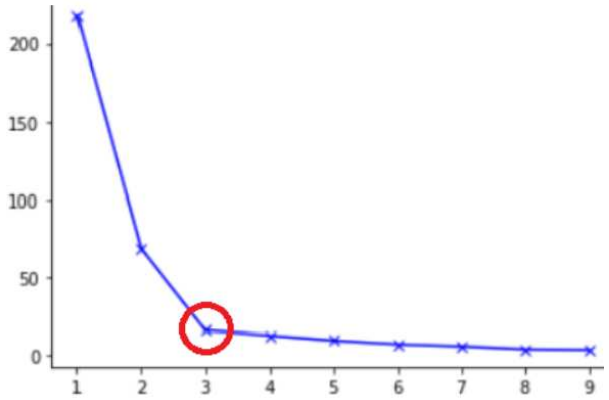
$$f^{(1)} \geq f^{(2)} \geq ... \geq 0.$$

To ukazuje da će algoritam dosegnuti minimum (od $f$) u konačnom broju koraka, ali i na to da će minimum često biti lokalan.

Ovaj problem očituje se kroz osjetljivost k-means algoritma na početne uvjete kao što su početni odabir klastera. U primjeni se to često rješava pokretanjem algoritma nekoliko puta s različitim početnim uvjetima i biranjem najboljeg rješenja. Druga mogućnost je da se na bolji način konstruiraju početni uvjeti.

Postoji unaprijeđena verzija k-means algoritma, a to je k-means++. Taj algoritam na pažljiviji način bira početna središta i tako rješava problem lokalnog minimuma pa je zbog toga korišten u ovom diplomskom radu. K-means++ algoritam ima kao cilj vidjeti kako se grupiraju podaci. Kada se algoritam izvrši, idući korak je procijeniti broj klastera (grupa).

Prvi način procjene broja klastera koji je korišten u ovom diplomskom radu je metoda lakta (*elbow method*). Iz definicije funkcije cilja, koja je usko vezana uz varijancu, vidljivo je da se dodavanjem još jednog novog klastera smanjuje varijanca. Povećanjem broja grupa vrijednost funkcije cilja monotono pada, te metodom lakta tražimo kada će taj pad prestati biti značajan. Zapravo, traži se za koji broj klastera $k$, ako se doda još jedan klaster, se neće značajno poboljšati podjela podataka. Na grafu se taj trenutak vidi po kutu koji podsjeća na lakat, odakle i ime same metode.




Slika 1.1: Prva metoda procjene broja klasterac(metoda lakta)

Na slici nam je crvenom bojom prikazan taj trenutak pa u ovom slučaju bi za broj klastera uzeli $k = 3$.

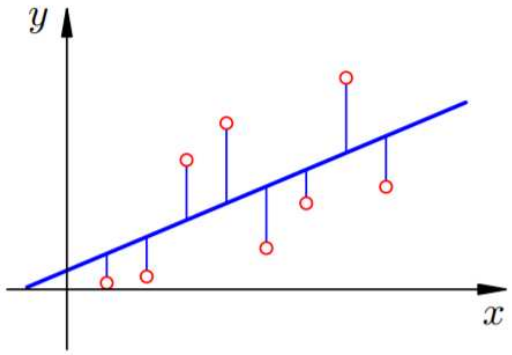
Druga metoda procjene broja klastera je preuzeta iz [7]. U toj je metodi cilj podatke opisati nekim linearnim modelom. Želimo povući pravac tako da suma kvadrata između stvarnih vrijednosti i vrijednosti predviđenih pravcem bude najmanja. Uzimamo da model izgleda ovako:

$$f(k) = \frac{\alpha}{k} + \frac{\beta}{k^2},$$

te procjenjujemo $\alpha$ i $\beta$. Transformiramo formulu za $f$ i dobivamo:

$$f(k) \cdot \ k^2 = \alpha \cdot k + \beta.$$

Pomoću funkcije lm u R-u mogu se dobiti aproksimacije $\alpha$ i $\beta$, a to su redom $\hat{\alpha}$ i $\hat{\beta}$ te dobivamo $\hat{f}(k)$. Računamo razliku između procijenjene vrijednosti i pravih vrijednosti: $f(k) - \hat{f}(k)$ te za najmanju vrijednost očitamo $k$ i taj $k$ gledamo kao optimalan broj klastera za grupiranje podataka.



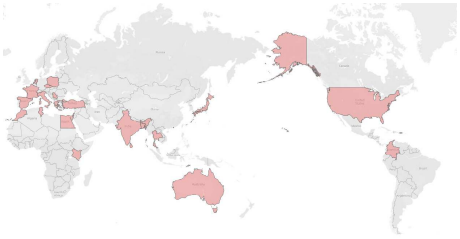

Slika 1.2: Druga metoda procjene broja klastera

# Poglavlje 2

# Opis problema

## 2.1 Struktura podataka

Prije nego se počne s analizom nizova proteina, navest će se na kakvim je podacima rađeno te interpretirati struktura proteina. Podaci na kojima je rađen diplomski rad su nizovi proteina M, N, S i E koji su iz razdoblja od ožujka 2020. godine do lipnja 2020. godine i lokacijski su iz SAD-a, Bangladesha, Japana, Indije, Italije i još mnogo drugih lokacija. U nastavku je slika koja pokazuje sve lokacije iz kojih su prikupljeni nizovi proteina.



Slika 2.1: Lokacije svih nizova

3592 su niza proteina E, 3564 niza proteina M, 3547 nizova proteina S i 3584 niza proteina N. Na primjer, prvi niz iz proteina E nije zapis istog koronavirusa kao prvi niz proteina M, N ili S. Odnosno, već se po nejednakom broju nizova vidi da proteini nisu vezani pa je to jedan od razloga zašto je rađena odvojena analiza za svaki protein. Niz aminokiselina u proteinu je određen nizom gena koji je zapisan u genetskom kodu i koji određuje 20 ”osnovnih” aminokiselina. Aminokiseline se označavaju skraćenicama od jednog ili triju slova.

U ovom radu se radi na podacima u kojima su aminokiseline prikazane s jednim od narednih slova: A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W i Y. S obzirom na to

da se radi o nizu slova, a ne numeričkim vrijednostima, teško je raditi statističke analize na takvim podacima zbog nedostatka metrike za usporedbu nenumeričkih vrijednosti. Taj problem je riješen iz [1] i [3] tako da je definirano preslikavanje u $\mathbb{R}^5$ koje "čuva" sve važne informacije o fizikalno-kemijskim i ostalim svojstvima aminokiselina. Odnosno, ako su podaci poravnati nizovi, aminokiseline možemo pretvoriti u 5-dimenzionalne vektore numeričkih vrijednosti. Tih 5 koordinata nazivamo faktorima. Faktor I (prva koordinata vektora) je bipolaran, faktor II (druga koordinata) je faktor sekundarne strukture, faktor III (treća koordinata) se odnosi na molekularni volumen ili veličinu aminokiseline, faktor IV (četvrta koordinata) odražava raznolikost kodona (relativnu kompoziciju aminokiselina u različitim proteinima), te faktor V (peta koordinata) označava elektrostatski naboj aminokiselina.

| AMINOKISELINE | 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 2.1: Faktori

Protein E duljine je 75, odnosno tih 20 različitih aminokiselina tvore niz od 75 aminokiselina. Protein M je duljine 222, protein N duljine 419, dok je protein S najdulji i duljine je 1273.

Sada kada je navedena strukturu virusa, proteina i način na koji se aminokiseline mogu preslikati u vektore, preostaje transformirati podatke na način koji je prethodno opisan.

## 2.2 Priprema podataka

Na početku je provjerena duljina nizova. Na primjer, navedeno je da je protein E duljine 75. Svi nizovi koji nisu jednake duljine, koji sadrže slovo koje nije oznaka jedne od 20 aminokiselina su izbačeni, te je na tim višestruko poravnatim nizovima rađena prethodno spomenuta transformacija pretvorbe aminokiselina u 5-dimenzionalne vektore. Analogno je "čišćenje" podataka rađeno i na preostalim proteinima.

Kada su podaci pročišćeni, dobivena su 3522 niza proteina E, 3463 nizova proteina M, 3117 nizova proteina N i 2731 nizova proteina S koji su sudjelovali u daljnjoj analizi. Razumljivo je da je najviše nizova izbačeno u proteinu S jer je on i najveće duljine, što nam i ukazuje da je bilo lakše da pri prikupljanju podataka dođe do greške u zapisu niza od preko 1000 slova nego na primjer u proteinu E gdje je svaki niz 75 slova.

Kada je to napravljeno, dobiveni su nizovi čiji su elementi 5-dimenzionalni vektori. Ako je niz duljine 75, sada je to 375-dimenzionalan vektor jer je svako slovo (aminokiselina) prikazano kao 5-dimenzionalan numerički vektor. Da bih se lakše analizirali podatci, nizovi su spremljeni u matrice. Te matrice imaju 5 stupaca, a broj redaka ovisi o duljini proteina. Broj matrica ovisi o broju nizova koji su ostavljeni nakon čišćenja podataka jer je očito da jedna matrica zapravo reprezentira jedan niz pa je broj matrica ekvivalentan broju nizova koji su ostali nakon sređivanja podataka.

U E proteinu to su matrice od 75 redaka i 5 stupaca, u M proteinu od 5 stupaca i 222 retka, u N proteinu su dimenzije od 5 stupaca i 419 redaka, te su u S proteinu matrice s 5 stupaca i 1273 retka. Odnosno, umnožak broja redaka i stupaca zapravo je dimenzija vektora od kojih smo krenuli.

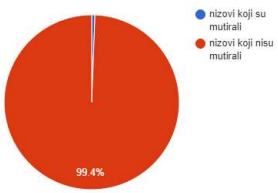
Nakon pretvorbe aminokiselina u vektore, te spremanja nizova u matrice, moguće je analizirati nizove. Prvo se gleda na kojim su se sve pozicijama dogodile promjene. Odnosno, pod pozicijama se misli na indeks aminokiseline u kojoj je došlo do neke mutacije. Pozicije su bile pronađene matricom varijance. To je sve moguće jer su aminokiseline prikazane kao numerički vektori. Za matricu varijanca, prvo je napravljena matrica srednjih vrijednosti gdje se zapravo računa srednja vrijednost po svakoj poziciji pojedinačno. Broj pozicija jednak je duljini proteina. Dobivene su matrice s 5 stupaca, a broj redaka je jednak duljini određenog proteina.

Nakon toga moguće je izračunati matricu varijance. Iz matrice varijance, moglo se zaključiti na kojim se sve pozicijama dogodila neka od promjena. Matrica varijance također je istih dimenzija kao i matrica srednjih vrijednosti. Prvi redak matrice varijance je prikaz prve aminokiseline, drugi druge aminokiseline i tako redom. Na mjestima gdje je varijanca različita od 0, to jest, u matrici nemamo nul-redak se dogodilo neko raspršenje, to jest, neka promjena. Indekse (pozicije) tih promjena se spremaju i iz svih nizova izdvojene su samo te pozicije i samo te su promatrane u daljnjoj analizi.

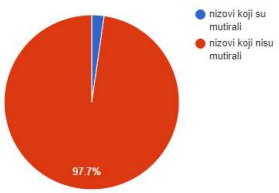
Nakon toga radi se provjera u kojim se sve nizovima dogodila neka od navedenih promjena. To nije bilo neophodno ali na taj način lakše je raditi analizu jer se ne promatraju nizovi koji su jednaki i u kojima nema promjena. Zapravo, namjerno se smanjuje broj nizova tako da se promatraju samo oni nizovi koju su mutirali. Da bi se lakše razumjelo, u nastavku će biti navedeno koliko je nizova mutiralo u kojem proteinu i koliko je pozicija mutiralo.

U proteinu E pronađene su promjene na 11 pozicija, u 21-om nizu. U proteinu M dogodile su se promjene na 28 pozicija, u 79 nizova, a protein N promijenio se na 101 pozicija, u 693 nizova te u proteinu S nalazimo neku od mutacija na 153 pozicija, u 943 nizova. Kada su dobivene te pozicije, nizovi se vraćaju u vektore te se u nastavku navode dimenzije novih vektora po različitim proteinima.

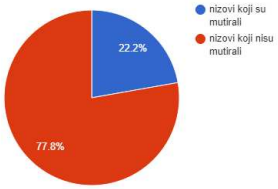
U E proteinu su dobiveni 55-dimenzionalni vektori (11 pozicija (aminokiselina) je mutiralo), a svaka aminokiselina ima 5 numeričkih koordinata tako da je broj 55 rezultat umnoška tih dvaju brojeva. Koliko se nizova promatra u nastavku je zapravo broj nizova u kojima je pronađena neka mutaciju. Stoga slijedi da je u E proteinu dobivamo 21 55-dimenzionalnih vektora. U M proteinu dobiveni su 140-dimenzionalni vektori (28 pozicija, a svaka pozicija je 5-dimenzionalan vektor) tako da umnoškom tih dvaju brojeva dobivamo dimenziju vektora. Analogno se dobiva i u ostalim proteinima. M protein tako ima 79 140-dimenzionalnih vektora, N protein 693 505-dimenzionalnih vektora i na samom kraju, u S proteinu pronađeni je 153 pozicija u 943 niza tako da je dobiveno 943 765-dimenzionalnih vektora.



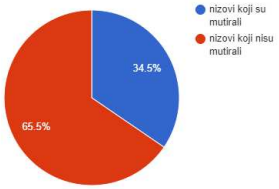

Slika 2.2: E protein




Slika 2.3: M protein




Slika 2.4: N protein




Slika 2.5: S protein

Slike prikazuju postotak mutiranih nizova po proteinima. Vidljivo je da je mali broj nizova mutirao u M i E proteinu i navedeno je da je pronađeno manje promjena u navedenim proteinima pa to ukazuje da se moramo usredotočiti u grupiranju podataka na N i S protein.

Na ovaj način, kroz postepenu analizu proteina, dobiveni su podaci u vektorskom prostoru koji su spremni za primjenu strojnog učenja, odnosno za k-means algoritam.

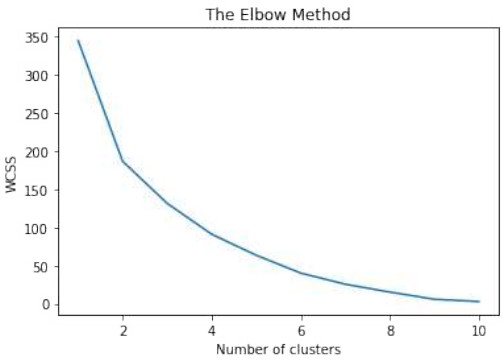
# Poglavlje 3

# Grupiranje podataka k-means algoritmom

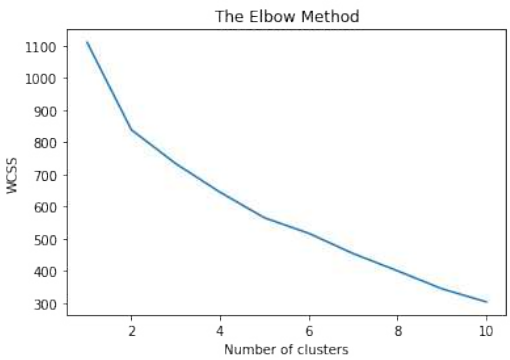
Kada su pripremljeni podaci za k-means algoritam, uz pomoć biblioteke Scikit-Learn u Pythonu, napravljen je k-means++ algoritam iz kojeg nam je cilj otkriti grupiraju li se podaci i ako da, kako.

## 3.1 Određivanje broja klastera

Jedan od načina procjene je pomoću funkcije cilja, metodom lakta (*elbow method*) koja je navedena u 1.3. U nastavku slijede vizualizacija rezultata za sva 4 različita proteina.

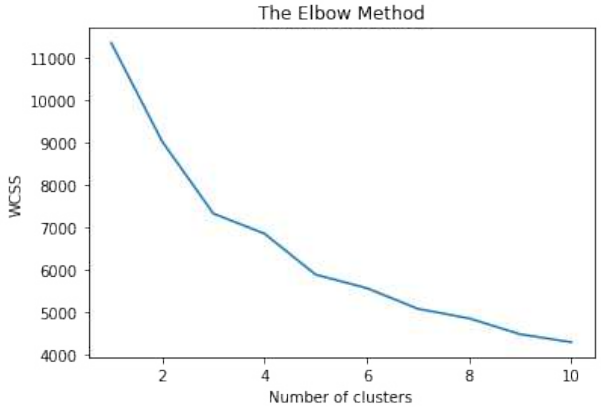



Slika 3.1: Metoda lakta za E protein

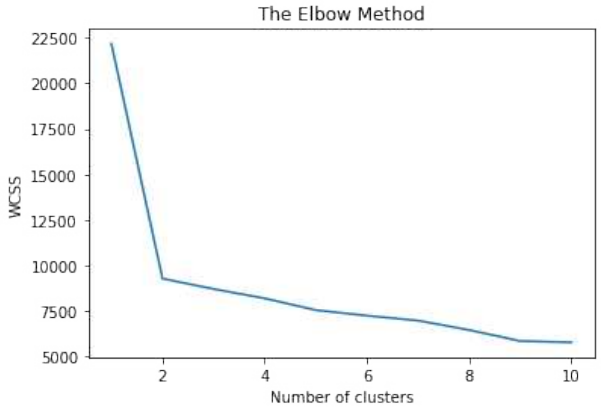



Slika 3.2: Metoda lakta za M protein

Kao što je navedeno u prethodnom poglavlju, ovdje zbog malog broja pozicija i nizova koji su mutirali, u nastavku promatramo samo proteine N i S.

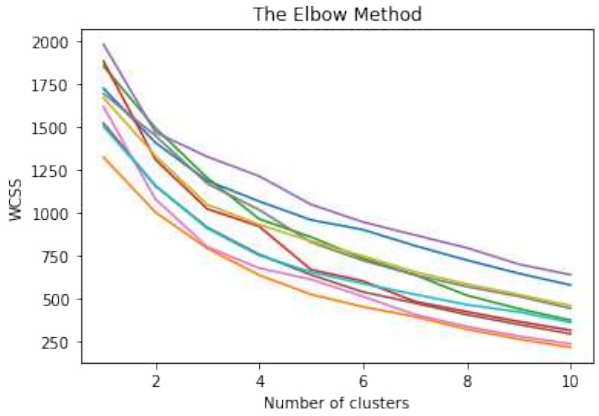



Slika 3.3: Metoda lakta za N protein

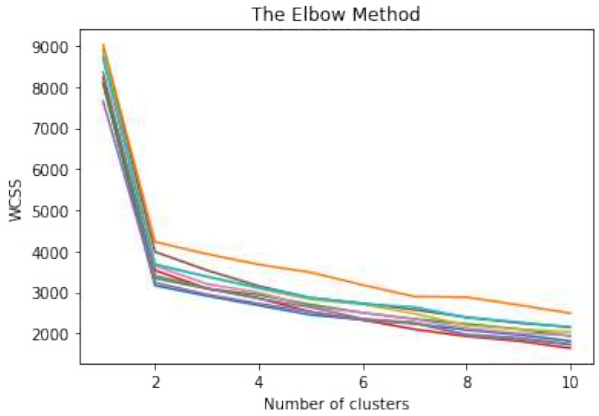



Slika 3.4: Metoda lakta za S protein

Iz slika se vidi kako se podaci drugačije ponašaju u sva 4 proteina, odnosno da svaki protein ima svoj mutacijski uzorak. Nakon toga je provjerena stabilnost klastera u proteinima N i S. Od ukupnog broja nizova, deset puta je nasumično uzimano po 1000 nizova i na njima je napravljena jednaka detaljna analiza. Sve što je navedeno do sada je ponovno obavljeno na tim nasumično odabranim nizovima. U nastavku slijede slike iz kojih je moguće vidjeti klasteri kojeg proteina se ponašaju stabilno.




Slika 3.5: Provjera stabilnosti klastera za N protein




Slika 3.6: Provjera stabilnosti klastera za S protein

Iz slika se može vidjeti kako se klasteri stabilno ponašaju jedino u S proteinu. Zbog nepravilnog grupiranja podataka u N proteinu se može zaključiti kako nema smisla gledati koje pozicije su najviše pridonijele takvom grupiranju podataka, pa preostaje zapravo donijeti zaključke o S proteinu. Iz slike metode lakta za S protein da se zaključiti u koliko se klastera grupiraju podaci. Nakon što je metodom lakta procijenjeno da su 2 klastera, provjera je napravljena i drugom metodom iz [6] koja je također navedena u 1.3. Tom metodom podaci se pokušavaju opisati linearnim modelom. Traži se u kojem trenutku je suma kvadrata između stvarnih vrijednosti i vrijednosti predviđenih pravcem bila najmanja. Tom metodom se također dolazi do zaključka kako je optimalan broj klastera za S protein dva. Iz ove dvije metode i provjere stabilnosti klastera S proteina se zaista može tvrditi kako se S protein grupira u 2 klastera.

Sada kada je donesena odluka koji protein ima smisla gledati i koji broj klastera najbolje opisuje podatke, može se krenuti u traženje "najznačajnijih" pozicija za takvo klasteriranje podataka.

## 3.2 Određivanje najznačajnijih pozicija za klasteriranje podataka

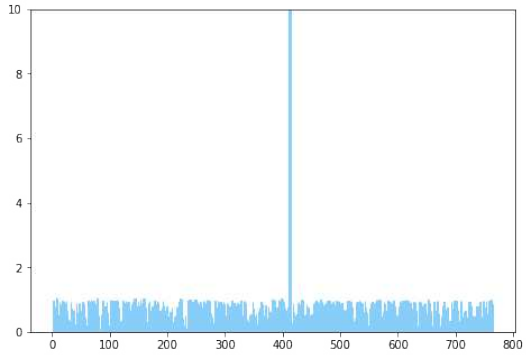
Kao što je i prethodno pokazano, pozicije koje su najviše utjecale na grupiranje podataka ima smisla gledati samo u S proteinu pa se samo na tom proteinu rade iduće korake. Rangiranje koje pokazuje koja koordinata (aminokiselina) najviše utječe na podjelu podataka radimo preko omjera. Taj omjer gledamo kao statistiku koja mjeri je li određena koordinata bolje opisana s jednim ili dva centra. Postupak se radi dva puta. Prvo po svakoj koordinati vektora posebno, a nakon toga po nizu od 5 koordinata. Drugi omjer radi se zato što svakoj aminokiselini pripada niz od 5 koordinata (prethodno je već objašnjeno kako je svaka aminokiselina zapravo prikazana kao 5-dimenzionalan vektor).

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

U omjeru je $l$ duljina vektora, a $br$ je oznaka za broj vektora koji su sudjelovali u k-means++ algoritmu. $K_1$ je oznaka prvog klastera, a $K_2$ je oznaka drugog klastera. Na ovaj način, napravljen je omjer za sve koordinate vektora S proteina. U slučaju S proteina, formula (3.1) izgleda ovako:

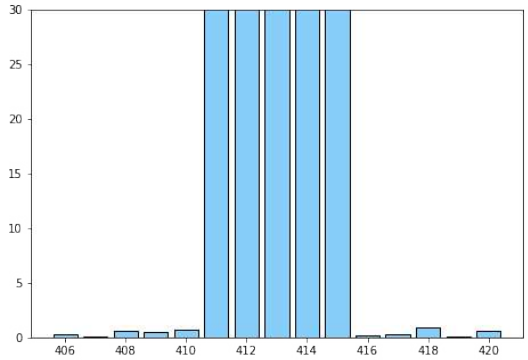
$$O(j) = \frac{\sum_{i=1}^{943}( x_{i,j} - \overline{x_j})^2}{\sum_{k_1 \in K_1}( x_{k_1,j} - \overline{x_{j,\,k_1}})^2 + \sum_{k_2 \in K_2}( x_{k_2,j} - \overline{x_{j,\,k_2}})^2}, j = 1, 2, ..., 765 \qquad (3.2)$$

Primjenom k-means++ algoritma dobivena su 603 vektora u prvom klasteru i 340 vektora u drugom klasteru. U nastavku su vizualizacije rangiranja koordinata.




Slika 3.7: Vrijednosti omjera za sve koordinate

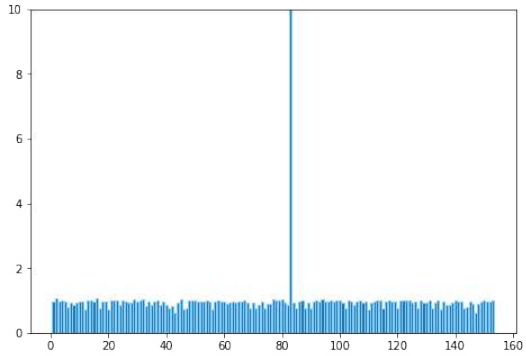
Iz slike je vidljivo da se negdje oko 400. koordinate događa velik skok u vrijednosti omjera. Iduća slika prikazuje u kojim točno koordinatama omjer ima najveću vrijednost.




Slika 3.8: Najznačajnije koordinate

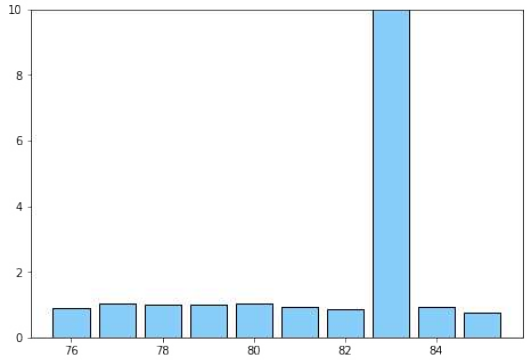
Slika nam pokazuje da omjer ima najveću vrijednost u 411., 412., 413., 414. i 415. koordinati. Naveli smo da omjer gledamo kao statistiku za rangiranje koordinata pa nam to povlači da su tih 5 koordinata najviše utjecale na stvaranje dva klastera.

Analogno formulu 3.1 prilagodimo aminokiselinama gdje zapravo paralelno sumiramo po 5-koordinata (jednoj aminokiselini).



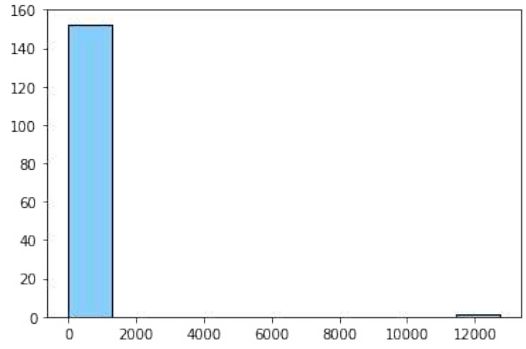

Slika 3.9: Vrijednosti omjera za sve aminokiselina

Iz slike također vidimo skok vrijednosti omjera, a u ovom slučaju vidimo da je to oko 80. aminokiseline. Iduća slika pokazuje koja je to aminokiselina najviše utjecala na stvaranje dvije grupe.



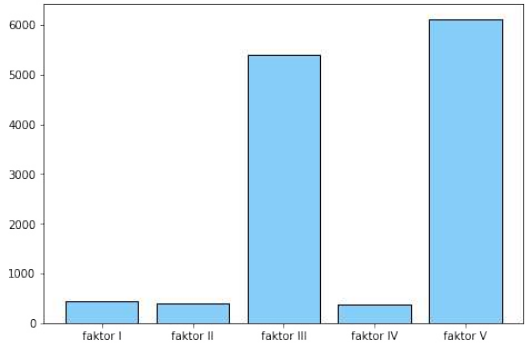

Slika 3.10: Vrijednosti omjera za dominantne aminokiseline

Oba omjera dala su jednake rezultate. Naime, vidimo da je najveći skok od 411. do 415. koordinate a to je zapravo 83. aminokiselina u našim podacima.




Slika 3.11: Histogram frekvencija za omjer aminokiselina

Da se radi o točno jednoj aminokiselini, vidimo iz histograma frekvencija. Kada se vratimo u početne nizove i pogledamo što se nalazi na tom mjestu, to je zapravo 614. aminokiselina. Oni nizovi koji nisu mutirali na toj poziciji imaju aminokiselinu G, u prvom klasteru na 614. poziciji je aminokiselina D, a u drugom klasteru aminokiselina G. Sada kada je određeno mutacije koje aminokiseline i na kojoj poziciji zapravo rade podjelu, preostaje pokazati koji su faktori najviše pridonijeli tome.




Slika 3.12: Faktori za 614. aminokiselinu

Slika nam pokazuje da su najviše pridonijeli razdvajanju grupa III. i V. faktor. Kao što je navedeno u 2.1, to su redom faktori koji se odnose na molekularnu veličini ili volumen aminokiselina i elektrostatski naboj.

Na kraju, zaključak je sljedeći. Cilj nam je bio pronaći značajne mutacije u proteinskim nizovima koronavirusa. Naime, S protein nam se pokazao jedini relevantan za traženje najznačajnijih pozicija za grupiranje podataka, a u istom nam se mutacija aminokiseline G-D na 614. poziciji S-proteina pokazala dovoljno dominantnom da podatke i podijeli na dvije grupe.

Slijedi interpretacija grupiranja podataka u S-proteinu i po državama. Od bitnijih lokacija to su Indiju i Sjedinjene Američke Države. Od 2731 nizova koje smo promatrali, 2127 je iz SAD-a, a 280 iz Indije. Nadalje, od 943 niza koji su mutirali, 732 niza su iz SAD-a, dok je 111 iz Indije. Od tih 732 niza iz SAD-a koji su mutirali, 523 je upalo u prvi klaster, dok je ostatak upao u drugi. Iz toga možemo zaključiti da na 614. poziciji u nizovima SAD-a je zastupljenija aminokiselina D, odnosno da je baš u većini nizova iz SAD-a prisutna presudna mutacija. Ako se želi gledati još preciznije, 46% od ukupnog SAD-a čini savezna država Washington, tj. njih 339. Od tog broja, 298 se smjestilo u prvi klaster, a 41 u drugi. Iduća lokacija za koju možemo zaključiti da je ima smisla analizirati jer je Indija. Naime, od 111 nizova koji su mutirali iz Indije, 21 se nalazi u prvom klasteru, dok se 90 nalazi u drugom. To nam govori da Indija na 614. poziciji ima aminokiselinu G, odnosno kako mutacija koju smo naveli nije karakteristična za tu državu.

| | |
|---|---|
| Australia | 19 |
| Bangladesh | 3 |
| France | 20 |
| Greece: Athens | 2 |
| India | 21 |
| Japan | 1 |
| Netherlands | 6 |
| Sri Lanka | 1 |
| Taiwan | 3 |
| Thailand | 1 |
| Timor-Leste | 3 |
| USA | 22 |
| USA: CA | 167 |
| USA: FL | 2 |
| USA: GA | 4 |
| USA: Illinois | 1 |
| USA: Massachusetts | 2 |
| USA: NC | 1 |
| USA: VA | 24 |
| USA: WA | 298 |
| USA: Wisconsin | 2 |

Slika 3.13: 1. klaster

| | |
|---|---|
| Australia | 7 |
| Bangladesh | 14 |
| Egypt | 2 |
| France | 10 |
| India | 90 |
| Japan | 1 |
| Netherlands | 2 |
| Poland | 1 |
| Serbia: Kraljevo | 1 |
| Spain: ASTURIAS | 1 |
| Thailand | 1 |
| Tunisia | 1 |
| USA | 58 |
| USA: AK | 1 |
| USA: CA | 35 |
| USA: CT | 3 |
| USA: FL | 3 |
| USA: ID | 2 |
| USA: LA | 3 |
| USA: Massachusetts | 18 |
| USA: Michigan | 9 |
| USA: VA | 34 |
| USA: WA | 41 |
| USA: Wisconsin | 2 |

Slika 3.14: 2. klaster

# 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. Glad, *Neke statističke metode u predikciji tercijarne strukture proteina*, Diplomski rad, Sveučilište u Zagrebu, Prirodoslovno-matematički fakultet, Matematički odsjek, 2012.

[4] M. Gregorić, *Strojno učenje kao alat za zaključivanje*, Završni rad, Sveučilište u Zagrebu, Filozofski fakultet, Odsjek za informacijske i komunikacijske znanosti, 2019.

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

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

[7] S. Šeperić, *Kriteriji kompleksnosti za k-means algoritam*, Diplomski rad, Sveučilište u Zagrebu, Prirodoslovno-matematički fakultet, Matematički odsjek, 2014.

[8] T. Šmuc, *Strojno učenje: Uvod u strojno učenje*, predavanja, Sveučilište u Zagrebu, Prirodoslovno-matematički fakultet, Matematički odsjek, 2019./2020., dostupno na `https://web.math.pmf.unizg.hr/nastava/su/`

# Sažetak

Tema ovog diplomskog rada je analiza proteinskih nizova iz koronavirusa, tj. analiza proteina: E, M, N i S. U radu se statističkim analizama i primjenom tehnika strojnog učenja na višestruko poravnatim nizovima navedenih proteina pokušava pronaći dominantne mutacije.

Na početku su dani matematički pojmovi potrebni za razumijevanje ostatka rada. Nakon toga, uvodi se struktura podataka na kojima su rađene analize te se podaci pripremaju za primjenu tehnika klasteriranja. Na kraju se primjenjuje jedna od tehnika klasteriranja (k-means++ algoritam) i analiziraju se rezultati. Pri tome, analizira se svaki protein koronavirusa zasebno i traže se najznačajnije pozicije za upravo takvo klasteriranje koje je dobiveno.

Diplomski rad je većinom napravljen u programskom jeziku Python. Uz njega korišten je programski jezik R te za vizualizaciju rezultata Tableau.

# Summary

The topic of this thesis is the analysis of the Coronavirus protein sequences, i.e. study of proteins E, M, N, and S. By applying statistical analysis and machine learning techniques to multiple sequence alignment of a given protein – we aim to determine dominant mutations.

The introduction contains a description of the mathematical framework. After that, a data structure used for the analysis is introduced, and data is prepared for the application of the clustering method. In the final part of the paper, one of the clustering methods (k-means++ algorithm) is applied, and the results are analyzed. In doing so, each Coronavirus protein is analyzed separately, and the goal is to find the most significant positions responsible for the obtained clustering.

The thesis has been pre-developed using programming languages Python and R, while Tableau was used for the visualization of the data.

# Životopis

Rođena sam 17. veljače 1995. godine u Zagrebu. Svoje školovanje započinjem u Osnovnoj školi Bukovac koju završavam 2009. godine i nakon koje upisujem II. gimnaziju u Zagrebu. Nakon završetka II. gimnazije upisujem preddiplomski studij matematike na Prirodoslovno-matematičkom fakultetu u Zagrebu. Po završetku preddiplomskog studija upisujem diplomski sveučilišni studij Matematičke statistike na istom odsjeku.