Clustering proteinskih poravnanja

Buljan, Milan

Master's thesis / Diplomski rad

2024

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **University of Zagreb, Faculty of Science / Sveučilište u Zagrebu, Prirodoslovno-matematički fakultet**

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:217:771064

Rights / Prava: In copyright / Zaštićeno autorskim pravom.

Download date / Datum preuzimanja: **2025-03-29**

Repository / Repozitorij:

Repository of the Faculty of Science - University of Zagreb

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Milan Buljan

CLUSTERING PROTEINSKIH PORAVNANJA

Diplomski rad

Voditelj rada:
doc. dr. sc. Pavle Goldstein

Zagreb, studeni 2024.

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. _____

Velike zahvale upućujem mom mentoru, doc. dr. sc. Pavlu Goldsteinu s kojim je izrada ovog diplomskog rada izgledala toliko jednostavno, a uz opuštenu atmosferu i kvalitetno vodstvo postigao sam cilj i završio svoj studentski put. Od srca hvala mojoj obitelji na iznimnoj i bezuvjetnoj podršci u svakom koraku mog odrastanja. Hvala im što su bili i jesu moj najveći vjetar u leđa. Hvala svim mojim sportskim prijateljima, a posebno onima iz NK Junak Sinj s kojima sam kroz sve ove godine studiranja kao potpredsjednik Kluba gradio vrhunsku priču. Udvostručili smo broj upisane djece, udvostručili smo broj sretnih osmijeha, a ta pozitivna energija stvorila je vrhunsku atmosferu. U pamćenju posebno ostaje sezona 2023./24. kada juniori samo zbog gol razlike nisu ušli u 1.HNL. Mala smo sredina, ali veliki smo Klub! Biti jedan od vas velika je privilegija, čast i ponos! Hvala svim prijateljima i poznanicima koji su vjerovali u mene, kao i onima koji se nisu pronašli u ovim zahvalama, a zaslužili su čuti riječ hvala. Ovo je samo početak jedne nove priče, još ljepše i sjajnije. Veselim se budućnosti okružen s toliko divnih i časnih ljudi!

Sadržaj

Sadržaj **iv**

Uvod **1**

1 Matematički pojmovi **2**
1.1 Linearna algebra 2
1.2 Vjerojatnost i statistika 5

2 Bioinformatika **10**
2.1 Biološki pojmovi 10
2.2 Prelazak u vektorski prostor 13

3 Analiza problema i rezultati **14**
3.1 Opis problema i ideja 14
3.2 Rezultati 16
3.3 Usporedba s prethodnim istraživanjima 28

Bibliografija **30**

Uvod

Proteini su makromolekule sastavljene od niza aminokiselina povezanih peptidnim vezama. Njihova uloga u biologiji je od iznimnog značaja, prisutni su u svim živim organizmima te su stoga od posebnog interesa i predmet su brojnih istraživanja kako bi se bolje razumijeli razni biološki procesi. U ovom radu razmatrati ćemo enzime, proteine koji djeluju kao biokatalizatori, a svojom funkcijom vezani su uz ubrzavanje kemijskih reakcija unutar stanica. Razmatranja se odnose na kinaze, ciklaze, acil transferaze te MLDH.

Svaka od navedenih familija proteina dijeli se na dvije podfamilije, a jedno od važnih otvorenih bioloških pitanja je koje su pozicije u proteinskim poravnanjima pojedine familije proteina značajne, odnosno najviše utječu na podjelu proteina u svaku od podfamilija. U ovom radu, korištena su proteinska poravnanja prethodno navedene četiri familije proteina kao i zadana podjela u podfamilije, objašnjen je prelazak sa niza aminokiselina u vektorski prostor te je na tako pripremljenim podacima izvršena obrada računanjem niza F-faktora za svaku od promatranih pozicija. Provjeravamo je li tako dobijen niz F-faktora F-distribuiran te računamo $90\%, 95\%$ i 99% kvantile procijenjene F distribucije koja najbolje opisuje dane podatke i na temelju njih određujemo značajne pozicije u svakoj od familija.

Konceptualno, rad je podijeljen u tri poglavlja. Prvo poglavlje donosi potrebne matematičke pojmove za razumijevanje rada, a vezano je uz područja linearne algebre, vjerojatnosti i statistike. Drugo poglavlje daje uvid u potrebne pojmove iz bioinformatike te je u njemu dan važan korak u obradi podataka - prelazak sa niza aminokiselina u vektorski prostor. Posljednje poglavlje opisuje problem kojim se bavi ovaj diplomski rad, daje se ideja i pristup njegovom rješavanju te se navode rezultati dobijeni na tri različita načina kroz brojne grafičke i tablične prikaze. Također, daju se pripadni komentari i u konačnici vrši se usporedba s prethodnim istraživanjima na ovu temu.

Poglavlje 1

Matematički pojmovi

U ovom poglavlju proći ćemo kroz najvažnije definicije, teoreme, propozicije i napomene potrebne za razumijevanje rada. Tematski, poglavlje obrađuje znanja linearne algebre te vjerojatnosti i statistike. Pojmovi su dijelom ili u cijelosti preuzeti iz izvora [2], [5], [11], [13], [3], [7], [14], [10], [20], [17].

1.1 Linearna algebra

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

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

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

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. $\langle x, x \rangle \geq 0, \quad \forall x \in V;$
2. $\langle x, x \rangle = 0 \Leftrightarrow x = 0;$
3. $\langle x_1 + x_2, y \rangle = \langle x_1, y \rangle + \langle x_2, y \rangle, \quad \forall x_1, x_2, y \in V;$
4. $\langle \alpha x, y \rangle = \alpha \langle x, y \rangle, \quad \forall \alpha \in \mathbb{F}, \forall x, y \in V;$
5. $\langle x, y \rangle = \overline{\langle y, x \rangle}, \quad \forall x, y \in V.$

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

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

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

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

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

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

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

Definicija 1.1.10. *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 $\mathbb{R}^n$, definirana u napomeni 1.1.6, dana je formulom*

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

Ova norma se 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. $d(x, z) \leq d(x, y) + d(y, z)$, $\forall x, y, z \in V$.

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.

1.2 Vjerojatnost i statistika

Vjerojatnosni prostor

Definicija 1.2.1. *Slučajni pokus ili* slučajni eksperiment *je pokus čiji ishodi nisu jednoznačno određeni.*

Definicija 1.2.2. *Prostor elementarnih događaja* Ω *je neprazan skup koji reprezentira skup svih ishoda slučajnog pokusa. Elemente* ω *skupa* Ω *nazivamo* elementarni događaji.

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

1. $\mathbb{P}(A) \geq 0$, $\forall A \in \mathcal{F}$;
2. $\mathbb{P}(\Omega) = 1$;
3. $A_i \in \mathcal{F}$, $i \in \mathbb{N}$ i $A_i \cap A_j = \emptyset$ *za* $i \neq j \implies \mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$.

Definicija 1.2.6. *Uređena trojka* $(\Omega, \mathcal{F}, \mathbb{P})$, *gdje je* $\mathcal{F}$ σ*-algebra na* Ω, *a* $\mathbb{P}$ *je vjerojatnost na* $\mathcal{F}$, *zove se* vjerojatnosni prostor.

Slučajna varijabla

Definicija 1.2.7. *Neka je S proizvoljan neprazan skup i $\mathcal{A}$ familija podskupova od S ($\mathcal{A} \subset \mathcal{P}(S)$). Sa $\sigma(\mathcal{A})$ označimo najmanju σ-algebru podskupova od S koja sadrži $\mathcal{A}$. Nju nazivamo* ***σ-algebra generirana sa $\mathcal{A}$.***

Definicija 1.2.8. *Neka je $\mathcal{B}$ označena σ-algebra generirana familijom svih otvorenih skupova na $\mathbb{R}$. $\mathcal{B}$ zovemo* ***σ-algebra Borelovih skupova*** *na $\mathbb{R}$, a elemente σ-algebre $\mathcal{B}$ zovemo* ***Borelovi skupovi****.*

Definicija 1.2.9. *Neka je $(\Omega, \mathcal{F}, \mathbb{P})$ vjerojatnosni prostor. Funkcija $X : \Omega \to \mathbb{R}$ je* ***slučajna varijabla (na Ω)*** *ako je $X^{-1}(B) \in \mathcal{F}$ za proizvoljno $B \in \mathcal{B}$, tj. $X^{-1}(\mathcal{B}) \subset \mathcal{F}$.*

Definicija 1.2.10. *Neka je $(\Omega, \mathcal{F}, \mathbb{P})$ vjerojatnosni prostor i $X : \Omega \to \mathbb{R}^n$. Kažemo da je X* ***n-dimenzionalan slučajni vektor*** *(ili, kraće,* ***slučajni 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 $(\Omega, \mathcal{F}, \mathbb{P})$. X je* ***jednostavna slučajna varijabla*** *ako je njeno područje vrijednosti konačan skup.*

X je jednostavna slučajna varijabla ako i samo ako je

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

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

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

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

i

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

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

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

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).*

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 s $\mathcal{K}$ skup svih jednostavnih slučajnih varijabli definiranih na Ω, a s $\mathcal{K}_+$ skup svih nenegativnih funkcija iz $\mathcal{K}$.

Neka je $X \in \mathcal{K}$, $X = \sum_{k=1}^{n} x_k \mathcal{X}_{A_k}$, gdje su $A_1, A_2, \ldots, A_n \in \mathcal{F}$ međusobno disjunktni.

Definicija 1.2.14. ***Matematičko očekivanje*** *od X ili, kraće,* ***očekivanje*** *od X označavamo s* $\mathbb{E}[X]$ *i definira se s:*

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

Propozicija 1.2.15. *1. Neka je* $c \in \mathbb{R}$ *i* $X \in \mathcal{K}$*. Tada je* $\mathbb{E}(cX) = c\mathbb{E}X$*.*

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

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

Neka je sada X **nenegativna slučajna varijabla** definirana na Ω. Prema teoremu 1.2.13 postoji rastući niz $(X_n)_{n\in\mathbb{N}}$ nenegativnih jednostavnih slučajnih varijabli takav da je $X = \lim_{n\to\infty} X_n$. Iz prethodne propozicije 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 1.2.16. ***Matematičko očekivanje*** *od X ili, kraće,* ***očekivanje*** *od X definira se s*

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

Neka je sada X proizvoljna slučajna varijabla na Ω. Vrijedi $X = X^+ - X^-$, gdje su X^+, X^- slučajne varijable i $X^+, X^- \geq 0$.

Definicija 1.2.17. *Kažemo da* ***matematičko očekivanje*** *od X ili, kraće,* ***očekivanje*** *od X postoji (ili je definirano) ako je barem jedna od veličina* $\mathbb{E}[X^+]$, $\mathbb{E}[X^-]$ *konačna, tj. vrijedi* $\min(\mathbb{E}[X^+], \mathbb{E}[X^-]) < +\infty$*. Tada po definiciji stavljamo*

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

Definicija 1.2.18. *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 s* $Var(X)$ *ili* σ_X^2 *na sljedeći način:*

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

Napomena 1.2.19. *Pozitivan drugi korijen iz varijance nazivamo* ***standardna devijacija*** *i označavamo sa* σ_X*.*

Funkcija distribucije

Definicija 1.2.20. *Neka je X slučajna varijabla na Ω.* ***Funkcija distribucije*** *od X je funkcija $F_X : \mathbb{R} \to [0,1]$ definirana s:*

$$F_X(x) = \mathbb{P}(X^{-1}((-\infty, x])) = \mathbb{P}\{\omega \in \Omega : X(\omega) \leq x\} = \mathbb{P}\{X \leq x\}, \quad x \in \mathbb{R}.$$

Napomena 1.2.21. *Ako je jasno o kojoj se slučajnoj varijabli, odnosno njenoj funkciji distribucije, radi piše se F umjesto F_X.*

Teorem 1.2.22. *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 1.2.23. *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 1.2.24. *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 $\mathbb{P}\{X \in D\} = 1$.*

Definicija 1.2.25. *Neka je X slučajna varijabla na vjerojatnosnom prostoru $(\Omega, \mathcal{F}, \mathbb{P})$ i neka je F_X njena funkcija distribucije. Kažemo da je X* ***apsolutno neprekidna*** *ili, kraće,* ***neprekidna slučajna varijabla*** *ako postoji nenegativna realna Borelova funkcija f na $\mathbb{R}$ (tj. $f : \mathbb{R} \to \mathbb{R}_+$) takva da je*

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

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

Definicija 1.2.26. *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*

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

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

Definicija 1.2.27. *Neka su $Z_1, Z_2, \ldots, Z_n$ nezavisne standardne normalne slučajne varijable, tj. $Z_i \sim N(0,1)$ za $i = 1, 2, \ldots, n$. Tada je slučajna varijabla*

$$\chi^2 = Z_1^2 + Z_2^2 + \cdots + Z_n^2$$

χ^2-distribuirana s n stupnjeva slobode, što zapisujemo kao $\chi^2 \sim \chi^2(n)$.

Definicija 1.2.28. *Neka su U_1 i U_2 nezavisne slučajne varijable koje prate χ^2-distribuciju s d_1 i d_2 stupnjeva slobode, redom. Tada je slučajna varijabla*

$$F = \frac{(U_1/d_1)}{(U_2/d_2)}$$

F-distribuirana s d_1 i d_2 stupnjeva slobode, što zapisujemo kao $F \sim F(d_1, d_2)$.

Opisna analiza podataka

U ovom dijelu ćemo se podsjetiti definicija iz deskriptivne statistike koje će nam biti potrebne u daljnjem razumijevanju rada. Navodimo pojmove kao što su aritmetička sredina, standardna devijacija uzorka te varijanca uzorka i standardizacija podataka.

Neka su

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

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

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

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

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

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

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

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

Poglavlje 2

Bioinformatika

2.1 Biološki pojmovi

Proteini (bjelančevine) su makromolekule koje se sastoje od lanaca aminokiselina povezanih peptidnim vezama. Prisutni su u svim živim organizmima i igraju ključnu ulogu u gotovo svim biološkim procesima. Proteini čine oko 20% mase ljudskog tijela što ih čini esencijalnim komponentama stanica i tkiva. Njihova struktura i funkcija su od temeljne važnosti za razumijevanje brojnih bioloških procesa.

Proteini su izgrađeni od 20 različitih aminokiselina prikazanih u tablici 2.1, koje se međusobno kombiniraju u specifičnim nizovima.

Oznaka	**Naziv**	**Oznaka**	**Naziv**
A	Alanin	M	Metionin
C	Cistein	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

Struktura proteina izravno utječe na njegovu funkciju, a svaka promjena u primarnoj strukturi može dovesti do ozbiljnih posljedica za funkciju proteina. Struktura proteina može se opisati na četiri razine:

1) Primarna struktura – niz aminokiselina.

2) Sekundarna struktura – lokalna organizacija lanca u oblike poput alfa-uzvojnica (spiralni oblik) i beta-ploča (oblik sličan nabranom listu).

3) Tercijarna struktura – trodimenzionalna struktura cjelokupnog proteina.

4) Kvartarna struktura – složenost nastala udruživanjem više proteinskih lanaca.

Proteini imaju širok spektar funkcija u organizmu, a u ovom radu ćemo promatrati **enzime** (proteine koji djeluju kao biokatalizatori, ubrzavajući kemijske reakcije unutar stanica).

Kinaze su enzimi koji kataliziraju prijenos fosfatne grupe s molekule ATP-a na određenu aminokiselinu u proteinskom supstratu, proces poznat kao fosforilacija. Ova modifikacija mijenja aktivnost proteina, stabilnost, lokalizaciju i interakcije s drugim molekulama, što je ključni mehanizam regulacije unutar stanice. Fosforilacija je posebno važna u prijenosu signala, staničnoj proliferaciji, diferencijaciji i apoptozi. Najvažnije vrste kinaza su tirozin kinaze i serin/treonin kinaze. Zbog njihove ključne uloge u regulaciji staničnih procesa, mnoge kinaze su povezane s razvojem raka i drugih bolesti, pa su česta meta istraživanja i terapija.

Ciklaze su enzimi koji kataliziraju formiranje cikličkih molekula iz neorganskih molekula poput ATP-a ili GTP-a. Najpoznatije su adenilat ciklaze i guanilat ciklaze, koje pretvaraju ATP u ciklički AMP (cAMP) i GTP u ciklički GMP (cGMP), redom. cAMP i cGMP su važni sekundarni glasnici koji prenose signale unutar stanice i reguliraju različite biološke procese, uključujući metabolizam, kontrakciju mišića, i neurotransmisiju. Adenilat ciklaze, primjerice, igraju ključnu ulogu u signalizaciji receptora povezanih s G-proteinom, što ima važnu ulogu u imunološkom odgovoru i regulaciji krvnog tlaka.

Acil transferaze (iz poliketidnih sintetaza) su enzimi koji prenose acilne grupe između molekula. One igraju ključnu ulogu u metabolizmu lipida i sintezi masnih kiselina, gdje kataliziraju prijenos acilnih skupina s koenzima A na druge molekule. Acil transferaze su važne za stvaranje raznih lipida, poput fosfolipida i triglicerida, koji su esencijalni za strukturu staničnih membrana i energetsku pohranu. Disfunkcija acil transferaza može utjecati na metabolizam lipida i povezati se s bolestima poput ateroskleroze.

Mitohondrijska laktat dehidrogenaza (MLDH) je enzim koji katalizira reverzibilnu konverziju laktata u piruvat u prisutnosti NAD^+. Ovaj proces omogućava recikliranje $NADH$ u NAD^+, što je ključno za održavanje anaerobne glikolize. MLDH igra ključnu ulogu u energetskom metabolizmu, posebno u tkivima s visokim energetskim zahtjevima, kao što su srčani i skeletni mišići. U aerobnim uvjetima, MLDH omogućuje iskorištavanje laktata kao izvora energije, čime doprinosi regulaciji razine laktata i održavanju homeostaze. Zbog svoje uloge u oksidaciji laktata, MLDH je važan za razumijevanje metaboličkih procesa, posebno kod stanja poput hipoksije i srčanih bolesti.

Proteinsko poravnanje (*engl. protein alignment*) je metoda u bioinformatici kojom se uspoređuju dva ili više proteinskih nizova kako bi se identificirale regije sličnosti. Te regije sličnosti često ukazuju na evolucijsku povezanost, zajedničke funkcije ili strukturne karakteristike među proteinima. Poravnanje može otkriti koje aminokiseline u različitim proteinima obavljaju slične ili iste uloge te tako pomoći u razumijevanju njihove biološke funkcije.

Neki od najpoznatijih alata za proteinska poravnanja su BLAST (*Basic Local Alignment Search Tool*), Clustal Omega i MUSCLE.

Pojmovi iz ovog potpoglavlja preuzeti su iz izvora: [8], [19], [4], [22], [15], [12], [21], [18] i [10].

2.2 Prelazak u vektorski prostor

Budući da se proteini sastoje od niza aminokiselina reprezentiranih slovima nad kojima nemamo neki uređen sustav, niti prirodnu metriku, u svrhu lakše obrade podataka i lakšeg manevra prelazimo u vektorski prostor. Kao što je opisano u radu [1], poznato je da svaku aminokiselinu možemo reprezentirati 5-dimenzionalnim vektorom realnih brojeva, a svaka od tih dimenzija predstavlja jedan faktor koji opisuje neko (ili više njih) svojstvo aminokiseline. Faktori I, II, III, IV, V redom opisuju: polaritet aminokiseline, sekundarnu strukturu, molekularni volumen, raznolikost kodona i elektrostatički naboj aminokiseline. Pregled faktora po aminokiselinama dan je u tablici 2.2.

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

Na ovaj način nizu koji se sastoji od n aminokiselina možemo pridružiti $5n$-dimenzionalni vektor nad kojim jednostavno možemo vršiti sve potrebne izračune.

Pojmovi iz ovog potpoglavlja preuzeti su iz izvora [1].

Poglavlje 3

Analiza problema i rezultati

Nakon upoznavanja s potrebnom matematičkom i biološkom pozadinom, u ovom poglavlju opisujemo konkretan problem te navodimo ideju pristupa, kao i dobivene rezultate. U konačnici, vršimo usporedbu s rezultatima iz dosadašnjih istraživanja na ovu temu u kojima su korišteni bitno drugačiji i kompleksniji pristupi.

3.1 Opis problema i ideja

Cilj ovog diplomskog rada je u raznim proteinskim familijama, promatrajući njihova poravnanja i zadanu podjelu u dvije podfamilije koje su biološki ustanovljene, pronaći statistički značajne pozicije koje određuju pripadnost pojedinog proteina nekoj od dvije podfamilije.

Dostupni podaci i priprema za obradu

U prethodnom poglavlju naveli smo bitne biološke karakteristike proteina te četiri familije od interesa: ciklaze, kinaze, AT i MLDH. U ovom poglavlju promatrati ćemo njihova poravnanja na način da ćemo konstruirati 4 matrice znakova (za svaku od familija po jednu), a koje će biti građene tako da poravnati proteini pripadnih familija predstavljaju retke matrice. Konkretno, u ovom radu, korišteni su podaci u kojima se nalazi 215 kinaza, 75 ciklaza, 177 AT i 183 MLDH. Ti brojevi predstavljati će broj redaka pojedine matrice. Poravnate kinaze imaju 600 znakova (slova i crtica), ciklaze 2040, AT 324 i MLDH njih 418. Ti brojevi predstavljaju broj stupaca pojedine matrice. Sljedeći korak je iz ovih matrica izbaciti sve one stupce koji u sebi sadrže znak '-', tj. crticu. Razlog toga je opisan u radu [4] gdje je navedeno da takvi stupci imaju zanemariv utjecaj. Time smo reducirali broj stupaca u matricama i za kinaze nam je ostao 161 stupac, za ciklaze 459, za AT 189 i za MLDH 245. U ovom trenutku imamo četiri matrice znakova, ali kao što je navedeno u prethodnom poglavlju, prelaskom u vektorski prostor, svaku aminokiselinu možemo reprezentirati sa

pet faktora, odnosno 5-dimenzionalnim vektorom, pa to i činimo. Time smo sa matrica znakova prešli na matrice realnih brojeva kojima ćemo u nastavku baratati. Drugi set dostupnih podataka su nam podfamilije (grupe) označene brojevima 1 i 2. Za svaku od četiri promatrane familije imamo vektor koji se sastoji od onoliko članova koliko u toj familiji imamo proteina (dakle, broj članova u vektoru jednak je broju redaka pripadne matrice).

Računanje F-faktora

Za svaku od promatranih familija proteina kreirati ćemo niz F-faktora. Neka je M neka matrica znakova sa k stupaca opisana u prethodnom potpoglavlju i promotrimo njen i-ti stupac gdje je $i \in \{1, 2, ..., k\}$. Svaki takav stupac zamijenili smo sa 5 stupaca brojeva, nazovimo ih $S_{i,1}, S_{i,2}, S_{i,3}, S_{i,4}, S_{i,5}$. F-faktor pridružen stupcu i računamo po formuli:

$$F_i = \sum_{j=1}^{5} \frac{Var(S_{i,j})}{Var(S_{i,j}|g = 1) + Var(S_{i,j}|g = 2) + 0.1} \tag{3.1}$$

gdje $Var(S_{i,j}|g = 1)$ predstavlja varijancu onih elemenata u stupcu $S_{i,j}$ koji pripadaju podfamiliji 1, dok $Var(S_{i,j}|g = 2)$ predstavlja varijancu onih elemenata u stupcu $S_{i,j}$ koji pripadaju podfamiliji 2.

Budući da je na ovakav način moguće dobijati brojeve vrlo bliske nuli, u radu je korištena i sljedeća formula:

$$F_i = \sum_{j=1}^{5} \frac{c + Var(S_{i,j})}{Var(S_{i,j}|g = 1) + Var(S_{i,j}|g = 2) + c^2} \tag{3.2}$$

gdje je $c \in \mathbb{R}$. Konkretno, korištene su vrijednosti $c = 1.5$ i $c = 2$.

Na ovakav način, za svaku familiju proteina, dobili smo niz $(F_i)_{i\in\{1,2,...,k\}}$. Sljedeći korak je prikazati podatke iz tog niza u histogramu te pronaći F distribuciju koja najbolje opisuje dane podatke i u konačnici Kolmogorov-Smirnovljevim testom provjeriti dolaze li ti podaci iz procijenjene F-distribucije. Pretpostavka na koju se naslanja ovaj rad je da bi prethodno definirani F-faktori trebali slijediti F-distribuciju. Tu pretpostavku ćemo provjeriti posebno na formuli 3.1 kao i na formuli 3.2. Na tako procijenjenoj F-distribuciji koja najbolje opisuje dane podatke odrediti ćemo 90%, 95% i 99% kvantile i potražiti one stupce kojima je pridružena vrijednost F-faktora veća od graničnih vrijednosti. Reći ćemo da su ti stupci značajni. Prikazujemo i t-SNE grafove temeljene na top 10 pozicija po vrijednosti F-faktora kako bi vizualno prezentirali uspješnost modela. Naime, prethodnim biološkim istraživanjima pokazano je kako sve vrste proteina nemaju jednak broj značajnih pozicija, one se uglavnom kreću od jedne (rijetko) do njih pet, šest, pa stoga u ovom razmatranju uzimamo njih 10 kako bi pokrili sve te slučajeve.

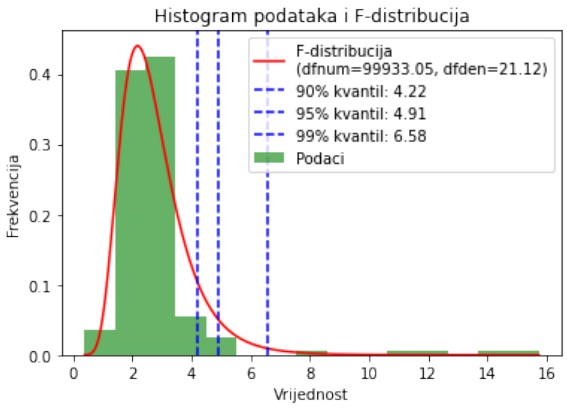
3.2 Rezultati

Kinaze

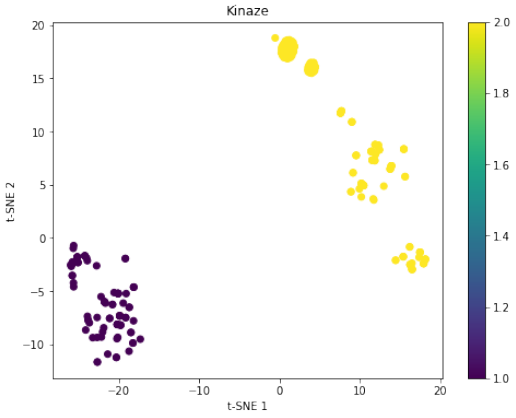
U nastavku navodimo rezultate dobijene promatranjem kinaza. U formuli 3.2 korištenjem konstanti $c = 1.5$ i $c = 2$ dobijeni su podaci koji s velikom vjerojatnosti prate F-distribuciju budući da su dobivene p-vrijednosti Kolmogorov-Smirnovljevog testa iznosile 0.3762 za $c = 1.5$ i 0.8173 za $c = 2$. Također, valja istaknuti kako smo korištenjem konstante $c = 1.5$ dobili pozicije koje su nam značajne s vjerojatnosti od preko 99%, a one su vidljive u tablici 3.2. Riječ je redom o pozicijama 532, 241, 65 i 343. Prelaskom na veće konstante c uočeno je da se p-vrijednost povećava, ali i da gubimo na značajnosti pozicija. Korištenjem formule 3.1 dobijamo da podaci s velikom vjerojatnosti ne prate F-distribuciju (p-vrijednost Kolmogorov-Smirnovljevog testa je manja od 10^{-4}). Ipak, zanimljivo je istaknuti kako su i ovakvim pristupom dobivene iste prethodno navedene pozicije (kao za konstantu $c = 1.5$) sa preko 99%-tnom značajnosti.

Korištenjem formule 3.1, $p < 10^{-4}$

Na svim standardnim razinama značajnosti odbacujemo pretpostavku o tome da pripadni podaci prate F-distribuciju. Korištenjem formule 3.1 dobivamo da je p-vrijednost Kolmogorov-Smirnovljevog testa $p < 10^{-4}$. U nastavku slijede grafički i tablični prikazi:



Slika 3.1: F-faktori



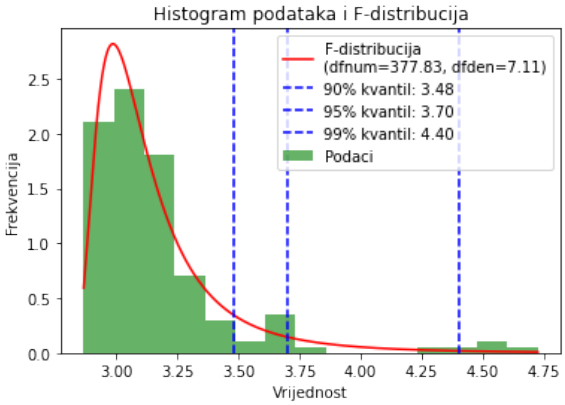
Slika 3.2: t-SNE prikaz

Pozicija	65	343	345	532	241	261	101	376	542	245
F-faktor	15.75	14.28	12.24	11.20	7.99	5.02	4.97	4.89	4.70	4.39

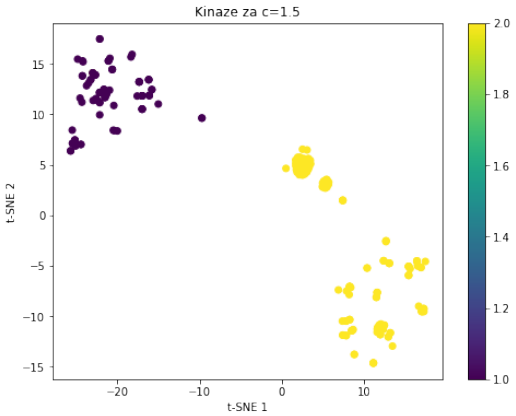
Tablica 3.1: Top 10 pozicija

Korištenjem formule 3.2, $c = 1.5, p = 0.3762$

Na razini značajnosti od 1% ne možemo odbaciti pretpostavku o tome da pripadni podaci prate F-distribuciju. U formuli 3.2 korištenjem $c = 1.5$ dobijamo da je p-vrijednost Kolmogorov-Smirnovljevog testa $p = 0.3762$. U nastavku slijede grafički i tablični prikazi:



Slika 3.3: F-faktori



Slika 3.4: t-SNE prikaz

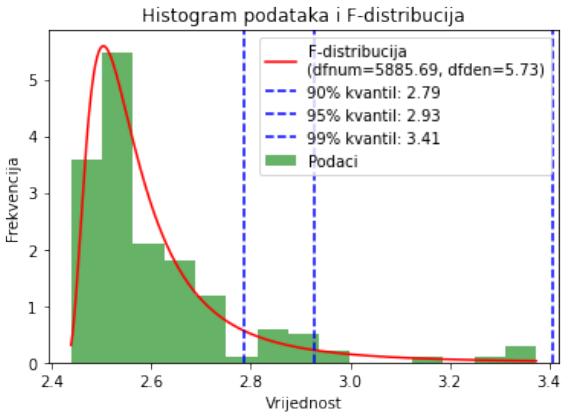
Pozicija	532	241	65	343	345	261	18	342	344	528
F-faktor	4.73	4.57	4.53	4.41	4.29	3.79	3.72	3.71	3.69	3.67

Tablica 3.2: Top 10 pozicija

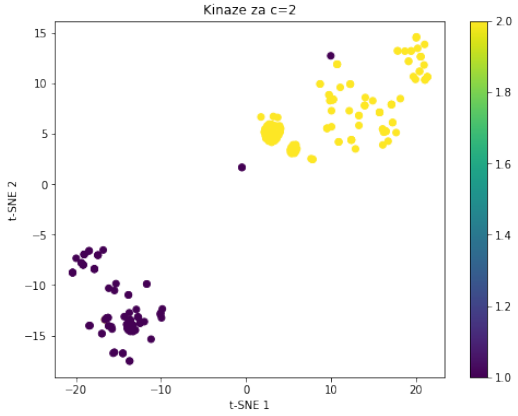
Iz danog t-SNE prikaza na slici 3.4 vidimo da se podaci grupiraju u gornjem lijevom kutu za podfamiliju 1, a u donjem desnom kutu za podfamiliju 2. Dio podataka koji pripada podfamiliji 2 nešto je bliži onima iz podfamilije 1, a to može biti povezano uz evoluciju samih proteina i način njihovog nastanka, kao i uz samu svrhu, odnosno zadaću tih proteina. Ipak, i unatoč tome vidljiva je jasna separacija proteina temeljena na top 10 pozicija prema F-faktoru, a koji su prikazani u tablici 3.2. Istaknimo još jednom kako smo dobili četiri pozicije koje su značajne s preko 99% vjerojatnosti, a to su redom pozicije $532, 241, 65$ i 343. Primijetimo da su sve navedene pozicije ispale značajne i u prethodnom razmatranju primjenom formule 3.1. Ono što je posebno interesantno je da smo korištenjem formule 3.1 dobili nešto bolji t-SNE prikaz, odnosno podaci su se bolje razdvojili prema podfamilijama.

Korištenjem formule 3.2, $c = 2, p = 0.8173$

Na razini značajnosti od 1% ne možemo odbaciti pretpostavku o tome da pripadni podaci prate F-distribuciju. U formuli 3.2 korištenjem $c = 2$ dobijamo da je p-vrijednost Kolmogorov-Smirnovljevog testa $p = 0.8173$. U nastavku slijede grafički i tablični prikazi:



Slika 3.5: F-faktori



Slika 3.6: t-SNE prikaz

Pozicija	241	532	65	343	345	18	377	261	245	528
F-faktor	3.37	3.37	3.32	3.28	3.14	2.97	2.95	2.93	2.89	2.89

Tablica 3.3: Top 10 pozicija

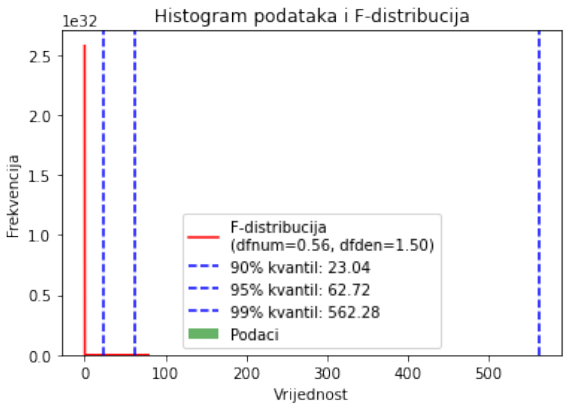
U ovom pristupu posebno je zanimljivo kako smo u t-SNE prikazu dobili dva kriva grupiranja, odnosno dva proteina koja pripadaju podfamiliji 1 prikazana su u blizini proteina koji pripadaju podfamiliji 2. Razlog tome može biti prevelika konstanta c koja je korištena u formuli 3.1. Primijetimo kako u ovom pristupu nismo dobili niti jednu poziciju koja je značajna sa 99% vjerojatnosti što je vidljivo u tablici 3.3. Imamo ukupno 8 pozicija koje su značajne sa 95% vjerojatnosti i među njima se nalaze najznačajnije pozicije iz prethodih razmatranja kada smo koristili formulu 3.1 i formulu 3.2 za $c = 1.5$ Također, istaknimo kako su u svim priloženim slučajevima proteini iz podfamilije 1 manje raspršeni od onih iz podfamilije 2.

Ciklaze

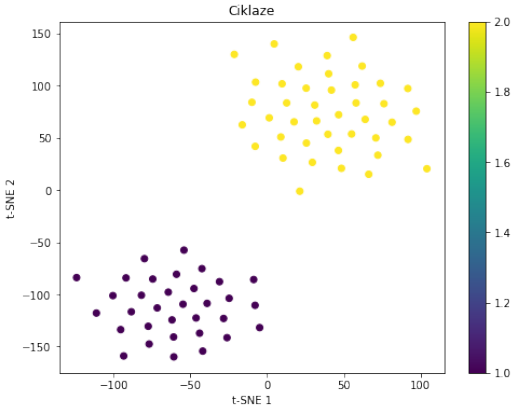
U nastavku navodimo rezultate dobijene promatranjem ciklaza. U formuli 3.2 korištenjem konstanti $c = 1.5$ i $c = 2$ dobijeni su podaci koji s velikom vjerojatnosti prate F-distribuciju budući da su dobivene p-vrijednosti Kolmogorov-Smirnovljevog testa iznosile 0.4250 za $c = 1.5$ i 0.1392 za $c = 2$. Također, valja istaknuti kako smo korištenjem konstante $c = 1.5$ dobili pozicije koje su nam značajne s vjerojatnosti od preko 99%, a one su vidljive u tablici 3.5. Riječ je redom o pozicijama $1632, 1488, 1634, 1497, 654$ i 1635. Prelaskom na veće konstante c uočeno je da se p-vrijednost smanjuje, ali i da gubimo na značajnosti pozicija. Korištenjem formule 3.1 dobijamo da podaci s velikom vjerojatnosti ne prate F-distribuciju (p-vrijednost Kolmogorov-Smirnovljevog testa je manja od 10^{-4}). Ipak, zanimljivo je istaknuti kako u ciklazama (za razliku od kinaza) ovakav pristup ne daje niti jednu 95% značajnu poziciju, već samo dvije 90% značajne pozicije, a to su pozicije 1488 i 1634 (primijetimo da su te pozicije već navedene korištenjem $c = 1.5$).

Korištenjem formule 3.1, $p < 10^{-4}$

Na svim standardnim razinama značajnosti odbacujemo pretpostavku o tome da pripadni podaci prate F-distribuciju. Korištenjem formule 3.1 dobivamo da je p-vrijednost Kolmogorov-Smirnovljevog testa $p < 10^{-4}$. U nastavku slijede grafički i tablični prikazi:



Slika 3.7: F-faktori



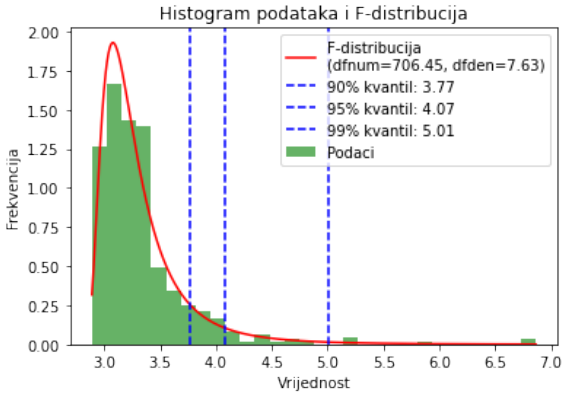
Slika 3.8: t-SNE prikaz

Pozicija	1488	1634	1635	1535	1632	1630	1536	1517	1541	1636
F-faktor	79.07	56.98	42.08	30.33	23.26	20.44	20.36	17.50	15.24	13.26

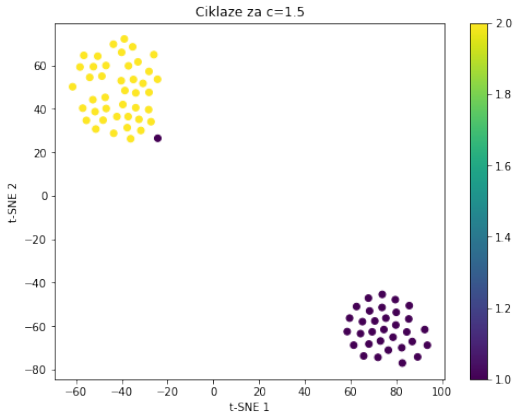
Tablica 3.4: Top 10 pozicija

Korištenjem formule 3.2, $c = 1.5, p = 0.4250$

Na razini značajnosti od 1% ne možemo odbaciti pretpostavku o tome da pripadni podaci prate F-distribuciju. U formuli 3.2 korištenjem $c = 1.5$ dobijamo da je p-vrijednost Kolmogorov-Smirnovljevog testa $p = 0.4250$. U nastavku slijede grafički i tablični prikazi:



Slika 3.9: F-faktori



Slika 3.10: t-SNE prikaz

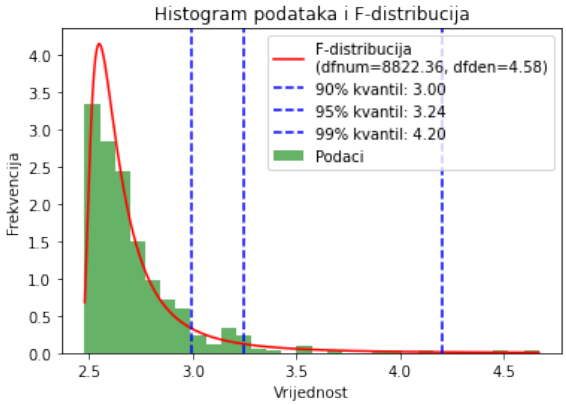
Pozicija	1632	1488	1634	1497	654	1635	1515	1535	1476	1541
F-faktor	6.86	6.85	5.87	5.27	5.21	5.20	4.78	4.68	4.62	4.55

Tablica 3.5: Top 10 pozicija

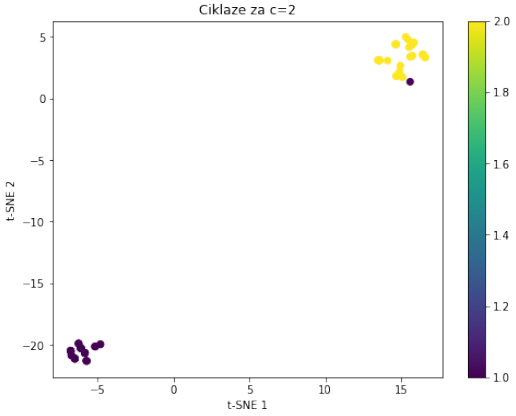
Iz t-SNE prikaza na slici 3.10 jasno se vidi separacija dvije podfamilije ciklaza, s time da se primijeti i jedna kriva separacija budući da je protein koji pripada podfamiliji 1 prikazan u blizini podataka koji pripadaju podfamiliji 2. Podaci su jasnije razdijeljeni nego u slučaju s kinazama, formiraju oblik nalik kugli i nisu toliko raspršeni kao kod kinaza. Zanimljivo je kako u t-SNE prikazu na slici 3.8 nemamo pogrešnih separacija, ali isto tako podaci su znatno raspršeniji i bliži jedni drugima što može biti posljedica toga da nemamo niti jednu 95% značajnu poziciju.

Korištenjem formule 3.2, $c = 2$, $p = 0.1392$

Na razini značajnosti od 1% ne možemo odbaciti pretpostavku o tome da pripadni podaci prate F-distribuciju. U formuli 3.2 korištenjem $c = 2$ dobijamo da je p-vrijednost Kolmogorov-Smirnovljevog testa $p = 0.1392$. U nastavku slijede grafički i tablični prikazi:



Slika 3.11: F-faktori



Slika 3.12: t-SNE prikaz

Pozicija	1632	1488	1497	654	1634	1515	1635	1476	1528	1519
F-faktor	4.67	4.48	4.10	3.97	3.92	3.65	3.55	3.53	3.53	3.38

Tablica 3.6: Top 10 pozicija

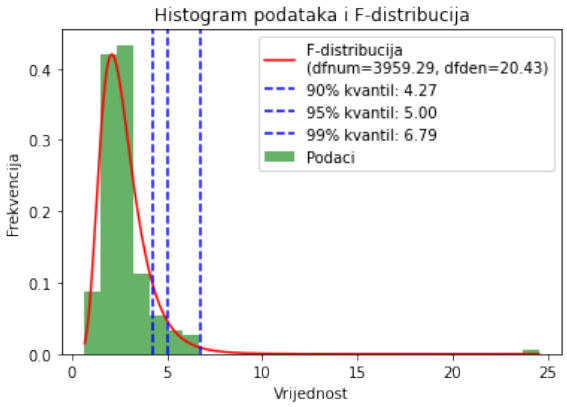
Iz tablice 3.6 vidljivo je kako korištenjem konstante $c = 2$ dobijamo dvije pozicije koje su značajne s vjerojatnošću od preko 99%, a to su pozicije 1632 i 1488 koje su nam ispale značajne na istoj razini i korištenjem konstante $c = 1.5$. Ponovno smo, kao kod kinaza, povećanjem konstante c smanjili broj značajnih pozicija, a razlika je u tome što smo kod ciklaza ujedno smanjili i p-vrijednost Kolmogorov-Smirnovljevog testa o pripadnosti F-distribuciji, što nam nije bio slučaj kod kinaza. Na t-SNE prikazu na slici 3.12 vidljivo je kako imamo jednu pogrešnu separaciju kao što je bio slučaj i korištenjem konstante $c = 1.5$ u formuli 3.2. Podaci su još jasnije separirani u odnosu na prije, i dalje formiraju oblik nalik kugli, a zanimljivo je da je ona bitno manjeg radijusa nego što je to slučaj u prethodnim razmatranjima korištenjem formule 3.1 i 3.2 za konstantu $c = 1.5$.

Acil transferaze (AT)

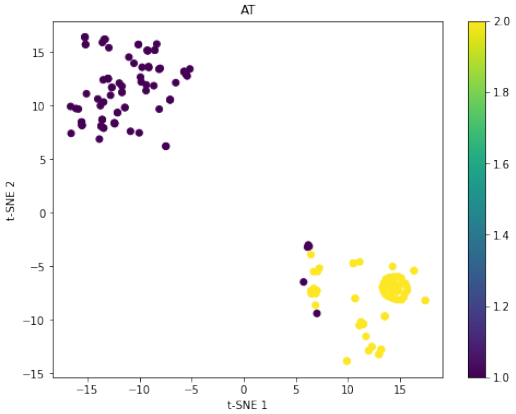
U nastavku navodimo rezultate dobijene promatranjem AT-a. U formuli 3.2 korištenjem konstanti $c = 1.5$ i $c = 2$ dobijeni su podaci koji s velikom vjerojatnosti prate F-distribuciju budući da su dobivene p-vrijednosti Kolmogorov-Smirnovljevog testa iznosile 0.3024 za $c = 1.5$ i 0.1856 za $c = 2$. Također, valja istaknuti kako smo korištenjem konstante $c = 1.5$ dobili poziciju koja nam je značajna s vjerojatnosti od preko 99%, a ona je vidljiva u tablici 3.8. Riječ je o poziciji 212. Prelaskom na veće konstante c uočeno je da se p-vrijednost smanjuje, ali i da gubimo na značajnosti pozicija. Korištenjem formule 3.1 dobijamo da podaci s velikom vjerojatnosti ne prate F-distribuciju (p-vrijednost Kolmogorov-Smirnovljevog testa je 0.0021). Ipak, zanimljivo je istaknuti kako u AT-u ovakav pristup daje istu značajnu poziciju kao i pristupi kada koristimo formulu 3.2 za konstante $c = 1.5$ i $c = 2$.

Korištenjem formule 3.1, $p = 0.0021$

Na svim standardnim razinama značajnosti odbacujemo pretpostavku o tome da pripadni podaci prate F-distribuciju. Korištenjem formule 3.1 dobivamo da je p-vrijednost Kolmogorov-Smirnovljevog testa $p = 0.0021$. U nastavku slijede grafički i tablični prikazi:



Slika 3.13: F-faktori



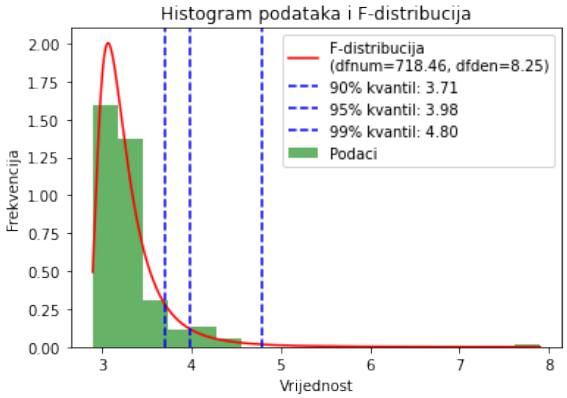
Slika 3.14: t-SNE prikaz

Pozicija	212	143	73	264	71	221	224	105	72	12
F-faktor	24.54	6.37	6.34	6.25	6.10	5.45	5.40	5.32	5.29	5.00

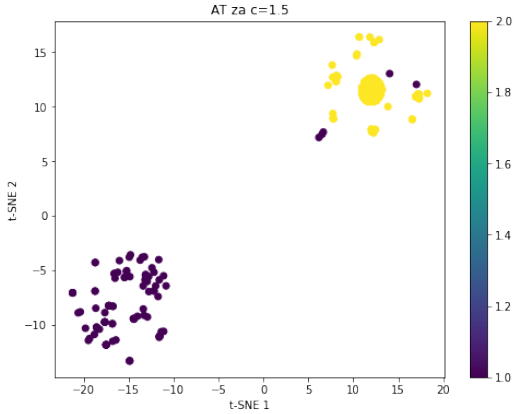
Tablica 3.7: Top 10 pozicija

Korištenjem formule 3.2, $c = 1.5, p = 0.3024$

Na razini značajnosti od 1% ne možemo odbaciti pretpostavku o tome da pripadni podaci prate F-distribuciju. U formuli 3.2 korištenjem $c = 1.5$ dobijamo da je p-vrijednost Kolmogorov-Smirnovljevog testa $p = 0.3024$. U nastavku slijede grafički i tablični prikazi:



Slika 3.15: F-faktori



Slika 3.16: t-SNE prikaz

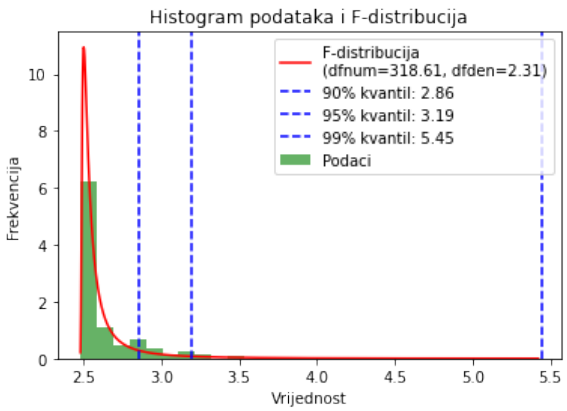
Pozicija	212	72	73	130	143	264	224	234	12	210
F-faktor	7.90	4.37	4.36	4.35	4.28	4.17	4.17	4.16	4.08	4.05

Tablica 3.8: Top 10 pozicija

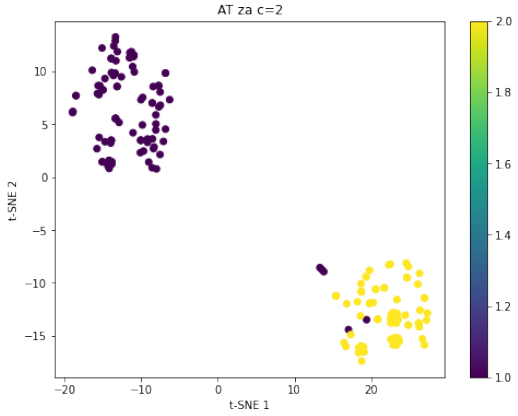
Kao što je i uvodno navedeno, jedina 99% značajna pozicija je 212. Na priloženim t-SNE prikazima vidljivo je da se podaci grupiraju dijagonalno i da su vidljivo separirani, kao i da imamo četiri krive separacija budući da su neki proteini koji pripadaju podfamiliji 1 prikazani u blizini onih koji pripadaju podfamiliji 2. Zanimljivo je primijetiti kako su u oba t-SNE prikaza podaci podjednako raspršeni što nije bio slučaj u dosadašnjem razmatranju kinaza i ciklaza.

Korištenjem formule 3.2, $c = 2$, $p = 0.1856$

Na razini značajnosti od 1% ne možemo odbaciti pretpostavku o tome da pripadni podaci prate F-distribuciju. U formuli 3.2 korištenjem $c = 2$ dobijamo da je p-vrijednost Kolmogorov-Smirnovljevog testa $p = 0.1856$. U nastavku slijede grafički i tablični prikazi:



Slika 3.17: F-faktori



Slika 3.18: t-SNE prikaz

Pozicija	212	72	130	234	73	143	210	264	221	149
F-faktor	5.42	3.53	3.48	3.32	3.29	3.23	3.18	3.15	3.15	3.14

Tablica 3.9: Top 10 pozicija

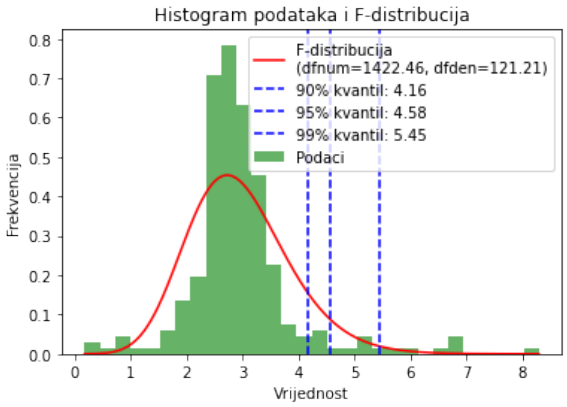
Korištenjem konstante $c = 2$ prema F-faktoru i dalje odskače pozicija 212, ali ona nam u ovom razmatranju nije 99% značajna (iako je dosta blizu te značajnosti). Ponovno smo povećanjem konstante c izgubili na značajnosti i opala nam je p-vrijednost u Kolmogorov-Smirnovljevom testu. Ipak, istaknimo kako i dalje imamo četiri krive separacije što je i očekivano s obzirom da pozicija 212 u sva tri slučaja značajno odskače po vrijednosti F-faktora od ostalih pozicija. Na kraju, raspršenost je i u ovom slučaju podjednaka kao u prethodnim razmatranjima AT-a što bi se isto moglo argumentirati time što jedna pozicija vidno odskače od ostalih po vrijednosti F-faktora.

MLDH

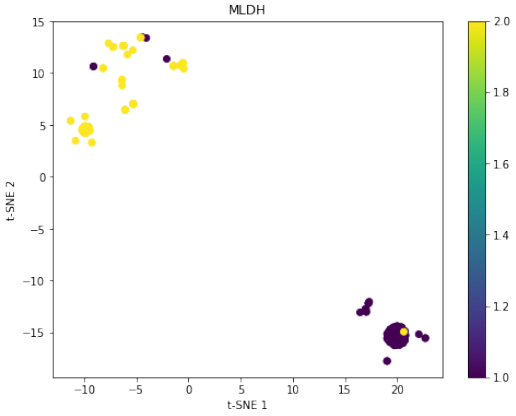
U nastavku navodimo rezultate dobijene promatranjem MLDH-a. U formuli 3.2 korištenjem konstanti $c = 1.5$ i $c = 2$ dobijeni su podaci koji s velikom vjerojatnosti prate F-distribuciju budući da su dobivene p-vrijednosti Kolmogorov-Smirnovljevog testa iznosile 0.4443 za $c = 1.5$ i 0.4003 za $c = 2$. Također, valja istaknuti kako smo korištenjem konstanti $c = 1.5$ i $c = 2$ dobili pozicije koje su nam značajne s vjerojatnosti od preko 99%, a one su vidljive u tablicama 3.11 i 3.12. Riječ je o pozicijama $144, 308, 169, 214$ koje se pojavljuju kao značajne za obe konstante c. Prelaskom na veće konstante c uočeno je da se p-vrijednost smanjuje, ali ne i da gubimo na značajnosti pozicija. Korištenjem formule 3.1 dobijamo da podaci s velikom vjerojatnosti ne prate F-distribuciju (p-vrijednost Kolmogorov-Smirnovljevog testa je 0.0002). Ipak, zanimljivo je istaknuti kako u MLDH-u ovakav pristup daje tri od četiri prethodno navedene značajne pozicije.

Korištenjem formule 3.1, $p = 0.0002$

Na svim standardnim razinama značajnosti odbacujemo pretpostavku o tome da pripadni podaci prate F-distribuciju. Korištenjem formule 3.1 dobivamo da je p-vrijednost Kolmogorov-Smirnovljevog testa $p = 0.0002$. U nastavku slijede grafički i tablični prikazi:



Slika 3.19: F-faktori



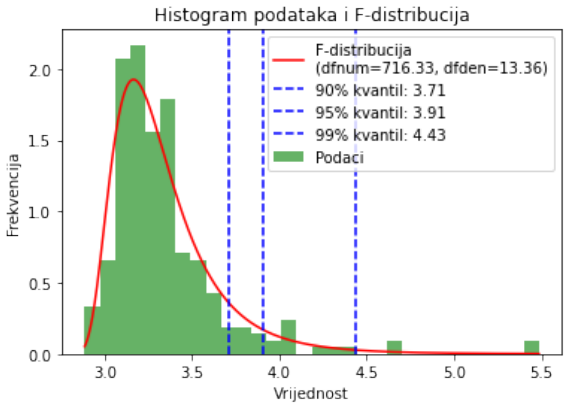
Slika 3.20: t-SNE prikaz

Pozicija	144	148	169	308	300	193	143	246	214	215
F-faktor	8.30	6.84	6.71	6.70	6.66	5.93	5.73	5.55	5.32	5.24

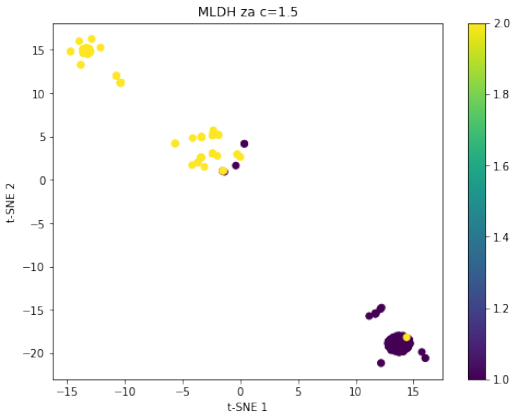
Tablica 3.10: Top 10 pozicija

Korištenjem formule 3.2, $c = 1.5, p = 0.4443$

Na razini značajnosti od 1% ne možemo odbaciti pretpostavku o tome da pripadni podaci prate F-distribuciju. U formuli 3.2 korištenjem $c = 1.5$ dobijamo da je p-vrijednost Kolmogorov-Smirnovljevog testa $p = 0.4443$. U nastavku slijede grafički i tablični prikazi:



Slika 3.21: F-faktori



Slika 3.22: t-SNE prikaz

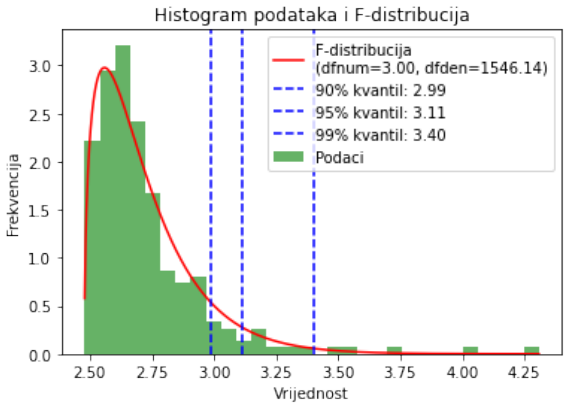
Pozicija	144	308	169	214	143	309	193	215	148	58
F-faktor	5.48	5.47	4.66	4.63	4.37	4.34	4.22	4.08	4.06	4.04

Tablica 3.11: Top 10 pozicija

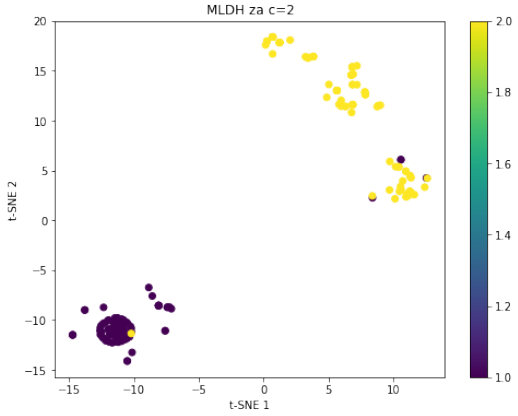
Na priloženim t-SNE prikazima vidljivo je da su proteini vezani za podfamiliju 2 raspršeniji od onih vezanih za podfamiliju 1. U oba prikaza imamo četiri krive separacije što je i očekivano s obzirom na podatke prikazane u tablicama i budući da se većina značajnih pozicija ponavlja. Ipak, istaknimo kako je u dosadašnjim razmatranjima ovo prvi put da imamo i pogrešno separiran protein koji pripada podfamiliji 2 (to nije bio slučaj u kinazama, ciklazama i AT-u).

Korištenjem formule 3.2, $c = 2$, $p = 0.4003$

Na razini značajnosti od 1% ne možemo odbaciti pretpostavku o tome da pripadni podaci prate F-distribuciju. U formuli 3.2 korištenjem $c = 2$ dobijamo da je p-vrijednost Kolmogorov-Smirnovljevog testa $p = 0.4003$. U nastavku slijede grafički i tablični prikazi:



Slika 3.23: F-faktori



Slika 3.24: t-SNE prikaz

Pozicija	308	144	214	169	309	143	119	193	256	79
F-faktor	4.31	4.06	3.71	3.54	3.49	3.39	3.29	3.23	3.18	3.18

Tablica 3.12: Top 10 pozicija

Kao i u dosadašnjim razmatranjima imamo četiri krivo separirana proteina, tri iz podfamilije 1 i jednoga iz podfamilije 2. U značajnim pozicijama nije došlo do velikih promjena, a usporedba se može napraviti preko tablice 3.12. Povećanjem konstante c nismo izgubili na značajnosti pozicija, već smo dapače dobili jednu novu 99% značajnu poziciju u odnosu na korištenje konstante $c = 1.5$ u formuli 3.2.

U ovom poglavlju korišteni su t-SNE grafovi koji su bolje opisani u [9] i [6].

3.3 Usporedba s prethodnim istraživanjima

U ovom poglavlju, za svaku od četiri promatrane familije proteina, donosimo usporedbu top 10 pozicija po vrijednosti F-faktora sa top 10 pozicija po značajnosti u prethodnim istraživanjima, a posebno onima vezanim uz članak [4].

U dosadašnjim istraživanjima postoje dva rangiranja pozicija po značajnosti, a usko su vezana uz dva različita pristupa u kojima su korištene različite supstitucijske matrice - Blosum (vidi [16]) i PAM (vidi [23]). Razlika u supstitucijskim matricama dovodi do razlika u rangiranju, a u nastavku navodimo usporedbu s oba od navedenih pristupa.

Redni broj	**3.1**	**3.2,** $c = 1.5$	**3.2,** $c = 2$	**Blosum**	**PAM**
1	65	532	241	65	343
2	343	241	532	343	542
3	345	65	65	542	65
4	532	343	343	532	381
5	241	345	345	63	345
6	261	261	18	381	532
7	101	18	377	345	344
8	376	342	261	344	63
9	542	344	245	337	245
10	245	528	528	342	376

Tablica 3.13: Kinaze, usporedba top 10 značajnih pozicija

Redni broj	**3.1**	**3.2,** $c = 1.5$	**3.2,** $c = 2$	**Blosum**	**PAM**
1	1488	1632	1632	1634	1634
2	1634	1488	1488	1630	1517
3	1635	1634	1497	1517	1533
4	1535	1497	654	1533	1440
5	1632	654	1634	1636	1536
6	1630	1635	1515	1440	1636
7	1536	1515	1635	1536	1580
8	1517	1535	1476	1497	1489
9	1541	1476	1528	1580	1476
10	1636	1541	1519	1617	1540

Tablica 3.14: Ciklaze, usporedba top 10 značajnih pozicija

Redni broj	**3.1**	**3.2,** $c = 1.5$	**3.2,** $c = 2$	**Blosum**	**PAM**
1	212	212	212	212	212
2	143	72	72	73	82
3	73	73	130	105	210
4	264	130	234	71	73
5	71	143	73	210	105
6	221	264	143	264	209
7	224	224	210	221	71
8	105	234	264	209	234
9	72	12	221	82	264
10	12	210	149	234	221

Tablica 3.15: AT, usporedba top 10 značajnih pozicija

Redni broj	**3.1**	**3.2,** $c = 1.5$	**3.2,** $c = 2$	**Blosum**	**PAM**
1	144	144	308	148	148
2	148	308	144	300	300
3	169	169	214	308	308
4	308	214	169	144	144
5	300	143	309	296	296
6	193	309	143	143	62
7	143	193	119	303	143
8	246	215	193	170	97
9	214	148	256	246	215
10	215	58	79	192	60

Tablica 3.16: MLDH, usporedba top 10 značajnih pozicija

U priloženim tablicama, stupac s oznakom **3.1** predstavlja rangiranje pozicija po vrijednosti F-faktora gdje smo za računanje F-faktora koristili formulu 3.1, dok stupci sa oznakama **3.2** i pripadnom konstantom c predstavljaju rangiranje pozicija gdje smo za računanje F-faktora koristili formulu 3.2.

Bibliografija

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

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

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

[4] P. Goldstein, J. Zucko, D. Vujaklija, A. Kriško, D. Hranueli, P.F. Long, C. Etchebest, B. Basrak i J. Cullum, *Clustering of protein domains for functional and evolutionary studies*, BioMed Central (2009), http://www.biomedcentral.com/1471-2105/10/335.

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

[6] G. Hinton i S.Roweis, *Stochastic Neighbor Embedding*, Neural Information Processing Systems (2002).

[7] M. Iveković, *Traženje proteinskih motiva i klasifikacija*, 2022, Diplomski rad, Sveučilište u Zagrebu, Prirodoslovno-matematički fakultet (Matematički odsjek).

[8] Leksikografski zavod Miroslav Krleža, *Bjelančevine*, Hrvatska enciklopedija, mrežno izdanje, 2013-2024, https://www.enciklopedija.hr/clanak/bjelancevine, [Pristupljeno 7.11.2024.].

[9] M. Pathak, *Introduction to t-SNE*, 2018, https://www.datacamp.com/community/tutorials/introduction-t-sne.

[10] J. Radnić, *Klasifikacija proteinskih fragmenata*, Diplomski rad, Sveučilište u Zagrebu, Prirodoslovno-matematički fakultet (Matematički odsjek), 2023.

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

[12] Science direct, *Acyltransferase*, https://www.sciencedirect.com/topics/medicine-and-dentistry/acyltransferase, (Studeni 2024).

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

[14] I. Višek, *Clustering i klasifikacija proteinskih nizova*, 2022, Diplomski rad, Sveučilište u Zagrebu, Prirodoslovno-matematički fakultet (Matematički odsjek).

[15] Wikipedia, *Aminokiselina*, https://bs.wikipedia.org/wiki/Aminokiselina, (Studeni 2024).

[16] ———, *BLOSUM*, https://en.wikipedia.org/wiki/BLOSUM, (Studeni 2024).

[17] ———, *Chi-squared distribution*, https://en.wikipedia.org/wiki/Chi-squared_distribution, (Studeni 2024).

[18] ———, *Ciklaza*, https://en.wikipedia.org/wiki/Adenylyl_cyclase, (Studeni 2024).

[19] ———, *Enzim*, https://hr.wikipedia.org/wiki/Enzim, (Studeni 2024).

[20] ———, *F-distribution*, https://en.wikipedia.org/wiki/F-distribution, (Studeni 2024).

[21] ———, *Kinaza*, https://bs.wikipedia.org/wiki/Kinaza, (Studeni 2024).

[22] ———, *Multiple sequence alignment*, https://en.wikipedia.org/wiki/Multiple_sequence_alignment, (Studeni 2024).

[23] ———, *Point accepted mutation*, https://en.wikipedia.org/wiki/Point_accepted_mutation, (Studeni 2024).

Sažetak

U ovom diplomskom radu promatrane su četiri proteinske familije: kinaze, ciklaze, acil transferaze (AT) i MLDH. Problem koji se obrađuje u ovom radu vezan je za pronalazak značajnih pozicija u proteinskim poravnanjima unutar pojedine proteinske familije, a koje su odlučujuće u klasifikaciji tih proteina na dvije podfamilije. Pripadnost podfamilijima u ovom radu nam je bila poznata, cilj je bio dobiti rangiranje pozicija prema vrijednostima F-faktora, a njegovo računanje vršili smo na tri različita načina. Primjenom dvije od te tri metode niz dobijenih F-faktora bio je F distribuiran, dok u trećoj nismo dobijali F distribuiranost podataka, ali ipak su proizlazili slični zaključci. Kroz rad je opisano kako smo sa niza aminokiselina prešli u vektorski prostor i koji je pristup korišten u pripremi podataka za obradu. U konačnici, usporedbom s dosadašnjim istraživanjima utvrđene su brojne sličnosti, ali i neka nova zapažanja, kao i drugačija rangiranja. Također, istaknuli smo pozicije koje su značajne s vjerojatnostima od 99%, 95% i 90%. Prikazom na t-SNE grafovima utvrđeno je kako smo promatranjem top 10 pozicija prema rangiranju dobijali jasne separacije sa svega nekoliko krivo pridruženih proteina.

Summary

In this thesis, four protein families were studied: kinases, cyclases, acyl transferases (AT), and MLDH. The primary focus of this work is identifying significant positions in protein alignments within each protein family that are crucial for classifying these proteins into two subfamilies. The subfamily membership was known in advance, and the goal was to rank the positions based on F-factor values, calculated using three different methods. Using two of these methods, the resulting F-factors followed an F-distribution, while in the third method, the F-distribution was not observed; however, similar conclusions were still reached. The thesis describes the process of transforming amino acid sequences into a vector space and the approach used for preparing the data for analysis. Finally, a comparison with previous research revealed numerous similarities, as well as some new observations and different rankings. We also highlighted positions that are significant with probabilities of 99%, 95%, and 90%. Visualization using t-SNE graphs showed that by focusing on the top 10 positions according to the rankings, clear separations were achieved, with only a few proteins misclassified.

Životopis

Rođen sam u Sinju, 20. kolovoza 1999. godine. Pohađao sam OŠ Marka Marulića nakon čega sam srednjoškolsko obrazovanje nastavio u splitskoj III. gimnaziji, popularno zvanom MIOC-u. Sudjelovao sam na brojnim županijskim i državnim natjecanjima iz prirodoslovnih predmeta gdje sam osjetio sve veći interes prema matematici, pa se samim time prirodno nametnuo moj sljedeći korak, a to je upisivanje zagrebačkog PMF-a. Zvanje sveučilišnog prvostupnika stekao sam 2021. godine, a iste godine sam upisao diplomski sveučilišni studij Matematička statistika na istom fakultetu.

Tijekom cijelog svog obrazovanja usko sam vezan uz sport, a za vrijeme diplomskog studija obnašao sam dužnost potpredsjednika NK Junak Sinj gdje provodim veliku većinu slobodnog vremena. Tijekom cijelog vremena studiranja radio sam kao instruktor iz matematike i fizike na portalu eMatematika (2200 sati održanih instrukcija), a u posljednjih godinu dana u RBA banci kao analitičar.