Funkcionalna specifičnost za proteinske familije

Šimić, Nino

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Nino Šimić

FUNKCIONALNA SPECIFIČNOST ZA PROTEINSKE FAMILIJE

Diplomski rad

Voditelj rada:
doc. dr. sc. Pavle Goldstein

Zagreb, siječanj, 2025.

Ovaj diplomski rad obranjen je dana _____ pred ispitnim povjerenstvom u sastavu:

1. _____, predsjednik
2. _____, član
3. _____, član

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Zahvaljujem mentoru doc. dr. sc. Pavlu Goldsteinu na vrijednom prenesenom znanju i savjetima, laganim dogovorima i kvalitetnom vodstvu. Od srca hvala mojim roditeljima, bratu, djevojci, cimeru i svim prijateljima i rođacima koji su bili uz mene, pružali mi veliku podršku, i uljepšali mi studentsko vrijeme.

Sadržaj

Uvod

U ovom radu analizirana su poravnanja 5 proteinskih, tj. enzimskih familija. To su redom Acil transferaze (preciznije, AT-domene u poliketidnim sintetazama), familija malatnih i laktatnih dehidrogenaza, ciklaze, kinaze i ketoreduktaze (preciznije, KR-domene u poliketidnim sintetazama). Svaka od tih 5 familija dijeli se na 2 grupe, odnosno podfamilije. Analiza provedena u ovom radu pokušava dati odgovor na važno otvoreno biološko, odnosno bioinformatičko pitanje, a to je pitanje utvrđivanja ključnih pozicija u proteinskom poravnanju koje značajno utječu na klasifikaciju pripadnih proteina u podfamilije. Rad je strukturiran u 5 velikih cjelina.

U prvom poglavlju uvode se osnovni matematički koncepti vezani uz linearnu algebru, vjerojatnost i statistiku.

Drugo poglavlje predstavlja i objašnjava ključne pojmove iz biologije i bioinformatike, te se u njemu prilažu rezultati faktorske analize iz [13] u vidu vektorskih reprezentacija aminokiselina, na čijim temeljima je omogućena daljnja statistička analiza.

Treće poglavlje donosi objašnjenja pripreme i obrade danih podataka, te se u njemu definira metrika značajnosti pozicija u razdvajanju proteina na podfamilije, nazvana razdvajajućom (split) statistikom (S-statistika). Navodi se i koncept dodavanja buke u podatke, radi povećanja broja analiziranih nizova te posljedično i boljih statističkih svojstava S-statistike.

Četvrto poglavlje obuhvaća rezultate odabranih statističkih metoda na poravnanjima svih 5 promatranih familija. Za svaku od familija prikazat će se najbolje rangirane pozicije a zatim će se ispitati pripadnost S-statistike F-distribuciji, i na temelju nje zaključiti koje su od prikazanih pozicija statistički značajne u separiranju dviju podfamilija. Vizualno će se ocijeniti moć najbolje rangiranih pozicija u klasifikaciji pripadnih proteina te će se usporediti dobiveni rezultati s prethodnim analizama i istraživanjima na istim podacima.

Peto poglavlje će rezimirati rezultate i zaključke iz četvrtog poglavlja.

Poglavlje 1

Matematički pojmovi

U ovom poglavlju definirat ćemo neke osnovne matematičke pojmove i zapisati neke važnije rezultate iz područja linearne algebre, vjerojatnosti i statistike. Definicije i rezultati u obliku propozicija, teorema i korolara su djelomično ili u cijelosti preuzeti iz izvora [14], [18], [24], [20], [16].

1.1 Linearna algebra

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

1. $\alpha + (\beta + \gamma) = (\alpha + \beta) + \gamma, \quad \forall \alpha, \beta, \gamma \in \mathbb{F}$;
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. *Jedan od primjera polja je skup realnih brojeva $\mathbb{R}$ s uobičajenim operacijama zbrajanja i množenja.*

Definicija 1.1.3. *Neka je V neprazan skup na kojem su zadane binarna operacija 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. $\mathbf{u} + \mathbf{v} = \mathbf{v} + \mathbf{u}, \quad \forall \mathbf{u}, \mathbf{v} \in V;$
2. $(\mathbf{u} + \mathbf{v}) + \mathbf{w} = \mathbf{u} + (\mathbf{v} + \mathbf{w}), \quad \forall \mathbf{u}, \mathbf{v}, \mathbf{w} \in V;$
3. *postoji* $\mathbf{0} \in V$ *takav da je* $\mathbf{u} + \mathbf{0} = \mathbf{0} + \mathbf{u} = \mathbf{u}, \quad \forall \mathbf{u} \in V;$
4. *za svaki* $\mathbf{u} \in V$, *postoji* $-\mathbf{u} \in V$ *takav da je* $\mathbf{u} + (-\mathbf{u}) = (-\mathbf{u}) + \mathbf{u} = \mathbf{0};$
5. $\alpha \cdot (\beta \cdot \mathbf{u}) = (\alpha \cdot \beta) \cdot \mathbf{u}, \quad \forall \alpha, \beta \in \mathbb{F}, \forall \mathbf{u} \in V;$
6. $1 \cdot \mathbf{u} = \mathbf{u}, \quad \forall \mathbf{u} \in V;$
7. $\alpha \cdot (\mathbf{u} + \mathbf{v}) = \alpha \cdot \mathbf{u} + \alpha \cdot \mathbf{v}, \quad \forall \alpha \in \mathbb{F}, \forall \mathbf{u}, \mathbf{v} \in V;$
8. $(\alpha + \beta) \cdot \mathbf{u} = \alpha \cdot \mathbf{u} + \beta \cdot \mathbf{u}, \quad \forall \alpha, \beta \in \mathbb{F}, \forall \mathbf{u} \in V.$

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. $\langle x, x \rangle = 0 \iff x = 0;$
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. *Skalarni produkt $\langle \cdot, \cdot \rangle : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ definiran s: $\langle (x_1, ..., x_n), (y_1, ..., y_n) \rangle = \sum_{i=1}^{n} x_i y_i$ zovemo* ***kanonskim*** *skalarnim produktom u $\mathbb{R}^n$.*

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 unitaranom prostoru V ima sljedeća svojstva:*

1. $\|x\| \geq 0$, $\forall x \in V$;
2. $\|x\| = 0 \iff x = 0$;
3. $\|\alpha x\| = |\alpha| \cdot \|x\|$, $\forall \alpha \in \mathbb{F}, \forall x \in V$;
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 $\mathbb{R}^n$, definiranog u napomeni 1.1.6, dana je formulom*

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

Ova norma zove se ***euklidska norma****.*

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

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

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

1. $d(x, y) \geq 0$, $\forall x, y \in V$;
2. $d(x, y) = 0 \iff x = y$, $\forall x, y \in V$;
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 $x = (x_1, ..., x_n)$ i $y = (y_1, ..., y_n)$ proizvoljni vektori u $\mathbb{R}^n$. Metrika na $\mathbb{R}^n$, inducirana euklidskom normom iz definicije 1.1.11, dana je s*

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

*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 ($\mathcal{F} \subseteq \mathcal{P}(\Omega)$) je **$\sigma$-algebra** skupova na Ω ako je:*

1. $\emptyset \in \mathcal{F}$;
2. $A \in \mathcal{F} \implies A^c \in \mathcal{F}$;
3. $A_i \in \mathcal{F}, i \in \mathbb{N} \implies \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

Definicija 1.2.4. *Neka je $\mathcal{F}$ σ-algebra na skupu Ω. Uređen par $(\Omega, \mathcal{F})$ zove se **izmjeriv prostor**.*

Definicija 1.2.5. *Neka je $(\Omega, \mathcal{F})$ izmjeriv prostor. Funkcija $\mathbb{P} : \mathcal{F} \to \mathbb{R}$ je **vjerojatnost** (na $\mathcal{F}$, na Ω) ako vrijedi:*

1. $\mathbb{P}(A) \geq 0$, $\forall A \in \mathcal{F}$;
2. $\mathbb{P}(\Omega) = 1$;
3. $A_i \in \mathcal{F}, i \in \mathbb{N}; A_i \cap A_j = \emptyset$ *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 $\mathbb{P}$ 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 ($\mathcal{A} \subseteq \mathcal{P}(S)$). Sa $\sigma(\mathcal{A})$ označimo najmanju σ-algebru podskupova od S koja sadrži $\mathcal{A}$. Nju nazivamo* ***σ-algebra generirana s $\mathcal{A}$****.*

Definicija 1.2.8. *Neka je s $\mathcal{B}$ 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 proizvoljan skup $B \in \mathcal{B}$, tj. $X^{-1}(\mathcal{B}) \subseteq \mathcal{F}$.*

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****) ako je $X^{-1}(B) \in \mathcal{F}$ za svaki skup $B \in \mathcal{B}^n$, tj. $X^{-1}(\mathcal{B}^n) \subseteq \mathcal{F}$.*

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

$$X = \sum_{i=1}^{n} x_i \mathbb{1}_{A_i},$$

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

$$X = X^+ - X^-, \quad |X| = X^+ + X^-.$$

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

Matematičko očekivanje se definira tzv. Lebesgueovom indukcijom, tj. prvo se definira za jednostavnu slučajnu varijablu, zatim za nenegativnu slučajnu varijablu i konačno, za općenitu slučajnu varijablu.

Neka je $(\Omega, \mathcal{F}, \mathbb{P})$ vjerojatnosni prostor. Označimo s $\mathcal{K}$ skup svih jednostavnih slučajnih varijabli definiranih na Ω. Neka je $X \in \mathcal{K}$ te $X = \sum_{k=1}^{n} x_k \mathbb{1}_{A_k}$, gdje su $A_1, A_2, \ldots, A_n$ 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 konačan ili jednak $+\infty$.

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

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

Definicija 1.2.18. *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. *Sa σ_X ćemo označavati* ***standardnu devijaciju****, tj. pozitivan drugi korijen iz varijance.*

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. *Ponekad se funkcija distribucije, ako je poznato na koju slučajnu varijablu se odnosi, označava samo s F.*

Teorem 1.2.22. *Funkcija distribucije F slučajne varijable X je rastuća i neprekidna zdesna na $\mathbb{R}$ te zadovoljava:*

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

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 $g^{-1}(\mathcal{B}) \subseteq \mathcal{B}$.*

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 $P(X \in D) = 1$.*

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*

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

Ako je X neprekidna slučajna varijabla, tada se funkcija f iz definicije 1.2.25 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 $\mu, \sigma \in \mathbb{R}$ i $\sigma > 0$. 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 $X \sim \mathcal{N}(\mu, \sigma^2)$.

Napomena 1.2.27. *Neprekidna slučajna varijabla X ima* ***standardnu (ili jediničnu) normalnu distribuciju*** *ako je* $X \sim \mathcal{N}(0,1)$*, dakle*

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

Definicija 1.2.28. *Neka su* $X_1, ..., X_n$ *slučajne varijable na vjerojatnosnom prostoru* $(\Omega, \mathcal{F}, \mathbb{P})$*. Kažemo da su* $X_1, ..., X_n$ ***nezavisne*** *ako za proizvoljne* $B_i \in \mathcal{B}$ $(i = 1, ..., n)$ *vrijedi*

$$\mathbb{P}(\{X_1 \in B_1, ..., X_n \in B_n\}) = \mathbb{P}(\bigcap_{i=1}^{n} \{X_i \in B_i\}) = \prod_{i=1}^{n} \mathbb{P}(\{X_i \in B_i\}).$$

Definicija 1.2.29. *Neka su* $Z_1, ..., Z_n$ *nezavisne standardne normalne slučajne varijable, tj.* $Z_i \sim \mathcal{N}(0,1)$ *za i=1,2,...,n. Tada je slučajna varijabla*

$$X = Z_1^2 + ... + Z_n^2$$

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

Definicija 1.2.30. *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*

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

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

Opisna statistika

Ovdje će se definirati neke temeljne statistike koje su potrebne za analizu stvarnih podataka na kojoj će se i temeljiti ovaj rad. Neka su

$$x_1, ..., x_n \tag{1.2}$$

n vrijednosti (opažanja) varijable X koje čine skup podataka. Pretpostavimo da je X numerička varijabla.

Definicija 1.2.31. ***Aritmetička sredina*** *podataka ili uzorka (1.2) je mjera centralne tendencije i definirana je kao*

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

Definicija 1.2.32. ***Varijanca*** *uzorka ili podataka (1.2) je mjera raspršenja podataka i predstavlja prosječno kvadratno odstupanje podataka od njihove aritmetičke sredine i dana je formulom:*

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

Definicija 1.2.33. ***Standardna devijacija*** *uzorka je 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 i podaci

2.1 Biološka pozadina

Protein je izuzetno složena tvar, tj. makromolekula koja se sastoji od aminokiselina, međusobno spojenih peptidnom vezom u lance. Ti lanci se mogu sastojati od nekoliko desetaka do mnogo tisuća aminokiselina.

Aminokiseline ili **aminokarboksilne kiseline** organski su spojevi koji u svojim molekulama sadrže karboksilne skupine (– COOH) i amino-skupine (– NH_2) između kojih je centralni ugljikov atom, a ono po čemu se međusobno razlikuju je bočni lanac vezan na taj centralni ugljikov atom. Aminokiseline su važne hranjive tvari, bitne za život stanice. U proteinima se prirodno javlja 20 različitih (standardnih) aminokiselina:

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

U procesu formiranja peptidne veze između svake dvije aminokiseline, ispušta se po jedna molekula vode, i u polipeptidnom lancu ostaju, sada peptidnom vezom spojeni, tzv. **os-**

taci aminokiselina (engl. *aminoacid residues*). To zapravo znači da kadgod pričamo o pozicijama aminokiselina u proteinu, na svakoj od tih pozicija u proteinu se nalazi ostatak pojedine aminokiseline.

Niz aminokiselina naziva se **primarnom strukturom proteina**, međusobni raspored polipeptidnih lanaca u spiralni oblik (alfa-zavojnica) ili u oblik sličan nabranom listu (beta-struktura proteina) **sekundarnom** strukturom, a sveukupni točan razmještaj atoma u trodimenzionalnom prostoru **tercijarnom** strukturom. Neki se proteini sastoje od više nekovalentno vezanih podjedinica (oligomerni proteini), a međusobni razmještaj podjedinica označava se kao **kvartarna** struktura proteina.

Proteini ili bjelančevine su prirodno prisutni u svim živim organizmima i uključuju mnoge esencijalne biološke spojeve poput enzima, hormona i antitijela. Bjelančevine kataliziraju reakcije u živom organizmu (enzimi), služe za pohranu i prijenos manjih molekula i iona, za pokretanje mišića, za obranu organizma (antitijela) te obavljaju mnoge druge vitalne funkcije. Čak i najjednostavnija bakterijska stanica sadržava oko tisuću, a ljudski organizam više od sto tisuća različitih vrsta bjelančevina.

Biološka funkcija svake bjelančevine vezana je uz njezinu tercijarnu strukturu (odnosno, *nativnu konformaciju*). Gubitak nativne konformacije (denaturacija) ima za posljedicu gubitak biološke funkcije. Dijelovi primarne strukture često su organizirani u prostorno cjelovite domene, kojima je struktura određena dijelovima gena. **Proteinska domena** je regija proteinskog polipeptidnog lanca koja se sama stabilizira i smotava (*engl. protein folding*) neovisno o ostatku proteina i koja često ima specifičnu funkciju. Npr. kod kinaza, jedna domena veže ATP, dok druga prepoznaje ciljni protein za fosforilaciju. U pravilu, domene variraju u duljini od otprilike 50 do 250 aminokiselina. Budući da su neovisno stabilne, domene dvaju proteina mogu biti "zamijenjene" genetičkim inženjeringom. Tako dobivamo proteine s funkcionalnim svojstvima izvedenim iz svakog od originalnih proteina. Takvi proteini nazivaju se **fuzijskim** ili **himernim proteinima**. Srodne domene pojavljuju se u različitim bjelančevinama sličnih funkcija. Tako se čini da su i bjelančevine i geni koji određuju njihovu strukturu građeni modularno, tj. da su sastavljeni od različitih funkcionalnih jedinica (domena) što olakšava, ali i otežava, objašnjenje njihove evolucije. Dakle, proteini sa sličnim funkcijama imaju sličan sastav i slijed aminokiselina. Iako još uvijek nije moguće u potpunosti objasniti sve funkcije proteina iz njegova slijeda aminokiselina, utvrđene korelacije između strukture i funkcije mogu se pripisati svojstvima aminokiselina koje čine proteine.

Enzimi su biološki katalizatori, tvari koje enormno ubrzavaju biokemijske procese u živim organizmima. Enzimi su proizvodi žive stanice, a po kemijskoj su naravi najčešće bjelančevine. Po reakcijama koje kataliziraju, enzimi se (prema broju enzimske komisije, tj. EC broju (*engl. Enzyme Comission number)*) svrstavaju u sedam skupina:

EC broj	Klasa enzima	Reakcije koje kataliziraju
1	Oksidoreduktaze	Reakcije oksidacije i redukcije
2	Transferaze	Prijenos skupina atoma
3	Hidrolaze	Hidrolitičke reakcije
4	Liaze	Eliminacija skupina atoma uz nastanak dvostruke veze
5	Izomeraze	Reakcije izomerizacije
6	Ligaze	Reakcije stvaranja kovalentne veze uz istodobnu hidrolizu ATP-a
7	Translokaze	Kretanje iona ili molekula kroz stanične membrane

Tablica 2.2: Klase enzima

Unutar svake skupine enzimi se dijele po decimalnoj enzimskoj klasifikaciji u dodatne razine. Tvari na koje enzim djeluje nazivaju se njegovim **supstratima**, a naziv enzima obično se tvori dodatkom sufiksa "-aza" nazivu supstrata ili nazivu reakcije, kao što vidimo i u prethodnoj tablici.

Multidomenski proteini su proteini koji na svom polipeptidnom lancu imaju više domena. Kod **multidomenskih enzima** svaka domena odrađuje jedan korak u odvijanju / kataliziranju određenog biokemijskog procesa. Npr. poliketidne sintetaze su familija multidomenskih enzima, odnosno sintetaza (enzima koji kataliziraju sintezu neke biološke tvari). Preciznije, poliketidne sintetaze kataliziraju sintezu poliketida, velike klase sekundarnih metabolita, kod bakterija, gljiva, biljki i nekoliko životinjskih skupina. Kod poliketidnih sintetaza tipa I možemo primijetiti više domena s jasno definiranim funkcijama. Neke od tih domena su AT-domene (acil transferaze) i KR-domene (ketoreduktaze) koje su neke od proteinskih familija koje će se promatrati u ovom radu.

Aktivno mjesto enzima njegov je dio koji direktno sudjeluje u stvaranju veze enzima sa supstratom i u samom katalitičkom procesu. Aktivno mjesto enzima sastoji se od aminokiselina koje formiraju veze sa supstratom (*mjesto vezanja*), i od aminokiselina koje kataliziraju reakciju tog supstrata (*katalitičko mjesto*). Premda aktivno mjesto zauzima samo oko 10 do 20 % volumena enzima, najvažniji je dio jer direktno katalizira kemijsku reakciju. Obično se sastoji od 3 do 4 aminokiseline, dok su ostale aminokiseline unutar proteina potrebne da bi održale tercijarnu strukturu enzima. **Ligand** je bilo koja molekula koja se veže na neko posebno mjesto u proteinu, enzimu ili receptoru i tako formira kompleks s tom makromolekulom u svrhu nekog biokemijskog procesa. U kontekstu enzima, supstrat se može interpretirati kao posebni tip liganda koji se veže na aktivno mjesto enzima. **Kemijska specifičnost** je sposobnost mjesta vezanja (odnosno aktivnog mjesta) enzima da veže specifične ligande (supstrate). Većina enzima ima aktivno mjesto visoke kemijske specifičnosti, što znači da je skup liganada (supstrata) za koje se to mjesto veže prilično limitiran. Drugim riječima, to aktivno mjesto će se htjeti vezati samo za neke, posebne, supstrate, i tako katalizirati posebnu reakciju.

Sada će se navesti i opisati 5 proteinskih, odnosno enzimskih, familija, koje će se

promatrati u radu.

1. **Acil transferaze** su enzimi koji kataliziraju reakcije *acilacije* gdje prenose acil grupu do supstrata. Ključna su klasa enzima koja sudjeluje u biosintezi važnih organskih spojeva, kao što su masne kiseline i lipidi, koji igraju bitnu ulogu u skladištenju energije i izgradnji staničnih membrana. Poremećaji u funkciji acil transferaza mogu nepovoljno utjecati na lipidni metabolizam i u nekim slučajevima dovesti do ateroskleroze (zadebljanje i oštećenje stijenke krvnih žila). Acil transferaze koje će se promatrati su iz poliketidnih sintetaza tipa I gdje dolaze u obliku integriranih domena (**AT-domene**) kao što smo i naveli ranije. Većina domena bira ili **C2** ili **C3** jedinicu, ovisno o tome je li koenzim A reagirao s malonskom kiselinom (→ supstrat malonil-CoA) ili s metilmalonskom kiselinom (→ supstrat metilmalonil-CoA), redom. Upravo po tome je familija AT-domena i podijeljena u dvije podfamilije.

2. **Kinaze** su također podklasa transferaza i sudjeluju u transferu fosfatne grupe s visokoenergijskih molekula adenozin trifosfata (ATP) koje doniraju fosfat, na specifične supstrate. Ovaj proces se naziva *fosforilacija*, a njezin rezultat je fosforiliziran supstrat i adenozin difosfat (ADP). U radu će se promatrati **protein-kinaze**, tj. kinaze koje prenose fosfatnu grupu na proteine i tako ih selektivno modificiraju. Fosforilacija proteina kinazama važan je mehanizam za komunikaciju signala unutar stanice (*transdukcija*) i regulaciju stanične aktivnosti (npr. stanične diobe). Familija protein-kinaza može se podijeliti u **serin/treonin kinaze**, koje fosforiliraju hidroksilne grupe serina i treonina, i **tirozin kinaze**, koje fosforiliraju ostatke aminokiseline tirozin u sklopu specifičnog proteina unutar stanice.

3. **Malatne dehidrogenaze (MDH)** i **laktatne dehidrogenaze (LDH)** dvije su podfamilije velike familije dehidrogenaza koja pripada enzimskoj klasi oksidoreduktaza. Malatna dehidrogenaza je enzim koji reverzibilno katalizira oksidaciju malata u oksaloctenu kiselinu koristeći redukciju NAD^+ u NADH. Ova reakcija je dio puno metaboličkih puteva, uključujući Krebsov ciklus. Laktatna dehidrogenaza je enzim prisutan u skoro svim ljudskim stanicama. Ona katalizira promjenu piruvata u laktat, i obratno, tako transformirajući NAD^+ u NADH, i obratno.

4. **Ciklaze** su podklasa liaza koja katalizira neku kemijsku reakciju da bi dobila ciklički spoj kao rezultat. U ovom radu promatrat će se dvije podfamilije ciklaza koje obje kao pripadajuće supstrate imaju nukleozid trifosfate. Jedna podfamilija pretvara adenozin trifosfat (**ATP**) u ciklički adenozin monofosfat (cAMP), a druga pretvara gvanozin trifosfat (**GTP**) u ciklički gvanozin monofosfat (cGMP). Produkti tih reakcija (cAMP i cGMP) sekundarni su glasnici unutar stanice. Ciklički AMP je odgovoran za intrastaničnu transdukciju, dok je cGMP zaslužan za aktivaciju intrastaničnih protein-kinaza i time relaksira mišiće.

5. **Ketoreduktaze** su podklasa oksidoreduktaza, a u ovom radu će se promatrati ketoreduktaze u domenama poliketidnih sintetaza tipa I (**KR-domene**), otkuda promatramo i gore spomenute AT-domene. Analizirat će se dvije podfamilije KR-domena koje koriste NADPH da bi stereospecifično reducirali prvotno formiranu keto grupu u hidroksilnu grupu.

Proteinsko poravnanje (engl. *protein alignment*) predstavlja potencijalnu rekonstrukciju evolucijske povezanosti. Naime, proteinsko poravnanje omogućuje usporedbu dva ili više proteina (niza aminokiselina) radi prepoznavanja sličnih područja. Ta zajednička područja često sugeriraju evolucijsku povezanost, slične funkcije ili strukturna obilježja među promatranim proteinima. Ova analiza može otkriti koje aminokiseline u različitim proteinima imaju slične ili istovjetne uloge, te koje pozicije ili područja imaju najveći značaj u razjedinjavanju ili klasificiranju promatranih proteina iz iste familije u pojedine podfamilije. Sve to olakšava razumijevanje biološkog značenja i funkcija tih proteina.

Praznina ("**crtica**") u biološkom (pa i proteinskom) nizu može se opisati kao odsutnost (odnosno prisutnost) regije koja je (odnosno nije) prisutna u drugom nizu. Praznine su prirodna pojava u biološkim nizovima. U mnogim biološkim primjenama, jedan mutacijski događaj može uzrokovati umetanje ili brisanje cijele regije (posebno u DNA), pa je precizno otkrivanje praznina u biološkim nizovima važan problem. Stvaranje praznina u DNA sekvencama (pa onda i u proteinskim nizovima) može biti uzrokovano brojnim biološkim procesima, uključujući sljedeće: dugi dijelovi DNA mogu biti kopirani i umetnuti jednim mutacijskim događajem; klizanje tijekom replikacije DNA može uzrokovati da se isto područje ponovi više puta jer replikacijski mehanizam izgubi svoje mjesto na predlošku; umetanje u jednu sekvencu u kombinaciji s recipročnim brisanjem u drugoj može biti uzrokovano nejednakom razmjenom tijekom mejoze; umetanje transponibilnih elemenata — skakačkih gena — u DNA sekvencu; umetanje DNA od strane retrovirusa; te translokacije DNA između kromosoma.

Neki od softverskih alata za proteinsko poravnanje uključuju BLAST (*Basic Local Alignment Search Tool*), Clustal Omega i MUSCLE.

Izvori korišteni u istraživanju, definiranju i opisivanju svih pojmova ovoga potpoglavlja (2.1) su sljedeći: [19], [22], [21], [23], [12], [8], [3], [7], [26], [4], [10], [11], [17], [25], [6], [5], [1], [2], [9], [15].

2.2 Aminokiseline kao petorke u $\mathbb{R}^5$

Nizovi aminokiselina u multiproteinskom poravnanju su zapravo dugi nizovi abecednih slova i "crtica". Nedostatak prirodne metrike i uređaja za usporedbu takvih abecednih podataka značajno otežava sofisticirane statističke analize tih nizova a samim time i modeliranje strukturnih i funkcionalnih aspekata proteina te srodnih problema.

AAindex (kratica za engl. *aminoacid index*), tj. preciznije *AAindeks1*, je *online* baza (dostupna na web adresi `https://www.genome.jp/aaindex/`) **aminokiselinskih indeksa**, tj. skupova od po 20 numeričkih vrijednosti koje reprezentiraju određeno fizikalnokemijsko ili biološko svojstvo svake od 20 standardnih aminokiselina. Ti indeksi uključuju generalna svojstva aminokiselina, kao npr. volumen ili veličinu molekule pojedine aminokiseline, hidrofobnost ili naboj, ali također i specifičnija mjerenja, kao npr. količinu nevezane energije po atomu ili orijentacijski kut bočnog lanca aminokiseline. Ti aminokiselinski indeksi su zapravo atributi, tj. numerički deskriptori aminokiselina.

U članku [13] je na 494 takva aminokiselinska indeksa, preuzeta s *AAindex*, napravljena **faktorska analiza** kojom se pokazalo da su podaci bili jako redundantni.

Faktorska analiza je statistička metoda koja se koristi da bi se producirao manji podskup novih, tzv. *zajedničkih faktora* koji bi sumirali i saželi originalni veći skup promatranih varijabli. U modelu faktorske analize, promatrane varijable su linearne kombinacije tih novoproduciranih faktora, a koeficijenti ispred faktora u određenoj linearnoj kombinaciji, (tzv. *faktorski koeficijenti*), pojednostavljeno, na neki način govore koliko pojedini faktor utječe na promatranu varijablu. Kod faktorske analize, nastoji se postići da za svaku pojedinu promatranu varijablu udio varijance te varijable koji faktori objašnjavaju bude što veći. Istovremeno, želi se postići i to da jednu originalnu varijablu "objašnjava" samo jedan faktor, tj. da samo koeficijent ispred tog jednog faktora bude značajno velik po apsolutnoj vrijednosti, a da ostali budu što manji (po apsolutnoj vrijednosti).

Vraćajući se na članak [13], zbog spomenute redundantnosti većeg broja aminokiselinskih atributa, od 494 izabrana su 54 aminokiselinska atributa na osnovu relativne veličine faktorskih koeficijenata, statističkih distribucijskih svojstava, relativne lakoće interpretacije i percipiranog strukturalnog značaja. Kao rezultat faktorske analize tih 54 atributa, dobiveno je **5 faktora**, tj. "klastera" od kojih svaki obuhvaća međusobno visokokorelirane fizikalnokemijske značajke za svaku od 20 aminokiselina. Također, atribute koje "objašnjava" jedan faktor, *većinom* ne odražavaju drugi faktori, tj. drugi faktori većinom tada imaju koeficijente relativno manje apsolutne vrijednosti za taj atribut.

Faktor I odražava istodobnu kovarijaciju između udjela eksponiranih naspram uvučenih ("zakopanih" - engl. *buried*) ostataka aminokiselina, nevezane naspram slobodne energije, broja donora vodikovih veza, polariteta te hidrofobnosti nasuprot hidrofilnosti. Radi jednostavnosti, faktor I nazivat ćemo **faktorom polariteta**.

Faktor II predstavlja **faktor sekundarne strukture**. On je povezan s dva "tipa" strukturalnih konformacija između kojih postoji inverzan odnos. To su relativna sklonost različitih aminokiselina prema strukturama kao što su zavojnice, skretanja ili pregibi i, s druge strane, učestalost pojavljivanja u obliku α-heliksa.

Faktor III povezan je s **molekularnom veličinom** ili **volumenom**, s visokim faktorskim koeficijentima za glomaznost, volumen ostatka aminokiseline, prosječni volumen "zakopanog" ostatka, volumen bočnog lanca i molekulsku masu. Velik negativan koefici-

jent faktora pojavljuje se za normaliziranu učestalost lijevozakrenutog α-heliksa.

Faktor IV odražava relativni sastav aminokiselina u raznim proteinima, **broj kodona** koji kodiraju aminokiselinu i sastav aminokiselina. Ovi atributi variraju inverzno s refraktivnošću i toplinskim kapacitetom.

Faktor V odnosi se na **elektrostatski naboj** s visokim koeficijentima za izoelektričnu točku i neto naboj. Očekivano, postoji inverzan odnos između pozitivnog i negativnog naboja.

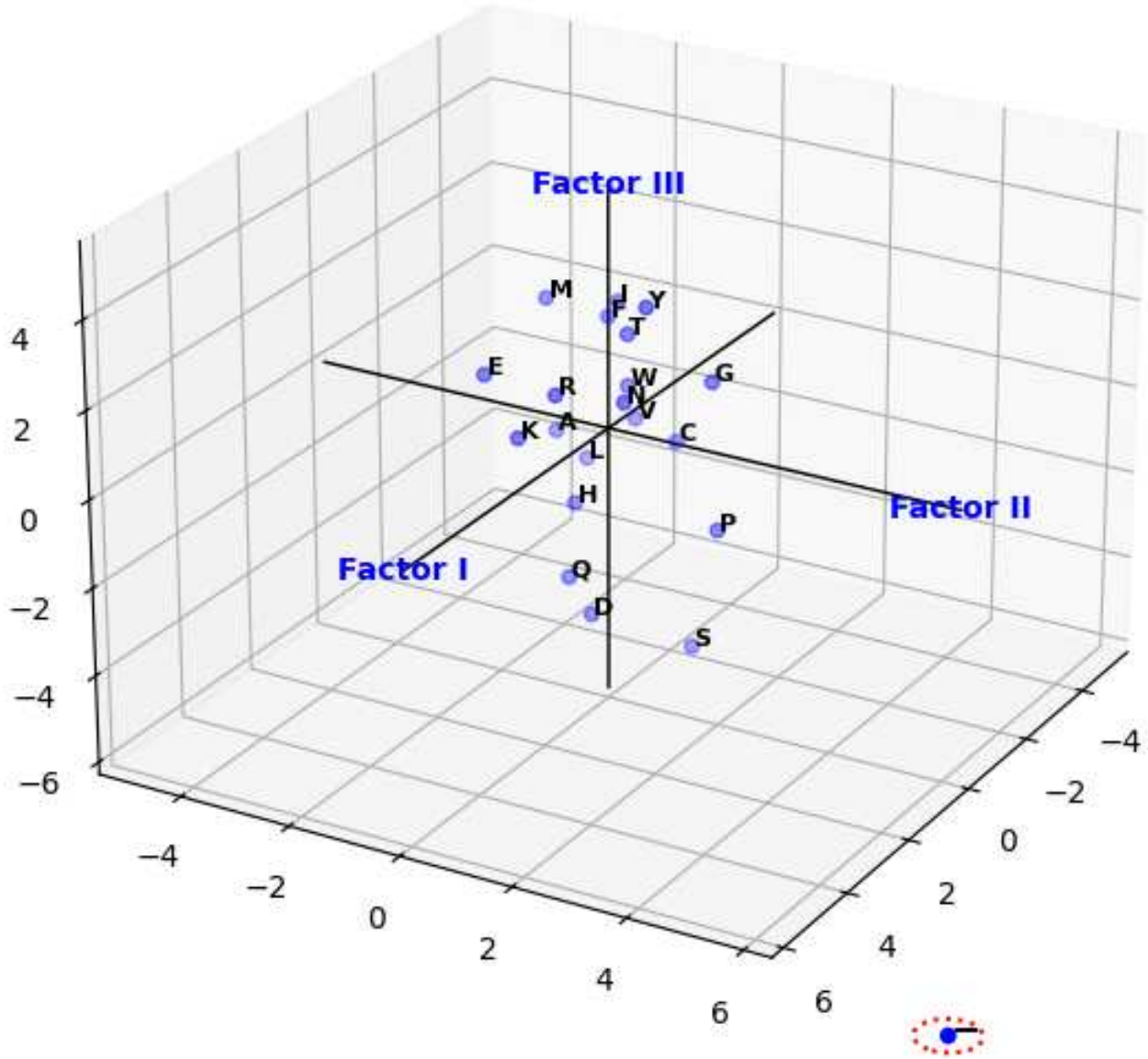
Ovim postupkom, svaka aminokiselina reprezentirana je s 5 brojeva. Posebno, petdimenzionalni vektor koji predstavlja "crticu", tj. izostanak aminokiselina na određenoj poziciji u proteinskom poravnanju, konstruiran je tako da taj vektor bude jako udaljen od svih ostalih vektora koji reprezentiraju aminokiseline.

Sada je sve spremno za statističku analizu proteinskih poravnanja, jer smo slova i "crtice" pretvorili u numeričke petdimenzionalne vektore. Ti vektori zapisani su u sljedećoj tablici:

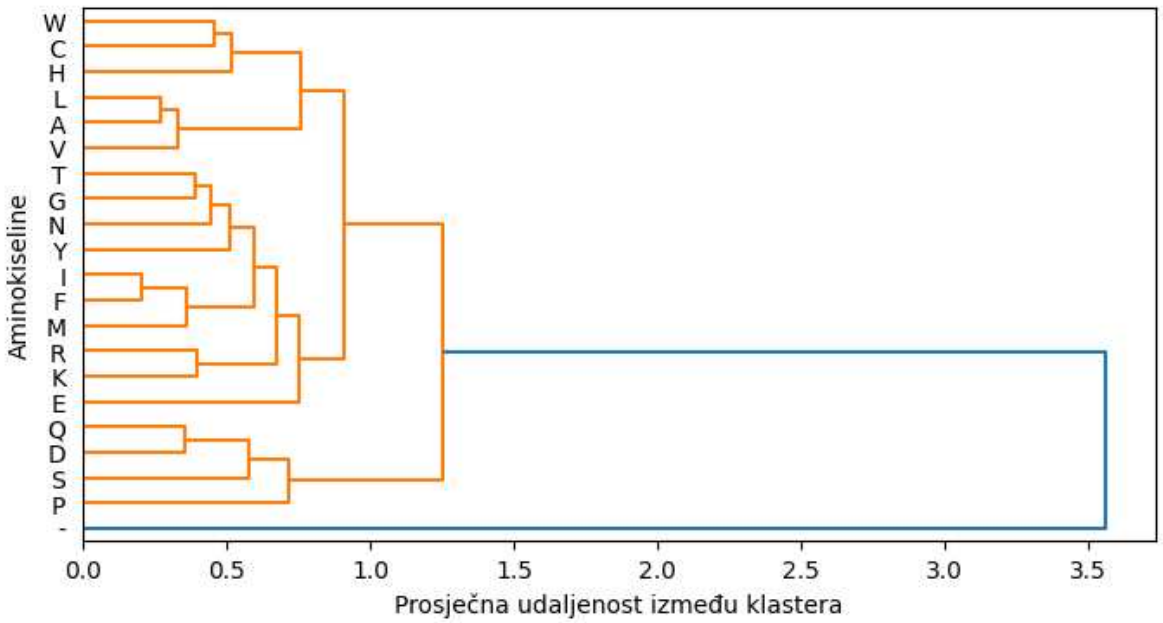
AMINOKISELINA	Faktor I	Faktor II	Faktor III	Faktor IV	Faktor V
A	-0.591	-1.302	-0.733	1.570	-0.146
C	-1.343	0.465	-0.862	-1.020	-0.255
D	1.050	0.302	-3.656	-0.259	-3.242
E	1.357	-1.453	1.477	-0.113	-0.837
F	-1.006	-0.590	1.891	-0.397	0.412
G	-0.384	1.652	1.330	1.045	2.064
H	0.336	-0.417	-1.673	-1.474	-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	-2.128	-0.184
W	-0.595	0.009	0.672	-2.128	-0.184
Y	0.260	0.830	3.097	-0.838	1.512
-	7.5	10.0	-6.0	-8.0	-1.0

Tablica 2.3: Faktori

Kako bi se dobio bolji dojam o rasporedu tih aminokiselina u prostoru i njihovim međusobnim udaljenostima, sada će se vizualno prikazati sve aminokiseline i "crtica" kao vektori u $\mathbb{R}^5$. Zbog nemogućnosti direktnog vizualiziranja vektora u $\mathbb{R}^5$, na prvoj slici 2.1 prikazat će se trodimenzionalni prikaz aminokiselina, koristeći prva tri faktora. Označena crvenom isprekidanom elipsom, daleko od svih, na dnu slike, leži "crtica". Zatim slijedi dendrogram (slika 2.2) dobiven *UPGMA* (engl. *unweighted pair group method with arithmetic mean*) metodom hijerarhijskog klasteriranja koja uključuje sve faktore. Na njemu se vidi koje aminokiseline čine klastere i međusobno su blizu, a koje su daleko. Opet je najupečatljivija udaljenost "crtice" od ostalih aminokiselina.



Slika 2.1: 3D prikaz aminokiselina



Slika 2.2: UPGMA dendrogram aminokiselina

Poglavlje 3

Ideja i pristup problemu

3.1 Priprema i obrada podataka

Podaci koji su temelj istraživanja u ovom radu su proteinska poravnanja spomenuta u [17], preuzeta s web stranice `https://www.biomedcentral.com/`. Konkretno, to su poravnanja 5 enzimskih familija: AT-domena, kinaza, malat i laktat dehidrogenaza (MDH i LDH), ciklaza i KR-domena.

Dodatno, uslijed bioloških eksperimentalnih istraživanja nad tim proteinskim familijama, poznata je i pripadnost svakog od proteina određenoj podfamiliji unutar familije kojoj taj protein pripada. Kako je svaka od 5 proteinskih familija podijeljena u dvije podfamilije, svakoj od familija dodijeljen je vektor grupa g koji poprima samo 2 vrijednosti. Dimenzija vektora g jednaka je broju redaka (broju proteina) u poravnanju, a on govori pripada li pojedini protein jednoj podfamiliji ili drugoj.

Cilj diplomskog rada je za svaku spomenutu proteinsku familiju, tj. njeno poravnanje, pronaći separacijske pozicije aminokiselina u poravnanju, odnosno one pozicije koje su statistički najznačajnije za raspodjelu proteina u jednu od dvije podfamilije.

Dakle, želi se vidjeti koje pozicije ili regije, zajedničke svim enzimima u familiji, određuju kakvu će specifičnu funkciju odrađivati pojedini enzimi, te, po tipu te funkcije, svrstavaju enzim u jednu od dvije podfamilije. Za razliku od kemijske specifičnosti, definirane u 2.1 koja se odnosi na pozicije jednog enzima, ova **funkcionalna specifičnost** koja će se istražiti u ovom radu, odnosi se na cijelu familiju enzima i treba odrediti pozicije ili regije specifične za sve enzime u familiji. Ipak, ako aktivna mjesta svih enzima u familiji upadaju u neku regiju, tj. skup pozicija, onda bi te pozicije intuitivno trebale biti i funkcionalno specifične. Kasnije će se vidjeti da će to i biti slučaj kod nekih familija (AT-domene).

Slova i "crtice" iz poravnanja ćemo, jasno, prikazati petdimenzionalnim vektorima, kako je opisano u 2.2. No, prije nego računamo s njima, valjalo bi ih **standardizirati**,

tj. svakom vektoru oduzeti aritmetičku sredinu svih vektora i rezultat podijeliti sa standardnom devijacijom svih vektora. Standardizacija je bitan korak u pripremanju podataka za statističku analizu jer bez standardizacije vektora, faktori (tj. skupovi vrijednosti pod određenim indeksom tih vektora) s većim rasponom mogu dominirati i dobiti veći statistički značaj nego što zaslužuju. Od sada pa nadalje svaki put kad spomenemo vektor koji reprezentira neku aminokiselinu ili prazninu, podrazumijevat ćemo da je standardiziran.

Zadana poravnanja su zapravo matrice znakova (slova i "crtica" koja redom označavaju aminokiseline i praznine u nizu). Znakove ćemo zamijeniti s pripadnim reprezentativnim standardiziranim petorkama koje ćemo jednostavno naslagati jednu iza druge u redu. Ako poravnanje familije ima m nizova od po n pozicija (znakova), onda ćemo tim postupkom konstruirati matricu tipa $m \times 5n$. Takvu matricu ćemo nazivati *matricom poravnanja*.

Konkretno, za familiju AT-domena, imamo 177 nizova duljine 324, i pritom 99 proteina potiče iz prve a preostalih 78 iz druge podfamilije. Stoga konstruiramo matricu poravnanja tipa 177×1620. U toj matrici prvih 99 redaka numerički predstavlja proteine koji pripadaju prvoj podfamiliji, a preostalih 78 proteine iz druge podfamilije.

Potpuno analogno, za kinaze konstruiramo matricu poravnanja dimenzija $(215, 5 \cdot 600)$, za MDH/LDH familiju matricu dimenzija $(183, 5 \cdot 418)$, dok su matrične dimenzije za ciklaze i KR-domene redom $(75, 5 \cdot 2040)$ i $(72, 5 \cdot 218)$.

3.2 Razdvajajuća (split) S-statistika

Budući da nam je cilj kvantificirati važnost neke pozicije u separiranju proteina u neku od dvije grupe (podfamilije), čini se intuitivnim gledati varijabilnost te pozicije unutar svake od dviju grupa (engl. *within-group variability - WGV*), i varijabilnost između grupa (engl. *between-group variability - BGV*). Ako je pozicija značajna za klasifikaciju proteina, onda će proteini iz pojedine grupe na tom mjestu imati manji broj aminokiselina koje su međusobno numerički blizu (čine jedan klaster), a udaljenost između klastera prve podfamilije i klastera druge podfamilije će biti velika. To točno podrazumijeva da je BGV velik, a WGV malen. To nas motivira da preko omjera vrijednosti intergrupne varijabilnosti (BGV) i intragrupne varijabilnosti (WGV) definiramo našu statistiku kojom ćemo mjeriti koliko je koja pozicija specifična, tj. značajna u razdvajanju dviju podfamilija (razdvajajuća statistika). Naravno, BGV i WGV možemo računati samo na numeričkim podacima, pa ćemo zato za svih 5 stupaca matrice poravnanja koji numerički predstavljaju poziciju i, izračunati omjer BGV-a i WGV-a.

Neka je zadana familija od m proteina od po n pozicija, pri čemu je m_1 proteina iz grupe 1, a m_2 proteina iz grupe 2. Neka je P pripadna reprezentativna numerička matrica poravnanja tipa $m \times 5n$. Označimo stupce matrice P s $p_1, p_2, ...p_{5n}$. Neka je $i \in \{1, ..., n\}$ broj proizvoljne pozicije u poravnanju. Za taj i definirajmo $k_j = 5(i-1) + j$, pri čemu

je $j \in \{1, 2, 3, 4, 5\}$. Tada su $p_{k_1}, p_{k_2}, p_{k_3}, p_{k_4}, p_{k_5}$ stupci matrice P koji odgovaraju poziciji i. Neka su $p_{k_j}^{(1)}$ i $p_{k_j}^{(2)}$ podskupi stupca p_{k_j} koji odgovaraju grupi 1, odnosno grupi 2, redom, gdje $j \in \{1, ..., 5\}$. Za $j \in \{1, ..., 5\}$ označimo s $\overline{p_{k_j}}$, $\overline{p_{k_j}^{(1)}}$ i $\overline{p_{k_j}^{(2)}}$ aritmetičke sredine pripadnih vektora, a s $Var[p_{k_j}]$, $Var[p_{k_j}^{(1)}]$ i $Var[p_{k_j}^{(2)}]$ njihove varijance. Tada definiramo **razdvajajuću (split) S-statistiku** sa:

$$S_i = \sum_{j=1}^{5} \frac{m_1 \cdot (\overline{p_{k_j}^{(1)}} - \overline{p_{k_j}})^2 + m_2 \cdot (\overline{p_{k_j}^{(2)}} - \overline{p_{k_j}})^2 + c}{(m_1 - 1) \cdot Var[p_{k_j}^{(1)}] + (m_2 - 1) \cdot Var[p_{k_j}^{(2)}] + c^2}, \tag{3.1}$$

za $i \in \{1, ..., n\}$, pri čemu je $c \in \mathbb{R}$ stabilizirajuća konstanta koju ćemo objasniti malo kasnije. Ova statistika (3.1) zapravo za svaku poziciju i računa 5 omjera intergrupne varijabilnosti (BGV) i intragrupne varijabilnosti (WGV) za svaki od 5 stupaca koji predstavljaju tu poziciju te ih zbraja kako bi dobila ukupni indikator specifičnosti te pozicije.

Kako uspoređujemo varijabilnost objašnjenu pripadanjem proteina dvama grupama, tj, intergrupnu objašnjenu varijabilnost (BGV) i neobjašnjenu varijabilnost unutar grupa (WGV), a takvi omjeri (s dodatnim korekcijama u vidu dijeljenja s brojem stupnjeva slobode) inače u ANOVI prate F-distribuciju, i mi se nadamo da će naša, heuristički smislena, razdvajajuća statistika pratiti F-distribuciju. U formuli 3.1 zbrajamo te omjere varijabilnosti pa se i po tome statistika razlikuje od one iz ANOVE, no to zbrajanje je smisleno, budući da se radi o povezanim stupcima. Razlog izostavljanja korekcije sa stupnjevima slobode je taj da dodatno dijeljenje u nazivniku prilično smanjuje nazivnik te velik broj omjera tada poprimi prevelike vrijednosti, i statistika tada više nikako ne prati F-distribuciju (čak i kad ne zbrajamo omjere, nego ih spremamo pojedinačno kao vrijednosti statistike — tada imamo doslovno istu statistiku kao što je F-statistika iz ANOVE za dvije grupe).

Konstanta c u brojniku i u nazivniku (u obliku c^2) pojedinih omjera, dodana je u formulu radi uspostavljanja numeričke stabilnosti. Naime, npr. ako za poziciju i imamo stupac koji gotovo svugdje ima istu aminokiselinu, onda će, za svaki od 5 omjera, varijanca u nazivniku, a i kvadrati u brojniku biti jako mali pa će omjeri numerički biti jako nestabilni. To znači da bi bez konstante potencijalno imali jako veliku vrijednost statistike koja bi ukazivala na visoku specifičnost te pozicije, što je potpuno netočno, s obzirom da svugdje imamo istu aminokiselinu. Dodavanje konstante c koja će u tom slučaju prevladati i omjerima dati manje i stabilnije vrijednosti donekle pomaže. Koliko male vrijednosti će tada poprimiti ti omjeri, ovisi o veličini konstante c. Najtočnije bi bilo tim pozicijama dodijeliti što nižu vrijednost S-statistike jer znamo da one uopće nisu važne za klasifikaciju. To se postiže povećavanjem konstante c. Međutim, tako formula gubi svoj heuristički značaj pa se zato i distribucija statistike jako mijenja i nikako ne prati F-distribuciju.

Probavanjem različitih vrijednosti konstante c, kako bi izbjegli navedene probleme, kao najbolja vrijednost konstante pokazala se $c = 1.5$.

Dodajmo još da smo i koristili jako sličnu statistiku, koja umjesto intergrupne varijabilnosti u brojniku j-tog sumanda ima ukupnu varijabilnost, tj. $m \cdot Var[p_{k_j}]$. S njom smo u konačnici dobili skoro identične rezultate.

3.3 Dodavanje buke

Neki zadani podaci, tj. poravnanja, imaju premalo proteina (redaka). Taj manjak podataka povlači manjak brojeva u stupcima matrice poravnanja, pa samim time prosjeci i varijance korištene u formuli 3.1 možda ne prikazuju pravo stanje u toj proteinskoj familiji. Zato ćemo umjetno generirati nove proteine, koji su bazirani na postojećim, ali s dozom šuma. Taj šum je dobiven iz distribucije standardnih aminokiselina uz randomizirani model, gdje su aminokiseline nezavisne:

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

(3.2)

Primijetimo da za "crticu" nemamo pridruženu vjerojatnost, i da brojevi u sumi daju 1. Zato stavljamo da "crtica" ima vjerojatnost 0.

Pomoću distribucije 3.2 računamo prosječnu vektorsku reprezentaciju aminokiseline (prosječnu s obzirom na distribuciju 3.2). Označimo tu reprezentaciju s $\overline{v}$. Neka je v_k reprezentativni standardizirani vektor aminokiseline zapisane k-te po redu u distribuciji 3.2 te neka je p_k vjerojatnost iz 3.2 pridružena aminokiselini reprezentiranoj s v_k. Definiramo $\overline{v}$ s:

$$\overline{v} = \sum_{k=1}^{20} p_k v_k$$

Uzmimo proizvoljno poravnanje s m proteina duljine n. Sada za svaki protein iz tog poravnanja, možemo generirati po još jedan sljedećim postupkom. Neka je $i \in \{1, ..., n\}$ bilo koja pozicija u poravnanju. Uzmimo prvi protein poravnanja i pogledajmo mu aminokiselinu (ili "crticu") na i-toj poziciji i označimo njezin reprezentativni vektor s $v_{1,i}$. Generirajmo novi vektor koji će predstavljati i-tu poziciju novog proteina sljedećom formulom:

$$v_{1',i} = \alpha \cdot v_{1,i} + (1 - \alpha) \cdot \overline{v},$$

gdje je α tzv. *koeficijent očuvanja* i iznosi $\alpha = 0.8$.
Ovo možemo napraviti za svaki $i \in \{1, ..., n\}$ pa dobivamo vektore $(v_{1',i})_{i \in \{1,...,n\}}$ koji, međusobno poredani po redu kako se i povećava, daju novi "protein", tj njegovu reprezentaciju.

Ako pustimo α da varira od 0.8 do 1 s pomakom od 0.1, onda na osnovu retka prvog proteina iz originalnog poravnanja možemo dobiti još 20 novih nizova. S N označimo broj nizova koji su rezultat ovog procesa, uključujući i niz originalnog proteina. Dakle, $N = 21$.

Primijetimo da ovaj cijeli postupak možemo ponoviti i za svaki drugi protein iz originalnog poravnanja, te tako dobiti ukupno $21 \cdot m$ numeričkih redaka koje možemo spremiti u posve novu matricu poravnanja, i odatle tvoriti novu razdvajajuću statistiku $(S_i)_{i\in\{1,\dots,n\}}$, kao što je opisano u 3.2.

Na ovaj način smo proširili opseg podataka i spremni smo za računanje S-statistike na konkretnim podacima.

Poglavlje 4

Rezultati

Ovdje prikazujemo rezultate statističke analize temeljene na razdvajajućoj S-statistici definiranoj u potpoglavlju 3.2 (s konstantom $c = 1.5$), koristeći i umjetno generirane retke u poravnanjima dobivene postupkom iz potpoglavlja 3.3, uz $N = 21$. Za svaku familiju navest ćemo najznačajnije pozicije za podjelu proteina u neku od dvije podfamilije. Nacrtat ćemo histograme distribucija S-statistike i qq-vjerojatnosne grafove te testirati njenu pripadnost Fisherovoj F-distribuciji Kolmogorov-Smirnovljevim testom. Zatim ćemo prikazati t-SNE (engl. *t-distributed stochastic neighbour embedding*) grafove koji vizualiziraju sve proteine neke familije u dvodimenzionalnom koordinatnom sustavu. Svaki protein iz originalnog poravnanja će biti reprezentiran s 10 pozicija s najvećom S-statistikom. Dakle, svaki protein iz originalnog poravnanja bit će prikazan s 10 brojeva, od kojih je svaki broj prosjek 5 faktora koji opisuju aminokiselinu ili "crticu" na pojedinoj poziciji. Bojajući točke koje prikazuju proteine jedne podfamilije u jednu boju, a točke, tj. proteine druge podfamilije u drugu, moći ćemo, na temelju (ne)vidljivog klasteriranja ocijeniti kako najznačajnijih 10 pozicija separira proteine u grupe. Konačno, usporedit ćemo ove rezultate s rezultatima sličnih istraživanja iz [17] i [16].

4.1 AT-domene

Za proteinsku familiju AT-domena, originalno smo imali poravnanje 177 proteina duljine 324, pri čemu je 99 proteina pripadalo jednoj podfamiliji AT-domena, a 77 proteina drugoj. Zatim smo numeričkim reprezentiranjem konstruirali matricu poravnanja dimenzija (177, 1600). Konačno, nakon generiranja novih numeričkih redaka, dobivamo novu matricu poravnanja P dimenzija (3717, 1600), gdje smo pritom novogenerirane retke matrice P klasificirali u onu grupu kojoj je pripadao originalni protein iz kojeg smo generirali te nove retke. Dakle, m_1 i m_2 koje ubacujemo u formulu 3.1 su redom $99 \cdot 21$ i $78 \cdot 21$, odnosno 2079 i 1638. Nad stupcima matrice P izvršili smo izračun S-statistike i sada pri-

kazujemo najvećih 25 vrijednosti S-statistike, zajedno s pripadnim pozicijama originalnog poravnanja:

Redni broj	**Pozicija**	**S-statistika**
1	212	52.98
2	156	15.96
3	282	13.74
4	71	13.18
5	291	12.18
6	157	11.73
7	155	10.56
8	264	10.45
9	167	10.09
10	277	9.81
11	64	8.95
12	48	8.56
13	221	8.33
14	224	8.12
15	143	7.95
16	105	7.90
17	73	7.86
18	133	7.64
19	209	6.79
20	168	5.92
21	72	5.91
22	210	5.56
23	130	5.22
24	234	5.13
25	217	5.03

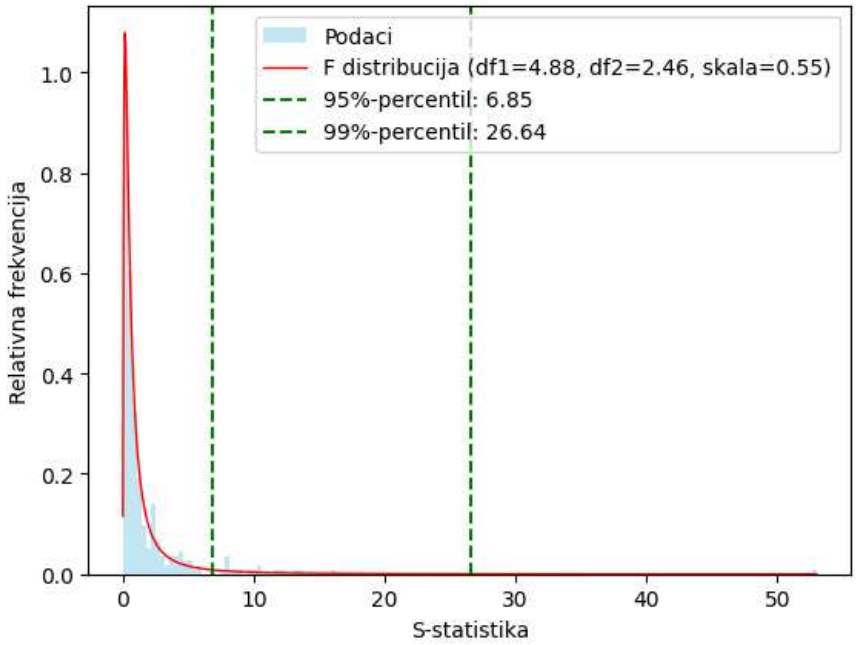
Tablica 4.1: Najspecifičnije pozicije AT-domena

Ispod je prikazan histogram vrijednosti S-statistike zajedno s crvenom linijom koja predstavlja teorijsku funkciju gustoće F-distribucije s parametrima procijenjenima iz podataka metodom maksimalne vjerodostojnosti (engl. Maximum Likelihood Estimation — MLE).

Procijenjena je F-distribucija sa stupnjevima slobode od otprilike 4.88 i 2.46, te parametrom skaliranja od približno 0.55. Parametar skaliranja u iznosu od ≈ 0.55 znači da vrijednosti F-distribucije s procijenjenim stupnjevima slobode od otprilike 4.88 i 2.46, tek

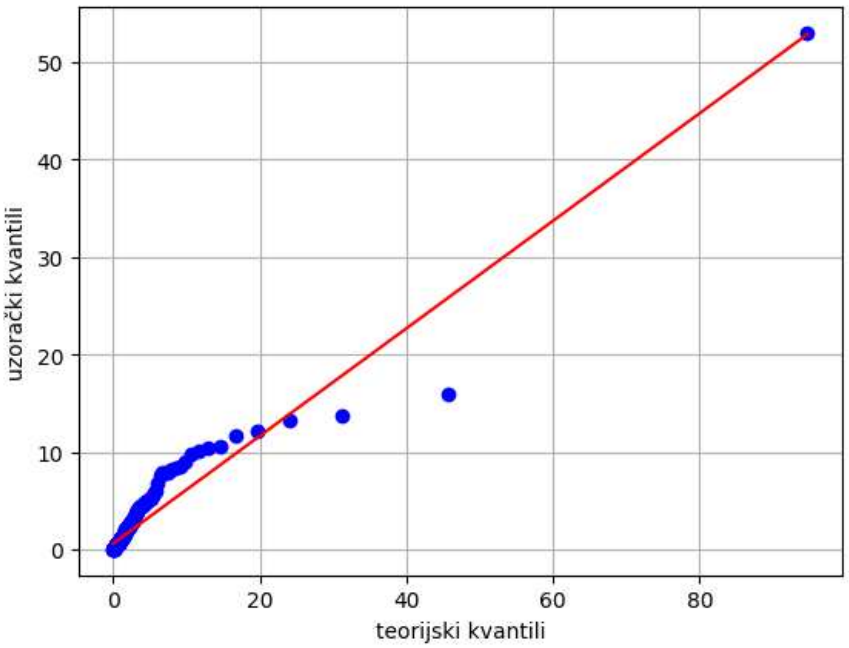
kad se pomnože s otprilike 0.55, onda dobro (u smislu metode MLE) opisuju naše podatke.

Ekvivalentno, mogli smo i transformirati vrijednosti S-statistike dijeleći ih s približno 0.55, i onda te podatke procijeniti F-distribucijom uzimajući u obzir samo stupnjeve slobode kao parametre za procjenu. Tako procijenjena F-distribucija imala bi iste stupnjeve slobode kao i ranije, ali naravno, jer su se podaci promijenili, i kvantili, tj. percentili bi bili drugačiji, no odgovarali bi identičnom rangiranju vrijednosti i odvajali iste skupove pozicija kao statistički značajne. K tome, i histogram bi izgledao identično, samo s drugom skalom na x-osi koja bi tada mjerila transformiranu S-statistiku. Kako ćemo kasnije prikazivati i qq-vjerojatnosne grafove i testirati pripadnost originalnih vrijednosti S-statistike F-distribuciji Kolmogorov-Smirnovljevim (KS) testom, spomenimo da bi normalni vjerojatnosni grafovi izgledali jednako, do na skalu, i da bi provedbom KS testa dobili jednaku p-vrijednost kao što ćemo dobiti s netransformiranim S-statistikama, i navedenim parametrom skaliranja. Dakako, ni t-SNE graf se ne bi razlikovao od onog kojeg ćemo priložiti, do na skalu.



Slika 4.1: Histogram S-statistike s procijenjenom F-distribucijom za AT-domene

Prikažimo i qq-vjerojatnosni graf teorijskih kvantila F-distribucije (df1=4.88, df2=2.46, skala=0.55) na x-osi i kvantila S-statistike na y-osi. Iako "fit" nije baš najbolji, možemo primijetiti da su podaci, bar za manje kvantile, blizu pravca.



Slika 4.2: qq-vjerojatnosni graf za AT-domene

Potom je Kolmogorov-Smirnovljevim (KS) testom testirana pripadnost podataka (vrijednosti S-statistike) F-distribuciji. Testna statistika iznosi 0.0431, a p-vrijednost 0.57. Zaključujemo da ne možemo odbaciti pripadnost podataka F-distribuciji na razinama značajnosti od 1%, 5%, ni na bilo kojim razumnim razinama značajnosti.

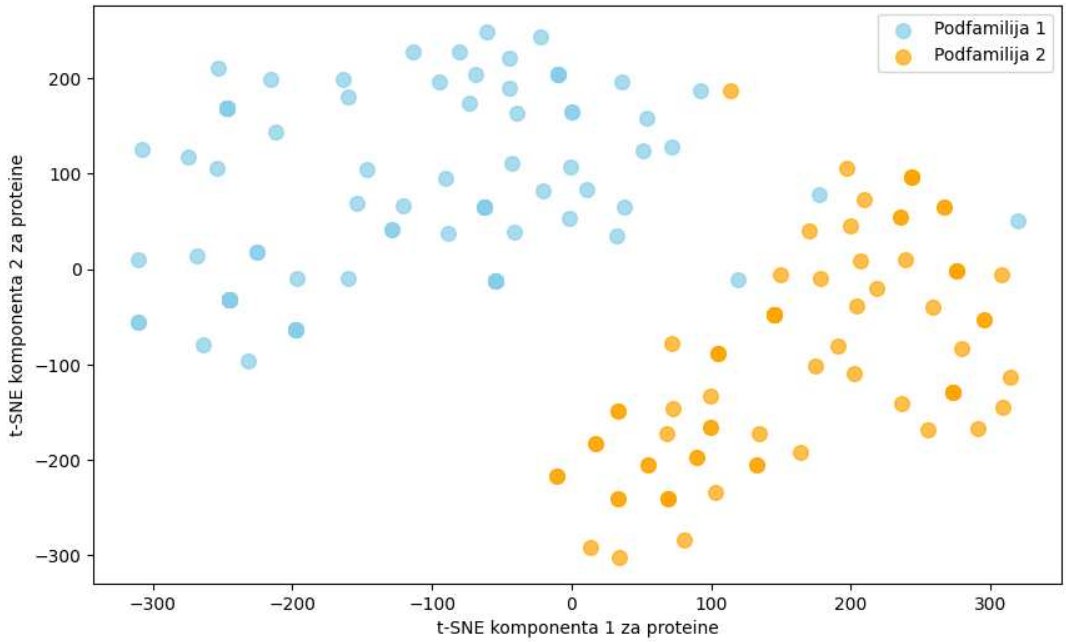
Kako možemo pretpostaviti da naša razdvajajuća S-statistika prati F-distribuciju, možemo odrediti one pozicije koje su statistički značajne za raspodjelu proteina u podfamilije na razini od 5% i 1%. Kako 95%-percentil F-dstribucije iznosi 6.85, na razini od 5%, statistički je značajno prvih 18 pozicija s najvećom S-statistikom. 99%-percentil F-distribucije iznosi 26.64 pa je na razini od 1% statistički značajna samo pozicija 212. Vraćajući se na tablicu 4.1, sada možemo objasniti zelene linije. Naime, one upravo označuju granice iznad kojih slijede statistički značajne pozicije, na razinama značajnosti od 5% i 1%.

Ono što možemo primijetiti detaljnijim promatranjem originalnih podataka je to da su pozicije 155-157, 167-168, 277, 282, 64-65 i 48 takve da za jednu od dvije grupe ima hrpetina “crtica”, a za drugu grupu neka mješavina drugih aminokiselina. Kako je vektor koji enkodira “crticu” jako udaljen od svih ostalih vektora, ne čudi da su te pozicije ispale dijagnostičke u diskriminaciji proteina u grupe. Postavlja se pitanje jesu li te pozicije gdje “crtice” prevladavaju stvarno toliko značajne.

S druge strane, za pozicije 212, 105 i 210 znamo da su i eksperimentalno potvrđene

kao značajne u diskriminiranju dviju grupa (C2 i C3). Za poziciju 212 se zna da je ključna za funkcionalnu specifičnost proteina, što odgovara i našim rezultatima u kojima je 212 daleko statistički najznačajnija pozicija, čak na razini od 1%. Za pozicije 209 i 210 postoje dokazi da su dio regije aktivnog mjesta koja je zajednička za sve AT-domene iz poravnanja. Gledajući naše rezultate, pozicije 209 i 210 osim što su u prvih 22 najspecifičnije pozicije, jako su blizu i da budu statistički značajne na razini od 5%, kao što je to pozicija 105.

Prikažimo sada t-SNE graf. Možemo vidjeti jasne klastere i svega 4 pogrešne klasifikacije proteina.



Slika 4.3: t-SNE graf za AT-domene

U sljedećoj tablici usporedit ćemo naše rezultate s rezultatima nekih drugih sličnih istraživanja. U prvom stupcu tablice zapisano je 10 naših najspecifičnijih pozicija, a u drugom isto toliko naših najvažnijih pozicija, ali isključujući one gdje "crtice" prevladavaju. Zatim slijede 3 stupca od 10 najbitnijih pozicija u radu [16] u kojem se koriste slične tehnike za analizu nizova poravnanja kao u ovom radu. Svaki od ta tri stupca dobiven je koristeći različite konstante stabilizacije, na način sličan kao u ovom radu. Također ćemo navesti i 10 najvažnijih pozicija dobivenih stohastičkom analizom istih podataka, iz članka [17], koristeći pristup s BLOSUM (engl. *BLOcks SUbstitution Matrix*) i PAM (engl. *Point Accepted Mutation*) supstitucijskim matricama. Valja napomenuti da ni analiza u [16] niti analiza u [17] nije uključivala one pozicije na kojima je bar jedan protein imao prazninu,

tj. “crticu”. Upravo zato smo prikazali drugi stupac u ovoj tablici.

4.1	**4.1(bez '-')**	**[16](1)**	**[16](2)**	**[16](3)**	**[17](BLOSUM)**	**[17](PAM)**
212	212	212	212	212	212	212
156	71	143	72	72	73	82
282	291	73	73	130	105	210
71	264	264	130	234	71	73
291	221	71	143	73	210	105
157	224	221	264	143	264	209
155	143	224	224	210	221	71
264	105	105	234	264	209	234
167	73	72	12	221	82	264
277	133	12	210	149	234	221

Tablica 4.2: Usporedba rezultata za AT-domene

Može se primijetiti da su rezultati u drugom stupcu jako slični onima iz ostalih istraživanja. Naime, svaku poziciju iz drugog stupca možemo naći u bar jednom od stupaca koji slijede, osim pozicija 291 i 133. Pozicija 133 se može naći u rezultatima stohastičke analize iz [17] na 35. i 26. mjestu, koristeći BLOSUM i PAM redom.

Posebno je zanimljiva pozicija 291 u kojoj ne prevladavaju “crtice” (na toj poziciji samo dva proteina iz originalnog poravnanja imaju prazninu), a ispala je kao peta najznačajnija pozicija. S razlogom je tako visoko, jer gotovo svi proteini prve grupe na tom mjestu imaju aminokiselinu G (glicin), i gotovo svi proteini druge grupe na tom mjestu imaju aminokiselinu S (serin), a te dvije aminokiseline su jako različite te su njihove vektorske reprezentacije jako udaljene (2.2). Poziciju 291 ne možemo naći u rezultatima drugih analiza ([16] ili [17]) jer je tamo ta pozicija u startu izbačena, s obzirom da dva proteina na tom mjestu imaju praznine.

4.2 MDH / LDH familija

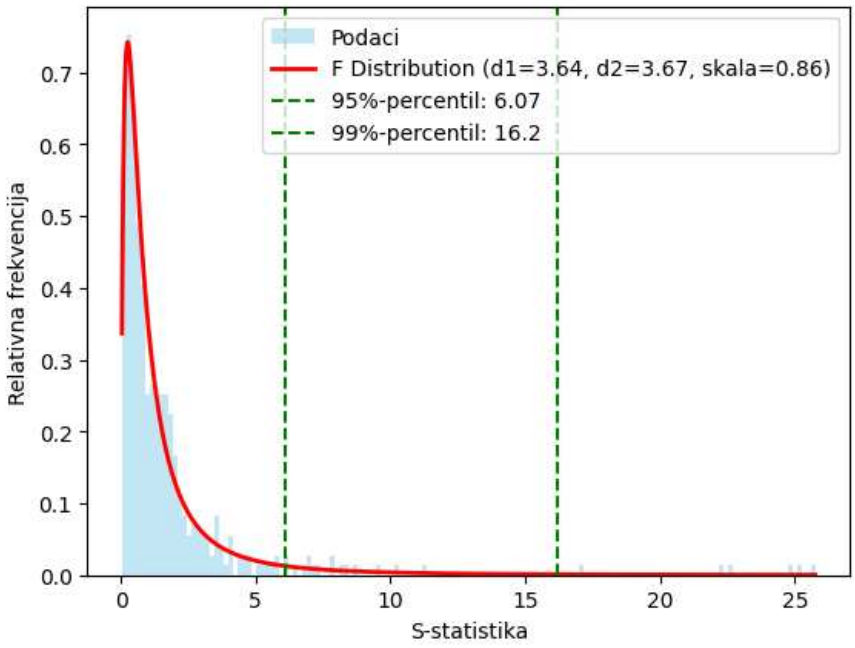
Za proteinsku familiju koju čine malatne dehidrogenaze (MDH) i laktatne dehidrogenaze (LDH), originalno smo imali poravnanje 183 proteina duljine 418, pri čemu je bilo 74 dehidrogenaze jednog tipa, a 109 dehidrogenaza drugog tipa. Zatim smo numeričkim reprezentiranjem konstruirali matricu poravnanja dimenzija (183, 2090). Konačno, nakon generiranja novih numeričkih nizova, dobivamo novu matricu poravnanja P dimenzija (3843, 2090), gdje smo pritom novogenerirane retke matrice P klasificirali u onu grupu kojoj je pripadao originalni protein iz kojeg smo generirali te nove retke. Dakle, m_1 i m_2

koje ubacujemo u formulu 3.1 su redom 1554 i 2289. Nad stupcima matrice P izvršili smo izračun S-statistike i sada prikazujemo najvećih 25 vrijednosti S-statistike, zajedno s pripadnim pozicijama originalnog poravnanja:

Redni broj	**Pozicija**	**S-statistika**
1	125	25.77
2	276	25.15
3	289	24.88
4	354	22.54
5	144	22.25
6	148	17.04
7	300	11.34
8	143	10.19
9	308	9.47
10	31	8.64
11	296	8.41
12	33	8.14
13	29	7.89
14	32	7.83
15	35	7.28
16	404	7.12
17	28	7.05
18	193	6.94
19	34	6.76
20	169	6.48
21	30	6.08
22	58	6.07
23	279	5.84
24	297	5.77
25	214	5.65

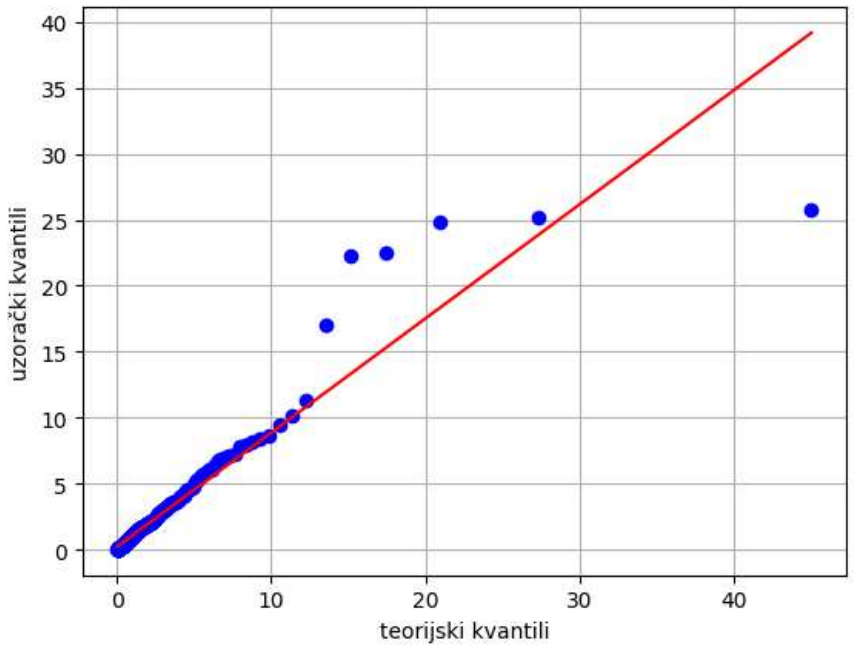
Tablica 4.3: Najspecifičnije pozicije MDH/LDH familije

Ispod je prikazan histogram vrijednosti S-statistike s crvenom linijom koja predstavlja teorijsku funkciju gustoće F-distribucije s parametrima procijenjenima iz podataka metodom maksimalne vjerodostojnosti. Procijenjena je F-distribucija sa stupnjevima slobode od otprilike 3.64 i 3.67, te parametrom skaliranja od približno 0.86.



Slika 4.4: Histogram S-statistike s procijenjenom F-distribucijom za MDH/LDH familiju

Prikažimo i qq-vjerojatnosni graf teorijskih kvantila F-distribucije (df1=3.64, df2=3.67, skala=0.86) na x-osi i kvantila S-statistike na y-osi. Možemo primijetiti da skoro svi podaci leže na pravcu. Samo 5 točki je osjetnije udaljeno od pravca.



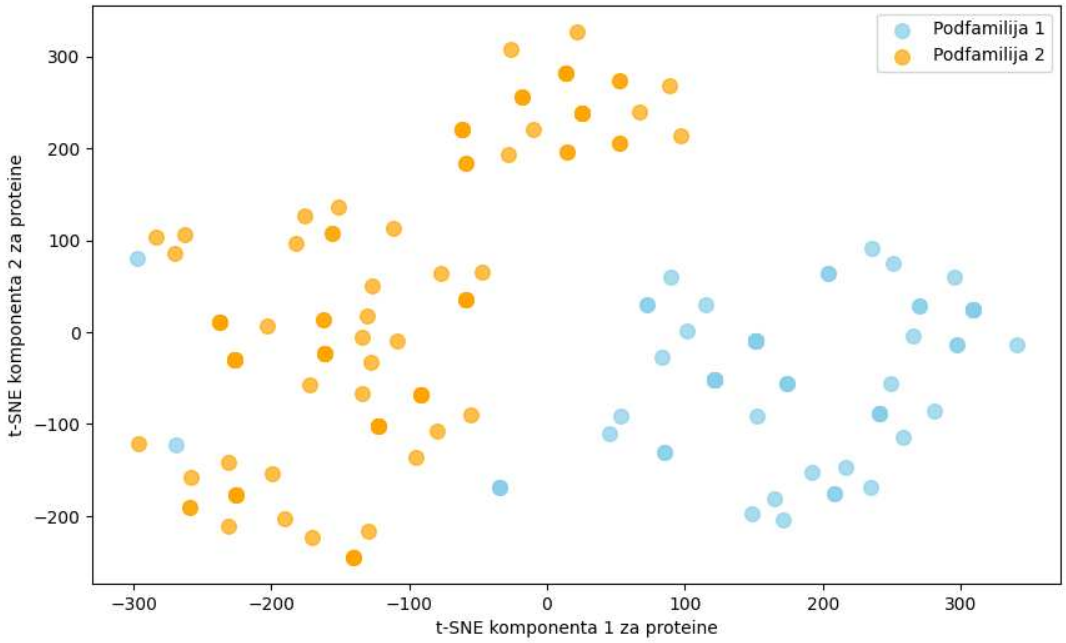
Slika 4.5: qq-vjerojatnosni graf za MDH/LDH familiju

Zatim je Kolmogorov-Smirnovljevim (KS) testom testirana pripadnost podataka (vrijednosti S-statistike) procijenjenoj F-distribuciji. Testna statistika iznosi 0.0322, a p-vrijednost 0.77. Zaključujemo da ne možemo odbaciti pripadnost podataka F-distribuciji na razinama značajnosti od 1%, 5%, ni na bilo kojim razumnim razinama značajnosti.

Kako možemo pretpostaviti da naša razdvajajuća S-statistika prati F-distribuciju, možemo odrediti one pozicije koje su statistički značajne za raspodjelu proteina u podfamilije na razini od 5% i 1%. Kako 95%-percentil F-distribucije iznosi 6.07, na razini od 5% statistički je značajno prvih 21 pozicija s najvećom S-statistikom. 99%-percentil F-dstribucije iznosi 16.2 pa je na razini od 1% statistički značajno prvih 6 pozicija s najvećom S-statistikom. Vraćajući se na tablicu 4.3, sada možemo objasniti zelene linije. Naime, kao i kod AT-domena, one upravo označuju granice iznad kojih slijede statistički značajne pozicije, na razinama značajnosti od 5% i 1%.

Ono što možemo primijetiti detaljnijim promatranjem originalnih podataka je to da su top 4 pozicije (125, 276, 289 i 354), kao i nekolicina pozicija malo niže u 4.3 (pozicije 29-35), takve da za jednu od dvije grupe prevladavaju “crtice”, a za drugu grupu neka mješavina drugih aminokiselina koje su sve jako udaljene od “crtice”. Opet, postavlja se pitanje jesu li te pozicije gdje “crtice” prevladavaju stvarno toliko značajne.

Prikažimo sada t-SNE graf. Možemo vidjeti jasne klastere i svega 2 pogrešne klasifikacije proteina.



Slika 4.6: t-SNE graf za MDH/LDH familiju

U sljedećoj tablici usporedit ćemo naše rezultate s rezultatima nekih drugih sličnih istraživanja, sasvim analogno kao kod AT-domena.

4.3	**4.3(bez '-')**	**[16](1)**	**[16](2)**	**[16](3)**	**[17](BLOSUM)**	**[17](PAM)**
125	144	144	144	308	148	148
276	148	148	308	144	300	300
289	300	169	169	214	308	308
354	143	308	214	169	144	144
144	308	300	143	309	296	296
148	296	193	309	143	143	62
300	193	143	193	119	303	143
143	169	246	215	193	170	97
308	58	214	148	256	246	215
31	297	215	58	79	192	60

Tablica 4.4: Usporedba rezultata za MDH/LDH familiju

Opet, drugi stupac je jako sličan stupcima drugih istraživanja. Jedini uljez, koji se ne pojavljuje još nigdje u top 10 pozicija je pozicija 297. Ona se može naći u rezultatima

stohastičke analize iz [17] na 45. i 47. mjestu, koristeći BLOSUM i PAM redom.

4.3 Ciklaze

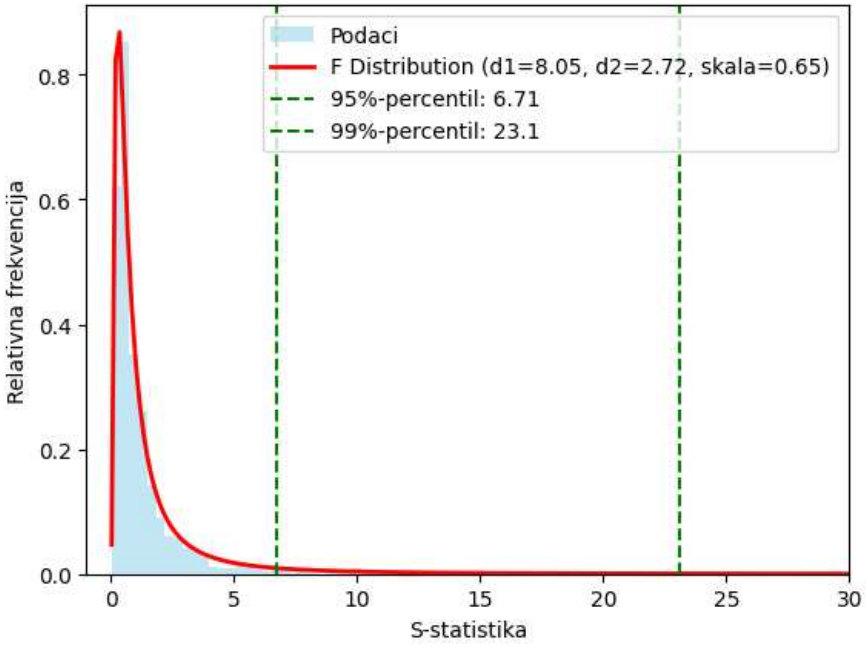
Za c=1.5, N=21

Za proteinsku familiju ciklaza, originalno smo imali poravnanje 75 proteina duljine 2040, pri čemu je bilo 33 enzima iz prve podfamilije, a 42 iz druge. Zatim smo numeričkim reprezentiranjem konstruirali matricu poravnanja dimenzija (75, 10200). Konačno, nakon generiranja novih numeričkih nizova, dobivamo novu matricu poravnanja P dimenzija (1575, 10200), gdje smo pritom novogenerirane retke matrice P klasificirali u onu grupu kojoj je pripadao originalni protein iz kojeg smo generirali te nove retke. Dakle, m_1 i m_2 koje ubacujemo u formulu 3.1 su redom 693 i 882. Nad stupcima matrice P izvršili smo izračun S-statistike i sada prikazujemo najvećih 50 vrijednosti S-statistike, zajedno s pripadnim pozicijama originalnog poravnanja:

Redni broj	Pozicija	S-statistika
1	1297	162.74
2	1474	154.89
3	1256	133.80
4	1265	126.17
5	883	119.05
6	1258	110.13
7	882	109.53
8	1254	109.29
9	1263	108.19
10	1634	108.01
11	1259	107.97
12	1296	107.28
13	1255	107.16
14	1257	105.41
15	1264	105.33
16	1260	101.04
17	881	94.80
18	1298	94.12
19	1253	89.35
20	1262	85.93
21	1488	84.75
22	1535	69.93
23	1536	48.60
24	1466	47.79
25	1568	43.71
26	1630	42.18
27	1517	40.89
28	583	40.42
29	1473	36.55
30	1467	36.09
31	1635	35.93
32	1566	35.83
33	1572	35.75
34	1567	35.58
35	1636	30.51
36	1137	30.18
37	1243	29.58
38	1632	28.46
39	1244	28.46
40	651	24.50
41	272	24.45
42	585	23.42
43	584	23.39
44	1506	22.35
45	315	21.59
46	1529	20.98
47	317	20.71
48	271	20.41
49	1533	18.68
50	1247	18.29

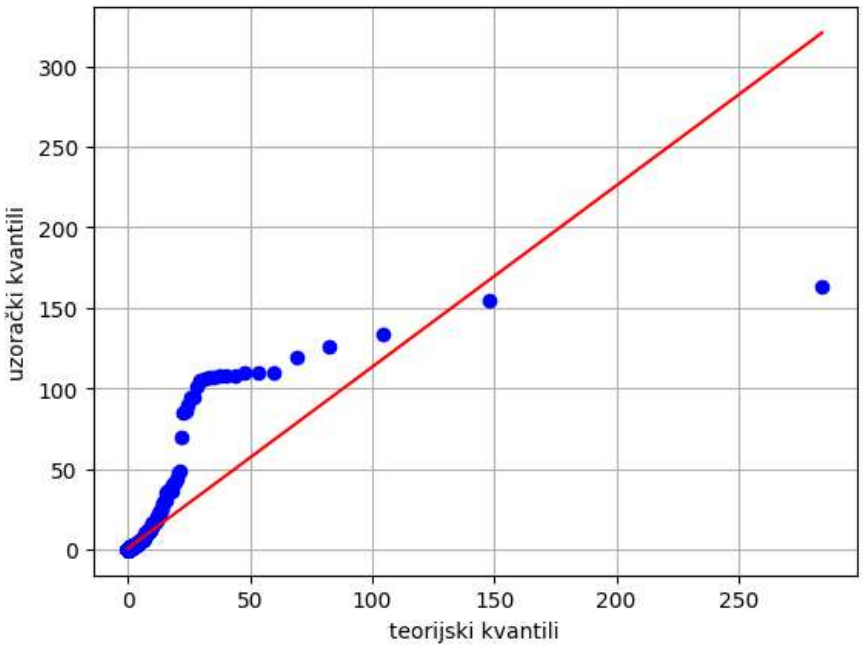
Tablica 4.5: Najspecifičnije pozicije ciklaza ($c = 1.5, N = 21$)

Ispod je prikazan histogram vrijednosti S-statistike s crvenom linijom koja predstavlja teorijsku funkciju gustoće F-distribucije s parametrima procijenjenima iz podataka metodom maksimalne vjerodostojnosti. Procijenjena je F-distribucija sa stupnjevima slobode od otprilike 8.05 i 2.72, te parametrom skaliranja od približno 0.65. Kako je jako mala relativna frekvencija S-statistika koje su veće od 30, prikazujemo skraćenu verziju histograma na kojoj se puno bolje vidi distribucija.



Slika 4.7: Histogram S-statistike s procijenjenom F-distribucijom za ciklaze ($c = 1.5, N = 21$)

Prikažimo i qq-vjerojatnosni graf teorijskih kvantila F-distribucije (df1=8.05, df2=2.72, skala=0.65) na x-osi i kvantila S-statistike na y-osi. Možemo primijetiti veliko odstupanje koje kreće otprilike nakon što S-statistika prijeđe vrijednost od 25.



Slika 4.8: qq-vjerojatnosni graf za ciklaze ($c = 1.5, N = 21$)

Potom je Kolmogorov-Smirnovljevim (KS) testom testirana pripadnost podataka (vrijednosti S-statistike) procijenjenoj F-distribuciji. Testna statistika iznosi 0.04545, a p-vrijednost 0.0004. Odbacujemo hipotezu da S-statistika prati F-distribuciju na razinama značajnosti od 1% i 5%. Ipak, radi analize pozicija, u nastavku ćemo pretpostaviti da je S-statistika F-distribuirana. Napomenimo da sljedeći rezultati u vidu statističke značajnosti pozicija zasada nisu pouzdani.

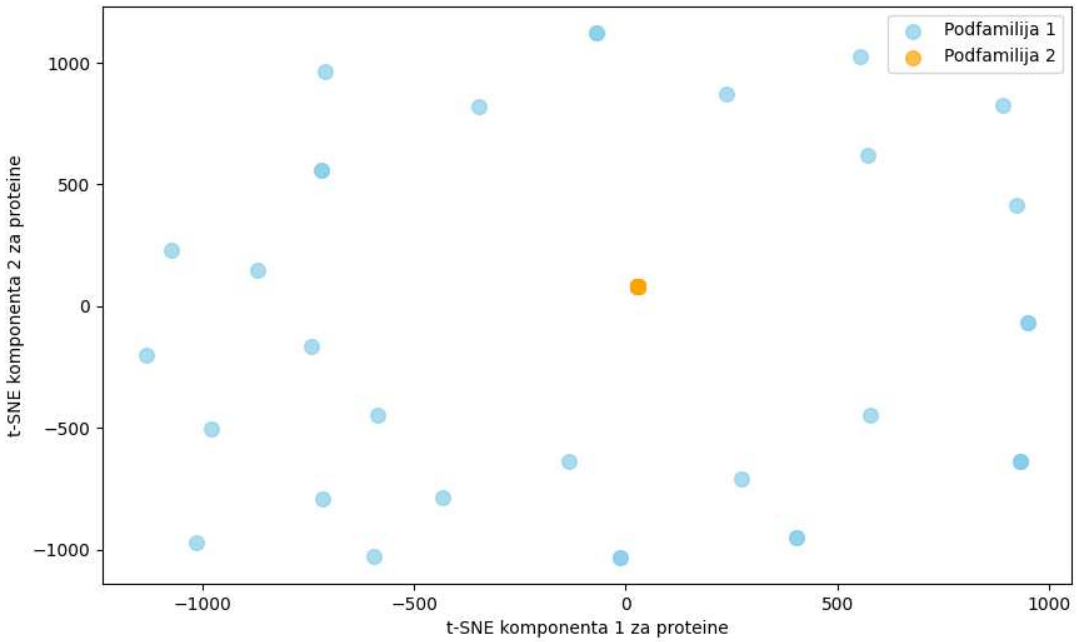
Pod jako labavom pretpostavkom F-distribuiranosti, odredimo one pozicije koje su statistički značajne za raspodjelu proteina u podfamilije na razini od 5% i 1%. Kako 95%-percentil F-distribucije otprilike iznosi 6.71, na razini od 5% statistički je značajno prvih 112 pozicija s najvećom S-statistikom. 99%-percentil F-dstribucije otprilike iznosi 23.1 pa su na razini od 1% statistički značajne prve 43 pozicije s najvećom S-statistikom. Na tablici 4.5 zelenom linijom označili smo granicu iznad koje slijede statistički značajne pozicije, na razini značajnosti od 1%.

Detaljnijim pregledavanjem zadanog poravnanja primjećujemo da ima jako puno pozicija gdje za jednu od grupa prevladavaju “crtice”. Naime, to su pozicije 1243-1298, 1476-1474, 881-883, 1466, 1467, 1473, 1474, 1566-1572, 583-585, 313-317 itd. Velika većina njih se pojavljuje visoko u tablici 4.5.

S druge strane, pogotovo za pozicije 1634 i 1536, a i za pozicije 1630, 1488 i 1517, znamo da su i eksperimentalno potvrđene kao značajne u diskriminiranju dviju grupa. To

se slaže s našim rezultatima, jer su te pozicije pri samom vrhu rang ljestvice.

Prikažimo sada t-SNE graf. Možemo primijetiti da su proteini podfamilije 2 zapanjujuće dobro klasterirani, do te mjere da se na grafu vidi samo jedna, malo veća i gušća točka. Proteini podfamilije 1 u osjetnom razmaku okružuju taj jako zgusnuti klaster podfamilije 2. Stoga je točnost klasificiranja proteina na temelju najznačajnijih 10 pozicija 100%-tna.



Slika 4.9: t-SNE graf za ciklaze

U sljedećoj tablici usporedit ćemo naše rezultate s rezultatima nekih drugih sličnih istraživanja, sasvim analogno kao i u drugim familijama.

4.5	4.5(bez '-')	[16](1)	[16](2)	[16](3)	[17](BLOSUM)	[17](PAM)
1297	1634	1488	1632	1632	1634	1634
1474	1488	1634	1488	1488	1630	1517
1256	1535	1635	1634	1497	1517	1533
1265	1536	1535	1497	654	1533	1440
883	1630	1632	654	1634	1636	1536
1258	1517	1630	1635	1515	1440	1636
882	1635	1536	1515	1635	1536	1580
1254	1636	1517	1535	1476	1497	1489
1263	1137	1541	1476	1528	1580	1476
1634	1632	1636	1541	1519	1617	1540

Tablica 4.6: Usporedba rezultata za ciklaze

Još jednom možemo vidjeti da je drugi stupac jako sličan stupcima drugih istraživanja. Jedini uljez ovaj put je pozicija 1137.

Pozicija 1137, kao i 291 kod AT-domena, posebno je zanimljiva. U njoj isto tako ne prevladavaju "crtice", a ispala je statistički značajna na razini značajnosti od 5%. Doduše, na toj poziciji 14 proteina od 42 iz druge grupe originalnog poravnanja ima prazninu. U usporedbi s 2 proteina od njih 99 iz prve grupe AT-domena koji su imali "crticu" na poziciji 291, ipak možemo zaključiti da ovdje u ciklazama kod pozicije 1137 "crtica" igra dosta veću ulogu. No, ne možemo reći da "crtica" kod pozicije 1137 prevladava u toj drugoj podfamiliji ciklaza pa je i pozicija 1137 kod ciklaza zaslužila spomen. Što se tiče aminokiselina koje se pojavljuju na tom mjestu, u drugoj grupi imamo 14 "crtica" i dvostruko više serina (S), dok kod prve grupe dominira fenilalanin (F) prisutan kod čak 24 proteina, a ima i po 4 tirozina (Y) i lizina (K) te jedan asparagin (N). Na grafu (2.2) možemo vidjeti da su vektorske reprezentacije serina koji dominira u drugoj skupini i klastera kojeg čine fenilalanin, tirozin, lizin i asparagin zajedno s još nekim aminokiselinama jako udaljene, što objašnjava značajnost pozicije 1137. Kao ni poziciju 291 u AT-domenama, niti poziciju 1137 kod ciklaza ne možemo naći u rezultatima drugih analiza ([16] ili [17]) jer je tamo ta pozicija u startu izbačena.

Za c=2.5, N=81

Dosada smo, općenito kod sve 3 proučavane familije, imali fiksne hiperparametre u borbi protiv numeričke nestabilnosti (konstanta c=1.5 u formuli S-statistike (3.1)) i u borbi protiv manjka proteina u poravnanju dodavanjem buke (konstruirali bismo matrice poravnanja dimenzija $(21 \cdot m, 5 \cdot n)$, gdje je m broj proteina u originalnom poravnanju, a n duljina proteina). Drugim riječima, osim što nam je c bio uvijek jednak 1.5, dodavali smo

po 20 novih nizova u poravnanje, za svaki protein iz originalnog poravnanja, rezultirajući u ukupno $21 \cdot m$ nizova u matrici poravnanja.

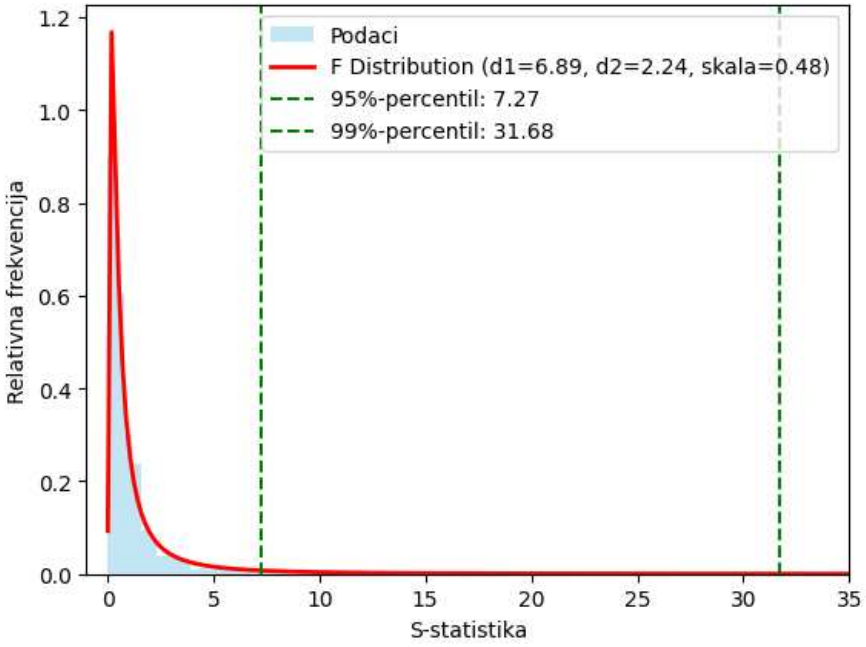
S ciljem dobivanja veće sličnosti distribucije S-statistike s nekom F-distribucijom i rezultirajuće veće pouzdanosti u zaključcima, radimo sljedeće. S nadom u bolju p-vrijednost KS testa, u poravnanje ćemo dodati po 80 novih nizova na jedan originalni. Dakle, sada će nam N biti jednak 81.

Stoga, konstruiramo novu matricu poravnanja P dimenzija (6075, 10200). Sada su novi m_1 i m_2 jednaki 2673 i 3402 redom. Eksperimentirajući s konstantom stabilizacije c, za $c = 2.5$ smo dobili zadovoljavajuće rezultate. Dakle, nad stupcima matrice P izvršili smo izračun S-statistike s konstantom $c = 2.5$ i sada prikazujemo najvećih 50 vrijednosti S-statistike, zajedno s pripadnim pozicijama poravnanja:

Redni broj	Pozicija	S-statistika
1	1297	174.41
2	1474	166.97
3	1256	141.96
4	1634	136.16
5	1265	133.57
6	883	125.77
7	882	115.41
8	1258	114.81
9	1254	114.21
10	1263	113.82
11	1259	113.00
12	1296	112.36
13	1255	111.16
14	1264	110.08
15	1257	110.02
16	1260	105.15
17	1488	100.34
18	881	98.95
19	1298	98.22
20	1253	92.62
21	1535	90.21
22	1262	89.09
23	1536	62.26
24	1517	52.22
25	1630	50.58
26	1466	48.65
27	1568	44.46
28	583	40.90
29	1635	39.23
30	1636	37.76
31	1473	37.14
32	1467	36.58
33	1566	36.30
34	1572	36.28
35	1567	36.09
36	1137	31.04
37	1243	29.82
38	1632	29.16
39	1244	28.69
40	272	24.71
41	651	24.65
42	585	23.60
43	584	23.53
44	1506	22.91
45	315	21.71
46	1529	21.10
47	317	20.81
48	271	20.57
49	1533	19.78
50	1247	18.37

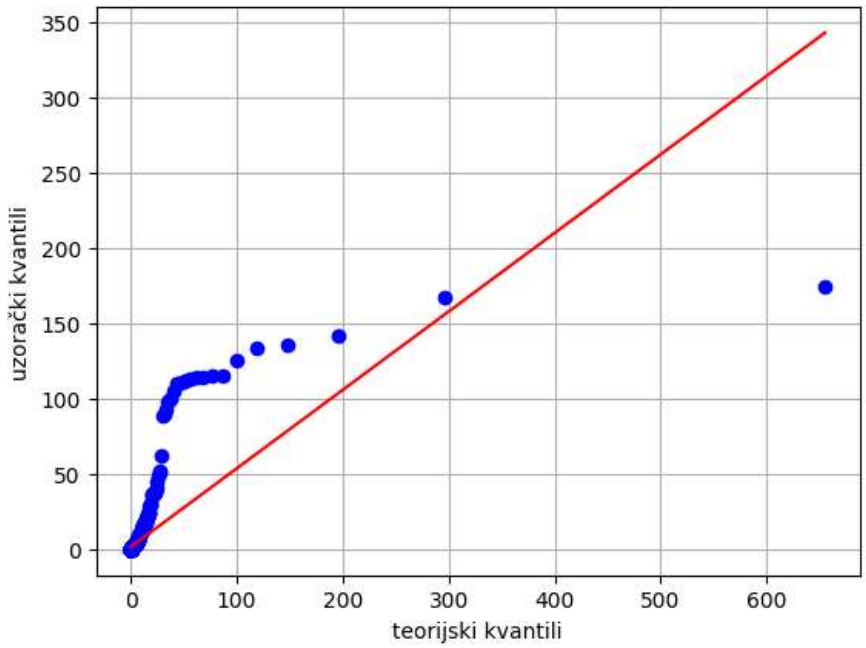
Tablica 4.7: Najspecifičnije pozicije ciklaza ($c = 2.5, N = 81$)

Ispod je prikazan histogram vrijednosti S-statistike s crvenom linijom koja predstavlja teorijsku funkciju gustoće F-distribucije s parametrima procijenjenima iz podataka metodom maksimalne vjerodostojnosti. Procijenjena je F-distribucija sa stupnjevima slobode od otprilike 6.89 i 2.24, te parametrom skaliranja od približno 0.48. Kako je jako mala relativna frekvencija S-statistika koje su veće od 35, prikazujemo skraćenu verziju histograma na kojoj se puno bolje vidi distribucija.



Slika 4.10: Histogram S-statistike s procijenjenom F-distribucijom za ciklaze ($c = 2.5, N = 81$))

Prikažimo i qq-vjerojatnosni graf teorijskih kvantila F-distribucije (df1=6.89, df2=2.24, skala=0.48) na x-osi i kvantila S-statistike na y-osi. Nažalost ne primjećujemo nikakvo poboljšanje u odnosu na sliku 4.8.



Slika 4.11: qq-vjerojatnosni graf za ciklaze ($c = 2.5, N = 81$)

Potom je Kolmogorov-Smirnovljevim (KS) testom testirana pripadnost podataka (vrijednosti S-statistike) procijenjenoj F-distribuciji. Testna statistika iznosi 0.02845, a p-vrijednost 0.07. Zaključujemo da ne možemo odbaciti pripadnost podataka F-distribuciji na razinama značajnosti od 1% i 5%.

Pod pretpostavkom F-distribuiranosti, odredimo one pozicije koje su statistički značajne za raspodjelu proteina u podfamilije na razini od 5% i 1%. Kako 95%-percentil F-distribucije otprilike iznosi 7.27, na razini od 5% statistički su značajne prve 103 pozicije iz 4.7. 99%-percentil F-dstribucije otprilike iznosi 31.68 pa je na razini od 1% statistički značajno prvih 35 pozicija iz 4.7. Na tablici 4.7 zelenom linijom označili smo granicu iznad koje slijede statistički značajne pozicije, na razini značajnosti od 1%.

U usporedbi s rezultatima S-statistike za $c = 1.5, N = 21$ i pripadnim rangiranjem, vidimo da su rezultati skoro identični, jedino su se neke pozicije u kojima nema uopće "crtica" a statistički su značajne, popele u rangiranju (1634, 1488, 1535, 1536, 1630, 1517, 1635, 1636).

Što se tiče top 10 pozicija, sa i bez pozicija s prevladavajućim "crticama", situacija je skoro pa identična kao ranije s $c = 1.5, N = 21$ (prva dva stupca u tablici 4.6). Jedina razlika je to što bi se pozicija 1634 popela sada na 4. mjesto, a pozicija 882 na 7. mjesto u prvom stupcu. To znači da bi prvih 10 najznačajnijih pozicija ostalo isto. Upravo zato i t-SNE graf izgleda identično kao ranije, pa smo ga zato izostavili. Napomenimo da je sada

uspješnost klasificiranja od 100% iščitana iz grafa 4.9 statistički opravdana. Drugi stupac bi bio identičan, osim što bi se pozicije 1517 i 1630 međusobno zamijenile. Kako smo sve zaključke koje smo napisali ranije komentirajući tablicu 4.6 donijeli na osnovu drugog stupca koji se ne bi skoro uopće promijenio, sve te zaključke potvrđujemo, ali sada uz veću pouzdanost, s obzirom na rezultate KS testa.

4.4 Kinaze

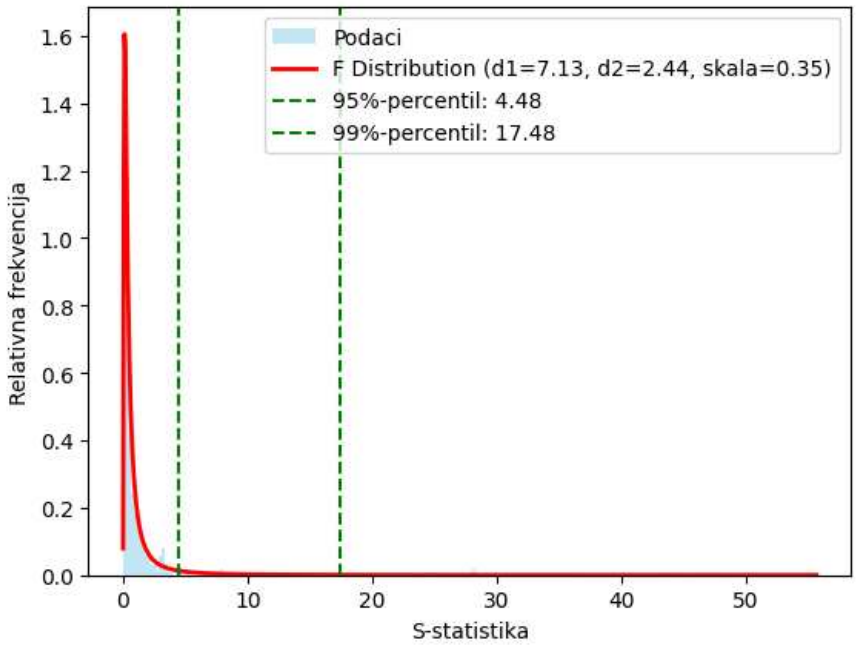
Za c=1.5, N=21

Za proteinsku familiju kinaza, originalno smo imali poravnanje 215 proteina duljine 600, pri čemu je bilo 85 enzima iz prve podfamilije, a 130 iz druge. Zatim smo numeričkim reprezentiranjem konstruirali matricu poravnanja dimenzija (215, 3000). Konačno, nakon generiranja novih numeričkih nizova, dobivamo novu matricu poravnanja P dimenzija (4515, 3000), gdje smo pritom novogenerirane retke matrice P klasificirali u onu grupu kojoj je pripadao originalni protein iz kojeg smo generirali te nove retke. Dakle, m_1 i m_2 koje ubacujemo u formulu 3.1 su redom 1785 i 2730. Nad stupcima matrice P izvršili smo izračun S-statistike i sada prikazujemo najvećih 35 vrijednosti S-statistike, zajedno s pripadnim pozicijama originalnog poravnanja:

Redni broj	Pozicija	S-statistika
1	378	55.70
2	443	50.51
3	327	46.87
4	133	44.70
5	343	43.52
6	444	35.09
7	326	28.12
8	65	28.08
9	532	25.94
10	345	24.78
11	352	21.40
12	355	19.27
13	353	17.56
14	241	14.99
15	354	13.56
16	383	12.24
17	109	11.45
18	527	10.40
19	108	9.98
20	542	8.69
21	376	8.42
22	101	7.96
23	342	7.85
24	445	6.37
25	365	5.45
26	528	5.29
27	344	5.27
28	261	4.85
29	59	4.79
30	360	4.65
31	18	4.53
32	381	4.50
33	262	3.93
34	245	3.70
35	446	3.55

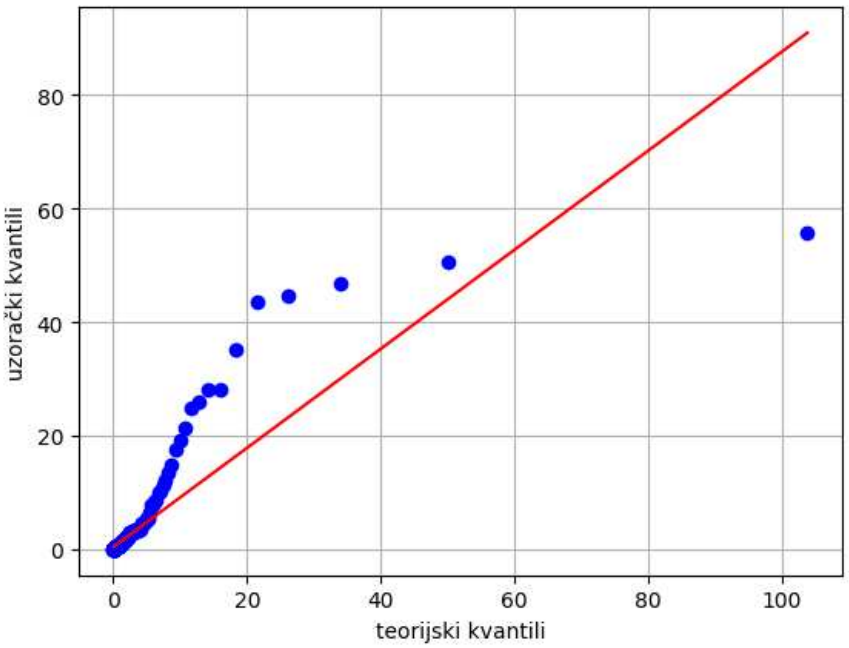
Tablica 4.8: Najspecifičnije pozicije kinaza ($c = 1.5, N = 21$)

Ispod je prikazan histogram vrijednosti S-statistike s crvenom linijom koja predstavlja teorijsku funkciju gustoće F-distribucije s parametrima procijenjenima iz podataka metodom maksimalne vjerodostojnosti. Procijenjena je F-distribucija sa stupnjevima slobode od otprilike 7.13 i 2.44, te parametrom skaliranja od približno 0.35.



Slika 4.12: Histogram S-statistike s procijenjenom F-distribucijom za kinaze ($c = 1.5, N = 21$)

Prikažimo i qq-vjerojatnosni graf teorijskih kvantila F-distribucije (df1=7.13, df2=2.44, skala=0.35) na x-osi i kvantila S-statistike na y-osi. Slično kao kod ciklaza, možemo primijetiti veliko odstupanje koje kreće otprilike nakon što S-statistika prijeđe vrijednost od oko 10. No, vidimo i da ondje gdje su podaci najgušći, a to su S-statistike vrijednosti do oko 10, pravac većinom jako dobro opisuje podatke.



Slika 4.13: qq-vjerojatnosni graf za kinaze ($c = 1.5, N = 21$)

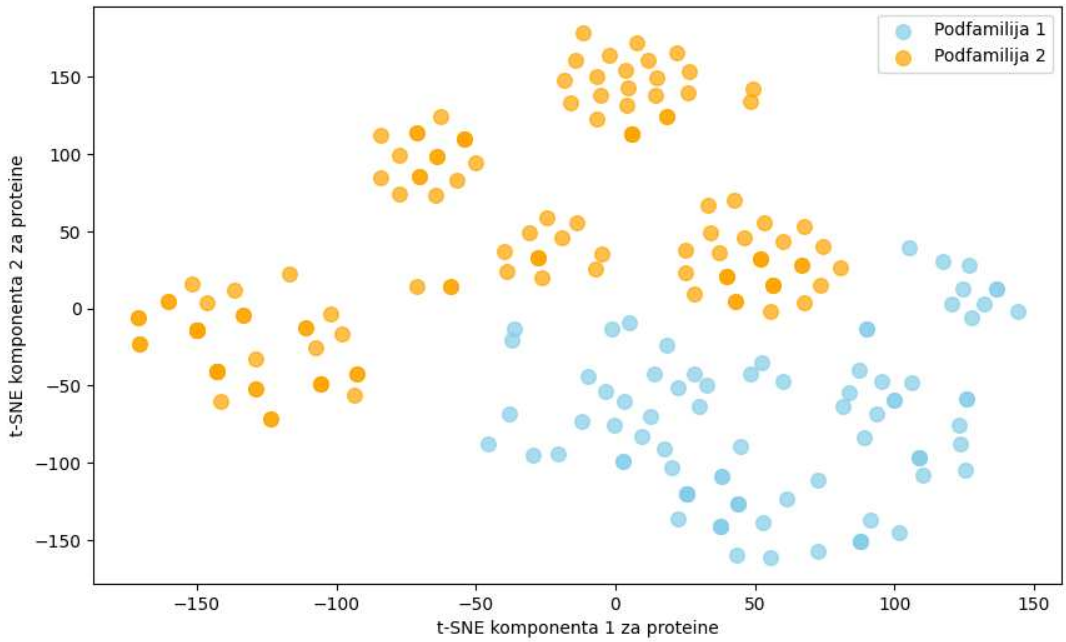
Potom je Kolmogorov-Smirnovljevim (KS) testom testirana pripadnost podataka (vrijednosti S-statistike) procijenjenoj F-distribuciji. Testna statistika iznosi 0.062, a p-vrijednost 0.02. Odbacujemo hipotezu da S-statistika prati F-distribuciju na razini značajnosti od 5%. Ipak, radi analize pozicija, u nastavku ćemo pretpostaviti da je S-statistika F-distribuirana. Napomenimo da sljedeći rezultati u vidu statističke značajnosti pozicija nisu pouzdani.

Pod labavom pretpostavkom F-distribuiranosti, odredimo one pozicije koje su statistički značajne za raspodjelu proteina u podfamilije na razini od 5% i 1%. Kako 95%-percentil F-distribucije otprilike iznosi 4.48, na razini od 5% statistički su značajne prve 32 pozicije s najvećom S-statistikom. 99%-percentil F-dstribucije otprilike iznosi 17.48 pa je na razini od 1% statistički značajno prvih 13 pozicija s najvećom S-statistikom. Na tablici 4.8 zelenim linijama označili smo granice iznad koje slijede statistički značajne pozicije, na razinama značajnosti od 5% i 1%.

Detaljnijim pregledavanjem zadanog poravnanja primjećujemo da ima dosta pozicija gdje za jednu od grupa prevladavaju “crtice”. Naime, to su pozicije 378, 383, 443-445, 326, 327, 133, 352-355, 527, 445, 108, 109, 59 itd. Velika većina njih se pojavljuje visoko u tablici 4.8.

Prikažimo sada t-SNE graf. Možemo primijetiti da su proteini dviju podfamilija jasno separirani. Također, unutar podfamilije 2, možemo uočiti 5 različitih manjih klastera, dok su proteini podfamilije 1, svi povezani u jedan veliki klaster. Stoga možemo reći

da je točnost klasificiranja proteina na temelju najznačajnijih 10 pozicija opet 100%-tna.



Slika 4.14: t-SNE graf za kinaze

U sljedećoj tablici usporedit ćemo naše rezultate s rezultatima nekih drugih sličnih istraživanja, sasvim analogno kao i u drugim familijama.

4.8	**4.8(bez '-')**	**[16](1)**	**[16](2)**	**[16](3)**	**[17](BLOSUM)**	**[17](PAM)**
378	343	65	532	241	65	343
443	65	343	241	532	343	542
327	532	345	65	65	542	65
133	345	532	343	343	532	381
343	241	241	345	345	63	345
444	542	261	261	18	381	532
326	376	101	18	377	345	344
65	101	376	342	261	344	63
532	342	542	344	245	337	245
345	365	245	528	528	342	376

Tablica 4.9: Usporedba rezultata za kinaze

Još jednom možemo vidjeti da je drugi stupac jako sličan stupcima drugih istraživanja. Je-

dini uljezi ovaj put su pozicije 101 i 365. Pozicija 101 se može naći u rezultatima [17] pod rednim brojevima 35. i 31., koristeći BLOSUM i PAM redom.

Pozicije 365 i 360, posebno su zanimljive. U njima ne prevladavaju "crtice", a ispale su statistički značajne na razini značajnosti od 5%. Na poziciji 365 u grupi 2 dominira fenilalanin (F), a u grupi 1 leucin (L) i valin (V), dok na poziciji 360 u grupi 2 dominira serin (S), a u grupi 1 valin (V) i izoleucin (I). Brzim pogledom na graf 2.2, odmah vidimo da što se tiče udaljenosti vektora, to ima smisla. Istaknimo da kod obje pozicije postoji samo jedan protein s "crticom" na tom mjestu, i to se baš radi o istom proteinu. Ni 365 niti 360 ne možemo naći u rezultatima drugih analiza ([16] ili [17]) jer je tamo ta pozicija u startu izbačena.

Za $\mathbf{c = 0.5, N = 81}$

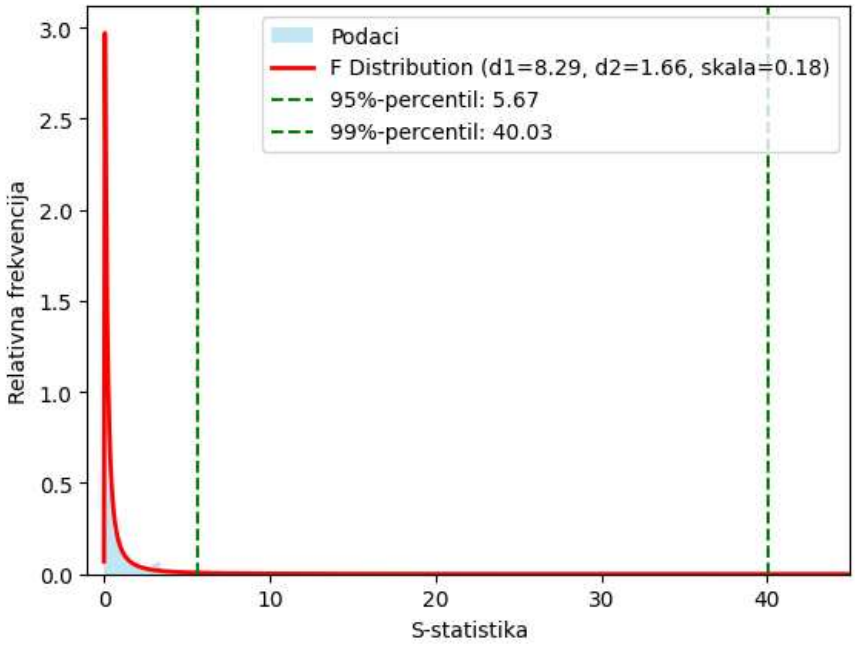
Kao kod ciklaza, motivirani smo da probamo prilagoditi hiperparametre c i N kako bi za proširenu skup podataka (veći N) dobili veću p-vrijednost u KS testu. Za N=81, dobivamo novu matricu poravnanja dimenzija (17415, 3000), gdje je $m_1 = 6885, m_2 = 10530$. Sada na tim novim podacima, za $c = 0.5$ dobivamo p-vrijednost KS testa za pripadnost novih vrijednosti S-statistike F-distribuciji od 0.06. Opet, top 35 pozicija se ne razlikuje skoro uopće od tablice 4.8:

Redni broj	Pozicija	S-statistika
1	343	66.22
2	378	57.22
3	443	51.74
4	327	47.94
5	133	45.44
6	444	35.71
7	65	32.81
8	326	28.52
9	532	28.30
10	345	28.04
11	352	21.56
12	355	19.39
13	353	17.68
14	241	15.84
15	354	13.62
16	383	12.29
17	109	11.51
18	527	10.43
19	108	10.02
20	542	9.39
21	376	8.77
22	101	8.65
23	342	8.54
24	445	6.38
25	365	5.51
26	344	5.35
27	528	5.34
28	261	5.02
29	59	4.79
30	381	4.68
31	360	4.66
32	18	4.60
33	262	4.06
34	245	3.80
35	446	3.56

Tablica 4.10: Najspecifičnije pozicije kinaza ($c = 0.5, N = 81$)

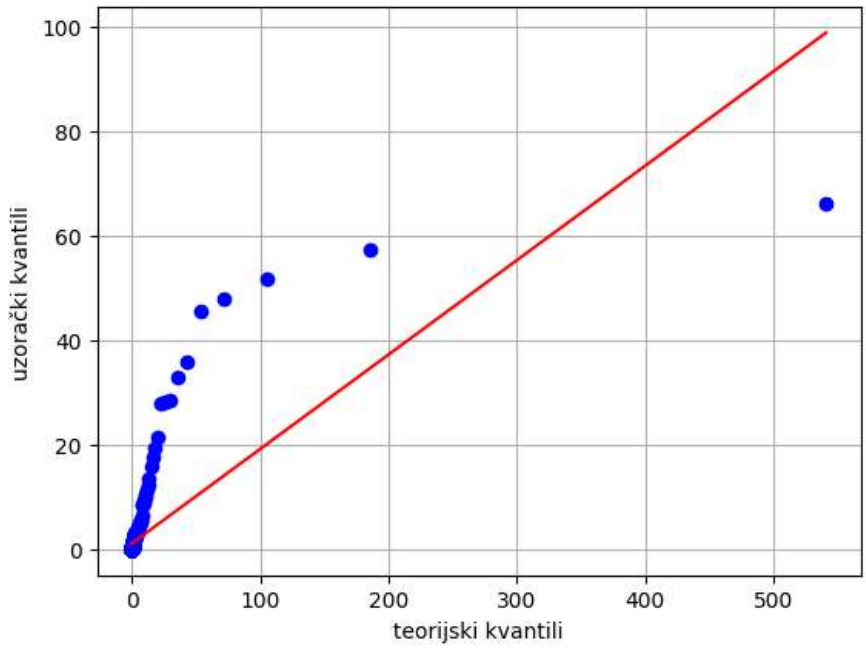
Ispod je prikazan histogram vrijednosti S-statistike s crvenom linijom koja predstavlja teorijsku funkciju gustoće F-distribucije s parametrima procijenjenima iz podataka metodom maksimalne vjerodostojnosti. Procijenjena je F-distribucija sa stupnjevima slobode od otprilike 8.29 i 1.66, te parametrom skaliranja od 0.18. Kako je pozicija za koje je S-statistika veća od 45 samo petero, prikazujemo skraćenu verziju histograma na kojoj se

bolje vidi distribucija.



Slika 4.15: Histogram S-statistike s procijenjenom F-distribucijom za kinaze ($c = 0.5, N = 81$))

Prikažimo i qq-vjerojatnosni graf teorijskih kvantila F-distribucije (df1=8.29, df2=1.66, skala=0.18) na x-osi i kvantila S-statistike na y-osi. Nažalost ne primjećujemo nikakvo poboljšanje u odnosu na sliku 4.13. Štoviše, sada se točke još ranije počinju odvajati od pravca.



Slika 4.16: qq-vjerojatnosni graf za kinaze ($c = 0.5, N = 81$)

Potom je Kolmogorov-Smirnovljevim (KS) testom testirana pripadnost podataka (vrijednosti S-statistike) procijenjenoj F-distribuciji. Testna statistika iznosi 0.05346, a p-vrijednost 0.06. Zaključujemo da ne možemo odbaciti pripadnost podataka F-distribuciji na razinama značajnosti od 1% i 5%.

Pod pretpostavkom F-distribuiranosti, odredimo one pozicije koje su statistički značajne za raspodjelu proteina u podfamilije na razini od 5% i 1%. Kako 95%-percentil F-distribucije otprilike iznosi 5.67, na razini od 5% statistički su značajne prve 24 pozicije iz tablice 4.10. 99%-percentil F-distribucije otprilike iznosi 40.03 pa je na razini od 1% statistički značajno prvih 5 pozicija iz tablice 4.10. Na istoj tablici zelenom linijom označili smo granicu iznad koje slijede statistički značajne pozicije, na razini značajnosti od 5% 1%.

U usporedbi s rezultatima S-statistike za $c = 1.5, N = 21$ i pripadnim rangiranjem, vidimo da su rezultati opet skoro identični. Može se primijetiti da su se neke pozicije na kojima nijedan protein nema "crticu", a statistički su značajne, popele u rangiranju (343 i 65).

Što se tiče top 10 pozicija, sa i bez pozicija s prevladavajućim "crticama", situacija je skoro pa identična kao ranije s $c = 1.5, N = 21$ (prva dva stupca u tablici 4.9). Jedina razlika je to što bi se pozicija 343 popela na 1. mjesto, a pozicija 65 na 7. mjesto u prvom stupcu. To opet znači da bi prvih 10 najznačajnijih pozicija ostalo isto. Upravo zato i t-SNE graf izgleda identično kao ranije, pa smo ga zato izostavili. Napomenimo da je sada

uspješnost klasificiranja od 100% iščitana iz grafa 4.14 statistički opravdana. Drugi stupac bi bio skroz identičan. Kako smo sve zaključke koje smo napisali ranije komentirajući tablicu 4.9 donijeli na osnovu drugog stupca koji se ne bi uopće promijenio, sve te zaključke potvrđujemo, ali sada uz veću pouzdanost, s obzirom na rezultate KS testa.

4.5 KR-domene

Za c = 1.5

Za proteinsku familiju KR-domena, originalno smo imali poravnanje 72 proteina duljine 218, pri čemu je 39 proteina pripadalo jednoj podfamiliji AT-domena, a 33 proteina drugoj. Zatim smo numeričkim reprezentiranjem konstruirali matricu poravnanja dimenzija (72, 1090). Konačno, nakon generiranja novih numeričkih nizova, dobivamo novu matricu poravnanja P dimenzija (1512, 1090), gdje smo pritom novogenerirane retke matrice P klasificirali u onu grupu kojoj je pripadao originalni protein iz kojeg smo generirali te nove retke. Dakle, m_1 i m_2 koje ubacujemo u formulu 3.1 su redom 819 i 693. Nad stupcima matrice P izvršili smo izračun S-statistike i sada prikazujemo najvećih 30 vrijednosti S-statistike, zajedno s pripadnim pozicijama originalnog poravnanja:

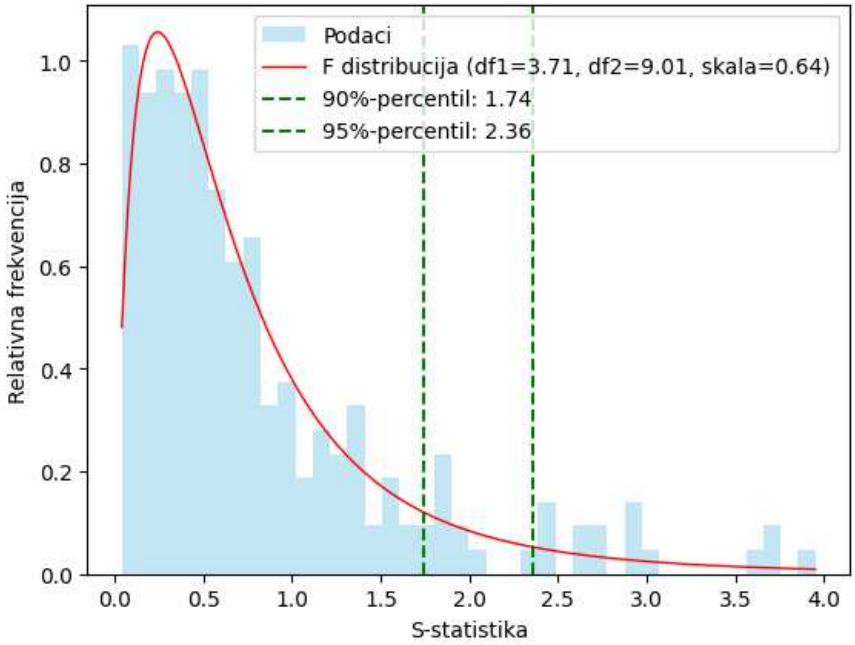
Redni broj	Pozicija	S-statistika
1	176	3.95
2	183	3.71
3	125	3.70
4	155	3.64
5	118	2.99
6	20	2.97
7	151	2.97
8	144	2.88
9	132	2.78
10	23	2.75
11	50	2.68
12	184	2.65
13	92	2.47
14	171	2.43
15	164	2.41
16	188	2.29
17	210	2.05
18	123	2.00
19	191	1.92
20	211	1.90
21	124	1.87
22	17	1.84
23	127	1.83
24	147	1.81
25	116	1.76
26	47	1.74
27	198	1.63
28	26	1.62
29	163	1.58
30	22	1.53

Tablica 4.11: Najspecifičnije pozicije KR-domena ($c = 1.5$)

U poravnanju ove familije imamo jako puno **konzerviranih** pozicija. Konzervirana pozicija u poravnanju je ona na kojoj svi proteini imaju istu aminokiselinu. Takve su pozicije 118, 20, 151, 144, 23, 50, 184, 92, 171, dok na pozicijama 188, 210, 191, 211, 17, 47 svi proteini osim jednog imaju istu aminokiselinu. Sve takve konzervirane ili skoro konzervirane pozicije će imati visoku S-statistiku ako ne koristimo konstantu c u formuli 3.1, kao što smo i objasnili u poglavlju 3.2. U tablici 4.11 takve pozicije smo označili

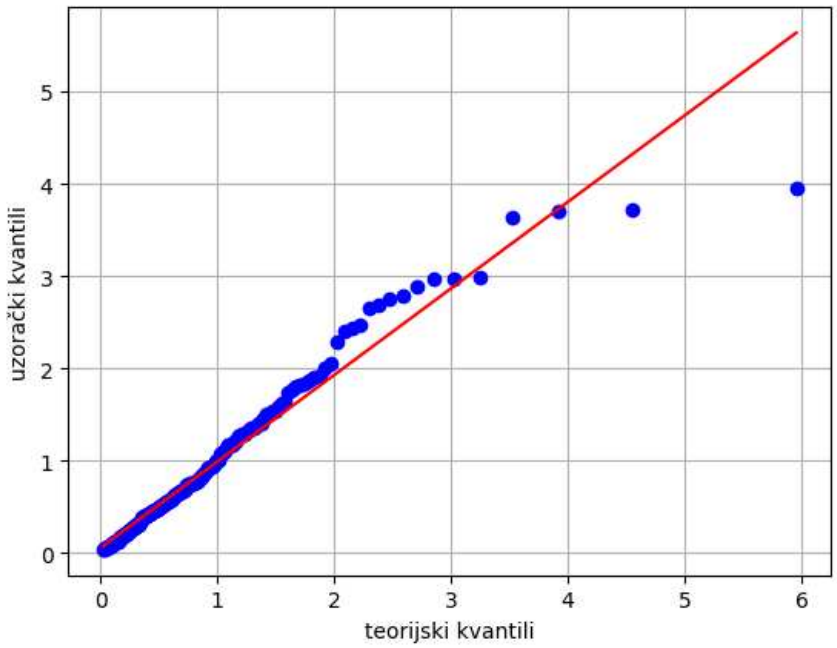
crvenom bojom. Dosada u poravnanjima ostalih familija nismo imali problem s takvim pozicijama jer ih u nekim poravnanjima nije ni bilo, a negdje ih je bilo malo i raspon S-statistike je bio veći. Ovdje, kod KR-domena, za S-statistiku uz $c = 1.5$ imamo jako mali raspon (0.04 - 3.95), a za konzervirane pozicije dobivamo vrijednosti koje dosežu skoro 3.

Ispod je prikazan histogram vrijednosti S-statistike s crvenom linijom koja predstavlja teorijsku funkciju gustoće F-distribucije s parametrima procijenjenima iz podataka metodom maksimalne vjerodostojnosti. Procijenjena je F-distribucija sa stupnjevima slobode od otprilike 3.71 i 9.01, te parametrom skaliranja od približno 0.64.



Slika 4.17: Histogram S-statistike s procijenjenom F-distribucijom za KR-domene ($c = 1.5$)

Prikažimo i qq-vjerojatnosni graf teorijskih kvantila F-distribucije (df1=3.71, df2=9.01, skala=0.64) na x-osi i kvantila S-statistike na y-osi. Ovdje imamo najbolje prijanjanje točaka uz pravac dosada, te uz odličan "fit" spomenute F-distribucije na histogramu 4.17, čini nam se da ova S-statistika prati specificiranu F-distribuciju.

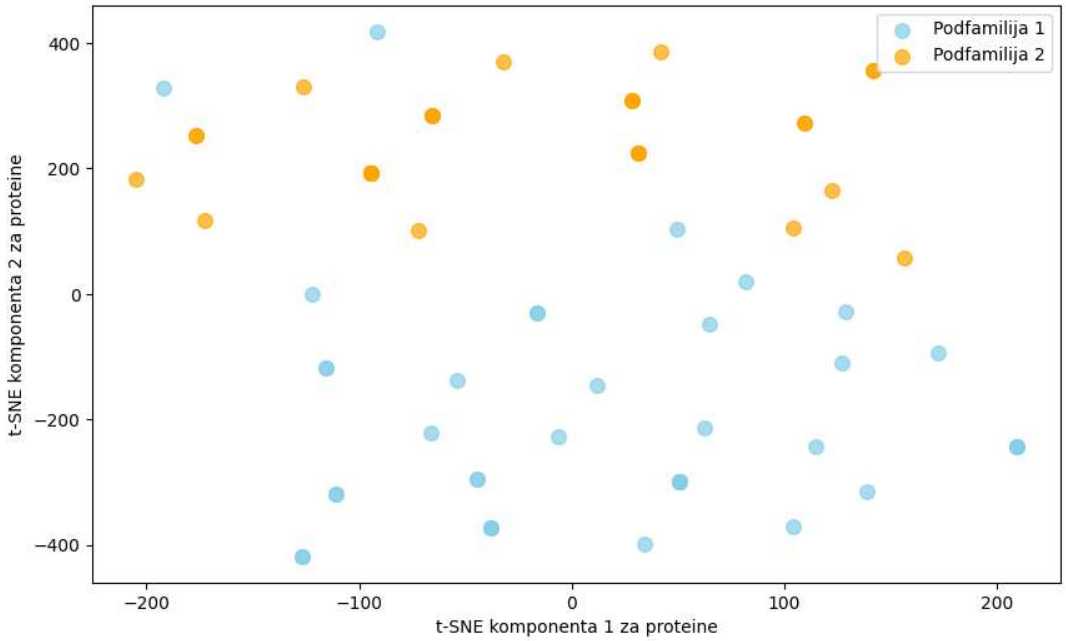


Slika 4.18: qq-vjerojatnosni graf za KR-domene ($c = 1.5$)

Potom je Kolmogorov-Smirnovljevim (KS) testom testirana pripadnost podataka (vrijednosti S-statistike) procijenjenoj F-distribuciji. Testna statistika iznosi 0.0336, a p-vrijednost 0.96. Nikako ne možemo odbaciti hipotezu da S-statistika prati F-distribuciju na svim razumnim razinama značajnosti. Naša slutnja o mogućnosti pretpostavljanja F-distribuiranosti se obistinila.

Pod čvrstom pretpostavkom F-distribuiranosti, odredimo one pozicije koje su statistički značajne za raspodjelu proteina u podfamilije na razini od 5% i 1%. Kako 99%-percentil F-distribucije otprilike iznosi 4.2, na razini od 1% statistički značajna nije nijedna pozicija. 95%-percentil F-dstribucije otprilike iznosi 2.36 pa je na razini od 5% statistički značajno prvih 15 pozicija s najvećom S-statistikom. Izračunajmo i 90%-percentil procijenjene F-distribucije. On iznosi 1.74 pa je na razini od 10% statistički značajno prvih 25 pozicija s najvećom S-statistikom. Na tablici 4.8 zelenim linijama označili smo granice iznad koje slijede statistički značajne pozicije, na razinama značajnosti od 10% i 5%.

Prikažimo sada i t-SNE graf. Možemo primijetiti da su proteini obiju podfamilija nekako široko postavljeni i nema očitih zgusnutih klastera. Pretpostavljamo da je tako zato što u prvih 10 najznačajnijih pozicija S-statistike ima čak 5 konzerviranih pozicija.



Slika 4.19: t-SNE graf za KR-domene ($c = 1.5$)

Za c = 5.5

Da bi se smanjio broj visokorangiranih konzerviranih pozicija, potrebno je povećati konstantu stabilizacije c. Povećanjem gubimo na praćenju F-distribucije. Kao konstanta koja potvrđuje ovo razmatranje, ali i dalje zadržava pouzdanim pretpostavku o F-distribuiranosti S-statistike, odabrana je $c = 5.5$. Koristeći nju u formuli S-statistike 3.1, dobiveno je sljedećih 30 najspecifičnijih pozicija:

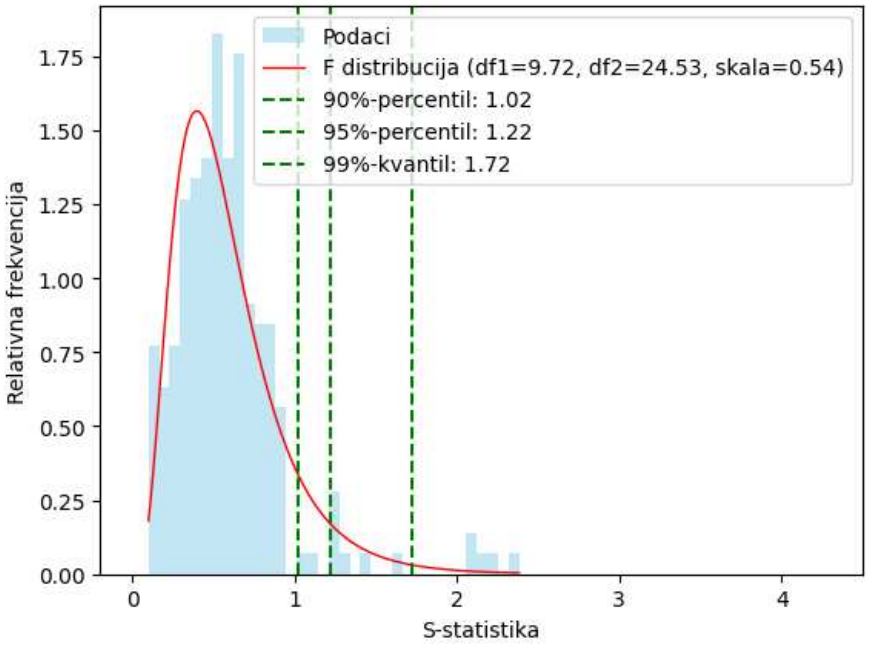
Redni broj	Pozicija	S-statistika
1	125	2.38
2	155	2.22
3	176	2.13
4	183	2.11
5	132	2.11
6	124	1.65
7	164	1.45
8	142	1.31
9	127	1.27
10	123	1.27
11	38	1.26
12	147	1.25
13	172	1.10
14	162	1.08
15	133	0.91
16	20	0.90
17	151	0.90
18	118	0.90
19	144	0.90
20	23	0.89
21	50	0.89
22	184	0.89
23	188	0.88
24	92	0.87
25	171	0.87
26	191	0.86
27	136	0.85
28	217	0.85
29	210	0.84
30	113	0.83

Tablica 4.12: Najspecifičnije pozicije KR-domena (c=5.5)

Možemo primijetiti da smo osjetno “raščistili od crvene boje” čak prvih 15 najznačajnijih pozicija. Naime, konzervirane pozicije u poretku sada dolaze tek nakon 15. mjesta.

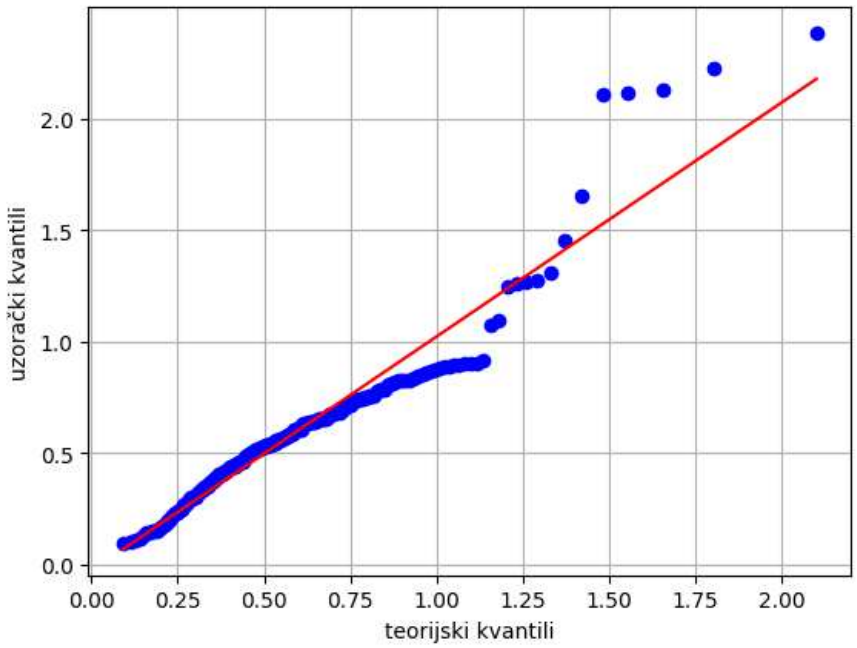
Ispod je prikazan histogram vrijednosti S-statistike s crvenom linijom koja predstavlja teorijsku funkciju gustoće F-distribucije s parametrima procijenjenima iz podataka metodom maksimalne vjerodostojnosti. Procijenjena je F-distribucija sa stupnjevima slobode

od otprilike 9.72 i 24.53, te parametrom skaliranja od približno 0.54.



Slika 4.20: Histogram S-statistike s procijenjenom F-distribucijom za KR-domene (c = 5.5)

Prikažimo i qq-vjerojatnosni graf teorijskih kvantila F-distribucije (df1=9.72, df2=24.53, skala=0.54) na x-osi i kvantila S-statistike na y-osi. Sada primjećujemo da točke malo lošije prate pravac, nego što je to bilo u 4.18, no i dalje većina točaka pada u blizini pravca, pogotovo do vrijednosti od približno 0.75 na y-osi.



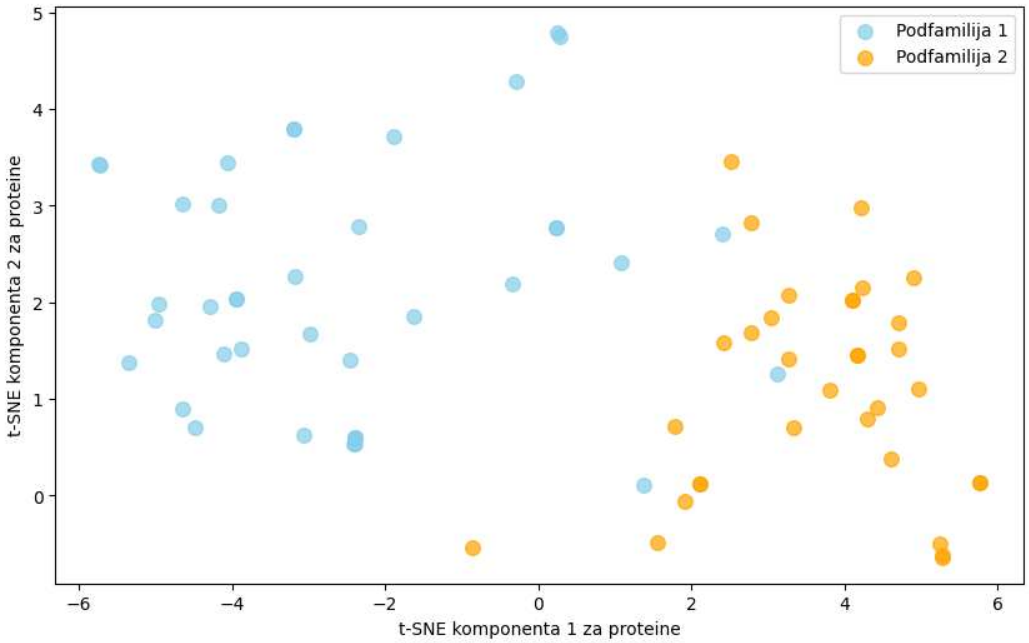
Slika 4.21: qq-vjerojatnosni graf za KR-domene ($c = 5.5$)

Kolmogorov-Smirnovljevim (KS) testom testirana je pripadnost podataka (vrijednosti S-statistike) procijenjenoj F-distribuciji. Testna statistika iznosi 0.0804, a p-vrijednost 0.11. Dakle, ne možemo odbaciti hipotezu da S-statistika prati F-distribuciju na razinama značajnosti od 10%, 5% i 1%.

Pod pretpostavkom F-distribuiranosti, odredimo one pozicije koje su statistički značajne za raspodjelu proteina u podfamilije na razini od 10%, 5% i 1%. Kako 99%-percentil F-distribucije otprilike iznosi 1.72, na razini od 1% statistički je značajno prvih 5 pozicija s najvećom S-statistikom. 95%-percentil F-dstribucije otprilike iznosi 1.22 pa je na razini od 5% statistički značajno prvih 12 pozicija s najvećom S-statistikom. Izračunajmo i 90%-percentil procijenjene F-distribucije. On iznosi 1.02 pa je na razini značajnosti od 10% statistički značajno prvih 14 pozicija s najvećom S-statistikom. Na tablici 4.12 zelenim linijama označili smo granice iznad koje slijede statistički značajne pozicije, na razinama značajnosti od 10%, 5% i 1%. Primijetimo da sve pozicije koje su iole statistički značajne upadaju u prvih 15 pozicija s najvećom S-statistikom, otkuda smo uspjeli maknuti sve konzervirane pozicije, a pritom i pretpostavku o F-distribuiranosti uspjeli sačuvati relativno čvrstom.

Prikažimo sada t-SNE graf. Možemo vidjeti veću kondenziranost proteina i jedne i druge podfamilije oko pripadajućih klastera. Klasteri su sada jasno odvojivi, a imamo samo 3 pogrešne klasifikacije. I očekivali smo ovakvu, bolju situaciju od slike 4.19, s ob-

zirom da u top 10 pozicija S-statistike sada više nemamo niti jednu konzerviranu poziciju.



Slika 4.22: t-SNE graf za ciklaze ($c = 5.5$)

Što se tiče drugih istraživanja koja su uključivala KR-domene, u [17] je uzet raspon pozicija 114-155 kao onaj u kojem se nalazi većina specifičnih pozicija. Kad pogledamo naše rezultate, posebno tablicu 4.12, vidimo da tom rasponu pripada 9 od top 15 pozicija, odnosno 8 od top 12 pozicija koje su statistički značajne na razini od 5%.

Poglavlje 5

Zaključak

U prošlom poglavlju za svaku smo od promatranih proteinskih familija priložili rezultate statističke analize. Glavni problem kojim se ovaj rad bavio je definiranje onih pozicija iz poravnanja proteinske familije koje rade razliku između dvije podgrupe te familije, odnosno koje određuju kojoj od te dvije podfamilije pripada svaki protein iz poravnanja. Kako bi imali i statističku pouzdanost za značajnost rezultata, vrijednosti razdvajajuće S-statistike koja kvantificira značajnost pozicija na razini jedne familije, modelirali smo procijenjenom F-distribucijom.

Za familiju AT-domena, pretpostavka da S-statistika prati F-distribuciju pokazala se validnom. Kao daleko statistički najznačajnija pozicija ispala je 212, čak na razini značajnosti od 1%. Na razini značajnosti od 5%, statistički je značajno još 17 pozicija, a navodimo 156, 282, 71 i 291. Pripadni t-SNE graf prikazuje dobru separiranost dvije podfamilije.

Kod MDH/LDH familije, na osnovu i qq-vjerojatnosnog grafa i histograma i KS testa, čvrsto možemo pretpostaviti F-distribuiranost. Statistički značajno na razini značajnosti od 1% ispalo je 6 pozicija, a to su redom 125, 276, 289, 354, 144 i 148. Njih još 15 ispalo je statistički značajno na razini značajnosti od 5%. Očekivano, i t-SNE graf je dobro odvojio dvije podfamilije, uz samo 2 pogrešne klasifikacije.

Za ciklaze, uobičajeni hiperparametri $c = 1.5, N = 21$ nisu bili dovoljni da S-statistika na tim podacima i s tom konstantom stabilizacije c prati F-distribuciju. Dodatnim povećanjem broja nizova u poravnanju što odgovara $N = 81$, dobili smo podatke na kojima F-distribuiranost S-statistike, uz novu vrijednost konstante stabilizacije od $c = 2.5$, nismo mogli odbaciti KS testom. Statistički najznačajnije pozicije su ispale 1297, 1474, 1256, 1634 i 1265, koje su i pri uobičajenim hiperparametrima bile pri samom vrhu. Statistički značajno na razini značajnosti od 1%, ispalo je čak 35 pozicija, uključujući naravno i gore spomenute. Graf t-SNE, identičan za prvi i drugi odabir hiperparametara, odlično je centrirao jednu podfamiliju, što je rezultiralo dobrim odvajanjem dvije podfamilije.

Ni za kinaze, uobičajeni hiperparametri $c = 1.5, N = 21$ nisu bili dovoljni da S-statistika na tim podacima i s tom konstantom stabilizacije c prati F-distribuciju. Opet, dodatnim povećanjem broja nizova u poravnanju što odgovara $N = 81$, dobili smo podatke na kojima F-distribuiranost S-statistike, sada uz novu vrijednost konstante stabilizacije od $c = 0.5$, nismo mogli odbaciti KS testom. Kao statistički najznačajnije, pokazale su se pozicije 343, 378, 443, 327 i 133, koje su ujedno i jedine pozicije statistički značajne na razini značajnosti od 1%. Iste te pozicije, u malo drukčijem poretku, bile su u prvih 5 pozicija rangiranja i pri korištenju uobičajenih hiperparametara. Statistički značajne na razini značajnosti od 5%, ispale su prve 32 pozicije s najvećom S-statistikom. Na t-SNE grafu, opet identičnom za prvi i drugi odabir hiperparametara, vidi se lijepo odvajanje dvije podfamilije. Prilično vidljivo je i grupiranje u 5 klastera unutar jedne od podfamilija.

Konačno, u slučaju KR-domena, imali smo puno konzerviranih pozicija koje su, zbog numeričke nestabilnosti S-statistike u tom slučaju, završavale pri vrhu rangiranja. Tome smo donekle stali na kraj tako što smo povećali konstantu stabilizacije na $c = 5.5$, održavajući pretpostavku o F-distribuiranosti S-statistike i dalje validnom, kao što je bila pri uobičajenim hiperparametrima. Prelaskom na $c = 5.5$, kao statistički značajne pozicije, na razini značajnosti od 1%, dobili smo pozicije 125, 155, 176, 183 i 132. Konzervirane pozicije smo potisnuli dolje u rangiranju, iza 15. mjesta, dok je prvih 14, odnosno, prvih 12 pozicija u rangiranju, statistički značajno na razini značajnosti od 10%, odnosno, 5%. Što se tiče t-SNE prikaza, prelaskom na $c = 5.5$, dobili smo graf na kojem se vidi puno bolje separiranje dviju podfamilija, nego što je to bio slučaj kod uobičajenih hiperparametara.

Navedeni rezultati posljedica su novog matematičkog pristupa rješavanju ovog problema. Potencijalno daju važnu informaciju za daljnja biološko-eksperimentalna i bioinformatička istraživanja.

Bibliografija

[1] _____, *Cyclic adenosine monophosphate — Wikipedia, The Free Encyclopedia*, https://en.wikipedia.org/w/index.php?title=Cyclic_adenosine_monophosphate&oldid=1272538245, (datum pristupanja: siječanj 2025.).

[2] _____, *Cyclic guanosine monophosphate — Wikipedia, The Free Encyclopedia*, https://en.wikipedia.org/w/index.php?title=Cyclic_guanosine_monophosphate&oldid=1263060756, (datum pristupanja: siječanj 2025.).

[3] _____, *Fusion protein — Wikipedia, The Free Encyclopedia*, https://en.wikipedia.org/w/index.php?title=Fusion_protein&oldid=1223920679, (datum pristupanja: siječanj 2025.).

[4] _____, *Kinase — Wikipedia, The Free Encyclopedia*, https://en.wikipedia.org/w/index.php?title=Kinase&oldid=1271478707, (datum pristupanja: siječanj 2025.).

[5] _____, *Lactate dehydrogenase — Wikipedia, The Free Encyclopedia*, https://en.wikipedia.org/w/index.php?title=Lactate_dehydrogenase&oldid=1228548891, (datum pristupanja: siječanj 2025.).

[6] _____, *Malate dehydrogenase — Wikipedia, The Free Encyclopedia*, https://en.wikipedia.org/w/index.php?title=Malate_dehydrogenase&oldid=1184035067, (datum pristupanja: siječanj 2025.).

[7] _____, *Polyketide synthase — Wikipedia, The Free Encyclopedia*, https://en.wikipedia.org/w/index.php?title=Polyketide_synthase&oldid=1238010067, (datum pristupanja: siječanj 2025.).

[8] _____, *Protein domain — Wikipedia, The Free Encyclopedia*, https://en.wikipedia.org/w/index.php?title=Protein_domain&oldid=1240507033, (datum pristupanja: siječanj 2025.).

[9] _____, *Sequence alignment — Wikipedia, The Free Encyclopedia*, https://en.wikipedia.org/w/index.php?title=Sequence_alignment&oldid=1267781271, (datum pristupanja: siječanj 2025.).

[10] _____, *Serin/treonin-specifična protein-kinaza — Wikipedia, Slobodna enciklopedija*, https://bs.wikipedia.org/wiki/Serin/treonin-specifi%C4%8Dna_protein-kinaza, (datum pristupanja: siječanj 2025.).

[11] _____, *Tirozin-kinaza — Wikipedia, Slobodna enciklopedija*, https://bs.wikipedia.org/wiki/Tirozin-kinaza, (datum pristupanja: siječanj 2025.).

[12] _____, *Translocase — Wikipedia, The Free Encyclopedia*, https://en.wikipedia.org/wiki/Translocase, (datum pristupanja: siječanj 2025.).

[13] W. R. Atchley, J. Zhao, A. D. Fernandes i T. Drüke, *Solving the protein sequence metric problem*, Proceedings of the National Academy of Sciences, National Acad Sciences **102** (2005), br. 18, 6395–6400.

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

[15] C. Barton, T. Flouri, C. S. Iliopoulos i S. P. Pissis, *Global and local sequence alignment with a bounded number of gaps*, Theoretical Computer Science, Elsevier **582** (2015), 1–16.

[16] M. Buljan, *Clustering proteinskih poravnanja*, Diplomski rad, Sveučilište u Zagrebu, Prirodoslovno-matematički fakultet (Matematički odsjek), 2024.

[17] P. Goldstein, J. Zucko, D. Vujaklija, A. Križko, D. Hranueli, P. F. Long, C. Etchebest, B. Basrak i J. Cullum, *Clustering of protein domains for functional and evolutionary studies*, BMC bioinformatics, Springer **10** (2009), 1–11.

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

[19] D. E. Koshland i F. Haurowitz, *"protein", Encyclopedia Britannica*, https://www.britannica.com/science/protein, (datum pristupanja: siječanj 2025.).

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

[21] _____, *"aminokiseline", Hrvatska enciklopedija, mrežno izdanje, Leksikografski zavod Miroslav Krleža, 2013. – 2025.*, https://www.enciklopedija.hr/clanak/bjelancevine, (datum pristupanja: siječanj 2025.).

[22] ———, *"bjelančevine", Hrvatska enciklopedija, mrežno izdanje, Leksikografski zavod Miroslav Krleža, 2013. – 2025.*, https://www.enciklopedija.hr/clanak/bjelancevine, (datum pristupanja: siječanj 2025.).

[23] ———, *Enzyme Nomenclature, International Union of Biochemistry and Molecular Biology*, https://iubmb.qmul.ac.uk/enzyme/, (datum pristupanja: siječanj 2025.).

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

[25] G. Yadav, R. S. Gokhale i D. Mohanty, *Computational approach for prediction of domain organization and substrate specificity of modular polyketide synthases*, Journal of molecular biology, Elsevier **328** (2003), br. 2, 335–363.

[26] H. Zhou, X. Xie i Y. Tang, *Engineering natural products using combinatorial biosynthesis and biocatalysis*, Current Opinion in Biotechnology, Elsevier **19** (2008), br. 6, 590–596.

Sažetak

U ovom diplomskom radu promatrana su poravnanja pet proteinskih familija od kojih je svaka podijeljena u dvije podfamilije. Za svaku familiju, cilj je bio pronaći one pozicije poravnanja koje su najvažnije u klasifikaciji proteina te familije u njene podfamilije, specijalizirane za različite funkcije. Proučavane proteinske familije uključuju: acil transferaze (tj. AT-domene), familiju malatnih i laktatnih dehidrogenaza (MDH/LDH), ciklaze, kinaze, te ketoreduktaze (tj. KR-domene). Statistička analiza provedena nad nizovima iz poravnanja moguća je jer se svakoj aminokiselini (i praznini) u poravnanju pridružio petdimenzionalni numerički vektor. Definirana je razdvajajuća (split) S-statistika koja sumira omjere intergrupne i intragrupne varijabilnosti po svakoj koordinati aminokiselinskog vektora. Podacima se dodao šum dobiven iz poznate prosječne distribucije svih aminokiselina. Po vrijednostima S-statistike rangirane su pozicije za svako od 5 poravnanja, dok je distribucija S-statistike procijenjena nekom F-distribucijom. U većini slučajeva F-distribuiranost S-statistike nismo mogli odbaciti KS testom, pa su izdvojene statistički značajne pozicije za svaku familiju, na razinama značajnosti od 1, 5 ili 10%. Prikazani su i t-SNE grafovi koji vizualiziraju originalne proteine iz poravnanja, koristeći samo 10 najznačajnijih pozicija tog poravnanja. Iz tih ilustrativnih grafova moglo se uočiti da, za svaku familiju, pripadne podfamilije tvore međusobno odvojene klastere, uz jako malo ili nimalo pogrešnih klasifikacija proteina. Konačno, usporedilo se rangiranje pozicija s rangiranjima u nekim sličnim prošlim istraživanjima. Dobivene značajne pozicije u ovom radu potencijalno daju vrijednu informaciju za buduća eksperimentalna biološka istraživanja, posebno u vidu mogućih mutacija enzima baš na tim pozicijama s ciljem postizanja drugačije, preferabilnije funkcije enzima.

Summary

In this thesis, alignments of five protein families were studied, where each family is split into two subfamilies. The goal was to find, for each protein family, the most important alignment positions in terms of separation of certain family into its subfamilies, specialized for different functions. Protein families that were studied include: acyl tranferases (AT-domains), a family of malate and lactate dehydrogenases (MDH/LDH), cyclases, kinases, and ketoreductases (KR-domains). Statistical analysis implemented on sequences of the alignment is possible because each aminoacid (and gap) in the alignment was given a five-dimensional numeric vector. Split statistic (S-statistic) was defined, which sums up ratios of between group variability and within group variability per each coordinate of aminoacid's vector. The noise produced from known random distribution of all aminoacids was added to the data. According to the values of S-statistic, the positions were ranked, for each of the 5 alignments, while the distribution of S-statistic was estimated by some F-distribution. In the majority of cases, the F-distribution of S-statistic could not be rejected with the KS test, so statistically significant positions for each family were selected, at significance levels of 1, 5 or 10%. Also shown are t-SNE graphs that visualize the original proteins from each alignment, solely using their aminoacid residues on the ten most important positions of that alignment. From those illustrative graphs it can be observed that for each family, corresponding subfamilies make up mutually separated clusters, with very few or zero protein misclassifications. Finally, the ranking of positions was compared with rankings in similar past research. The significant positions found in this thesis potentially provide valuable information for future experimental biological research, especially in the form of possible enzyme mutations at those exact positions, with the aim of achieving a different, more preferable enzyme function.

Životopis

Rođen sam u Bihaću, 29. lipnja 2000. godine. Osnovnoškolsko obrazovanje sam završio u Karlovcu, u Osnovnoj školi Dragojle Jarnević. 2015. godine upisujem prirodoslovno-matematički smjer Gimnazije Karlovac. Paralelno s navedenim osnovnoškolskim i srednjoškolskim obrazovanjem, pohađam osnovnu i srednju Glazbenu školu u Karlovcu, nakon čijeg završetka 2018. godine stječem kvalifikaciju glazbenik gitarist.

Gimnaziju Karlovac završavam 2019. godine a iste godine upisujem preddiplomski sveučilišni studij Matematika na Prirodoslovno-matematičkom fakultetu u Zagrebu. Godine 2022. stječem zvanje sveučilišnog prvostupnika matematike (*univ. bacc. math.*) a iste godine upisujem diplomski sveučilišni studij Matematička statistika na istom fakultetu.

U slobodno vrijeme volim svirati gitaru, igrati nogomet, košarku ili stolni tenis, natjecati se na pub kvizovima, i putovati.