Preciznost i klasifikacija

Mirković, Monika

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Monika Mirković

PRECIZNOST I KLASIFIKACIJA

Diplomski rad

Voditelj rada:
doc. dr. sc. Pavle Goldstein

Zagreb, veljača, 2020.

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 se mentoru doc. dr. sc. Pavlu Goldsteinu na posvećenom vremenu, pomoći i savjetima pri izradi ovog diplomskog rada. Posebno hvala mojoj obitelji na podršci koju su mi pružili tijekom studiranja.

Sadržaj

Uvod

Napretkom tehnologije u posljednjih nekoliko desetljeća omogućena je pohrana, obrada i analiza velike količine podataka. Istovremeno, došlo je do velikog napretka u biološkim područjima kao što su molekularna biologija, genetičko inženjerstvo i biotehnologija. Time su se povećali zahtjevi za statističkom obradom i analizom bioloških podataka dobivenih u znanstvenim istraživanjima što je dovelo do razvoja bioinformatike. Bioinformatika je interdisciplinarna znanost koja primjenjuje tehnike iz primijenjene matematike, statistike i računarstva u obradi i analizi bioloških podataka s ciljem boljeg razumijevanja bioloških struktura i procesa.

Jedno od važnijih pitanja u bioinformatici je pitanje pripadnosti proteina nekoj proteinskoj familiji. Proteini su biološke makromolekule građene od dugih lanaca međusobno povezanih aminokiselina. Nalaze se u svim živim bićima i sudjeluju u različitim procesima važnima za život. Proteinske familije su skupine evolucijski povezanih proteina. Sličnost kraćih nizova aminokiselina jedan je od najjasnijih pokazatelja pripadnosti istoj proteinskoj familiji. Budući da se evolucijom genetski materijal živih bića mijenja i dolazi do promjena u proteinima, u ovom diplomskom radu promatramo kraće nizove aminokiselina s karakterističnim mutacijama koje nazivamo motivi. Opisujemo i testiramo algoritam za traženje karakterističnih motiva neke proteinske familije na temelju kojih se može provesti klasifikacija novih, dosad neopisanih proteina.

Ovaj diplomski rad podijeljen je u četiri poglavlja. U prvom poglavlju navodimo i objašnjavamo pojmove iz linearne algebre, vjerojatnosti i statistike potrebne za razumijevanje daljnjeg rada. U drugom poglavlju detaljnije se objašnjava problem klasifikacije proteina u proteinske familije. Opisuje se algoritam za traženje karakterističnih motiva neke proteinske familije. U trećem poglavlju objašnjavamo princip ocjenjivanja nekog proteina u odnosu na dobivene karakteristične motive i način pridruživanja proteina nekoj od proteinskih familija od interesa. Nadalje, opisujemo način podjele podataka kod problema klasifikacije. U posljednjem poglavlju objašnjeni algoritam traženja karakterističnih motiva primjenjujemo na dvije proteinske familije, lipoproteinske lipaze i Walkerove motive. Na kraju navodimo i analiziramo dobivene rezultate. Programski jezici korišteni pri izradi ovog diplomskog rada su Python, R i C.

Poglavlje 1

Matematička podloga

U ovom poglavlju definirat ćemo pojmove i matematičke strukture iz linearne algebre, vjerojatnosti i statistike koji će nam biti potrebni za razumijevanje daljnjeg rada. Pojmovi iz linearne algebre su velikim dijelom preuzeti iz [1]. Za definiranje pojmova iz vjerojatnosti i statistike korištene su bilješke s predavanja iz kolegija Statistika, kao i izvori [12], [14], [7] i [4].

1.1 Linearna algebra

Definicija 1.1.1. *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. $a + (b + c) = (a + b) + c,\ \forall a, b, c \in V$;
2. $\exists\, 0 \in V$ *sa svojstvom* $a + 0 = 0 + a = a,\ \forall a \in V$;
3. $\forall a \in V, \exists - a \in V$ *tako da je* $a + (-a) = (-a) + a = 0$;
4. $a + b = b + a,\ \forall a, b \in V$;
5. $\alpha(\beta a) = (\alpha\beta)a, \forall \alpha, \beta \in \mathbb{F},\ \forall a \in V$;
6. $(\alpha + \beta)a = \alpha a + \beta a, \forall \alpha, \beta \in \mathbb{F},\ \forall a \in V$;
7. $\alpha(a + b) = \alpha a + \alpha b,\ \forall \alpha \in \mathbb{F}, \forall a, b \in V$;
8. $1 \cdot a = a,\ \forall a \in V$.

Vektori *su elementi vektorskog prostora.*

Definicija 1.1.2. *Neka je V vektorski prostor nad poljem $\mathbb{F}$ i $S \subseteq V$, $S \neq \emptyset$.* ***Linearna ljuska skupa*** *S označava se simbolom $[S]$ i definira kao*

$$[S] = \left\{ \sum_{i=1}^{k} \alpha_i a_i : \alpha_i \in \mathbb{F}, a_i \in S, k \in \mathbb{N} \right\}.$$

Dodatno, definira se $[\emptyset] = \{0\}$.

Definicija 1.1.3. *Neka je V vektorski prostor i $S \subseteq V$. Kaže se da je S* ***sustav izvodnica za*** *V (ili da S generira V) ako vrijedi $[S] = V$.*

Definicija 1.1.4. *Konačan skup $B = \{b_1, b_2, ..., b_n\}$, $n \in \mathbb{N}$, u vektorskom prostoru V se naziva* ***baza za*** *V ako je B linearno nezavisan sustav izvodnica za V.*

Definicija 1.1.5. *Neka je V vektorski prostor nad poljem $\mathbb{F}$ i neka je $M \subseteq V$, $M \neq \emptyset$. Ako je i $(M, +, \cdot)$ vektorski prostor nad poljem $\mathbb{F}$ uz iste operacije iz V, kažemo da je M* ***potprostor od*** *V.*

Definicija 1.1.6. *Neka su V i W vektorski prostori nad istim poljem $\mathbb{F}$. Preslikavanje $A : V \to W$ zove se* ***linearan operator*** *ako vrijedi*

$$A(\alpha x + \beta y) = \alpha Ax + \beta Ay, \qquad \forall x, y \in V, \quad \forall \alpha, \beta \in \mathbb{F}.$$

Napomena 1.1.7. *Vektorski prostor linearnih operatora s V u W označava se s $L(V, W)$. Kada je $V = W$ umjesto $L(V, V)$ piše se $L(V)$.*

Definicija 1.1.8. *Neka je V vektorski prostor nad poljem $\mathbb{F}$.* ***Skalarni produkt na*** *V je preslikavanje $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{F}$ koje ima sljedeća svojstva:*

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

Definicija 1.1.9. *Vektorski prostor na kojem je definiran skalarni produkt zove se* ***unitaran prostor****.*

Primjer 1.1.10. *U $\mathbb{R}^n$ skalarni produkt definiran je s*

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

Definicija 1.1.11. *Neka je V konačnodimenzionalan unitaran prostor i $A \in L(V)$. Operator $A^* \in L(V)$ sa svojstvom $\langle Ax, y\rangle = \langle x, A^*y\rangle$, $\forall x, y \in V$ zove se* ***hermitski adjungiran operator operatoru*** *A.*

Propozicija 1.1.12. *Neka je V konačnodimenzionalan unitaran prostor i $P \in L(V)$. Operator P je* ***ortogonalni projektor*** *ako i samo ako vrijedi $P^2 = P = P^*$.*

Definicija 1.1.13. *Neka je V unitaran prostor.* ***Norma na*** *V je funkcija $\|\cdot\| : V \to \mathbb{R}$ definirana s $\|x\| = \sqrt{\langle x, x\rangle}$.*

Propozicija 1.1.14. *Norma na unitarnom prostoru V ima sljedeća* ***svojstva****:*

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

Napomena 1.1.15. *Svaka funkcija $\|\cdot\| : V \to \mathbb{R}$ na vektorskom prostoru V sa svojstvima iz 1.1.14 naziva se* ***norma****. Tada $(V, \|\cdot\|)$ zovemo* ***normirani prostor****.*

Navedimo **primjere normi u** $\mathbb{R}^n$ koje ćemo koristiti.

Primjer 1.1.16.
1-norma *je funkcija $\|\cdot\|_1 : \mathbb{R}^n \to \mathbb{R}$ definirana s*

$$\|x\|_1 = \sum_{i=1}^{n} |x_i|. \tag{1.1}$$

2-norma *ili* ***Euklidska norma*** *je funkcija $\|\cdot\|_2 : \mathbb{R}^n \to \mathbb{R}$ definirana s*

$$\|x\|_2 = \sqrt{\langle x, x\rangle} = \sqrt{\sum_{i=1}^{n} x_i{}^2}. \tag{1.2}$$

max-norma *ili* ***∞-norma*** *je funkcija $\|\cdot\|_\infty : \mathbb{R}^n \to \mathbb{R}$ definirana s*

$$\|x\|_\infty = \max\Big\{|x_1|, |x_2|, ..., |x_n|\Big\}. \tag{1.3}$$

1.2 Vjerojatnost i statistika

Definicija 1.2.1. *Neka je Ω proizvoljan neprazan skup, $\mathcal{F}$ σ-algebra na skupu Ω. Uređeni par $(\Omega, \mathcal{F})$ zove se **izmjeriv prostor**. Funkcija $\mathbb{P}\colon \mathcal{F} \to \mathbb{R}$ je **vjerojatnost** na $\mathcal{F}$ ako vrijedi:*

(i) $\mathbb{P}(A) \geq 0$, $\forall A \in \mathcal{F}$

(ii) $\mathbb{P}(\Omega) = 1$

(iii) $A_i \in \mathcal{F}$, $i \in \mathbb{N}$ $\;i\;$ $A_i \cap A_j = \emptyset$ $\;za\; i \neq j \Rightarrow \mathbb{P}(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$

Napomena 1.2.2. *1. Uređena trojka $(\Omega, \mathcal{F}, \mathbb{P})$, gdje je $\mathcal{F}$ σ-algebra na Ω i $\mathbb{P}$ vjerojatnost na $\mathcal{F}$, zove se **vjerojatnosni prostor**.*

*2. Neka je $(\Omega, \mathcal{F}, \mathbb{P})$ vjerojatnosni prostor. Elemente σ-algebre $\mathcal{F}$ zovemo **događaji**, a broj $\mathbb{P}(A)$, $A \in \mathcal{F}$ zove se **vjerojatnost događaja A**.*

Definicija 1.2.3. *Neka je $(\Omega, \mathcal{F}, \mathbb{P})$ vjerojatnosni prostor. Funkcija $X\colon \Omega \to \mathbb{R}$ je **slučajna varijabla** (na Ω ili na $\mathcal{F}$) ako je $X^{-1}(B) \in \mathcal{F}$ za proizvoljno $B \in \mathcal{B}$, odnosno $X^{-1}(\mathcal{B}) \subseteq \mathcal{F}$.*

Napomena 1.2.4. *1. $\mathcal{B}$ iz prethodne definicije je **Borelova σ-algebra skupova na** $\mathbb{R}$. Elemente od $\mathcal{B}$ zovemo **Borelovi skupovi**.*

2. Kada se bavimo problemima vezanima za određenu slučajnu varijablu X pogodnije je operirati s vjerojatnosnim prostorom induciranim s X. Za $B \in \mathcal{B}$ stavimo:

$$\mathbb{P}_X(B) = \mathbb{P}(X^{-1}(B)) = \mathbb{P}\{\omega \in \Omega\colon X(\omega) \in B\}$$

*Time je definirana funkcija $\mathbb{P}_X\colon \mathcal{B} \to [0, 1]$ koja je vjerojatnosna mjera na $\mathcal{B}$ i zovemo ju **vjerojatnost inducirana slučajnom varijablom** X. Svakoj slučajnoj varijabli X je na prirodan način pridružen vjerojatnosni prostor $(\mathbb{R}, \mathcal{B}, \mathbb{P}_X)$ induciran slučajnom varijablom X.*

Definicija 1.2.5. ***Funkcija distribucije slučajne varijable** X je funkcija $F_X = F\colon \mathbb{R} \to [0, 1]$ definirana s:*

$$F(x) = \mathbb{P}_X(< -\infty, x]) = \mathbb{P}(X \leq x) = \mathbb{P}\{\omega \in \Omega\colon X(\omega) \leq x\}, \quad x \in \mathbb{R}.$$

Postoje dvije glavne vrste slučajnih varijabli: diskretne i neprekidne.

Definicija 1.2.6. *Slučajna varijabla X je **diskretna** ako postoji konačan ili prebrojiv skup $D \subset \mathbb{R}$ takav da je $\mathbb{P}(X \in D) = 1$.*

Napomena 1.2.7. *Diskretna slučajna varijabla X najčešće se označava s:*

$$X = \begin{pmatrix} x_1 & x_2 & \dots & x_n & \dots \\ p_1 & p_2 & \dots & p_n & \dots \end{pmatrix} \tag{1.4}$$

U prvom retku tablice 1.4 stoje sve moguće različite vrijednosti slučajne varijable X, dakle međusobno različiti realni brojevi, a u drugom retku su pripadne vjerojatnosti da X poprimi te vrijednosti, dakle nenegativni realni brojevi sa sumom 1. Prema tome, svakoj slučajnoj varijabli na diskretnom vjerojatnosnom prostoru se na jednoznačan način pridružuje tablica 1.4 koju zovemo ***distribucija slučajne varijable*** *X ili* ***zakon razdiobe od*** *X. Obratno, ako je zadana tablica 1.4 takva da je* $x_i \neq x_j$ *za* $i \neq j$, $p_i \geq 0$, $\sum_i p_i = 1$, *tada postoji diskretan vjerojatnosni prostor i slučajna varijabla X na tom vjerojatnosnom prostoru tako da je 1.4 distribucija od X.*

Definicija 1.2.8. *Kažemo da je slučajna varijabla X* ***apsolutno neprekidna*** *ili, kraće,* ***neprekidna slučajna varijabla*** *ako postoji nenegativna Borelova funkcija f na* $\mathbb{R}$ *takva da je*

$$F_X(x) = \int_{-\infty}^{x} f(t)d\lambda(t), \quad x \in \mathbb{R}. \tag{1.5}$$

Za funkciju distribucije F_X *slučajne varijable X kažemo da je* ***apsolutno neprekidna funkcija distribucije****. U tom slučaju se funkcija f iz 1.5 naziva* ***funkcija gustoće vjerojatnosti od*** *X.*

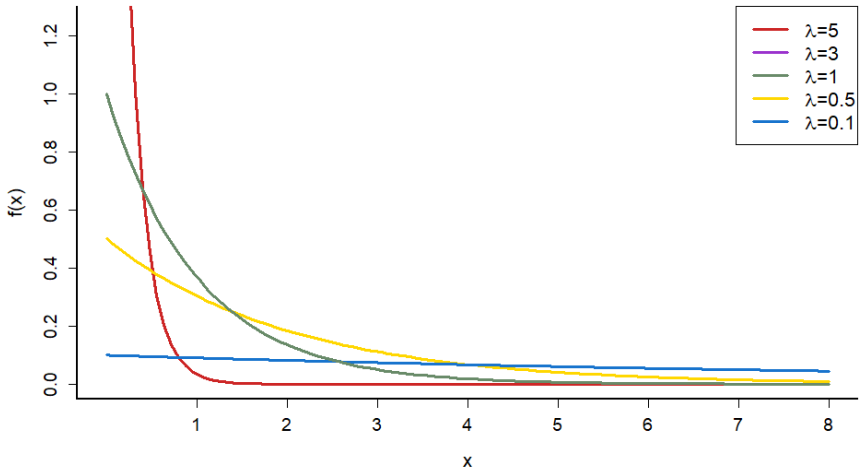
Navedimo neke **primjere slučajnih varijabli**.

Primjer 1.2.9. *Neprekidna slučajna varijabla X ima* ***eksponencijalnu distribuciju*** *s parametrom* $\lambda > 0$ *ako joj je funkcija gustoće dana s:*

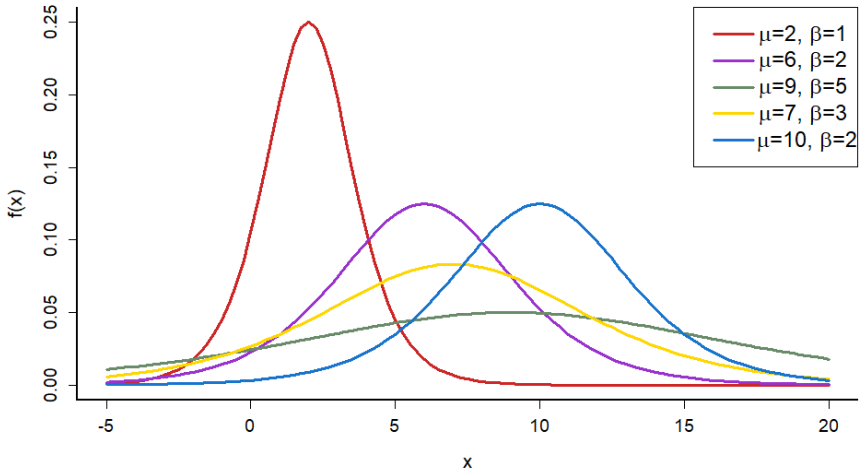
$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x > 0 \\ 0, & x \leq 0. \end{cases}$$

Primjer 1.2.10. *Neka je* $\mu \in \mathbb{R}$ *i* $\beta > 0$*. Neprekidna slučajna varijabla X ima* ***logističku distribuciju*** *s parametrima* μ *i* β *ako joj je funkcija gustoće dana s:*

$$f(x) = \frac{e^{-\frac{x-\mu}{\beta}}}{\beta\left(1 + e^{-\frac{x-\mu}{\beta}}\right)^2}, \quad x \in \mathbb{R}.$$



Slika 1.1: Graf funkcije gustoće eksponencijalne distribucije za različite vrijednosti λ



Slika 1.2: Graf funkcije gustoće logističke distribucije za različite vrijednosti μ i β

Primjer 1.2.11. *Neka su $p, q > 0$. Slučajna varijabla X ima* ***generaliziranu logističku distribuciju*** *ako joj je funkcija gustoće dana s:*

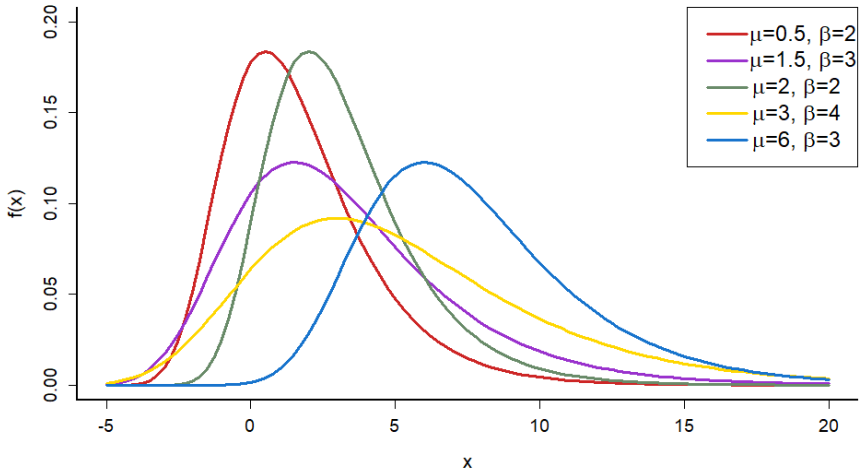
$$f(x) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} \frac{e^{-qx}}{(1+e^{-x})^{p+q}}, \quad x \in \mathbb{R},$$

pri čemu je Γ gama funkcija definirana formulom $\Gamma(x) = \int_0^\infty e^{-t} t^{x-1} dt$, $x > 0$.

Teorija ekstremnih vrijednosti

Definicija 1.2.12. *Neka je $\mu \in \mathbb{R}$ i $\beta > 0$. Neprekidna slučajna varijabla X ima* ***Gumbelovu distribuciju*** *s parametrima μ i β ako joj je funkcija gustoće dana s:*

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



Slika 1.3: Graf funkcije gustoće Gumbelove distribucije za različite vrijednosti μ i β

Definicija 1.2.13. *Neka je $p > 0$. Slučajna varijabla X ima* ***generaliziranu Gumbelovu distribuciju*** *ako joj je funkcija gustoće dana s:*

$$f(x) = \frac{1}{\Gamma(p)} e^{-px} e^{e^{-px}}, \quad x \in \mathbb{R}.$$

Korolar 1.2.14. *Neka su X_1 i X_2 nezavisne generalizirane Gumbel distribuirane slučajne varijable s parametrima p i q, respektivno. Tada slučajna varijabla $Y = X_1 - X_2$ ima generaliziranu logističku distribuciju s parametrima p i q.*

Teorem 1.2.15. ***(Fisher–Tippett–Gnedenko)***
Neka je $(X_n, n \in \mathbb{N})$ niz nezavisnih i jednako distribuiranih slučajnih varijabli i neka je $M_n = \max\{X_1, X_2, ..., X_n\}$, za $n \in \mathbb{N}$. Ako postoje realni nizovi (a_n) i (b_n) takvi da je $a_n > 0$, za svaki $n \in \mathbb{N}$ i $\lim_{n\to\infty} \mathbb{P}\left(\frac{M_n - b_n}{a_n} \leq x\right) = F(x)$, gdje je F nedegenerirana distribucija, tada granična distribucija F pripada Gumbelovoj, Fréchetovoj ili Weibullovoj distribuciji.

Napomena 1.2.16. *Gumbelova, Fréchetova i Weibullova distribucija su distribucije ekstremnih vrijednosti.*

Teorija informacija

Definicija 1.2.17. *Neka je X diskretna slučajna varijabla na vjerojatnosnom prostoru $(\Omega, \mathcal{F}, \mathbb{P})$, gdje je $\Omega = \{x_1, x_2, ..., x_n\}$. Neka je $P = [p_1, p_2, ..., p_n]$ distribucija slučajne varijable X, gdje je $\mathbb{P}(x_i) = p_i$, za svaki $i = 1, 2, ..., n$.*
Entropija slučajne varijable *X označava se s $H(X)$ ili $H(P)$ i definira formulom:*

$$H(P) = -\sum_{i=1}^{n} p_i \log_b(p_i) \tag{1.6}$$

gdje je b baza logaritma.
Ukoliko je $p_i = 0$ za neki $i \in \{1, 2, ..., n\}$ koristi se konvencija $0 \log_b 0 = 0$.

Neka **svojstva entropije** H su:

1. H je uvijek nenegativna.
2. H poprima minimum u 0 u slučaju kada je distribucija P koncentrirana u jednoj točki, tj. $p_i = 1$, za neki $i \in \{1, 2, ..., n\}$.
3. H poprima maksimum u $\log_b n$ u slučaju kada je distribucija P uniformna, tj, $p_i = \frac{1}{n}$, za $i = 1, 2, ..., n$.

Napomena 1.2.18. *1. Intuitivno, entropija diskretne slučajne varijable X je mjera nesigurnosti (neodređenosti) ishoda slučajne varijable X. Što je entropija manja, to je nesigurnost ishoda slučajne varijable X manja.*

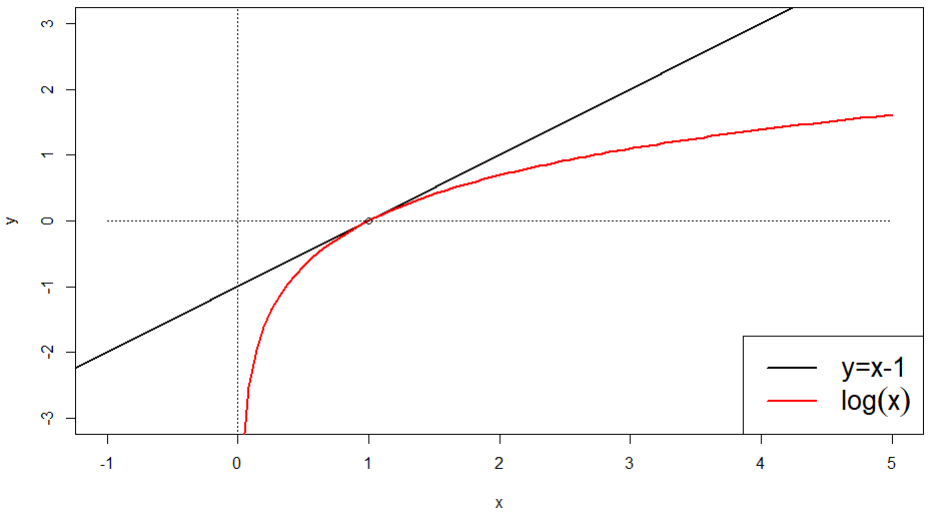
2. *Možemo uočiti da definicija dozvoljava proizvoljnu bazu logaritma. U ovom radu koristit ćemo* $b = 20$ *jer postoji ukupno* 20 *standardnih aminokiselina (*$n = 20$*) pa će se vrijednosti entropije* H *nalaziti u* $[0, 1]$.

Definicija 1.2.19. *Neka su* X *i* Y *dvije diskretne slučajne varijable definirane na istom vjerojatnosnom prostoru* $(\Omega, \mathcal{F}, \mathbb{P})$*, gdje je* $\Omega = \{x_1, x_2, ..., x_n\}$*. Neka je* $P = [p_1, p_2, ..., p_n]$ *distribucija od* X *i* $Q = [q_1, q_2, ..., q_n]$ *distribucija od* Y*. Vrijedi* $\mathbb{P}(X = x_i) = p_i$*, za* $i = 1, 2, ..., n$ *i* $\mathbb{P}(Y = x_i) = q_i$*, za* $i = 1, 2, ..., n$
Relativna entropija od P ***u odnosu na*** Q *označava se s* $H(P \parallel Q)$ *i definira formulom:*

$$H(P \parallel Q) = \sum_{i=1}^{n} p_i \log\left(\frac{p_i}{q_i}\right) \tag{1.7}$$

Koristi se konvencija $0 \log\left(\frac{0}{q}\right) = 0$ *i* $p \log\left(\frac{p}{0}\right) = +\infty$ *za* $p, q \in \mathbb{R}$.

Propozicija 1.2.20. *Relativna entropija* H *je uvijek nenegativna.*



Slika 1.4: Grafovi funkcija $y = x - 1$ i $y = \log(x)$

Dokaz. Iz slike 1.4 vidimo da vrijedi: $\log(x) \leq x - 1$, $\forall x \in \mathbb{R}$. Jednakost se postiže samo za $x = 1$. Iz toga slijedi:

$$-H(P \parallel Q) = \sum_{i=1}^{n} p_i \log\left(\frac{q_i}{p_i}\right) \leq \sum_{i=1}^{n} p_i \left(\frac{q_i}{p_i} - 1\right) = \sum_{i=1}^{n} q_i - \sum_{i=1}^{n} p_i = 1 - 1 = 0 \tag{1.8}$$

Množenjem 1.8 s -1 dobijemo traženu tvrdnju: $H(P \parallel Q) \geq 0$, $\forall\, P, Q$. □

Neka **svojstva relativne entropije** H su:

1. H je uvijek nenegativna (vidi propoziciju 1.2.20).
2. H poprima minimum u 0 u slučaju kada su distribucije P i Q identične. Tada vrijedi $H(P \parallel Q) = H(Q \parallel P) = 0$.
3. Ne postoji maksimum od H.
4. Općenito, H nije simetrična, tj. $H(P \parallel Q) \neq H(Q \parallel P)$ pa H nije ni metrika.

Napomena 1.2.21. *1. Intuitivno, relativna entropija je mjera sličnosti dviju diskretnih distribucija (jednakih duljina). Što je relativna entropija manja, dvije distribucije su sličnije.*

2. Relativna entropija se još naziva i Kullback-Leiblerova divergencija.

3. Entropija i relativna entropija mogu se definirati i za neprekidne slučajne varijable pomoću integrala.

Kvantili

Definicija 1.2.22. *Podatke $x_1, x_2, ..., x_n$ poredane po veličini označavamo s $x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}$. Pri tome za $s = k + r$, $k \in \mathbb{Z}$, $r \in [0,\ 1\rangle$ vrijedi formula:*

$$x_s = x_k + r \cdot (x_{k+1} - x_k)$$

***Medijan** uzorka je $x_{\frac{n+1}{2}}$, a 75% **kvantil** je $x_{\frac{75(n+1)}{100}}$.*

Dakle, za brojčani niz podataka $x_1, x_2, ..., x_n$ medijan M je realan broj sa svojstvom da je pola podataka manje ili jednako od M. Također, 75% kvantil je realan broj $q_{0.75}$ sa svojstvom da je udio podataka manjih ili jednakih $q_{0.75}$ približno 75%.

Osjetljivost i specifičnost testa

U ovom odjeljku definirat ćemo pojmove koje ćemo koristiti kasnije kod analize uspješnosti algoritma za traženje karakterističnih motiva. Osjetljivost i specifičnost testa su statističke mjere pomoću kojih mjerimo uspješnost provedenog testa.
Osjetljivost testa, stopa stvarno pozitivnih, mjeri proporciju pozitivnih elemenata uzorka ispravno prepoznatih testom u odnosu na ukupni broj pozitivnih elemenata.

$$\begin{aligned}\text{osjetljivost} &= \frac{\text{broj stvarno pozitivnih}}{\text{broj stvarno pozitivnih + broj lažno negativnih}} \\ &= \frac{\text{TP}}{\text{CP}} = \frac{\text{TP}}{\text{TP + FN}}\end{aligned}$$

Specifičnost testa, stopa stvarno negativnih, mjeri proporciju negativnih elemenata uzorka ispravno prepoznatih testom u odnosu na ukupni broj negativnih elemenata.

$$\begin{aligned}\text{specifičnost} &= \frac{\text{broj stvarno negativnih}}{\text{broj stvarno negativnih + broj lažno pozitivnih}} \\ &= \frac{\text{TN}}{\text{CN}} = \frac{\text{TN}}{\text{TN + FP}}\end{aligned}$$

Kod ocjenjivanja uspješnosti testa također je korisno definirati pozitivnu prediktivnu vrijednost (PPV) i negativnu prediktivnu vrijednost (NPV).
Pozitivna prediktivna vrijednost (PPV) mjeri u kojem postotku pozitivno identificirani elementi zaista jesu stvarno pozitivni.

$$\begin{aligned}\text{PPV} &= \frac{\text{broj stvarno pozitivnih}}{\text{broj stvarno pozitivnih + broj lažno pozitivnih}} \\ &= \frac{\text{TP}}{\text{P}} = \frac{\text{TP}}{\text{TP + FP}}\end{aligned}$$

Negativna prediktivna vrijednost (NPV) mjeri u kojem postotku negativno identificirani elementi zaista jesu stvarno negativni.

$$\begin{aligned}\text{NPV} &= \frac{\text{broj stvarno negativnih}}{\text{broj stvarno negativnih + broj lažno negativnih}} \\ &= \frac{\text{TN}}{\text{N}} = \frac{\text{TN}}{\text{TN + FN}}\end{aligned}$$

Rezultati testa često se prikazuju tablicom 1.1 u kojoj su zapisane ranije navedene veličine uspješnosti testa. Takva tablica naziva se još i **matrica konfuzije**.

		predviđeno stanje		
		pozitivno stanje (P)	**negativno stanje (N)**	
stvarno stanje	**pozitivno stanje (CP)**	TP (stvarno pozitivni)	FN (lažno negativni)	osjetljivost
	negativno stanje (CN)	FP (lažno pozitivni)	TN (stvarno negativni)	specifičnost
		PPV (pozitivna prediktivna vrijednost)	NPV (negativna prediktivna vrijednost)	

Tablica 1.1: Matrica konfuzije

Poglavlje 2

Problem klasifikacije proteina

2.1 Biološki koncepti

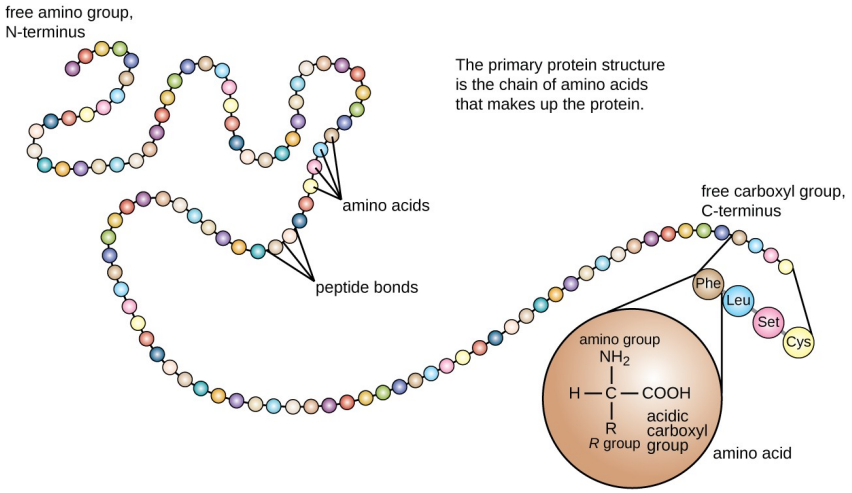
Proteini ili bjelančevine osnovni su sastavni dijelovi svake stanice što ih čini osnovom života na Zemlji. Proteom je skup svih proteina određenog organizma ili stanice koji nastaju kao posljedica ekspresije gena u određenom trenutku. Proteini, ovisno o svojoj građi, unutar organizma obavljaju ili pomažu različite procese važne za život kao što je proces rasta, razvoja i regeneracije tjelesnih stanica. Dvadeset različitih standardnih aminokiselina međusobno se povezuju peptidnom vezom u duge lance, a lanci u biološke makromolekule koje grade protein. Način povezivanja i struktura aminokiselina može se vidjeti na slici 2.1 preuzetoj iz izvora [9].

Proteini, kao i RNA i DNA, spadaju u biološke nizove. Biološki nizovi su nizovi bez separatora u odgovarajućem biološkom alfabetu. RNA i DNA su nizovi nukleotida, dok su proteini nizovi aminokiselina. Proteini se prema zajedničkom porijeklu grupiraju u proteinske familije. Proteini iz iste proteinske familije imaju slične funkcije, sličnu trodimenzionalnu strukturu i slične nizove aminokiselina. Sličnost nizova aminokiselina jedan je od najjasnijih pokazatelja zajedničkog porijekla, tj. pripadnosti istoj proteinskoj familiji.

Genom je genetski materijal organizma. Prema [13] sinteza proteina započinje u jezgri gdje se genska uputa prenosi s kodirajuće DNA na RNA procesom transkripcije. Unutar ribosoma koji su izgrađeni od RNA odvija se proces translacije, tj. prijepis nukleotida u aminokiseline. Kodon je triplet nukleotida koji kodira jednu aminokiselinu prema skupu pravila koje nazivamo genetski kod. Prikaz sinteze proteina može se vidjeti na slici 2.5 preuzetoj iz izvora [5]. Genetski materijal živih bića se kroz generacije mijenja, odnosno dolazi do mutacija gena, iz čega dolazi do mutacija u proteinu. Osnovni mutacijski procesi su zamjena ili supstitucija (engl. *supstitution*), umetanje ili adicija (engl. *insertion*) i brisanje ili delecija (engl. *deletion*). Mi ćemo se baviti samo zamjenom ili supstitucijom.

U tablici 2.2 su dane standardne aminokiseline i njihove oznake, a u tablicama 2.3 i

2.4 nukleotidne baze RNA i DNA i njihove oznake. U tablici 2.6 prikazan je genetski kod iz kojeg možemo vidjeti da kodon AUG (aminokiselina M) označava početak translacije, dok kodoni UAA, UGA, UAG označavaju kraj translacije. Također, možemo uočiti da jedan kodon kodira najviše jednu aminokiselinu, dok jednu aminokiselinu može kodirati više različitih kodona. Iz tog razloga je moguće da neka mutacija promijeni kodon, ali on i dalje kodira istu aminokiselinu. Takve mutacije nazivamo neutralnim mutacijama.



Slika 2.1: Način povezivanja i struktura 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

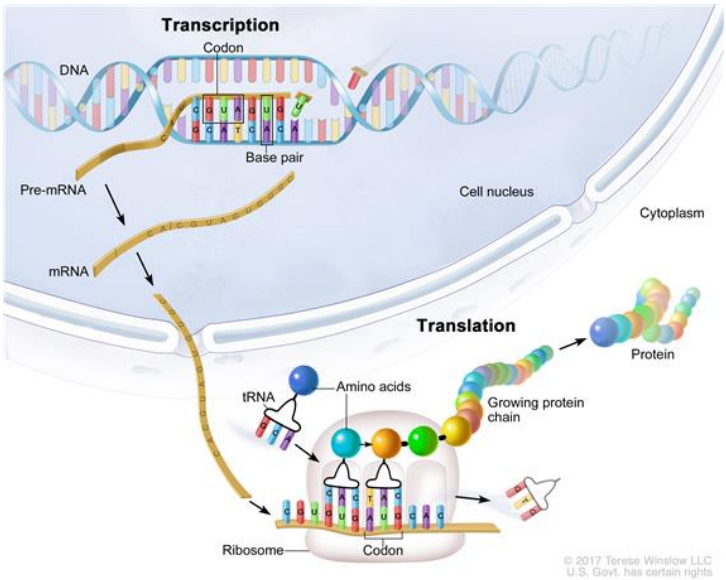
Slika 2.2: Tablica standardnih aminokiselina

Oznaka	Naziv
A	Adenin
G	Gvanin
C	Citozin
U	Uracil

Slika 2.3: Tablica nukleotidnih baza RNA

Oznaka	Naziv
A	Adenin
G	Gvanin
C	Citozin
T	Timin

Slika 2.4: Tablica nukleotidnih baza DNA



Slika 2.5: Sinteza proteina

Amino.	Kodoni
A	GCU, GCC, GCA, GCG
C	UGU, UGC
D	GAU, GAC
E	GAA, GAG
F	UUU, UUC
G	GGU, GGC, GGA, GGG
H	CAU, CAC
I	AUU, AUC, AUA
K	AAA, AAG
L	UUA, UUG, CUU, CUC, CUA, CUG
M	AUG
N	AAU, AAC
P	CCU, CCC, CCA, CCG
Q	CAA, CAG
R	CGU, CGC, CGA, CGG, AGA, AGG
S	UCU, UCC, UCA, UCG, AGU, AGC
T	ACU, ACC, ACA, ACG
V	GUU, GUC, GUA, GUG
W	UGG
Y	UAU, UAC
START	AUG
STOP	UAA, UGA, UAG

Slika 2.6: Genetski kod

2.2 Klasifikacija proteina kao klasifikacija teksta

Klasifikacija dokumenata je problem koji se javlja u mnogim područjima ljudskoga života. Dokumenti koje klasificiramo mogu biti različitog tipa, kao na primjer tekstovi, slike i zvukovi. Glavni zadatak klasifikacije je pridružiti dokument jednoj ili više unaprijed definiranih klasa. Klasom smatramo skup objekata koji imaju iste ili slične karakteristike. U ovom radu bavimo se klasifikacijom proteina u proteinske familije koje nam predstavljaju klase. Klasifikaciju provodimo na osnovi proteinskog niza, tj. niza aminokiselina.

Sekvenciranje u genetici je metoda kojom je moguće utvrditi redoslijed pojedinih gra-

divnih elemenata u makromolekuli. Kod DNA utvrđuje se redoslijed nukleotida. Moderne tehnologije sekvenciranja sljedeće generacije (engl. *next generation sequencing*, NGS) imaju velik potencijal biološke, medicinske, pa čak i kliničke primjene. Sekvenciranjem ljudskog genoma modernim tehnologijama sekvenciranja sljedeće generacije dobijemo velik broj nukleotidnih nizova koji se zatim prema određenim pravilima i raznim tehnikama pretrage i klasifikacije prepisuju u nizove aminokiselina (proteine) i pridružuju proteinskim familijama. Iz tog razloga je od velike važnosti unaprijediti postojeće i razviti nove tehnike pretrage i klasifikacije bioloških nizova. Prema [3] bolje razumijevanje našeg jedinstvenog genetskog materijala u budućnosti bi omogućilo da terapija bude prilagođena potrebama pojedinca. Kolika je važnost toga možemo vidjeti na primjeru razvoja novih lijekova za liječenje raka koji su specifično usmjereni na mutacije koje se mogu identificirati sekvenciranjem genoma karcinoma kod pojedinih pacijenata. Također, sekvenciranje genoma se već počinje značajno koristiti u dijagnostici rizika od nasljednih bolesti jer identificiranjem specifičnih mutacija gena za određenu bolest, kao što je rak dojke kod žena, kod zdravih ljudi mogle bi se poduzeti odgovarajuće mjere za prevenciju bolesti.

Budući da proteine promatramo kao nizove aminokiselina označene slovima engleskog alfabeta, problem klasifikacije proteina možemo smjestiti u problem klasifikacije tekstova. Jedna od mogućih tehnika klasifikacije tekstova je modifikacija tehnike semantičkog indeksiranja. Semantičko indeksiranje temelji se na traženju karakterističnih pojmova (riječi) u dokumentima. Općenito, kod semantičkog indeksiranja određuje se matrica frekvencija karakterističnih pojmova u dokumentima (engl. *document term matrix*) na temelju koje se provodi klasifikacija. Više o tome u izvoru [11]. U našem slučaju klasifikacije proteina u proteinske familije jedan protein, tj. jedan niz aminokiselina, nam predstavlja dokument, a kraći nizovi aminokiselina duljine n koje nazivamo n-grami nam predstavljaju pojmove. Umjesto standardne matrice frekvencija koristit ćemo generaliziranu matricu frekvencija. U daljnjem radu objasnit ćemo na koji način dolazimo do generaliziranih frekvencija.

Za svaki protein iz neke proteinske familije možemo pronaći n-grame koji su bolje očuvani i specifični za njega pa vjerojatno imaju biološki značaj. Budući da evolucijom dolazi do mutacija u proteinima umjesto standardnih n-grama fiksne duljine tražit ćemo tzv. motive. Motiv je n-gram s karakterističnim mutacijama (varijacijama). Preciznije, motivom smatramo niz aminokiselina duljine n koji ima karakterističnu supstituciju, odnosno specifične zakone mutacija koji čuvaju određenu funkciju. Upit je n-gram iz nekog proteina iz proteinske familije od interesa.

Motiv ćemo određivati na način da uzmemo upit koji nam predstavlja ulazni motiv i "prolazimo" kroz sve proteine iz iste proteinske familije. Kada nađemo n-gram koji je "dovoljno sličan" zadanom ulaznom motivu, odnosno čija je sličnost značajna, dodat ćemo novi n-gram u popis varijanti ulaznog motiva. Taj postupak ćemo ponovljati za novodobivene motive sve dok pronalazimo značajne n-grame ili ne dostignemo zadano "ograničenje". Zatim ćemo određenim metodama "filtrirati" i "produžiti" motive tako da

nisu svi iste duljine n. Kod nas je n jednak 10.

Cilj ovog rada je generirati pojmove, odnosno motive karakteristične za neku proteinsku familiju i provesti klasifikaciju koja će nam biti pokazatelj točnosti našeg algoritma za traženje karakterističnih motiva. Ovaj dio rada usko je vezan uz diplomske radove [2], [7] i [14]. Klasifikacija na temelju n-grama fiksne duljine i njihovih frekvencija unutar proteinskih familija od interesa može se pronaći u diplomskim radovima [6] i [8]. Rezultate i grafove u pojedinim koracima algoritma prikazujemo za proteinske familije lipoproteinske lipaze i Walkerove motive opisane u poglavlju 4, ali algoritam ne ovisi o izboru i broju proteinskih familija koje klasificiramo.

2.3 Traženje motiva

Profil motiva

Profil motiva ćemo definirati matricom PSSM (engl. *position specific scoring matrix*) koja se koristi za prikaz razdiobe aminokiselina na svakoj poziciji motiva. Neka je $\mathcal{A} = [A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V]$ vektor standardnih aminokiselina. Distribuciju motiva ili osnovnu PSSM matricu relativnih frekvencija za zadani motiv sastavljen od m n-grama dobijemo tako da za svaku od n pozicija motiva, tj. za svaki "stupac motiva", izračunamo relativnu frekvenciju pojavljivanja svake od 20 standardnih aminokiselina iz $\mathcal{A}$. Time dobijemo niz vektora distribucije $f_i = (f_{i1}, f_{i2}, ..., f_{i20})$, $i = 1, 2, ..., n$. Matrica $F = [f_{ij}]$, $i = 1, 2, ..., n$, $j = 1, 2, ..., 20$ predstavlja distribuciju zadanog motiva. Duljina distribucije motiva je broj redaka matrice, tj. n.
Problem malog uzorka riješavamo tako da relativnim frekvencijama dodamo mali pseudozbroj, kao na primjer 0.01. Time dobijemo da je vjerojatnost pojave svake aminokiseline iz $\mathcal{A}$ na svakoj poziciji motiva veća od 0. Označimo nove vektore distribucija s $g_i = (g_{i1}, g_{i2}, ..., g_{i20})$, i = 1,2, ..., n, gdje je

$$g_{ij} = \frac{f_{ij} + 0.02}{1.4}, \qquad i = 1, 2, ..., n \quad j = 1, 2, ..., 20. \tag{2.1}$$

Neka je $A = [a_{ij}]$ PAM matrica (engl. *point accepted mutation matrix*). Ona sadrži vjerojatnosti da pojedina aminokiselina mutira u drugu aminokiselinu za aminokiseline iz $\mathcal{A}$. Neka je sada $B = [b_{ij}] = A^k$, za $k = 120$. Vektor $b_i = (b_{i1}, b_{i2}, ..., b_{i20})$, za $i \in \{1, 2, ..., 20\}$ predstavlja očekivani vektor mutacije za i-tu aminokiselinu iz $\mathcal{A}$ nakon k milijuna godina evolucije.

$$\widetilde{g}_{ij} = \sum_{k=1}^{20} g_{ik} b_{kj}, \qquad i = 1, 2, ..., n \quad j = 1, 2, ..., 20 \tag{2.2}$$

$\widetilde{g}_{ij}$ je vjerojatnost da se na i-toj poziciji motiva pojavi j-ta aminokiselina iz $\mathcal{A}$.

Definirajmo vjerojatnosni vektor q kojim je određena prosječna distribucija aminokiselina u proteinskim familijama.

$$q = [0.078,\ 0.051,\ 0.043,\ 0.053,\ 0.019,\ 0.043,\ 0.063,\ 0.072,\ 0.023,\ 0.053,\\ 0.091,\ 0.059,\ 0.022,\ 0.039,\ 0.052,\ 0.068,\ 0.059,\ 0.014,\ 0.032,\ 0.066]$$

Distribucija q dobivena je računanjem relativnih frekvencija aminokiselina iz $\mathcal{A}$ u proteomima velikog broja organizama. Nazivamo je još i *bio-background*.

Budući da raspolažemo malom količinom podataka, imamo malo informacija o proteinskoj familiji koja nas zanima. Zbog toga radimo *log-odds* omjera vjerojatnosti iz (2.2) i vjerojatnosti iz *bio-background*:

$$p_{ij} = \log\left(\frac{\widetilde{g}_{ij}}{q_j}\right), \qquad i = 1, 2, ..., n \quad j = 1, 2, ..., 20. \tag{2.3}$$

Vektor $p_i = (p_{i1}, p_{i2}, ..., p_{i20})$, za $i \in \{1, 2, ..., n\}$ označava vjerojatnost pojave aminokiselina iz $\mathcal{A}$ na i-toj poziciji motiva, a matrica $P = [p_{ij}]$ predstavlja profil zadanog motiva. Duljina profila je broj redaka od P, tj. n.

Ocjena sličnosti (score)

Kako bismo odredili koji nam je od n-grama "dovoljno dobar" kao jedna od varijanti (mutacija) zadanog motiva moramo definirati ocjenu sličnosti, odnosno odrediti način dodjeljivanja određenog *scora* tom n-gramu u usporedbi sa zadanim motivom.
Neka je $P = [p_{ij}]$ profil zadanog motiva i neka je $x^{(k)} = x_k x_{k+1} ... x_{k+n-1}$ n-gram na k-toj poziciji u proteinu duljine l, za $k \in \{1, 2, ..., l - n + 1\}$. Ocjenu sličnosti ili *score* definiramo kao evaluaciju n-grama $x^{(k)}$ u odnosu na profil P formulom:

$$s_k = \sum_{h=0}^{n-1} \log\left(\frac{\mathbb{P}(x_{k+h}|p_{h+1})}{\mathbb{P}(x_{k+h}|q)}\right) \tag{2.4}$$

Sliding window i značajnost ocjene sličnosti

Pretpostavimo da imamo zadanu proteinsku familiju za koju želimo odrediti karakteristične motive. Prvo uzmemo upit koji nam predstavlja ulazni motiv. Metodom klizećeg prozora (engl. *sliding window*) usporedimo naš upit sa svakim od n-grama u svim proteinima zadane proteinske familije. Grafički ćemo prikazati metodu *sliding window* za upit

$y = y_1y_2...y_n$ i protein $x = x_1x_2...x_l$, gdje je $n \leq l$.

$$
\begin{array}{cccccccccc}
x_1 & x_2 & x_3 & \dots & x_{n-1} & x_n & x_{n+1} & \dots & x_l \\
y_1 & y_2 & y_3 & \dots & y_{n-1} & y_n & & & \\
 & & & & & & & & \\
x_1 & x_2 & x_3 & \dots & x_n & x_{n+1} & x_{n+2} & \dots & x_l \\
 & y_1 & y_2 & \dots & y_{n-1} & y_n & & & \\
 & & & & & & & & \\
\vdots & & & & & & & & \\
 & & & & & & & & \\
x_1 & x_2 & \dots & x_n & \dots & x_{l-n+1} & x_{l-n+2} & \dots & x_l \\
 & & & & & y_1 & y_2 & \dots & y_n
\end{array}
$$

Kod svakog uspoređivanja upita y s n-gramom u proteinu x računamo ocjenu sličnosti definiranu formulom (2.4). Što je ocjena sličnosti veća n-gram je sličniji danom upitu y. Za svaki protein iz zadane proteinske familije pamtimo maksimalnu ocjenu sličnosti S definiranu formulom:

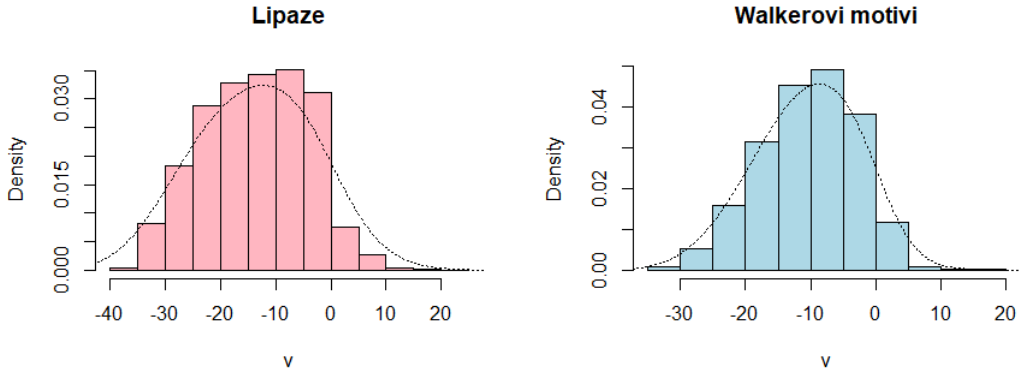
$$S = \max_{k=1,2,...,l-n+1} s_k \tag{2.5}$$

Kada smo popisali maksimalne ocjene sličnosti za sve proteine iz zadane proteinske familije željeli bismo odrediti najznačajnije maksimalne ocjene, odnosno definirati prag iznad kojeg ćemo maksimalnu ocjenu smatrati značajnom. Kako bi to odredili moramo prvo odrediti distribuciju maksimalnih ocjena. Budući da tražimo najznačajnije maksimalne ocjene dovoljno je pogledati samo desni rep distribucije jer se u njemu nalaze maksimumi. Iz histograma se lako može vidjeti da ocjene sličnosti unutar jednog proteina prate neku od distribucija s eksponencijalnim repom (vidi 2.7). Desni rep je upravo ono što nas zanima jer se tu nalaze maksimalne ocjene sličnosti. Uz pretpostavku nezavisnosti i jednake distribuiranosti proteina iz teorema 1.2.15 slijedi da maksimalne ocjene sličnosti prate neku od distribucija ekstremnih vrijednosti. Prema [2] simulacijama se lako može provjeriti da je za nizove jednakih duljina to upravo Gumbelova distribucija. Međutim, mi nemamo ispunjenu traženu pretpostavku jednake distribuiranosti jer naši proteini nisu jednakih duljina. Očito, što je protein dulji to je veća vjerojatnost da ćemo u njemu pronaći n-gram s većom ocjenom sličnosti. Stoga je prirodno pretpostaviti da je distribucija maksimalnih ocjena ovisna o duljini proteina. Iz histograma duljina proteina 2.8 uočavamo da bi duljine proteina mogle pratiti Gumbelovu distribuciju. Također, iz histograma maksimalnih ocjena 2.9 možemo vidjeti da maksimalne ocjene prate logističku distribuciju. Budući da dolazi do korekcije s obzirom na duljinu proteina, pojavu logističke distribucije možemo opravdati korolarom 1.2.14 prema kojem razlika dvije Gumbel distribuirane slučajne varijable ima logističku distribuciju. Više o distribucijama maksimalnih ocjena u [7].

Sada možemo definirati prag tako da se za određeni broj β udaljimo od prosječne maksimalne ocjene sličnosti, tj. formulom:

$$prag = \mu + skala \cdot \beta. \tag{2.6}$$

U formuli (2.6) μ je očekivanje neprekidne slučajne varijable s logističkom distribucijom, β je parametar logističke distribucije definiran s $\frac{\sqrt{3}}{\pi}\sigma$, a skala je proizvoljan pozitivan broj. Eksperimentalno odredimo "dovoljno dobru" skalu koja ne smije biti prevelika jer u tom slučaju dobijemo motive koji se sastoje od samo nekoliko n-grama i ne smije biti premala jer tada dobijemo motive koji se sastoje od velikog broja n-grama s jako malom međusobnom sličnošću. U ovom radu koristit ćemo skalu 4.

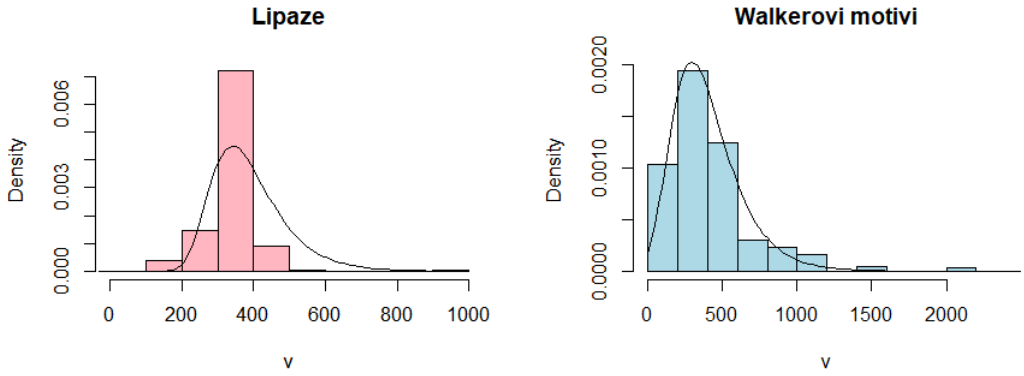


Slika 2.7: Histogrami ocjena sličnosti s procijenjenim funkcijama gustoće za jedan protein s obzirom na jedan ulazni motiv

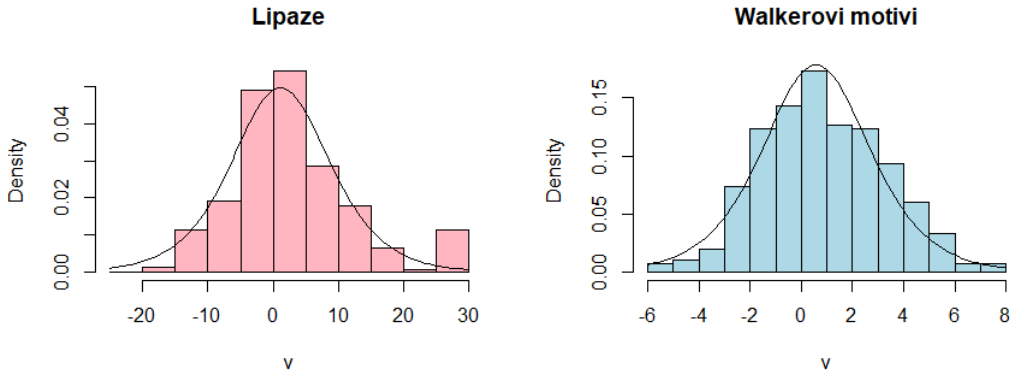
Iteriranje

Postupak traženja motiva provodimo iterativnim putem. Prvo odredimo upit i redni broj proteina iz kojeg smo uzeli upit. U prvoj iteraciji upit je ulazni motiv koji se sastoji od jednog n-grama, a kasnije se može sastojati i od više n-grama. Zatim odredimo prag i opisanom metodom *sliding window* tražimo n-grame "dovoljno slične" upitu. One n-grame čija je ocjena sličnosti veća ili jednaka od praga smatrat ćemo "dovoljno sličnima" i njih ćemo dodati kao varijantu upita (ulaznog motiva) i zapamtiti redni broj proteina u kojem smo pronašli taj n-gram.

Novodobiveni motiv nam zatim postaje ulazni motiv. Određujemo njegov profil i prag, a zatim ponavljamo postupak traženja i ocjenjivanja n-grama metodom *sliding window* na



Slika 2.8: Histogrami duljina proteina i funkcije gustoće Gumbelove distribucije s procijenjenim parametrima



Slika 2.9: Histogrami maksimalnih ocjena za jedan ulazni motiv i funkcije gustoće logističke distribucije s procijenjenim parametrima

temelju tog motiva. S iteriranjem stajemo kada se lista n-grama koji čine motiv ne promijeni ili kada dostignemo zadani maksimalni broj iteracija. Za maksimalni broj iteracija u ovom radu uzimamo 10.

Kada završimo cijeli proces iteracije za jedan upit u proteinu se pomičemo jedno mjesto udesno, uzimamo novi upit i ponavljamo isti postupak iteriranja. Postupak ponavljamo dok nismo iskoristili sve upite u zadanoj proteinskoj familiji. Na kraju postupka imamo listu motiva zadane proteinske familije. Kako bismo smanjili broj motiva uzimamo u obzir samo one koji su građeni od više od 15 n-grama.

Filtriranje motiva s istim prvim n-gramom

U dobivenoj listi motiva zadane proteinske familije možemo uočiti velik broj motiva koji su sastavljeni od istih ili gotovo istih n-grama. U daljnjem radu n-grame ćemo promatrati zajedno s rednim brojem proteina u kojem su pronađeni. Kako ne bismo za svaka dva motiva u listi uspoređivali sve n-grame od kojih su sastavljeni i tražili razlike, filtriranje ćemo provesti samo na temelju prvog n-grama. Sve motive kojima su prvi n-grami isti i pronađeni u istom proteinu spojimo u jedan motiv koji se sastoji od svih n-grama koji se nalaze u svakom od tih motiva, ali bez ponavljanja onih n-grama koji su isti i pronađeni u istom proteinu. Time znatno reduciramo broj motiva u ukupnoj listi motiva zadane proteinske familije. Objasnimo postupak filtriranja motiva s istim prvim n-gramom na primjeru. Neka su redom zadani motivi M_1, M_2 i M_3.

M_1	M_2	M_3
MADGLVKGPY 15	MADGLVKGPY 15	MADGLVKGPY 15
MAQGLLNGRY 17	MAQGLLNGRY 17	MAQGLLNGRY 17
MAEGILNGPY 19	MAEGILNGPY 19	MAEGILNGPY 19
IADGILNGPY 20	IADGILNGPY 20	IADGILNGPY 20
MAQVILNGTY 21	MTEGILNGPY 23	VSSQILTGKY 22
MTEGILNGPY 23	MAEGILNGPY 28	MTEGILNGPY 23
MAEGILNGPY 28	MAEGILKGPY 29	MAEGILNGPY 28
MAEGILKGPY 29	ISEGLLKGPY 30	MAEGILKGPY 29
ISEGLLKGPY 30	MAEGILNGPY 41	ISEGLLKGPY 30
MAEGILNGPY 41	MAEGILNGPY 43	MAEGILNGPY 41
MAEGILNGPY 43	IADGLLKGPY 44	VSSQILTGKY 42
IADGLLKGPY 44	MAEGILKGPY 46	MAEGILNGPY 43
MAEGILKGPY 46	MADGLVKGPY 47	IADGLLKGPY 44
MADGLVKGPY 47	IANSILNGPY 126	MAEGILKGPY 46
IANSILNGPY 126	IAIGLLRGSY 144	MADGLVKGPY 47
IADGVLNGPF 154	IAIGLLRGSY 145	ISEGVLTGPY 63
	IAIGLLQGPY 146	VALHILTGKY 81
	IAIGLLQGSY 147	IANSILNGPY 126
	IAIGLLQGSY 149	IADGVLNGPF 154
	VASQILTGKY 281	IASQILTGRY 211
		VASQILTGKY 281

Odmah možemo uočiti da je prvi n-gram zajedno s rednim brojem proteina u kojem je pronađen “MADGLVKGPY 15” isti u sva tri motiva. Nadalje, “MAQGLLNGRY 17”, “MAEGILNGPY 19”, “IADGILNGPY 20”, “MTEGILNGPY 23”, “MAEGILNGPY 28”, “MAEGILKGPY 29”, “ISEGLLKGPY 30”, “MAEGILNGPY 41”, “MAEGILNGPY 43”, “IADGLLKGPY 44”, “MAEGILKGPY 46”, “MADGLVKGPY 47” i “IANSILN-

GPY 126" se javljaju u sva tri motiva. "IADGVLNGPF 154" se javlja u motivima M_1 i M_3, "VASQILTGKY 281" se javlja u motivima M_2 i M_3, dok se ostali n-grami zajedno s rednim brojem proteina u kojem su pronađeni "MAQVILNGTY 21", "IAIGLLRGSY 144", "IAIGLLRGSY 145", "IAIGLLQGPY 146", "IAIGLLQGSY 147", "IAIGLLQGSY 149", "VSSQILTGKY 22", "VSSQILTGKY 42", "ISEGVLTGPY 63", "VALHILTGKY 81", "IASQILTGRY 211" javljaju samo u jednom od ta tri motiva. Motiv koji dobijemo filtriranjem na opisani način je:

MADGLVKGPY 15
MAQGLLNGRY 17
MAEGILNGPY 19
IADGILNGPY 20
MAQVILNGTY 21
MTEGILNGPY 23
MAEGILNGPY 28
MAEGILKGPY 29
ISEGLLKGPY 30
MAEGILNGPY 41
MAEGILNGPY 43
IADGLLKGPY 44
MAEGILKGPY 46
MADGLVKGPY 47
IANSILNGPY 126
IADGVLNGPF 154
IAIGLLRGSY 144
IAIGLLRGSY 145
IAIGLLQGPY 146
IAIGLLQGSY 147
IAIGLLQGSY 149
VSSQILTGKY 22
VSSQILTGKY 42
ISEGVLTGPY 63
VALHILTGKY 81
IASQILTGRY 211
VASQILTGKY 281

Produživanje motiva

Nakon filtriranja motiva s istim prvim n-gramom u našoj listi motiva možemo uočiti motive kod kojih se prvi n-grami međusobno nadovezuju s pomakom za c pozicija, $c \in \{1, 2, ..., n-2\}$. To je posljedica načina na koji smo birali motive metodom *sliding window*. Objasnimo što znači da se dva n-grama međusobno nadovezuju s pomakom za c pozicija. Pretpostavimo da imamo zadana dva n-grama N_1 i N_2 zajedno s rednim brojem proteina u kojem su pronađeni. Reći ćemo da se N_1 i N_2 međusobno nadovezuju s pomakom za c pozicija ako se posljednjih $n-c$ aminokiselina n-grama N_1 poklapa s prvih $n-c$ aminokiselina n-grama N_2 i oba n-grama su pronađena u istom proteinu zadane proteinske familije. Tada bi mogli na n-gram N_1 dodati posljednjih c aminokiselina n-grama N_2 i dobiti novi (n+c)-gram. Pokažimo opisano na dva primjera.
Ako za N_1 i N_2 imamo redom "ISTISSAITE 15" i "STISSAITEL 15", tada se N_1 i N_2 međusobno nadovezuju s pomakom za 1 poziciju jer su oba 10-grama pronađena u 15. proteinu zadane proteinske familije i 2. – 10. aminokiselina prvog 10-grama se poklapa s 1. – 9. aminokiselinom drugog 10-grama.

```
ISTISSAITE
 STISSAITEL
```

Ako za N_1 i N_2 imamo redom "ISTISSAITE 15" i "TISSAITELI 15", tada se N_1 i N_2 međusobno nadovezuju s pomakom za 2 pozicije jer su oba 10-grama pronađena u 15. proteinu zadane proteinske familije i 3. – 10. aminokiselina prvog 10-grama se poklapa s 1. – 8. aminokiselinom drugog 10-grama.

```
ISTISSAITE
  TISSAITELI
```

Željeli bismo spojiti one motive kod kojih se prvi n-grami međusobno nadovezuju s pomakom za c pozicija, za $c \in \{1, 2, ..., n-2\}$, ali tako da dopuštamo spajanje, odnosno produživanje samo jednu po jednu poziciju. Preciznije, ako imamo motive M_1 i M_2 čiji se prvi n-grami međusobno nadovezuju s pomakom za c pozicija, gdje je $c > 1$, spojit ćemo ih samo ako je u nekom od prethodnih koraka motiv M_1 dobiven spajanjem neka dva motiva kod kojih su se prvi n-grami međusobno nadovezivali s pomakom za $c-1$ poziciju. Zbog načina na koji smo prethodno filtrirali motive na temelju prvih n-grama, prvi n-grami svih motiva u listi su međusobno različiti.

Intuitivno, kada uočimo dva motiva kod kojih se prvi n-grami međusobno nadovezuju s pomakom za c pozicija željeli bismo spojiti ta dva motiva, tj. svakom od n-grama prvog motiva željeli bismo dodati c posljednjih aminokiselina odgovarajućeg n-grama drugog motiva. Međutim, to u većini slučajeva nećemo moći. Jedan od razloga je i taj što naši motivi nisu sastavljeni od jednakog broja n-grama. Stoga spajanje, odnosno produživanje motiva nećemo raditi na razini n-grama (nizova aminokiselina), nego na razini distribucija

motiva. Svim motivima u listi motiva zadane proteinske familije odredimo distribuciju na način objašnjen ranije u ovom potpoglavlju i spremimo u novu listu. Kako je objašnjeno, i-ti redak distribucije motiva predstavlja distribuciju aminokiselina iz $\mathcal{A}$ na i-toj poziciji motiva, za $i = 1, 2, ..., n$. Objasnimo postupak produživanja na primjeru dvaju motiva M_1 i M_2 s distribucijama D_1 i D_2 duljine n.

M1	M2
ISTISSAITE 15	STISSAITEL 15
IASISSTITE 20	ASISSTITEL 20
ITTISSAITE 28	TTISSAITEL 28
ITTISSAITE 43	ETISSAITEL 29
ITKISSAIVE 44	ATISSAITEL 30
IETISSAITE 46	TTISSAITEL 43
ISTISSAITE 47	TKISSAIVEL 44
IASISSTITE 48	ETISSAITEL 46
ITTISSAISE 63	STISSAITEL 47
IAKISSTITE 144	ASISSTITEL 48
IAKISSTITE 145	TTISSAISEL 63
IAKISSIITE 147	GTITSAITEL 143
IAKISSIITE 148	AKISSTITEL 144
IAKISSIITE 149	AKISSTITEL 145
IAKISSTITE 151	AKISSIITEL 147
ITKISSAIVV 18	AKISSIITEL 148
IAKISSTITL 146	AKISSIITEL 149
IAKISSTITV 150	AKISSTITLS 146
	AKISSTITVS 150
	ASISSTVQEL 154
	KAISSAIVDL 17
	KAISSAIVDL 21
	KAISSAIVDL 23
	KAISSAIVDL 41

Uočimo da se prvi n-grami motiva M_1 i M_2 međusobno nadovezuju s pomakom za 1 poziciju. Spajanje motiva M_1 i M_2 radimo tako da 10. redak distribucije D_2 nadodamo kao posljednji redak distribucije D_1, a zatim izbacimo distribuciju D_2 iz liste distribucija svih motiva i motiv M_2 iz liste svih motiva. Sada se distribucija D_1 sastoji od 11 redaka. Time smo na neki način svaki od 10-grama motiva M_1 proširili do 11-grama pomoću aminokiselina na 10. poziciji motiva M_2.

Budući da gledamo samo prve n-grame, pri spajanju motiva na opisani način javlja se određena greška. Kako greška ne bi bila prevelika i spajanje motiva na temelju prvih n-

grama neopravdano, grešku ćemo "regulirati" pomoću relativne entropije distribucija onih pozicija motiva na kojima se prvi n-grami podudaraju. U prethodno navedenom primjeru promatramo $2.-10.$ poziciju motiva M_1 i $1.-9.$ poziciju motiva M_2. Kako se aminokiselina na 2. poziciji prvog n-grama motiva M_1 podudara s aminokiselinom na 1. poziciji prvog n-grama motiva M_2, računamo relativnu entropiju distribucije 2. pozicije motiva M_1 u odnosu na distribuciju 1. pozicije motiva M_2 i obrnuto. Zatim zbrojimo ta dva dobivena broja kako bi na neki način "simetrizirali" relativnu entropiju, tj. računamo

$$H(D_1[2,] \,||\, D_2[1,]) + H(D_2[1,] \,||\, D_1[2,])$$

prema formuli 1.7. Označimo dobiveni broj s RE_1. Ponovimo postupak za sve tražene pozicije i dobijemo redom brojeve RE_2, RE_3, ..., RE_9. Zbrajanjem svih dobivenih brojeva dobijemo ukupnu relativnu entropiju pozicija motiva na kojima se prvi n-grami motiva M_1 i M_2 podudaraju.

$$RE = RE_1 + RE_2 + ... + RE_9$$

Željeli bismo da su distribucije onih pozicija motiva na kojima se prvi n-grami podudaraju što sličnije, stoga želimo da je dobivena ukupna relativna entropija RE što manja, tj. "dovoljno blizu 0". Potrebno je odrediti granicu ispod koje ćemo vrijednosti RE smatrati "dovoljno blizu 0". Promotrimo motive jedne proteinske familije. Ako za svaki $c \in \{1, 2, ..., n-2\}$ pogledamo histogram ukupnih relativnih entropija izračunatih za sve parove motiva kod kojih se prvi n-grami međusobno nadovezuju s pomakom za c pozicija možemo uočiti da je gotovo uvijek najmanje $\frac{3}{4}$ vrijednosti "jako blizu 0" pa možemo pretpostaviti da je oko $\frac{1}{4}$ parova motiva "krivo" spojeno, odnosno spojeno s "velikom" greškom (vidi 2.11). Iz tog razloga granicu ćemo odrediti na način da za traženi $c = 1$ izračunamo ukupnu relativnu entropiju svih parova motiva u listi motiva zadane proteinske familije kod kojih se prvi n-grami međusobno nadovezuju s pomakom za c pozicija i uzmemo 75% kvantil dobivenih vrijednosti. Pamtimo dobivenu granicu za pomak od c pozicija kako je ne bismo svaki put iznova računali. U slučaju da je dobivena ukupna relativna entropija RE manja ili jednaka dobivenoj granici izvršit ćemo spajanje motiva M_1 i M_2 na ranije opisan način.

Pretpostavimo da smo u daljnjem traženju pronašli motiv M_3 s distribucijom D_3.

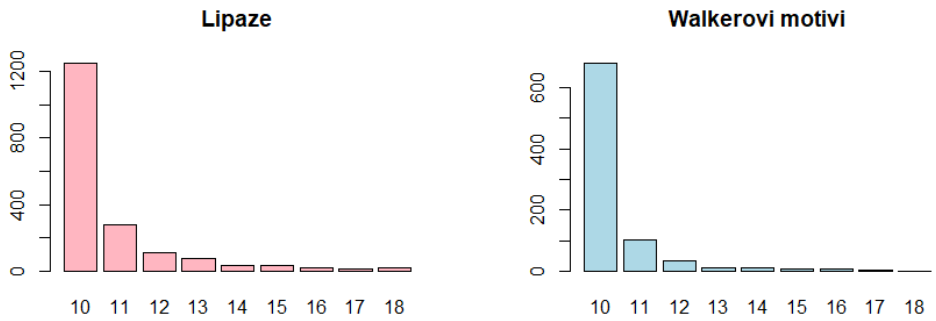
M_3

```
TISSAITELI 15
SISSTITELI 20
TISSAITELI 28
TISSAITELI 29
TISSAITELV 30
TISSAITELI 43
TISSAITELI 46
TISSAITELI 47
```

```
SISSTITELI 48
TISSAISELV 63
TITSAITELI 143
KISSTITELI 144
KISSTITELI 145
KISSIITELI 147
KISSIITELI 149
KISSTITELI 151
AISSAIVDLI 17
AISSAIVDLI 21
AISSAIVDLI 23
AISSAIVDLI 41
KISSAIVELV 44
SISSTVQELI 154
KISSIITELR 148
```

Vidimo da se prvi n-grami motiva M_1 i M_3 međusobno nadovezuju s pomakom za 2 pozicije. Budući da je $c = 2$ što je veće od 1, prvo ćemo provjeriti jesmo li u nekom od prethodnih koraka algoritma motiv M_1 spojili s nekim motivom tako da su im se prvi n-grami međusobno nadovezivali s pomakom za $c - 1$, tj. 1 poziciju. Provjeravamo tako da pogledamo sastoji li se distribucija motiva M_1 od $n + (c - 1)$, tj. 11 redaka. Ukoliko je to istina, izračunamo RE, odnosno ukupnu relativnu entropiju onih pozicija motiva M_1 i M_3 na kojima se njihovi prvi n-grami podudaraju. Zatim, ukoliko već nije određena, odredimo granicu za pomak od 2 pozicije na ranije opisan način. Ukoliko je RE manja ili jednaka granici izvršit ćemo spajanje motiva M_1 i M_3. Spajanje, odnosno produživanje motiva radimo tako da distribuciju 10. pozicije motiva M_3, tj. 10. redak distribucije D_3, nadodamo kao posljednji redak distribucije D_1. Naposljetku, izbacimo distribuciju D_3 iz liste svih distribucija i motiv M_3 iz liste svih motiva. Tada se distribucija D_1 sastoji od 12 redaka i na neki način smo motiv M_1 produžili od duljine 11 do duljine 12 pomoću aminokiselina na 10. poziciji motiva M_3. Budući da nismo imali uvjeta na duljinu distribucije D_3, time smo možda izbacili motiv čija je duljina veća od n, ali smo dobili novi, dulji motiv koji je "jako sličan" izbačenom motivu na $n - c + 1$ pozicija.

Postupak ponavljamo za svaki od pomaka iz $\{1, 2, ..., n - 2\}$ tako da usporedimo prve n-grame svaka dva motiva iz "preostale" liste motiva zadane proteinske familije i izvršimo sva moguća produživanja za zadani pomak na opisan način. Tako reduciramo broj motiva u ukupnoj listi motiva zadane proteinske familije i dobivamo motive koji nisu uvijek fiksne duljine n, već im je duljina od n do $2n - 2$. Duljine novodobivenih motiva, preciznije njihovih distribucija, možemo prikazati stupčastim dijagramom (vidi 2.10). U daljnjem radu motive promatramo na razini njihovih distribucija, a ne n-grama s mutacijama.



Slika 2.10: Stupčasti dijagrami duljina motiva

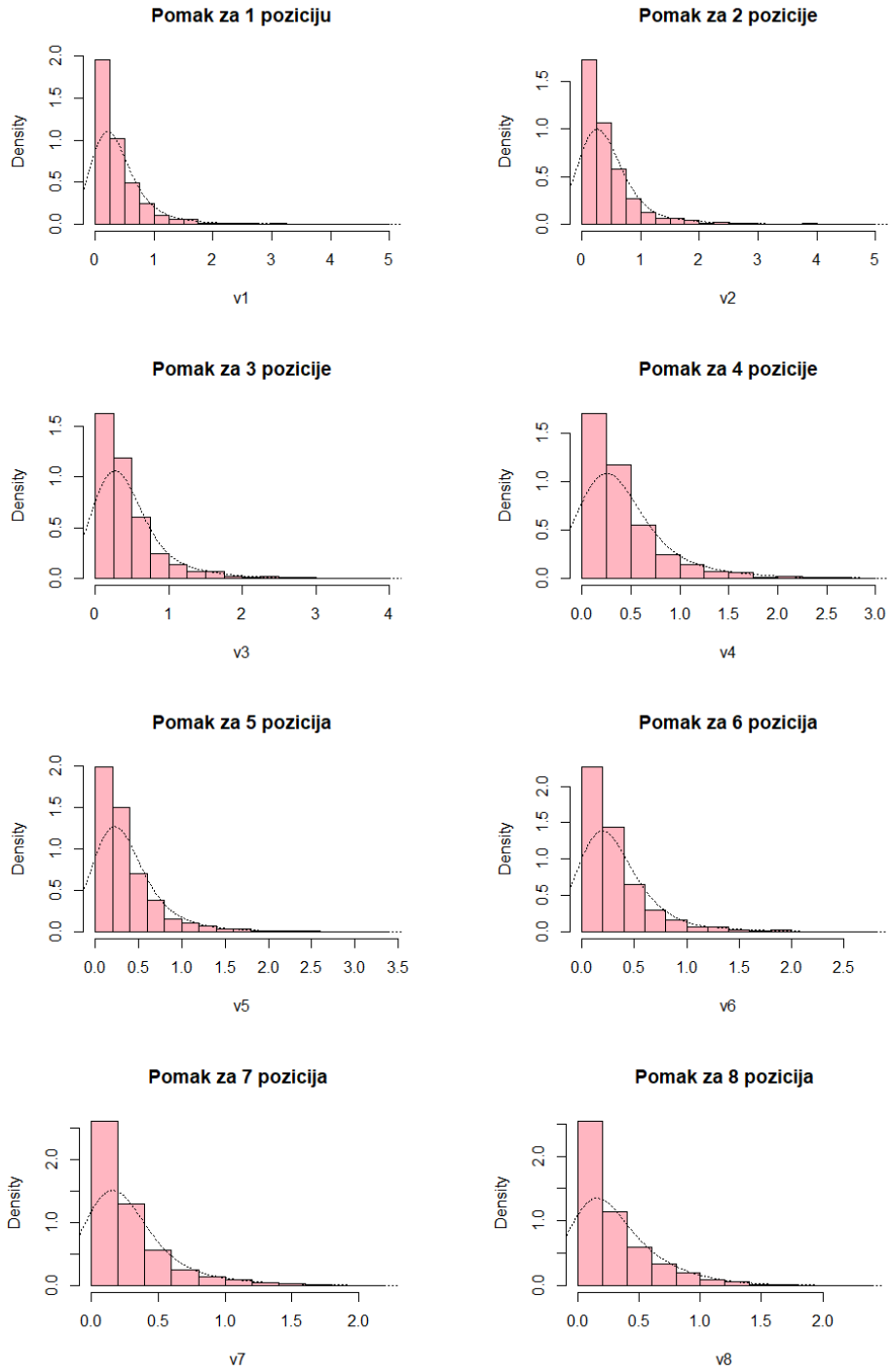
Redukcija broja motiva entropijom

Željeli bismo dodatno reducirati broj motiva tako da od svih distribucija motiva koje smo dobili odredimo njih 100 koje će "najbolje" opisivati proteine zadane proteinske familije. Kako ne želimo da su nam distribucije motiva "previše raspršene" koristit ćemo entropiju definiranu u poglavlju 1. Za 100 najboljih motiva uzet ćemo 100 distribucija motiva koje imaju najmanju prosječnu entropiju distribucija od kojih su sastavljene. Ukupnu entropiju distribucije motiva dobijemo tako da za distribuciju svake pozicije motiva izračunamo entropiju prema formuli 1.6 i zbrojimo dobivene vrijednosti. Kako bi dobili prosječnu entropiju distribucija od kojih je sastavljena distribucija motiva podijelimo dobivenu ukupnu entropiju s njihovim brojem, tj. duljinom distribucije motiva.

Pretpostavimo da imamo distribuciju motiva D duljine d, gdje je $d \in \{n, n+1, ..., 2n-2\}$. Prosječna entropija distribucija od kojih je sastavljena distribucija motiva D iznosi:

$$E = \frac{1}{d} \sum_{i=1}^{d} H(D[i,])$$

gdje je $H(D[i,]) = -\sum_{j=1}^{20} D[i,j] \log_{20}(D[i,j])$.



Slika 2.11: Histogrami ukupnih relativnih entropija za svaki pomak

Poglavlje 3

Validacija

3.1 Ocjena proteina u odnosu na profile motiva

U prethodnom poglavlju opisali smo algoritam traženja karakterističnih motiva kojim dolazimo do liste od 100 motiva iz zadane proteinske familije od interesa. Svakom od tih motiva odredimo profil na način opisan u potpoglavlju 2.3. Označimo s $P = \{P_1, P_2, ..., P_{100}\}$ skup profila dobivene liste motiva. Neka je $B = [b_1, b_2, ..., b_{20}]$ proizvoljan profil iz skupa P duljine d, gdje je $d \in \{n, n+1, ..., 2n-2\}$. Želimo li protein duljine l ocijeniti u odnosu na profil B potrebno je evaluirati svaki d-gram u proteinu s obzirom na profil B i uzeti maksimum dobivenih vrijednosti. Preciznije, ako je za $k \in \{1, 2, ..., l-d+1\}$ $x^{(k)} = x_k x_{k+1} ... x_{k+d-1}$ d-gram na k-toj poziciji zadanog proteina duljine l evaluaciju d-grama $x^{(k)}$ u odnosu na profil B računamo pomoću formule 2.4, tj. vrijedi:

$$s_k = \sum_{h=0}^{d-1} \log\left(\frac{\mathbb{P}(x_{k+h}|b_{h+1})}{\mathbb{P}(x_{k+h}|q)}\right)$$

Nakon što odredimo s_k za svaki $k \in \{1, 2, ..., l-d+1\}$ ocjenu proteina u odnosu na profil B koju označavamo sa S dobijemo tako da odredimo maksimum dobivenih vrijednosti.

$$S = \max_{k=1,2,...,l-d+1} s_k$$

Zadani protein ocijenimo na opisan način u odnosu na sve profile iz skupa P i vrijednosti stavimo redom u vektor v_P. Na taj način iz proteina dobijemo 100-dimenzionalni vektor s kojim dalje možemo računati. Računajući normu vektora v_P možemo odrediti koliko je dobro zadani protein opisan listom od 100 motiva iz zadane proteinske familije. Računamo 1-normu, 2-normu i max-normu vektora v_P pomoću formula 1.1, 1.2, 1.3 iz poglavlja 1. Što je norma veća protein je bolje opisan profilima iz skupa P. Na taj način moći ćemo

usporediti koliko dobro je neki protein opisan profilima motiva za različite proteinske familije od interesa i na temelju toga provesti klasifikaciju. Klasifikaciju ćemo provesti tako da protein pridružimo onoj proteinskoj familiji čiji vektor ocjena tog proteina u odnosu na profile karakterističnih motiva ima maksimalnu normu.

3.2 Podjela podataka

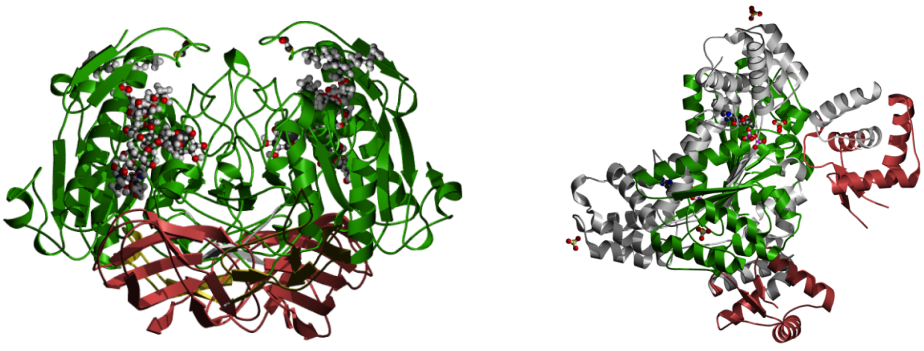
Kod problema klasifikacije običaj je da se skup podataka podijeli na dva dijela. Veći dio podataka nazivamo *trening set* i on nam služi za analizu algoritma i izgradnju rješenja. U našem slučaju koristit ćemo ga kod traženja liste od 100 karakterističnih motiva zadane proteinske familije. Drugi dio podataka nazivamo *test set* i služi nam za testiranje rješenja i procjenu točnosti. U našem slučaju koristit ćemo ga kod ocjenjivanja proteina u odnosu na profile motiva proteinskih familija koje klasificiramo i analize dobivenih rezultata.

Poglavlje 4

Analiza algoritma na dvije proteinske familije

4.1 Opis familija

Promatramo dvije proteinske familije, lipoproteinske lipaze i Walkerove motive. Prema [8] lipoproteinske lipaze su enzimi koji ubrzavaju kemijsku reakciju razgradnje masti u doticaju s vodom, dok Walkerovi motivi imaju ulogu u vezanju molekula ATP-a. Iz svake familije uzeli smo po 400 proteina koje smo podijelili u omjeru 3 : 1 na *trening set* i *test set*, odnosno po 300 proteina za *trening set* i po 100 proteina za *test set*. Proteinsku familiju lipoproteinskih lipaza nazivat ćemo još i prva familija, a familiju Walkerovih motiva druga familija.



Slika 4.1: 3D struktura nekih proteina iz familija lipoproteinskih lipaza i Walkerovih motiva (vidi [10])

4.2 Primjena algoritma

Za svaku familiju odredimo listu od 100 motiva i njihove profile kako je opisano u potpoglavlju 2.3. Označimo skup profila motiva prve familije s $P^{(1)}$, a druge familije s $P^{(2)}$. Na temelju profila motiva iz $P^{(1)}$ i $P^{(2)}$ klasificiramo proteine iz *test setova* obje familije. Objasnit ćemo postupak ocjenjivanja iz potpoglavlja 3.1 na primjeru jednog proizvoljnog proteina iz *test seta* prve familije. Za taj protein odredimo vektore $v_{P^{(1)}}$ i $v_{P^{(2)}}$, gdje je $v_{P^{(1)}}$ vektor ocjena zadanog proteina u odnosu na profile iz skupa $P^{(1)}$ i $v_{P^{(2)}}$ vektor ocjena zadanog proteina u odnosu na profile iz skupa $P^{(2)}$. Zatim odredimo 1-normu, 2-normu i max-normu vektora $v_{P^{(1)}}$ i $v_{P^{(2)}}$. Svaku normu gledamo posebno. Uspoređujemo odgovarajuće norme vektora $v_{P^{(1)}}$ i $v_{P^{(2)}}$ i ukoliko je norma vektora $v_{P^{(1)}}$ veća od norme vektora $v_{P^{(2)}}$ zadani protein pridružujemo prvoj familiji, inače zadani protein pridružujemo drugoj familiji. Opisani postupak radimo za sve proteine iz *test seta* prve familije. Potpuno analogno klasificiramo proteine iz *test seta* druge familije.

4.3 Analiza rezultata

Točnost algoritma za traženje karakterističnih motiva procjenjujemo na temelju toga koliki broj proteina iz *test setova* je pridružen familiji kojoj stvarno pripada. Dobiveni rezultati prikazani su u tablici 4.2.

	1. familija		**2. familija**	
	Točno	**Netočno**	**Točno**	**Netočno**
1-norma	98	2	100	0
2-norma	98	2	99	1
max-norma	100	0	98	2

Slika 4.2: Tablica rezultata

Vidimo da su dobiveni rezultati jako dobri. Za sve tri norme klasifikacija proteina iz *test setova* je veća ili jednaka 98%, stoga možemo zaključiti da su dobiveni motivi zaista karakteristični za pripadajuće proteinske familije. Ovakvi rezultati su zanimljivi jer su dobiveni uz korištenje jako male razine biološkog znanja o proteinskim familijama koje klasificiramo.

Za svaku normu posebno izračunamo osjetljivost, specifičnost, PPV i NPV definirane u poglavlju 1 kako bi odredili uspješnost algoritma (vidi tablice 4.1, 4.2 i 4.3). Prvu familiju gledamo kao pozitivno stanje, a drugu kao negativno.

		predviđeno stanje		
		pozitivno stanje	**negativno stanje**	
stvarno stanje	**pozitivno stanje**	TP = 98	FN = 2	osjetljivost = 0.98
	negativno stanje	FP = 0	TN = 100	specifičnost = 1
		PPV = 1	NPV = 0.9804	

Tablica 4.1: Matrica konfuzije, 1-norma

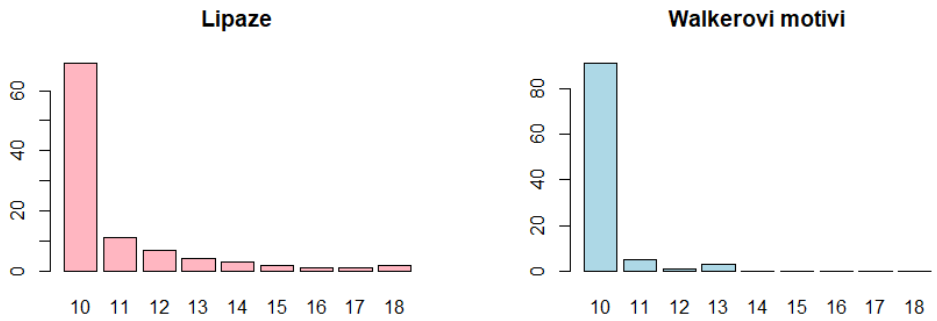
		predviđeno stanje		
		pozitivno stanje	**negativno stanje**	
stvarno stanje	**pozitivno stanje**	TP = 98	FN = 2	osjetljivost = 0.98
	negativno stanje	FP = 1	TN = 99	specifičnost = 0.99
		PPV = 0.9899	NPV = 0.9802	

Tablica 4.2: Matrica konfuzije, 2-norma

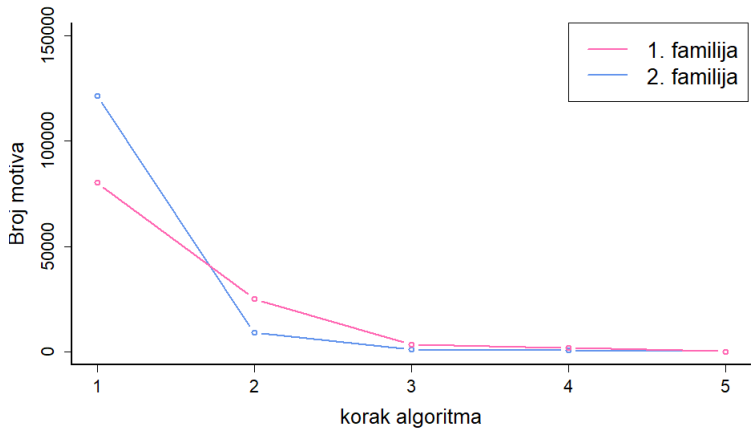
		predviđeno stanje		
		pozitivno stanje	**negativno stanje**	
stvarno stanje	**pozitivno stanje**	TP = 100	FN = 0	osjetljivost = 1
	negativno stanje	FP = 2	TN = 98	specifičnost = 0.98
		PPV = 0.9804	NPV = 1	

Tablica 4.3: Matrica konfuzije, max-norma

Duljine motiva u dobivenim listama od 100 motiva za svaku od promatranih familija možemo vidjeti na stupčastim dijagramima 4.3. Možemo primijetiti da je raspon duljina motiva za prvu familiju od 10 do 18, dok je za drugu familiju od 10 do 13. Dakle, u prosjeku su motivi prve familije dulji od motiva druge familije. Također, u obje familije većina motiva je i dalje duljine 10.



Slika 4.3: Stupčasti dijagrami duljina motiva



Slika 4.4: Redukcija broja motiva u različitim koracima algoritma

korak algoritma	**ukupno 10-torki**	**iteriranje**	**filtriranje motiva**	**produživanje motiva**	**entropija**
1. familija	80121	25020	3385	1834	100
2. familija	121237	9198	1216	857	100

Slika 4.5: Broj motiva u različitim koracima algoritma

U različitim koracima algoritma radili smo redukciju broja motiva. Broj motiva u pojedinim koracima prikazan je na grafu 4.4 i u tablici 4.5. Možemo vidjeti da smo od ukupno 80121 različite 10-torke prve familije dobili 100 karakterističnih motiva i od ukupno 121237 različitih 10-torki druge familije 100 karakterističnih motiva. Time smo smanjili dimenziju prostora u kojoj klasificiramo na 100 elemenata (motiva) po familiji, a i dalje dobili jako dobru klasifikaciju.

Zanimljivo pitanje je i pronalazimo li među dobivenim karakterističnim motivima one koji su od prije identificirani kao značajni za te dvije familije. Za familiju lipoproteinskih lipaza značajan je 10-gram *FVFGDSLSDA* koji još nazivamo i *Block 1*. Pretraživanjem dobivenih karakterističnih motiva prve familije i rekonstrukcijom njihovih distribucija možemo uočiti distribuciju duljine 18 kod koje se s "velikom" vjerojatnošću pojavljuje 18-gram *PAVFNFGDSNSDTGGLSA*. Vjerojatnosti pojave pojedinih aminokiselina 18-grama *PAVFNFGDSNSDTGGLSA* na odgovarajućim pozicijama karakterističnog motiva mogu se vidjeti u tablici 4.6. *Block 1* se poklapa s 18-gramom *PAVFNFGDSNSDTGGLSA* na 7 od 10 pozicija počevši od 4. pozicije što je jako dobro poklapanje.

FVFGDSLSDA
PAVFNFGDSNSDTGGLSA

Za familiju Walkerovih motiva značajan tzv. *Walkerov motiv*, 8-gram *GXXXXGKT/S*. *X* označava bilo koju standardnu aminokiselinu, a / znači "ili". Pretraživanjem dobivenih karakterističnih motiva druge familije i rekonstrukcijom njihovih distribucija na isti način možemo uočiti distribuciju duljine 10 kod koje postoji mogućnost pojave 10-grama *GGGGGGGGKS*. Vjerojatnosti pojave pojedinih aminokiselina 10-grama *GGGGGGGGKS* na odgovarajućim pozicijama karakterističnog motiva mogu se vidjeti u tablici 4.7. *Walkerov motiv* se poklapa s dobivenim 10-gramom *GGGGGGGGKS* na 4 od 4 pozicije počevši od 3. pozicije.

GXXXXGKS
GGGGGGGGKS

Klasifikaciju dviju proteinskih familija na temelju 100 karakterističnih motiva opisanu u ovom diplomskom radu možemo usporediti s klasifikacijom na temelju 100 najfrekventnijih četvorki opisanom u diplomskom radu [8]. U oba rada za klasifikaciju dviju proteinskih familija korišten je isti skup podataka, po 400 proteina iz familija lipoproteinskih lipaza i Walkerovih motiva. U oba rada dobivena je jako dobra klasifikacija proteina iz *test setova*. Rezultat klasifikacije prve familije u oba slučaja je 98%, dok je rezultat klasifikacije druge familije 97% na temelju najfrekventnijih četvorki, a 99% na temelju karakterističnih motiva. Klasifikacije uspoređujemo na temelju Euklidske norme. Postupak

određivanja najfrekventnijih četvorki opisan u diplomskom radu [8] je jednostavniji od algoritma traženja karakterističnih motiva opisanog u ovom diplomskom radu, ali karakteristični motivi su "puno" dulji od četvorki (duljine od 10 do 18) i nisu fiksni, nego sadrže specifične mutacije koje se javljaju evolucijom čime je u obzir uzet i biološki značaj.

Pozicija	**Amino.**	**Vjerojatnost**
1.	P	0.826087
2.	A	0.521739
3.	V	0.478261
4.	F	0.956522
5.	N	0.913043
6.	F	1.0
7.	G	1.0
8.	D	1.0
9.	S	1.0
10.	N	1.0
11.	S	0.956522
12.	D	1.0
13.	T	1.0
14.	G	1.0
15.	G	1.0
16.	L	0.538462
17.	S	0.307692
18.	A	0.923077

Slika 4.6: Vjerojatnosti pojave pojedinih aminokiselina 18-grama *PAVFNFGDSNSDTGGLSA* na odgovarajućim pozicijama karakterističnog motiva

Pozicija	**Amino.**	**Vjerojatnost**
1.	G	0.731707
2.	G	0.902439
3.	G	0.902439
4.	G	0.829268
5.	G	0.682927
6.	G	0.585366
7.	G	0.463415
8.	G	0.512195
9.	K	0.487805
10.	S	0.439024

Slika 4.7: Vjerojatnosti pojave pojedinih aminokiselina 10-grama *GGGGGGGGKS* na odgovarajućim pozicijama karakterističnog motiva

Bibliografija

[1] D. Bakić, *Linearna algebra*, Sveučilište u Zagrebu, Matematički odsjek PMF-a, 2008.

[2] M. Cigula, *Iterativna optimizacija modela i pretraživanje proteoma*, Diplomski rad, Sveučilište u Zagrebu, Matematički odsjek PMF-a, 2016.

[3] G. M. Cooper i R. E. Hausman, *The cell: A molecular approach*, 7th edition, Sinauer Associates, Oxford, 2015.

[4] R. Durbin, S. Eddy, A. Krogh i G. Mitchison, *Biological sequence analysis*, Cambridge University Press, 1998.

[5] National Cancer Institute, *NCI Dictionary of Cancer Terms*, `https://www.cancer.gov/publications/dictionaries/cancer-terms/def/translation`.

[6] F. Janjić, *Semantičko indeksiranje i klasifikacija dokumenata*, Diplomski rad, Sveučilište u Zagrebu, Matematički odsjek PMF-a, 2019.

[7] M. Kobovac, *Neki aspekti iterativnog pretraživanja proteoma*, Diplomski rad, Sveučilište u Zagrebu, Matematički odsjek PMF-a, 2017.

[8] S. Mavrek, *Iterativno traženje fraza i statistika semantičkog indeksiranja*, Diplomski rad, Sveučilište u Zagrebu, Matematički odsjek PMF-a, 2019.

[9] Lumen Microbiology, *Proteins*, `https://courses.lumenlearning.com/microbiology/chapter/proteins`.

[10] Pfam, `https://pfam.xfam.org`.

[11] G. M. Salton i C. Buckley, *Term-weighting approaches in automatic text retrieval*, In Information Processing and Management, Volume 24, Issue 5, 1988.

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

[13] Wikipedia, *Protein biosynthesis*, https://en.wikipedia.org/wiki/Protein_biosynthesis.

[14] A. Đurić, *Analiza tehnika traženja proteinskih motiva*, Diplomski rad, Sveučilište u Zagrebu, Matematički odsjek PMF-a, 2018.

Sažetak

Razvojem bioinformatike i otkrivanjem novih proteina javlja se potreba za razvojem novih metoda klasifikacije proteina. U ovom diplomskom radu promatrali smo kraće nizove aminokiselina s karakterističnim mutacijama koje nazivamo motivi. Opisali smo postupak traženja karakterističnih motiva za neku proteinsku familiju. Posebno je važno istaknuti da traženi motivi nisu fiksne duljine. Zatim smo opisali način ocjenjivanja proteina u odnosu na profile dobivenih karakterističnih motiva pomoću kojeg provodimo klasifikaciju proteina u proteinske familije. Kako bi provjerili točnost našeg algoritma proveli smo postupak traženja karakterističnih motiva na dvije proteinske familije, lipoproteinskim lipazama i Walkerovim motivima. Klasifikacijom proteina iz test setova obje familije dobiveni su iznenađujuće dobri rezultati.

Summary

With the advancement of bioinformatics and the discovery of new proteins, a need for new protein classification methods has arisen. This master's thesis focuses on the short sequences of amino acids with characteristic mutations called motifs. We describe a process of characteristic motif discovery for a given protein family. It is worth mentioning that the observed motifs are of variable length. Further, we describe the evaluation of proteins with respect to the profiles of the discovered characteristic motifs. Based on this, we classify the proteins in protein families. In order to validate the algorithm, we apply it to two protein families: lipoprotein lipase family and Walker motifs. The results of the protein classification for the test sets from both families are surprisingly good.

Životopis

Rođena sam 4. travnja 1995. godine u Slavonskom Brodu. Svoje školovanje započinjem 2002. godine u Osnovnoj školi "Antun Mihanović" u Slavonskom Brodu te ga nastavljam 2010. godine u istom mjestu u Prirodoslovno-matematičkoj gimnaziji "Matija Mesić". Po završetku srednjoškolskog obrazovanja, 2014. godine upisujem Preddiplomski sveučilišni studij Matematika na Prirodoslovno-matematičkom fakultetu Sveučilišta u Zagrebu. Preddiplomski studij završavam 2017. godine i stječem akademski naziv sveučilišne prvostupnice matematike. Iste godine upisujem Diplomski sveučilišni studij Matematička statistika na istom fakultetu.