Iterativno traženje motiva i vreća fraza u genomu i proteomu

Ramov, Laura

Master's thesis / Diplomski rad

2021

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

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:217:862542

Rights / Prava: In copyright / Zaštićeno autorskim pravom.

Download date / Datum preuzimanja: **2024-03-29**

Repository / Repozitorij:

Repository of the Faculty of Science - University of Zagreb

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Laura Ramov

ITERATIVNO TRAŽENJE MOTIVA I VREĆA FRAZA U GENOMU I PROTEOMU

Diplomski rad

Voditelj rada:
doc. dr. sc. Pavle Goldstein

Zagreb, veljača, 2021.

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

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

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Zahvaljujem mentoru doc. dr. sc. Pavlu Goldsteinu na iskazanom strpljenju, motivaciji i vođenju pri izradi ovog diplomskog rada.
Hvala mojoj obitelji i dečku na podršci koju su mi pružili tijekom cijelog studiranja, a posebno mojoj majci na svim toplim riječima i ohrabrenju.

Sadržaj

Sadržaj **iv**

Uvod **1**

1 Matematički pojmovi **2**
1.1 Teorija grafova 2
1.2 Linearna algebra 3
1.3 Vjerojatnost i statistika 5

2 Biološki pojmovi **7**

3 Reducirani alfabet **9**
3.1 Stvaranje alfabeta 9
3.2 Pretraživanje 6-grama u reduciranom alfabetu 12
3.3 Određivanje motiva 13
3.4 Profil motiva 15
3.5 Računanje distribucije i vjerojatnosti motiva 16
3.6 Računanje norme motiva s proteinima familije 17
3.7 Klasifikacija 17
3.8 Rezultati i usporedba 18

4 All against all **22**
4.1 Određivanje motiva 22
4.2 Rezultati i usporedba 24

5 Zaključak **30**

Bibliografija **31**

Uvod

Razvojem interneta i napretkom tehnologije omogućila se pohrana, obrada i analiza velike količine podataka. Ujedno došlo je i do napretka u biološkim područjima kao što su biotehnologija, molekularna biologija i genetičko iznženjerstvo koji su povećali potrebu za statističkom obradom i analizom bioloških podataka u znanstvenim istraživanjima. To je dovelo do razvoja bioinformatike.

Bioinformatika je interdisciplinarna znanost koja se bavi istraživanjem genetičkih i drugih bioloških informacija uz pomoć računalne tehnologije, statistike i primijenjene matematike.

Jedno od važnijih pitanja u bioinformatici pitanje je pripadnosti proteina nekoj proteinskoj familiji. Proteini su biološke makromolekule građene od aminokiselina povezanih peptidnom vezom. Grade sva živa bića te sudjeluju u raznim procesima potrebnima za život. Proteinske familije skupine su evolucijski povezanih proteina. Pripadnost istoj proteinskoj familiji najjasnije pokazuje sličnost njihovih kraćih nizova aminokiselina. Budući da evolucijom dolazi do promjena u proteinima živih bića, u ovom diplomskom radu promatramo kraće nizove (duljine 6-15) aminokiselina s karakterističnim mutacijama koje nazivamo motivi. Opisujemo i testiramo algoritam za traženje karakterističnih motiva neke proteinske familije te na temelju toga radimo klasifikaciju proteina. Uspoređujemo dvije provedene metode, metodu reduciranog alfabeta te metodu all against all s metodom iz [4].

Poglavlje 1

Matematički pojmovi

1.1 Teorija grafova

Definicija 1.1.1. *Neka je $A \neq \emptyset$ proizvoljan skup. Kažemo da je familija skupova $\mathcal{F}$ particija skupa A ako vrijedi:*

a) za svaki $x \in \mathcal{F}$ je $x \subseteq A$;

b) za svaki $x \in \mathcal{F}$ je $x \neq \emptyset$;

c) za sve $x, y \in \mathcal{F}$, $x \neq y$, vrijedi $x \cap y = \emptyset$;

d) $\bigcup_{x \in \mathcal{F}} x = A$.

Definicija 1.1.2. *Graf G je uređeni par vrhova (V, E), gdje je V skup vrhova, a E skup 2-podskupova od V, koje zovemo bridovi.*

Napomena 1.1.3. *Katkada gornju definiciju proširujemo tako da dopustimo petlje (bridove koje spajaju vrh sa samim sobom), višestruke bridove (više bridova između para vrhova) i usmjerene bridove (bridovi koji imaju orijentaciju tako da idu od jednog vrha prema drugome). Usmjereni bridovi se reprezentiraju uređenim parovima, a ne 2-podskupovima, dok kod višestrukih bridova E postaje multiskup.*

Definicija 1.1.4. *Graf koji ima usmjerene bridove zvat ćemo usmjereni graf ili digraf.*

Definicija 1.1.5. *Put od v_0 do v_n u grafu $G = (V, E)$ je niz $(v_0, e_1, v_1, e_2, v_2, \ldots, e_n, v_n)$,gdje su vrhovi različiti (osim eventualno prvog i zadnjeg vrha), a e_i je brid $\{v_{i-1}, v_i\}$, za $i = 1, 2, \ldots, n$.*

Definicija 1.1.6. *Relacija ekvivalencije $\equiv$ na skupu vrhova V grafa $G = (V, E)$:*
vrhovi x i y su u relaciji, odnosno $x \equiv y$, ako postoji put u grafu od x do y.

Definicija 1.1.7. *Komponenta povezanosti grafa $G = (V, E)$ je podgraf induciran klasom ekvivalencije gore definirane relacije ekvivalencije $\equiv$.*

Definicija 1.1.8. *Kažemo da je graf $G = (V, E)$ povezan ako postoji samo jedna komponenta povezanosti.*

1.2 Linearna algebra

Definicija 1.2.1. *Neka je V neprazan skup i neka je $\mathbb{F}$ polje. Neka su zadane operacije zbrajanja vektora*

$$+ : V \times V \to V$$

i operacija množenja vektora skalarom

$$\cdot : F \times V \to V.$$

Kažemo da je uređena trojka $(V, +, \cdot)$ vektorski prostor nad poljem $\mathbb{F}$ ako vrijedi:

(1) $(x + y) + z = x + (y + z), \quad \forall x, y, z \in V,$

(2) $\exists 0 \in V, \quad 0 + x = x + 0 = x, \quad \forall x \in V,$

(3) $\forall x \in V, \exists -x \in V, \quad (-x) + x = x + (-x) = 0,$

(4) $x + y = y + x, \quad \forall x, y \in V,$

(5) $\alpha(\beta x) = (\alpha\beta)x, \quad \forall \alpha, \beta \in \mathbb{F}, \forall x \in V,$

(6) $(\alpha + \beta)x = \alpha x + \beta x, \quad \forall \alpha, \beta \in \mathbb{F}, \forall x \in V,$

(7) $\alpha(x + y) = \alpha x + \alpha y, \quad \forall \alpha \in \mathbb{F}, \forall x, y \in V,$

(8) $\exists 1 \in \mathbb{F}, 1 \cdot x = x, \quad \forall x \in V.$

U tom slučaju elemente skupa V zovemo vektori, dok elemente polja $\mathbb{F}$ zovemo skalari. Skup $\mathbb{R}^n = \{(x_1, \ldots, x_n) : x_1, \ldots, x_n \in \mathbb{R}\}$ s operacijama

$$(x_1, \ldots, x_n) + (y_1, \ldots, y_n) = (x_1 + y_1, \ldots, x_n + y_n)$$
$$\alpha(x_1, \ldots, x_n) = (\alpha x_1, \ldots, \alpha x_n), \alpha \in \mathbb{R}$$

je vektorski prostor nad $\mathbb{R}$.

Definicija 1.2.2. *Neka je $(V, +, \cdot)$ vektorski prostor nad poljem $\mathbb{F}$. Skalarni produkt na V je preslikavanje $\langle \cdot , \cdot \rangle : V \times V \to \mathbb{F}$ sa sljedećim svojstvima:*

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

(2) $\langle x, x\rangle = 0 \Longleftrightarrow x = 0,$

(3) $\langle x + y, z\rangle = \langle x, z\rangle + \langle y, z\rangle, \forall x, y, z \in V,$

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

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

U tom slučaju vektorski prostor $(V, +, \cdot)$ *zajedno sa skalarnim produktom zovemo unitarni prostor.*

Euklidski skalarni produkt na vektorskom prostoru $\mathbb{R}^n$ zadan je s

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

Definicija 1.2.3. *Neka je* V *vektorski prostor nad poljem* $\mathbb{F}$. *Funkciju* $\|\cdot\| : V \to R$ *koja zadovoljava svojstva:*

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

(2) $\|x\| = 0 \Longleftrightarrow x = 0,$

(3) $\|\alpha x\| = |\alpha| \, \|x\|, \quad \forall x, y \in V,$

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

zovemo norma na V.

Propozicija 1.2.4. *Neka je* V *unitarni prostor. Funkcija* $\|\cdot\| : V \to \mathbb{R}$ *definirana s*

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

je norma.

Euklidska norma inducirana euklidskim skalarnim produktom na $\mathbb{R}^n$ zadana je s

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

1.3 Vjerojatnost i statistika

Definicija 1.3.1. *Pod **slučajnim pokusom** podrazumijevamo takav pokus čiji **ishodi**, odnosno **rezultati** nisu jednoznačno određeni uvjetima u kojima izvodimo pokus. Rezultate slučajnog pokusa nazivamo **događajima**.*

Definicija 1.3.2. *Neka je A događaj vezan uz neki slučajni pokus. Pretpostavimo da smo taj pokus ponovili n puta i da se u tih n ponavljanja događaj A pojavio točno n_A puta. Tada broj n_A zovemo **frekvencija dogadaja** A, a broj $\frac{n_A}{n}$ **relativna frekvencija događaja** A.*

Definicija 1.3.3. *Neka je $(\Omega, \mathcal{F})$ izmjeriv prostor. Funkcija $\mathbb{P} : \mathbb{F} \to \mathbb{R}$ jest vjerojatnost ako vrijedi*

- $\mathbb{P}(\Omega) = 1$
- $\mathbb{P}(A) \geq 0, \quad A \in \mathcal{F}$
- $A_i \in \mathcal{F}_i, i \in \mathbb{N} \quad A_i \cap A_j = \emptyset, i \neq j \Rightarrow \mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i).$

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

Definicija 1.3.4. *Neka je $(\Omega, \mathcal{F}, \mathbb{P})$ proizvoljan vjerojatnosni prostor i $A \in \mathcal{F}$ takav da je $\mathbb{P}(A) > 0$. Definiramo funkciju $\mathbb{P}_A : \mathcal{F} \to [0, 1]$ s*

$$\mathbb{P}_A(B) = \mathbb{P}(B|A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}, B \in \mathcal{F}$$

*Lako je provjeriti da je $\mathbb{P}_A$ vjerojatnost na $\mathcal{F}$ i nju zovemo **vjerojatnost od B uz uvjet A**.*

Definicija 1.3.5. *Neka je $(\Omega, \mathcal{F}, \mathbb{P})$ vjerojatnosni prostor. Funkcija $X : \Omega \to \mathbb{R}$ je **slučajna varijabla** (na Ω) ako je $X^{-1}(B) \in F$ za proizvoljno $B \in \mathcal{B}$, odnosno $X^{-1}(B) \subset F$.*

Definicija 1.3.6. *Neka je X slučajna varijabla na Ω. **Funkcija distribucije** od X je funkcija $F_X = F : \mathbb{R} \to [0, 1]$ definirana s*

$$F(x) = \mathbb{P}\{X \leq x\} = \mathbb{P}\{\omega : X(\omega) \leq x\}, x \in \mathbb{R}.$$

Definicija 1.3.7. *Slučajna varijabla X je diskretna ako postoji konačan ili prebrojiv skup $D \subset \mathbb{R}$ takav da je $\mathbb{P}\{X \in D\} = 1$.*

Definicija 1.3.8. *Neka je $(\Omega, \mathcal{F}, \mathbb{P})$ vjerojatnostni prostor i $X : \Omega \to \mathbb{R}^n$. Kažemo da je X **n-dimenzionalan slučajni vektor** (ili, kraće, **slučajan vektor**) (na Ω) ako je $X^{-1}(B) \in \mathcal{F}$ za svako $B \in \mathcal{B}^n$, tj. $X^{-1}(\mathcal{B}^n) \subset \mathcal{F}$.*

Propozicija 1.3.9. *Neka je* $X : \Omega \to \mathbb{R}^n$, $X = (X_1, X_2, \ldots, X_n)$. *Tada je* X *slučajni vektor ako i samo ako je* X_k *slučajna varijabla za svaki* $k = 1, \ldots, n$.

Pojmovi ovog poglavlja preuzeti su iz [1], [3], [5], [6].

Poglavlje 2

Biološki pojmovi

Proteini su biološke makromolekule sastavljene od aminokiselina. Aminokiselinski niz proteina određuje njegovu trodimenzionalnu strukturu i funkciju. U tablici 2.1 dan je popis standardnih aminokiselina.

Motiv proteinske familije kratak je niz aminokiselina, u pravilu 5 do 20, koji je ostao djelomično sačuvan selekcijskim pročišćavanjem ili evolucijom i ima neko biološko značenje. U principu, motiv prepoznajemo time što ima specifičan supstitucijski uzorak.

Niz od n aminokiselina nazivamo **n-gramom**.

Tablica 2.1: Standardne aminokiseline

Oznaka	Naziv	Oznaka	Naziv
A	Alanin	M	Metionin
C	Cistein	N	Asparagin
D	Asparginska kiselina	P	Prolin
E	Glutaminska kiselina	Q	Glutamin
F	Fenilalanin	R	Arganin
G	Glicin	S	Serin
H	Histidin	T	Treonin
I	Izoleucin	V	Valin
K	Lizin	W	Triptofan
L	Leucin	Y	Tirozin

Definicija 2.0.1. *Blosum matrica B je* 20×20 *matrica,* $B = (b_{ij}) \in M_20(\mathbb{Z})$*, koja na* (i, j)*-tom mjestu sadrži koeficijente sličnosti i-te i j-te aminokiseline. Ukratko, bazirana je na sljedećoj formuli:*

$$B(i, j) = \left\lfloor log \frac{\mathbb{P}(a_i \leftrightarrow b_j \mid M)}{\mathbb{P}(a_i, b_j \mid R)} \right\rfloor, a_i, b_j \in \mathcal{A} \tag{2.1}$$

gdje su a_i *i* b_j *aminokiseline pridružene respektivno, i-tom i j-tom mjestu, a* $\mathcal{A}$ *je skup svih standardnih aminokiselina. M je model koji pretpostavlja da aminokiseline* a_i *i* b_j *imaju zajedničkog pretka, a R je random model koji pretpostavlja nezavisnost aminokiselina.*

Neki pojmovi ovog poglavlja preuzeti su iz [3] i [7] .

Proteinske familije u ovom radu

Na početku promatramo problem klasifikacije dvije familije proteina, lipoproteinske lipaze i Walkerove motive. Lipoproteinske lipaze su enzimi koji ubrzavaju kemijsku reakciju razgradnje masti u doticaju s vodom, a Walkerovi motivi imaju ulogu u vezanju molekula ATP-a. Lipoproteinske lipaze u ovom radu zvat ćemo skraćeno lipaze ili familija lipaza.

Zatim promatramo klasifikaciju četiri proteinske familije. Cytochrome bc1 complex koji se nalazi u stanicama svih životinja i aerobnih eukariota, a ima ulogu u transportu elektrona. Radi jednostavnosti, nju ćemo zvati prvom familijom ili familijom 1. Familiju proteina Upf2 iz kvasca Saccharomyces cerevisiae koji sudjeluju u procesu nadziranja i ispravljanja mRNA zovemo drugom familijom ili familijom 2. Nadalje, Endoribonukleaze XendoU koje se nalaze u vrsti žaba Xenopus, a imaju ulogu u sintezi RNA predstavljat će našu treću familiju, odnosno familiju 3. Posljednja familija koju promatramo su peptidaze koje se nalaze u zivotinjama, biljkama, bakterijama i virusima, a kataliziraju proces razgradnje proteina na manje gradivne jedinice. Zvat ćemo je četvrtom familijom ili familijom 4.

Poglavlje 3

Reducirani alfabet

Reducirani alfabet definiramo tako da skup od 20 slova postojećeg alfabeta (koji predstavlja 20 standardnih aminokiselina) particioniramo u n skupova, $n \in \{1, 2, \ldots, 19\}$, te za svaki skup odaberemo njegovog reprezentanta. Kada proteinsku familiju želimo prepisati u reducirani alfabet, za svako slovo postojećeg alfbeta trebamo pogledati u kojem se skupu particije nalazi te ga zamijeniti reprezentantom tog skupa. Na taj način familiju koja je prethodno bila zapisana s 20 slova, zapišemo pomoću njih n.

Cilj korištenja reduciranog alfabeta bio je istražiti može li se pronalazak karakterističnih motiva familije pojednostaviti kada je ona zapisana s manje slova.

3.1 Stvaranje alfabeta

Htjeli smo stvoriti particiju za koju će vrijediti da elementi unutar svakog skupa budu slova koja označavaju proteine slične po nekom kriteriju. Kao što smo rekli u poglavlju 2, vrijednosti blosum50 matrice koeficijenti su sličnosti između dvije aminokiseline te ćemo upravo nju koristiti za definiranje naše particije.

Koeficijenti u matrici blosum50 su cijeli brojevi između -5 i 15. Mi ćemo tu matricu modificirati tako da imamo samo 2 broja koja označavaju 2 stanja, 0 - nisu slični i 1 - slični. To smo napravili na način da smo njene elemente strogo veće od 1 zamijenili s 1, a ostale pretvorili u nule. Novu modificiranu matricu nazvali smo B (Slika 3.1).

Slika 3.1: Modificirana blosum50 matrica

$$B = \begin{array}{c} \\ A \\ R \\ N \\ D \\ C \\ Q \\ E \\ G \\ H \\ I \\ L \\ K \\ M \\ F \\ P \\ S \\ T \\ W \\ Y \\ V \end{array} \begin{array}{c} \begin{array}{cccccccccccccccccccc} A & R & N & D & C & Q & E & G & H & I & L & K & M & F & P & S & T & W & Y & V \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \end{array}$$

Neka je $G = (V, E)$ graf gdje nam je skup vrhova V jednak skupu 20 standardnih aminokiselina $\mathcal{A}$, a bridovi E iščitani iz modificirane blosum50 matrice $B = [b_{ij}]$ na način:

$$b_{ij} > 0 \Rightarrow (A(i), A(j)) \in E,$$

gdje je A=[A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V] vektor 20 standardnih aminokiselina.

Particiju reduciranog alfabeta definiramo kao komponente povezanosti grafa G. Na taj način osiguravamo sličnost aminokiselina unutar skupova particije.

Matrica B nam svojim pozitivnim vrijednostima označava postojanje brida između vrhova. Matrica B^2 svojim pozitivnim vrijednostima označava između kojih vrhova postoji put duljine 2. Primijetimo, u matrici B^2 će biti sadržane i sve pozitivne vrijednosti iz B jer ako je postojao brid između 2 vrha, tada postoji put duljine 2 koji ide iz prvog vrha u prvi vrh te onda u drugi vrh.

Kako bismo dobili komponente povezanosti grafa G, matricu B dižemo na 20. potenciju jer će nam ona pokazati sve postojeće puteve iz matrice B. Tako dobivamo matricu $N = [n_{ij}]$ za koju vrijedi:

$$n_{ij} \geq 0 \Rightarrow \textit{postoji put između } A(i) \textit{ i } A(j).$$

Kako bismo lakše iščitali puteve, sve pozitivne vrijednosti matrice N preslikali smo u 1, a ostale u 0. Tako smo dobili matricu M (Slika 3.2) iz koje lako možemo očitati 8 komponenti povezanosti jer su redovi matrice isti za one vrhove koji su u istoj komponenti povezanosti.

Slika 3.2: Matrica M

$$M = \begin{array}{c} \\ A \\ R \\ N \\ D \\ C \\ Q \\ E \\ G \\ H \\ I \\ L \\ K \\ M \\ F \\ P \\ S \\ T \\ W \\ Y \\ V \end{array} \begin{array}{c} \begin{array}{ccccccccccccccccccccc} A & R & N & D & C & Q & E & G & H & I & L & K & M & F & P & S & T & W & Y & V \end{array} \\ \begin{bmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 1 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{bmatrix} \end{array}$$

Dobivena particija za reducirani alfabet:

1. A
2. R, N, D, Q, E, K
3. C
4. G
5. H, F, W, Y
6. I, L, M, V
7. P
8. S, T

Odabrani reprezentani prva su slova u prethodnom zapisu.

3.2 Pretraživanje 6-grama u reduciranom alfabetu

Familiju proteina (čije motive želimo dobiti) zapišemo u našem reduciranom alfabetu, tj. pomoću 8 različitih slova (A, R, C, G, H, I, P, S). Nakon toga nađemo frekvencije svih mogućih 6-grama (nizova od 6 aminokiselina). Izbacimo niskofrekventne 6-grame iz razmatranja te nastavljamo s najčešćih $\frac{1}{6} \times$ duljina familije. S obzirom da sve naše trening familije imaju po 300 proteina, to je značilo da smo gledali najfrekventnijih 50 6-grama.

Slika 3.3: Primjer 6-grama s frekvencijama

Frekvencija	Ime
130	PIGCIP
130	GRHSRG
127	SGRHSR
124	RRIRRI
121	IRRIRR
112	RRHIRR
101	RHIHHR
100	RIRRII
97	RHSRGR
94	RIIIRH
94	RHIRRH
94	IIIRHI

6-grame označimo sa S_i, $i \in \{1, \ldots, 50\}$, a njihove frekvencije s F_i, $i \in \{1, \ldots, 50\}$.

Pretragom 6-grama u reduciranom alfabetu dobili smo osnovu od koje ćemo napraviti motiv, jer se na jednoj poziciji mogu pojavljivati samo dovoljno slične aminokiseline. Za očekivati je da, ako pravi motiv ima više mogućih aminokiselina na jednoj poziciji, one budu slične po nekom kriteriju.

3.3 Određivanje motiva

Postupak započinjemo tako da svaki 6-gram S_i, $i \in \{1, \ldots, 50\}$ vratimo u standardni alfabet pomoću funkcije *izrada*, koja kaže da se svako slovo 6-grama može zamijeniti bilo kojim slovom iz pripadnog skupa particije reduciranog alfabeta.

Slika 3.4: Funkcija *izrada*

```
def izrada(string):
    kon=""
    for slo in string:
        if slo=='A':
            kon=kon+'A'
        if slo=='R':
            kon=kon+'[R|D|N|Q|E|K]'
        if slo=='C':
            kon=kon+'C'
        if slo=='G':
            kon=kon+'G'
        if slo=='P':
            kon=kon+'P'
        if slo=='I':
            kon=kon+'[I|L|M|V]'
        if slo=='H':
            kon=kon+'[H|F|W|Y]'
        if slo=='S':
            kon=kon+'[S|T]'
    return(kon)
```

$$M_i := izrada(S_i), i \in \{1, 2, \ldots, 50\}$$

Za ilustraciju, u reduciranom alfabetu 6-gram $S_k = PIGCIP$, $k \in \{1, \ldots, 50\}$ *sada bio bi*

$$M_k = P[I|L|M|V]GC[I|L|M|V]P,$$

gdje okomite linije označavaju riječ „ili”.

Za svaki $i \in \{1, \ldots, 50\}$, u pripadnoj familiji pretražimo sva pojavljivanja od M_i te ih spremimo ih u listu. Zatim za svaku od 6 pozicija 6-grama prebrojimo koliko se puta koja aminokiselina ponavlja na njoj u elementima te liste. Tako smo dobili frekvencije mogućih aminokiselina na svakoj poziciji pa spremimo rezultate u varijablu $frekv_i$.

Koristeći naš prethodni ilustrativni primjer, recimo da se 6-gram P[I|L|M|V]GC[I|L|M|V]P pojavio 10 puta u familiji te da je lista motiva jednaka [PVGCLP, PIGCMP, PIGCLP, PIGCVP, PIGCVP, PIGCVP, PIGCVP, PIGCVP, PIGCVP, PIGCVP]. Tada imamo $frekv_k$*= [(P:10),(I:9, V:1),(G:10),(C:10),(V:7, L:2, M:1),(P:10)]*

Frekvencija nam otkriva postoje li aminokiseline koje se pojavljuju proporcionalno malo (s obzirom na broj ukupnih pojavljivanja tog 6-grama u familiji) na određenoj poziciji, te zbog toga ne bi trebale biti uključene kao varijacija potencijalnog motiva.

Kako bismo sa svake pozicije izbacili slabo frekventne aminokiseline iskoristimo funkciju *skrati* (Slika 3.5). Ona radi tako da čita varijablu $frekv_i$ te izbaci sve one aminokiseline koje se na svojoj poziciji pojavljuju manje ili jednako $0.1 \times F_i$ puta, s tim da je F_i frekvencija, odnosno broj pojavljivanja 6-grama M_i.

Slika 3.5: Funkcija *skrati*

```
def skrati(motiv,frekv):
    stvoreni=''

    for i in range(len(frekv)):
        slova=list(set(frekv[i]))
        novi='['
        if frekv[i][slova[0]]>0.1*np.size(motiv):
            novi=novi+slova[0]
        for j in range(1,len(slova)):
            if frekv[i][slova[j]]>0.1*np.size(motiv):
                novi=novi+'|'+slova[j]
        novi=novi+']'
        stvoreni=stvoreni+novi
    return(stvoreni)
```

$$M_i = skrati(M_i, frekv_i)$$

Na našem primjeru izbacili bismo V s druge pozicije te M s pete pozicije, jer su im frekvencije 1 što je jednako $0.1 \times F_k = 0.1 \times 10$. *Tako bismo dobili novi* $M_k = P[I|L|M]GC[I|L|V]P$.

Pretražimo koliko sada imamo pojavljivanja promijenjenog M_i u familiji te zapišemo novi broj pojavljivanja F_i.

Nastavljamo istim postupkom skraćivanja sve dok više nemamo frekvencija aminokiselina manjih od $0.1 \times F_i$ te se M_i uopće ne promijeni nakon upotrebe funkcije *skrati*. Tada je metoda konvergirala i imamo potpuno skraćeni 6-gram, onaj u kojem se svaka aminokiselina na svojoj poziciji pojavljuje u više od 10% pojavljivanja.

Naš prethodni primjer nastavio bi se s pretragama novog $M_k = P[I|L|M]GC[I|L|V]P$. *Njegova lista pojavljivanja bila bi [PIGCLP, PIGCVP, PIGCVP, PIGCVP, PIGCVP, PIGCVP, PIGCVP, PIGCVP], a* $frekv_k = [(P : 8), (I : 8), (G : 8), (C : 8), (V : 7, L : 1), (P : 8)]$. *To bio bi potpuno skraćeni 6-gram jer nemamo aminokiselina u* $frekv_k$ *koje se pojavljuju manje od* $0.1 \times F_i = 0.1 \times 8 = 0.8$ *puta.*

Sada prebrojimo koliko se puta novonastali skraćeni 6-grami pojavljuju u cijeloj familiji te uzmemo samo najfrekventnijih $q := 10$. Na taj način odabrali smo 10 motiva pripadne proteinske familije. Dakle, naši motivi izabrani su na način da se dovoljno često pojavljuju u proteinskoj familiji te da na svakoj poziciji imaju dovoljno slične aminokiseline.

3.4 Profil motiva

Profil motiva dobijemo dodavajući dovoljno dobre n-grame (u ovom slučaju n=6) iz pripadne proteinske familije na prethodno napravljenu listu pojavljivanja tog motiva.

Kako bismo odredili koji je od n-grama „dovoljno dobar" kao jedna od varijanti (mutacija) zadanog motiva morali smo definirati ocjenu sličnosti, odnosno odrediti način dodjeljivanja određenog score-a tom n-gramu u usporedbi sa zadanim motivom.

Ocjenu sličnosti računali smo koristeći metodu klizećeg prozora (grafički prikaz dolje). Dakle, po svakom proteinu u familiji kliže prozor koji obuhvaća jedan n-gram te ga uspoređuje s našim motivom. Usporedbu vrši gledanjem imaju li n-gram i motiv istu aminokiselinu na jednakoj poziciji.

Ocjena sličnosti n-grama i motiva je broj između 0 i n koji označava na koliko pozicija se nalazi ista aminokiselina.

Tako bi npr. ocjena sličnosti između PIGCIP i PAGCIM bila 4 jer imaju 4 iste aminokiseline na jednakoj poziciji.

$$\begin{array}{cccccc} P & I & G & C & I & P \\ | & & | & | & | & \\ P & A & G & C & I & M \end{array}$$

Grafički prikaz metode klizećeg prozora:
Neka je $y_1 y_2 \dots y_n$ naš motiv, a $x_1 x_2 \dots x_k$ protein familije, $n < k$

$$
\begin{array}{cccccccccc}
x_1 & x_2 & x_3 & \dots & x_{n-1} & x_n & x_{n+1} & x_{n+2} & \dots & x_k \\
y_1 & y_2 & y_3 & \dots & y_{n-1} & y_n & & & & \\
\\
x_1 & x_2 & x_3 & \dots & x_n & x_{n+1} & x_{n+2} & x_{n+3} & \dots & x_k \\
 & y_1 & y_2 & \dots & y_{n-1} & y_n & & & & \\
\\
x_1 & x_2 & x_3 & \dots & x_{n+1} & x_{n+2} & x_{n+3} & x_{n+4} & \dots & x_k \\
 & & y_1 & \dots & y_{n-1} & y_n & & & & \\
 & & & & \vdots & & & & & \\
x_1 & x_2 & x_3 & \dots & x_{k-n+1} & x_{k-n+2} & x_{k-n+3} & & \dots & x_k \\
 & & & & y_1 & y_2 & y_3 & & \dots & y_n
\end{array}
$$

Na kraju imamo listu svih pojavljivanja motiva te njihovih dovoljno sličnih varijanti. Označimo duljinu te liste, tj. broj elemenata liste s m. Elemente te liste koristit ćemo dalje za izračun distribucije i vjerojatnosti motiva.

3.5 Računanje distribucije i vjerojatnosti motiva

Motiv gledamo kao slučajni vektor, čije su sve slučajne varijable diskretne s prostorom događaja od 20 slova, za 20 standardnih aminokiselina. Tada možemo izračunati njegovu distribuciju.

Distribuciju motiva definirali smo pomoću matrice PSSM (eng. Position Specific Scoring Matrix) koja se koristi za prikaz razdiobe aminokiselina na svakoj poziciji motiva. Neka je A=[A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V] vektor 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 izračunamo relativnu frekvenciju pojavljivanja svake od 20 standardnih aminokiselina iz A. Time dobijemo niz vektora distribucija $f_k = (f_{k1}, \dots, f_{k20})$, $k = 1, 2, \dots, n$. Matrica $F = [f_{kj}]$, $k = 1, 2, \dots, n$, $j = 1, 2, \dots, 20$ predstavlja distribuciju zadanog motiva. Zatim relativnim frekvencijama dodamo mali broj, u našem slučaju 0.01, te ih sve podijelimo s 1.2 kako bi im zbroj i dalje

bio 1. Time dobijemo da je vjerojatnost pojave svake aminokiseline iz A na svakoj poziciji motiva veća od 0.

$$p_{kj} = \frac{f_{kj} + 0.01}{1.2} \tag{3.1}$$

Vektor $p_k = (p_{k1}, p_{k2}, \ldots, p_{k20})$, za $k \in \{1, 2, \ldots, n\}$ označava vjerojatnost pojave aminokiselina iz A na k-toj poziciji motiva, a matrica $P = [p_{kj}]$ predstavlja profil zadanog motiva.

3.6 Računanje norme motiva s proteinima familije

Za svaki motiv M_i, $i \in \{1, \ldots, q\}$, te njegovu vjerojatnost $p_k^{(i)}$ (dobivenu kao u prethodnom potpoglavlju) metodom klizećeg prozora krećemo se po proteinu neke familije. Onaj n-gram koji je obuhvaćen prozorom označimo s x. Izračunamo vjerojatnost svakog n-grama x od tog proteina uz zadanu distribuciju,

$$v^i = \sum_{k=1}^{n} \mathbb{P}(x \mid p_k^{(i)}). \tag{3.2}$$

Definirajmo V_i kao maksimalan v^i od svih n-grama tog proteina te pripadni vektor $V = [V_i]$, $i \in 1, ..., q$ u kojem su maksimalne vjerojatnosti za svaki motiv. Sada izračunamo euklidsku normu vektora V te na taj način dobivamo brojčanu vrijednost koja nam govori koliko dobro određenih q motiva opisuje protein familije s kojom se uspoređuje.

3.7 Klasifikacija

Sad kad imamo metodu računanja norme karakterističnih motiva i proteina, lako je napraviti klasifikaciju. Sve što trebamo je za neki protein izračunati norme karakterističnih motiva svake familije koje želimo ispitati. Tada protein klasificiramo u onu familiju čija je norma karakterističnih motiva s njim bila najveća.

Dakle, ako imamo jedan protein te tri familije u koje ga pokušavamo klasificirati, trebamo izračunati normu proteina s karakterističnim motivima prve familije, normu proteina s karakterističnim motivima druge familije te normu proteina s karakterističnim motivima treće familije. Norme se, naravno, računaju kao u prethodnom potpoglavlju. Nađemo koja je od te tri norme najveća te protein klasificiramo upravo u tu familiju čiji su karakteristični motivi postigli najveću normu s proteinom. Na taj način smo dobili rezultate.

3.8 Rezultati i usporedba

Za svaku familiju imali smo po 300 podataka za trening te 100 podataka za testiranje. Dakle, motive bismo dobili na temelju tih 300 podataka, a ispitali smo koliko dobro model radi na 100 podataka svake familije. Naše rezultate usporedili smo s rezultatima Silvestra Mavreka u radu Iterativno traženje fraza i statistika semantičkog indeksiranja (vidi [4]).

Lipaze i Walker

Dobiveni motivi za familiju lipaza:

$$
\begin{aligned}
[&'[P][L|M|I|V][G][C][L|I|V][P]',\\
&'[G][R][F][T|S][D|N][G]',\\
&'[T][G][R][F][T|S][D|N]',\\
&'[R][Y|F][T|S][D|N][G][K|R]',\\
&'[P][S|T][G][R][Y|F][T|S]',\\
&'[I|V][F][G][D][S][L|I|V]',\\
&'[G][R][L|V][I|V][I|V][D]',\\
&'[G][N][N][D|N][F|Y][L|I]',\\
&'[D|N][G][R][L|V][I|V][I|V]',\\
&'[Y|F][I|V][F][G][D][S]']
\end{aligned}
$$

Dobiveni motivi za familiju walker:

$$
\begin{aligned}
[&'[D|K|E|N|R][R|K|D|E][K|D|R|E][K|D|R|E][K|D|R|E][D|K|E|N|R]',\\
&'[D|K|E|N|R][D|K|E|R|Q][V|I|L][D|K|E|R|Q][N|K|R|E][I|V|L|M]',\\
&'[I|V|L][D|K|E|R|Q][D|K|E|N|R|Q][D|K|E|R|Q][D|K|E|N|R|Q][D|K|E|R|Q]',\\
&'[D|K|E|R|Q][D|K|E|R|Q][D|K|E|R|Q][D|K|E|R|Q][K|D|R|E][I|V|L|M]',\\
&'[D|K|E|R|Q][I|V|L][D|K|E|N|R|Q][D|K|E|R|Q][D|K|E|R|Q][D|K|E|R|Q]',\\
&'[D|K|E|N|R][D|K|E|N|R][K|D|R|E][L|V|I|M][R|K|D|E][N|K|R|E]',\\
&'[D|K|E|N|R][D|K|E|N|R][L|V|I|M][N|K|R|E][N|K|R|E][D|K|E|R|Q]',\\
&'[D|K|E|N|R|Q][D|K|E|N|R][D|K|E|R|Q][V|I|L][I|V|L|M][D|K|E|R|Q]',\\
&'[D|K|E|N|R][D|K|E|R|Q][I|V|L][I|V|L|M][K|Q|R|E][D|K|E|R|Q]',\\
&'[L|V|I|M][D|K|E|N|R|Q][D|K|E|N|R|Q][D|K|E|N|R][D|K|E|N|R|Q][I|V|L]']
\end{aligned}
$$

Slika 3.6: Rezultati naše klasifikacije

Test set \ Svrstano u	Lipaze	Walker
Lipaze	99	1
Walker	36	64

Slika 3.7: Rezultati najbolje klasifikacije iz [4]

Test set \ Svrstano u	Lipaze	Walker
Lipaze	99	1
Walker	3	97

Dakle, lipaze smo uspjeli jednako dobro klasificirati, ali familiju walker ne.

4 familije

Dobiveni motivi prve familije

$$
\begin{aligned}
[\,'&[F][N][T][F|W][R][R]',\\
&'[A][I|V][F][N][T][F|W]',\\
&'[I|V][F][N][T][F|W][R]',\\
&'[I|L][S\,][P][F|Y][E|R][Q]',\\
&'[S\,][P][F|Y][E|R][Q][K|R]',\\
&'[Y][S\,|T][I|V|L][S\,][P][F|Y]',\\
&'[S\,|T][Y][S\,|T][V|L][S\,][P]',\\
&'[E][F|Y][L][N][S\,][K]',\\
&'[K|R][G][I|V][I|V][S\,|T][Y]',\\
&'[S\,|T][I|L][S\,][P][F|Y][Q|E|R]'\,]
\end{aligned}
$$

Dobiveni motivi druge familije

$$
\begin{aligned}
[\,'&[K|D|E][K|E|R][Q|D|E|K][K|Q|E|R][K|E|R][K|E|R]',\\
&'[F][I|V|L][K|Q][K|E|R][I|L][K|R]',\\
&'[D][S][S][I|L][K][K|R]',\\
&'[I|L][K][K|R][N][T][A]',\\
&'[I|L][K|Q][E|R][I|L][R][T]',\\
&'[I|L|M][D|E|N][D|E|N|R][Q|D|E|K][Q|R|E|K][K|R]',\\
&'[L][D][S][S][I|L][K]',\\
&'[K][K][N][T][A][F]',\\
&'[K][N][T][A][F][I|V]',\\
&'[T][A][F][I|V][K][K|R]']
\end{aligned}
$$

Dobiveni motivi treće familije

$$
\begin{aligned}
[\,'&[D|E|N|R][E|D|N|Q|R][K|D|Q|E][E|D|N|Q|R][D|E|N|R][D|E|R]',\\
&'[L][L][D][N][Y][E|N]',\\
&'[E|D|N|K|R][K|D|E][K|E|N|R][K|D|E|R][I|V|L][K|D|E|N]',\\
&'[K|D|E|N][V|L][K|D|E|N][D|E][D|E|N][K|Q|E|R]',\\
&'[D|E|N][K|Q|N][W|Y|F][Q|E|N][K|R][Q|D|E|K]',\\
&'[W][A][Q|E|N][V][V][S|T]',\\
&'[S|T][W][A][Q|E|N][V][V]',\\
&'[S][S|T][W][A][Q|E|N][V]',\\
&'[S][M|L][V|L][D|N][K|N][F|Y]',\\
&'[Q|D|E|R][Q|D|E|N][Q|D|E][Q|E|R][S][Q|N|R]']
\end{aligned}
$$

Dobiveni motivi četvrte familije

$$
\begin{aligned}
&[\,'[K|Q|R][Q|E|R][I|V|L][K|E][E|R][D|R]',\\
&\,'[D|N][H|Y][Y][H][F][W|H|Y|F]',\\
&\,'[H|Y][Y][H][F][W|H|Y|F][H]',\\
&\,'[K|Q][K|Q|R][E|R][I|V][K][E|R]',\\
&\,'[I|V|L][A][E|Q|N|K|R][K|R][H|Y][G]',\\
&\,'[D|E][P][Q|K][V][K|Q|E][W]',\\
&\,'[K|Q|E][W][I|L][Q|E][Q][Q]',\\
&\,'[V][K|Q|E][W][I|L][Q|E][Q]',\\
&\,'[Q|K][V][K|Q|E][W][I|L][Q|E]',\\
&\,'[W][I|V|L][Q|E][Q][Q][V]']
\end{aligned}
$$

Slika 3.8: Rezultati naše klasifikacije

Test set \ Svrstano u	Familija 1	Familija 2	Familija 3	Familija 4
Familija 1	98	1		1
Familija 2	2	94	1	3
Familija 3	23	33	34	10
Familija 4	4	6		90

Slika 3.9: Rezultati najbolje klasifikacije iz [4]

Test set \ Svrstano u	Familija 1	Familija 2	Familija 3	Familija 4
Familija 1	100			1
Familija 2	42	57		1
Familija 3	60	2	38	
Familija 4	35	1	2	62

Naša klasifikacija bila je samo malo gora za familiju 1, ali dosta bolja za familiju 2 i familiju 4. Familiju 3 nismo dobro uspjeli klasificirati ni mi ni [4].

Poglavlje 4

All against all

All against all metodom (vidi [2]) uspoređujemo sve 10-grame u proteinima neke familije koristeći određeni kriterij sličnosti kojeg možemo prilagođavati. Značajnost score-a sličnosti procjenjuje se pomoću logističke distribucije.

Neka je R_i, $i \in \{1, ..., n\}$ neki protein famlije s n proteina. Klizećim prozorom odabiremo njegov 10-gram d_i.

Neka je $j \in \{1, ..., n\}$, $j \neq i$. Sada d_i uspoređujemo sa svim 10-gramima proteina R_j te uzmemo onog najsličnijeg i nazovemo ga d_j. Ako d_j zadovoljava prethodno određen kriterij sličnosti, dodamo ga u listu 10-grama dovoljno sličnih d_i. Ponovimo postupak za sve proteine familije, tj. za sve $j \neq i$ te zatim za sve ostale desetorke u R_i dobivene metodom klizećeg prozora. Nakon toga isti postupak provedemo za sve ostale $i \in \{1, ..., n\}$.

Tako smo usporedili sve 10-grame sa svim ostalim 10-gramima u određenoj proteinskoj familiji te sačuvali liste dovoljno sličnih. Neka je m broj lista koje smo dobili.

Dakle, na kraju imamo m lista dovoljno sličnih 10-grama, s tim da za svaku listu vrijedi da su njezini elementi 10-grami iz različitih proteina iste familije.

4.1 Određivanje motiva

Od prethodno dobivenih m lista, odabrali smo 12-15 najduljih, tj. onih s najviše elemenata. Na taj način osigurali smo da se 10-grami s kojima nastavljamo pojavljuju u što više proteina te familije, jer kao što smo prije rekli, unutar svake liste elementi su iz različitih proteina te familije.

Kako bismo dobili 12-15 najduljih lista, trebalo je odrediti broj h koji nam označava duljinu liste koja nam je prihvatljiva za tu familiju. Odabrali smo h na način da broj lista te familije duljih od h bude što manji broj između 12 i 15.

h za naše familije:

lipaze	$h = 51$
walker	$h = 39$
familija 1	$h = 81$
familija 2	$h = 76$
familija 3	$h = 51$
familija 4	$h = 69$

Dakle, za lipaze gledali smo sve liste dulje od 51, za walker sve liste dulje od 39 itd. Na taj način smo dobili $t \leq m$ lista desetorki za svaku familiju:

lipaze	$t = 12$
walker	$t = 14$
familija 1	$t = 12$
familija 2	$t = 13$
familija 3	$t = 13$
familija 4	$t = 13$

Sada za svaku familiju imamo t lista 10-grama, koje ćemo proglasiti pojavljivanjima t motiva te familije. S obzirom da su svi elementi liste 10-grami, tako će i naši novonastali motivi biti duljine 10.

Kako bismo formalno zapisali motiv od njegove liste pojavljivanja, dalje nastavljamo računati distribuciju kao u poglavlju 3. Kako bismo uskladili oznake:

$$\begin{aligned} m &= \text{duljina liste pojavljivanja motiva} \\ \text{n-grami} &= \text{10-grami} \\ q &= t \end{aligned}$$

Distibuciju sada imamo zapisanu u matrici F te sve aminokiseline u vektoru A. Motiv definiramo pomoću njih na sljedeći način. Matrica $F = [f_{ij}]$ $i = 1, .., 10$, $j = 1, .., 20$ ima 10 redaka koji označavaju 10 pozicija u motivu. Poziciju i, $i = 1, .., 10$ motiva definiramo kao sve aminokiseline $A[j]$ takve da je $f_{ij} > 0$.

Jednostavnije rečeno, motiv definiramo tako da se na i-toj poziciji motiva može nalaziti svaka aminokiselina s i-te pozicije svakog 10-grama u listi pojavljivanja. Na taj način od svake liste pojavljivanja zapisali smo motiv.

Nakon što smo odredili motive nastavljamo kao u poglavlju 3 te dobivamo rezultate klasifikacije.

4.2 Rezultati i usporedba

Kao i prije, uspoređujemo se s radom [4], ali i s našom metodom reduciranog alfabeta.

Lipaze i Walker

Dobiveni motivi za familiju lipaza:

$$
\begin{aligned}
&[\,'Y[N|R|V|E|Q|G|A|S|D|K|T][N|M|Q|H|Y|A|L][D|G][A|G|C]R[K|R][V|I|F][V|G|S|A|I|L][V|I|L|F]',\\
&'[K|M|V|Q|C|Y|I|L|F][M|V|L|F][R|N|E|G|S|A|K|P|T][R|N|Q|H|S|K]S[V|I|L|F][V|F]\\
&[V|I|L|M][M|V|I|L|F][S|V|E|G]',\\
&'A[S|E|Q|D][R|D|E|H|S|A|K|L|F][V|L|F|M][E|G|N][V|I|L|F]P[Y|H|L|F][V|S|I|L|P][N|Q|S|P|T]',\\
&'[T|W|I|M|V|Y|K|L|P|F][R|F|M|V|I|L|P|T][R|V|I|L|P][V|S|I|L|P|F][T|M|A|L|F]\\
&[V|C|Y|A|S|L|F][F|I|V|Q|G|C|Y|A|S|L|P|T][N|F|V|Q|G|C|S|A|K|L|T]\\
&[N|M|V|Q|G|C|S|A|I|L|T][T|N|R|V|G|C|H|S|I|L|F]',\\
&'[F|M|V|Y|A|K|L|T][R|M|V|Q|A|I|L|P|T][R|M|V|S|A|K|L|P|T][V|A|L|M][V|A|L|C]\\
&[V|S|A|L|P|F][W|R|V|Q|G|S|A|L|F][V|Q|A|I|L|F][S|L|F][R|V|E|Q|G|I|L|F]',\\
&'[T|R|K|M|Q|C|S|I|L|F][F|I|V|Y|E|S|A|K|L|P|T][R|M|Q|Y|A|L|P][M|V|Y|A|L|F]\\
&[R|V|C|S|A|I|L|P][W|V|A|L|F][V|Q|A|I|L|F][T|Q|A|L|F][V|Q|H|I|L|F][R|F|V|G|I|L|T]',\\
&'[R|I|M|V|E|S|A|K|L|P|T][R|M|V|Q|S|A|I|L|P|T][T|N|R|M|V|Q|H|A|I|L|P|F]\\
&[R|V|H|S|A|L|P][R|V|S|A|L|F][V|I|L|F][V|L|F][V|I|L|F][V|Q|G|C|A|I|L][N|R|M|V|Q|G|C|S|A|L|T]',\\
&'[T|R|M|Y|C|S|A|L|P|F][F|M|V|Y|I|L|P|T][R|V|H|S|A|I|L|P|F][V|S|D|I|L|P|F]\\
&[M|V|Q|S|L|F][M|V|G|C|A|I|L|F][V|Q|H|C|S|A|L|F][T|N|R|V|Q|G|C|H|S|A|L|F]\\
&[N|R|V|Q|G|C|H|S|A|I|L|F][T|N|V|Y|G|C|H|S|I|L|F]',\\
&'[R|F|M|V|C|Y|I|L|T][T|R|V|E|H|A|S|D|I|L|P|F][R|M|A|L|P|T][R|M|V|Y|I|L|F]\\
&[F|M|V|A|I|L|P|T][R|V|Q|C|S|A|L|P|T][V|Q|G|C|S|A|L|F]\\
&[N|R|M|V|Q|G|C|S|A|L|F][N|F|M|V|Q|C|I|L|T][T|N|R|V|G|H|S|A|I|L|F]',\\
&'GC[M|V|A|I|L][R|V|A|K|P][R|N|V|E|Q|G|H|S|A|K|P|T][T|N|S|I|Q|Y|K|P|F]\\
&[N|M|V|L|T][R|D|E|S|A|K|T][T|M|V|A|I|L|F][R|V|Y|S|A|P|F]',\\
&'[C|Y|I|L|F][V|I|L][A|P][S|E|Q|D][R|D|E|S|A|K|F][M|I|L|F][D|G|S|A|T][M|V|I|L|T]\\
&[R|V|A|K|P][R|E|Q|G|Y|D|K|L|F]',\\
&'[R|N|V|G|A|S|D|L|P|T][T|D|V|E|G|H|S|A|L|P|F][N|G|S|A|T][R|N|V|G|S|A|K|L|T]\\
&[N|D|V|G|S|A|P|T][V|G|S|A|T][R|V|G|S|A|T][R|G|S|A|K][M|V|G|Y|A|L|F][R|D|G|C|S|A|T]']
\end{aligned}
$$

Dobiveni motivi za familiju walker:

[′[E|G|P|V|Q|N|H|K|D|A|T|R|S|Y][G|P|N|H|K|D|A|T|R|S|Y][G|P|C|N|H|K|D|A|T|R|S|Y]
[G|P|C|N|H|K|D|A|T|S][G|C|V|T|P|N|K|D|A|F|R|S][G|P|V|C|N|F|K|A|T|R|S][L|E|G|P|N|F|K|D|A|T|R|S]
[E|G|P|T|V|N|K|D|A|F|R|S][E|G|P|C|V|N|K|D|A|T|R|S][L|E|G|P|V|C|T|N|K|D|A|F|R|S]′,
′[G|P|V|Q|N|H|K|D|A|T|I|S|Y][G|P|V|C|N|H|K|D|A|T|R|S|Y][G|P|N|H|K|D|A|T|S][G|P|V|N|F|K|D|A|T|S]
[G|P|V|C|N|K|D|A|T|R|S][G|P|V|N|K|D|A|T|R|S][L|E|G|P|V|N|K|D|A|T|R|I|S][E|G|C|P|H|N|K|D|A|T|R|S]
[L|E|G|P|T|C|V|H|N|K|D|A|F|R|S][E|G|P|V|Q|C|H|N|K|M|D|A|T|R|S]′,
′[G|P|V|T|Q|N|K|M|A|F|R|I|S][G|P|V|N|H|K|A|T|R|I|S][L|G|P|V|N|K|A|T|R|I|S][L|G|N|H|K|A|F|S]
[L|G|P|H|N|A|T|R|S][L|G|P|V|F|D|A|T|I|S][L|E|G|P|V|Q|H|N|D|A|T|R|I|S][L|E|G|P|C|V|N|H|M|D|A|T|R|S]
[L|E|G|P|C|Q|N|H|F|K|D|A|T|R|S][E|G|P|V|Q|C|N|H|K|F|D|A|T|R|S]′,
′[G|P|V|Q|K|A|R|I|S][L|G|P|V|N|K|A|I|S][L|G|P|V|N|K|A|I|S][L|G|T|V|Q|H|N|K|A|F|I|S]
[L|G|V|N|H|D|A|T|S][L|G|Q|N|F|K|D|T|R|S][L|G|P|N|F|D|A|T|S][L|E|G|P|C|Q|N|H|D|A|T|R|S]
[E|G|P|V|C|H|F|N|A|T|R|S][L|E|P|Q|H|F|K|N|D|A|T|R|I|S]′,
′[E|G|P|V|N|D|A|T|S][T|D|S|L][E|G|V|Q|N|K|D|R|I|S][G|V|N|M|D|A|T|I|S][E|G|Q|H|N|D]
[E|G|V|N|D|I|S][L|Q|D|A|S|Y][E|G|P|C|Q|H|D|A|S][L|E|C|V|P|A|F|I|S][L|G|V|N|A|T|I|S]′,
′[E|G|N|D|A|S][L|E|P|V|Q|N|D|T|S][E|G|P|V|Q|N|D|A|I|S][L|V|A|F|R|I][E|V|D|A|F|I|Y]
[L|E|G|V|K|R|I][L|G|M|A|F|I][L|G|C|V|Q|N|K|A|T|I|S][G|Q|H|K|D|A|R|S][E|Q|H|K|D|A|T|S]′,
′[L|E|P|H|F|T|I|S][L|E|P|V|D|A|R|I|S][L|E|G|V|Q|N|K|D|A|F|S][L|E|V|N|H|I][L|E|K|M|A|I]
[C|H|D|T|I|S][E|C|Q|H|K|D|R][E|V|N|K|D|I|S][L|Q|W|D|A|T|R|I|S][L|E|V|Q|K|D|R]′,
′[L|E|G|Q|K|M|D|A|T|R|I|S][L|E|T|V|N|K|D|A|F|R|I|S][E|G|P|V|Q|K|D|A|T|I][L|E|G|K|D|A|R|I]
[E|G|C|V|P|Q|K|D|A|T|I][E|N|K|D|T|R|S][E|V|K|D|A|T|R][E|G|V|Q|N|W|K|D|A|T|R|S]
[E|G|N|K|D|A|R|S|Y][L|E|G|V|Q|N|H|K|D|A|T|R|I|Y]′,
′[L|E|P|V|T|Q|N|H|K|D|A|F|R|I][L|E|V|K|T][G|C|K|A|T|R|I|S][L|E|G|Q|K|A|T|R|S][E|V|K|D|A|I|S]
[L|V|N|K|A|T|R|S][L|G|N|K|M|D|A|S][E|G|V|Q|N|H|K|D|T][M|V|I|L][L|V|Q|K|R|I|S]′,
′[L|V|H|T|S][L|G|C|P|T|I|S][E|Q|H|K|D|T|R|S][L|E|G|V|Q|H|K|D|F|I][L|P|V|N|K|R|S|Y]
[L|V|Q|N|K|M|D|S][E|G|Q|H|K|D|T|S][V|I|L][L|E|G|P|V|Q|A|R][L|E|Q|H|N|K|D|R|S]′,
′[E|G|C|K|D|A|R|S][E|G|C|K|A|R|S][E|G|V|N|K|D|Y][L|T|V|K|F|R|S][L|E|G|V|Q|K|M|D|T|R]
[E|Q|K|D|A|T|R][L|N|W|K|D|R][L|E|V|D|A|F|R|S][P|Q|N|K|D|R|S][L|E|G|P|V|Q|D|A|S|Y]′,
′[L|E|Q|K|T|R|S][E|Q|K|D|R|S][L|E|T|Q|N|K|D|A|F|R|S|Y][E|V|Q|K|M|A|R|S][E|G|V|Q|H|K|D|T|R|I|S]
[L|E|Q|N|W|K|R|I][L|E|G|P|V|Q|K|D|A|R|S][L|E|G|Q|N|K|M|A|T|R|S][L|E|G|V|Q|K|D|A|I|S]
[L|E|P|V|N|H|K|D|A|R|S]′,
′[E|G|V|N|K|D|I|S][L|C|V|Q|N|R|I|S][E|G|Q|K|M|R][E|G|P|V|K|D|R][L|H|W|K|R|I][L|E|V|N|A|F|I|S]
[E|P|Q|K|M|A|R|S][E|P|V|Q|W|D|A|F|I|S][L|E|G|N|M|K|A|R|I|S][L|E|G|P|N|W|M|K|D|T|R|S]′,
′[G|P|V|K|A|T|R|I|S][L|G|P|V|K|A|T|R|I|S][L|E|G|P|V|K|A|T|S][L|E|G|P|H|K|M|A|T|R|I|S]
[L|G|H|N|A|T|I|S][L|G|P|V|F|A|T|R|I|S][L|E|G|P|N|H|D|A|T|S][L|E|G|P|V|C|H|D|A|T|R|I|S|Y]
[L|G|P|C|Q|H|F|K|N|A|T|R|S][E|G|P|V|Q|H|N|K|F|D|A|T|R|S]′]

Slika 4.1: Rezultati all against all klasifikacije

Test set \ Svrstano u	Lipaze	Walker
Lipaze	93	7
Walker	21	79

Lipaze svrstali smo lošije nego s metodom reduciranog alfabeta, ali walker familiju bolje. Ipak i dalje lošije od rezultata iz [4].

4 familije

Dobiveni motivi prve familije:

$[{}'G[Y|H|C|N|D|A|S][L|F|I|M]G[T|G|S][Q|T|L|P|A|S][K|T|V|P|A]Q[K|R][G|N|Y]',$
$'[Q|T|V|D|I|S][D|E|H|Q][K|R][N|S][E|H][F|Y]L[N|Y]SK',$
$'[W|N|L|F|M][G|A]G[A|T][L|F|G|A|M][N|H][N|A|D|S]A[F|V|I]F',$
$'[T|N|D|A|S]A[F|V|I]FN[V|T][F|W]RR[F|T]',$
$'Y[E|Y|Q|T|V|L|F|G|A|S][A|S|T|C|V|L|G|I|M]M[E|H|Q|N|D|S]W[G|A][Q|T|V|I|S][H|E|D|Q][K|R]',$
$'[E|Y|Q|T|V|L|F|G|A|S][A|T|C|V|L|G|I|M]M[E|H|Q|N|D|S]W[G|A][Q|T|V|I|S][E|K|H|Q|D][K|R][W|N|S]',$
$'[A|T|C|V|L|G|I|M]M[E|H|Q|N|D|S]W[G|A][Q|T|V|I|S][E|K|H|Q|D][K|R][N|S][E|H]',$
$'M[E|H|Q|N|D|A|S]W[G|A][Q|T|V|I|S][E|K|H|Q|D][K|R][N|S][E|H][F|Y]',$
$'[E|H|Q|N|D|A|S]WA[Q|T|V|I|S][E|K|H|Q|D][K|R][N|S][E|H][F|Y]L',$
$'WA[Q|T|V|D|F|I|S][D|E|H|Q][K|R][N|S][E|H][F|Y]LN',$
$'A[Q|T|V|D|F|I|S][D|E|H|Q][K|R][N|S][E|H][F|Y]LNS',$
$'[Q|T|V|D|F|I|S][D|E|H|Q][K|R][N|S][E|H][F|Y]LNSK']$

Dobiveni motivi druge familije:

$$
\begin{aligned}
&[\,'[E|K|S|Q|D|L|R][M|K|S|H|Q|V|L|I|R][E|M|K|H|Q|V|I|D|L|F|A|R][E|M|K|S|H|Q|V|I|D|L|F|A|R]\\
&\ [E|R|K|H|Q|T|I|D|L|F|A|M][K|E|D|Q][E|K|Q|V|D|L|A|R][E|K|S|Q|V|I|D|L|A|R]\\
&\ [E|M|K|Q|V|I|N|L|A|R][E|M|K|H|Q|V|I|N|L|A|R]',\\
&\ '[K|R][K|Q|R][N|H|R][E|H|T|Q|D][L|I][K|R][E|K|T|V|N|D|A|S][Q|T|V|L|S]N[E|S|Q|T|N|D|L|I|A]',\\
&\ '[Q|H|S]KYL[T|A|S]E[V|I][V|I][A|S]A',\\
&\ '[E|A|D|T][A|T][G|A][V|A]E[V|I][T|V|L|I|A][C|S]AL',\\
&\ '[E|R|M|K|H|Q|V|I|D|L|P|A|S][E|K|Q|D|I][E|K|S|Q|I|D|L|A|R][E|K|S|Q|V|I|D|L|A|R][E|M|K|S|Q|V|I|L|A|R]\\
&\ [E|M|K|H|Q|V|L|F|A|R][E|M|K|S|H|Q|V|I|D|L|F|A|R][E|M|K|H|Q|T|I|D|L|F|A|R][E|K|H|Q|D|L|A|R]\\
&\ [E|K|Q|V|D|L|A]',\\
&\ '[V|I][V|I|A|T][C|S]ALHQRF[T|G|A|S]',\\
&\ 'L[Q|T|H|S]KYL[T|A|S]E[V|I][V|I][A|S]',\\
&\ '[Q|H|S]KY[L|V][T|A|S]E[V|I][V|I][A|S]A',\\
&\ 'KYL[T|A|S]E[V|I][V|I][A|S]A[T|C|F|A|S]',\\
&\ '[T|S]L[E|S|H|Q]KY[L|V][T|A|S]E[V|I][V|I]',\\
&\ '[Q|H|S]KY[L|V][T|A|S]E[V|I][V|I][A|S]A',\\
&\ '[E|Q|V|N|D|L|P|A|S][V|I|T][H|Q|N|G|A|S][R|K|H|T|Q|N|D|G|A|S][T|K|A|S]LDS[A|N|S|T][L|M]',\\
&\ '[E|K|S|Q|V|D|L|I|R][M|K|S|H|Q|V|L|F|I|R][E|M|K|S|H|Q|I|D|L|F|A|R][E|M|K|S|H|Q|V|I|D|L|F|A|R]\\
&\ [E|R|K|H|Q|I|D|L|F|A|M][K|E|D|Q][E|K|S|Q|V|D|L|A|R][E|K|S|Q|V|I|D|L|A|R][E|M|K|Q|V|I|N|L|A|R]\\
&\ [E|M|K|H|Q|V|N|L|A|R]']
\end{aligned}
$$

Dobiveni motivi treće familije:

$$
\begin{aligned}
&[\,'[L|M][L|A|M][G|N|D]N[F|Y][E|K|H|Q|N|D|I|A][R|K|S|T|P|A|M][E|M|H|Q|N|D|A|S][V|A|T][G|S]',\\
&'[L|F|A|M][G|N|D|R][K|N|H|R][F|Y][E|H|Q|N|D|A|S][R|K|S|T|A|M][E|Q|D|A|S][R|T][G|S]\\
&[A|E|K|H|Q|T|V|W|I|R]',\\
&'[G|N|D|R][K|N][F|Y][E|Y|Q|N|D|A|S][R|K|S|T|A|M][E|H|Q|D|A|S]T[G|S][A|E|K|H|Q|V|N|I|R]\\
&[E|K|T|C|L|P|A|S]',\\
&'[E|R|K|Q|T|V|N|D|P|G|A|S][E|R|Y|K|H|T|Q|V|N|D|P|G|A|S][E|Y|K|T|Q|V|I|D|P|A|S][E|K|T|V|I|P|G|A|S]\\
&[K|T|V|I|L|P|G|A|S][A|R|K|T|V|N|L|P|G|I|S][R|K|T|Q|V|N|L|P|G|A|S][R|K|T|V|N|L|P|A|S]\\
&[K|T|V|D|L|P|A|S][K|T|V|D|L|P|A|S]',\\
&'[E|M|K|Q|N|D|P|A|S][E|K|H|T|Q|N|D|G|A|S][L|V|I][Y|K|L|F|A|R][E|R|K|S|H|Q|N|D|A|M]\\
&[R|K|T|Q|L|G|I|S][P|E|T][V|T|S][F|Y][E|K|H|Q|V|N|A|S]',\\
&'[L|V|F|I][R|K|L|F|S][R|K|H|T|N|A|S][K|T|L|I|R][E|K|T|D|P][T|V|S][F|Y][E|K|T|V|N|A|S]\\
&[R|Y|K|H|T|Q|L|F|A|S][L|F|V|A]',\\
&'[E|K|S|Q|T|V|I|N|D|L|F|G|A|R][E|K|S|H|Q|V|I|N|D|L|F|P|G|A|R][E|K|S|H|Q|T|N|D|P|G|A|R]\\
&[E|K|S|H|Q|N|D|P|G|R][E|R|H|Q|N|D|G|S][E|R|H|Q|N|D|G|S][E|R|Q|N|D|G|S][E|R|Q|W|N|D|P|G|S]\\
&[E|R|H|Q|N|D|P|G|A|S][E|R|Y|Q|W|N|D|P|G|A|S]',\\
&'[E|K|S|H|Q|T|V|I|N|D|L|F|P|G|A|R][E|K|S|H|T|V|N|D|L|P|G|A|R][E|R|K|H|Q|N|D|L|P|G|A|S]\\
&[E|S|H|Q|T|N|D|P|G|R][E|S|H|Q|N|D|P|G|R][E|R|K|Q|N|D|G|S][E|R|K|H|Q|N|D|G|S][E|R|Q|W|N|D|G|S]\\
&[E|R|Q|N|D|P|G|A|S][E|R|Q|N|D|P|G|A|S]',\\
&'[E|K|S|H|Q|T|I|N|D|L|G|A|R][E|Y|K|H|Q|I|N|D|L|P|G|A|R][E|K|S|H|Q|N|D|G|R]\\
&[E|Y|K|S|H|Q|N|D|P|G|R][E|R|K|H|Q|N|D|G|S][E|R|H|Q|W|N|D|G|A|S][E|R|H|Q|N|D|G|A|S]\\
&[E|R|Q|W|N|D|P|G|A|S][E|R|Q|T|N|D|P|G|A|S][E|R|Q|W|N|D|P|G|A|S]',\\
&'[E|S|H|Q|T|V|N|D|P|G|A|R][E|R|K|Q|T|N|D|F|P|G|A|S][E|H|Q|N|D|P|G|R][E|R|H|Q|N|D|F|P|G|A|S]\\
&[E|K|S|H|Q|N|D|G|R][E|R|H|Q|N|D|P|G|S][E|R|H|Q|N|D|P|G|A|S][E|S|H|Q|T|W|N|D|A|R][E|Q|N|D|P|S]\\
&[E|R|H|T|Q|N|D|P|G|A|S]',\\
&'[E|K|S|H|Q|N|D|L|G|A|R][E|K|S|H|Q|W|N|D|P|G|A|R][E|Y|K|S|H|Q|N|D|P|G|A|R][E|R|K|H|Q|N|D|G|S]\\
&[E|K|S|H|Q|W|N|D|P|A|R][E|R|K|Q|N|D|P|G|A|S][E|R|K|Q|N|D|P|G|A|S][E|R|K|H|Q|W|N|D|G|S]\\
&[E|R|Q|V|N|D|P|G|S][E|R|K|Q|T|V|W|N|D|G|A|S]',\\
&'[E|K|S|Q|W|N|D|P|G|A|R][E|Y|K|S|H|T|N|D|P|G|R][E|K|S|H|Q|C|N|D|G|R][E|K|H|Q|W|N|D|P|G|A|R]\\
&[E|R|K|H|Q|T|N|D|G|A|S][E|K|S|H|Q|V|W|N|D|P|G|A|R][E|Y|K|H|Q|N|D|G|A|R][E|Q|V|N|D|G|A|R]\\
&[E|R|K|Q|N|D|P|G|A|S][E|R|H|Q|T|N|D|P|G|A|S]',\\
&'[E|K|S|H|Q|T|N|D|P|G|R][E|K|S|H|Q|N|D|P|G|R][E|Y|S|H|Q|V|N|D|P|G|A|R][E|R|K|Q|N|D|P|G|A|S]\\
&[E|R|K|H|Q|N|D|P|G|A|S][E|S|H|Q|N|D|L|G|A|R][E|H|Q|N|D|G|A|S][E|H|Q|T|N|D|G|A|R]\\
&[E|R|K|H|Q|W|N|D|P|G|S][E|R|K|H|Q|W|D|P|G|A|S]']
\end{aligned}
$$

Dobiveni motivi četvrte familije:

$$
\begin{aligned}
&[\,'[A|E|K|Q|T|V|D|F|I|R][E|K|S|T|V|N|D|F|P|R][A|Q|T|V|D|L|I|M][Y|C|D|F|M][T|C|V|L|F|M]\\
&\ [C|F|N|I][E|H|T|F|P|I][P|W|I|Y][P|S|A|Y][R|Y|V|A|S]',\\
&\ '[R|K|T|P|S][T|Q|V|I|L|A][F|Y][L|V|S|T]N[E|H|Q|T|R]WA[V|A][E|K|H|Q|R]',\\
&\ '[A|R|K|T|V|D|L|I|S][L|Y|N|S][R|H|Q|V|L|I|M][L|G][K|Q][V|I]G[T|N|D|G|A|S]L[E|K|D|P|R]',\\
&\ '[V|S|I|Q][V|G|D|S][T|N|D|G|A|S][L|K|D|T][E|R|K|T|Q|D|M][K|C|N|D|M][Y|H|C|L|F|M][C|L|F|Y]\\
&\ [L|F|H][F|I]',\\
&\ '[M|S|V|L|P|G|A|R][S|C|V|L|P|G|A|R][E|S|Q|C|V|T|P|G|A|R][S|T|C|P|G|A|R][S|H|T|L|P|G|A|R]\\
&\ [E|S|Q|C|L|P|G|A|R][E|R|T|L|P|G|A|S][S|T|W|L|P|G|A|R][E|R|S|C|V|D|L|P|G|I|A][S|H|T|Q|V|I|L|P|G|A|R]',\\
&\ 'WA[V|A|M][E|K|H|Q|R][L|V|I][E|R|T|N|P|G|A|S][G|D]G[A|E|K|Q|L|F|P|I|S][E|Q|V|N|D|P|A|S]',\\
&\ '[V|A][E|K|H|Q|R][V|I][E|T|V|N|P|G|A|S]GG[E|Q|N|L|F|P|I|S][E|Q|D|P|A|S][E|K|V|D|A][E|A]',\\
&\ '[V|I][E|T|V|N|L|P|A|S]GG[E|Q|N|L|F|P|S][E|T|P|G|A][V|E|A|D]A[E|Q|N|D|S][E|R|K|Q|L|G|A|S]',\\
&\ '[K|G|E|Q]IG[T|N|D|G|A|S]L[E|K|D|P|R][G|N|D][F|H|Y][F|Y][L|H]',\\
&\ '[V|I]G[T|N|D|G|A|S]L[E|K|Q|D|R][N|D][H|Y]Y[L|H|T]F',\\
&\ 'G[T|N|D|G|A|S]L[K|E|D|R][N|D][H|Y]Y[L|H|T]F[Y|K|H|F|R]',\\
&\ '[T|N|D|G|A|S]L[K|E|D|R][N|D][H|Y]Y[L|H]F[Y|K|H|F|R]H',\\
&\ 'H[P|S][K|Q|R][T|R][F|I|S][K|Q|R]RS[V|A|T][Y|V|L|F|P|I]']
\end{aligned}
$$

Slika 4.2: Rezultati all against all klasifikacije

Test set \ Svrstano u	Familija 1	Familija 2	Familija 3	Familija 4
Familija 1	72	28		1
Familija 2		100		
Familija 3		80	20	
Familija 4		29		71

Familiju 1,3 i 4 lošije smo svrstali nego s metodom reduciranog alfabeta. Familija 2 bolje je svrstana, ali sve u svemu ova metoda ipak nije donijela poboljšanje u usporedbi s metodom reduciranog alfabeta. Ipak, familije 2 i 4 i dalje su bolje svrstane ovom metodom nego metodom iz [4].

Poglavlje 5

Zaključak

Metodom reduciranog alfabeta u 4/6 slučajeva dobili smo klasifikaciju bolju od 90%. U usporedbi s radom [4], za dvije familije dobili smo bolju klasifikaciju, za jednu familiju jednako dobru klasifikaciju, a za tri familije lošiju. Iako ova metoda nije donijela veću uspješnost u konačnici, njezina prednost je znatno manji broj motiva. Odabrano ih je samo 10 u usporedbi sa njih 100 iz [4] što znači manje potrebe za računalnom snagom i vremenom. Metoda bi se mogla upotrijebiti za potrebe klasifikacije puno većeg broja familija.

All against all metoda također koristi manje motiva, njih 12-15, ali nije pokazala veću uspješnost od metode reduciranog alfabeta.

Bibliografija

[1] D. Bakić, *Linearna algebra*, Školska Knjiga, 2008.

[2] Rabar et al., *IGLOSS: Iterative Gapless Local Similarity Search*, (2018), `https://arxiv.org/pdf/1807.11862.pdf`.

[3] K. Martinić, *Maksimalne klike u analizi sličnosti proteinskih motiva*, Diplomski rad, Prirodoslovno-matmatički fakultet, Matematički odsjek, 2018.

[4] S. Mavrek, *Iterativno traženje fraza i statistika semantičkog indeksiranja*, Diplomski rad, Prirodoslovno-matmatički fakultet, Matematički odsjek, 2019.

[5] N. Sarapa, *Teorija vjerojatnosti*, Školska Knjiga, 2002.

[6] M. Vuković, *Teorija Skupova*, Skripta, Sveučilište u Zagrebu, PMF–Matematički odsjek, 2015.

[7] A. Čavka, *Iterativno traženje motiva i vreća fraza u genomu i proteomu*, Diplomski rad, Prirodoslovno-matmatički fakultet, Matematički odsjek, 2019.

Sažetak

Napretkom bioinformatike i pronalaskom novih neopisanih proteina javlja se sve veća potreba za novim metodama klasifikacije proteina. U ovom diplomskom radu prezentirali smo dvije metode klasifikacije. Razvili smo i opisali metodu reduciranog alfabeta te proveli all against all metodu. Svaka od tih metoda proizvela je 10-15 motiva za svaku proteinsku familiju pomoću kojih smo klasificirali članove familija.

Summary

Advances in bioinformatics and the discovery of new protein families increases the need for new protein classification methods. In this thesis we presented two classification methods. We developed and described the reduced alphabet method, and we carried out an all-against-all search. Both methods produced 10-15 motifs for each protein family, which were then used to classify members.

Životopis

Rođena sam u Zadru, 5. listopada 1995. godine. U rodnom gradu završila sam opći smjer Gimnazije Franje Petrića nakon čega sam odselila u Zagreb gdje 2014. godine upisujem Preddiplomski sveučilišni studij Matematika na Prirodoslovno-Matematičkom fakultetu Sveučilišta u Zagrebu. Zatim na istom fakultetu 2018. godine upisujem Diplomski sveučilišni studij Matematička statistika.