Analiza točnosti pretraživanja

Strmečki, Darija

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO-MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Darija Strmečki

ANALIZA TOČNOSTI PRETRAŽIVANJA

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. _____

Ovaj rad posvećujem svojim roditeljima, bratu i baki. Zahvaljujem vam na pruženoj ljubavi, potpori i razumijevanju tijekom cijelog mog školovanja.
Posebno se želim zahvaliti svom mentoru doc. dr. sc. Pavlu Goldsteinu na zanimljivoj temi, te velikoj pomoći, uloženom trudu i strpljenju pri izradi ovog rada.

Sadržaj

Uvod

Bioinformatika je znanstvena disciplina koja se bavi analizom raznih bioloških podataka pomoću primijenjene matematike, statistike i računarstva. Neki od zadataka bioinformatike su sekvenciranje genoma i klasifikacija proteina (odnosno nizova aminokiselina) u proteinske familije. Razvojem tehnologije pronađeno je mnogo dosad nepoznatih nizova aminokiselina, što je rezultiralo stvaranjem velikih bioloških skupova podataka. Uspoređivanje i klasifikacija tih podataka predstavlja velik izazov u bioinformatici.

U ovom radu, bavit ćemo se pretraživanjem proteoma četiri različite biljke: talijinog uročnjaka, rajčice, azijske riže i šećerne repe. Proteom čine svi proteini nekog organizma nastali kao posljedica ekspresije gena. Želimo identificirati sve proteine u proteomu koje pripadaju određenoj proteinskoj familiji, u našem slučaju GDSL lipazama.

Jedna od tehnika korištenih za pretraživanje proteoma je iterativno pretraživanje u odnosu na kraći niz aminokiselina karakterističan za određenu proteinsku familiju. Takav niz nazivamo motiv (upit). Pomoću iterativnog pretraživanja, kroz određeni broj iteracija dobiva se skup nizova aminokiselina koji su dovoljno slični zadanom motivu.

Tehnika koju ćemo koristiti u ovom radu sastoji se od određivanja očekivanih parametara za funkciju sličnosti proteina iz određene proteinske familije i određivanja očekivane vrijednosti te funkcije za svaki od proteina unutar promatranog skupa. Svakom elementu tog skupa pridružujemo $0-1$ graf, te pomoću najveće maksimalne klike pokušavamo pronaći proteine iz određene proteinske familije. Na kraju, analiziramo točnost provedenog pretraživanja pomoću maksimalne klike u odnosu na rezultate dobivene iterativnim pretraživanjem proteoma.

Ovaj rad podijeljen je u pet poglavlja. U prvom poglavlju definirani su osnovni matematički pojmovi iz vjerojatnosti i statistike. U drugom poglavlju navedeni su pojmovi iz teorije grafova, a u trećem poglavlju pojmovi iz bioinformatike potrebni za razumijevanje rada. U četvrtom poglavlju opisana je metoda kojom smo odredili očekivane parametre za funkciju sličnosti, te metoda za pretraživanje skupa proteina. Konačno, u zadnjem poglavlju navedeni su rezultati i dobiveni zaključci.

Poglavlje 1

Vjerojatnost i statistika

1.1 Vjerojatnosni prostor

Definicija 1.1.1. ***Slučajni pokus*** *ili* ***slučajni eksperiment*** *je pokus čiji ishodi, tj. rezultati nisu jednoznačno određeni uvjetima u kojima izvodimo pokus.*

Definicija 1.1.2. ***Prostor elementarnih događaja*** Ω *je neprazan skup koji reprezentira skup svih ishoda slučajnog pokusa. Elemente* ω *skupa* Ω *nazivamo* ***elementarni događaji****.*

Definicija 1.1.3. *Familija* $\mathcal{A}$ *podskupova od* Ω ($\mathcal{A} \subset \mathcal{P}(\Omega)$) *je* ***algebra skupova*** *(na* Ω*) ako je:*

(i) $\emptyset \in \mathcal{A}$

(ii) $A \in \mathcal{A} \Rightarrow A^c \in \mathcal{A}$

(iii) $A_1, A_2, ..., A_n \in \mathcal{A} \Rightarrow \bigcup_{i=1}^{n} A_i \in \mathcal{A}.$

Definicija 1.1.4. *Familija* $\mathcal{F}$ *podskupova od* Ω ($\mathcal{F} \subset \mathcal{P}(\Omega)$) *je* ***$\sigma$-algebra skupova*** *(na* Ω*) ako je:*

(i) $\emptyset \in \mathcal{F}$

(ii) $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$

(iii) $A_i \in \mathcal{F}, i \in \mathbb{N} \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}.$

Definicija 1.1.5. *Neka je* $\mathcal{F}$ *σ-algebra na skupu* Ω*. Uređen par* $(\Omega, \mathcal{F})$ *zove se* ***izmjeriv prostor****.*

Definicija 1.1.6. *Neka je $(\Omega, \mathcal{F})$ izmjeriv prostor. Funkcija $\mathbb{P} : \mathcal{F} \to \mathbb{R}$ je **vjerojatnost** (na $\mathcal{F}$, na Ω) ako vrijedi:*

(i) $\mathbb{P}(A) \geqslant 0, A \in \mathcal{F}$ *(nenegativnost)*

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

(iii) $A_i \in \mathcal{F}, i \in \mathbb{N}$ *i* $A_i \cap A_j = \emptyset$ *za* $i \neq j \Rightarrow \mathbb{P}\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$ *(prebrojiva ili σ-aditivnost).*

Definicija 1.1.7. *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 $\mathcal{F}$ zovemo **događaji**, a broj $\mathbb{P}(A), A \in \mathcal{F}$ zove se **vjerojatnost događaja** A.

1.2 Uvjetna vjerojatnost. Nezavisnost

Definicija 1.2.1. *Neka je $(\Omega, \mathcal{F}, \mathbb{P})$ proizvoljan vjerojatnosni prostor i $A \in \mathcal{F}$ takav da je $\mathbb{P}(A) > 0$. Definirajmo funkciju $\mathbb{P}_A : \mathcal{F} \to [0, 1]$:*

$$\mathbb{P}_A(B) = \mathbb{P}(B \,|\, A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}, \quad B \in \mathcal{F}. \tag{1.1}$$

*$\mathbb{P}_A$ je vjerojatnost na $\mathcal{F}$ i nazivamo je **uvjetna vjerojatnost uz uvjet A**. Broj $\mathbb{P}(B \,|\, A)$ zovemo **vjerojatnost od B uz uvjet A**.*

Definicija 1.2.2. *Neka je $(\Omega, \mathcal{F}, \mathbb{P})$ vjerojatnosni prostor i $A, B \in \mathcal{F}$. Događaji A i B su **nezavisni** ako vrijedi:*

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B). \tag{1.2}$$

Napomena 1.2.3. *Ako su događaji A i B nezavisni i $\mathbb{P}(A) > 0$, $\mathbb{P}(B) > 0$, tada iz (1.2) i (1.1) slijedi:*

$$\mathbb{P}(B|A) = \mathbb{P}(B), \quad \mathbb{P}(A|B) = \mathbb{P}(A). \tag{1.3}$$

Definicija 1.2.4. *Neka je $(\Omega, \mathcal{F}, \mathbb{P})$ vjerojatnosni prostor i $A_i \in \mathcal{F}, i \in I$ proizvoljna familija događaja. Kažemo da je to **familija nezavisnih događaja** ako za svaki konačan podskup različitih indeksa $i_1, i_2, \ldots, i_k \in I$ vrijedi*

$$\mathbb{P}\Big(\bigcap_{j=1}^{k} A_{i_j}\Big) = \prod_{j=1}^{k} \mathbb{P}(A_{i_j}). \tag{1.4}$$

1.3 Slučajne varijable

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

Definicija 1.3.1. *Neka je $\mathcal{B}$ σ-algebra generirana familijom svih otvorenih skupova na $\mathbb{R}$. $\mathcal{B}$ zovemo **σ-algebra Borelovih skupova** na $\mathbb{R}$, a elemente σ-algebre $\mathcal{B}$ zovemo **Borelovi skupovi**.*

Definicija 1.3.2. *Neka je $(\Omega, \mathcal{F}, \mathbb{P})$ vjerojatnosni prostor. Funkcija $X : \Omega \to \mathbb{R}$ je **slučajna varijabla** (na Ω) ako je $X^{-1}(B) \in \mathcal{F}$ za proizvoljno $B \in \mathcal{B}$, tj. $X^{-1}(\mathcal{B}) \subset \mathcal{F}$.*

Definicija 1.3.3. *Funkcija $g : \mathbb{R} \to \mathbb{R}$ je **Borelova funkcija** ako je $g^{-1}(B) \in \mathcal{B}$ za svako $B \in \mathcal{B}$, tj. ako je $g^{-1}(\mathcal{B}) \subset \mathcal{B}$.*

Definicija 1.3.4. *Neka je $(\Omega, \mathcal{F}, \mathbb{P})$ vjerojatnosni prostor i X slučajna varijabla na Ω. Za $B \in \mathcal{B}$, definiramo funkciju $\mathbb{P}_X : \mathcal{B} \to [0, 1]$ relacijom:*

$$\mathbb{P}_X(B) = \mathbb{P}(X^{-1}(B)) = \mathbb{P}\{\omega \in \Omega : X(\omega) \in B\} = \mathbb{P}\{X \in B\}. \tag{1.5}$$

*$\mathbb{P}_X$ zovemo **vjerojatnosna mjera inducirana sa X**, a vjerojatnosni prostor $(\mathbb{R}, \mathcal{B}, \mathbb{P}_X)$ zovemo **vjerojatnosni prostor induciran sa X**. $\mathbb{P}_X$ često zovemo i **zakon razdiobe od X**.*

Definicija 1.3.5. *Slučajna varijabla X je **diskretna** ako postoji konačan ili prebrojiv skup $D \subset \mathbb{R}$ takav da je $\mathbb{P}\{X \in D\} = 1$.*

1.4 Funkcija distribucije. Funkcija gustoće

Definicija 1.4.1. *Neka je X slučajna varijabla na Ω. **Funkcija distribucije od X** je funkcija $F_X : \mathbb{R} \to [0, 1]$ definirana sa*

$$\begin{aligned} F_X(x) &= \mathbb{P}_X((-\infty, x]) = \mathbb{P}(X^{-1}(-\infty, x]) = \\ &= \mathbb{P}\{\omega \in \Omega : X(\omega) \leqslant x\} = \mathbb{P}\{X \leqslant x\}, \quad x \in \mathbb{R}. \end{aligned} \tag{1.6}$$

Napomena 1.4.2. *Ukoliko je jasno o kojoj se slučajnoj varijabli radi, umjesto F_X pisat ćemo F.*

Teorem 1.4.3. *Funkcija distribucije F slučajne varijable X je rastuća i neprekidna zdesna na $\mathbb{R}$, te zadovoljava*

$$\begin{aligned} F(-\infty) &= \lim_{x \to -\infty} F(x) = 0 \\ F(+\infty) &= \lim_{x \to +\infty} F(x) = 1. \end{aligned} \tag{1.7}$$

Funkciju $F : \mathbb{R} \to [0, 1]$ koja ima svojstva iz prethodnog teorema zvat ćemo **vjerojatnosna funkcija distribucije** (na $\mathbb{R}$) ili, kraće, **funkcija distribucije**.

Definicija 1.4.4. *Neka je X slučajna varijabla na vjerojatnosnom prostoru $(\Omega, \mathcal{F}, \mathbb{P})$ i neka je F_X njezina funkcija distribucije. Kažemo da je X **apsolutno neprekidna** ili, kraće, **neprekidna slučajna varijabla** ako postoji nenegativna realna Borelova funkcija f na $\mathbb{R}$ $(f : \mathbb{R} \to \mathbb{R}_+)$ takva da je*

$$F_X(x) = \int_{-\infty}^{x} f(t)d\lambda(t), \quad x \in \mathbb{R}. \tag{1.8}$$

Ako je X neprekidna slučajna varijabla, tada se funkcija f iz (1.8) zove **funkcija gustoće vjerojatnosti od X** ili, kraće, **gustoća od X**, i ponekad je označavamo s f_X.

1.5 Matematičko očekivanje. Varijanca

Definicija matematičkog očekivanja provodi se u tri koraka. Prvo se definira matematičko očekivanje jednostavne slučajne varijable, zatim nenegativne slučajne varijable, te na kraju općenite slučajne varijable.

Definicija 1.5.1. *Neka je X slučajna varijabla na $(\Omega, \mathcal{F}, \mathbb{P})$. X je **jednostavna slučajna varijabla** ako je njezino područje vrijednosti konačan skup.*

Neka je $(\Omega, \mathcal{F}, \mathbb{P})$ vjerojatnosni prostor. Označimo sa $\mathcal{K}$ skup svih jednostavnih slučajnih varijabli definiranih na Ω, a sa $\mathcal{K}_+$ skup svih nenegativnih funkcija iz $\mathcal{K}$.

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

Definicija 1.5.2. ***Matematičko očekivanje** od X ili, kraće, **očekivanje** od X koje označavamo sa $\mathbb{E}[X]$ definira se sa*

$$\mathbb{E}[X] = \sum_{k=1}^{n} x_k \mathbb{P}(A_k). \tag{1.9}$$

Neka je X nenegativna slučajna varijabla definirana na Ω. Tada postoji rastući niz $(X_n)_{n\in\mathbb{N}}$ nenegativnih jednostavnih slučajnih varijabli takav da je $X = \lim_{n\to\infty} X_n$. Niz $(\mathbb{E}[X_n])_{n\in\mathbb{N}}$ je rastući niz u $\mathbb{R}_+$, dakle postoji $\lim_{n\to\infty} \mathbb{E}[X_n]$ koji može biti jednak i $+\infty$.

Definicija 1.5.3. ***Matematičko očekivanje** od X ili, kraće, **očekivanje** od X definira se sa*

$$\mathbb{E}[X] = \lim_{n\to\infty} X_n. \tag{1.10}$$

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

Definicija 1.5.4. *Kažemo da* ***matematičko očekivanje*** *od X ili, kraće,* ***očekivanje*** *od X postoji ili da je definirano ako je barem jedna od veličina $\mathbb{E}[X^+]$, $\mathbb{E}[X^-]$ konačna, tj. vrijedi* $\min\{\mathbb{E}[X^+], \mathbb{E}[X^+]\} < +\infty$. *Tada po definiciji stavljamo*

$$\mathbb{E}[X] = \mathbb{E}[X^+] + \mathbb{E}[X^-]. \tag{1.11}$$

Neka je X slučajna varijabla na vjerojatnosnom prostoru $(\Omega, \mathcal{F}, \mathbb{P})$ i $r > 0$.

Definicija 1.5.5. $\mathbb{E}[X^r]$ *zovemo r-**ti moment** od X, a $\mathbb{E}[|X|^r]$ zovemo r-**ti apsolutni moment** od X.*

Po dogovoru stavljamo $\mathbb{E}[X^0] = \mathbb{E}[|X|^0] = 1$.

Definicija 1.5.6. *Neka $\mathbb{E}[X]$ postoji (tj. konačno je). Tada $\mathbb{E}[(X - \mathbb{E}[X])^r]$ zovemo r-**ti centralni moment** od X, a $\mathbb{E}[|X - \mathbb{E}[X]|^r]$ zovemo r-**ti apsolutni centralni moment** od X.*

Definicija 1.5.7. ***Varijanca*** *od X koju označavamo sa $VarX$ ili σ_x^2 jest drugi centralni moment od X, dakle*

$$VarX = \mathbb{E}[(X - \mathbb{E}[X])^2]. \tag{1.12}$$

Pozitivan drugi korijen iz varijance zovemo ***standardna devijacija*** *od X i označavamo sa σ_x.*

1.6 Primjeri slučajnih varijabli

Eksponencijalna distribucija

Neprekidna slučajna varijabla X ima **eksponencijalnu distribuciju** s parametrom $\lambda > 0$ ako joj je funkcija gustoće f dana sa

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

Logistička distribucija

Neka su $\alpha, \beta \in \mathbb{R}, \beta > 0$. Neprekidna slučajna varijabla X ima **logističku distribuciju** s parametrima α i β ako joj je funkcija gustoće f dana sa

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

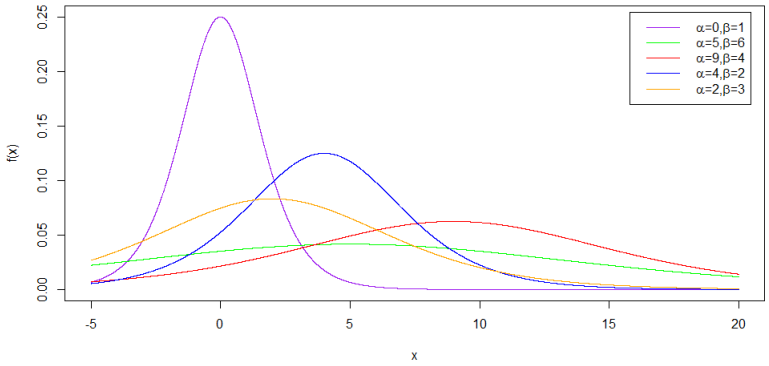
Za parametre $\alpha = 0$ i $\beta = 1$ kažemo da X ima **standardnu logističku distribuciju**.

Generalizirana logistička distribucija

Neka su $\alpha, \beta > 0$. Slučajna varijabla X ima **generaliziranu logističku distribuciju** ako joj je funkcija gustoće f dana sa

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

gdje je funkcija B definirana sa $B(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}dt, \quad x, y > 0.$



Slika 1.1: Graf funkcije gustoće logističke distribucije za razne vrijednosti parametara

Gumbelova distribucija

Neka su $\alpha \in \mathbb{R}$ i $\beta > 0$. Slučajna varijabla X ima **Gumbelovu distribuciju** s parametrima α i β ako joj je funkcija gustoće f dana sa

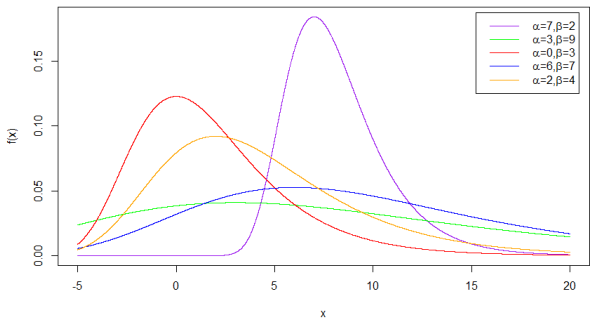
$$f(x) = \frac{1}{\beta} e^{-\frac{x-\alpha}{\beta} - e^{-\frac{x-\alpha}{\beta}}}, \quad x \in \mathbb{R},$$

Generalizirana Gumbelova distribucija

Neka je $p > 0$. Slučajna varijabla X ima **generaliziranu Gumbelovu distribuciju** ako joj je funkcija gustoće f dana sa

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

gdje je funkcija Γ definirana sa $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt, \quad x > 0.$



Slika 1.2: Graf funkcije gustoće Gumbelove distribucije za razne vrijednosti parametara

1.7 Linearna regresija

Neka su $x^{(1)}, x^{(2)}, ..., x^{(k)}$ kontrolirane (neslučajne) varijable i Y slučajna varijabla mjerena u ovisnosti o $x = (x^{(1)}, x^{(2)}, ..., x^{(k)})$, odnosno $Y = Y(x)$. **Linearni model** ovisnosti veličine Y o x zadan je sa

$$Y = \theta_0 + \theta_1 x^1 + ... + \theta_k x^k + \varepsilon,$$

gdje je ε slučajna pogreška, a $\theta_0, \theta_1, ..., \theta_k$ parametri modela.

Općenitiji zapis je

$$Y = \theta_0 + \theta_1 p_1(x) + ... + \theta_k p_k(x) + \varepsilon,$$

gdje su $1, p_1, p_2, ..., p_k$ linearno nezavisne realne funkcije. Cilj linearnog modela je odrediti parametre $\theta_0, \theta_1, ..., \theta_k$ koji će najbolje opisati traženi model.

Neka su $x_i = (x_i^{(1)}, x_i^{(2)}, ..., x_i^{(k)})$, za $i = 1, ..., n$, zadane vrijednosti od x takve da su barem dvije od njih različite, a $y_1, y_2, ..., y_n$ realizacije slučajne varijable Y. Minimiziramo funkciju

$$L(\theta_0, \theta_1, ..., \theta_k) = \sum_{i=1}^{n} [y_i - \theta_0 - \theta_1 x_i^{(1)} - ... - \theta_k x_i^{(k)}]^2$$

metodom najmanjih kvadrata.

Neka su $Y_i = \theta_0 + \theta_1 x_i^{(1)} + ... + \theta_k x_i^{(k)} + \varepsilon_i$, za $i = 1, 2, ..., n$, slučajne varijable. Pretpostavljamo da vrijede **Gauss-Markovljevi uvjeti**:

(i) $\mathbb{E}[\varepsilon_i] = 0, \forall i = 1, 2, ..., n$

(ii) $\mathbb{E}[\varepsilon_i \varepsilon_j] = 0, \forall i, j = 1, 2, ..., n$ takve da je $i \neq j$ (nekoreliranost)

(iii) $Var[\varepsilon_i] = \sigma^2 > 0, \forall i = 1, 2, ..., n.$

Neka je X **matrica dizajna**:

$$X = \begin{pmatrix} 1 & x_1^{(1)} & x_1^{(2)} & \dots & x_1^{(k)} \\ 1 & x_2^{(1)} & x_2^{(2)} & \dots & x_2^{(k)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n^{(1)} & x_n^{(2)} & \dots & x_n^{(k)} \end{pmatrix}$$

Označimo

$$\theta = \begin{pmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \vdots \\ \theta_k \end{pmatrix}, y = \begin{pmatrix} y_0 \\ y_1 \\ y_2 \\ \vdots \\ y_k \end{pmatrix}, Y = \begin{pmatrix} Y_0 \\ Y_1 \\ Y_2 \\ \vdots \\ Y_k \end{pmatrix}, \varepsilon = \begin{pmatrix} \varepsilon_0 \\ \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_k \end{pmatrix}.$$

Ovime smo dobili model

$$Y = X\theta + \varepsilon,$$

i tražimo minimum funkcije

$$L(\theta) = \|y - X\theta\|^2 .$$

Nepristrani procjenitelj za θ metodom najmanjih kvadrata je $\hat{\theta} = (X^T X)^{-1} X^T Y$, uz procjenu $(X^T X)^{-1} X^T y$, a procjenitelji za Y_i su $\hat{Y}_i = \hat{\theta}_0 + \hat{\theta}_1 x_i^{(1)} + ... + \hat{\theta}_k x_i^{(k)}$, za $i = 1, ..., n$.

Koeficijent determinacije je

$$R^2 = \frac{SSR}{S_{YY}} = 1 - \frac{SSE}{S_{YY}} \in [0, 1],$$

gdje su:

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

$$SSR = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$$

$$S_{YY} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

Koeficijent determinacije govori koliko je ukupne varijabilnosti objašnjeno našim regresijskim modelom. Što je R^2 bliži 1, to je prilagodba modelu bolja.

1.8 Mjere uspješnosti modela

Neka je X neko statističko obilježje. **Statističkom hipotezom** nazivamo neku pretpostavku o obilježju X, kao što je vrijednost parametra u modelu ili pripadnost određenoj distribuciji.

Pretpostavimo da nas zanima je li određena hipoteza točna. Tu hipotezu nazivamo **nultom hipotezom** i označavamo s H_0. Njoj suprotnu hipotezu nazivamo **alternativnom hipotezom** i označavamo s H_1. Odluke o odbacivanju nulte hipoteze provodimo statističkim testovima.

Ponekad se može dogoditi da je zaključak testa pogrešan. Pogrešku koju činimo kada odbacimo hipotezu H_0 ukoliko je ona istinita nazivamo **pogreškom prve vrste**, a pogrešku koju činimo kada ne odbacimo hipotezu H_0 ukoliko je istinita hipoteza H_1 nazivamo **pogreškom druge vrste**.

	Zaključak	
Točno je	ne odbaciti H_0	odbaciti H_0
H_0	✓	pogreška (I.)
H_1	pogreška (II.)	✓

Tablica 1.1: Pogreška prve i druge vrste

Definirajmo sada neke od mjera uspješnosti modela:

Osjetljivost ili **TPR** (eng. *true positive rate*) je postotak pozitivnih elemenata uzorka u odnosu na određeno stanje, odnosno CP (eng. *condition positive*) elemenata uzorka, koji su ispravno prepoznati kao pozitivni.

$$\text{TPR} = \frac{\text{broj stvarno pozitivnih}}{\text{broj stvarno pozitivnih + broj lažno negativnih}} = \frac{\text{TP}}{\text{TP+FN}} = \frac{\text{TP}}{\text{CP}}$$

Specifičnost ili **TNR** (eng. *true negative rate*) je postotak negativnih elemenata uzorka u odnosu na određeno stanje, odnosno CN (eng. *condition negative*) elemenata uzoraka, koji su ispravno prepoznati kao negativni.

$$\text{TNR} = \frac{\text{broj stvarno negativnih}}{\text{broj stvarno negativnih + broj lažno pozitivnih}} = \frac{\text{TN}}{\text{TN+FP}} = \frac{\text{TN}}{\text{CN}}$$

Preciznost ili **PPV** (eng. *positive predictive value*) je omjer broja stvarno pozitivnih elemenata uzorka i broja elemenata uzorka koji su modelom prepoznati kao pozitivni.

$$\text{PPV} = \frac{\text{broj stvarno pozitivnih}}{\text{broj stvarno pozitivnih + broj lažno pozitivnih}} = \frac{\text{TP}}{\text{TP+FP}}$$

Negativna prediktivna vrijednost ili **NPV** (eng. *negative predictive value*) je omjer broja stvarno negativnih elemenata uzorka i broja elemenata uzorka koji su modelom prepoznati kao negativni.

$$\text{NPV} = \frac{\text{broj stvarno negativnih}}{\text{broj stvarno negativnih + broj lažno negativnih}} = \frac{\text{TN}}{\text{TN+FN}}$$

		Predviđeno stanje		
	Ukupna populacija	Ocijenjeni pozitivno	Ocijenjeni negativno	
Stvarno stanje	Pozitivno stanje (CP)	TP (stvarno pozitivni)	FN (lažno negativni)	Osjetljivost (TPR)
	Negativno stanje (CN)	FP (lažno pozitivni)	TN (stvarno negativni)	Specifičnost (NPR)
		Preciznost (PPV)	Negativna prediktivna vrijednost (NPV)	

Tablica 1.2: Mjere uspješnosti modela

F_β**-score** jedna je od mjera uspješnosti modela. Računa se kao harmonijska sredina osjetljivosti i preciznosti modela, uz težinski faktor β.

$$F_\beta = \frac{1}{\alpha\frac{1}{PPV} + (1-\alpha)\frac{1}{TPR}} = \frac{(\beta^2+1)\cdot PPV \cdot TPR}{\beta^2 \cdot PPV + TPR}$$

Najčešće se koristi F_1**-score** ($\beta = 1$ ili $\alpha = \frac{1}{2}$):

$$F_1 = \frac{2}{\frac{1}{PPV} + \frac{1}{TPR}} = 2 \cdot \frac{PPV \cdot TPR}{PPV + TPR}$$

Sve navedene mjere uspješnosti modela postižu vrijednosti iz intervala $[0, 1]$, gdje 1 predstavlja najbolje ocijenjen model, a 0 najlošije ocijenjen model.

Pojmovi iz ovog poglavlja preuzeti su iz [1], [7], [8], [11], [12] i [14].

Poglavlje 2

Teorija grafova

2.1 Osnovni pojmovi

Definicija 2.1.1. ***Graf*** *G je uređeni par $G = (V, E)$, gdje je $V \neq 0$ skup* ***vrhova****, a E skup $2-$podskupova od V, koje zovemo* ***bridovi****.*

Napomena 2.1.2. *Prethodnu definiciju ponekad proširujemo tako da dopustimo* ***petlje*** *(bridove koji spajaju vrh sa samim sobom),* ***višestruke bridove*** *(više bridova između istog para vrhova) i* ***usmjerene bridove*** *(bridove s orijentacijom tako da idu od jednog vrha prema drugome).*

Usmjereni bridovi su, umjesto 2-podskupovima, reprezentirani uređenim parovima, dok kod višestrukih bridova skup E postaje multiskup.

Definicija 2.1.3. *Kažemo da je graf G:*

(i) ***usmjereni graf*** *ili* ***digraf*** *ako su mu bridovi usmjereni.*

(ii) ***multigraf*** *ako u njemu postoji višestruki brid.*

Napomena 2.1.4. *Ako želimo naglasiti da u grafu nema višestrukih ili usmjerenih bridova niti petlji, nazivamo ga* ***jednostavnim grafom****. U ovom radu, jednostavan graf nazivamo* ***0-1 graf****.*

Definicija 2.1.5. *Kažemo da su vrhovi $u, v \in V$ grafa $G = (V, E)$* ***susjedni*** *ako postoji brid $e = \{u, v\} \in E$.*

Definicija 2.1.6. *Kažemo da je graf $G = (V, E)$:*

(i) ***potpun*** *ako svaki par vrhova iz V čini brid.*

(ii) ***nulgraf*** *ako je skup bridova E prazan.*

Definicija 2.1.7. *Kažemo da je graf $G' = (V', E')$:*

(i) ***podgraf*** *grafa $G = (V, E)$ ako je $V' \subseteq V$ i $E' \subseteq E$.*

(ii) ***inducirani podgraf*** *grafa $G = (V, E)$ ako je G' podgraf od G i vrijedi da se skup E' sastoji od svih bridova iz G čija oba kraja leže u V'.*

(iii) ***razapinjući podgraf*** *grafa $G = (V, E)$ ako je $V' = V$.*

Definicija 2.1.8. ***Težinski graf*** *$G = (V, E)$ je graf s težinskom funkcijom $f : E \to \mathbb{R}(\mathbb{R}_0^+)$ na skupu bridova E.*

2.2 Put i povezanost

Definicija 2.2.1. *("kretanje" po grafu)*

(i) ***Šetnja*** *u grafu $G = (V, E)$ je niz $(v_0, e_1, v_1, e_2, v_2, \ldots, e_n, v_n)$, gdje je $e_i \in E$ brid $\{v_{i-1}, v_i\}$, $v_i \in V$, za $i = 1, 2, \ldots, n$. Kažemo da je to šetnja od v_0 do v_n.*

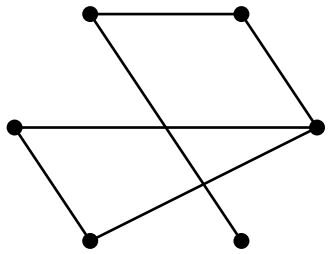
(ii) ***Staza*** *je šetnja u kojoj su svi bridovi različiti.*

(iii) ***Put*** *je šetnja u kojoj su svi vrhovi različiti (osim eventualno prvog i zadnjeg).*

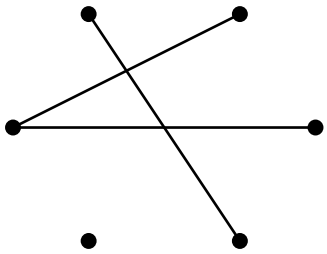
Definicija 2.2.2. *(Relacija ekvivalencije $\equiv$ na skupu vrhova V grafa $G = (V, E)$)*
Vrhovi $x, y \in V$ grafa $G = (V, E)$ su ***u relaciji****, odnosno $x \equiv y$, ako postoji put u grafu G od x do y.*

Definicija 2.2.3. ***Komponenta povezanosti*** *grafa $G = (V, E)$ je podgraf induciran klasom ekvivalencije $\equiv$ iz prethodne definicije.*

Definicija 2.2.4. *Kažemo da je graf $G = (V, E)$* ***povezan*** *ako postoji samo jedna komponenta povezanosti u grafu.*



Slika 2.1: Povezan graf



Slika 2.2: Graf s tri komponente povezanosti

2.3 Maksimalna klika

Definicija 2.3.1. ***Klika*** *u grafu $G = (V, E)$ je njegov potpuni podgraf koji se sastoji od barem dva vrha.*

Definicija 2.3.2. ***Maksimalna klika*** *u grafu $G = (V, E)$ je klika koja nije sadržana niti u jednoj većoj kliki, odnosno dodavanjem nekog vrha ona prestaje biti klika.*

Definicija 2.3.3. ***Najveća klika*** *u grafu $G = (V, E)$ je klika koja ima najveći broj vrhova.*

Napomena 2.3.4. *Najveća maksimalna klika u grafu G ujedno je i najveća klika u grafu G.*

Bron-Kerbosch algoritam

U ovom radu, za računanje maksimalnih klika korišten je Bron-Kerbosch algoritam (verzija s pivotiranjem). To je rekurzivni algoritam koji vraća sve maksimalne klike neusmjerenog grafa. Osmislili su ga nizozemski programeri Coenraad Bron i Joep Kerbosch, a objavljen je 1973. godine. Korišten algoritam prikazan je sljedećim pseudokodom (za više detalja pogledati [8]).

```
bronkerbosch(R, P, X):
    if P = ∅ and X = ∅:
        return R
    odaberi pivotni vrh p ∈ P ∪ X
    for v ∈ P \ N(p):          #N(p) su susjedni vrhovi vrha p
        bronkerbosch(R ∪ {v}, P ∩ N(v), X ∩ N(v))
        P := P \ {v}
        X := X ∪ {v}
```

Listing 2.1: Bron-Kerbosch pseudokod

Pojmovi iz ovog poglavlja preuzeti su iz [8], [9] i [13].

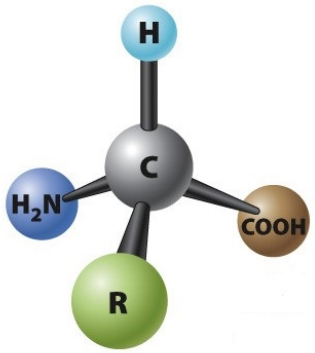
Poglavlje 3

Pojmovi iz bioinformatike

3.1 Osnovni biološki pojmovi

Proteini ili bjelančevine biološki su važni organski spojevi. Sastavni su dijelovi svake stanice, što ih čini osnovom života na Zemlji. Glavni su izvor tvari za izgradnju mišića, krvi, kože, kose, noktiju i unutarnjih organa. Također, odgovorni su i za kataliziranje metaboličkih reakcija, prijenos molekula unutar stanice, umnažanje i prepisivanje DNA, odgovaranje na podražaje i za razne druge funkcije u organizmu. Svi proteini nekog organizma koji nastaju kao posljedica ekspresije gena u određenom trenutku pod određenim uvjetima čine proteom.

Aminokiseline su organski spojevi sastavljeni od karboksilne skupine (-COOH), amino skupine (-NH_2) i bočnog lanca (R) po kojemu se one međusobno razlikuju. One su osnovne građevne jedinice svakog proteina.



Slika 3.1: Građa aminokiseline

Aminokiseline se međusobno spajaju u lance peptidnom vezom, pri čemu se izdvaja jedna molekula vode (H_2O). Peptidna veza je veza između amino skupine jedne aminokiseline i karboksilne skupine druge aminokiseline. U izgradnji proteina sudjeluje 20 različitih aminokiselina koje nazivamo standardnima.

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

Tablica 3.1: Standardne aminokiseline

Definicija 3.1.1. *Neka je* $\mathcal{A} = \{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\}$ *skup standardnih aminokiselina.* ***Protein*** X *je konačna uređena n-torka elemenata iz* $\mathcal{A}$*, odnosno* $X = (x_1, x_2, ..., x_n)$*, gdje je* $x_i \in \mathcal{A}$ *za svaki* $i = 1, ..., n$*,* $n \in \mathbb{N}$.

Napomena 3.1.2. *Zbog jednostavnosti, protein zapisujemo kao konačan niz slova iz* $\mathcal{A}$*. Neka je* $X = (M, G, S, Q, H, F, Y, V, Y, H, C, E, Q, R, I, S)$ *protein. Tada je jednostavni zapis proteina X jednak* `MGSQHFYVYHCEQRIS`.

GDSL lipaze

Lipaze su skupine enzima koji kataliziraju hidrolizu lipida, odnosno kataliziraju razgradnju molekula lipida u reakciji s vodom. Jedan primjer su upravo GDSL lipaze. One su podskupina lipotičkih enzima koja se razlikuje po tome što ne sadrži uobičajeni GxSxG motiv, gdje je x neka standardna aminokiselina. Karakteristične su po fleksibilnom katalitičkom mjestu koje mijenja svoju strukturu u prisutnosti različitih supstrata. To objašnjava njihovu katalitičku multifunkcionalnost i čini ih zanimljivima za istraživanja i primjene.

GDSL lipaze su nađene u raznim živim organizmima, a najviše u kopnenim biljkama. Otkriveno je da bi biljke mogle biti dobar izvor obećavajućih enzima za primjenu u hidrolizi i sintezi važnih spojeva od interesa u biotehnologiji, stoga je važno tražiti nove GDSL lipaze. No GDSL lipaze u sebi ne sadrže nužno GDSL motiv, što otežava njihov pronalazak te je bitno pronaći metodu pretraživanja koja bi olakšala taj problem.

3.2 Evolucija proteina. Poravnanje

Proučavanje evolucije proteina od velikog je interesa kako bismo bolje razumjeli njihovu ulogu u organizmu. Skup proteina koji potječu od istog pretka nazivamo **familija proteina**. **Mutacija** je trajna promjena genetskog materijala, a **evolucijskim procesom** nazivamo mutacijske događaje na slučajnom mjestu u proteinskom nizu. Ti mutacijski događaji dijele se u tri vrste:

- **insercija** - dodavanje jedne ili više aminokiselina postojećem proteinu
- **supstitucija** - zamjena jedne aminokiseline u proteinu drugom
- **delecija** - izostavljanje jedne ili više aminokiselina u postojećem proteinu

Poravnanje nizova je najtočniji prikaz evolucijskih procesa. Ono može biti globalno ili lokalno. Kod globalnog poravnanja poravnavamo cijele nizove (odnosno cijeli protein), dok kod lokalnog poravnanja poravnavamo samo dijelove s najvećom vjerojatnošću da dolaze od istog pretka.

Primjer 3.2.1. *Primjer mutacije na nizu aminokiselina* `PETAK`*:*

- `PRETAK` - *insercija: aminokiselina R dodana je na drugo mjesto u nizu*
- `LETAK` - *supstitucija: aminokiselina P zamijenjena je aminokiselinom L*
- `PETA` - *delecija: aminokiselina K izbačena je s kraja niza*

Označimo s _ mjesto delecije aminokiseline u jednom nizu, odnosno mjesto insercije aminokiseline na tom mjestu u drugom nizu. Simbol _ nazivamo ***praznina*** *(eng. gap).*

U ovom slučaju, za svaki od nizova znamo mutacijski događaj koji je doveo do njegova nastanka. Na taj način dobivamo nedvosmislena poravnanja:

- *insercija:*
  `P_ETAK`
  `PRETAK`
- *supstitucija:*
  `PETAK`
  `LETAK`
- *delecija:*
  `PETAK`
  `PETA_`

3.3 BLOSUM matrica. BLOSUM score

Definicija 3.3.1. *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. (O BLOSUM matrici više u [5]). Bazirana je na sljedećoj formuli:*

$$B(i, j) = \left\lfloor \log \frac{P(a_i \leftrightarrow b_j \,|\, M)}{P(a_i, b_j \,|\, R)} \right\rfloor, \quad a_i, b_j \in \mathcal{A}, \tag{3.1}$$

gdje je $\mathcal{A}$ skup svih standardnih aminokiselina, dok su a_i i b_j aminokiseline pridružene i-tom i j-tom mjestu niza. M je model koji pretpostavlja da aminokiseline a_i i b_j imaju zajedničkog pretka, a R je random model koji pretpostavlja nezavisnost aminokiselina pa vrijedi $P(a_i, b_j \,|\, R) = P(a_i \,|\, R) \cdot P(b_j \,|\, R)$. Distribucija standardnih aminokiselina uz model R dana je sa:

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

Definicija 3.3.2. *BLOSUM score s je rezultat koji odgovara sličnosti (ili povezanosti) dvaju nizova aminokiselina. Što je BLOSUM score veći, nizovi aminokiselina su sličniji.*

Neka su X_1 i X_2 nizovi aminokiselina i $\bar{X}_1 = (\bar{x}_{11}, \bar{x}_{21}, ..., \bar{x}_{n1})$, $\bar{X}_2(\bar{x}_{12}, \bar{x}_{22}, ..., \bar{x}_{n2})$ njihovo poravnanje duljine n. Tada je $s(\bar{X}_1, \bar{X}_2)$ jednak sumi score-ova po komponentama, gdje je

$$s(\bar{x}_{i1}, \bar{x}_{i2}) = \begin{cases} B(\bar{x}_{i1}, \bar{x}_{i2}), & \bar{x}_{i1}, \bar{x}_{i2} \in \mathcal{A} \\ -8, & \textit{inače}. \end{cases}$$

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	5	−2	−1	−2	−1	−1	−1	0	−2	−1	−2	−1	−1	−3	−1	1	0	−3	−2	0
R	−2	7	−1	−2	−4	1	0	−3	0	−4	−3	3	−2	−3	−3	−1	−1	−3	−1	−3
N	−1	−1	7	2	−2	0	0	0	1	−3	−4	0	−2	−4	−2	1	0	−4	−2	−3
D	−2	−2	2	8	−4	0	2	−1	−1	−4	−4	−1	−4	−5	−1	0	−1	−5	−3	−4
C	−1	−4	−2	−4	13	−3	−3	−3	−3	−2	−2	−3	−2	−2	−4	−1	−1	−5	−3	−1
Q	−1	1	0	0	−3	7	2	−2	1	−3	−2	2	0	−4	−1	0	−1	−1	−1	−3
E	−1	0	0	2	−3	2	6	−3	0	−4	−3	1	−2	−3	−1	−1	−1	−3	−2	−3
G	0	−3	0	−1	−3	−2	−3	8	−2	−4	−4	−2	−3	−4	−2	0	−2	−3	−3	−4
H	−2	0	1	−1	−3	1	0	−2	10	−4	−3	0	−1	−1	−2	−1	−2	−3	2	−4
I	−1	−4	−3	−4	−2	−3	−4	−4	−4	5	2	−3	2	0	−3	−3	−1	−3	−1	4
L	−2	−3	−4	−4	−2	−2	−3	−4	−3	2	5	−3	3	1	−4	−3	−1	−2	−1	1
K	−1	3	0	−1	−3	2	1	−2	0	−3	−3	6	−2	−4	−1	0	−1	−3	−2	−3
M	−1	−2	−2	−4	−2	0	−2	−3	−1	2	3	−2	7	0	−3	−2	−1	−1	0	1
F	−3	−3	−4	−5	−2	−4	−3	−4	−1	0	1	−4	0	8	−4	−3	−2	1	4	−1
P	−1	−3	−2	−1	−4	−1	−1	−2	−2	−3	−4	−1	−3	−4	10	−1	−1	−4	−3	−3
S	1	−1	1	0	−1	0	−1	0	−1	−3	−3	0	−2	−3	−1	5	2	−4	−2	−2
T	0	−1	0	−1	−1	−1	−1	−2	−2	−1	−1	−1	−1	−2	−1	2	5	−3	−2	0
W	−3	−3	−4	−5	−5	−1	−3	−3	−3	−3	−2	−3	−1	1	−4	−4	−3	15	2	−3
Y	−2	−1	−2	−3	−3	−1	−2	−3	2	−1	−1	−2	0	4	−3	−2	−2	2	8	−1
V	0	−3	−3	−4	−1	−3	−3	−4	−4	4	1	−3	1	−1	−3	−2	0	−3	−1	5

Slika 3.2: BLOSUM matrica

Primjer 3.3.3. *(Računanje BLOSUM score-a)*

(i) Neka su zadani nizovi aminokiselina FVFGDSL *i* DHILKGQ*. Želimo izračunati sličnost, odnosno score tih nizova. Radi jednostavnosti, zapišimo ih jedan ispod drugog:*

FVFGDSL
DHILKGQ

$s(\mathtt{FVFGDSL}, \mathtt{DHILKGQ}) =$
$= s(\mathtt{F},\mathtt{D}) + s(\mathtt{V},\mathtt{H}) + s(\mathtt{F},\mathtt{I}) + s(\mathtt{G},\mathtt{L}) + s(\mathtt{D},\mathtt{K}) + s(\mathtt{S},\mathtt{G}) + s(\mathtt{L},\mathtt{Q})$
$= -5 - 4 + 0 - 4 - 1 + 0 - 2 = -16$

(ii) Izračunajmo score poravnanja nizova aminokiselina PETAK *i* PETA *iz Primjera 3.2.1.*

PETAK
PETA_

$s(\mathtt{PETAK}, \mathtt{PETA\_}) =$
$= s(\mathtt{P},\mathtt{P}) + s(\mathtt{E},\mathtt{E}) + s(\mathtt{T},\mathtt{T}) + s(\mathtt{A},\mathtt{A}) + s(\mathtt{K},\_)$
$= 10 + 6 + 5 + 5 - 8 = 18$

3.4 Optimalno poravnanje nizova

Kako bismo odredili optimalno poravnanje ukoliko ne znamo događaje koji su doveli do nastanka nekog proteina, promatramo score funkciju dvaju poravnatih nizova. Optimalno poravnanje bit će ono koje ima najveći score.

Neka je $\bar{X}_1$, $\bar{X}_2$ optimalno poravnanje nizova X_1 i X_2. Kako je score poravnanja jednak sumi score-ova po komponentama, ukoliko "prerežemo" to poravnanje, oba dobivena dijela će i dalje biti optimalno poravnata.

Pretpostavimo da je $\bar{X}_1 = (\bar{x}_{11}, \bar{x}_{21}, ..., \bar{x}_{k1})$, $\bar{X}_2 = (\bar{x}_{12}, \bar{x}_{22}, ..., \bar{x}_{j2})$ optimalno poravnanje nizova X_1 i X_2 . Prerežimo to poravnanje na zadnjem mjestu. Tada je

$$F(k,j) = max \begin{cases} s(x_k, y_j) + F(k-1, j-1) \\ -8 + F(k, j-1) \\ -8 + F(k-1, j), \end{cases}$$

gdje je $F(k, j)$ score optimalnog poravnanja.

Vrijednosti funkcije F zapisujemo u obliku tablice, te da bismo odredili optimalno poravnanje, radimo traceback - vraćamo se istim putem kojim smo i došli.

Ovaj algoritam naziva se Needleman-Wunsch algoritam i njime dobivamo optimalno globalno poravnanje dvaju nizova.

Primjer 3.4.1. *Neka su zadani nizovi aminokiselina* `PGIW` *i* `GW`*. Izračunajmo optimalno globalno poravnanje tih nizova.*

Zapišimo nizove u tablicu i dodajmo početne uvjete na sljedeći način:

		P	G	I	W
	0	-8	-16	-24	-32
G	-8				
W	-16				

Sada, izračunajmo vrijednost sljedećeg elementa tablice po prethodnoj formuli:

$$F(1,1) = max\begin{cases} s(P,G) + F(0,0) \\ -8 + F(1,0) \\ -8 + F(0,1) \end{cases} = max\begin{cases} -2 + 0 \\ -8 - 8 \\ -8 - 8 \end{cases} = -2$$

Unesimo dobiveni rezultat u tablicu:

		P	G	I	W
	0	-8	-16	-24	-32
G	-8	-2			
W	-16				

Nastavljamo dalje istim postupkom dok ne ispunimo cijelu tablicu, a zatim radimo traceback: krećemo unazad od donjeg desnog ruba i označavamo put kojim smo došli do njega.

		P	G	I	W
	0	-8	-16	-24	-32
G	-8	-2	0	-8	-16
W	-16	-10	-5	-3	7

Ovime smo dobili score optimalnog poravnanja (u donjem desnom kutu tablice)

$$s(\texttt{PGIW}, \texttt{GW}) = 7$$

i optimalno globalno poravnanje:

```
PGIW
_G_W
```

Smith-Waterman algoritam

Smith-Waterman algoritam je verzija Needleman-Wunsch algoritma kojom se traži optimalno lokalno poravnanje dvaju nizova aminokiselina. Razlika je u tome što u funkciji F ne dopuštamo negativne vrijednosti, tj. funkcija je oblika

$$F(k,j) = max \begin{cases} s(x_k, y_j) + F(k-1, j-1) \\ -8 + F(k, j-1) \\ -8 + F(k-1, j) \\ 0 \end{cases}$$

Da bismo dobili optimalno lokalno poravnanje, krećemo od najvećeg elementa tablice i radimo traceback dok ne naiđemo na element jednak nuli.

Primjer 3.4.2. *Neka su zadani nizovi aminokiselina* `LIGPEWA` *i* `PW`*. Izračunajmo optimalno lokalno poravnanje tih nizova.*

Zapišimo nizove u tablicu kao i u prethodnom primjeru. U ovom slučaju, svi početni uvjeti jednaki su 0.

		L	I	G	P	E	W	A
	0	0	0	0	0	0	0	0
P	0							
W	0							

Izračunajmo ostale vrijednosti u tablici po gornjoj formuli i napravimo traceback od elementa tablice s najvećom vrijednošću do prvog elementa s vrijednošću nula.

		L	I	G	P	E	W	A
	0	0	0	0	0	0	0	0
P	0	0	0	0	10	2	0	0
W	0	0	0	0	2	7	17	9

Dobili smo da je score optimalnog lokalnog poravnanja (u donjem desnom kutu tablice)

$$s(\texttt{LIGPEWA}, \texttt{GW}) = 9,$$

a optimalno lokalno poravnanje:

```
PEW
P_W
```

3.5 Iterativno pretraživanje proteoma

Da bismo dobili nizove aminokiselina za analizu, iterativno ćemo pretražiti željeni proteom. Želimo, za neki upit, dobiti sve nizove aminokiselina koji su mu dovoljno slični s obzirom na neku funkciju sličnosti. To radimo pomoću IGLOSS servera, čiji je princip rada detaljno opisan u izvoru [10]. Na serveru ćemo unijeti proteom, željeni upit, skalu pretraživanja i maksimalni broj iteracija, te ćemo dobiti odgovor, tj. skup nizova aminokiselina koji zadovoljavaju tražene uvjete.

Upit (ili motiv) je kraći niz aminokiselina (najčešće duljine od 5 do 20). Za ocjenu sličnosti korištena je funkcija LLR (log likelihood ratio) ocijenjena pomoću logističke distribucije, a za traženi skup uzimaju se u obzir oni nizovi čija je ocjena veća ili jednaka od skale pretraživanja. U prvoj iteraciji, pretražujemo proteom kako bismo dobili nizove aminokiselina dovoljno slične zadanom upitu. Proces se nastavlja pretraživanjem proteoma sa cijelim skupom nizova dobivenim iz prethodnih iteracija. Iteriranje završava kada se skup nizova aminokiselina ne promijeni u odnosu na skup iz prethodne iteracije, ili kada je postignut maksimalni broj iteracija.

Skala pretraživanja je parametar koji određuje koji nizovi aminokiselina će biti odabrani kao dovoljno slični u jednoj iteraciji pretraživanja. Što je skala veća, odabrani su sličniji nizovi. Maksimalne ocjene sličnosti u iteraciji pretraživanja su logistički distribuirane. Skala je parametar te distribucije koji određuje koliko puta ćemo se odmaknuti za parametar β od prosječne ocjene sličnosti. Sve ocjene koje su veće od $\alpha + skala \cdot \beta$ smatraju se statistički značajnima. U standardnoj logističkoj distribuciji statistički su značajne sve ocjene sličnosti veće od skale.

Poglavlje 4

Traženje očekivane sličnosti proteina

U ovom poglavlju analizirat ćemo sličnost između proteina iz iste proteinske familije. Želimo pronaći očekivane parametre za funkciju sličnosti kako bismo te parametre mogli koristiti za računanje očekivanog score-a poravnanja proteina. Očekivani score ćemo računati za sve parove proteina iz odgovora dobivenog iterativnim pretraživanjem te metodom najveće maksimalne klike odrediti pripadnost određenoj proteinskoj familiji, u našem slučaju GDSL lipazama.

4.1 Motivacija

Pretraživanje motiva bitan je dio analize proteina. Koristi se u razne svrhe, kao što je klasifikacija proteina u proteinske familije te predviđanje strukture proteina. Cilj te pretrage motiva je minimizirati pogrešku prve i druge vrste, odnosno maksimizirati F_1-score pretraživanja. Drugim riječima, cilj je pronaći što više relevantnih motiva (motiva s liste pozitivaca), a da smo pri tome pronašli što manje lažno pozitivnih motiva.

Dobri rezultati dobiveni su analizom pomoću najveće maksimalne klike $0-1$ grafa pridruženog motivima duljine 10 tipičnima za GDSL lipaze. Ti motivi dobiveni su iterativnim pretraživanjem proteoma, a njihova međusobna sličnost iznosila je oko 68%. Pokazano je da očekivana vrijednost funkcije sličnosti dvaju takvih motiva iznosi 25 (za više detalja, pogledati [8]). Rezultati dobiveni najvećom maksimalnom klikom značajno su povećali PPV, što znači da je u kliki sadržan veći udio stvarnih pozitivaca nego u odgovoru dobivenom iterativnim pretraživanjem.

Ideja u ovom radu je proširiti prethodno opisanu metodu traženja najveće maksimalne klike $0-1$ grafa. Umjesto motiva duljine 10, grafu ćemo pridružiti cijele proteine, a očekivanu vrijednost funkcije sličnosti odredit ćemo pomoću optimalnog lokalnog poravnanja proteina i linearne regresije.

4.2 Određivanje očekivanih parametara za funkciju sličnosti

Kako bismo odredili očekivane parametre za funkciju sličnosti dvaju proteina, koristimo listu od 118 proteina iz iste proteinske familije, te listu od 100 nasumično odabranih proteina koji se ne nalaze u toj familiji.

Za početak, odredili smo optimalno lokalno poravnanje za svaka dva proteina iz iste familije koristeći Smith-Waterman algoritam. Iz toga smo izvukli duljinu poravnanja, broj praznina u poravnanju i dobiveni score poravnanja, kao što je pokazano u sljedećem primjeru. Podatke dobivene poravnanjima želimo iskoristiti kako bismo linearnom regresijom dobili vrijednost očekivanog score-a poravnanja dvije aminokiseline $s_{ocek}(a_i, a_j)$, za $a_i, a_j \in \mathcal{A}$, te očekivani score poravnanja aminokiseline s prazninom $s_{ocek}(a_i, \_)$, za $a_i \in \mathcal{A}$.

Primjer 4.2.1. *Neka su* $X_1 = (..., G, D, S, L, I, L, M, G, E, I, G, G, N, D, Y, N, Y, P, F, F, E, G, K, S, I, N, E, I, K, E, L, V, P, L, I, V, K, A, I, S, S, A, I, V, D, L, I, D, L, G, G, K, T, F, L, V, P...)$ *i* $X_2 = (..., P, S, L, V, I, V, Y, F, G, G, N, D, S, M, A, P, H, S, S, G, L, G, P, H, V, P, L, T, E, Y, V, D, N, M, K, K, I, A, L, H, L, Q, S, L, S, D, F, T, R, I, I, F, L, S, S, P, ...)$ *proteini, gdje je* X_1 *duljine 384, a* X_2 *duljine 256.*

Tada je optimalno lokalno poravnanje dobiveno Smith-Watermanovim algoritmom

```
SLILMGEIGGNDYNYPFFEGKSIN_EIKELVPLIVKAISSAIVDLIDLGGKTFL
SLVIV_YFGGNDSMAPHSSGLGPHVPLTEYVD_NMKKIALHLQSLSDFTRIIFL
```

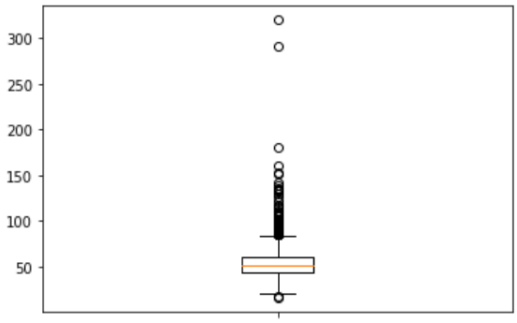
Iz tog poravnanja čitamo:

- ***duljinu poravnanja*** $n_{uk} = 54$
- ***broj praznina*** $n_g = 3$
- ***score poravnanja*** $s = 51$

Sada iz ovih podataka možemo dobiti i broj mjesta na kojima su poravnate dvije aminokiseline kojeg ćemo označiti sa n_a.

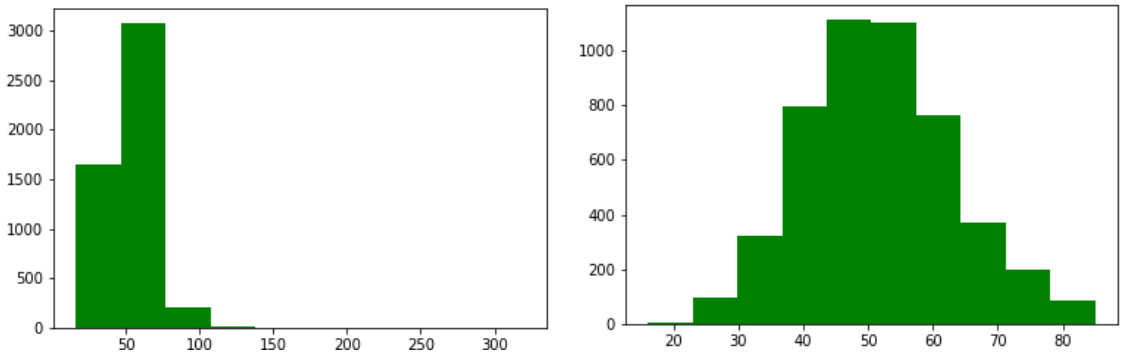
U daljnjem istraživanju, želimo pomoću dobivenih očekivanih parametara odrediti koji proteini pripadaju promatranoj proteinskoj familiji. Pri tome pretpostavljamo da su proteini iz iste familije međusobno sličniji nego proteini iz različitih familija. Dakle, iz podataka dobivenih optimalnim lokalnim poravnanjima proteina iz iste proteinske familije želimo izbaciti poravnanja čiji je score dovoljno velik da bismo za bilo koji protein bili gotovo sigurni da pripada toj proteinskoj familiji. To ćemo postići analizom nasumično odabranih proteina.

Promotrimo optimalna lokalna poravnanja 100 nasumično odabranih proteina. Deskriptivnom statistikom ustvrdili smo da je najveći score poravnanja bio 320, a očekivanje 53.2 sa standardnom devijacijom 31.4.



Slika 4.1: Boxplot nasumično odabranih proteina

Promatrat ćemo dva skupa podataka: jedan dobiven poravnanjima proteina iz iste proteinske familije bez elemenata čiji je score veći od 320, i drugi dobiven poravnanjima bez elemenata čiji je score veći od 85 (suma očekivanja i standardne devijacije).



Slika 4.2: Histogrami promatranih skupova podataka

Primijenimo sada linearnu regresiju na očišćene skupove podataka. Naš model je oblika

$$s = n_a \cdot s_a + n_g \cdot s_g, \tag{4.1}$$

gdje smo sa s_a označili očekivani score poravnanja dviju aminokiselina, a sa s_g očekivani score poravnanja aminokiseline s prazninom. Nakon uvrštavanja naših podataka, dobivamo da su očekivani parametri $s_a = 1.2465$ i $s_g = -2.5573$, ukoliko promatramo poravnanja sa score-om do 320, te $s_a = 0.835$ i $s_g = -2.66$ ukoliko promatramo poravnanja sa score-om do 85.

4.3 Pretraživanje odgovora

Računanje očekivanog score-a

Promotrimo odgovor dobiven iterativnim pretraživanjem proteoma s obzirom na neki zadani upit. Želimo za svaki par nizova iz odgovora pronaći očekivani score pomoću parametara dobivenih u prethodnom poglavlju.

Ponovno ćemo određivati optimalna lokalna poravnanja i pamtiti njihov score (s), duljinu (n_{uk}) i broj praznina (n_g), no ovaj put za nizove iz odgovora. Očekivani score poravnanja računamo pomoću formule:

$$s_{ocek} = n_a \cdot s_a + n_g \cdot s_g, \tag{4.2}$$

gdje su s_a i s_g parametri dobiveni u prethodnom poglavlju, a $n_a = n_{uk} - n_g$ broj mjesta na kojima su poravnate dvije aminokiseline.

Maksimalna klika

Koristeći izračunati score poravnanja, našem odgovoru pridružit ćemo potpun težinski graf. Skup vrhova bit će cijeli odgovor, a težinska funkcija bit će izračunati score poravnanja s između proteina koji čine vrhove tog brida.

Dobivenom težinskom grafu pridružujemo $0-1$ graf. Na bridove čiji je izračunati score veći od očekivanog, stavljamo težinu 1, a na ostale 0. Dakle, iz grafa smo uklonili bridove čiji je score manji od očekivanog. Sada se naš problem svodi na traženje najvećeg potpunog podgrafa dobivenog $0-1$ grafa, odnosno na traženje najveće klike $0-1$ grafa.

Pomoću Bron-Kerbosch algoritma odredit ćemo sve maksimalne klike našeg $0-1$ grafa, a zatim ćemo od dobivenih klika naći najveću (ili jednu od najvećih, ukoliko ne postoji jedinstvena najveća maksimalna klika).

Cilj

Naša pretpostavka je da će proteini sadržani u najvećoj maksimalnoj kliki biti iz iste proteinske familije. No, ukoliko je izračunati prag previsok, dobiveni $0-1$ graf imat će puno komponenata povezanosti, te će najveća maksimalna klika biti mala. Na taj način izgubit ćemo velik broj proteina iz tražene familije. S druge strane, ukoliko je izračunati prag prenizak, graf će biti povezaniji te će najveća maksimalna klika obuhvatiti i velik broj proteina koji ne pripadaju toj familiji. Time ćemo dobiti veliku pogrešku prve vrste.

Dakle, cilj je pronaći optimalne očekivane parametre kako bi najveća maksimalna klika sadržavala što više proteina iz promatrane familije, ali i što manje proteina koji ne pripadaju toj familiji.

Poglavlje 5

Rezultati

5.1 Korišteni proteomi

Biološka lista pozitivaca nekog proteoma je lista GDSL lipaza u tom proteomu. Proteine, odnosno GDSL lipaze, s te liste označavamo sa CP (eng. *condition positive*), a sve ostale proteine u proteomu označavamo sa CN (eng. *condition negative*).

Proteine iz odgovora dobivenog iterativnim pretraživanjem smatramo **pozitivcima**. Dijelimo ih na TP (eng. *true positives*), odnosno one koji se nalaze na listi CP-ova, i FP (eng. *false positives*), odnosno one koji se ne nalaze na listi CP-ova. Skala korištena u iterativnom pretraživanju je 7.5.

Talijin uročnjak

Talijin uročnjak (lat. *Arabidopsis thaliana*) je malena biljka koja se često koristi kao organizam za modeliranje u genetici i biologiji. Pripada porodici *Brassicaceae*, koju čine i kultivirane vrste kao što su kupus i rotkvica.

Slika 5.1: Talijin uročnjak

Talijin uročnjak pogodan je za razna istraživanja kao prva biljka s potpuno sekvencioniranim genomom. Za svaki protein u njezinom proteomu poznato je kojoj familiji pripada. U prethodnom poglavlju koristili smo upravo proteom *Arabidopsis thaliana* i njezinu biološku listu pozitivaca kako bismo odredili očekivane parametre za score poravnanja pomoću linearne regresije. Ta biološka lista pozitivaca sastoji se od 118 proteina.

Ostali proteomi

Rajčica (lat. *Solanum lycopersicum*) pripada porodici *Solanaceae* i jedna je od najvažnijih prehrambenih biljaka. Bogat je izvor vitamina i minerala. Njezina biološka lista pozitivaca sastoji se od 108 proteina.

Slika 5.2: Cvijet rajčice

Azijska riža (lat. *Oryza sativa*) jedna je od najvažnijih prehrambenih namirnica koju koristi više od trećine svjetske populacije. Pripada porodici *Poaceae*, a njenu biološku listu pozitivaca čini 155 proteina.

Slika 5.3: Azijska riža

Šećerna repa (lat. *Beta vulgaris*) je prehrambena biljka koja pripada porodici *Amaranthaceae*. Bogata je C vitaminom i šećerom, a uzgaja se za proizvodnju šećera. Njezinu biološku listu pozitivaca čini 86 proteina.

Slika 5.4: Šećerna repa

5.2 Rezultati - *Arabidopsis thaliana*

Usporedba parametara

Pogledajmo rezultate dobivene iterativnim pretraživanjem i klikom za upit `FVFGDSLSDA`.

CP=118	Broj pozitivaca	TP	PPV	TPR	F_1-score
Odgovor	145	102	0.703	0.864	0.775

Tablica 5.1: Rezultati dobiveni iterativnim pretraživanjem

Usporedili smo rezultate dobivene klikom za razne parametre:

Parametri	Broj pozitivaca	TP	PPV	TPR	F_1-score
2.47, −11.47	33	31	0.939	0.263	0.411
1.2465, −2.5573	72	70	0.972	0.593	0.737
1.13, −3.27	93	84	0.903	0.712	0.796
0.835, −2.66	110	99	0.900	0.839	0.868

Tablica 5.2: Rezultati dobiveni najvećom maksimalnom klikom

Vidimo da je F_1-score najbolji kod parametara dobivenih linearnom regresijom za podatke čija je sličnost manja od 85. U usporedbi s iterativnim pretraživanjem, povećali smo PPV za 0.197, a smanjili TPR tek za 0.025, čime je F_1-score porastao za 0.093. Također, vidimo da je u najvećoj kliki ostalo sačuvano čak 97.06% stvarnih pozitivaca iz odgovora.

S druge strane, parametri dobiveni linearnom regresijom čija je sličnost do 320 značajno su poboljšali PPV (za 0.269), ali je i TPR znatno pao (za 0.271) pa je F_1-score manji nego kod iterativnog pretraživanja. Klika je sačuvala 68.63% stvarnih pozitivaca iz odgovora.

Upit FVFNDSLSDA

Usporedimo rezultate za odgovor dobiven iterativnim pretraživanjem proteoma za zadani upit `FVFNDSLSDA` sa rezultatima dobivenim najvećom maksimalnom klikom za oba para izračunatih parametara. Vidimo da smo sa obje najveće maksimalne klike popravili F_1-score. PPV za kliku dobivenu parametrima 1.2465 i −2.5573 iznosi 1, što znači da se u odgovoru dobivenom klikom nalaze samo elementi s biološke liste pozitivaca, no TPR je pao za 0.152. U kliki je ostalo sačuvano 64.71% stvarnih pozitivaca iz odgovora.

Uočimo da je za parametre 0.835 i −2.66 TPR ostao isti, dok je PPV porastao za više od dvostruke vrijednosti dobivene iterativnim pretraživanjem, čime smo popravili F_1-score za 0.151. Također, svi stvarni pozitivci iz odgovora su ostali sačuvani u najvećoj maksimalnoj kliki.

CP=118	Broj pozitivaca	TP	PPV	TPR	F_1-score
Odgovor	122	51	0.418	0.432	0.425
Najveća klika (0.835, −2.66)	59	51	0.864	0.432	0.576
Najveća klika (1.2465, −2.5573)	33	33	1.000	0.280	0.438

Tablica 5.3: Rezultati za upit FVFNDSLSDA

Upit VFFGDSLSDN

Promotrimo rezultate za upit `VFFGDSLSDN`. Vidimo da najveća maksimalna klika za parametre 1.2465 i −2.5573 nije našla niti jednog stvarnog pozitivca iz odgovora. No najveća maksimalna klika dobivena parametrima 0.835 i −2.66 povećala je PPV za više od 0.6, dok je TPR pao za samo 0.008. Time je F_1-score porastao za 0.172. Uočimo i da je u kliki ostalo sačuvano 97.5% stvarnih pozitivaca iz odgovora.

CP=118	Broj pozitivaca	TP	PPV	TPR	F_1-score
Odgovor	138	40	0.290	0.339	0.313
Najveća klika (0.835, −2.66)	43	39	0.907	0.331	0.485
Najveća klika (1.2465, −2.5573)	34	0	0	0	–

Tablica 5.4: Rezultati za upit VFFGDSLSDN

5.3 Rezultati na ostalim proteomima za upit FVFGDSLSDA

Azijska riža

Pogledajmo rezultate dobivene pretraživanjem proteoma azijske riže. Za kliku dobivenu parametrima 1.2465 i −2.5573 PPV je porastao za više od 0.3, no TPR je značajno pao pa je F_1-score manji za 0.111. Najveća maksimalna klika sačuvala je 54.78% stvarnih pozitivaca iz odgovora.

Koristeći parametre 0.835 i −2.66 također smo povećali PPV za više od 0.3, ali je pri tome TPR pao za samo 0.007. Time se F_1-score značajno povećao, za čak 0.146. Također, u najvećoj maksimalnoj kliki ostalo je sačuvano 99.13% stvarnih pozitivaca iz odgovora.

CP=155	Broj pozitivaca	TP	PPV	TPR	F_1-score
Odgovor	182	115	0.632	0.742	0.683
Najveća klika (0.835, −2.66)	120	114	0.950	0.735	0.829
Najveća klika (1.2465, −2.5573)	65	63	0.969	0.406	0.572

Tablica 5.5: Rezultati - Azijska riža

Rajčica

I u ovom slučaju vidimo da nam najveća maksimalna klika za parametre 0.835, −2.66 daje najbolje rezultate. PPV je porastao za 0.272, dok se TPR smanjio za samo 0.01, što je popravilo F_1-score za 0.129. Nadalje, u dobivenoj kliki ostalo je sačuvano 98.9% stvarnih pozitivaca iz odgovora.

S parametrima 1.2465 i −2.5573 PPV je porastao za 0.268, no TPR je pao za 0.241, te se F_1-score smanjio za 0.028 u odnosu na rezultat dobiven iterativnim pretraživanjem. U kliki je ostalo sačuvano 71.43% stvarnih pozitivaca iz odgovora.

CP=108	Broj pozitivaca	TP	PPV	TPR	F_1-score
Odgovor	127	91	0.717	0.843	0.775
Najveća klika (0.835, −2.66)	91	90	0.989	0.833	0.904
Najveća klika (1.2465, −2.5573)	66	65	0.985	0.602	0.747

Tablica 5.6: Rezultati - Rajčica

Šećerna repa

Kao što vidimo u donjoj tablici, najbolje rezultate ponovno dobivamo za parametre 0.835 i −2.66. U najvećoj maksimalnoj kliki ostali su sačuvani svi nizovi iz odgovora, a PPV je porastao za 0.29. Time je F_1-score porastao za 0.125 u odnosu na rezultate dobivene iterativnim pretraživanjem.

Parametri 1.2465 i −2.5573 digli su PPV na 1, no klika je sačuvala tek 60.66% stvarnih pozitivaca iz odgovora. TPR se smanjio za 0.28, a F_1-score za 0.093 u odnosu na rezultate dobivene iterativnim pretraživanjem.

CP=86	Broj pozitivaca	TP	PPV	TPR	F_1-score
Odgovor	90	61	0.678	0.710	0.694
Najveća klika (0.835, −2.66)	63	61	0.968	0.710	0.819
Najveća klika (1.2465, −2.5573)	37	37	1	0.430	0.601

Tablica 5.7: Rezultati - Šećerna repa

5.4 Analiza rezultata

Nakon analize strukture odgovora dobivenog iterativnim pretraživanjem proteoma talijinog uročnjaka u odnosu na upite `FVFGDSLSDA`, `FVFNDSLSDA` i `VFFGDSLSDN`, te proteoma rajčice, azijske riže i šećerne repe u odnosu na upit `FVFGDSLSDA`, možemo donijeti slične zaključke.

U svim slučajevima, najveća maksimalna klika dobivena za parametre 0.835 i −2.66 znatno je popravila F_1-score u odnosu na iterativno pretraživanje. Također, u svim primjerima klika je sačuvala više od 97% odgovora dobivenog iterativnim pretraživanjem proteoma.

Najveća maksimalna klika dobivena parametrima 1.2465 i −2.5573 u većini je slučajeva znatno povećala PPV, ali i znatno smanjila TPR, što je dovelo i do smanjenja F_1-scorea. Možemo zaključiti da su ti parametri previsoki te da zbog toga klika zanemaruje dosta stvarnih pozitivaca iz upita.

Literatura

[1] *Statistički praktikum: linearna regresija*, vježbe, Sveučilište u Zagrebu, Prirodoslovno matematički fakultet, Matematički odsjek, 2018./2019.

[2] C. C. Akoh, G. C. Lee, Y. C. Liaw, T. H. Huang i J. F. Shaw, *GDSL family of serine esterases/lipases*, Abstract, Prog. Lipid Res. 43, 534-52 (2004), https://www.ncbi.nlm.nih.gov/pubmed/15522763.

[3] A. Berko, *Iterativno traženje motiva i nekodirajuća DNA*, Diplomski rad, Sveučilište u Zagrebu, Prirodoslovno matematički fakultet, Matematički odsjek, 2019.

[4] P. Goldstein, *Bioinformatika 1*, https://web.math.pmf.unizg.hr/nastava/bioinformatika/?Bioinformatika_I_2018, 2018./2019.

[5] S. Henikoff i J. Henikoff, *Amino acid substitution matrices from protein blocks*, Proc. Natl. Acad. Sci. USA, Vol. 89, pp. 10915-10919 (1992), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC50453/pdf/pnas01096-0363.pdf.

[6] K. Holenda i A. Dragobratović, *Bjelančevine - Kemija 8*, Digitalni obrazovni sadržaj za 8. razred osnovne škole, https://edutorij.e-skole.hr/share/proxy/alfresco-noauth/edutorij/api/proxy-guest/7b5e1fe5-86e2-4142-af6c-5197c4a08148/kemija-8/m04/j10/index.html.

[7] M. Huzak, *Statistika: statistički testovi 1*, predavanja, Sveučilište u Zagrebu, Prirodoslovno matematički fakultet, Matematički odsjek, 2018./2019.

[8] K. Martinić, *Maksimalne klike u analizi sličnosti proteinskih motiva*, Diplomski rad, Sveučilište u Zagrebu, Prirodoslovno matematički fakultet, Matematički odsjek, 2018.

[9] I. Nakić, *Diskretna matematika*, predavanja, Sveučilište u Zagrebu, Prirodoslovno matematički fakultet, Matematički odsjek, 2011./2012.

[10] B. Rabar, S. Ristov, M. Zagorščak, M. Rosenzweig i P. Goldstein, *Igloss: Iterative gapless local similarity search*, arXiv:1807.11862v1 [q-bio.QM] (2018), `https://arxiv.org/pdf/1807.11862.pdf`.

[11] A. Relja, *Neki statistički aspekti prepoznavanja motiva*, Diplomski rad, Sveučilište u Zagrebu, Prirodoslovno matematički fakultet, Matematički odsjek, 2014.

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

[13] D. Veljan, *Kombinatorna i diskretna matematika*, Algoritam, Zagreb, 2001.

[14] T. Šmuc, *Strojno učenje: evaluacija modela*, predavanja, Sveučilište u Zagrebu, Prirodoslovno matematički fakultet, Matematički odsjek, 2018./2019.

Sažetak

U ovom radu koristili smo tehniku analiziranja određenog skupa proteina pomoću najveće maksimalne klike kako bismo pronašli proteine iz iste proteinske familije. Analizirali smo točnost te pretrage, te smo usporedili dobivene rezultate s rezultatima dobivenim iterativnim pretraživanjem proteoma.

Pomoću očekivanih parametara dobivenih linearnom regresijom na proteinima iz iste familije, izračunali smo očekivanu funkciju sličnosti i na taj način pridružili promatranom skupu proteina težinski, a zatim i $0-1$ graf kojem smo odredili najveću maksimalnu kliku. Tehniku smo proveli na proteomima talijinog uročnjaka, rajčice, riže i šećerne repe.

Dobiveni zaključak je da, uz optimalne parametre, metodom najveće maksimalne klike poboljšavamo točnost pretrage. Za ocjenu modela koristili smo F_1-score, koji je u odnosu na iterativno pretraživanje uz izabrane parametre značajno porastao na svim odgovorima dobivenima najvećom maksimalnom klikom. Također, u kliki je ostalo sačuvano više od 97% stvarnih pozitivaca iz uzorka.

Summary

In this thesis, the largest maximal clique analyzing method has been used on a specific collection of proteins to find proteins belonging to the same protein family. The accuracy of that search has been analyzed and the results have been compared with the results obtained by the iterative search of the proteome.

Using expected parameters obtained by the linear regression on the proteins from the same protein family, the expected similarity function has been calculated and associated with the weight graph and $0-1$ graph on the specific collection of proteins. Then the largest maximal clique for a specific collection of proteins in the $0-1$ graph has been found. The method was used on proteomes of *Arabidopsis thaliana*, *Solanum lycopersicum*, *Oryza sativa* and *Beta vulgaris*.

The conclusion is that the largest maximal clique method, with the optimal parameters, improves the accuracy of the search. The F_1-score was used to assess the results, and the largest maximal clique model yielded a significant improvement compared to the plain iterative search method. Furthermore, the clique preserved more than 97% of true positives in the sample.

Životopis

Rođena sam 28. kolovoza 1995. godine u Zagrebu. Školovanje sam započela u osnovnoj školi Malešnica, te nastavila u V. gimnaziji u Zagrebu. Paralelno sam pohađala glazbenu školu Blagoja Berse, smjer violina, u kojoj sam maturirala 2013. godine. Nakon završetka srednjoškolskog obrazovanja 2014. godine, upisala sam preddiplomski studij Matematika na matematičkom odsjeku PMF-a u Zagrebu. Preddiplomski studij završila sam 2017. godine, čime sam stekla titulu prvostupnice matematike. Iste sam godine upisala diplomski studij Matematička statistika, također na matematičkom odsjeku PMF-a u Zagrebu.

Dana 7.4.2017. godine sudjelovala sam na Otvorenim danima Matematičkog odsjeka PMF-a kao jedan od voditelja radionice Matematička igraonica (Matematički pictionary, Mancala, Mastermind). Osvojila sam 6. mjesto na studentskom natjecanju Mozgalo 2019. godine kao dio tima Outliers. Iste sam godine volontirala kao jedan od mentora na Ljetnoj tvornici znanosti (Avengers: Bjelovar - radionica o statistici i kriptografiji za 5. i 6. razred osnovne škole).