Neki statistički aspekti prepoznavanja motiva

Relja, Ajka

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Ajka Relja

NEKI STATISTIČKI ASPEKTI PREPOZNAVANJA MOTIVA

Diplomski rad

Voditelj rada:
doc. dr. sc. Pavle Goldstein

Zagreb, srpanj 2014.

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

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

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Osobito mi je zadovoljstvo zahvaliti se mom uvaženom i cijenjenom mentoru, doc. dr. sc. Pavlu Goldsteinu, koji me strpljivo i nesebično pomagao prilikom izrade ovog diplomskoga rada. Bilo je nemoguće odoljeti njegovom radnom optimizmu, zaraznom entuzijazmu i pozitivnoj energiji koja me nadahnjivala i motivirala.:)

Sadržaj

Uvod

Bioinformatika je najmlađa znanost o dekodiranju najstarijeg jezika: jezika gena. To je znanost koja se bavi analizom bioloških nizova uz pomoć tehnika iz primijenjene matematike, statistike i računarstva. Sve veća dostupnost tehnologije sekvenciranja rezultirala je stvaranjem velikih skupova bioloških podataka. Glavna su istraživanja ovog područja poravnanje nizova proteina, predviđanje strukuture proteina te pronalaženje varijanti određenih enzima u organizmima.

U ovom diplomskom radu, nakon kratkog biološkog uvoda i navoda osnovnih matematičkih pojmova, prelazimo na statistički i računarski dio rada gdje analiziramo distribuciju maksimalnih scorova za određeni enzim predstavljen upitom. Za reprezentaciju motiva u biološkim nizovima koristimo matricu pozicijsko specifičnih težina, još poznatu kao i PSSM matricu, gdje ćemo scorove računati metodom klizećeg prozora nad nizovima proteoma biljke *A. thaliana*. Posebnu pažnju posvećujemo upravo distribuciji tako izračunatih maksimalnih scorova koja bi po teorijskoj osnovi trebala pripadati familiji distribucija ekstremih vrijednosti. Kako to nije slučaj kod nas, navest ćemo konkretnu korekciju na duljinu niza uz koju se scorovi ponašaju po teorijskoj osnovi.

Iz biološke perspektive, tema rada je usko povezana s metodom traženja varijanti nekog enzima u biološkim organizmima, dok je sa statističke strane povezana s intepretacijom distribucije scorova.

Poglavlje 1

Osnovni pojmovi

1.1 Proteini

Proteini ili *bjelančevine*, uz vodu, najvažnije su tvari u ljudskom organizmu. Glavni su izvor tvari za izgradnju mišića, krvi, kože, kose, noktiju i unutarnjih organa, uključujući srce i mozak. Sastavni su dijelovi svake stanice, što ih čini osnovom života na Zemlji. Izgrađene su od aminokiselina koje su međusobno povezane peptidnom vezom.

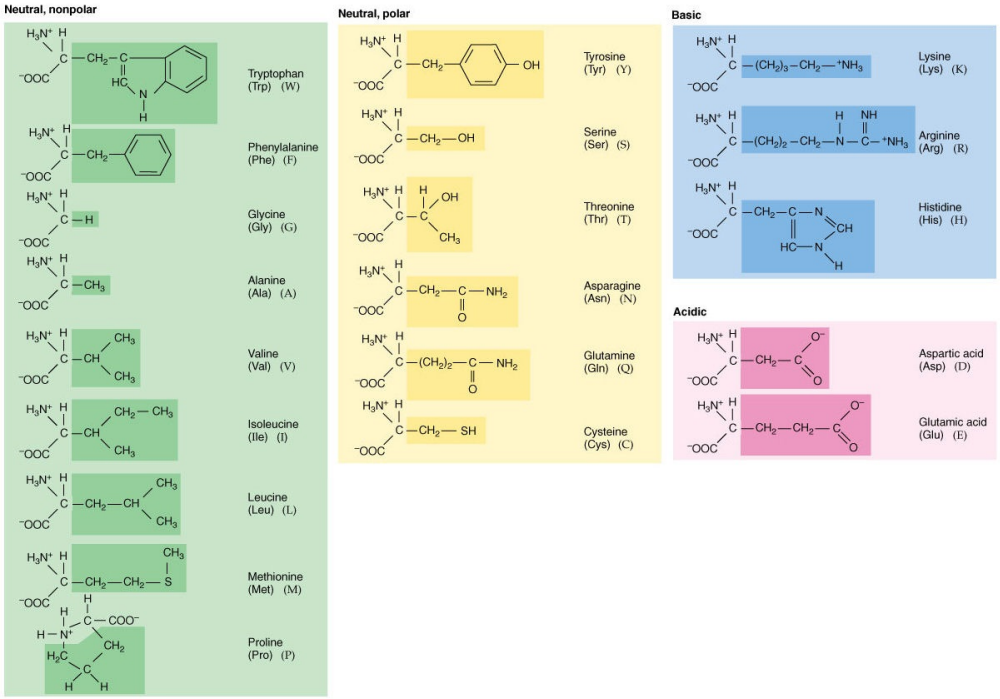
Aminokiselina (eng. *aminoacid*) je osnovna građevna jedinica svakog proteina. Postoji 20 standardnih aminokiselina, a njihov redoslijed u proteinu određuje funkciju proteina. Promjenom redoslijeda aminokiselina u lancu, mijenjaju se njegove karakteristike. Aminokiseline obično označavamo skraćenicama od jednog ili tri slova, ali u ovom radu koristimo se skraćenicama od jednog slova i to u abecednom poretku:

$$\mathcal{A} = \{ \text{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y} \}.$$

Na slici (1.1) prikazan je njihov pun naziv, često korištene skraćenice te podjela aminokiselina s obzirom na *R-skupinu*:

- Aminokiseline nepolarne R-skupine: alanin, valin, leucin, izoleucin, prolin, fenilalanin, triptofan, metionin.
- Aminokiseline polarne R-skupine (-ON, -SN, - SONH2). Ove skupine omogućavaju formiranje vodikovih veza koje su osnova za formiranje viših oblika organizacije aminokiselina u molekulama proteina: glicin, serin, treonin, cistein, tirozin, asparagin, glutamin.
- Aminokiseline s kiselom R-skupinom (s negativno naelektriziranim bočnim lancima): asparaginska kiselina i glutaminska kiselina.

- Aminokiseline s bazičnom R-skupinom (s pozitivno naelektriziranim bočnim lancima): lizin, arginin, histidin.



Slika 1.1: Podjela aminokiselina

1.1.1 Struktura proteina

Primarna struktura proteina (eng. *primary structure*) točan je niz aminokiselina u proteinu kojeg često nazivamo i *aminokiselinska niz.*

Sekundarna struktura proteina (eng. *secondary structure*) prostorna je organizacija aminokiselina koje su blizu u primarnoj strukturi. Standardni su elementi sekundarne strukture α***-zavojnice*** (eng. α*-helix*) β***-ploče*** (eng. β*-sheet*) i ***okreti*** (eng. *turn*).

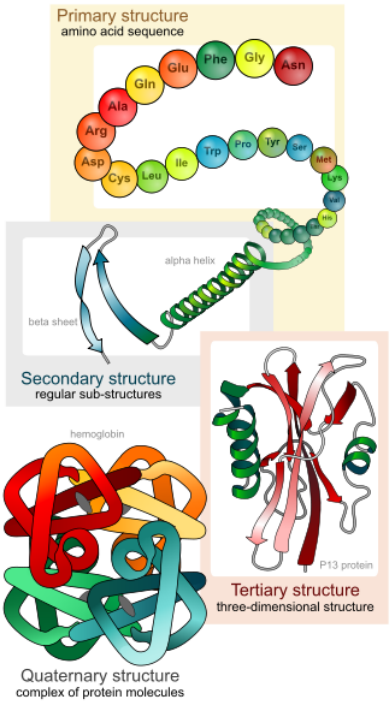
- α**-zavojnica** ima spiralnu strukturu i geometrijski je najregularniji i najčešći element sekundarne strukture.

- **β-ploča** nije jednako učestala struktura kao što je α-zavojnica i nastaje tako da se uspostavljaju vodikove veze između dva susjedna aminokiselinska niza. Može biti *paralelna* ili *antiparalelna* ovisno o smjeru pružanja lanca.
- **Okret** je dio proteina koji nema strukturu ni α-zavojnice ni β-ploče.

Tercijarna struktura proteina (eng. *tertiary structure*) prostorna je organizacija proteina nastala interakcijama među aminokiselinama koje nisu blizu u primarnoj strukturi. Rezultat je savijen lanac aminokiselina. Informacija o tercijarnoj strukturi proteina je ključna za razumijevanje evolucije i funkcije proteina. Predviđanje tercijarne strukture proteina danas je poznatije kao *protein folding problem* i jedan je od najvećih neriješenih problema molekularne biologije.

Kvaterna struktura proteina (eng. *quaternary structure*) nastaje spajanjem više aminokiselinskih nizova u jednu strukturalnu jedinicu.

Na Slici (1.2) grafički je prikazana prethodno iznesena shema strukture proteina.



Slika 1.2: Struktura proteina

1.1.2 Evolucija proteina

Zbog velike važnosti proteina i njihove raznolikosti, od velikog je interesa proučavanje evolucije proteina radi boljeg razumijevanja njihove uloge, funkcije i strukture u organizmu. Kako proteini ne nastaju *de novo*, već su posljedica evolucije postojećih nizova, uvodimo pojam *familije proteina*. To je skup proteina koji potječu od istog pretka. U ovom radu, pod evolucijom proteina, smatra se mutacijski događaj na slučajnom mjestu u proteinskom nizu. **Mutacijski događaj** može biti jedno od sljedećeg:

- **insercija** - ubacivanje jedne ili više aminokiselina
- **delecija** - izostavljanje jedne ili više aminokiselina
- **supstitucija**- zamjena jedne aminokiseline drugom

Primjer 1.1.1. *Pokažimo na nizu aminokiselina **KAMEN** primjer mutacije:*

- ***KAMIN** - aminokiselina E zamijenjena je aminokiselinom I (*supstitucija*)*
- ***KAMENA** - aminokiselina A dodana je na kraj niza (*insercija*)*
- ***AMEN** - iz niza je izbačena aminokiselina K (*delecija*)*

*Kako nam je poznata primarna struktura "pretka" iz kojeg su gornji nizovi nastali (**KAMEN**), moguće ih je nedvosmisleno poravnati:*

```
KAMEN-
KAMENA
-AMEN-
```

Gdje sa ' - ' označavamo mjesto delecije aminokiseline u jednom ili mjesto insercije na tom mjestu u drugom nizu. Simbol ' - ' nazivamo **praznina** *(eng.* gap*).*

1.1.3 Motiv

Neki biolozi smatraju da je ***motiv*** najmanja strukturalna jedinica koja opstaje ili nestaje evolucijom proteina. **Motiv** je niz od 10-ak do 20-ak aminokiselina u proteinu na kojem se vidi jasan supstitucijski uzorak tj. shema. **Motiv suprasekundarne strukture** je podniz aminokiselina iz proteina koji odgovaraju nekoj od suprasekundarnih struktura. **Suprasekundarnu strukturu proteina** (eng. *suprasecondary protein structure*) čini više povezanih elemenata sekundarne strukture.

Po mišljenju dijela biologa, suprasekundarna struktura ima važnu ulogu pri proučavanju evolucije proteina pa se javlja interes za popisivanjem svih mogućih suprasekundarnih struktura opisivanjem ukupnog prostora odgovarajućih motiva. Oblik u kojemu se motivi zapisuju nazivamo *profil motiva*.

Profil motiva razdioba je vjerojatnosti pojavljivanja 20 aminokiselina na svakoj poziciji u motivu. Dakle, ako se motiv sastoji od deset aminokiselina, profil tog motiva bit će zadan s deset vjerojatnosnih distribucija. Svaka pozicija je vjerojatnosni vektor duljine 20 što znači da je profil tog motiva matrica 10×20. Vjerojatnosti pojavljivanja aminokiselina na pojedinoj poziciji u motivu gotovo nikada nisu jednake. Najčešće očekujemo da distribucija vjerojatnosti aminokiselina za svaku poziciju bude unimodalna ili bimodalna. Smatra se da je popis svih takvih motiva suprasekundarne strukture, tzv. *baza motiva suprasekundarne strukture*, konačan.

1.2 Teorija vjerojatnosti

1.2.1 Zadavanje vjerojatnosnog prostora

Definicija 1.2.1. ***Slučajni pokus*** *ili* ***slučajni eksperiment*** *takav je pokus čiji ishodi, tj. rezultati nisu jednoznačno određeni uvjetima u kojima izvodimo pokus.*

Kod svakog slučajnog pokusa osnovno je da se ustanovi odnos između uzroka i posljedice. Poznavanje tog odnosa omogućuje definiranje **uvjeta pokusa** i predviđanje **ishoda** pri svakom realiziranju pokusa.

Najčešći primjer slučajnog pokusa bacanje igraće (simetrične) kocke.

Definicija 1.2.2. ***Prostor elementarnih događaja*** *neprazan je skup Ω koji reprezentira skup svih ishoda slučajnog pokusa. Elemente od Ω označavamo s ω i zovemo* ***elementarni događaji****.*

Definicija 1.2.3. *Familija $\mathcal{A}$ podskupova od Ω jest* ***algebra skupova*** *(na Ω) ako je:*

1. $\emptyset \in \mathcal{A}$
2. $A \in \mathcal{A} \Longrightarrow A^c \in \mathcal{A}$
3. $A_1, A_2, \ldots, A_n \in \mathcal{A} \Longrightarrow \bigcup_{i=1}^{n} A_i \in \mathcal{A}$

Definicija 1.2.4. *Familija $\mathcal{F}$ podskupova od Ω ($\mathcal{F} \subset \mathcal{P}(\Omega)$) je* ***$\sigma$-algebra skupova*** *(na Ω) ako je:*

1. $\emptyset \in \mathcal{F}$

2. $A \in \mathcal{F} \Longrightarrow A^c \in \mathcal{F}$

3. $A_i \in \mathcal{F} \Longrightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$

Definicija 1.2.5. *Neka je $\mathcal{F}$ σ-algebra na skupu Ω. Uređen par $(\Omega, \mathcal{F})$ zove se **izmjeriv prostor**.*

Definicija 1.2.6. *Neka je $(\Omega, \mathcal{F})$ izmjeriv prostor. Funkcija $P : \mathcal{F} \to \mathbb{R}$ je **vjerojatnost** (na $\mathcal{F}$, na Ω) ako vrijedi:*

1. $P(A) \geq 0,\ A \in \mathcal{F}$ *(nenegativnost)*

2. $P(\Omega) = 1$ *(normiranost)*

3. *Za svaki niz $(A_n,\ n \in \mathbb{N})$, $A_n \in \mathcal{F}$, takav da je $A_n \bigcap A_m = \emptyset$ za $m \neq n$, vrijedi*

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mathbb{P}(A_n)\ (\sigma\text{ -aditivnost ili prebrojiva aditivnost})$$

Definicija 1.2.7. *Uređena trojka $(\Omega, \mathcal{F}, P)$, gdje je $\mathcal{F}$ σ-algebra na Ω i P vjerojatnost na $\mathcal{F}$, zove se **vjerojatnostni prostor**.*

Ako je Ω konačan ili prebrojiv skup $(\Omega, \mathcal{F}, \mathbb{P})$ zovemo **diskretni vjerojatnosni prostor**.

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

Definicija 1.2.8. *Neka je $(\Omega, \mathcal{F}, \mathbb{P})$ proizvoljan vjerojatnosni prostor i $A \in \mathcal{F}$ takav da je $\mathbb{P}(A) > 0$. Definirajmo funkciju $\mathbb{P}_A \colon \mathcal{F} \to [0, 1]$:*

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

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

Definicija 1.2.9. *Konačna ili prebrojiva familija $(H_i, i = 1, 2, \ldots)$ događaja u vjerojatnosnom prostoru $(\Omega, \mathcal{F}, \mathbb{P})$ jest **potpun sistem događaja** ako je $H_i \neq \emptyset$ za $\forall i$, $H_i \bigcap H_j = \emptyset$ za $i \neq j$ (tj. događaji se uzajamno isključuju) i $\bigcup_i H_i = \Omega$. Drugim riječima, potpun sistem događaja konačna je ili prebrojiva particija skupa Ω s tim da su elementi particije događaji.*

Teorem 1.2.10. *(Formula potpune vjerojatnosti) Neka je* $(H_i, i = 1, 2, \ldots)$ *potpun sistem događaja u vjerojatnosnom prostoru* $(\Omega, \mathcal{F}, \mathbb{P})$*. Tada za proizvoljno* $A \in \mathcal{F}$ *vrijedi*

$$\mathbb{P}(A) = \sum_i \mathbb{P}(H_i)\mathbb{P}(A \mid H_i). \tag{1.2}$$

Teorem 1.2.11. *(Bayesova formula) Neka je* $(H_i, i = 1, 2, \ldots)$ *potpun sistem događaja u vjerojatnosnom prostoru* $(\Omega, \mathcal{F}, \mathbb{P})$ *i* $A \in \mathcal{F}$ *takav da je* $\mathbb{P}(A) > 0$*. Tada za svako i vrijedi:*

$$\mathbb{P}(H_i \mid A) = \frac{\mathbb{P}(H_i)\mathbb{P}(A \mid H_i)}{\sum_j \mathbb{P}(H_j)\mathbb{P}(A \mid H_j)}. \tag{1.3}$$

Neka je $\mathbb{R}$ skup realnih brojeva. Sa $\mathcal{B}$ označimo σ-algebru generiranu familijom svih otvorenih skupova u $\mathbb{R}$. $\mathcal{B}$ zovemo **σ-algebra Borelovih skupova na** $\mathbb{R}$, a elemente σ-algebre $\mathcal{B}$ zovemo **Borelovi skupovi**.

Definicija 1.2.12. *Neka je* $(\Omega, \mathcal{F}, P)$ *vjerojatnosni prostor. Funkcija* $X : \Omega \to \mathbb{R}$ *jest* ***slučajna varijabla*** *(na* Ω*) ako je* $X^{-1}(B) \in \mathcal{F}$ *za proizvoljno* $B \in \mathcal{B}$*, tj.* $X^{-1}(B) \subset \mathcal{F}$*.*

Za $B \in \mathcal{B}$ stavimo:

$$P_X(B) = P\left(X^{-1}\left(B\right)\right) = P\left\{\, \omega \in \Omega; X\left(\omega\right) \in B \,\right\} = P\left\{X \in B\,\right\}. \tag{1.4}$$

Relacijom (1.4) definirana je funkcija $P_X : \mathcal{B} \to [0, 1]$, za koju se lako pokaže da je vjerojatnosna mjera na $\mathcal{B}$. P_X zovemo **vjerojatnosna mjera inducirana sa X**, a vjerojatnosni prostor $(\mathbb{R}, \mathcal{B}, P_X)$ zovemo **vjerojatnosni prostor induciran sa X**.

1.2.2 Funkcija distribucije, funkcija gustoće

Definicija 1.2.13. *Neka je X slučajna varijabla na* Ω*.* ***Funkcija distribucije od X*** *jest funkcija* $F_X : \mathbb{R} \to [0, 1]$ *definirana sa:*

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

Definicija 1.2.14. *Neka je X slučajna varijabla na vjerojatnosnom prostoru (* $\Omega, \mathcal{F}, P$*) 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:*

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

Integral u (1.5) Lebesgueov je integral funkcije f u odnosu na Lebesgueovu mjeru λ. Za funkciju distribucije F_X neprekidne slučajne varijable X, dakle za funkciju oblika (1.5) kažemo da je **apsolutno neprekidna funkcija distribucije**.

Ako je X neprekidna slučajna varijabla, tada se funkcija f iz (1.5) zove **funkcija gustoće vjerojatnosti od X**, tj. od njezine funkcije distribucije F_X ili, kraće **gustoća od X**, ponekad je i označujemo sa f_x.

1.2.3 Matematičko očekivanje

Definicija 1.2.15. *Neka je X slučajna varijabla na ($\Omega, \mathcal{F}, P$). X je **jednostavna slučajna varijabla** ako je njezino područje vrijednosti konačan skup.*

Definicija matematičkog očekivanja provodi se u tri koraka. Prvo se definira matematičko očekivanje jednostavne slučajne varijable, zatim nenegativne slučajne varijable i na kraju opće slučajne varijable. Sa $\mathcal{K}$ označimo skup svih jednostavnih slučajnih varijabli definiranih na Ω , a sa $\mathcal{K}_+$ skup svih nenegativnih funkcija iz $\mathcal{K}$.

Neka je

$$X \in \mathcal{K}, \;\; X = \sum_{k=1}^{n} x_k K_{A_k}, \quad A_1, \ldots, A_n \in \mathcal{F} \text{ međusobno disjunktni.}$$

Definicija 1.2.16. ***Matematičko očekivanje od X** ili, kraće, **očekivanje od X** koje označavamo sa EX definira se sa*

$$EX = \sum_{k=1}^{n} x_k \, P(A_k)$$

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 $(EX_n, \; n \in \mathbb{N})$ rastući je niz u $\mathbb{R}_+$, pa postoji $\lim_{n\to\infty}$ koji može biti jednak i $+\infty$.

Definicija 1.2.17. ***Matematičko očekivanje od X** ili, kraće, **očekivanje od X** definira se sa*

$$EX = \lim_{n\to\infty} EX_n.$$

Neka je sada X proizvoljna slučajna varijabla na Ω. Vrijedi $X = X^+ - X^-$, za X^+ i X^- nenegativne slučajne varijable. $(X^+, X^- \geq 0)$.

Definicija 1.2.18. *Kažemo da **matematičko očekivanje od X** ili, kraće, **očekivanje od X**, koje označavamo sa EX, **postoji** ili da je **definirano** ako je barem jedna od veličina EX^+ ili EX^- konačna, tj. vrijedi*

$$min \, \{EX^+, EX^-\} < \infty.$$

Tada po definiciji stavljamo $EX = EX^+ - EX^-$.

1.2.4 Varijanca

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

Definicija 1.2.19. *$E[X^r]$ zovemo **r-ti moment od X**, a $E[\,|X^r|\,]$ zovemo **r-ti apsolutni moment od X**.*

Po dogovoru stavljamo $E(X^0) = E(|X^0|) = 1$.

Definicija 1.2.20. *Neka $E(X)$ postoji (tj. konačno je). Tada $E[(X - EX)^r]$ zovemo **r-ti centralni moment od X**, a $E[|X - EX|^r]$ zovemo **r-ti apsolutni centralni moment od X**.*

Definicija 1.2.21. ***Varijanca od X** koju označavamo sa $VarX$ ili σ_X^2 jest drugi centralni moment od X, dakle*

$$VarX = E\left[(X - EX)^2\right].$$

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

1.3 Regresijska analiza

Regresijska analiza je metoda ispitivanja i analize ovisnosti jedne varijable (zavisne) o jednoj ili više drugih (nezavisnih) varijabli, dok regresijski model je jednadžba koja kvantificira povezanost između zavisne varijable s nezavisnim varijablama. U promatranju njihove povezanosti, zanima nas polinomijalna veza, tj. polinom k-tog stupnja koji će najbolje opisati dane podatke.

1.3.1 Konstrukcija modela

Neka je Y zavisna varijabla (varijabla koju želimo opisati ili procijeniti), a x nezavisna varijabla (varijabla pomoću koje želimo opisati zavisnu varijablu). Promatramo model

$$Y = \theta_0 + \theta_1 p_1(x) + \theta_2 p_2(x) + \cdots + \theta_k p_k(x) + \varepsilon,$$

pri čemu su $\theta_0, \theta_1, \ldots \theta_k$ parametri modela, ε slučajna pogreška, a $1, p_1, p_2, \ldots, p_k$ linearno nezavisne realne funkcije. Domena interesa su nam polinomi *k-tog* stupnja, pa stavljamo

$$p_1(x) = x, \quad p_2(x) = x^2, \ \ldots, \ p_k(x) = x^k.$$

Neka su x_i, $i = 1, \ldots, n$ zadane vrijednosti od x, a $y_1, \ldots, y_n$ realizacije slučajne varijable Y. Metodom najmanjih kvadrata želimo pronaći parametre modela $\theta_0, \theta_1, \ldots \theta_k$ tj. tražimo minimum

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

Neka su

$$Y_i = \theta_o + \theta_1 x_i^1 + \cdots + \theta_k x_i^k + \varepsilon_i, \;\; i = 1, \ldots, n$$

slučajne varijable. Pretpostavljamo da vrijede sljedeći Gauss-Markovljevi uvjeti za slučajne greške:

1. $E[\varepsilon_i] = 0, \qquad \forall i = 1, \ldots, n$
2. $E[\varepsilon_i \varepsilon_j] = 0, \quad \forall i, j = 1, \ldots, n, \; i \neq j$
3. $Var[\varepsilon_i] = \sigma^2 > 0, \;\; \forall i = 1, \ldots, n.$

Matrica dizajna je sljedeća:

$$X = \begin{pmatrix} 1 & x_1^1 & x_1^2 & \ldots & x_1^k \\ 1 & x_2^1 & x_2^2 & \ldots & x_2^k \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ 1 & x_n^1 & x_n^2 & \ldots & x_n^n \end{pmatrix} \tag{1.6}$$

dok su

$$\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_3 \\ \vdots \\ \theta_k \end{bmatrix}, \quad y = \begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad Y = \begin{bmatrix} Y_0 \\ Y_1 \\ Y_2 \\ \vdots \\ Y_k \end{bmatrix}, \quad \varepsilon = \begin{bmatrix} \varepsilon_0 \\ \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_k \end{bmatrix} \tag{1.7}$$

Sada nam je model

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

dok tražimo minimum funkcije

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

Nepristran procjenitelj za θ metodom najmanjih kvadrata je $\hat{\theta} = (X^T X)^{-1} X^T Y$, dok mu je procjena $\hat{\theta} = (X^T X)^{-1} X^T y$.

Procjenitelji za Y_i su

$$\hat{Y}_i = \hat{\theta}_o + \hat{\theta}_1 x_i^1 + \cdots + \hat{\theta}_k x_i^k, \; i = 1, \ldots, n,$$

dok su **reziduali** slučajne varijable (tj. njihove realizacije $e_i, i = 1, \ldots, n$)

$$E_i = Y_i - \hat{Y}_i$$

.

- **neobjašnjena varijabilnost** $SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$
- **objašnjena varijabilnost** $SSR = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$
- **ukupna varijabilnost** $S_{YY} = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$

To nam je potrebno kako bi izračunali **koeficijent determinacije** (u oznaci R^2), odnosno proporciju ukupne varijabilnosti objašnjene našim regresijskim modelom. Ona se računa po sljedećoj formuli:

$$R^2 = \frac{\text{objašnjena varijabilnost}}{\text{ukupna varijabilnost}} = 1 - \frac{SSE}{S_{YY}} \in [0, 1]. \tag{1.8}$$

1.3.2 Statističko zaključivanje pri regresijskoj analizi

Neka praktična pitanja na koje treba odgovoriti pri ovakvom modeliranju su:

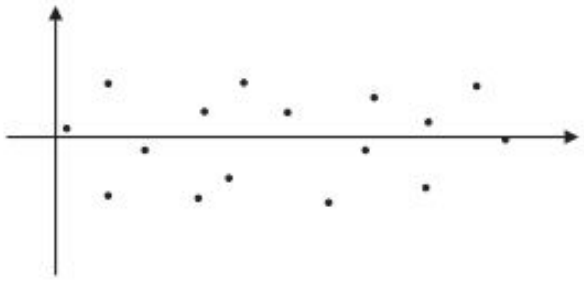
1. Koliki dio promjena u mjerenim vrijednostima zavisne varijable je objašnjen dobivenim modelom? (TEST JAKOSTI MODELA)

   Odgovor na ovo pitanje daje koeficijent determinacije R^2 koji nam zapravo daje informaciju o tome koliko rasipanja izlaznih podataka potječe od funkcijske ovisnosti, a koliko otpada na tzv. rezidualno ili neobjašnjeno rasipanje (tu informaciju očitavamo iz $1 - R^2$). Drugim riječima daje informaciju o tome koliko jaka funkcijska veza je između x i Y. Što je vrijednost koeficijenta R^2 bliža 1, zavisnost je jača i prilagodna modelu je bolja.

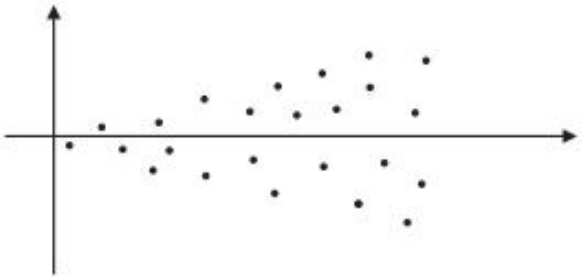
2. Analiza reziduala: utvrđujemo jesu li ispunjene sve početne pretpostavke koje reziduali trebaju ispunjavati, a to su:

   - Varijance grešaka (koje su, kako znamo, slučajnog karaktera) jednake su. Homogenost varijanci reziduala provjeravamo analizom grafičkog prikaza ovisnosti reziduala E_i o procijenjenim vrijednostima $\hat{Y}_i$.

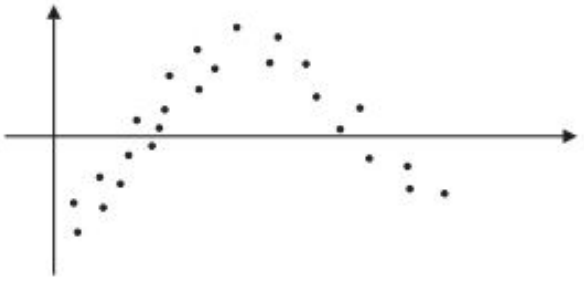
     Jednostavno donošenje zaključaka o varijanci dano je pomoću sličica (1.3), (1.4), (1.5).



Slika 1.3: Horizontalno raspoređene točke sugeriraju homogenost varijanci



Slika 1.4: Ovakav raspored točaka sugerira stalan rast varijance, dakle varijance nisu homogene



Slika 1.5: Ovakav raspored točaka sugerira neadekvatnost linearnog modela

- Reziduali su normalno distribuirani. Normalnost reziduala provjeravamo analizom histograma reziduala i p-plota reziduala.

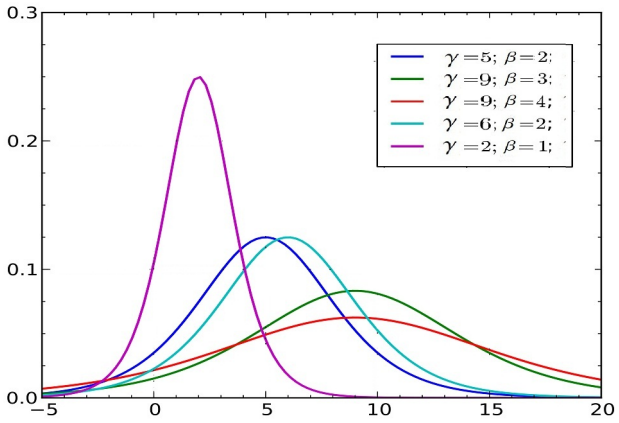
- Reziduali moraju biti međusobno nezavisni, tj. vrijednost reziduala koji se odnosi na realizaciju y_i slučajne varijable Y nema nikakvog utjecaja na vrijednost reziduala koji se odnosi na realizaciju y_j iste slučajne varijable.

Ako reziduali E_i zadovoljavaju prethodno navedene pretpostavke, smatramo ih dobrim procjenama stvarnih normalnih grješaka ε_i.

1.4 Razdioba od interesa

Ovdje ćemo navest samo logističku razdiobu, jer ostale, koje će se pojavljivati u ovom radu, detaljnije ćemo obraditi u idućem poglavlju.

1.4.1 Logistička distribucija



Slika 1.6: Logistička distribucija za različite parametre

Logistička distribucija neprekidna je vjerojatnosna distribucija sa sljedećom funkcijom gustoće:

$$f(x) = \frac{e^{-z}}{\beta(1+e^{z})^2},$$

gdje je $z = \frac{x-\gamma}{\beta}$, $\beta > 0$ parametar mjere i $\gamma \in \mathbb{R}$ parametar lokacije.

Očekivanje i varijanca neprekidne slučajne varijable s logističkom distribucijom su

$$EX = \gamma, \qquad VarX = \frac{\beta^2\pi^2}{3}. \tag{1.9}$$

Poglavlje 2

Teorija ekstremnih vrijednosti

Teorija ekstremnih vrijednosti posebna je grana statistike koja se bavi ekstremnim vrijednostima. Zasnovana na teoremu koji kaže kako postoje samo tri tipa distribucije potrebna za modeliranje maksimuma, odnosno minimuma skupa slučajnih uzoraka iz iste distribucije. Drugim riječima, ako generiramo N uzoraka iz iste distribucije, te za novi skup podataka uzmemo maksimume generiranih uzoraka, taj skup podataka može biti opisan samo jednim od sljedeća tri modela- konkretno, Gumbelovom, Fréchetovom, i Weibullovom distribucijom. Ovi modeli imaju široku primjenu u upravljanu rizicima, financijama, osiguranjima, ekonomji, hidrologiji, telekomunikacijama i mnogim drugim industrijama koje se bave ekstremnim događajima.

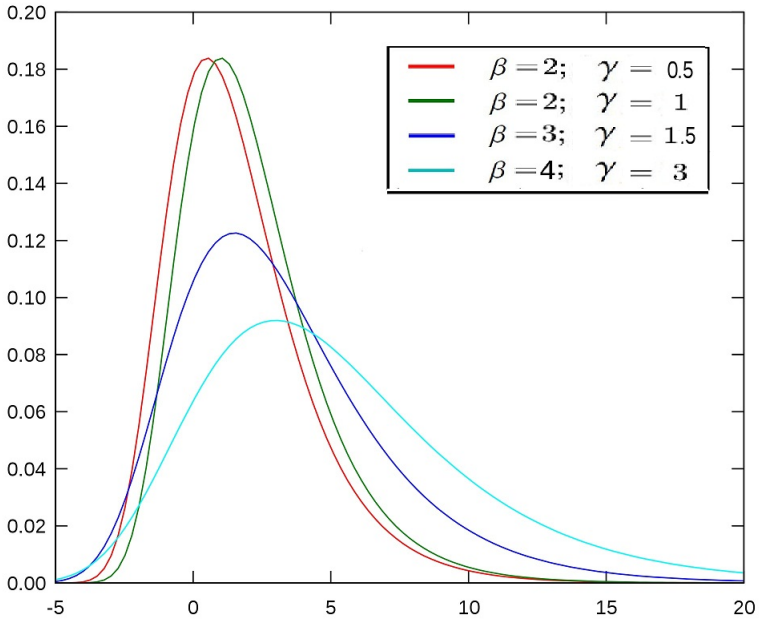
Pionir područja teorije ekstremnih vrijednosti bio je Leonard Tippett (1902. - 1985.). Udruga britanske industrije pamuka zaposlila ga je kako bi istraživao izdržljivost konca. U svojim studijama zaključio je kako snaga niti konca ovisi o izdržljivosti njenih najslabijih vlakana. Uz pomoć R. A. Fishera, Tippett dobio je tri asimptotske granice za opisivanje distribucija ekstrema.

2.1 Distribucije ektremnih vrijednosti

2.1.1 Gumbelova distribucija

Gumbelova distribucija dobila je ime po njemačkom matematičaru Emil Julius Gumbel-u (1891–1966), na temelju njegovih originalnih radova i knjige u kojoj opisuje distribuciju ekstremnih vrijednosti. Njegov glavni fokus bio je prvenstveno na primjeni teorije ekstremnih vrijednosti u inženjerskim problemima, posebno u modeliranju meteroloških fenomena poput godišnjih tokova poplava:

> "It seems that the rivers know the theory. It only remains to convince the engineers of the validity of this analysis."



Slika 2.1: Gumbelova distribucija za različite parametre

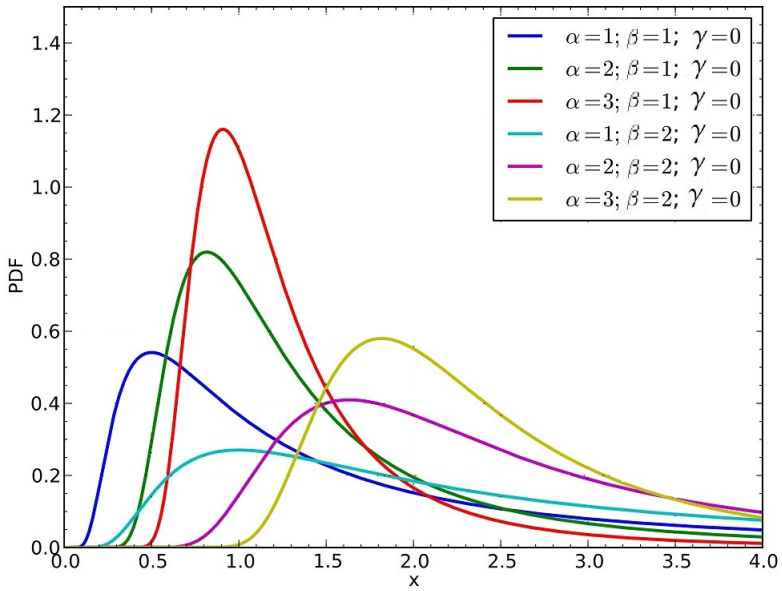
Gumbelova distribucija se koristi prilikom modeliranja maksimuma (odnosno minimuma) uzoraka raznih distribucija. Njezina potencijalna primjenjivost u reprezentaciji maksimuma povezana je s teorijom ekstremnih vrijednosti, posebno kada je distribucija temeljnih uzoraka normalnog ili eksponencijalnog tipa. Gumbelova distribucija poznata je kao i log-Weibullova odnosno dupla eksponencijalna distribucija, što čini posebni slučaj generalizirane distribucije ekstremnih vrijednosti (poznate kao Fisher-Tippett distribucije), odnosno *tip I distribucije ekstremnih vrijednosti*. Njena funkcija gustoće ima sljedeću formu:

$$f(x) = \frac{1}{\beta} e^{(-z-e^{-z})}, \tag{2.1}$$

gdje je $z = \frac{x-\gamma}{\beta}$, uz lokacijski parametar γ i parametar mjere $\beta > 0$. U slučaju kada je $\gamma = 0$ i $\beta = 1$, distribuciju zovemo standardnom Gumbelovom.

2.1.2 Fréchetova distribucija

Maurice Fréchet (1878. - 1973.) bio je francuski matematičar koji je indetificirao jednu moguću graničnu distribuciju za ekstreme uređenih statistika. Fréchetova distribucija također poznata je kao i *tip II distribucije ekstremnih*



Slika 2.2: Fréchetova distribucija za različite parametre

vrijednosti, definirana sa funkcijom gustoće na sljedeći način:

$$f(x) = \frac{\alpha}{\beta}\left(\frac{\beta}{x-\gamma}\right)^{\alpha+1} exp\left(-\left(\frac{\beta}{x-\gamma}\right)^{\alpha}\right) \tag{2.2}$$

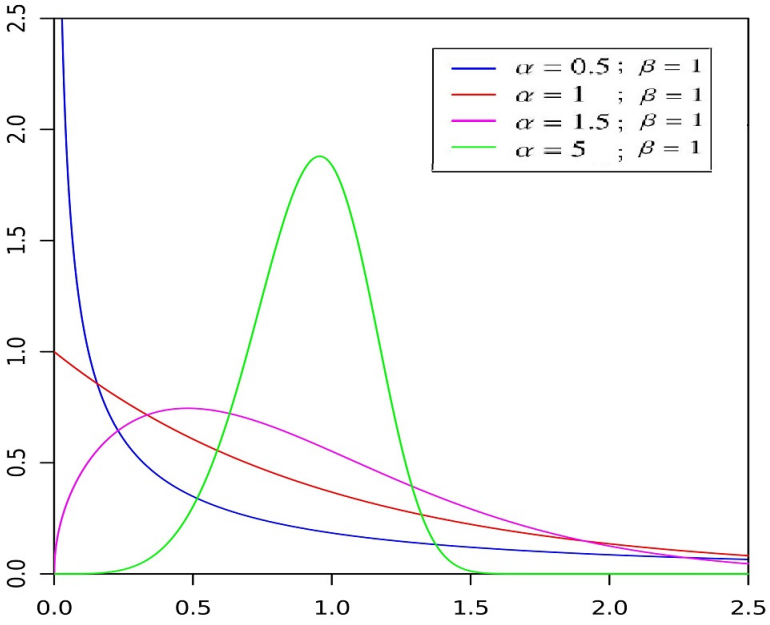
gdje je α parametar oblika ($\alpha > 0$), β parametar mjere ($\beta > 0$), a γ parametar lokacije. Za $\gamma \equiv 0$, dobivamo dvo-parametarsku Fréchetovu distribuciju. Ova distribucija omeđena je ($\gamma < x < \infty$) i ima teški gornji rep.

2.1.3 Weibullova distribucija

Waloddi Weibull (1887. - 1979.), bio je švedski inženjer, znanstvenik i matematičar, poznat po istraživanjima snage materijala i analizi umora. Weibullova distribucija, također poznata i kao *tip III distribucije ekstremnih vrijednosti*, prvi put spominje se u njegovim radovima 1939. godine. Dvo-parametarska verzija ove distribucije ima sljedeću funkciju gustoće :

$$f(x) = \frac{\alpha}{\beta}\left(\frac{x}{\beta}\right)^{\alpha-1} exp\left(-\left(\frac{x}{\beta}\right)^{\alpha}\right). \tag{2.3}$$

Distribucija je definirana za $x > 0$, gdje su α parametar oblika i β parametar mjere pozitivni. Ako dodamo još i parametar lokacije γ, dobivamo generaliza-



Slika 2.3: Weibullova distribucija za različite parametre

ciju sa sljedećom funkcijom gustoće:

$$f(x) = \frac{\alpha}{\beta}\left(\frac{x-\gamma}{\beta}\right)^{\alpha-1} exp\left(-\left(\frac{x-\gamma}{\beta}\right)^{\alpha}\right). \tag{2.4}$$

U ovom modelu, lokacijski parametar može poprimiti bilo koje realne vrijednosti, dok je distribucija definirana za $x > \gamma$.

Za $\alpha = 1$, distribucija se svodi na eksponencijalni model, a za $\alpha = 2$ oponaša Rayleigovu distribuciju koja se uglavnom koristi u telekomunikacijama. Osim toga, sliči na normalnu distribuciju za $\alpha = 3.5$.

Odnosno, sumirajmo sve u definiciju:

Definicija 2.1.1. *Sljedeće su standardne funkcije distribucije ekstremnih vrijednosti:*
(i) Gumbel (tip I): $\Lambda(x) = exp\{-exp\,(-x)\}, x \in \mathbb{R}$;
(ii) Fréchet (tip II): $\Phi_\alpha(x) = \begin{cases} 0, & x \leq 0; \\ exp\{-x^{-\alpha}\}, & x > 0, \alpha > 0; \end{cases}$

(iii) Weibull (tip III): $\Psi_\alpha(x) = \begin{cases} exp\{-(-x^\alpha)\}, & x \leq 0, \alpha > 0; \\ 1, & x > 0. \end{cases}$

EVD (Extreme value distribution) familije možemo generalizirati parametrima lokacije ($\gamma \in \mathbb{R}$) i mjere ($\beta > 0$):

$$\Lambda(x; \gamma, \beta) = \Lambda((x - \gamma)/\beta) \,,$$

$$\Phi_\alpha(x; \gamma, \beta) = \Phi_\alpha((x - \gamma)/\beta) \,,$$

$$\Psi_\alpha(x; \gamma, \beta) = \Psi_\alpha((x - \gamma)/\beta) \,.$$

Između ove tri familije funkcija distribucija, *tip I* tj. Gumbelova distribucija, $\{\Lambda(x; \gamma, \beta) = \Lambda((x - \gamma)\beta); \gamma \in \mathbb{R}, \beta > 0\}$ se najčešće spominje u diskusijama o ekstremnim vrijednostima gdje čini distribuciju ekstremnih vrijednosti.

Uočimo da je zbog relacije

$$min(X_1, X_2, \ldots, X_n) = -max(-X_1, -X_2, \ldots, -X_n),$$

dovoljno je promatrati distribucije maksimalnih vrijednosti sadržanih u prethodnoj definiciji.

2.2 Generalizirana distribucija ekstremnih vrijednosti

Generalizirana distribucija ekstremnih vrijednosti, još poznata kao i Fisher–Tippettova distribucija, fleksibilan je model sa tri parametra koji kombinira Gumbelovu, Frechétovu i Weibullovu distribuciju maksimalne ekstremne vrijednosti. Ona ima sljedeću vjerojatnosnu funkciju gustoće:

$$f(x) = \begin{cases} \frac{1}{\beta} exp\left(-1(1 + \alpha z)^{-1/\alpha}\right)(1 + \alpha z)^{-1-1/\alpha}, & \alpha \neq 0 \\ \frac{1}{\beta} exp(-z - exp(-z)), & \alpha = 0 \end{cases} \tag{2.5}$$

gdje je $z = \frac{x-\gamma}{\beta}$, te $\alpha \in \mathbb{R}$, $\beta > 0$ i $\gamma \in \mathbb{R}$ redom parametri oblika, mjere i lokacije.

Domena ovisi o parametru α:

$$1 + \alpha \frac{(x - \gamma)}{\beta}, \alpha \neq 0$$

$$-\infty < x < \infty, \alpha = 0$$

Za različite vrijednosti parametara, dobivaju se distribucije ekstremnih vrijednosti *tipa I* , *tipa II* ili *tipa III*. Točnije, tri slučaja kada je $\alpha = 0$, $\alpha > 0$ i $\alpha < 0$, odgovaraju redom Gumbelovoj, Fréchetovoj i 'obrnutoj' Weibullovoj distribuciji tako da ona ima gornju granicu koju u primjenama moramo procijeniti, za razliku od obične Weibullove gdje je donja granica nula. Također primijetimo kako Fréchetova ima donju granicu, dok je Gumbelova neograničena.

2.3 Fisher- Tippett- Gnedenkov teorem

Fisher- Tippett- Gnedenkov teorem predstavlja fundamentalan rezultat u teoriji ekstremnih vrijednosti vezanih za asimptotsko ponašanje distribucija ekstrema uređenih statistika. Maksimumi uzorka nezavisnih jednako distribuiranih varijabli nakon odgovarajuće normalizacije, mogu jedino konvergirati (po distribuciji) u jednu od tri moguće distribucije; Gumbelovu, Fréchetovu ili Weibullovu razdiobu.

Uloga teorema maksimalnih ekstremnih vrijednosti je slična centralnom teoremu za sredine, osim činjenice da je centralni teorem primjenjiv na uzorke bilo koje distribucije s konačnim varijancama, dok Fisher- Tippett- Gnedenkov teorem kaže da ako distribucija normaliziranih maksimuma konvergira, da će tada limes biti jedna od razdioba određene klase distribucija. Ne navodi da distribucija normaliziranih vrijednosti konvergira.

Teorem 2.3.1. *(Fisher-Tippett 1928; Gnedenko, 1943.)*
Neka su $X_1, X_2, \ldots, X_n$ jednako distribuirane slučajne varijable. Ako postoje normalizacijske konstante $a_n > 0$ i $b_n \in \mathbb{R}$ takve da

$$\lim_{n\to\infty} P\left\{a_n^{-1}\left(max(X_1, X_2, \ldots, X_n) - b_n\right) \leq x\right\} = G(x)$$

za neku ne-degeneričnu distribuciju G, tada je moguće redefinirati normalizacijske konstante tako da

$$G(x) = G_\alpha(x) := exp(-1(1 + \alpha x)^{-1/\alpha}) \,,$$

za sve x takve da $1 + \alpha x > 0$, s indeksom ekstremne vrijednosti $\alpha \in \mathbb{R}$. Za $\alpha \to 0$, $G_\alpha(x)$ poprima oblik $\Lambda(x)$ za sve $x \in \mathbb{R}$. Prema tome, distribucija F pripada domeni atrakcije od G_α koja je označena s $F \in D(G_\alpha)$.

Za $\alpha < 0$, $\alpha = 0$ i $\alpha > 0$, se distribucija G_α reducira na redom Weibullovu, Gumbelovu i Fréchetovu distribuciju. Točnije,

$$\Lambda(x) \equiv G_0\,(x)\,,$$
$$\Phi_\alpha(x) \equiv G_{1/\alpha}\,(\,\alpha\,(x-1)\,)\,,$$
$$\Psi_\alpha(x) \equiv G_{-1/\alpha}\,(\,\alpha\,(1+x)\,)\,.$$

Poglavlje 3

Poravnanje nizova. Scoring sistem. PSSM. Metoda klizećeg prozora

U ovom poglavlju objasnit ćemo osnovne pojmove koji su nam potrebni da bismo uveli PSSM matricu i adekvatan scoring sistem koji će nam bit potreban za analiziranje distribucije maksimalnih scorova.

3.1 Poravnanje nizova

Poravnanje nizova je način uređivanja nizova proteina u svrhu identificiranja mjesta sličnosti koja mogu biti posljedica funkcionalne, evolucijske ili strukturne veze između nizova. Poravnate nizove aminokiselina obično reprezentiramo redcima matrice, gdje nizovima dodajemo praznine da bi uspješno poravnali identične ili slične aminokiseline proteina.

Osnovna primjena analize nizova je u ispitivanju povezanosti dva niza. To se obično radi tako da najprije poravnamo nizove, a zatim odlučimo da li je vjerojatnije da se poravnanje dogodilo jer su nizovi povezani, ili je slučajno.

Primjer 3.1.1. *Poravnanje niza fragmenta human alpha globina s human beta globinom. (SWISS-PROT database)*

```
HBA_HUMAN      GSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKL
               G+ +VK+HGKKV  A+++++AH+D++ +++++LS+LH  KL
HBB_HUMAN      GNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKL
```

U srednjem retku je slovo ako su pozicije jednake, tj. znak '+' ako su one slične.

Ključna pitanja koja se nameću u analizi proteina su :

- koji algoritam i *scoring* sistem treba koristiti?
- kakva poravnanja treba promatrati?
- koje su statističke metode za ocjenu značajnosti scorova poravnanja?

Za poravnavanje nizova ćemo koristiti tzv. protokol klizećeg prozora (*eng. sliding window protocol*) čija je osnova poravnavanje upita tj. motiva sa 'prozorima' tj. dijelovima niza aminokiselina, te ćemo ga kasnije detaljnije objasniti.

3.2 Scorovi za poravnanja nizova upita

Svaka metoda za poravnavanje nizova zahtijeva metodu za računanje *scorova* koja će dati mjeru relativne šanse da su nizovi povezani nasuprot nepovezanosti. Stoga ćemo prije algoritma za traženje optimalnog poravnanja uvesti scoring sistem.

Supstitucijski scorovi mjere cijenu zamjene jedne aminokiseline drugom, dok *penali za praznine* mjere cijenu zamjene aminokiseline s nizom aminokiselina ili ni jednom aminokiselinom. **Scorovi za poravnavanje nizova** su suma supstitucijskih scorova i penala praznina nad svim poravnatim aminokiselinama, pa se za najbolje poravnanje uzima ono koje ima najveći score.

Supstitucijski scorovi za poravnavanje nizova iz proteinskih familija su bazirani na očuvanim modelima aminokiselina i njihovim svojstvima, pa su se takve matrice scorova poboljšale kroz godine.

U našem modelu ćemo promatrati one scorove koji će isključivati penale za praznine. Pretpostavimo da promatramo nizove x i y duljina n i m. Neka x_i predstavlja i-tu aminokiselinu niza x, a y_j j-tu aminokiselinu niza y. Za izračunavanje scorova koristimo dva modela koji pridružuju vjerojatnost svakom poravnanju, zatim računamo vrijednosti njihovih omjera.

- **Random model R** je jednostavan model koji pretpostavlja da se aminokiselina a u nizu pojavljuje nezavisno sa frekvencijom q_a. Stoga, vjerojatnost dva niza je jednostavno produkt vjerojatnosti svake aminokiseline:

$$P(x, y \mid R) = \prod_i q_{x_i} \prod_j q_{y_j}$$

- **Match model M** za svaki par poravnatih aminokiselina pojavljuje sa zajedničkom vjerojatnošću p_{ab}. Vrijednost p_{ab} možemo zamisliti kao vjerojatnost da su aminokiseline a i b izvedene iz neke nepoznate aminokiseline c u njihovom zajedničkom pretku (c može biti isti kao i a i/ili b). Ovo daje vjerojatnost cijelog poravnanja :

$$P(x, y \mid M) = \prod_i p_{x_i, y_i}$$

Omjer ove dvije šanse je poznat kao *odds ratio*:

$$\frac{P(x, y \mid M)}{P(x, y \mid R)} = \frac{\prod_i p_{x_i, y_i}}{\prod_i q_{x_i} \prod_i q_{y_i}} = \prod_i \frac{p_{x_i, y_i}}{q_{x_i} q_{y_i}}.$$

Kako bi došli do aditivnog scoring sistema, uzimamo logaritam tog omjera, poznatog još kao i *log-odds ratio*:

$$S = \sum_i log\left(\frac{p_{x_i, y_i}}{q_{x_i} q_{y_i}}\right) = \sum_i s(x_i, y_i), \tag{3.1}$$

gdje $s(x_i, y_i)$ predstavlja logaritam omjera šansi da se par aminokiselina x_i i y_i pojavi kao poravnati odnosno neporavnati par.

3.3 Matrica pozicijsko specifičnih težina (PSSM)

Matrica pozicijsko specifičnih težina, još poznata i kao *matrica specifičnih težina* ili *PSSM* matrica se uobičajeno koristi za reprezentaciju motiva u nekom biološkom nizu. Često se računa za skup poravnatih nizova za koje se smatra da su funkcionalno povezani, te je kao takva postala iznimno važan dio mnogih softvera za računalno otkrivanje motiva.

PSSM matricu prvi je uveo američki genetičar Gary Stormo s kolegama 1982. godine kao alternativu metode suglasnih nizova (*eng. consensus sequences*), koja se do tada koristila pri reprezentaciji obrazaca u biološkim uzorcima, gdje su postojale poteškoće u predviđanju novih pojava tih obrazaca.

Poljsko-američki matematičar Andrzej Eurenfeucht predložio je algoritam za kreiranje matrice težina koja je mogla razlikovati prava mjesta spajanja od drugih ne funkcionalnih mjesta sa sličnim nizovima. Korištenjem te matrice za pretraživanje novih nizova se pokazalo da je ova metoda osjetljivija i preciznija za razliku od metode suglasnih nizova. Upravo je ta prednost pridonijela

popularnosti PSSM matrica za reprezentaciju obrazaca u biološkim uzorcima, kao i njihovoj upotrebi u modernim algoritmima za otkrivanje motiva.

Specifične pozicijske scoring matrice omogućuju reprezentaciju familija nizova koje mogu detektirati suptilne sličnosti. Opsežne evaluacije mogu efektivno pomoći u odabiru scorova za poravnavanje nizova, kako i za predikciju proteiske strukture.

Konkretno, PSSM matricu za skup od n poravnatih nizova duljina L računamo na sljedeći način:

- Ako poravnate nizove označimo sa

  $$x_1, \ldots, x_n$$

  gdje $k - ti$ niz izgleda ovako

  $$x_k = x_{k1}, \ldots, x_{kL}, \quad k = 1, \ldots, n$$

  tada su elementi x_{kj}, $j = 1, \ldots, L$ aminokiseline.
- Elemente matrice specifičnih težina

  $$M = [M_{ij}] \quad i = 1, \ldots L, \quad j = 1, \ldots, 20$$

  računamo kao

  $$M_{ij} = \frac{1}{n} \sum_{k=1}^{n} \delta_i(x_{kj}) \tag{3.2}$$

  gdje je

  $$\delta_i(q) = \begin{cases} 1 & i = q \\ 0 & i \neq q. \end{cases}$$

Pretpostavljamo statističku nezavisnost između pozicija u nizovima za koje računamo PSSM matricu, tako da vjerojatnost računamo posebno za svako mjesto neovisno o drugim mjestima. To efektivno znači vjerojatnost da se određena aminokiselina pojavi na određenom mjestu u nizu ne ovisi o vjerojatnosti da se ta ista aminokiselina pojavi na nekom drugom mjestu u nizu.

Iz formule (3.2) slijedi da suma retka matrice za određenu poziciju u nizu iznosi jedan, stoga smatramo da svaki redak ima multinomijalnu distribuciju

neovisnu o distibuciji drugih redaka. Odnosno, vidimo kako PSSM predstavlja primjer profila motiva.

Na sljedećem primjeru ćemo ilustrirati kako izgleda osnovna PSSM matrica, gdje ćemo zbog jednostavnosti prikaza koristiti samo nizove koji se sastoje od četri aminokiseline (tj. DNA nizove).

Primjer 3.3.1. ***Osnovna PSSM matrica relativnih frekvencija***
Promatramo sljedećih deset funkcionalno povezanih i poravnatih nizova nukleotida (nizovi aminokiselina A,C,T,G):

```
GAGGTAAAC
TCCGTAAGT
CAGGTTGGA
ACAGTCAGT
TAGGTCATT
TAGGTACTG
ATGGTAACT
CAGGTATAC
TGTGTGAGT
AAGGTAAGT
```

Za njih računamo matricu frekvencija pojavljivanja pojedine aminokiseline po određenim mjestima u nizu, jer je jasno da frekvencija pojavljvanja određene aminokiseline nije jednaka za svako mjesto u nizu. Također ćemo zbog jednostavnijeg zapisa promatrati transponiranu matricu frekvencija, što će rezultirati transponiranom PSSM matricom.

Matrica frekvencija:

	1	2	3	4	5	6	7	8	9
A	3	6	1	0	0	6	7	2	1
C	2	2	1	0	0	2	1	1	2
G	1	1	7	10	0	1	1	5	1
T	4	1	1	0	10	1	1	2	6
	10	10	10	10	10	10	10	10	10

Matrica relativnih frekvencija:

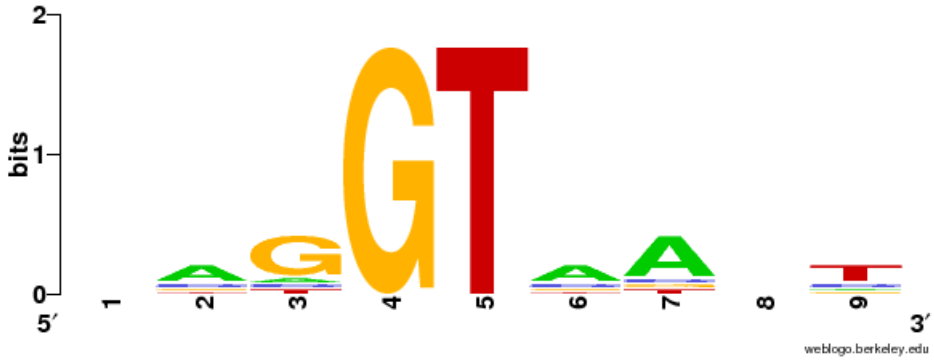
	1	2	3	4	5	6	7	8	9
A	0.3	0.6	0.1	0.0	0.0	0.6	0.7	0.2	0.1
C	0.2	0.2	0.1	0.0	0.0	0.2	0.1	0.1	0.2
G	0.1	0.1	0.7	1.0	0.0	0.1	0.1	0.5	0.1
T	0.4	0.1	0.1	0.0	1.0	0.1	0.1	0.2	0.6
	1	1	1	1	1	1	1	1	1

Kada transponiramo posljednje dobivenu matricu dobivamo osnovnu PSSM matricu gdje $i-ti$ redak označava vjerojatnosti pojavljivanja određenih aminokiselina u $i-tom$ mjestu u motivu tj. upitu čiju varijantu želimo naći u nekom novom organizmu.

Sada jednostavno izračunamo vjerojatnost niza uz pomoć dane PSSM matrice tako da pomnožimo odgovarajuće vjerojatnosti za svaku poziciju. Npr. vjerojatnost niza $X = GAGGTAAAC$ uz danu gornju PSSM marticu je:

$$P(X|M) = 0.1 \times 0.6 \times 0.7 \times 1 \times 1 \times 0.6 \times 0.7 \times 0.2 \times 0.2 = 0.0007056.$$

Tom Schneider predlaže grafički prikaz u kojem je sadržana važnost pojedinih aminokiselina na određenim mjestima u poravnatim nizovima. Pogledajmo kako to izgleda za naš primjer:



Slika 3.1: Schneiderov prikaz niza

3.4 PSSM i score

PSSM matrica za upit duljine L je dimenzije $L \times 20$ i sastoji od stupaca vjerojatnosti za svaku aminokiselinu. Prirodni vjerojatnosni model je specificiranje nezavisnih vjerojatnosti $e_i(a)$ za opaženu aminokiselinu a na mjestu i. Tada je vjerojatnost novog niza x po ovom modelu:

$$P(x \mid M) = \prod_{i=1}^{L} e_i(x_i), \tag{3.3}$$

gdje je L duljina upita.

Sada nas za izračunavanje scora zanima omjer vjerojatnosti iz (3.3) i vjerojatnosti niza x pod random modelom, pa računamo *log-odds ratio*:

$$S = \sum_{i=1}^{L} log\left(\frac{e_i(x_i)}{q_{x_i}}\right). \tag{3.4}$$

Uočimo da se vrijednosti

$$log\frac{e_i(a)}{q_a}$$

ponašaju isto kao i elementi $s(x_i, y_i)$ iz (3.1) gdje umjesto aminokiseline y_i uzimamo indeks i. Iz tog razloga je ovakav pristup poznat kao PSSM.

3.5 Metoda klizećeg prozora

PSSM matricu koristimo za izračunavanje scorova S_j pri traženju upita duljine L kod nizova veće duljine nego što je sam upit.

Sliding window protocol nam kaže da score S_j izračunamo nad nizom aminokiselina duljine N počevši od $j-te$ aminokiseline, gdje j ide od prvog do $N - L + 1$-og mjesta niza. Score S_j računamo po formuli (3.4), pa na ovakav način dobivamo $N - L + 1$ score.

Ovu metodu ćemo primijeniti na skupu od K nizova aminokiselina duljina L_i $i = 1, \ldots, K$, gdje ćemo za svaki niz pamtiti maksimalni score

$$S_i = max(S_1, \ldots, S_{N-L_i+1}), \quad i = 1, \ldots, K. \tag{3.5}$$

Po teoriji ekstremnih vrijednosti opisanoj u drugom poglavlju, ovakvi maksimumi bi trebali pratiti Gumbelovu distribuciju.

Poglavlje 4

Rezultati

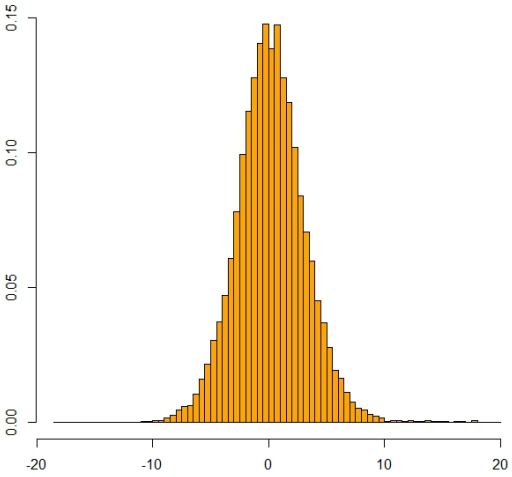
U prethodnom poglavlju smo opisali metodu klizaćeg prozora kojom se uz PSSM matricu služimo u ovom radu te smo joj pridružili potrebne parametre. Opišimo sad, na primjeru proteoma biljke *A. thaliane* (vidi dodatak 5.1), kako smo se služili tom metodom, odnosno opisanim scoring algoritmom i što smo zaključili.

4.1 Analiza distribuicije

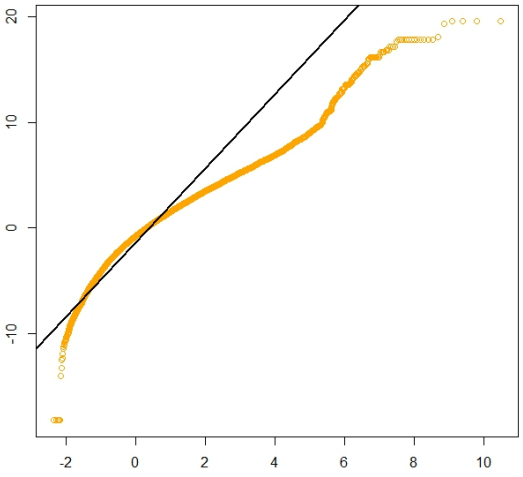
Za dani proteom biljke koji predstavlja skup proteina u određenom trenutku, pod određenim uvjetima, računamo scorove po formuli (3.4), odnosno maksimalne scorove po formuli (3.5). Odgovarajuća PSSM matrica tj. matrica profila motiva nam odgovara stvarnom upitu koji prestavlja enzim (vidi dodatak 5.2) čije varijante želimo naći u našem organizmu, tj. biljci *A. thaliani*.

Kako proteom biljke ima 35176 linija nizova različitih duljina od 16 do 5393 aminokiseline, na ovaj način smo dobili isto toliko scorova koji su maksimumi scorova za svaki redak. Njih ćemo prikazati histogramom (Slika 4.1).

Po teorijskoj podlozi navedenoj u drugom poglavlju, ti bi maksimumi trebali slijediti Gumbelovu distribuciju. No, ako pogledamo QQ-plot *(quantile-quantile plot)* usporedbe distribucije uzorka s teorijskom Gumbelovom distribucijom, nam je odmah jasno kako to nije slučaj. (Slika 4.2)

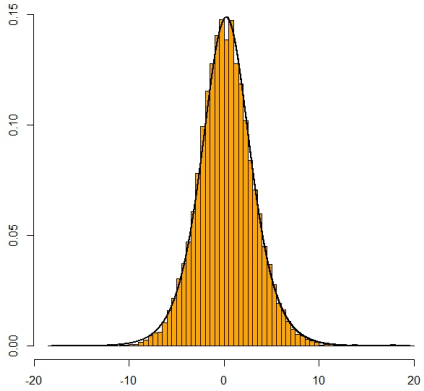


Slika 4.1: Histogram maksimuma



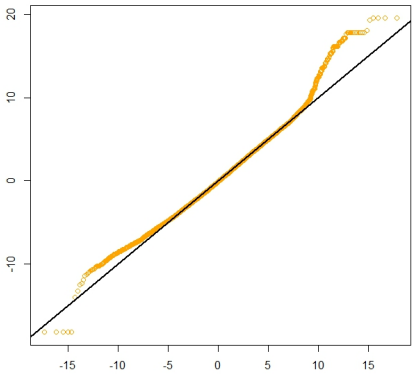
Slika 4.2: QQ-plot usporedbe, Gumbel

Na temelju histograma maksimuma i funkcije gustoće logističke distribucije s procijenjenim parametrima $\hat{\gamma} = E(X) = 0.23771$ i $\hat{\beta} = \frac{\sigma\sqrt{n}}{3} = 1.679812$, primjećujemo da bi uzorak mogao slijediti logističku distribuciju (vidi 1.4.1).



Slika 4.3: Histogram maksimuma i funkcija gustoće logističke distribucije

Našu slutnju potvrđujemo QQ-plotom usporedbe distribucije uzorka s teorijskom logističkom distribucijom gdje vidimo da nam velik dio podataka slijedi logističku distribuciju.

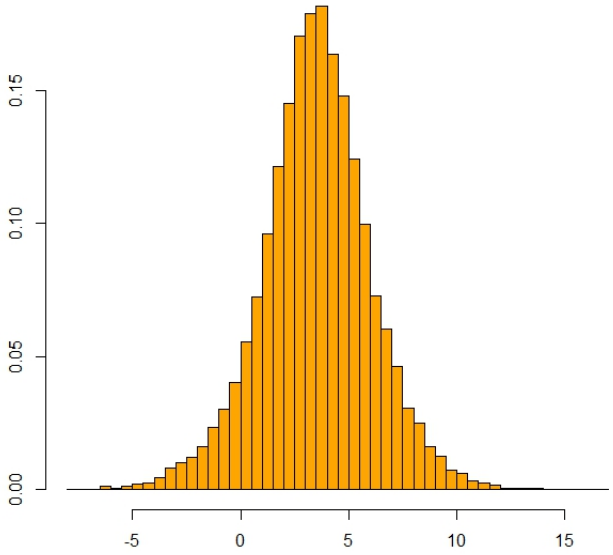


Slika 4.4: QQ plot usporedbe, logistička

4.2 Korekcija

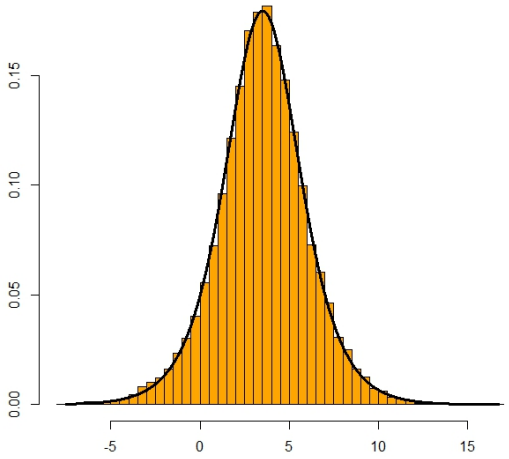
S obzirom da smo očekivali da maksimalni scorovi prate Gumbelovu distribuciju, prirodno se pitamo:“ Zašto naša metoda ne daje rezultate koji prate teorijsku podlogu?” U tu svrhu simuliramo proteom koji sadrži po 200 nizova za različite duljine niza $(100, 150, 250, \ldots, 5450)$ kako bi promotrili kako se ponašaju pripadni scorovi s obzirom na duljine niza.

Prikažimo izračunate maksimalne scorove u smislu formule (3.5) histogramom:

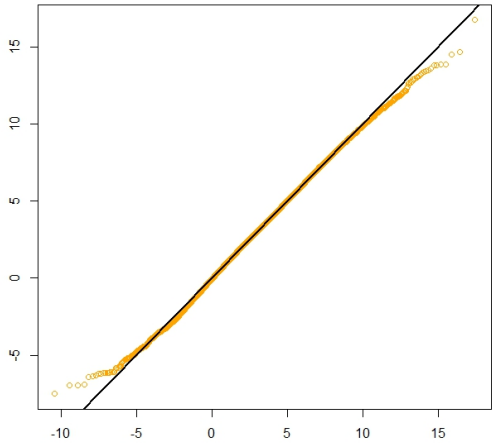


Slika 4.5: Histogram maksimuma za simulirani proteom

Opet, kao i za maksimalne scorove proteoma *A. thaliana* primjećujemo kako QQ-plot usporedbe distribucije uzorka s teorijskom distrbucijom ukazuje da maksimalni scorovi simuliranog proteoma slijede logističku distribuciju. Prikažimo QQ-plot usporedbe i histogram podataka, ali ovaj put na njega dodajmo funkciju gustoće logističke distribucije s procijenjenim parametrima lokacije $\hat{\gamma} = EX = 3.506724$ i mjere $\hat{\beta} = \frac{\sigma\sqrt{n}}{3} = 1.394566$.



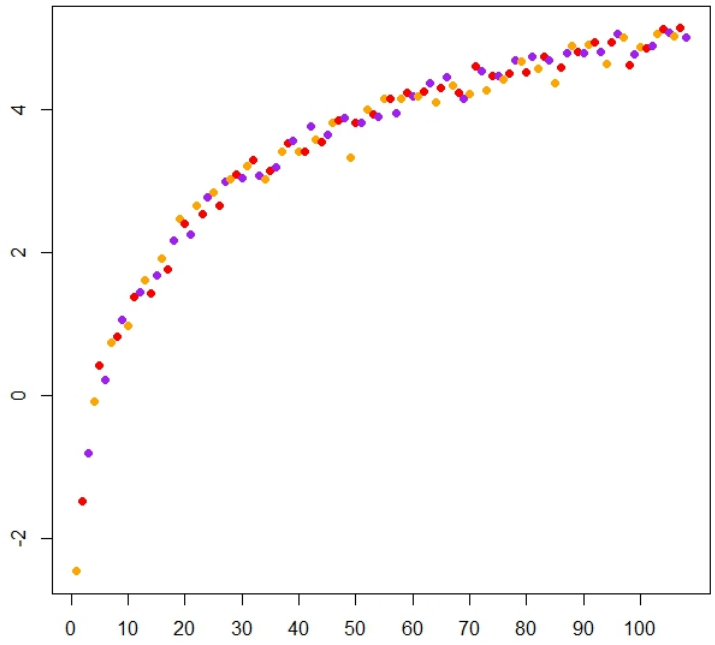
Slika 4.6: Histogram maksimuma za simulirani proteom i funkcija gustoće logističke distribucije



Slika 4.7: QQ-plot usporedbe, logistička distribucija

Promatranje nastavimo tako da uprosječimo maksimalne scorove simuliranog proteoma po duljinama te na taj način dobijemo 108 prosiječnih scorova. Simulirani proteom je skup od po 200 nizova za svaku od duljina $100, 150, \ldots, 5450$, smatramo kako će prosječni score biti dobro reprezentiran.

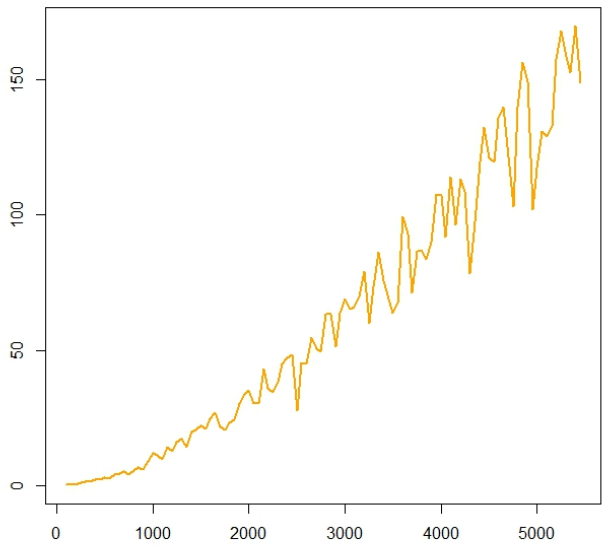
Prikažimo prosjeke sljedećom slikom:



Slika 4.8: Prosjeci maksimalnih scorova za simulirani proteom

Sada je jasno da će makismalni scorovi za nizove većih duljina biti veći, stoga želimo pronaći odgovarajuću korekciju scora za duljinu uz koju bi tako korigirani scorovi oponašali Gumbelovu distribuciju.

S obzirom da su scorovi po formuli (3.4) logaritmi, ekponencirajmo naše prosjeke kako bi linearnom regresijom tj. fitanjem na linearni model ocijenili mijenjanje scora s obzirom na duljinu.



Slika 4.9: Eksponencijalni prosjeci maksimalnih scorova za simulirani proteom

Koristit ćemo paket *lm* za regresiju iz *R-a*, besplatnog softverskog programerskog jezika za statističko računanje i grafiku. Ako linearnom regresijom fitamo eksponencijalne prosjeke na polinom prvog stupnja, dobijemo sljedeći model:

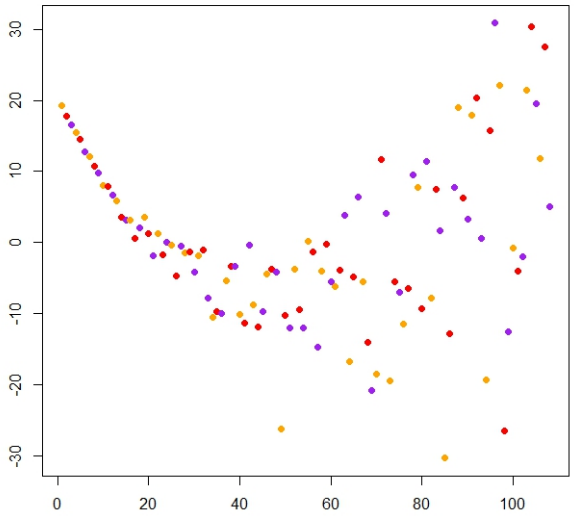
```
Call:
lm(formula = exp(prosjeci) ~ I(t))

Residuals:
    Min      1Q  Median      3Q     Max
-30.379  -7.851  -1.176   7.575  30.885

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.218e+01  2.363e+00  -9.388 1.34e-15 ***
I(t)         3.046e-02  7.424e-04  41.026  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 12.03 on 106 degrees of freedom
Multiple R-squared: 0.9408,     Adjusted R-squared: 0.9402
F-statistic:  1683 on 1 and 106 DF,  p-value: < 2.2e-16
```

Iako su koeficijenti modela ispali značajni te nam je model opisao 94.08% podataka (koeficijent determinacije $R - squared$ pokazuje koliko su dobro podaci opisani statističkim modelom), pogledom na reziduale tako dobivenom modela uočavamo kako varijance nisu homogene. Štoviše, mogli bi reći kako model nije adekvatan.



Slika 4.10: Reziduali linearnog modela prvog stupnja

Ako regresijom fitamo na polinom drugog stupnja, dobivamo sljedeći model:

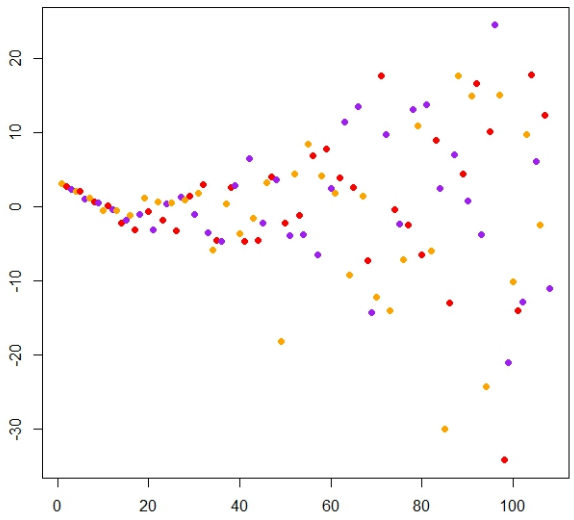
```
Call:
lm(formula = exp(prosjeci) ~ I(t) + I(t^2))

Residuals:
    Min      1Q  Median      3Q     Max
-34.281  -3.657   0.502   3.926  24.508
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.264e+00  2.892e+00  -1.474    0.143
I(t)         1.159e-02  2.402e-03   4.824 4.78e-06 ***
I(t^2)       3.400e-06  4.198e-07   8.099 1.06e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.48 on 105 degrees of freedom
Multiple R-squared: 0.9635,     Adjusted R-squared: 0.9628
F-statistic:  1387 on 2 and 105 DF,  p-value: < 2.2e-16
```

Opet ako pogledamo reziduale zaključujemo da iako su koeficijenti modela značajni i graf reziduala se neznatno popravio, model nema homogene varijance.



Slika 4.11: Reziduali linearnog modela drugog stupnja

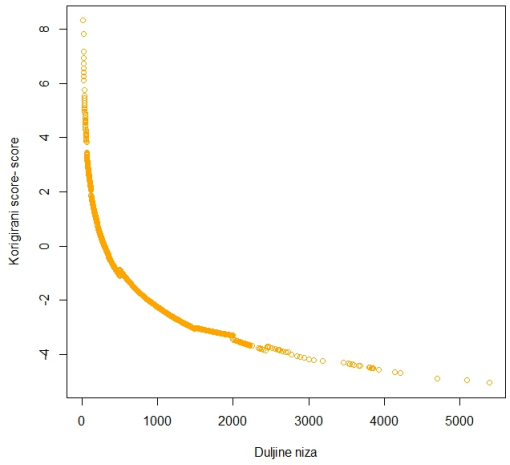
Kako regresijom na cijelom skupu uprosječenih scorova po duljinama simuliranog proteoma nismo uspjeli pronaći adekvatan model, pokušajmo napraviti regresiju po dijelovima, tj. različitim razredima u ovisnosti o duljini niza za

kojeg je score izračunat. Razrede biramo tako da nam modeli (polinomi prvog ili drugog stupnja) budu značajni i pripadni reziduali što simetričniji oko nule. Regresije po razredima provedemo na sličan način kao u gornjim primjerima, ali bez *Intercepta*. Za svaku regresiju dobivamo značajne koeficijente modela kojima korigiramo maksimume scorova dobivenih na proteomu *A.thaliane* s obzirom na duljine nizova proteoma. Jer su scorovi po formuli (3.4) logaritmi, tako dobivene koeficijente logaritmiramo prije nego ih oduzmemo od scora kojeg želimo korigirati. Ono što će se pokazati je da korigirani podaci slijede Gumbelovu distribuciju.

Konkretno, dane korekcije za nizove duljine N izgledaju ovako :

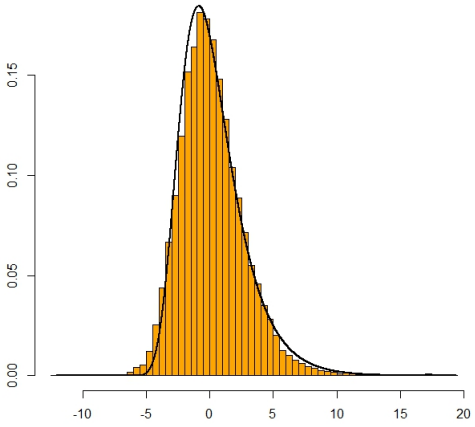
- $N \leq 60 \Longrightarrow S = S - log(\ 8.211 \times 10^{-06}\ \ N^2 - 1.161 \times 10^{-4}\ \ N)$
- $N \geq 60 \quad \& N \leq 120 \Longrightarrow S = S - log(\ 8.865 \times 10^{-06}\ \ N^2)$
- $N \geq 120 \ \& N \leq 200 \Longrightarrow S = S - log(\ 1.068 \times 10^{-05}\ \ N^2)$
- $N \geq 200 \ \& N \leq 350 \Longrightarrow S = S - log(\ 1.138 \times 10^{-05}\ \ N^2)$
- $N \geq 350 \ \& N \leq 500 \Longrightarrow S = S - log(\ 1.197 \times 10^{-05}\ \ N^2)$
- $N \geq\ 500 \ \& N \leq 900 \Longrightarrow S = S - log(\ 9.612 \times 10^{-06}\ \ N^2)$
- $N \geq\ 900 \ \& N \leq 1500 \Longrightarrow S = S - log(\ 9.455 \times 10^{-06}\ \ N^2)$
- $N \geq 1500 \ \& N \leq 2000 \Longrightarrow S = S - log(\ 0.013553\ \ N)$
- $N \geq 2000 \ \& N \leq 2450 \Longrightarrow S = S - log(\ 7.861 \times 10^{-06}\ \ N^2)$
- $N \geq 2450 \ \& N \leq 2750 \Longrightarrow S = S - log(\ 6.779 \times 10^{-06}\ \ N^2)$
- $N \geq 2750 \ \& N \leq 3100 \Longrightarrow S = S - log(\ 7.161 \times 10^{-06}\ \ N^2)$
- $N \geq 3100 \ \& N \leq 3450 \Longrightarrow S = S - log(\ 6.728 \times 10^{-06}\ \ N^2)$
- $N \geq 3450 \ \& N \leq 4050 \Longrightarrow S = S - log(\ 6.104 \times 10^{-06}\ \ N^2)$
- $N \geq 4050 \ \& N \leq 4250 \Longrightarrow S = S - log(\ 6.091 \times 10^{-06}\ \ N^2)$
- $N \geq 4250 \ \& N \leq 4650 \Longrightarrow S = S - log(\ 5.926 \times 10^{-06}\ \ N^2)$
- $N \geq 4650 \ \& N \leq 4900 \Longrightarrow S = S - log(\ 5.924 \times 10^{-06}\ \ N^2)$
- $N \geq 4900 \Longrightarrow S = S - log(\ 5.348 \times 10^{-06}\ \ N^2)$

Kako bi mogli dobiti bolju predodžbu na koji način je korekcija utjecala na naše scorove, grafički prikažimo njihovu razliku s obzirom na duljinu niza za koju je pripadni score izračunat. Sa slike (4.12) se vidi kako je korekcija za nizove duljina do 300 povećala, dok je za nizove duljina većih od 300 smanjivala score.



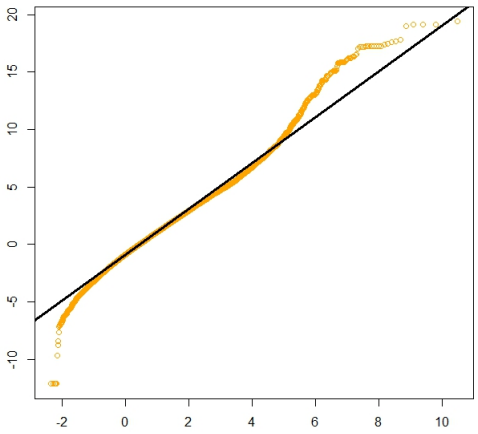
Slika 4.12: Razlike korigiranih i originalnih scorova

Na sljedećoj slici je histogram korigiranih maksimalnih scorova s funkcijom gustoće Gumbelove distribucije koja ima procjenjene parametre metodom maksimalne vjerodostojnosti ($\hat{\gamma} = 1.991326$ i $\hat{\beta} = -0.8821394$):



Slika 4.13: Histogram korigiranih scorova i funkcija gustoće Gumbelove distribucije

Iz QQ-plota usporedbe distribucije s teorijskom Gumbelovom distribucijom zaključujemo kako nam korigirani scorovi oponašaju Gumbelovu distribuibuciju.



Slika 4.14: QQ plot usporedbe, Gumbelova distribucija

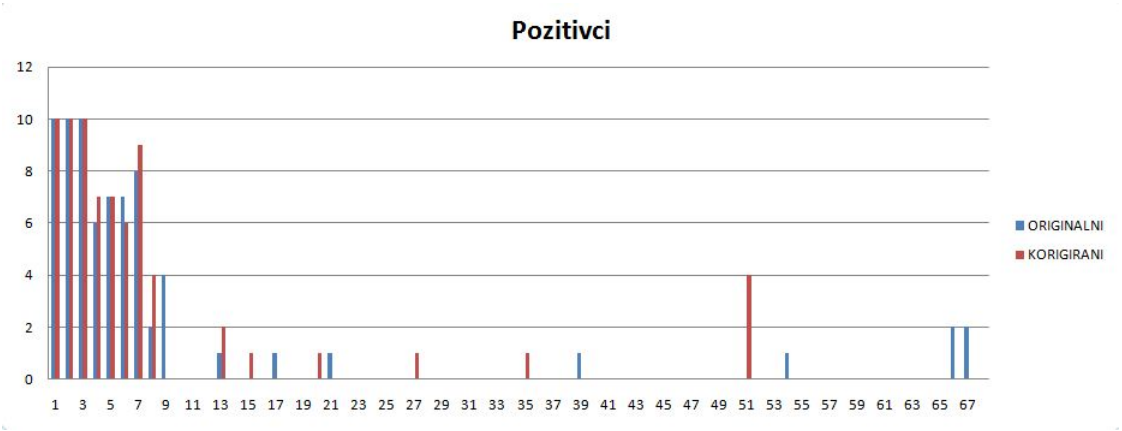
4.3 Pozitivci

Dostupna nam je lista pozitivaca za određeni upit (enzim) čije varijante želimo pronaći u *A. thaliani*. *Pozitivci* su oni enzimi koji su u funkcionalnoj, strukturalnoj i evolucijskoj vezi sa određenim enzimom. Zanima nas kako je prije navedena korekcija za scorove na *A.thaliani* utjecala na scorove pozitivaca. Štoviše, želimo proučiti da li je naša korekcija poboljšala njihovo rangiranje.

Za očekivati je da najveći scorovi budu dodjeljeni onim nizovima koji sadrže pozitivca. Ako u takvom rangiranju veliki score bude dodjeljen nizu koji ne sadrži pozitivca, njega zovemo *negativno pozitivnim* (eng. false positive).

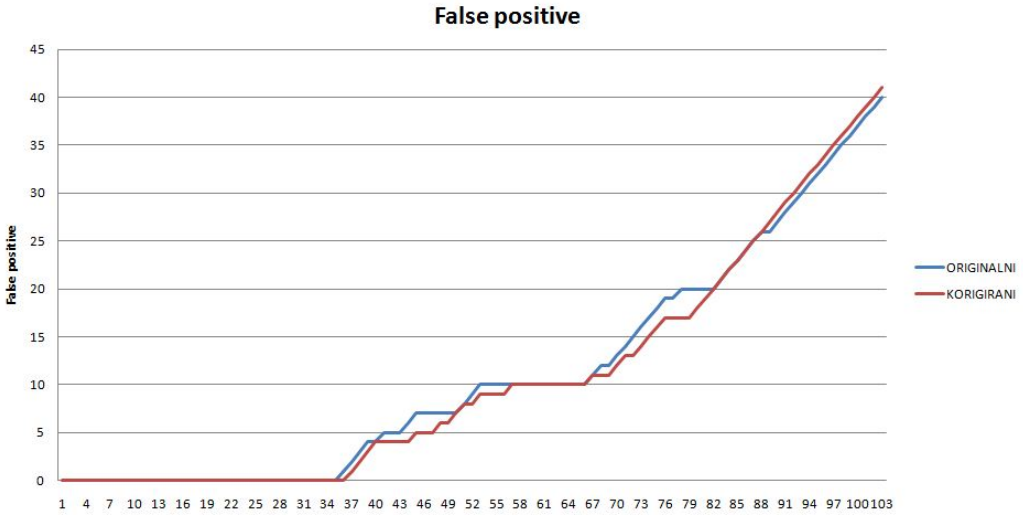
Za sortirane maksimalne scorove *A. thaliane* želimo vidjeti gdje se nalaze scorovi pripadnih nizova u kojima se nalaze naši pozitivci prije i poslije korekcije. Nacrtali smo *bar chart* usporedbe gdje smo zbog bolje preglednosti podijelili u razrede od deset nizova (Slika.(4.15)). Stupac je visok onoliko koliko ima pozitivaca u tom razredu. Jasno je da je maksimalni broj pozitivaca po razredu

iznosi 10, i da će se broj pozitivaca u razredima smanjivati kako se dodjeljeni score smanjuje.



Slika 4.15: Pozitivci po razredima

Prvih 36 maksimalnih scorova prije i poslije korekcije su scorovi izračunati nad nizovima koji su sadržavali pozitivca. Nakon toga se pojavljuju *false positive* scorovi.

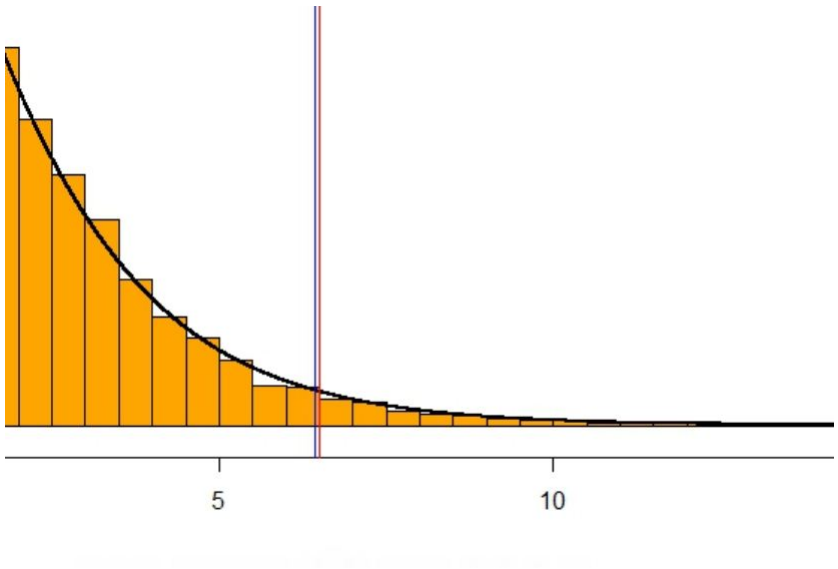


Slika 4.16: Negativni pozitivci

Sa slike vidimo da je u prvih 90 maksimalnih scorova *false pozitivnih* scorova manje ili jednako nego u nekorigiranom nizu. To predstavlja neznatno poboljšanje.

Zadnji pozitivac za korigirane scorove je u 4068. nizu koji je sortiran s obzirom na score i njegov score iznosi 2.663135, dok je za nekorigirane scorove na 3310. mjestu i score iznosi 3.5557. Ako na sličan način kao na slici(4.16) nacrtamo broj *false posivite* scorova i za ostale nizove, linije će se poklopiti pa vidimo da korekcija nije značajno utjecala na pozitivce, odnosno njihove scorove.

Pogledajmo još što se dogodilo sa 95% scorova pozitivaca. Nacrtajmo uvećani dio histograma zbog bolje preglednosti, pa se 95% scorova pozitivaca nalazi u desnom repu iza plave linije za originalne scorove, odnosno crvene za korigirane scorove.



Slika 4.17: Uvećani histogram scorova

Ono što je važno je broj negativno pozitivnih scorova u tih 95% scorova pozitivaca, što je njih 589 za nekorigirane scorova i 436 za korigirane scorove. Stoga, naša korekcija je poboljšala scorove za pozitivce.

Poglavlje 5

Dodatak

5.1 Biljka Arabidopsis thaliana

Arabidopsis thaliana je mala cvjetajuća biljka porijeklom iz Euroazije. Budući da je to jednogodišnja biljka s relativno kratkim životnim ciklusom, postala je popularan model u botanici i genetici. Među kompleksnim multićelijskim eukariotima, *A.thaliana* ima relativno mali genom od oko 135 megabaznih parova (Mbp). Dugo se smatralo da ima najmanji genom među cvjetajućim biljkama, dok je sada poznato da cvjetajuće biljke s najmanjim genomom dolaze iz roda biljaka *Genlisea*, red *Lamiales*, gdje mesožderka *Genlisea margaretae* ima genom veličine 63,5 Mpb.

To je bila prva biljka za koju su se laboratorijskim postupkom odredili svi nizovi aminokiselina genoma u jednom trenutku (tj. njen proteom).

Danas predstavlja popularan alat za razumijevanje genetike, evolucije, populacijske genetike i razvoja biljaka. Njenu ulogu možemo usporediti s onom vinskih mušica i miševa u zoologiji. Iako Athaliana ima neznatan utjecaj na poljoprivredu, ima nekoliko osobina koje su korisne za razumijevanje genetičke, stanične i molekularne biologije cvjetajućih biljaka uključujući razvitak cvijeta i opažanje svijetla.

5.2 Enzimi GDSL skupine

Prikazat ćemo listu pozitivaca, odnosno onih enzima iz *A. thaliane* koji su u funkcionalnoj, evolucijskoj i strukturalnoj vezi s enzimom FVFGDSLSDA čije var.ijante želimo pronaći. To su enzimi iz familije GDSL hidrolaza koja je relativno nova i neistražena skupina enzima s velikim potencijalom za primjenu u farmaceutskoj i prehrambenoj industriji, kao i za daljnja biokemijska istraživanja. Posebnost ove enzimske obitelji su izuzetna supstratna nespecifičnost te tipični GDSL motiv oko serina iz katalitičke trijade, koji se razlikuje od G-S-G motiva iz većine dosad istraženih lipaza.

FVFGDSLVDA	FVFGDSVFDA	FAFGDSLFEA	FVFGDSSVDS	FVSSNSLSDT
FVFGDSLVDA	FVFGDSVFDA	FVFGDSVFDN	FTFGDSNFDA	FVSSNSLSDT
FVFGDSLFDA	FVFGDSLIDN	FVFGDSVFDN	FTFGDSNFDA	FVSSNSLSDT
YVFGDSLVDA	FIFGDSLVDS	FVFGDSVFDN	FTFGDSNFDA	FVSSNSLSDT
FVFGDSLVDS	FIFGDSLVDS	FNFGDSNSDT	FIFGDSVVDV	YAFGGSLSDF
FVFGDSLVDS	FVFGDSLVEV	FNFGDSNSDT	FAFGDSILDT	YAFGDSFTDT
FVFGDSLVDS	FIFGDSLVDN	FNFGDSNSDT	FAFGDSILDT	FVWGESISDG
FIFGDSLVDA	FIFGDSLVDN	FNFGDSNSDT	YNFGDSNSDT	YIFGDSLTEV
FVFGDSLVDN	FIFGDSLVDN	FNFGDSNSDT	FLFGDSFLDA	FVISGSLNDA
FVFGDSLVDN	FVLGDSLVDA	FNFGDSNSDT	FAFGDSILDT	FVISGSLHDA
FVFGDSLVDN	FIFGDSLVDV	FNFGDSNSDT	YVIGDSLVDS	FVFSGSLVNG
FNFGDSLSDT	FVFGDSYADT	FNFGDSNSDT	YQFGDSISDT	FTFGDSSYDV
FIFGDSLSDV	FVFGDSYADT	FTFGDSIFDA	YVYGDSLLDG	YAFGDSTVDS
FVFGDSMSDN	FVFGDSYADT	FNFGDSNSDT	YVIGDSLVDP	FLFGDSITEE
FVFGDGLYDA	FVFGDSYADT	FTFGDSYYDA	FAFGDSVLDT	

Bibliografija

[1] Isabel Fraga Alves i Cláudia Neves, *Extreme Value Distributions*, International Encyclopedia of Statistical Science (Miodrag Lovric, ur.), Springer Berlin Heidelberg, 2011, str. 493–496, ISBN 978-3-642-04897-5.

[2] R. Durbin, *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press, 1998, ISBN 9780521629713.

[3] S.I. Resnick, *Extreme values, regular variation, and point processes*, Applied probability, Springer-Verlag, 1987, ISBN 9780387964812.

[4] N. Sarapa, *Teorija vjerojatnosti*, Udžbenici Sveučilišta u Zagrebu, Školska knjiga, 2002, ISBN 9789530308169.

[5] T. D. Schneider i R. M. Stephens, *Sequence Logos: A New Way to Display Consensus Sequences*, Nucleic Acids Res. **18** (1990), 6097–6100.

Sažetak

U ovom diplomskom radu analiziramo distribuciju maksimalnih scorova dobivenih za stvarni enzim iz familije GDSL hidrolaza na biljci *A. thaliani*. Tema nas je zaintrigirala zbog činjenice što takva distribucija ne odgovara Gumbelovoj distribuciji iz familije distribucija ekstremnih vrijednosti koja je teorijski opravdana.

Izračunavamo scorove za upit koristeći se metodom klizećeg prozora i specifičnom matricom težina i analiziramo njihovu distribuciju. Ono što primjećujemo je da uz određenu korekciju scorova s obzirom na duljinu niza nad kojim su izračunati, uspijevamo dobiti Gumbelovu razdiobu.

Na kraju provjeravamo kako je korekcija utjecala na scorove pozitivaca, odnosno onih enzima za koje znamo da su u funkcionalnoj vezi s našim enzimom čije smo varijante želili pronaći u proteomu biljke. To je od iznimne važnosti prvenstveno zbog činjenice što želimo da korekcijom oni enzimi koji su u vezi sa upitom i dalje imaju veći score od onih koji nisu. Pokaže se da naša korekcija ne kvari takav odnos.

Summary

In this thesis we have analysed the distribution of the maximum scores obtained for an actual enzyme from GDSL family of hydrolases on the plant *A. thaliana*. This is an intriguing topic, since such distribution does not match Gumbel distribution, as it should, according to relevant theoretical results.

For the purpose of analysing such distribution, we have calculated scores for query using the sliding window protocol and PSSM matrix. What we have noticed is that a certain correction of scores with respect to the length of the protein, gives Gumbel distribution.

Finally we check how the correction affected scores of positive matches, i.e. those enzymes that are in functional relation with our query whose variants we wanted to find in the proteome. That is rather important, primarily since we want enzymes which are related to the query to keep higher scores than those that are not. We show that our correction keeps such a relationship.

Životopis

Rođena sam 15.01.1991 godine u Splitu. Osnovno školovanje završavam u OŠ Dobri u Splitu, u razdoblju od 1997.do 2005. godine. Od 2005.do 2009. godine pohađam III. Gimnaziju u Splitu. Upisujem 2009. preddiplomski studij Matematike na PMF-MO u Zagebu koji završavam 2012. godine. Iste godine nastavljam školovanje upisivanjem diplomskog studija Matematičke statistike na PMF-MO u Zagrebu.