Iterativna optimizacija modela i pretraživanje proteoma

Cigula, Maja

Master's thesis / Diplomski rad

2016

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

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:217:877304

Rights / Prava: In copyright / Zaštićeno autorskim pravom.

Download date / Datum preuzimanja: **2024-07-10**

Repository / Repozitorij:

Repository of the Faculty of Science - University of Zagreb

SVEUČILIŠTE U ZAGREBU
PRIRODOSLOVNO–MATEMATIČKI FAKULTET
MATEMATIČKI ODSJEK

Maja Cigula

ITERATIVNA OPTIMIZACIJA MODELA I PRETRAŽIVANJE PROTEOMA

Diplomski rad

Zagreb, veljača, 2016.

SVEUČILIŠTE U ZAGREBU
PRIRODOSLOVNO–MATEMATIČKI FAKULTET
MATEMATIČKI ODSJEK

Maja Cigula

ITERATIVNA OPTIMIZACIJA MODELA I PRETRAŽIVANJE PROTEOMA

Diplomski rad

Voditelj rada:
doc. dr. sc. Pavle Goldstein

Zagreb, veljača, 2016.

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. _____

Mami i tati

Sadržaj

Uvod

U bioinformatici se često postavlja pitanje pripadnosti proteina nekoj familiji. Cilj rada je opisati metode čija je svrha identifikacija nizova od interesa u proteomu. Naime, uz dani proteom i zadani motiv želimo naći proteine s maksimalnim scorovima. To postižemo iterativnim stvaranjem modela. Koristimo dani inicijalni profil motiva pomoću kojega stvaramo listu pozitivno rangiranih nizova. Nju koristimo kako bismo izgradili novi model pomoću kojeg opet prolazimo kroz dani skup podataka, točnije proteom. Iteracije prestaju kada se lista pozitivaca prestane mijenjati. Objasnit ćemo dvije metode i usporediti ih. Obje metode testirane su na stvarnim biljnim proteomima.

Poglavlje 1

Uvodni pojmovi iz vjerojatnosti

1.1 Vjerojatnost. Slučajna varijabla.

Na početku ćemo se prisjetiti glavnih pojmova vezanih za teoriju vjerojatnosti.

Neka je Ω proizvoljan neprazan skup. Familija $\mathcal{F}$ podskupova od Ω je **σ-algebra** skupova (na Ω) ako je:
1) $\emptyset \in \mathcal{F}$
2) $A \in \mathcal{F} \implies A^c \in \mathcal{F}$
3) $A_i \in \mathcal{F} \implies \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

Uređen par $(\Omega, \mathcal{F})$ zove se **izmjeriv prostor**.

Definicija 1.1.1. *Neka je $(\Omega, \mathcal{F})$ izmjeriv prostor. Funkcija $\mathbb{P} : \mathcal{F} \to \mathbb{R}$ je **vjerojatnost** (na $\mathcal{F}$) ako vrijedi:*
1) $\mathbb{P}(A) \geqslant 0$, $\forall A \in \mathcal{F}$
2) $\mathbb{P}(\Omega) = 1$
3) $A_i \in \mathcal{F}$, $i \in \mathbb{N}$ *i* $A_i \cap A_j = 0$ *za* $i \neq j \Rightarrow \mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$.

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

Neka je $\mathcal{B}$ Borelova σ-algebra generirana familijom svih otvorenih skupova na $\mathbb{R}$.

Definicija 1.1.2. *Neka je $(\Omega, \mathcal{F}, \mathbb{P})$ vjerojatnosni prostor. Funkcija $X : \Omega \to \mathbb{R}$ je **slučajna varijabla** ako je $X^{-1}(B) \in \mathcal{F}$ za proizvoljno $B \in \mathcal{B}$ tj. $X^{-1}(\mathcal{B}) \subset \mathcal{F}$.*

Često nas zanima problem vezan za određenu slučajnu varijablu X. Tada je pogodnije operirati sa vjerojatnosnim prostorom induciranim s X. Za $B \in \mathcal{B}$ stavimo:

$$\mathbb{P}_X(B) = \mathbb{P}(X^{-1}(B)) = \mathbb{P}(\omega \in \Omega : X(\omega) \in B) = \mathbb{P}(X \in B) \tag{1.1}$$

Time je definirana funkcija $\mathbb{P}_X : \mathcal{B} \to [0,1]$ koju zovemo **vjerojatnosna mjera** na $\mathcal{B}$. Svakoj je slučajnoj varijabli X preko relacije (1.1) na prirodan način pridružen vjerojatnosni prostor $(\mathbb{R}, \mathcal{B}, \mathbb{P}_X)$ induciran slučajnom varijablom X.

Uvjetna vjerojatnost i nezavisnost

Neka je $A \in \mathcal{F}$ takav da je $\mathbb{P}(A) > 0$. Definirajmo funkciju $\mathbb{P}_A : \mathcal{F} \to [0,1]$:

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

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

Definicija 1.1.3. *Neka su $X_1, X_2, ..., X_n$ slučajne varijable na vjerojatnosnom prostoru $(\Omega, \mathcal{F}, \mathbb{P})$. Kažemo da su $X_1, X_2, ..., X_n$* ***nezavisne*** *ako za proizvoljne $B_i \in \mathcal{B}$, i = 1,2, ... , n vrijedi:*

$$\mathbb{P}(\bigcap_{i=1}^{n}(X_i \in B_i)) = \prod_{i=1}^{n} \mathbb{P}(X_i \in B_i)$$

1.2 Funkcija distribucije

Definicija 1.2.1. *Neka je X slučajna varijabla na Ω.* ***Funkcija distribucije*** *od X je funkcija $F_X : \mathbb{R} \to [0,1]$ definirana s*

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

Postoje dva glavna tipa slučajnih varijabli: diskretne i neprekidne.

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

Slučajna varijabla X je **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

$$F_X(x) = \int_{\infty}^{x} f(t) d\lambda(t), \quad x \in \mathbb{R}. \tag{1.2}$$

Za funkciju distribucije F_X neprekidne slučajne varijable X, dakle za funkciju oblika (1.2), kažemo da je apsolutno neprekidna funkcija distribucije. Ako je X neprekidna slučajna varijabla, tada se funkcija f iz (1.2) zove **funkcija gustoće** vjerojatnosti od X ili, kraće, gustoća od X i ponekad je označujemo s f_X.

Matematičko očekivanje i varijanca

Da bismo definirali matematičko očekivanje potrebno nam je više koraka. Prvo se definira matematičko očekivanje jednostavne slučajne varijable, zatim nenegativne i na kraju opće slučajne varijable. Navodimo samo očekivanje diskretne slučajne varijable i iskazujemo formulu za očekivanje koja se često koristi u teoriji vjerojatnosti.

Neka je X diskretna slučajna varijabla i neka je D skup iz definicije diskretne slučajne varijable, $D = \{x_1, x_2, ...\}$, te za svako k vrijedi $\mathbb{P}_{\mathbb{X}}(\{x_k\}) = p_k$. Tada je očekivanje slučajne varijable X dano s

$$\mathbb{E}[X] = \sum_k x_k p_k.$$

Neka je sada X neprekidna slučajna varijabla s funkcijom distribucije F_X. Očekivanje slučajne varijable X dano je s

$$\mathbb{E}[X] = \int_\Omega X \, d\mathbb{P} = \int_{\mathbb{R}} x \, dF_X(x).$$

Za Borelovu funkciju $g : \mathbb{R} \to \mathbb{R}$ vrijedi

$$\mathbb{E}[g(X)] = \int_\Omega g(X) \, d\mathbb{P} = \int_{\mathbb{R}} g(x) \, dF_X(x).$$

Neka $\mathbb{E}[X]$ postoji tj. konačno je. Tada $\mathbb{E}[(X - \mathbb{E}X)^r]$ zovemo r-ti centralni moment od X.
Varijanca od X, u oznaci $VarX$ ili σ_X^2, je drugi centralni moment od X, tj.

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

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

Konvergencija

Podsjetimo se osnovnih tipova konvergencije slučajnih varijabli. Neka je $(X_n, n \in \mathbb{N})$ niz slučajnih varijabli definiran na vjerojatnosnom prostoru $(\Omega, \mathcal{F}, \mathbb{P})$.

Kažemo da niz $(X_n, n \in \mathbb{N})$ slučajnih varijabli konvergira **gotovo sigurno (g.s.)** prema slučajnoj varijabli X ako je

$$\mathbb{P}\big(\omega \in \Omega : X(\omega) = \lim_{n\to\infty} X_n(\omega)\big) = 1$$

Oznaka je (g.s.) $\lim_{n\to\infty} X_n = X$ i takav je limes g.s. jedinstven.

Kažemo da niz $(X_n, n \in \mathbb{N})$ slučajnih varijabli konvergira **po vjerojatnosti** prema slučajnoj varijabli X ako za svaki $\epsilon > 0$ vrijedi

$$\lim_{n\to\infty} \mathbb{P}\big(|X_n - X| \geqslant \epsilon\big) = 0$$

Oznaka je (P) $\lim_{n\to\infty} X_n = X$ i takav je limes također (g.s.) jedinstven.

Kažemo da niz $(X_n, n \in \mathbb{N})$ slučajnih varijabli konvergira **po distribuciji** prema slučajnoj varijabli X ako je

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x), \quad x \in C(F_X)$$

gdje je $C(F_X)$ skup svih točaka neprekidnosti funkcije F_X. Oznaka je $(\mathcal{D})$ $\lim_{n\to\infty} X_n = X$.

Vrijede sljedeće implikacije među navedenim tipovima konvergencije:

$$(g.s.) \lim_{n\to\infty} X_n = X \implies (P) \lim_{n\to\infty} X_n = X$$

$$(P) \lim_{n\to\infty} X_n = X \implies (\mathcal{D}) \lim_{n\to\infty} X_n = X$$

1.3 Karakteristična funkcija i svojstva

Da bismo u radu lakše pokazali određene veze između slučajnih varijabli koristimo svojstva karakterističnih funkcija. Metoda karakterističnih fukncija osnovno je sredstvo analitičkog aparata teorije vjerojatnosti. Korisnost karakterističnih funkcija posljedica je činjenice da postoji 1-1 korespondencija između skupa karakterističnih funkcija i skupa funkcija distribucije.

Neka je F ograničena (poopćena) funkcija distribucije na $\mathbb{R}$.

Definicija 1.3.1. ***Karakteristična funkcija od*** *F jest funkcija ρ definirana sa*

$$\rho(t) = \int_{-\infty}^{\infty} e^{itx} dF(x) = \int_{-\infty}^{\infty} cos(tx) dF(x) + i \int_{-\infty}^{\infty} sin(tx) dF(x), \quad t \in \mathbb{R} \tag{1.3}$$

Za svako $t \in \mathbb{R}$ funkcija $x \mapsto e^{itx}$ je neprekidna i budući da je $|e^{itx}| = 1$, ρ je dobro definirana, tj. imamo $\rho : \mathbb{R} \to \mathbb{C}$.

Definicija 1.3.2. *Neka je X slučajna varijabla s funkcijom distribucije F_X.* ***Karakteristična funkcija*** *ρ_X* ***od*** *X je karakteristična funkcija od F_X.*

Ako je X neprekidna slučajna varijabla s gustoćom f_x, tada je

$$\rho_X(t) = \int_{-\infty}^{\infty} e^{itx} f_X(x) d(x) \tag{1.4}$$

.

Propozicija 1.3.3. *(a) Ako je X slučajna varijabla i $a, b \in \mathbb{R}$, tada vrijedi*

$$\rho_{aX+b}(t) = e^{ibt} \rho_X(at), \quad t \in \mathbb{R} \tag{1.5}$$

(b) Ako su $X_1, X_2, ..., X_n$ nezavisne slučajne varijable, tada vrijedi

$$\rho_{\sum_{i=1}^{n} X_i}(t) = \prod_{i=1}^{n} \rho_{X_i}(t), \quad t \in \mathbb{R} \tag{1.6}$$

Sljedeći teorem pokazuje da je korespondencija između funkcija distribucije i karakterističnih funkcija 1-1 korespondencija.

Teorem 1.3.4. *(Teorem jedinstvenosti.) Neka su F_1 i F_2 funkcije distribucije na $\mathbb{R}$ i neka one imaju istu karakterističnu funkciju, tj. za sve $x \in \mathbb{R}$ vrijedi*

$$\int_{-\infty}^{\infty} e^{itx} dF_{X_1}(x) = \int_{-\infty}^{\infty} e^{itx} dF_{X_2}(x) \tag{1.7}$$

Tada je $F_1 = F_2$.

1.4 Primjeri slučajnih varijabli

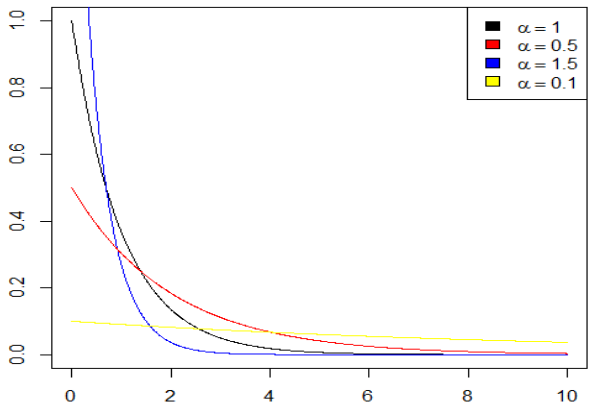
Iz teorije vjerojatnosti je poznato da ako imamo gustoću neprekidne slučajne varijable X, znamo vjerojatnosti svih događaja koji su u vezi sa tom slučajnom varijablom. Također, funkcija distribucije neprekidne slučajne varijable X u potpunosti je određena njezinom gustoćom. Navodimo primjere funkcija gustoće slučajnih varijabli koje će se pojavljivati u radu.

Eksponencijalna slučajna varijabla

Neprekidna slučajna varijabla X ima eksponencijalnu distribuciju ako joj je funkcija gustoće zadana s

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geqslant 0 \\ 0 & x < 0 \end{cases} \tag{1.8}$$

gdje je parametar $\lambda > 0$ fiksan.
Očekivanje eksponencijalne slučaje varijable jednako je $\mathbb{E}[X] = \frac{1}{\lambda}$, a varijanca $Var(X) = \frac{1}{\lambda^2}$.



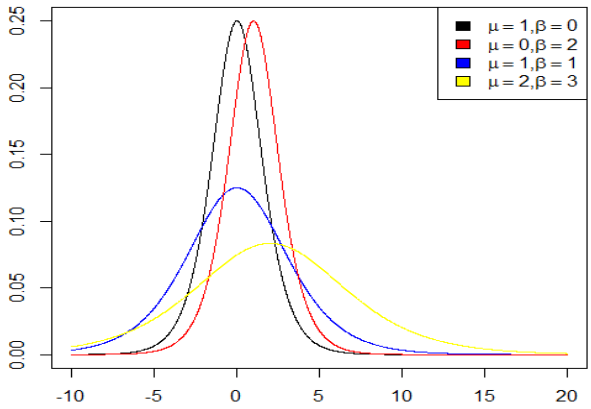
Slika 1.1: Funkcije gustoće eksponencijalne distribucije s različitim parametrima λ

Logistička distribucija

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

$$f(x) = \frac{e^{-\frac{x-\mu}{\beta}}}{\beta\left(1 + e^{-\frac{x-\mu}{\beta}}\right)^2}, \quad x \in \mathbb{R}. \tag{1.9}$$

Za očekivanje vrijedi $\mathbb{E}[X] = \mu$, a varijancu $Var(X) = \frac{\beta^2\pi^2}{3}$.



Slika 1.2: Funkcije gustoće logističke distribucije s različitim parametrima

Na kraju poglavlja, navedimo još generaliziranu formulu funkcije gustoće logističke slučajne varijable i njenu karakterističnu funkciju.
Označimo sa Y generaliziranu logističku slučajnu varijablu. Neka su $p, q > 0$. Generalizirana funkcija gustoće dana je s:

$$f_Y(y; p, q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} \frac{e^{py}}{(1+e^y)^{p+q}}, \quad y \in \mathbb{R} \tag{1.10}$$

Njena karakteristična funkcija jednaka je:

$$\rho_Y(t) = \frac{\Gamma(p+it)\Gamma(q-it)}{\Gamma(p)\Gamma(q)} \tag{1.11}$$

Poglavlje 2

Teorija ekstremnih vrijednosti

Teorija ili analiza ekstremnih vrijednosti grana je statistike koja, kako joj samo ime govori, proučava ekstremne vrijednosti. Najčešće nas zanima distribucija maksimalnih vrijednosti. Rezultati se iskazuju s obzirom na maksimum zbog relacije:

$$-max(-X) = min(X).$$

2.1 Distribucije ekstremnih vrijednosti

Neka je $X_1, X_2, ..., X_n, n \in \mathbb{N}$ niz nezavisnih jednako distribuiranih slučajnih varijabli sa funkcijom distribucije F. S M_n označimo $\max\{X_1, X_2, ..., X_n\}$. Znamo da je M_n slučajna varijabla i zanima nas funkcija distribucije od M_n.

$$F_{M_n}(x) = \mathbb{P}(M_n \leqslant x) = \mathbb{P}(\bigcap_{i=1}^{n}(X_i \leqslant x)) = \prod_{i=1}^{n}\mathbb{P}(X_i \leqslant x) = \prod_{i=1}^{n}F(x) = F^n(x), x \in \mathbb{R}$$

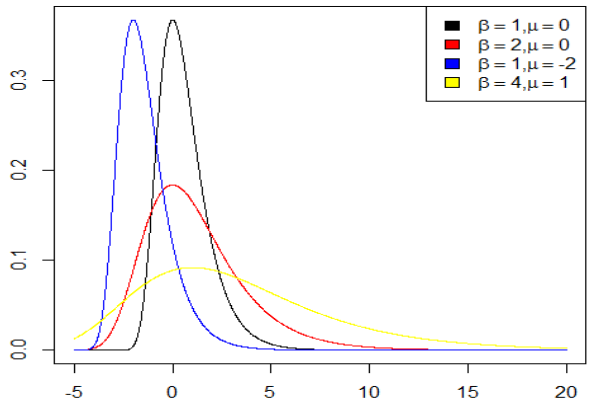
U primjenama je teško računati s funkcijom F^n. U tome nam pomaže teorem na kojem je zasnovana teorija ekstremnih vrijednosti koji kaže da postoje samo tri tipa distribucije potrebnih za modeliranje maksimuma.

Gumbelova distribucija

Gumbelova distribucija nazvana je u čast njemačkom matematičaru Emilu Juliusu Gumbelu (1891.-1966.). Ona je tip I distribucije ekstremnih vrijednosti. Neka su $\mu \in \mathbb{R}$ i $\beta > 0$. Neprekidna slučajna varijabla X ima Gumbelovu distribuciju sa parametrima μ i β ako joj je funkcija gustoće zadana s:

$$f(x) = \frac{1}{\beta}e^{-\frac{x-\mu}{\beta}-e^{-\frac{x-\mu}{\beta}}}, \quad x \in \mathbb{R}. \tag{2.1}$$

Za vrijednost parametra lokacije $\mu = 0$ i parametra mjere $\beta = 1$ nazivamo je standardnom Gumbelovom distribucijom. Očekivanje je izraženo s $\mathbb{E}[X] = \gamma + \sigma\gamma$, gdje je γ Eulerova konstanta. Varijanca je jednaka $Var(X) = \frac{\gamma^2\pi^2}{6}$.



Slika 2.1: Funkcije gustoće Gumbelove distribucije sa različitim parametima

S Y označimo generaliziranu Gumbelovu slučajnu varijablu. Navedimo njenu generaliziranu funkciju gustoće. Za $p > 0$ ona glasi:

$$g(y) = \frac{1}{\Gamma(p)} e^{-py} e^{-e^{py}}, \quad y \in \mathbb{R} \tag{2.2}$$

Karakteristična funkcija dana je s:

$$\rho_Y(t) = \frac{\Gamma(p - it)}{\Gamma(p)} \tag{2.3}$$

Zanimljiva je veza između dvije nezavisne Gumbel distribuirane slučajne varijable. Navodimo je u obliku korolara prethodno iskazanih tvrdnji.

Korolar 2.1.1. *Neka su X_1 i X_2 nezavisne generalizirane Gumbel distribuirane slučajne varijable. Slučajna varijabla $Y = X_1$ - X_2 ima generaliziranu logističku distribuciju s parametrima p i q.*

Dokaz. Odgovarajuće karakteristične funkcije jednake su:

$$\rho_{X_1}(t) = \frac{\Gamma(p-it)}{\Gamma(p)}, \quad \rho_{X_2}(t) = \frac{\Gamma(q-it)}{\Gamma(q)}$$

Koristeći (1.5) i (1.6) dobivamo:

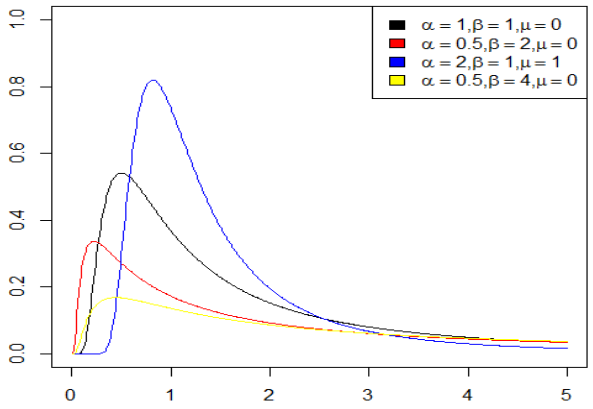
$$\rho_Y(t) = \rho_{X_1-X_2}(t) = \rho_{X_1}(t)\rho_{X_2}(-t) = \frac{\Gamma(p-it)\Gamma(q+it)}{\Gamma(p)\Gamma(q)}.$$

Primijetimo da smo dobili karakterističnu funkciju generalizirane logističke slučajne varijable. Iz teorema jedinstvenosti slijedi tvrdnja. □

Fréchetova distribucija

Tip II distribucije ekstremnih vrijednosti nazvana je po francuskom matematičaru Maurice Fréchetu (1878.-1973.).
Neka su $\alpha > 0, \beta > 0$ i $\mu \in \mathbb{R}$. Funkcija gustoće Fréchetove distribucije dana je s:

$$f(x) = \frac{\alpha}{\beta}\Big(\frac{\beta}{x-\mu}\Big)^{\alpha+1} e^{-(\frac{\beta}{x-\mu})^{\alpha}}, \quad x \in \mathbb{R} \tag{2.4}$$



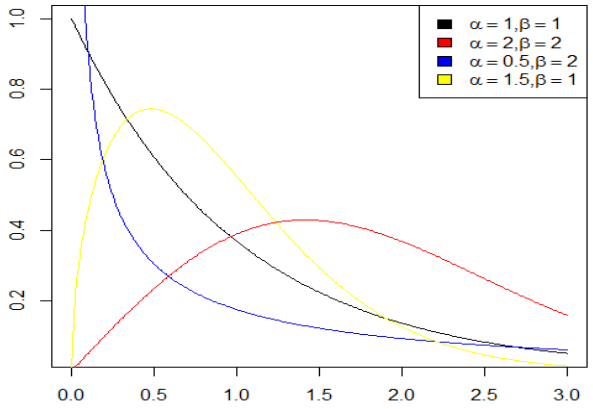
Slika 2.2: Funkcije gustoće Fréchetove distribucije s različitim parametrima

Očekivanje je definirano za $\alpha > 0$ i jednako je $\mathbb{E}[X] = \mu + \beta\Gamma(1+\frac{1}{\alpha})$. Varijanca je definirana za $\alpha > 2$ i vrijedi $Var(X) = \beta^2(\Gamma(1-\frac{2}{\alpha}) - (\Gamma(1-\frac{1}{\alpha}))^2)$

Weibullova distribucija

Waloddi Weibull (1887.-1979.), bio je švedski matematičar po kojemu je Tip III distribucije ekstremnih vrijednosti nazvana Weibulovom distribucijom.
Neka su $\alpha > 0, \beta > 0$. Parametar α nazivamo parametar oblika, a β parametar mjere. Funkcija gustoće Weibullove distribucije dana je s:

$$f(x) = \begin{cases} \frac{\alpha}{\beta}((\frac{x}{\beta})^{\alpha-1}e^{-(\frac{x}{\beta})^{\alpha}} & x \geqslant 0 \\ 0 & x < 0 \end{cases} \tag{2.5}$$



Slika 2.3: Funkcije gustoće Weibullove distribucije s različitim parametrima

Očekivanje je jednako $\mathbb{E}[X] = \alpha\Gamma(1 + \frac{1}{\beta})$. Varijanca je jednaka $Var(X) = \alpha^2(\Gamma(1 + \frac{2}{\beta}) - (\Gamma(1 + \frac{1}{\beta}))^2)$.
Očito je da je za vrijednosti parametara $\alpha = 1$ i $\beta = \frac{1}{\lambda}$ Weibullova distribucija zapravo eksponencijalna.

2.2 Fisher - Tippet - Gnedenkov teorem

Navodimo teorem bez dokaza.

Teorem 2.2.1. *(Fisher - Tippet, 1928; Gnedenko,1943.) Neka su $X_1, X_2, \ldots, X_n$ jednako distribuirane slučajne varijable i neka je $M_n = \max\{X_1, \ldots, X_n$. Ako postoji $a_n > 0$ i $b_n \in \mathbb{R}$ tako da $\lim_{n\to\infty} P\left(\frac{M_n - b_n}{a_n} \leqslant x\right) = F(x)$, gdje je F nedegenerična distribucija, tada granična distribucija F pripada Gumbel, Fréchet ili Weibull distribuciji.*

Distribucije tipa I, tipa II i tipa III možemo zapisati u generaliziranom obliku. Ona ima funkciju gustoće

$$f(x;\alpha,\beta,\mu) = \begin{cases} \frac{1}{\beta}(1+\alpha z)^{-1-\frac{1}{\alpha}} e^{-(1+\alpha z)^{-\frac{1}{\alpha}}} & \alpha \neq 0 \\ \frac{1}{\beta} e^{-z-e^{-}z} & \alpha = 0 \end{cases} \tag{2.6}$$

gdje je $z = \frac{x-\mu}{\beta}$, za $\alpha, \mu \in \mathbb{R}, \beta > 0$ parametre oblika, lokacije i mjere.

Spomenimo još vezu između generalizirane funkcije distribucije ekstremnih vrijednosti i logističke distribucije.
Neka su X i Y generalizirane slučajne varijable ekstremnih vrijdnosti s parametrima $(\alpha, \beta, 0)$, tada je njihova razlika $X - Y$ logistički distribuirana slučajna varijabla s parametrima $\alpha = 0$ i β.

Poglavlje 3

Opis modela

3.1 Općenito o proteomu

Proteom je skup svih proteina koje neki organizam ili stanica može proizvesti u određenom vremenu pod određenim uvjetima. Proteini su osnovne građevne jedinice živih bića. Sastavljeni su od aminokiselina. Aminokiseline su prirodni spojevi koji u prirodi rijetko dolaze u slobodnom stanju. Uglavnom su međusobno povezane u makromolekule, peptide i proteine. U proteinima se obično nalazi oko dvadeset različitih aminokiselina.

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

Tablica 3.1: Glavne aminokiseline i njihove kratice

Motivom se smatra najmanja strukturalna jedinica koja opstaje ili nestaje evolucijom proteina. To je niz od 10 do 20 aminokiselina u proteinu na kojem se vidi jasan supstitucijski uzorak. Oblik u kojem se motiv zapisuje naziva se profil motiva. Profil motiva zadan je nizom od n slučajnih varijabli. Primjerice, ako je duljina motiva 10 to znači da je karakteriziran matricom dimenzije 10×20.

U radu će od interesa biti motivi koji karakteriziraju enzime GDSL familije. Ona uključuje i hidrolitičke enzime s multifunkcionalnim svojstvima i velikim potencijalom za primjenu u raznim industrijama. Otkriveno je da bi biljke mogle biti dobar izvor novih enzima, pa je zato njihovo proučavanje od općeg interesa. Tipičan niz koji pripada ovoj familiji karakteriziran je s više motiva, ali u radu opisujemo metode pretraživanja samo jednim.

3.2 PSSM

PSSM (eng. *position specific scoring matrix*), još je poznata i kao matrica specifičnih težina. Koristi se za reprezentaciju motiva u nekom biološkom nizu. Detaljnije, osnovnu PSSM matricu relativnih frekvencija od n poravnatih nizova iste duljine L računamo tako da za svaku poziciju u motivu izbrojimo i sumiramo koliko puta se određeno slovo pojavilo i podijelimo s ukupnim brojem motiva. Primjerice, za 4 niza duljine 4, ako se slovo A pojavilo jednom na prvoj poziciji u motivu, onda će element na mjestu $M_{1,A}$ biti jednak $\frac{1}{4}$. Također, pretpostavljamo nezavisnost između pozicija u nizovima.
Da bismo parametrizirali model, nije nam dovoljna samo osnovna PSSM matrica. Uz stručno znanje biologa i matematičke alate parametriziran je model. Početni seed i parametrizacija preuzeta je iz članka [5] i ovdje je ukratko opisana.

3.3 Početni seed i emisije

Početni seed, tj. osnovni skup podataka za parametrizaciju čine 23 eksperimentalno utvrđena GDSL enzima. Između njih nema praznina tako da je poravnanje zapravo blok. Neka je $F_k^j = [f_k^j]$ vektor relativnih frekvencija aminokiseline u j-tom stupcu poravnanja motiva, $j \in \{1, ..., n\}$, $k \in \{1, ..., 20\}$, gdje je n duljina motiva. Da bismo ispravili moguć nedostatak nezavisnosti uzorka, računajući F koristimo blagu težinsku shemu opisanu u [4] i [5], a problem malog uzorka rješavamo tako da relativnim frekvencijama dodajemo pseudo-zbroj 10^{-2}. Točnije, dobivamo vektore

$$\hat{f}_k^j = \frac{f_k^j + 0.01}{1.2}.$$

Neka je $A = a_{ij}$, oznaka za PAM(eng. *point accepted mutation*) matricu. A je stohastička matrica, tj. $\sum_{j=1}^{20} a_{ij} = 1$. Neka je $B = b_{ij} = A^k$, gdje je k velik (računato je za $k = 120$). Napomenimo da redak $b_i = (b_{i1}, b_{i2}, ..., b_{i20})$ predstavlja očekivan mutacijski uzorak za i-tu aminokiselinu nakon k milijuna godina evolucije. U konačnosti, emisijska vjerojatnost aminokiseline a u j-tom stupcu motiva je

$$e_j(a) = \sum_{l=1}^{20} b_{la} \cdot \hat{f}_l^j.$$

3.4 Score

Kako bismo odredili koji nam je niz zanimljiviji, tj. potencijalno sadrži određeni enzim, potreban nam je način dodjeljivanja određenog scora tome nizu u usporedbi sa zadanim motivom. Matrica opisana u prethodnom poglavlju zapravo predstavlja profil zadanog motiva i služi nam za "evaluaciju prozora", tj. računanje scora niza. Neka je $x = x_1, x_2, ..., x_n$ niz i $y = y_1, y_2, ..., y_L$ vjerojatnost pojavljivanja određene aminokiseline na mjestu i u motivu. Neka je $q = (q_1, q_2, ..., q_n)$ vektor distribucija aminokiselina, često zvana "bio-background". Evaluacija prozora s početnim indeksom l duljine jednake duljini zadanog motiva je

$$s_l(m) = \sum_{i=0}^{L-1} log\Big(\frac{\mathbb{P}(x_{l+i}|y_{i+1})}{\mathbb{P}(x_{l+i}|q)}\Big). \tag{3.1}$$

Neka a predstavlja jednu aminokiselinu. Tada je $\mathbb{P}(x_{l+i}|y_{i+1})$ jednaka upravo $e_i(a)$ iz već spomenute matrice, a $\mathbb{P}(x_{l+i}|q)$ jednaka $q(a)$.

3.5 Metode traženja motiva

U radu ćemo opisati 2 načina pomoću kojih određujemo mjesta na kojima računamo scorove, te ih kasnije usporediti.

Sliding window

Neka je $x = x_1, ..., x_n$ niz i $y = y_1, ..., y_L$ motiv, točnije naš upit okarakteriziran gore spomenutom matricom. PSSM algoritam implementiramo kao metodu klizećeg prozora.
Na taj način evaluiramo svaki "prozor" duljine L duž cijelog niza formulom (3.1). Očito je da za svaki niz dobivamo $n - L + 1$ scorova, pa zato u svakom nizu uzimamo

njegov maksimalni

$$S_i(m) = \max_{j \in 0,1,...n-L+1} s_j(m),$$

gdje s i označavamo protein u proteomu, tj. niz u matrici.
To možemo prikazati ovako:

$x_1 \quad x_2 \quad x_3 \quad ... \quad x_L \quad ... \quad x_{N1}$
$y_1 \quad y_2 \quad y_3 \quad ... \quad y_L$

$x_1 \quad x_2 \quad x_3 \quad ... \quad x_L \quad x_{L+1} \quad ... \quad x_{N1}$
$\quad\quad y_1 \quad y_2 \quad y_3 \quad ... \quad y_L$

$\quad\quad\quad\quad\quad\quad .$
$\quad\quad\quad\quad\quad\quad .$
$\quad\quad\quad\quad\quad\quad .$

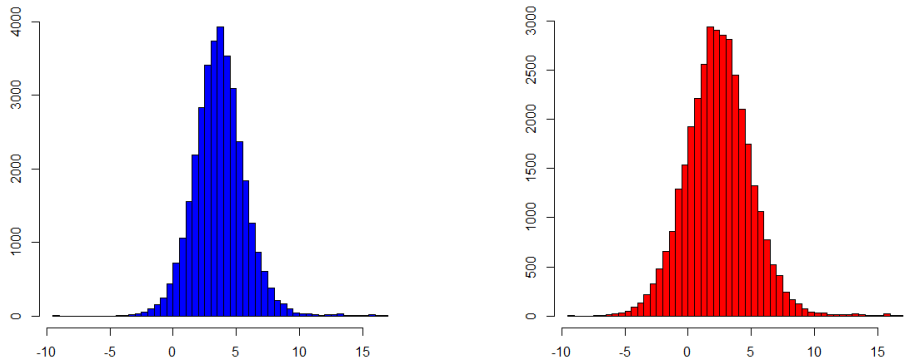
$x_1 \quad x_2 \quad x_3 \quad ... \quad x_L \quad ... \quad x_{N1-L} \quad x_{N1-L+1} \quad ... \quad x_{N1}$
$\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad y_1 \quad ... \quad ... \quad y_L$

String matching

Drugi način traženja motiva je tako da prvo izračunamo broj mjesta poklapanja znakova aminokiselina. Prolazimo kroz proteom kao da radimo sliding window, ali ovoga puta bilježimo mjesta maksimalnih poklapanja slova. Samo na mjestima maksimalnih poklapanja u nizu računamo scorove kao što je gore navedeno.
Sve to radimo kako bismo ubrzali proces traženja. U prvoj iteraciji jednak broj puta prolazimo kroz proteom. Već se tu događa razlika u brzini jer u prvoj metodi za svaki prozor računamo score, što je znatno sporije od samog uspoređivanja slova. U idućim iteracijama u prvoj metodi score računamo opet iznova za svaki prozor, $k \times (n-L+1)$ puta, što prosječno po nizu iznosi preko 300 puta. U drugoj metodi, u drugoj i svim narednim iteracijama, računamo vrijednost prozora na svega nekoliko pozicija, što je u prosjeku 100 puta manje za svaki niz.

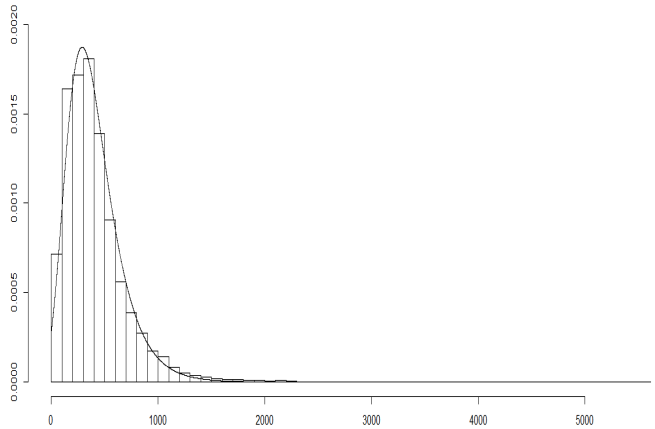
3.6 Distribucija scorova

Budući da za svaki niz pamtimo maksimalni score, prema teoriji opisanoj u 2. poglavlju za očekivati je da će scorovi pratiti neku od navedenih distribucija ekstremnih vrijednosti. Pogledamo li histogram scorova u obje metode odmah je jasno da ovdje to nije slučaj.



Slika 3.1: Histogrami scorova prve i druge metode

Iz nekog razloga scorovi prate logističku distribuciju.
Naime, proteom se sastoji od proteina različitih duljina i zbog toga ne možemo pretpostaviti da su jednako distribuirani. Zbog toga je prirodno pretpostaviti da je distribucija scorova ovisna o duljini niza. Simulacijom se može prikazati distribucija scorova za nizove jednake duljine. U tom slučaju scorovi doista jesu Gumbel distribuirani. Više o tome u [6]. Pogledamo li histogram duljina nizova, mogli bismo naslutiti da slijede možda eksponencijalnu ili čak Gumbelovu distribuciju.



Slika 3.2: Histogram duljine nizova i funkcija gustoće Gumbelove distribucije

Promatramo li duljine kao Gumbel distribuiranu slučajnu varijablu, prema korolaru 2.1.1. možemo barem intuitivno opravdati pojavu logističke distribucije koju koristimo za modeliranje.
Nakon što imamo scorove, u obje metode na isti način dalje postupamo. Metodom maksimalne vjerodostojnosti procjenjujemo parametre logističke distribucije.

3.7 Prag

Sada se nameće pitanje što uzeti za prag, tj. koje scorove ćemo proglasiti dovoljno dobrima. Ovdje je također potrebna suradnja s biolozima. Uzimati samo neki određeni percentil je teško jer u pravilu pozitivno rangirani nizovi čine tek oko $0,2 - 0,6\%$ proteina u proteomu. Primjerice, da su scorovi normalno distribuirani, uzeli bismo one scorove koji su 2 ili možda čak 3 standardne devijacije udaljeni od prosjeka.
Međutim, i za logističku distribuciju postoji nešto slično. Naime, parametar oblika β možemo izraziti u terminima standardne devijacije i obratno. Koristimo supstituciju $\sigma = \frac{\pi}{\sqrt{3}}\beta$. Korišten je kriterij tako da se za pozitivce proglase oni nizovi čiji je score veći $\mu + 8\beta$, što je otprilike malo više od 4 standardne devijacije.

Napomena 3.7.1. *Prag određuje koje nizove smatramo "pozitivcima", a koje "negativcima". U radu se ne spominju mjere osjetljivosti i specifičnosti, odnosno prava pozitivna vrijednost i prava negativna vrijednost. Ne analizira se učinkovitost metode iz jednostavnog razloga što nemamo pravu listu pravih pozitivaca pa nemamo s čime usporediti rezultat. Preciznije, TP (true positives) i FP (false positives) zajedno nazivamo pozitivcima, a TN (true negatives) i FN (false negatives) negativcima.*

3.8 Iteracija

Nakon što smo odabrali privremene pozitivce, pogledamo kojim motivima pripadaju odabrani scorovi. Novodobivenu listu motiva koristimo za novi seed. Ubacujemo je u gore opisani model, umjesto početnog, s time da početna matrica nosi polovicu težine novodobivene. Sada se prirodno nameće pitanje do kada ćemo iterirati cijeli model. Jednostavno, dok god postoji promjena u listi pozitivnih motiva iteriramo, te stajemo kada se prestane mijenjati.

3.9 Komentari

Primijetimo da u prvoj metodi svaki puta ponovno prolazimo kroz proteom, a u drugoj računamo scorove uvijek na istim mjestima. To znači da se motivi za koje računamo score ne mijenjaju već samo može doći do promjene njegovog rednog broja na listi.
Mogli bismo postaviti mnogo pitanja u svakom koraku. Na samom početku, koji motiv odabrati? Što ako ih uzmemo više, što ako krenemo s nekim "lošim"? Također, možemo se pitati jesmo li na dobar način odabrali prag. Neke varijante su isprobane u izradi rada, ali nisu značajnije utjecale na rezultat.

Poglavlje 4

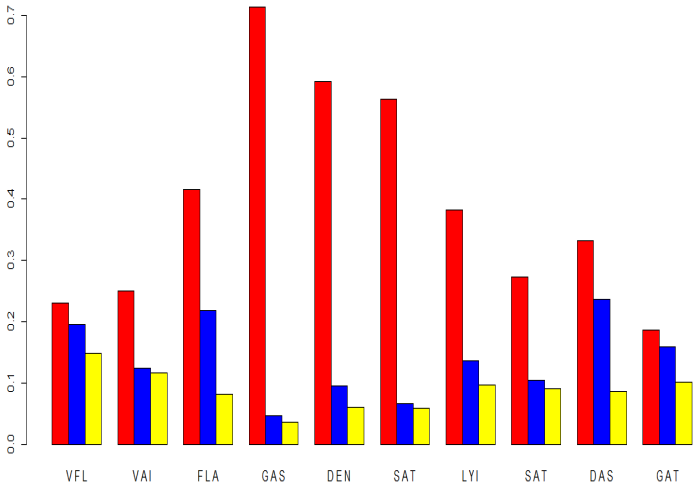
Proteomi biljaka

U ovome poglavlju ćemo proći kroz gore navedeni proces na primjerima. Prvi primjer opisujemo detaljno, dok za preostala 3 samo navodimo rezultate.

4.1 Arabidopsis thaliana

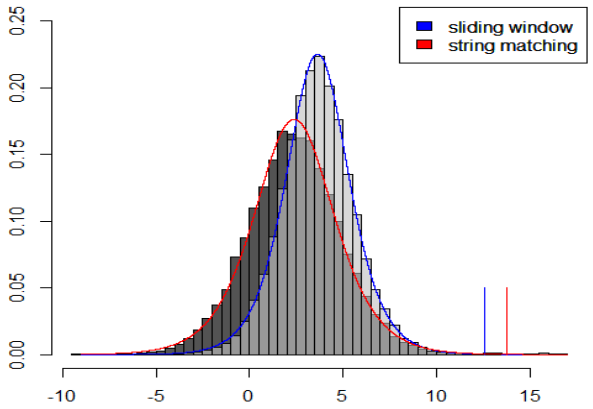
Arabidopsis thaliana mala je biljka s cvjetovima podrijetlom iz Euroazije. Jednogodišnja je biljka, pa je zbog svog kratkog životnog ciklusa pogodna za modeliranje u botanici i genetici. To je bila prva biljka za koju su laboratorijskim postupkom odredili sve nizove aminokiselina genoma u jednom trenutku. Njen proteom sastoji se od 35176 nizova proteina.
Kao što smo već spomenuli, sliding window metodom računamo score za svaki prozor, dok za string matching metodu samo za određene. Točnije, za proteom ove biljke, u samo jednoj iteraciji prvom metodom score računamo 70571650, a drugom nakon utvrđivanja sličnosti poklapanja "favoriziranim crvenim" motivom 'VVFGDSLSDG' 119598 puta. Početna matrica kojom pretražujemo "favorizira" određena slova na svakom mjestu u motivu. Radi jednostavnijeg prikaza barplotom su prikazana samo prve 3 maksimalne vjerojatnosti za svaku poziciju u motivu.

Slikom je djelomično prikazan profil motiva.



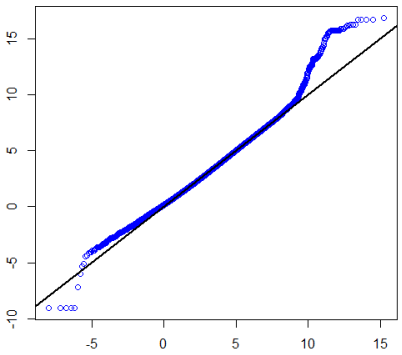
Slika 4.1: Vjerojatnosti aminokiselina u nizu

Histogramima je pridružena funkcija gustoće logističke razdiobe s parametrima procijenjenima iz uzorka.

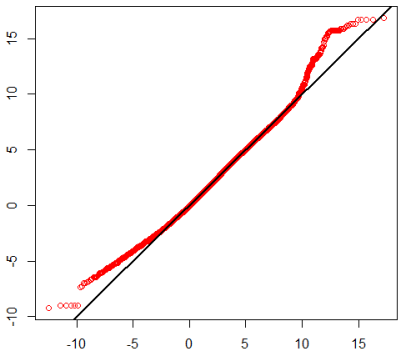


Slika 4.2: Histogrami, funkcije gustoće i pragovi u prvoj iteraciji

Usporedimo li kvantile logističke distribucije i dane podatke dobivamo sljedeće grafove.

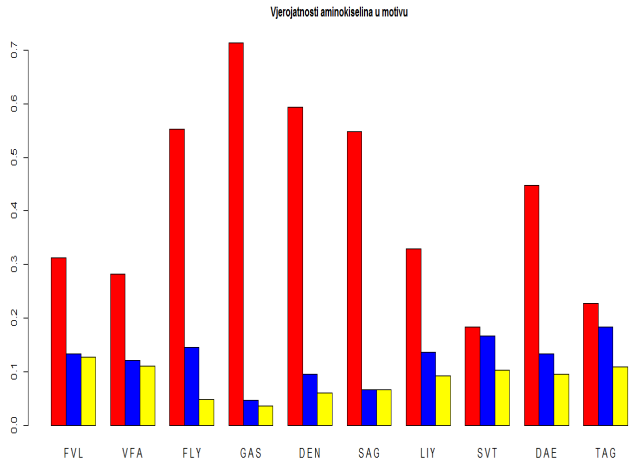


Slika 4.3: Q-Q graf

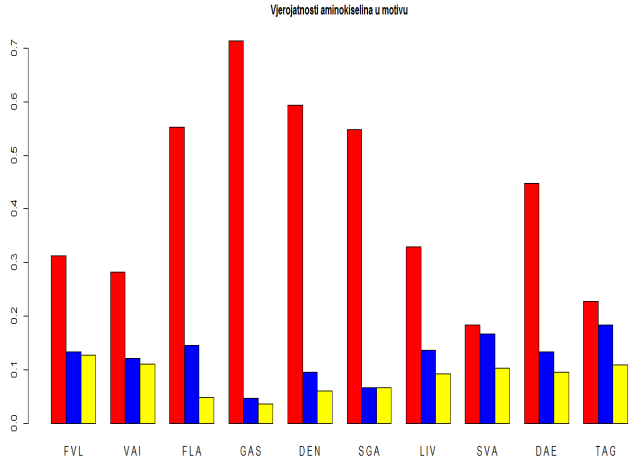


Slika 4.4: Q-Q graf

Vidimo da se u oba slučaja kvantili za veliku većinu podataka dobro grupiraju oko pravca $y = \hat{\mu} + \hat{\sigma}x$, pa možemo reći da maksimalni score-ovi slijede logističku distribuciju.
Iste grafove ponavljamo za zadnju iteraciju.

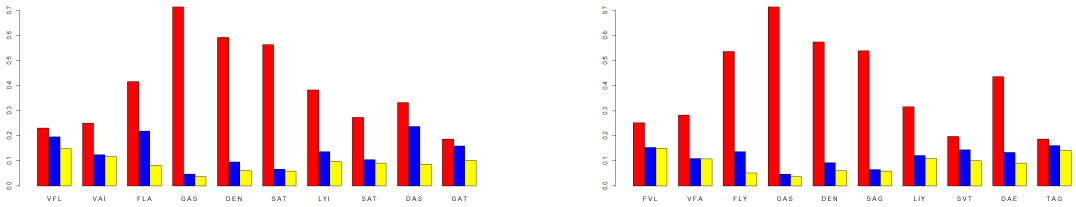


Slika 4.5: Profil motiva u zadnjoj iteraciji sliding window metode

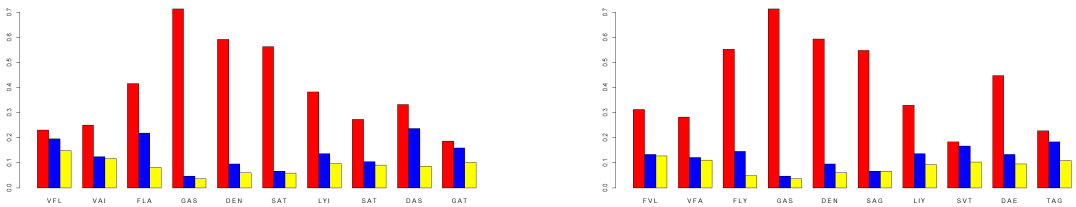


Slika 4.6: Profil motiva u zadnjoj iteraciji string matching metode

Mogli bismo reći kako nema nikakve razlike, međutim nije tako. Slične su vrijednosti, ali za različita slova. Konkretno, kada bismo gledali koji su motivi proglašeni pozitivcima u sliding window metodi bi ih bilo 63, a sa string matchingom 55. Na idućim slikama se lakše uočava razlika između modela u prvoj i zadnjim iteracijama.

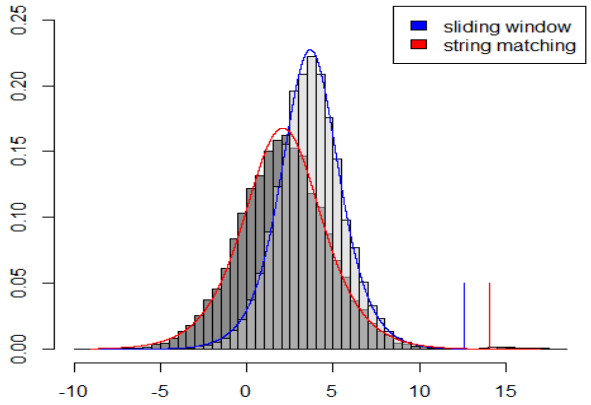


Slika 4.7: Usporedba prve i zadnje iteracije sliding window metode



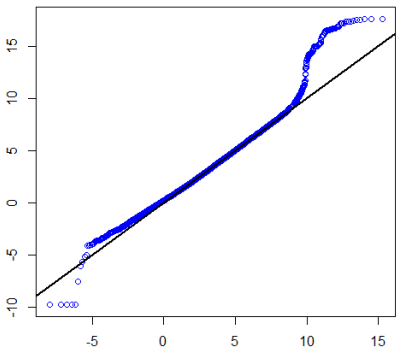
Slika 4.8: Usporedba prve i zadnje iteracije string matching metode

Sliding-window metodom je bilo potrebno 5, a string matchingom 7 iteracija.

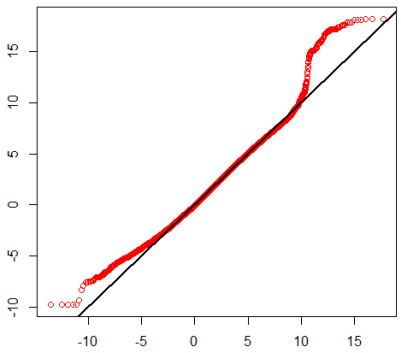


Slika 4.9: Histogrami, funkcije gustoće i pragovi u prvoj iteraciji

I ovdje, q-q grafom potvrđujemo da scorovi slijede logističku distribuciju.

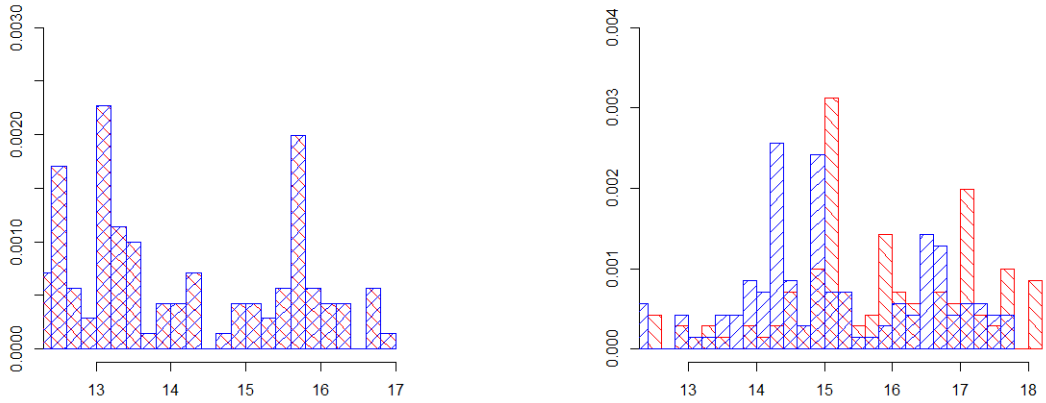


Slika 4.10: Q-Q graf



Slika 4.11: Q-Q graf

Naizgled se i ovdje može činiti da nema gotovo nikakve razlike u grafovima. Na uvećanim grafovima se razlika lakše uočava.



Slika 4.12: Desni rep u prvoj i zadnjim iteracijama

Na prvoj slici nema nikakve razlike, ali to je logično jer ulazimo s istom matricom u model. Malo preciznije, zapravo smo ovim kriterijem sliding-windowom sveukupno proglasili 122 pozitivca, a string matchingom 108. Svih 108 "pozitivaca crvene metode" su i "pozitivci plave metode".

Bitno je obratiti pozornost na izvršavanje programa u sekundama. Sliding window metoda je gotovo 20 puta sporija zbog ogromne količine operacija koje se izvršavaju.

Slični rezultati su dobiveni i u proteomima drugih biljaka.

4.2 Oryza Sativa

Oryza sativa je žitarica poznata kao azijska riža ili samo riža. Jednogodišnja je biljka, naraste do 1.8 m u visinu. Lako ju je genetski modificirati. Njen proteom sadrži 35629 proteina.

U slučaju ove biljke u obje metode je bilo potrebno samo 3 iteracije da bi se zaustavila. Sliding window metodom je proglašeno 149 pozitivaca, a string matching 140. Svih 140 pozitivaca prve metode su i pozitivci plave metode.

4.3 Populus trichocarpa

Populus trichocarpa tj. kalifornijska topola je širokolisno listopadno stablo iz Sjeverne Amerike. Može narasti 30-50 m u visinu. Njen puni genom objavljen je 2006. godine. To je prva vrsta stabla čiji je genom sekvenciran. Drvo se koristi u građevini, ali također zbog vrlo brzog rasta i ranog sazrijevanja je popularna među biolozima i genetičarima. Njen proteom se sastoji od 42345 proteina.

U njenom slučaju je plavoj metodi bilo potrebno 4 iteracije da bi pozitivcima proglasila 109 nizova, a crvenoj 5 koja je za pozitivce detektirala 96 nizova. I u ovom slučaju pozitivci se podudaraju.

4.4 Sorghum

Sorghum je rod brojnih biljnih vrsta iz porodice trava. Neke porodice podrijetlom su iz Australije, a neke iz Afrike. Neke od njih se uzgajaju kao žitarice, stočna hrana i za proizvodnju sirupa i alkoholnih pića. Zanimljivo je da ne sadrži gluten. Uglavnom se uzgaja radi žita u toplim klimama diljem svijeta. Proteom sadrži 32889 nizova proteina.

Plavoj metodi bilo je potrebno 3 iteracije da za pozitivce proglasi 104 nizova, a crvenoj 4, koja je za pozitivce detektirala 95 nizova. 92 pozitivca koje string matching metoda proglasi su i pozitivci sliding window metode.

Bibliografija

[1] Nikola Sarapa, *Teorija vjerojatnosti*, Školska knjiga, 2002.

[2] Sidney I Resnick, *Extreme values, regular variation, and point processes*, Springer, 2007.

[3] M.O.Ojo, *Some relationships between generalized Gumbel and other distributions*, Kragujevac J. Math, 23:101, 2001.

[4] I.M. Wallace, D.G.Higgins, *Supervised multivariate analysis of sequence groups to identify specificity determining residues*, BCM Bioinformatics, 8:135, 2007.

[5] Steven Henikoff, Jorja G. Henikoff *Position-based sequence weights*, Journal of Molecular Biology, 243:574, 1994.

[6] Silvija Vrbančić, *Lokalno poravnanje i prepoznavanje motiva*, Diplomski rad, Sveučilište u Zagrebu, Prirodoslovno-matematički fakultet, Matematički odsjek, 2014.

Sažetak

Nakon kratkog pregleda osnovnih pojmova teorije vjerojatnosti, upoznajemo se sa distribucijama ekstremnih vrijednosti. Opisujemo dvije metode pomoću kojih tražimo karakteristične članove GDSL familije proteina. String matching metoda je uspoređena sa sliding window metodom i opisan je model pomoću kojega određujemo pozitivno rangirane nizove. Objašnjeno je zašto pritom maksimalni scorovi prate logističku distribuciju. Obje metode daju vrlo slične rezultate na četiri biljna proteoma što znači da string matching metoda ima potencijal za korištenje u pretraživanju proteoma zadanim motivom.

Summary

After a brief introduction to elementary probability theory we introduce certain aspects of the theory of extreme value distribution. We describe and compare two algorithms for searching for strings in large datasets - sliding window and string matching algorithm. Using results from extreme value theory and either of the algorithms, we construct an iterative string searching method. The results of this procedure are compared between the algorithms as well as against expert knowledge, on four plant proteomes.

Životopis

Rođena sam 09.05.1991. godine u Kutini. Školovanje sam započela u rodnom gradu 1998. godine u osnovnoj školi "Stjepan Kefelja". Nakon toga pohađala sam prirodoslovno - matematičku gimnaziju "Tin Ujević", također u Kutini. 2010. godine upisujem preddiplomski studij matematike na Prirodoslovno - matematičkom fakultetu u Zagrebu. Završetkom preddiplomskog studija 2013. godine upisala sam diplomski studij Matematička statistika na istom fakultetu.