# Analiza točnosti iterativnog pretraživanja

**Mihovilčević, Marijana**

**Master's thesis / Diplomski rad**

**2017**

*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:217:769285

*Rights / Prava:* In copyright / Zaštićeno autorskim pravom.

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

*Repository / Repozitorij:*

Repository of the Faculty of Science - University of Zagreb

zir.nsk.hr

dabar
DIGITALNI AKADEMSKI ARHIVI I REPOZITORIJI

**SVEUČILIŠTE U ZAGREBU**

**PRIRODOSLOVNO–MATEMATIČKI FAKULTET**

**MATEMATIČKI ODSJEK**

Marijana Mihovilčević

# ANALIZA TOČNOSTI ITERATIVNOG PRETRAŽIVANJA

Diplomski rad

Voditelj rada:
doc. dr. sc. Pavle Goldstein

Zagreb, 2017.

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

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

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

*Zahvaljujem mentoru doc. dr. sc. Pavlu Goldsteinu na strpljenju i posvećenom vremenu tijekom izrade ovog rada. Najveće hvala Bogu, mojoj obitelji i prijateljima na bezuvjetnoj podršci i ljubavi koju su mi pružili tijekom studiranja.*

# Sadržaj

**Sadržaj** **iv**

**Uvod** **1**

**1 Osnovni pojmovi** **2**
1.1 Vjerojatnost . . . . . 2
1.2 Distribucije ekstremnih vrijednosti . . . . . 6
1.3 Osjetljivost i specifičnost testa . . . . . 8

**2 Traženje motiva** **9**
2.1 Osnovni biološki pojmovi . . . . . 9
2.2 PSSM algoritam . . . . . 11
2.3 Sliding window . . . . . 11
2.4 Značajnost ocjene . . . . . 12
2.5 Iterativni dio . . . . . 14

**3 Analiza točnosti** **15**
3.1 Entropija . . . . . 15
3.2 Varijanca . . . . . 21

**4 Arabidopsis thaliana** **23**

**Bibliografija** **26**

# Uvod

Ubrzanim razvojem tehnologije, dolazi do razvoja područja bioinformatike. Bioinformatika je interdisciplinarna znanost koja se bavi analizom, skladištenjem i organizacijom podataka dobivenih interpretacijom bioloških makromolekula. Ona primjenjuje postupke računarske znanosti, statistike i matematike za obradu podataka iz DNA, RNA i molekula proteina.

Jedna od važnih zadaća u bioinformatici je odgovoriti na pitanje o porijeklu proteina, to jest njihovim precima. Tragove zajedničkih korijena tražimo na molekularnom nivou, odnosno proučavajući promjene u nizovima proteoma koje su tijekom vremena nastale mutacijom. To radimo na način da iterativno pretražujemo bazu podataka bioloških nizova u potrazi za sličnim nizovima, to jest onim nizovima sa zajedničkim pretkom. Preciznije, za dani proteom želimo pronaći nizove s najboljom ocjenom poravnanja te ih spremamo u listu pozitivaca.

Kako bismo detaljnije mogli reći da li je neki niz u srodnosti s drugim, u prvom poglavlju uvodimo pojam vjerojatnosti. Razvoj i opis modela kojeg koristimo za pretraživanje detaljno je iznesen u drugom poglavlju. Da bismo ispitali sadrži li dobivena lista pozitivaca biološki smislene nizove ili samo slučajno povezane nizove, koristit ćemo se nekim od statističkih mjera raspršenosti kao što su entropija i varijanca. Postupci i rezultati opisani su u trećem poglavlju. Također, cilj nam je odrediti prag, odnosno granicu "točnosti" od koje ćemo nadalje dobiti najznačajnije nizove. U zadnjem poglavlju ćemo usporediti dobivene rezultate sa GDSL enzimima iz proteoma biljke *Arabidopsis thaliana*.

# Poglavlje 1

# Osnovni pojmovi

## 1.1 Vjerojatnost

Neka je $\Omega$ prostor elementarnih događaja tj. $\Omega$ je proizvoljan neprazan skup.

**Definicija 1.1.1.** *Familija $\mathcal{F}$ podskupova od $\Omega$ ($\mathcal{F} \subset \mathcal{P}(\Omega)$) jest $\sigma$-**algebra skupova** (na $\Omega$) ako je:*
$(F1)$ $\emptyset \in \mathcal{F}$
$(F2)$ $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$
$(F3)$ $A_i \in \mathcal{F}, i \in \mathbb{N} \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$

**Definicija 1.1.2.** *Neka je $\mathcal{F}$ $\sigma$-algebra na skupu $\Omega$. Uređen par $(\Omega, \mathcal{F})$ zove se **izmjeriv prostor**.*

**Definicija 1.1.3.** *Neka je $(\Omega, \mathcal{F})$ izmjeriv prostor. Funkcija $\mathbb{P} : \mathcal{F} \to \mathbb{R}$ jest **vjerojatnost** (na $\mathcal{F}$, na $\Omega$) ako vrijedi:*
$(P1)$ $\mathbb{P}(A) \geqslant 0$, $A \in \mathcal{F}$; $\quad \mathbb{P}(\Omega) = 1$
$(P2)$ $A_i \in \mathcal{F}, i \in \mathbb{N}$ *te* $A_i \bigcap A_j = \emptyset$ *za* $i \neq j \Rightarrow \mathbb{P}(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$

**Definicija 1.1.4.** *Uređena trojka $(\Omega, \mathcal{F}, \mathbb{P})$ gdje je $\mathcal{F}$ $\sigma$-algebra na $\Omega$ i $\mathbb{P}$ vjerojatnost na $\mathcal{F}$ zove se **vjerojatnosni prostor**.*

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

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

**Definicija 1.1.5.** *Neka je* $(\Omega, \mathcal{F}, \mathbb{P})$ *proizvoljan vjerojatnosni prostor i* $A \in \mathcal{F}$ *takav da je* $\mathbb{P}(A) > 0$. *Definiramo funkciju* $\mathbb{P} : \mathcal{F} \to [0,1]$ *ovako:*

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

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

## Funkcija distribucije

**Definicija 1.1.6.** *Neka je* $(\Omega, \mathcal{F}, \mathbb{P})$ *vjerojatnosni prostor. Funkcija* $X \colon \Omega \to \mathbb{R}$ *jest* ***slučajna varijabla*** *(na* $\Omega$*) ako je* $X^{-1}(B) \in \mathcal{F}$ *za proizvoljno* $B \in \mathcal{B}$*, tj.* $X^{-1}(\mathcal{B}) \subset \mathcal{F}$.

**Definicija 1.1.7.** *Neka su* $X_1, X_2, \ldots, X_n$ *slučajne varijable na vjerojatnosnom prostoru* $(\Omega, \mathcal{F}, \mathbb{P})$. *Kažemo da su* $X_1, X_2, \ldots, X_n$ *nezavisne ako za proizvoljne* $B_i \in \mathcal{B}\,(i = 1, 2, \ldots, n)$ *vrijedi*

$$\mathbb{P}\{X_1 \in B_1, \ldots, X_n \in B_n\} = \mathbb{P}\Big(\bigcap_{i=1}^{n} \{X_i \in B_i\}\Big) = \prod_{i=1}^{n} \mathbb{P}\{X_i \in B_i\}. \tag{1.1}$$

**Definicija 1.1.8.** *Neka je X slučajna varijabla na* $\Omega$. ***Funkcija distribucije*** *od X jest funkcija* $F_X : \mathbb{R} \to [0,1]$ *definirana sa*

$$F_X(x) = \mathbb{P}\{\,\omega \in \Omega;\, X(\omega) \leqslant x\,\} = \mathbb{P}\,\{X \leqslant x\}, \quad x \in \mathbb{R}.$$

Postoje dva glavna tipa slučajnih varijabli: diskretne i neprekidne.

**Definicija 1.1.9.** *Slučajna varijabla X je* ***diskretna*** *ako postoji konačan ili prebrojiv skup* $D \subset \mathbb{R}$ *takav da je* $\mathbb{P}\{X \in D\} = 1$.

**Definicija 1.1.10.** *Slučajna varijabla X je* ***apsolutno neprekidna*** *slučajna varijabla 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 $X$.

## Matematičko očekivanje i varijanca

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

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

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

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

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

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

**Definicija 1.1.11.** *Neka $\mathbb{E}X$ postoji (tj. konačno je). Tada $\mathbb{E}\left[(X - \mathbb{E}X)^r\right]$ zovemo r-ti centralni moment od X, a $\mathbb{E}\left[|X - \mathbb{E}X|^r\right]$ zovemo r-ti apsolutni centralni moment od X.*

**Definicija 1.1.12.** ***Varijanca*** *od X koju označavamo sa VarX ili ${\sigma_x}^2$ jest drugi centralni moment od X tj.*

$$VarX = \mathbb{E}\left[(X - \mathbb{E}X)^2\right]. \tag{1.3}$$

Pozitivan drugi korijen iz varijance zovemo **standarna devijacija** od $X$ i označavamo s $\sigma_X$.

## Logistička distribucija

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

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

Očekivanje slučajne varijable $X$ je $\mathbb{E}X = \mu$, a varijanca iznosi $VarX = \frac{\beta^2\pi^2}{3}$. Funkcija distribucije od $X$ je zadana s

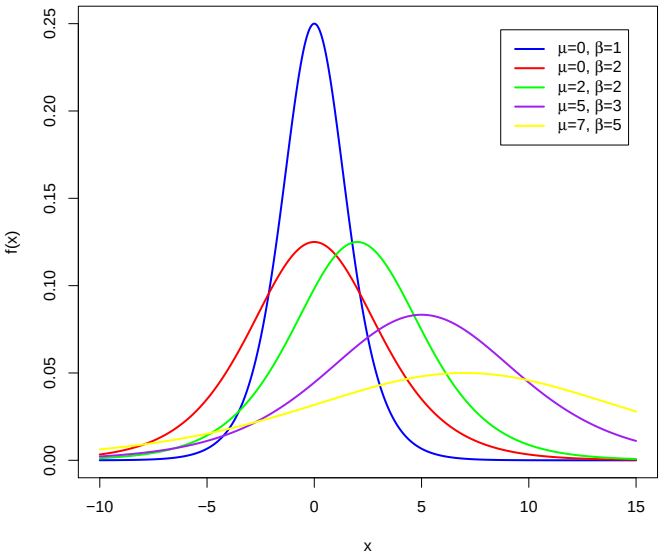
$$F(x) = \frac{e^{-\frac{x-\mu}{\beta}}}{1 + e^{-\frac{x-\mu}{\beta}}}.$$

Za $\mu = 0$ i $\beta = 1$ dobivamo standarnu logističku distribuciju s funkcijom distribucije

$$F(x) = \frac{e^{-x}}{1 + e^{-x}}, \quad x \in \mathbb{R}$$

i funkcijom gustoće

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




Slika 1.1: Funkcije gustoće logističke distribucije

## 1.2 Distribucije ekstremnih vrijednosti

Iz neke distribucije generiramo $n$ uzoraka. Ako za novi skup podataka uzmemo maksimume (minimume) generiranih uzoraka, taj skup podataka možemo prikazati nekom od sljedećih distribucija ekstremnih vrijednosti: Gumbelovom, Frechetovom ili Weibullovom distribucijom. Svi rezultati će biti iskazani za maksimum jer rezultate za minimum lako dobijemo sljedećom relacijom:

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

### Gumbelova distribucija

Gumbelova[1] funkcija distribucije u općem obliku je

$$F(x) = e^{-e^{-\frac{x-\mu}{\beta}}},\ \mu,\beta \in \mathbb{R}$$

gdje je $\mu$ lokacijski parametar, a $\beta > 0$ parametar mjere. Funkcija gustoće $f$ ima oblik

$$f(x) = \frac{1}{\beta} e^{-\frac{x-\mu}{\beta} - e^{-\frac{x-\mu}{\beta}}}.$$

Ako stavimo $\mu = 0$ i $\beta = 1$ dobivamo standarnu Gumbelovu funkciju distribucije.

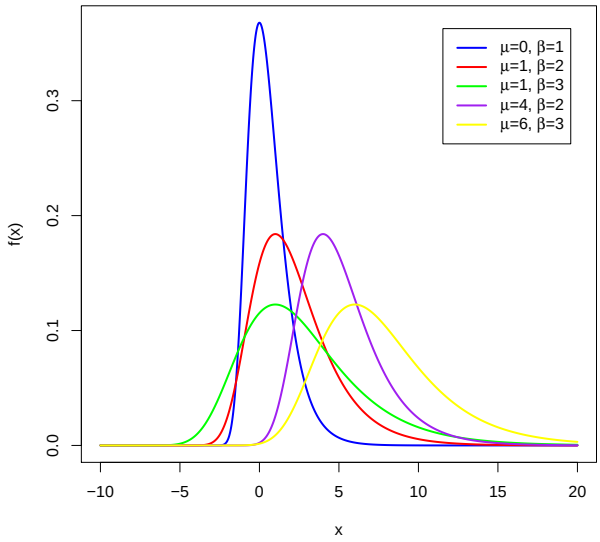
$$F(x) = e^{-e^{-x}},\ x \in \mathbb{R} \tag{1.4}$$

Funkcija gustoće tako definirane slučajne varijable je

$$f(x) = e^{-x-e^{-x}},\ x \in \mathbb{R}\,. \tag{1.5}$$

Neka je $X$ slučajna varijabla s Gumbelovom distribucijom. Očekivanje slučajne varijable $X$ je $\mathbb{E}X = \mu + \beta\gamma$, gdje je $\gamma \approx 0.5722$ Eulerova konstanta. Varijanca je jednaka $VarX = \frac{1}{6}\pi^2\beta^2$.

[1] Emil Gumbel (1891. - 1966.), njemački matematičar koji se bavio modeliranjem ekstremnih vrijednosti u području meteorologije.




Slika 1.2: Funkcije gustoće Gumbelove distribucije

**Korolar 1.2.1.** *Ako su $X$ i $Y$ nezavisne slučajne varijable sa standarnom Gumbelovom distribucijom, tada slučajna varijabla $Z = Y - X$ ima standarnu logističku distribuciju.*

*Dokaz.* $X$ i $Y$ imaju funkciju distribucije i funkciju gustoće prikazanu s (1.4) i (1.5). Neka je $z \in \mathbb{R}$.

$$\mathbb{P}(Z \leqslant z) = \mathbb{P}(Y \leqslant X + z) = \mathbb{E}[\mathbb{P}(Y \leqslant X + z \,|\, X)]$$
$$= \int_{-\infty}^{\infty} e^{-e^{-(x+z)}} e^{-x-e^{-x}} dx$$

Zamjenom $u = -e^{-(x+z)}$ dobivamo

$$\mathbb{P}(Z \leqslant z) = \int_{-\infty}^{0} e^{u} e^{e^{z} u} e^{z} du = e^{z} \int_{-\infty}^{0} e^{u(1+e^{z})} du$$
$$= \frac{e^{z}}{1 + e^{z}}$$

□

## 1.3 Osjetljivost i specifičnost testa

Kod analize uspješnosti metode koristimo sljedeće pojmove.

| | | predviđeno stanje | | |
|---|---|---|---|---|
| | ukupna populacija | predviđeno pozitivno stanje | predviđeno negativno stanje | |
| stvarno stanje | pozitivno stanje | pravi pozitivci (TP) | lažni negativci (FN) | osjetljivost |
| | negativno stanje | lažni pozitivci (FP) | pravi negativci (TN) | specifičnost |
| | | PPV | NPV | |

Tablica 1.1: Osjetljivost i specifičnost testa

Osjetljivost testa (stopa stvarno pozitivnih) mjeri proporciju pravih pozitivaca u odnosu na ukupan broj pozitivnih, dok specifičnost testa (stopa stvarno negativnih) mjeri proporciju pravih negativaca u odnosu na ukupan broj negativnih.

$$\text{osjetljivost} = \frac{TP}{TP + FN}$$

$$\text{specifičnost} = \frac{TN}{FP + TN}$$

Pozitivno predviđena vrijednost (PPV) pokazuje koliki je postotak pravih pozitivaca u odnosu na predviđeno pozitivno stanje, dok negativno predviđena vrijednost (NPV) otkriva postotak pravih negativaca u odnosa na predviđeno negativno stanje.

$$PPV = \frac{TP}{TP + FP}$$

$$NPV = \frac{TN}{TN + FN}$$

S obzirom na statističku prirodu testa, rezultati su u vrlo rijetkim slučajevima bez greške. Razlikujemo dvije vrste grešaka: greška I. vrste i greška II. vrste. Prvo zadajemo nultu hipotezu koja odgovara "stvarnom stanju". Greška I. vrste se pojavljuje kada je nulta hipoteza točna, ali je odbacujemo. Takve rezultate nazivamo "lažni pozitivci" (FP). Greška II. vrste je prisutna kada je nulta hipoteza netočna, ali je ne odbacujemo i takve rezultate nazivamo "lažni negativci" (FN).

# Poglavlje 2

# Traženje motiva

## 2.1 Osnovni biološki pojmovi

Proteini ili bjelančevine su makromolekule sastavljene od jednog ili više lanaca aminokiselina. Uz vodu, proteini su najvažnije tvari u tijelu. Ovisno o svojoj građi, izvode čitav niz različitih aktivnosti unutar organizma. Odgovorni su za ubrzanje metaboličkih reakcija, replikaciju DNA, transport molekula i brojne druge funkcije. Sastavni su dio svake stanice, što ih čini osnovom života na Zemlji.

Aminokiseline su prirodni spojevi koji u prirodi rijetko dolaze u slobodnom stanju. Uglavnom su međusobno povezane u makromolekule peptida i proteina. Postoji 20 standarnih aminokiselina.

| | |
|---|---|
| Alanin (A) | Arginin (R) |
| Asparagin (N) | Asparaginska kiselina (D) |
| Cistein (C) | Glutaminksa 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 2.1: Aminokiseline i njihove kratice

Proteom je skup svih proteina koje neki organizam proizvodi. Motiv proteinskog niza je kratak obrazac sastavljen od nekoliko aminokiselina, najčešće 5 do 20 aminokiselina, koji je ostao sačuvan selekcijskim pročišćavanjem i ima neko biološko značenje.

Kroz generacije, utjecajem različitih vanjskih faktora, motiv se mijenja. Taj proces nazivamo mutacijom. Na proteinskom nivou, mutacije se reflektiraju kao supstitucija (zamjena jedne aminokiseline drugom), insercija (umetanje aminokiseline) i delecija (brisanje aminokiseline).

Kako bi preciznije mogli ustanoviti sličnost dvaju nizova, uvodimo ocjenu poravnanja *(engl. score)* koja će svakom poravnanju dodijeliti realni broj S. Definiramo s $\mathcal{A}$ skup aminokiselina.

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

Neka su $X = x_1 x_2 \dots x_{n_1}$ i $Y = y_1 y_2 \dots y_{n_2}$ dva proteina, gdje su $x_i$ i $y_i$ aminokiseline iz skupa $\mathcal{A}, i = 1, 2, \dots, 20$. Pretpostavimo da je pojava svake aminokiseline u nizu nezavisna, odnosno vjerojatnost pojave jedne aminokiseline ne utječe na vjerojatnost pojavljivanja neke druge aminokiseline. Neka je R *(engl. random)* slučajni model u kojem su nizovi slučajno poravnati. S $q_{x_i}$ označavamo vjerojatnost da se $x_i$ pojavi u nizu $X$, a s $q_{y_i}$ vjerojatnost da se $y_i$ pojavi u nizu $Y$.

$$\mathbb{P}(X, Y \mid R) = \prod_i q_{x_i} \prod_i q_{y_i} \tag{2.1}$$

Neka je M *(engl. match)* model u kojem je vjerojatnost da se aminokiseline $x_i$ i $y_i$ pojave jednaka $p_{x_i y_i}$, za neki $i = 1, 2, \dots 20$. Vjerojatnost $p_{x_i y_i}$ možemo shvatiti kao vjerojatnost da te dvije aminokiseline imaju zajedničkog pretka. Tada je vjerojatnost poravnanja nizova $X$ i $Y$ jednaka:

$$\mathbb{P}(X, Y \mid M) = \prod_i p_{x_i y_i}\,. \tag{2.2}$$

Omjer formula (2.2) i (2.1) naziva se omjer šansi *(engl. odds ratio)*.

$$\frac{\mathbb{P}(X, Y \mid M)}{\mathbb{P}(X, Y \mid R)} = \frac{\prod_i p_{x_i y_i}}{\prod_i q_{x_i} \prod_i q_{y_i}} = \prod_i \frac{p_{x_i y_i}}{q_{x_i} q_{y_i}} \tag{2.3}$$

Radi lakšeg računanja gornji omjer logaritmiramo te dobivamo:

$$S = \sum_i s(x_i, y_i) \quad \text{gdje je} \quad s(x_i, y_i) = \log \frac{p_{x_i y_i}}{q_{x_i} q_{y_i}}. \tag{2.4}$$

Za model R uzimamo sljedeću distribuciju koja je dobivena računanjem relativnih frekvencija aminokiselina u proteomima nekog većeg skupa organizma.

$$\begin{aligned} q = (&0.078, 0.051, 0.043, 0.053, 0.019, 0.043, 0.063, 0.072, 0.023, 0.053,\\ &0.091, 0.059, 0.022, 0.039, 0.052, 0.068, 0.059, 0.014, 0.032, 0.066) \end{aligned}$$

## 2.2 PSSM algoritam

Model M gradimo na temelju ulaznog motiva. Parametri modela M dani su u matrici $P = (m_{ij})$, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, 20$ gdje je $n$ duljina ulaznog motiva. Promotrimo sljedeće nizove:

FGLSN
VGLSD
VSLYN
FGLSA

Računamo relativne frekvencije pojavljivanja aminokiselina. Na prvoj poziciji pojavljuju se aminokiseline F i V jednak broj puta, stoga će njihova relativna frekvencija biti $\frac{1}{2}$. Tada su elementi matrice $m_{1,14} = \frac{1}{2}$ i $m_{1,20} = \frac{1}{2}$, a ostali elementi su 0. Analogno, računajući ostale elemente dobivamo matricu P koju zovemo PSSM matrica *(engl. position specific scoring matrix)*.

Kako bi izbjegli da neki od elemenata $m_{ij}$ bude jednak 0, dodajemo pseudo zbroj,

$$f_{ij} = \frac{m_{ij} + 0.01}{1.2} \quad i = 1, 2, \ldots, n,\ j = 1, 2, \ldots, 20.$$

Dobivamo vektor $f_i = (f_{i1}, f_{i2}, \ldots, f_{i20})$ relativnih frekvencija aminokiseline u $j$-tom stupcu poravnanja motiva. PAM *(engl. point accepted mutation)* matricu označavamo s $A = (a_{ij})$. A je stohastička matrica, tj. vrijedi

$$\sum_{j=1}^{20} a_{ij} = 1.$$

U matrica $A$, $i$-ti redak daje vjerojatnosti da $i$-ta aminokiselina evolucijom pređe u neku drugu aminokiselinu ili u samu sebe. $B = (b_{ij}) = A^k$ je matrica koja opisuje očekivanu distribuciju nakon k milijuna godina.

## 2.3 Sliding window

Metoda klizećeg prozora *(engl. sliding window)* je metoda pomoću koje tražimo nizove najsličnije zadanom motivu. Rezultat ocjene podudaranja s modelom je ta vjerojatnost u odnosu na slučajni model R koji je određen vektorom $q$ tj. sljedeći log-odds ratio

$$\log \frac{\mathbb{P}(x|M)}{\mathbb{P}(x|R)}.$$

Neka je $x = x_1, x_2, \ldots, x_n$ niz i neka su $Y_1, Y_2, \ldots, Y_L$ slučajne varijable koje određuju zadani motiv. Koristeći (2.4) dobivamo

$$s_k = \sum_{i=0}^{L-1} log \frac{\mathbb{P}(x_{k+i}|Y_{i+1})}{\mathbb{P}(x_{k+i}|q)}, \quad k = 1, 2, \ldots, n - L + 1 \tag{2.5}$$

gdje s $\mathbb{P}(x_{k+i}|Y_{i+1})$ označavamo vjerojatnost poravnanja niza $x_k, x_{k+1}, \ldots, x_{k+L-1}$ s nizom koji je opisan modelom M. Kada niz dobro odgovara našem modelu M, ocjena $s_k$ je visoka, a kada odgovara slučajnom modelu R, ocjena je negativna. Evaluiramo svaki "prozor" duljine $L$ duž cijelog niza i na svakoj poziciji $k$ računamo ocjenu poravnanja koristeći (2.5). Dobivamo ukupno n-L+1 ocjena. U svakom nizu uzimamo maksimalnu ocjenu

$$s_i = \max_{j=0,1,\ldots,n-L+1} s_j \,,$$

gdje $i$ označava protein, odnosno redak proteoma. Shemu možemo ilustrativno prikazati na sljedeći način.

$$\begin{array}{cccccccc}
x_1 & x_2 & \ldots & x_L & x_{L+1} & \ldots & x_n & \\
Y_1 & Y_2 & \ldots & Y_L & & & &
\end{array}$$

$$\begin{array}{cccccccc}
x_1 & x_2 & \ldots & x_L & x_{L+1} & \ldots & x_n & \\
 & Y_1 & Y_2 & \ldots & Y_L & & &
\end{array}$$

$$\vdots$$

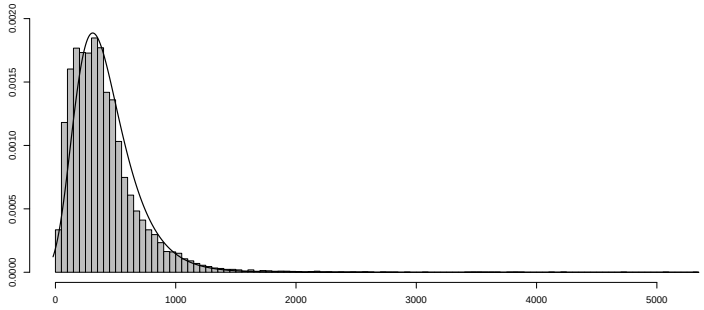
$$\begin{array}{cccccccc}
x_1 & x_2 & \ldots & x_L & \ldots & x_{n-L+1} & \ldots & x_n \\
 & & & & & Y_1 & Y_2 \ldots & Y_L
\end{array}$$

## 2.4 Značajnost ocjene

Kao što smo rekli, sličnost nizova izražena je ocjenom poravnanja. Ali, kako provjeriti jesu li maksimalne ocjene koje smo dobili statistički značajne?

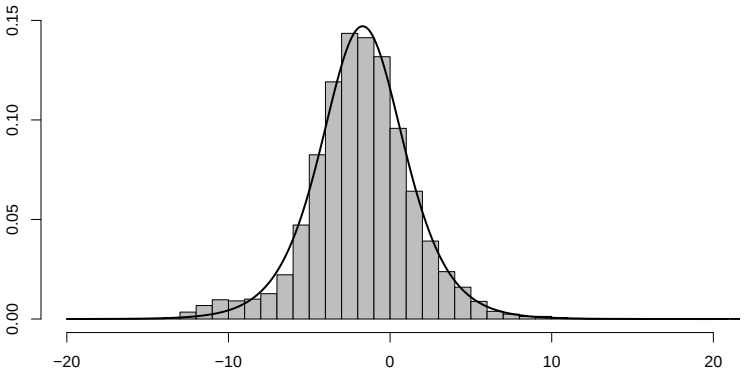
Za početak, potrebno je odrediti kako su distribuirane maksimalne ocjene svakog niza. Uočavamo da distribucija ocjena po svim pozicijima pojedinog niza proteoma ima eksponencijalni rep. Budući da nas zanimaju maksimumi ocjena svakog niza proteoma, dovoljno je da samo rep distribucije bude eksponencijalan jer se maksimum nalazi u repu.

Ako pretpostavimo da su svi nizovi jednake duljine, tada iz teorije ekstremnih vrijednosti slijedi da maksimalne ocjene imaju Gumbelovu distribuciju. Međutim, nizovi koje promatramo nisu jednake duljine, pa ne možemo tako nešto zaključiti.




Slika 2.1: Histogram duljina nizova i funkcija gustoće Gumbel distribucije

Iz histograma duljina svih nizova naslućujemo da bi mogli pratiti Gumbelovu distribuciju. Dakle, imamo dvije Gumbel distribuirane slučajne varijable, pa iz korolara (1.2.1) slijedi da razlika te dvije varijable ima logističku distribuciju.




Slika 2.2: Histogram maksimalnih ocjena poravnanja i funkcija gustoće logističke distribucije

Preostaje odrediti prag od kojeg nadalje svaki dobiveni podniz smatramo pogotkom. Uočimo da parametar $\beta$ logističke distribucije možemo izraziti preko standarne devijacije.

$$\beta = \frac{\sqrt{3}}{\pi}\sigma.$$

To nas motivira da prag definiramo na sljedeći način, od prosječne ocjene odmaknemo se nekoliko standarnih devijacija udesno odnosno:

$$\text{prag} = \mu + \text{skala} \cdot \beta$$

gdje je skala prirodan broj. Što je taj broj veći, može se dogoditi da distribuciju "odrežemo" predaleko i na taj način odbacimo TP (pravi pozitivci). Međutim, što je skala manja, raste broj FP (lažni pozitivci). Na koji način odrediti skalu za koju ćemo dobiti najbolje pozitivce? O tome ćemo više pričati u 3. poglavlju.

## 2.5 Iterativni dio

U ovom dijelu bavimo se iterativnim pretraživanjem motiva. Za zadani motiv želimo pronaći njemu slične podnizove u određenom proteomu. To postižemo tako da iterativno gradimo profil motiva. Ulazni podatak je motiv zadan s jednim ili više nizova aminokiselina.

U prvoj iteraciji pretražujemo proteom i dobivamo listu pogodaka koja sadrži motive slične ulaznom motivu. Iz te liste uzimamo one motive čija je ocjena poravnanja veća od zadanog praga i spremamo ih u listu pozitivaca. Pomoću njih gradimo novi profil motiva. Proteom ponovno pretražujemo te dobivamo novu listu pozitivaca i novi model. Iterativni proces završava kada nema promjena u listi pozitivaca ili kada je zadani broj iteracija postignut.

Zanima nas jesu li motivi koje smo dobili biološki smisleni ili samo slučajno povezani nizovi. Također, kako odrediti prag tj. vrijednost skale za koju su dobiveni motivi smisleni. U sljedećem poglavlju ćemo to detaljnije opisati.

# Poglavlje 3

# Analiza točnosti

Da bismo odredili koji su motivi u listi pozitivaca biološki smisleni, a koji samo slučajno povezani nizovi, koristimo se nekim od statističkih mjera raspršenosti kao što su entropija i varijanca.

## 3.1 Entropija

Entropija je mjera nesigurnosti slučajne varijable. Za slučajnu varijablu $X$ koja ima vjerojatnosti $\mathbb{P}(x_i)$ gdje je $x_1, \ldots, x_k$ niz diskretnih događaja, definirana je entropija na sljedeći način:

$$H(x) = -\sum_i \mathbb{P}(x_i) \log \mathbb{P}(x_i). \tag{3.1}$$

Ovako definiranu entropiju nazivamo još i Shannonova[1] entropija. U daljnim računanjima koristimo prirodni logaritam s bazom $e = 2.7138$. Entropija je minimalna kada imamo siguran događaj tj. ako je $\mathbb{P}(x_j) = 1$ za neki $j \in \{1, 2, \ldots, n\}$ i $\mathbb{P}(x_i) = 0$ za $i \in \{1, 2, \ldots, n\} \backslash \{j\}$. U tom slučaju vrijednost entropije $H$ je 0.

**Napomena 3.1.1.** *Neka je $x \in \mathbb{R}$. Vrijedi* $\lim_{x \to 0^+} x \ln x = 0$.

*Dokaz.*

$$\lim_{x \to 0^+} x \ln x = \lim_{x \to 0^+} \frac{\ln x}{\frac{1}{x}} = \left(\frac{\infty}{\infty}\right) = \text{(L'Hospitalovo pravilo)} =$$

$$= \lim_{x \to 0^+} \frac{\frac{1}{x}}{-\frac{1}{x^2}} = \lim_{x \to 0^+} (-x) = 0$$

□

[1] Claude Shannon (1916. - 2001.), američki matematičar, poznat kao "otac teorije informacija"

**Napomena 3.1.2.** *Neka je* $p_i = \mathbb{P}(x_i) > 0$, $i = 1, 2, \ldots, n$ *gdje je* $\{x_1, x_2, \ldots, x_n\}$ *niz diskretnih događaja i vrijedi* $\sum_{i=1}^{n} p_i = 1$. *Želimo maksimizirati Shannonovu entropiju, odnosno jednadžbu* $f(p_1, p_2, \ldots, p_n) = -\sum_{i=1}^{n} p_i \log p_i$.

*Dokaz.* Koristimo metodu Lagrangeovog multiplikatora.

$$g(p_1, p_2, \ldots, p_n) = \sum_{i=1}^{n} p_i - 1$$

$$\begin{aligned}\mathcal{L}(p_1, p_2, \ldots, p_n, \lambda) &= f(p_1, p_2, \ldots, p_n) + \lambda \cdot g(p_1, p_2, \ldots, p_n) \\ &= -\sum_{i=1}^{n} p_i \log p_i + \lambda \cdot (\sum_{i=1}^{n} p_i - 1)\end{aligned}$$

$$\begin{aligned}\nabla_{p_1, p_2, \ldots, p_n, \lambda} \mathcal{L}(p_1, p_2, \ldots, p_n, \lambda) &= \Big( \frac{\partial \mathcal{L}}{\partial p_1}, \frac{\partial \mathcal{L}}{\partial p_2}, \ldots, \frac{\partial \mathcal{L}}{\partial p_n}, \frac{\partial \mathcal{L}}{\partial \lambda} \Big) = \\ &= (-\log p_1 + \lambda, \ldots, -\log p_n + \lambda, \sum_{i=1}^{n} p_i - 1)\end{aligned}$$

$$\nabla_{p_1, p_2, \ldots, p_n, \lambda} \mathcal{L}(p_1, p_2, \ldots, p_n, \lambda) = 0$$

Rješavanjem sustava jednadžbi dobivamo

$$p_1 = p_2 = \ldots = p_n$$

Uvrštavajući u $\sum_{i=1}^{n} p_i = 1$ slijedi da je

$$p_i = \frac{1}{n}, \quad i = 1, 2, \ldots, n$$

□

Maksimalna entropija se postiže kada su svi događaji jednako vjerojatni, odnosno kada su sve vjerojatnosti $\mathbb{P}(x_i)$ jednake ($\mathbb{P}(x_i) = \frac{1}{n}$, $i = 1, 2, \ldots, n$). U tom slučaju vrijednost entropije iznosi

$$H = -\sum_{i=1}^{n} \frac{1}{n} \log \frac{1}{n} = \log n.$$

Pogledajmo na primjeru motiva duljine 10, 'FVFGDSLVDN'. Za skalu 14, lista pozitivaca sadrži sljedećih 10 motiva:

FVFGDSLVDN FIFGDSLVDN
FIFGDSLVDN FVFGDSLIDN
FVFGDSLVDN FVFGDSLVDS
FIFGDSLVDN FVFGDSLVDN
FVFGDSLVDS FVFGDSLVDS

Računamo relativne frekvencije tako da za svaku poziciju u motivu sumiramo koliko puta se određena aminokiselina pojavila i podijelimo s ukupnim brojem motiva. Primjetimo da svi gornji nizovi imaju aminokiselinu F na prvoj poziciji. Stoga, dobivamo sljedeći vektor relativnih frekvencija za prvu poziciju:

$$[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0]$$

Entropiju računamo koristeći formulu (3.1) na sljedeći način.

$$\begin{aligned} H &= -[0 \cdot \log 0 + \ldots + 0 \cdot \log 0 + 1 \cdot \log 1 + 0 \cdot \log 0 + \ldots + 0 \cdot \log 0] \\ &= (3.1.1) = 0 \end{aligned}$$

Na isti način dobivamo entropiju za ostalih 9 pozicija.

$$H = [0, 0.61, 0, 0, 0, 0, 0, 0.33, 0, 0.61]$$

Primjećujemo da su pozicije na kojima je entropija 0 konzervirane, tj. na tim pozicijama nalazi se ista aminokiselina. Promotrimo ponašanje entropije za druge vrijednosti skale.

| | **entropija po pozicijama** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **skala** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** |
| **15** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **14** | 0 | 0.61 | 0 | 0 | 0 | 0 | 0 | 0.33 | 0 | 0.61 |
| **13** | 0.38 | 0.62 | 0 | 0 | 0 | 0 | 0 | 0.23 | 0 | 0.98 |
| **12** | 0.56 | 0.64 | 0 | 0 | 0 | 0 | 0 | 0.21 | 0 | 1.18 |
| **11** | 0.4 | 0.55 | 0.15 | 0 | 0 | 0 | 0.79 | 0.79 | 0.15 | 1.25 |
| **10** | 0.47 | 0.57 | 0.28 | 0 | 0 | 0 | 0.76 | 0.85 | 0.14 | 1.3 |
| **9** | 0.94 | 0.55 | 0.27 | 0 | 0 | 0.1 | 1.47 | 1.06 | 0.1 | 1.64 |
| **8** | 1.06 | 1.46 | 0.17 | 0 | 0 | 0.06 | 1.88 | 1.62 | 0.11 | 1.52 |
| **7** | 1.27 | 1.77 | 0.17 | 0.08 | 0.22 | 0.05 | 1.87 | 1.79 | 0.11 | 1.51 |
| **6** | 1.49 | 1.78 | 0.4 | 0.65 | 0.88 | 0.34 | 2.07 | 1.87 | 0.54 | 1.75 |
| **5** | 1.66 | 1.84 | 0.84 | 1.07 | 1.5 | 0.87 | 2.27 | 2.03 | 1.03 | 1.96 |

Tablica 3.1: Entropija po pozicijama za motiv ‘FVFGDSLVDN’

Iz tablice 3.1 možemo primijetiti da entropija raste smanjenjem skale. Također, vidimo da su neke pozicije očuvanije od drugih. Za skalu 8, vrijednost entropije na 4. i 5. poziciji je 0, dok je na ostalim pozicijama jako mala. Za skale manje od 8, entropija poprima veće vrijednosti. Dakle, možemo zaključiti da bi skala 8 mogla biti dobar prag.

Kada bi gledali prosječnu entropiju po pozicijama, mogli bi primijetiti da je najmanja za 4., 5., 6. i 7. poziciju, odnosno tamo gdje se nalaze aminokiseline G, D, S i L. To znači da većina motiva iz liste pozitivaca sadrži te aminokiseline na pripadnim pozicijama.

| **skala** | **broj pozitivaca** | **suma entropije** | **log entropije** |
|---|---|---|---|
| **15** | 3 | 0 | 0 |
| **14** | 10 | 0.67 | 1.55 |
| **13** | 16 | 0.8 | 2.22 |
| **12** | 18 | 0.89 | 2.58 |
| **11** | 29 | 1.21 | 4.09 |
| **10** | 31 | 1.28 | 4.38 |
| **9** | 48 | 1.58 | 6.13 |
| **8** | 91 | 1.75 | 7.88 |
| **7** | 125 | 1.83 | 8.85 |
| **6** | 174 | 2.28 | 11.77 |
| **5** | 273 | 2.69 | 15.06 |

Tablica 3.2: Broj pozitivaca i vrijednosti entropije za motiv ‘FVFGDSLVDN’

U tablici 3.2 prikazan je broj pozitivaca za svaku skalu od 5 do 15. Primjećujemo da smanjenjem skale, broj pozitivaca postepeno raste.

Ako gledamo sumu entropije u odnosu na vrijednost skale, također uočavamo lagani rast smanjenjem skale. Budući da nam suma ne daje puno informacija, logaritmirati ćemo na način da sumu entropije podijelimo s logaritmom broja pozitivaca. Dobiveni brojevi su u 4. stupcu tablice 3.2. Ne uočavamo nikakve značajnije razlike promjenom skale.

Pogledajmo sada ponašanje entropije na drugom motivu duljine 10, ‘PEPLISEILF’. Za skalu 14, lista pozitivaca sadrži samo dva sljedeća motiva:

PEPLISEILF
PEPLISEILF

Računajući entropiju kao u prethodnom primjeru dobivamo sljedeću tablicu.

| skala | broj pozitivaca | suma entropije | log entropije |
|---|---|---|---|
| **15** | 2 | 0 | 0 |
| **14** | 2 | 0 | 0 |
| **13** | 2 | 0 | 0 |
| **12** | 2 | 0 | 0 |
| **11** | 2 | 0 | 0 |
| **10** | 2 | 0 | 0 |
| **9** | 8 | 2.02 | 4.2 |
| **8** | 39 | 2.08 | 7.62 |
| **7** | 65 | 2.3 | 9.6 |
| **6** | 178 | 2.84 | 12.11 |
| **5** | 674 | 2.69 | 17.55 |

Tablica 3.3: Broj pozitivaca i vrijednost entropije za motiv 'PEPLISEILF'

Kod ovog motiva uočavamo nagli skok za skalu 8. Broj pozitivaca (39) je skoro 5 puta veći u odnosu na broj pozitivaca za skalu 9 (8). Suma entropije nam tu ne daje mnogo informacija, ali kod logaritma entropije uočavamo mali skok.

| | **entropija po pozicijama** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **skala** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** |
| **15** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **14** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **13** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **12** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **11** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **10** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **9** | 0 | 0.97 | 0.9 | 0 | 0 | 0.69 | 0.66 | 0 | 0 | 0.97 |
| **8** | 0.12 | 0.78 | 1.71 | 0 | 0.79 | 1.3 | 1.15 | 0.32 | 0.24 | 1.21 |
| **7** | 0.16 | 1.17 | 2.07 | 0.22 | 1.05 | 1.64 | 1.46 | 0.35 | 0.24 | 1.25 |
| **6** | 0.2 | 2.32 | 1.78 | 0.23 | 1.7 | 2.25 | 1.34 | 0.38 | 0.25 | 1.65 |
| **5** | 0.88 | 2.3 | 1.89 | 0.45 | 1.79 | 2.42 | 1.73 | 0.6 | 0.74 | 2.03 |

Tablica 3.4: Entropija po pozicijama za motiv 'PEPLISEILF'

Vrijednosti entropije po pozicijama za skalu od 10 do 15 su 0, što je za očekivati jer lista pozitivaca za te vrijednosti skale sadrži samo dva ista motiva. Primjetimo da je za skalu 8 očuvana samo 4. pozicija na kojoj se nalazi aminokiselina L. Za skale manje od 8,

entropija poprima sve veće vrijednosti. Maksimalna vrijednost entropije je $\ln(20) \approx 3$, a na nekim pozicijama uočavamo vrijednosti 2.42 i 2.25 što je jako blizu maksimuma.

Ako usporedimo entropiju po pozicijama kod motiva 'FVFGDSLVDN' (tablica 3.1) i 'PEPLISEILF' (tablica 3.4) uočavamo neke različitosti. U prvoj tablici, vrijednosti entropije postepeno rastu smanjenjem skale, dok kod drugog motiva uočavamo nagli skok entropije. Možemo vidjeti da za vrijednost skale 8 motiv 'FVFGDSLVDN' ima dvije očuvane pozicije, dok kod motiva 'PEPLISEILF' uočavamo jednu očuvanu poziciju, a na ostalim pozicijama vrijednosti entropije su veće nego kod prvog motiva. Iz toga možemo zaključiti da lista pozitivaca koju dobijemo iterativnim pretraživanjem motiva 'FVFGDSLVDN' sadrži biološki smislenije nizove nego lista pozitivaca dobivena za motiv 'PEPLISEILF'.

## 3.2 Varijanca

Iz velikog broja aminokiselinskih svojstava izveden je manji skup numeričkih vrijednosti, faktora. Definirano je preslikavanje u $\mathbb{R}^5$ koje čuva sve važne informacije o svojstvima aminokiselina.

| **aminokiselina** | I. | II. | III. | IV. | V. |
|---|---|---|---|---|---|
| **A** | -0.591 | -1.302 | -0.733 | 1.57 | -0.146 |
| **C** | -1.343 | 0.465 | -0.862 | -1.02 | -0.255 |
| **D** | 1.05 | 0.302 | -3.656 | -0.259 | -3.242 |
| **E** | 1.357 | -1.453 | 1.477 | 0.113 | -0.837 |
| **F** | -1.006 | -0.59 | 1.891 | -0.397 | 0.412 |
| **G** | -0.384 | 1.652 | 1.330 | 1.045 | 2.064 |
| **H** | 0.336 | -0.417 | -1.673 | -1.474 | -0.078 |
| **I** | -1.239 | -0.547 | 2.131 | 0.393 | 0.816 |
| **K** | 1.831 | -0.561 | 0.533 | -0.277 | 1.648 |
| **L** | -1.019 | -0.987 | -1.505 | 1.266 | -0.912 |
| **M** | -0.663 | -1.524 | 2.219 | -1.005 | 1.212 |
| **N** | 0.945 | 0.828 | 1.299 | -0.169 | 0.933 |
| **P** | 0.189 | 2.081 | -1.628 | 0.421 | -1.392 |
| **Q** | 0.931 | -0.179 | -3.005 | -0.503 | -1.853 |
| **R** | 1.538 | -0.055 | 1.502 | 0.44 | 2.897 |
| **S** | -0.228 | 1.399 | -4.760 | 0.67 | -2.647 |
| **T** | -0.032 | 0.326 | 2.213 | 0.908 | 1.313 |
| **V** | -1.337 | -0.279 | -0.544 | 1.242 | -1.262 |
| **W** | -0.595 | 0.009 | 0.672 | -2.128 | -0.184 |
| **Y** | 0.260 | 0.830 | 3.097 | -0.838 | 1.512 |

Tablica 3.5: Faktori

Faktror I. označava polarnost ili hidrofobnost.
Faktor II. je faktor sekundarne strukture.
Faktor III. se odnosi na molekularnu veličinu ili volumen.
Faktor IV. odražava relativnu kompoziciju aminokiselina u različitim proteinima.
Faktor V. označava elektrostatski naboj.

Svaku aminokiselinu možemo zamijeniti nizom vrijednosti prikazanih u tablici 3.5. Na tako transformiranim podacima možemo dalje računati. Motiv duljine 10 postaje vektor duljine 50 te računamo varijancu za svaku poziciju.

Pogledajmo za skalu 8, ponašanje varijance kod motiva 'FVFGDSLVDN' i 'PEPLISEILF'.

| | **varijanca po pozicijama** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **faktori** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** |
| **I.** | 0.1 | 0.59 | 0.001 | 0 | 0 | 0.0003 | 0.6 | 0.21 | 0.002 | 0.32 |
| **II.** | 0.1 | 0.33 | 0.002 | 0 | 0 | 0.0007 | 0.7 | 0.64 | 0.07 | 0.86 |
| **III.** | 1.6 | 1.47 | 0.13 | 0 | 0 | 0.403 | 4.27 | 4.55 | 0.57 | 5.17 |
| **IV.** | 0.4 | 0.37 | 0.04 | 0 | 0 | 0.002 | 0.5 | 0.63 | 0.003 | 0.28 |
| **V.** | 0.35 | 1.07 | 0.02 | 0 | 0 | 0.24 | 1.4 | 1.33 | 0.12 | 1.82 |

Tablica 3.6: Varijanca po pozicijama za 'FVFGDSLVDN'

| | **varijanca po pozicijama** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **faktori** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** |
| **I.** | 0.004 | 0.02 | 0.48 | 0 | 0.002 | 0.36 | 0.25 | 0.002 | 0.003 | 0.25 |
| **II.** | 0.01 | 0.73 | 1.94 | 0 | 0.05 | 0.29 | 0.24 | 0.009 | 0.011 | 1.17 |
| **III.** | 0.25 | 6.32 | 5.3 | 0 | 3.24 | 6.35 | 2.03 | 0.65 | 0.62 | 7.06 |
| **IV.** | 0.003 | 0.04 | 0.59 | 0 | 0.2 | 0.84 | 0.5 | 0.05 | 0.19 | 0.29 |
| **V.** | 0.04 | 1.38 | 1.2 | 0 | 0.73 | 2.19 | 0.9 | 0.28 | 0.15 | 1.98 |

Tablica 3.7: Varijanca po pozicijama za 'PEPLISEILF'

Možemo primijetiti da u Tablici 3.7 imamo dvije pozicije na kojima je varijabilnost 0, dok je na 3. i 6. poziciji vrijednost varijance jako mala. Kod motiva 'PEPLISEILF' uočavamo također varijabilnost 0 na 4. poziciji, ali na drugim pozicijima te vrijednosti su nešto veće.

# Poglavlje 4

# Arabidopsis thaliana

*Arabidopsis thaliana* (L.) Heynh. je mala biljka s cvjetovima porijeklom iz Europe, Azije i sjeverozapadne Afrike. Zbog svoje velike rasprostranjenosti, često se smatra korovom. Kao i ostale biljke iz porodice Brassicaceae, ova jednogodišnja biljka je jestiva te se često koristi kao začin u salatama. *A. thaliana* je diploid i ima relativno kratak genom, što je čini prvom biljkom kojoj je sekvencioniran genom. Vrlo je pogodna za istraživanje u molekularnoj biologiji i genetici. Proteom ove biljke, na kojem ispitujemo metodu ima 33410 nizova proteina.

U proteomu biljke *A. thaliana* pronađeno je 127 GDSL enzima. GDSL familija uključuje hidrolitičke enzime s multifunkcionalnim svojstvima koji imaju veliku primjenu u prehrambenoj i farmaceutskoj industriji. Sve je češća potraga za novim enzimima s korisnim svojstvima, a obilje GDSL enzima u biljnom svijetu indicira da bi biljke mogle biti dobar izvor tih enzima.

Rezultate dobivene metodom objašnjenom u poglavlju 2 za motiv 'FVFGDSLVDN', usporedili smo s listom poznatih GDSL motiva iz ovog biljnog proteoma. Broj pogođenih GDSL motiva za različite vrijednosti skale, prikazan je u Tablici 4.1.

| **skala** | **broj pozitivaca** | **broj pogođenih** |
|---|---|---|
| **15** | 3 | 3 |
| **14** | 10 | 10 |
| **13** | 16 | 16 |
| **12** | 18 | 18 |
| **11** | 29 | 29 |
| **10** | 31 | 31 |
| **9** | 48 | 48 |
| **8** | 91 | 91 |
| **7** | 125 | 115 |
| **6** | 174 | 119 |
| **5** | 273 | 122 |

Tablica 4.1: Broj pogođenih GDSL motiva

Primjećujemo da je broj pogođenih GDSL motiva za vrijednosti skale veće od 8, jednak broju motiva u listi pozitivaca. Smanjenjem skale, broj pogođenih GDSL motiva je manji od broja pozitivaca, odnosno lista pozitivaca sadrži i neke druge motive koji nisu GDSL nizovi.

Ispitat ćemo osjetljivost i specifičnost testa koja je objašnjena u Poglavlju 1.3. Pravi pozitivci (TP) su oni nizovi koje je metoda točno identificirala kao GDSL nizove, a lažni negativci (TN) su oni nizovi koji ne pripadaju GDSL familiji i metoda ih je točno svrstala u "negativce". Lažne negativce (FN) definiramo kao one GDSL nizove koje metoda stavlja ispod zadanog praga, dok su lažni pozitivci (FP) oni nizovi koje je metoda prepoznala kao pozitivce, ali ne pripadaju GDSL familiji.

| | | |
|---|---|---|
| TP = 122 | FN = 5 | osjetljivost = 0.9606 |
| FP = 151 | TN = 33132 | specifičnost = 0.9955 |
| PPV = 0.4469 | NPV = 0.9998 | |

Tablica 4.2: Uspješnost testa za skalu 5

| | | |
|---|---|---|
| TP = 119 | FN = 8 | osjetljivost = 0.9370 |
| FP = 55 | TN = 33228 | specifičnost = 0.9983 |
| PPV = 0.6839 | NPV = 0.9998 | |

Tablica 4.3: Uspješnost testa za skalu 6

| | | |
|---|---|---|
| TP = 115 | FN = 12 | osjetljivost = 0.9055 |
| FP = 10 | TN = 33273 | specifičnost = 0.9997 |
| PPV = 0.92 | NPV = 0.9996 | |

Tablica 4.4: Uspješnost testa za skalu 7

| | | |
|---|---|---|
| TP = 91 | FN = 36 | osjetljivost = 0.7165 |
| FP = 0 | TN = 33383 | specifičnost = 1 |
| PPV = 1 | NPV = 0.9989 | |

Tablica 4.5: Uspješnost testa za skalu 8

Budući da je broj GDSL nizova izrazito mali u odnosu na broj nizova u biljnom proteomu, jako veliki broj nizova biti će "negativci". Zbog toga negativno predviđena vrijednost (NPV) nije pogodan parametar za ocjenu uspješnosti testa. Iz rezultata možemo uočiti da pozitivno predviđena vrijednost (PPV) koja pokazuje koliki postotak pozitivnih nizova ("pozitivaca") pripada GDSL familiji, raste porastom skale, dok je za skale veće od 8 maksimalna. Međutim, osjetljivost testa se smanjuje porastom skale pa ako želimo bolju osjetljivost testa odabrat ćemo manju skalu.

# Bibliografija


[1] R. Durbin, S. R. Eddy, A. Krogh, G. Mitchison, *Biological sequence analysis: probabilistic models of proteins and nucleic acids*, Cambridge university press, 1998.

[2] A. Medved, *Lokalno poravnanje i prepoznavanje motiva*, Diplomski rad, Sveučilište u Zagrebu, Prirodoslovno-matematički fakultet, Matematički odsjek, 2016.

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

[4] D. Vukajlija et al, *Deep scanning for GDSL motifs in plant proteoms*

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

# Sažetak

U ovom radu bavimo se metodom za iterativno traženje motiva u nekom proteomu ili velikom skupu proteinskih nizova. Ocjenu sličnosti motiva i podniza definirali smo kao logaritam omjera nekih vjerojatnosti i ustanovili da su dobivene vrijednosti logistički distribuirane. Koristeći mjere raspršenosti ili varijabilnosti, analizirali smo dobivene rezultate i povezali te mjere s preciznošću i osjetljivošću iterativne metode.

# Summary

In this work, we are concerned with an iterative motif scanning method. We defined the similarity score in terms of a log-odds ratio, and established that the scores tend to be logistically distributed. We analyzed our results in terms of dispersion measures and established a connection of these measures with sensitivity and specificity of the iterative procedure.

# Životopis

Rođena sam 01.04.1992. godine u Splitu. Školovanje sam započela u Osnovnoj školi Hvar, koju sam pohađala od 1998. do 2006. godine, i nastavila u III. Gimnaziji u Splitu, gdje sam maturirala 2010. godine. Nakon toga upisala sam Preddiplomski studij Matematika na Prirodoslovno-matematičkom fakultetu u Zagrebu. Završetkom preddiplomskog studija 2014. godine upisala sam Diplomski studij Matematička statistika također na Prirodoslovno-matematičkom fakultetu u Zagrebu.