SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Antonija Medved

LOKALNO PORAVNANJE I PREPOZNAVANJE MOTIVA

Diplomski rad

Voditelj rada:
doc. dr. sc. Pavle Goldstein

Zagreb, rujan 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. _____

Zahvaljujem mentoru doc. dr. sc. Pavlu Goldsteinu na posvećenom vremenu i pruženoj pomoći u izradi ovog diplomskog rada. Najveće hvala mojoj obitelji na bezuvjetnoj podršci i ljubavi koju su mi pružili tijekom studiranja.

Sadržaj

Uvod

U današnje vrijeme računala su postala nezaobilazan dio svakodnevnog života, pa tako i u znanosti. S ubrzanim razvojem tehnologije omogućeno je pohranjivanje i obrada velike količine podataka. Zahtjevi za sofisticiranom analizom bioloških nizova postali su sve jači, što je na kraju dovelo do razvoja područja bioinformatike. Primarni joj je cilj povećanje razumijevanja bioloških procesa primjenom računalne tehnologije i statističkih modela.

Razmatramo problem iterativnog pretraživanja baze podataka za varijantama zadanog tekstualnog obrasca (*engl. string matching*). Preciznije, za dani proteom, skup proteoma ili jednostavno skup proteinskih nizova želimo pronaći najbolja podudaranja sa danim motivom u svakom retku, odnosno proteinu. To se postiže iterativnom nadogradnjom profila motiva. Ideja je brzo pronaći mjesta mogućih podudaranja i time smanjiti prostor pretraživanja, ali svejedno optimizirati funkciju sličnosti.

Ovaj rad podijeljen je na 4 poglavlja. U prvom dajemo kratak osvrt na osnovne pojmove iz vjerojatnosti i statistike koji su nužni za razumijevanje analize koju provodimo. Detaljnija objašnjena te dokazi iznesenih tvrdnji mogu se pronaći u [3]. Razvoj i opis modela koji koristimo za pretraživanje precizno je iznesen u drugom poglavlju. Najprije se konstruira funkcija kojom mjerimo sličnost dvaju nizova aminokiselina te je dana njihova statistička analiza pomoću teorije ekstremnih vrijednosti. Nadalje objašnjen je princip na kojem pretražujemo određenu bazu podataka za danim motivom, te je na kraju opisana iteracija modela i kriterij zaustavljanja. U trećem poglavlju bavimo se ubrzavanjem izvršavanja programa uvođenjem indeksacije proteoma. Naposljetku, u zadnjem poglavlju dajemo i komentiramo rezultate na stvarnom proteomu biljke *Arabidopsis thaliana* u kojem smo tražili motive GDSL enzima.

Poglavlje 1

Pojmovi iz vjerojatnosti i statistike

1.1 Vjerojatnost

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

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

Osnovni objekt u teoriji vjerojatnosti jest neprazan skup Ω koji zovemo **prostor elementarnih događaja** i koji reprezentira skup svih ishoda slučajnih pokusa. Točke ω iz skupa Ω zvat ćemo **elementarni događaji**.

Označimo sa $\mathcal{P}(\Omega)$ partitivni skup od Ω.

Definicija 1.1.1. *Familija $\mathcal{F}$ podskupova od Ω ($\mathcal{F} \subset \mathcal{P}(\Omega)$) jest* ***$\sigma$-algebra skupova*** *(na Ω) 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 $\mathcal{F}$ σ-algebra na skupu Ω. Uređen par $(\Omega, \mathcal{F})$ se zove* ***izmjeriv prostor****.*

Sad možemo definirati vjerojatnost.

Definicija 1.1.3. *Neka je $(\Omega, \mathcal{F})$ izmjeriv prostor. Funkcija $\mathbb{P} : \mathcal{F} \to \mathbb{R}$ jest* ***vjerojatnost*** *ako vrijedi:*

(P1) $\mathbb{P}(\Omega) = 1$ *(normiranost vjerojatnosti)*

(P2) $\mathbb{P}(A) \geqslant 0,\ A \in \mathcal{F}$ *(nenegativnost vjerojatnosti)*

(P3) $A_i \in \mathcal{F},\ i \in \mathbb{N}$ *te* $A_i \cap A_j = \emptyset$ *za* $i \neq j \Rightarrow \mathbb{P}(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$ *(prebrojiva ili σ-aditivnost vjerojatnosti)*

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

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

Slučajna varijabla i funkcija distribucije

Označimo sa $\mathcal{B}$ σ-algebru generiranu familijom svih otvorenih skupova na skupu realnih brojeva $\mathbb{R}$. $\mathcal{B}$ zovemo **σ-algebra skupova na $\mathbb{R}$**, a elemente σ-algebre $\mathcal{B}$ zovemo **Borelovi skupovi**.

Definicija 1.1.4. *Neka je $(\Omega, \mathcal{F}, \mathbb{P})$ vjerojatnosni prostor. Funkcija $X : \Omega \to \mathbb{R}$ jest **slučajna varijabla** (na Ω) ako je $X^{-1}(B) \in \mathcal{F}$ za proizvoljno $B \in \mathcal{B}$, odnosno $X^{-1}(\mathcal{B}) \subset \mathcal{F}$.*

Definicija 1.1.5. ***Funkcija distribucije slučajne varijable*** *X jest funkcija $F_X = F : \mathbb{R} \to [0, 1]$ definirana sa:*

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

Uvjetna vjerojatnost i nezavisnost

Definicija 1.1.6. *Neka je $(\Omega, \mathcal{F}, \mathbb{P})$ proizvoljan vjerojatnosni prostor i $A \in \mathcal{F}$ takav da je $\mathbb{P}(A) > 0$. Definiramo funkciju $P_A : \mathcal{F} \to [0, 1]$ ovako:*

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

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

Definicija 1.1.7. *Neka je $(\Omega, \mathcal{F}, \mathbb{P})$ vjerojatnosni prostor i $A_i \in \mathcal{F}, i \in I$ proizvoljna familija događaja. Kažemo da je to **familija nezavisnih događaja** ako za svaki konačan podskup različitih indeksa $i_1, i_2, \ldots, i_k$ vrijedi*

$$\mathbb{P}\left(\bigcap_{j=1}^{k} A_{i_j}\right) = \prod_{j=1}^{k} \mathbb{P}(A_{i_j}). \tag{1.2}$$

Primjeri slučajnih varijabli

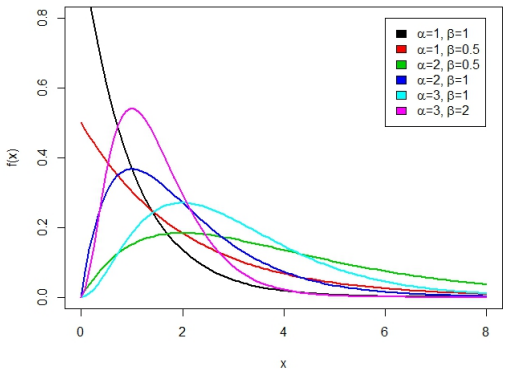
Gama distribucija. Eksponencijalna distribucija

Neka je $\alpha > 0, \beta > 0$ i $\Gamma(x) = \int_0^\infty e^{-t}t^{x-1}\,dt$, $x > 0$ gama funkcija. Neprekidna slučajna varijabila X ima **gama distribuciju** s parametrima α i β ako joj je funkcija gustoće f dana s:

$$f(x) = \begin{cases} \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{\frac{-x}{\beta}}, & x > 0 \\ 0, & x \leqslant 0. \end{cases} \tag{1.3}$$

Ako je $\alpha = 1$ i $\beta = \frac{1}{\lambda}$, tada kažemo da X ima **eksponencijalnu distribuciju** s parametrom λ. Funkcija gustoće eksponencijalne distribucije s parametrom λ jest

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x > 0 \\ 0, & x \leqslant 0. \end{cases} \tag{1.4}$$



Slika 1.1: Funkcije gustoće gama distribucije s različitim parametrima

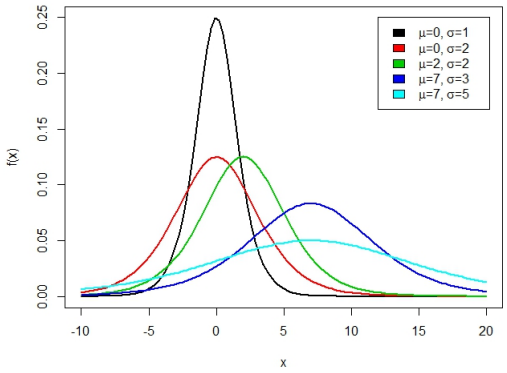
Logistička distribucija

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

$$f(x) = \frac{e^{-\frac{x-\mu}{\beta}}}{\beta\left(1 + -\frac{x-\mu}{\beta^2}\right)},\ x \in \mathbb{R}. \tag{1.5}$$

Neka je $p, q > 0$. Slučajna varijabla X ima **generaliziranu logističku distribuciju** ako joj je funkcija gustoće dana s:

$$f(x) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} \frac{e^{pq}}{(1+e^{y})^{p+q}}, \ x \in \mathbb{R}. \tag{1.6}$$



Slika 1.2: Funkcije gustoće logističke distribucije s različitim parametrima

1.2 Teorija ekstremnih vrijednosti

Sljedeći pojmovi bit će nam korisni kod analize distribucije ocjena poravnanja. Pojmovi su preuzeti iz [4] i [1], te su tamo može vidjeti njihova motivacija i dokaz.

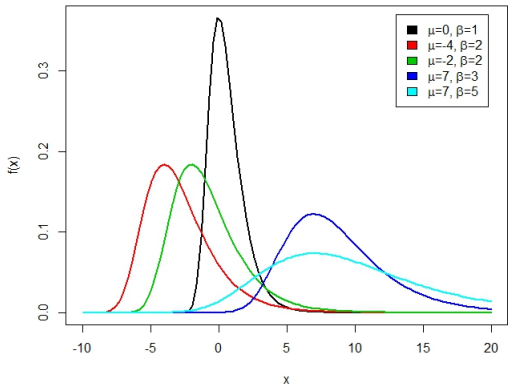
Gumbel distribucija

Neka su $\mu \in \mathbb{R}$ i $\beta > 0$. Neprekidna slučajna varijabla X ima **Gumbel distribuciju** sa parametrima μ i β ako joj je funkcija gustoće dana s:

$$f(x) = \frac{1}{\beta} e^{-\frac{x-\mu}{\beta} - e^{-\frac{x-\mu}{\beta}}}, \ x \in \mathbb{R}. \tag{1.7}$$

Neka je $p > 0$. Slučajna varijabla X ima **generaliziranu Gumbel distribuciju** ako joj je funkcija gustoće dana s:

$$f(x) = \frac{1}{\Gamma(p)} e^{-px} e^{e^{-px}}, \ x \in \mathbb{R}. \tag{1.8}$$



Slika 1.3: Funkcije gustoće Gumbel distribucije s različitim parametrima

Korolar 1.2.1. *Neka su X_1 i X_2 nezavisne generalizirane Gumbel distribuirane slučajne varijable s parametrima p i q, respektivno. Tada slučajna varijabla $Y = X_1 - X_2$ ima generaliziranu logističku distribuciju s parametrima p i q.*

Fréchetova distribucija

Neka su $\alpha > 0, \beta > 0$ i $\mu \in \mathbb{R}$. Slučajna varijabla X ima **Fréchetovu distribuciju** ako joj je funkcija gustoće dana s:

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

Weibullova distribucija

Neka su $\alpha > 0, \beta > 0$. Slučajna varijabla X ima **Weibullovu distribuciju** ako joj je funkcija gustoće dana s:

$$f(x) = \begin{cases} \frac{\alpha}{\beta}\left(\frac{x}{\beta}\right)^{\alpha-1} e^{-\left(\frac{x}{\beta}\right)^{\alpha}}, & x \geqslant 0 \\ 0, & x < 0. \end{cases} \tag{1.10}$$

Teorem 1.2.2. *Neka su $X_1, X_2, \ldots, X_n$ jednako distribuirane slučajne varijable i neka je $M_n = \max\{X_1, X_2, \ldots, X_n\}$. Ako postoji $a_n > 0$ i $b_n \in \mathbb{R}$ tako da je $\lim_{n\to\infty} \mathbb{P}\left(\frac{M_n - b_n}{a_n} \leqslant x\right) = F(x)$, gdje je F nedegenerirana distribucija, tada je granična distribucija F pripada Gumbel, Fréchet ili Weibull distribuciji.*

1.3 Osjetljivost i specifičnost testa

Sljedeći pojmovi trebat će nam kod analize uspješnosti metode.

		Predviđeno stanje		
	Ukupna populacija	predviđeno pozitivno stanje	predviđeno negativno stanje	
Stvarno stanje	pozitivno stanje	stvarno pozitivno (TP)	lažno negativno (FN)	osjetljivost
	negativno stanje	lažno pozitivno (FP)	stvarno negativno (TN)	specifičnost
		PPV	NPV	

Tablica 1.1: Veličine uspješnosti testa

Osjetljivost testa (stopa stvarno pozitivnih) mjeri proporciju pozitivnih ispravno prepoznatih testom od ukupnog broja pozitivnih, dok specifičnost testa (stopa stvarno negativnih) mjeri proporciju negativnih ispravno prepoznatih testom od ukupnog broja negativnih.

$$\text{osjetljivost} = \frac{\text{broj stvarno pozitivnih}}{\text{broj stvarno pozitivnih + broj lažno negativnih}}$$

$$\text{specifičnost} = \frac{\text{broj stvarno negativnih}}{\text{broj stvarno negativnih + broj lažno pozitivnih}}$$

Uz osjetljivost i specifičnost testa korisno je definirati veličine kojima opisujemo učinkovitost testa. Pozitivno prediktivna vrijednost (PPV) mjeri u kojem postotku pozitivno identificirani zaista jesu takvi, s druge strane negativna prediktivna vrijednost (NPV) mjeri postotak negativno identificiranih koji zaista jesu negativni.

$$\text{pozitivna prediktivna vrijednost} = \frac{\text{broj stvarno pozitivnih}}{\text{broj stvarno pozitivnih + broj lažno pozitivnih}}$$

$$\text{negativna prediktivna vrijednost} = \frac{\text{broj stvarno negativnih}}{\text{broj stvarno negativnih + broj lažno negativnih}}$$

Poglavlje 2

Model ocjenjivanja i traženje motiva

2.1 Osnovni biološki pojmovi

Najvažnije tvari u tijelu, uz vodu, su *proteini* ili *bjelančevine*. Te kemijske tvari upravljaju svim životnim procesima stanice, a kao dijelovi svake stanice čine osnovu života na Zemlji. Odgovorne su za proces rasta i razvoja, te nadomještanje oštećenih i odumrlih stanica, služe za ubrzavanje metaboličkih reakcija, omogućavaju komunikaciju i usklađivanje biokemijskih procesa između različitih tkiva i organa i brojne druge funkcije. U prirodi postoji više od 500 različitih aminokiselina ali proteini svih vrsta, od bakterija do ljudi, sastoje se od 20 aminokiselina popisanih u tablici (2.1).

Naziv	**Kratica**	**Naziv**	**Kratica**
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 2.1: Standardne aminokiseline i njihove oznake

Sveukupan skup proteina proizveden ili modificiran nekim organizmom, odnosno sustavom nazivamo *proteom*. On može varirati s obzirom na vrijeme ili različite zahtjeve i naprezanja kojima je stanica ili organizam podvrgnut.

Motiv proteinskog niza je kratak obrazac sastavljen od nekoliko aminokiselina, u pravilu sadrži 5 do 20 aminokiselina, koji je ostao sačuvan selekcijskim pročišćavanjem i ima neko biološko značenje. Može predstavljati aktivno mjesto enzima ili strukturnu jedinicu potrebnu za pravilno sklapanje proteina. Stoga je motiv jedna od osnovnih funkcionalnih jedinica molekularne evolucije.

2.2 Model ocjenjivanja

Jedno od temeljnih pitanja kojima se bioinformatika bavi je pretraživanje baze podataka bioloških nizova u potrazi za sličnim nizovima, nekog danog niza, takvih da sličnost nije slučajna. Uspoređuju se nizovi aminokiselina nekog organizma s bazom nizova specifičnih za protein koji je od interesa da bi se proširila ta baza u svrhu lakšeg otkrivanja tog proteina u nekim novim organizmima. Sličnost upućuje na zajedničko podrijetlo organizama, pa samim time i na sličnost biološke funkcije. Takvim analitičkim pristupom mogu se dijelom izbjeći skupi i dugotrajni eksperimenti u kojima bi se izravno utvrđivala funkcija pojedinog proteina pronađenog u novosekvencioniranim organizmima.

Genetički materijal svih bića kroz generacije se mijenja i tu promjenu nazivamo mutacijom. Osnovni mutacijski procesi su zamjena (*engl. supstitution*), umetanje (*engl. insertion*) i brisanje (*engl. deletion*) određene aminokiseline u nizu. Postavlja se pitanje kako za dva određena niza aminokiselina, bez poznavanja evolucijskih događaja, odrediti imaju li oni zajedničko podrijetlo, odnosno pripadaju li istoj proteinskoj familiji. Konkretno, za dani motiv želimo pronaći sve njegove varijante u nekom organizmu. Pojednostavljeno rečeno, dani niz slova želimo aproksimirati nekim drugim nizova slova jednake duljine koji mu je "dovoljno sličan". U ovom poglavlju opisan je matematički aparat kojim identificiramo sličnost dvaju nizova.

Neka su x = `FIFGDSLYDN` i y = `FVFGDSLYDD` dva niza koja želimo usporediti. Najprije ćemo ih potpisati jedan ispod drugoga

```
F I F G D S L Y D N
F V F G D S L Y D D
```

Taj postupak nazivamo poravnanje nizova. Općenito, poravnavati možemo nizove različitih duljina umetanjem praznina, koje grafički prikazujemo kao "-", na odgovarajuća mjesta, a predstavljaju mjesta na kojima se desilo umetanje ili brisanje određene aminokiseline. Međutim, za nas je irelevantno promatrati takav model jer polazimo od pretpostavke da je motiv predstavljen najočuvanijom regijom u nizu, a mi pretražujemo bazu podataka (dani proteom) za svim njemu sličnima i ne očekujemo umetanja i brisanja u motivu.

Nama je od interesa u svakom retku proteoma, koji predstavlja jedan protein duljine n, pronaći podniz duljine m koji se najviše poklapa s polaznim motivom duljine m. Najprije

ćemo uvesti funkciju sličnosti, koju ćemo zvati **ocjena poravnanja** (*engl. score*), a koja će svakom poravnanju dodijeliti realan broj S. Prirodno, što je poravnanje bolje, odnosno što ima više poklapanja, to će S biti veći. Jasno je da ćemo uzeti onaj podniz u retku koji ima najveću ocjenu u tom retku.

Razmatramo dva potpuno poravnata niza jednakih duljina $x = x_1 \dots x_n$ i $y = y_1 \dots y_n$. Simboli x_i i y_i elementi su $\mathcal{A} = \{\mathtt{A}, \mathtt{R}, \mathtt{N}, \mathtt{D}, \mathtt{C}, \mathtt{Q}, \mathtt{E}, \mathtt{G}, \mathtt{H}, \mathtt{I}, \mathtt{L}, \mathtt{K}, \mathtt{M}, \mathtt{F}, \mathtt{P}, \mathtt{S}, \mathtt{T}, \mathtt{W}, \mathtt{Y}, \mathtt{V}\}$, alfabeta 20 aminokiselina. Danom paru želimo dodijeliti ocjenu kojom iskazujemo u kojoj mjeri su nizovi povezani, u odnosu na to koliko nisu. Izračunat ćemo vjerojatnosti za svaki slučaj posebno, a zatim ćemo pogledati njihov omjer.

Opišimo prvo model u kojem dva niza nisu povezana, odnosno slučajni (*engl. random*) model R. U takvom modelu pretpostavlja se da se simbol a dogodi nezavisno od drugih s nekom vjerojatnošću q_a. Uzevši u obzir da je vjerojatnost dva nezavisna niza jednaka produktu vjerojatnosti da se dogodi svaka njihova pojedina aminokiselina slijedi:

$$\mathbb{P}(x, y \mid R) = \prod_i q_{x_i} \prod_j q_{y_j}. \tag{2.1}$$

U modelu M (*engl. match*) u kojem želimo opisati povezanost nizova, pretpostavljamo da se poravnati par simbola a i b pojavljuje sa zajedničkom vjerojatnošću $p_{a,b}$. 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*).

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

Htjeli bismo da sustav ocjenjivanja bude aditivan, pa gornji omjer još i logaritmiramo, te dobivamo:

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

Ocjene $s(x_i, y_i)$ zapisuju se u matricu oblika 20×20 koju zovemo **supstitucijska matrica**.

Za model R, umjesto uniformne distribucije uzimamo sljedeću distribuciju koja je dobivena računanjem relativnih frekvencija aminokiselina u proteomima nekog većeg skupa organizama:

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

Model M gradimo na temelju ulaznog motiva. Za i-ti stupac danog motiva izračuna se relativna frekvencija svih aminokiselina $f_i = (f_{i1}, f_{i2}, \ldots, f_{i20})$ za $i = 1, 2, \ldots, m,$ gdje je n širina motiva. Matricu f zovemo još **PSSM** (*engl. Position Specific Scoring Matrix*) matrica, odnosno matrica ocjena specifičnih za pojedinu poziciju. Zatim na njih primijenimo blagu težinsku shemu da bismo ispravili mogući nedostatak nezavisnosti kada se motiv sastoji od više nizova. Budući da je motiv relativno mali uzorak, kako bismo izbjegli da neka od f_{ij} bude jednaka 0 dodajemo pseudo-zbroj. Time dobivamo vektore distribucija $g_i = (g_{i1}, g_{i2}, \ldots, g_{i20})$ za $i = 1, 2, \ldots, m,$ pri čemu je:

$$g_{ij} = \frac{f_{ij} + 0.01}{1.2},\ i = 1, 2, \ldots, m,\ j = 1, 2, \ldots, 20.$$

Neka je $A = (a_{ij})$ PAM (*engl. Point Accepted Mutation*) matrica. Matrica A je stohastička, tj.

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

i stavimo $B = (b_{ij}) = A^k$, gdje je k dovoljno velik ($k = 120$). Redak $b_i = (b_{i1}, b_{i2}, \ldots, b_{i20})$ tako dobivene matrice predstavlja očekivanu mutaciju i-te aminokiseline nakon k milijuna godina evolucije. Sada možemo definirati vjerojatnost da se u i-tom stupcu motiva dogodi j-ta aminokiselina:

$$p_{ij} = \sum_{k=1}^{20} g_{ik} b_{kj}.$$

Sada samo dobivene p i q uvrstimo u formulu (2.4).

2.3 String matching

Računalni proces traženja najsličnijeg podniza ulaznom motivu u svakom retku proteoma vrlo je spor jer su proteomi ogromne baze podataka. Neka je x jedan redak proteoma duljine n i y motiv duljine m gdje je $m < n$. Grafički ćemo prikazati najjednostavniji princip uspoređivanja tih dvaju nizova s ciljem pronalaska najsličnijeg podniza. Riječ je o metodi klizećeg prozora (*engl. sliding window*):
Na k-toj poziciji evaluiramo formulu (2.4) za nizove $x^{(k)} = x_k x_{k+1} \ldots x_{k+m-1}$ i y, odnosno za svaki "prozor" duljine m računamo

$$s_k = \sum_{i=0}^{m-1} \log \frac{\mathbb{P}(x_{k+i}|M_i)}{\mathbb{P}(x_{k+i}|q)},\ k = 1, 2, \ldots, n - m + 1,$$

$$
\begin{array}{cccccccccc}
x_1 & x_2 & x_3 & \cdots & x_{m-1} & x_m & x_{m+1} & x_{m+2} & \cdots & x_n \\
y_1 & y_2 & y_3 & \cdots & y_{m-1} & y_m & & & &
\end{array}
$$

$$
\begin{array}{ccccccccc}
x_1 & x_2 & x_3 & x_4 & \cdots & x_m & x_{m+1} & x_{m+2} & \cdots & x_n \\
 & y_1 & y_2 & y_3 & \cdots & y_{m-1} & y_m & & &
\end{array}
$$

$$
\begin{array}{cccccccccc}
 & & & & & . & & & & \\
 & & & & & & . & & & \\
 & & & & & & & . & &
\end{array}
$$

$$
\begin{array}{ccccccccccc}
x_1 & x_2 & x_3 & \cdots & x_m & \cdots & x_{n-m+1} & x_{n-m+2} & x_{n-m+3} & \cdots & x_n \\
 & & & & & & y_1 & y_2 & y_3 & \cdots & y_m
\end{array}
$$

gdje $M_i,\ i = 0, 1, \ldots, m-1$ predstavlja model za i-tu poziciju motiva, te pogledamo koji od njih ima najveću ocjenu

$$S = \max_{k=0,1,\ldots,n-m+1} s_k.$$

Postupak ponavljamo za svaki redak proteoma, odnosno ako s d označimo broj redaka u proteomu, u prvoj iteraciji ocjenu poravnanja treba izračunati $d \times (n - m + 1)$, što je za proteom duljine oko 33000, motiv širine 10 te prosječne duljine niza u proteomu 400 otprilike 13 milijuna evaluacija formule (2.4).

Kao alternativa, ovakvom se problemu pristupa heurističkim rješenjima. Njihova je osobina značajno ubrzanje, međutim pronalazak optimalnog rješenja nije zajamčen. Takvim pristupom postižemo svojevrstan kompromis između brzine kojom se dolazi do rješenja te osjetljivosti dobivenog rješenja.

Razmotrimo najprije generalan problem čiji je cilj naći neku riječ ili kratak obrazac u nekom tekstu uzimajući u obzir da su i tekst i riječ bili izloženi nekim (neželjenim) promjenama. Formalno problem možemo postaviti na sljedeći način. Neka je $\mathcal{A}$ konačan alfabet. Za *tekst* $x \in \mathcal{A}$ duljine n i *riječ* $y \in \mathcal{A}$ duljine m takva da je $m < n$ uz danu maksimalnu dozvoljenu grešku $k \in \mathbb{R}$ i funkciju udaljenosti $d : \mathcal{A} \times \mathcal{A} \to \mathbb{R}$ treba naći sve pozicije i u x takve da vrijedi:

$$d(x_i \ldots x_{i+m-1}, y) < k.$$

Funkcija d još se naziva se i *Hammingova udaljenost*. Podudaranju dvaju simbola dodjeljuje iznos 0, dok se za svaku zamjenu odnosno različitost dodaje 1.

Prema tome, metodom klizećeg prozora prolazimo kroz svaki redak proteoma, te umjesto da odmah svaku poziciju ocjenjujemo formulom (2.4) pa gledamo maksimum svih, prvo ćemo izračunat Hammingovu udaljenost. Odnosno za svaku poziciju izračunat ćemo

broj mjesta na kojima se motiv poklapa s podnizom tog retka. Recimo na primjeru nizova

```
F I F G D S L Y D N
F V F G D S L Y D D
```

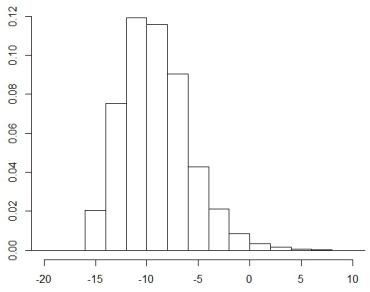
broj poklapanja iznosi 8. Slično kao kod računanja ocjena, za redak duljine n i motiv duljine m dobijemo $n - m + 1$ brojeva, od kojih zapamtimo samo najveće (može se desiti da više podnizova ima isti broj poklapanja). Počevši od druge iteracije pa nadalje model evaluiramo samo na tim pozicijama, što je u prosjeku 100 puta manje pozicija po nizu.

2.4 Značajnost ocjene

Postavlja se novo pitanje, na koji način ćemo procijeniti jesu li maksimalne ocjene koje smo izračunali statistički značajne? Odnosno, kako ćemo odlučiti da li je poravnanje koje smo dobili biološki smislen dokaz povezanosti ili samo najbolje poravnanje dva potpuno nepovezana niza.

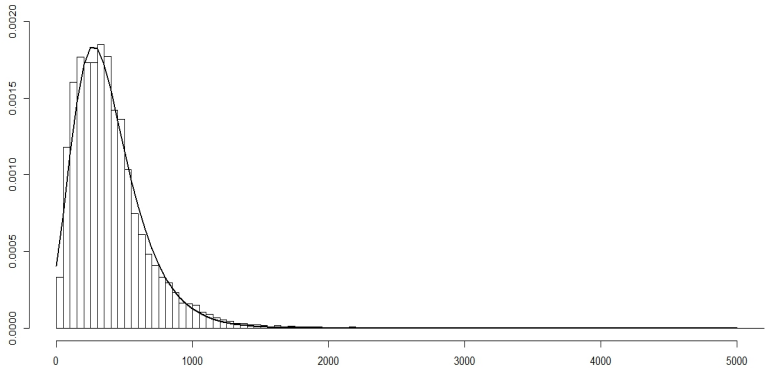
Vjerojatnost da je poravnanje nastalo slučajno ovisit će o nekoliko stvari. Jednostavno je za primijetiti da će u slučaju pretraživanja veće baze podataka i vjerojatnost da dođe do slučajnog "kvalitetnog" poravnanja s nekim od nizova te baze biti veća. Isti problem javlja se i s kratkim upitnim nizovima. Jasno je da za upit duljine 5 vjerojatnost da se pojavi u bazi duljine nekoliko milijuna velika, dok je za upit duljine 50 znatno manja. Eksperimentima se pokazalo da je praktično uzeti motive duljine 10.

Za početak potrebno je odrediti kako su distribuirane maksimalne ocjene svakog niza. Ako pogledamo distribuciju ocjena po svakoj poziciji pojedinog niza proteoma, koje su evaluirane kao omjer log-vjerodostojnosti, uočavamo da imaju eksponencijalni rep. Budući da nas zanima distribucija maksimuma ocjena po svakom nizu proteoma, za daljnje zaključke dovoljno je da samo rep bude eksponencijalan jer se maksimum nalazi baš u repu.



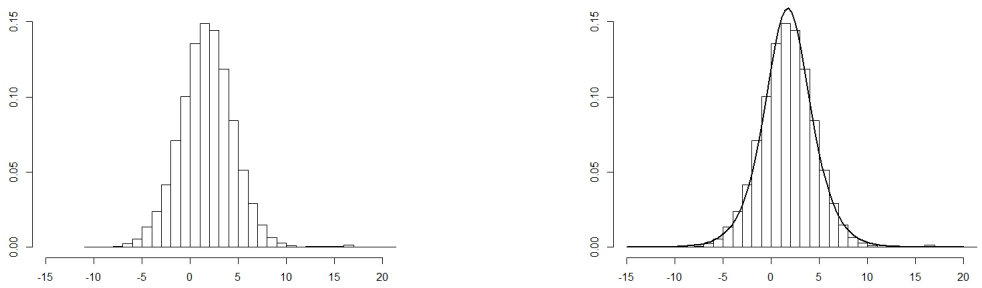
Slika 2.1: Histogram ocjena po svakoj poziciji jednog niza proteoma

Pretpostavimo li da su svi nizovi jednake duljine, odnosno jednaku distribuiranost nizova, tada bi iz teorije ekstremnih vrijednosti slijedilo da maksimalne ocjene po svakom nizu prate Gumbelovu distribuciju. Što se lako može provjeriti simulacijama. Detaljnije o tome u [4]. Međutim nizovi nisu jednake duljine, pa takvo što ne možemo zaključiti. Nacrtamo li histogram duljina svih nizova naslućujemo da bi mogli pratiti Gumbelovu distribuciju.



Slika 2.2: Histogram duljine nizova protema *A. thaliana* i funkcija gustoće Gumbel distribucije

Dakle promatramo dvije Gumbel distribuirane slučajne varijable, pa prema korolaru (1.2.1) razlika dviju takvih prati logističku distribuciju. Pogledajmo histogram maksimalnih ocjena.



Slika 2.3: Distribucija maksimalnih ocjena i funkcija gustoće logističke distribucije

Sada kada smo utvrdili da su maksimalne ocjene po nizovima logistički distribuirane, preostaje definirati na koji način ćemo odlučivati koje od tih ocjena su statistički značajne. Statistički formulirano testiramo sljedeću hipotezu:

$$H_0 : \text{nizovi nisu povezani}$$
$$H_1 : \text{povezani su}$$

Testna statistika je ocjena poravnanja. Po definiciji, p-vrijednost testa je vjerojatnost da, ako vrijedi nulta hipoteza, dobijemo broj veći ili jednak testnoj statistici, odnosno da je ocjena poravnanja veća ili jednaka opaženoj. Jasno da ako dobijemo veliku p vrijednost riječ je o slučajnom događaju, dok za malu p-vrijednost odbacujemo nultu hipotezu o nepovezanosti. Dakle, gledamo desni rep logističke distribucije. Preostaje odrediti prag od kojeg nadalje svaki podniz dobiven prethodnim izračunom smatramo pogotkom, odnosno koje od redaka proteoma . Uočimo da parametar β logističke distribucije možemo izraziti preko standardne devijacije distribucije. Vrijedi $\beta = \frac{\sqrt{3}}{\pi}\sigma$. To nas motivira da prag definiramo na sljedeći način, od prosječne ocjene odmaknemo se nekoliko standardnih devijacija udesno, odnosno:

$$prag = \mu + skala \cdot \beta,$$

gdje je *skala* prirodan broj i eksperimentalno je utvrđeno da se najbolji rezultati postižu za *skala* = 6, 7, 8. Za veće brojeve veće od 8, postoji mogućnost da distribuciju "odrežemo" predaleko, pa time odbacimo stvarno pozitivne. Dok za brojeve manje od 6, raste broj lažno pozitivnih.
Primijetimo da je funkcija distribucije logističke slučajne varijable X dana s:

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

Uzevši *skala* = 7, kao da smo pozitivno rangirali

$$1 - F(7) = 1 - \frac{e^7}{1 + e^7} = 0.0009,$$

oko 0.1% maksimalnih ocjena.

2.5 Iteriranje

Ulazni podatak procedure je motiv koji se može sastojati od jednog niza ili više poravnatih nizova aminokiselina. Kako smo već napomenuli, promatramo poravnanja bez praznina, pa upit čini blok. Najprije se gradi inicijalni profil motiva kako je opisano u odjeljku (2.2)

pomoću kojeg se u prvoj iteraciji pretražuje proteom metodom opisanom u odjeljku (2.3). Time dobivamo listu pogodaka iz koje uzimamo one koji su veći od praga određenog u odjeljku (2.4) i njih spremamo u listu pozitivnih pogodaka (*engl. hits*). Pomoću njih gradi se novi profil motiva, te se proteom iznova pretražuje, ali samo na pozicijama najboljeg poklapanja određenim pomoću inicijalnog motiva u prvoj iteraciji. To rezultira novom listom pogodaka, pa otuda proizlazi i novi model. Prirodno, iterativni proces staje kad nema promjene u listi pogodaka ili kad se dosegne zadani broj iteracija.
Na kraju, pogledajmo kratak pseudokod svega dosad izloženog.
Algoritam:

- Učitaj proteom
- Zadaj motiv
- Ponavljaj dok je broj iteracija manji od $max - iter$:
  - Odredi model M na temlju motiva
  - Izračunaj ocjene na mjestima najduljih poklapanja
  - S =najveće ocjene za svaki redak proteoma
  - Nađi parametre logističke distribucije na temelju podataka S i izračunaj prag
  - Sve podstringove s ocjenama većim od praga spremi u pogotke P
  - Ako nema novih pogodaka, s obzirom na prethodnu iteraciju, prekini izvršavanje
  - Inače, motiv=P
- Kraj

Poglavlje 3

Optimizacija

U prethodnom poglavlju opisali smo postupak kojim za dani motiv u određenom proteomu (bazi podataka) tražimo njemu slične podnizove. Međutim u praksi taj postupak vremenski može potrajati, stoga u ovom poglavlju iznosimo dorađeni model koji će isti problem riješiti u kraćem vremenu.

3.1 Indeksiranje proteoma

Polazeći od ideje da skratimo vrijeme izvršavanja pojedine iteracije, uočimo da je računalno pretraživanje za određenim motivom nad većom kolekcijom slova znatno sporije nego nad kolekcijom nenegativnih cijelih brojeva. Stoga je prvi korak bio proteom kao kolekciju slova prebaciti u oblik koji sadrži samo brojeve.

Svakom retku proteoma pridružimo listu koja sadrži 20 podlista od kojih svaka predstavlja jednu od mogućih 20 aminokiselina. U i-tu liste j-te podliste upisujemo nenegativne cijele brojeve koji označavaju sve pozicije na kojima se nalazi j-ta aminokiselina u i-tom retku proteoma.
Tablicom (3.1) dana je indeksacija sljedećeg isječka iz jednog retka proteoma

$$x = \texttt{MVGGKKKTKICDKVSHEEDRISQLPEPLISEILFHLSTKDSVRTSALSTK}$$

Zbog jednostavnosti redni brojevi pozicija u nizu kreću od 0, umjesto od 1. Duljina od x je $n = 50$.

3.2 Najdulje poklapanje u svakom retku proteoma

Sljedeći korak bio je za dani motiv izračunati mjesta maksimalnih poklapanja u svakom retku protema pomoću generiranih indeksa. Neka je x_i i-ti redak u proteomu duljine n_i.

Oznaka aminokiseline	Redni broj aminokiseline u retku						
A	45						
R	19	42					
N							
D	11	18	39				
C	10						
E	22						
Q	16	17	25	30			
G	2	3					
H	15	34					
I	9	20	28	31			
L	23	27	32	35	46		
K	4	5	6	8	12	38	49
M	0						
F	33						
P	24	26					
S	14	21	29	36	40	44	47
T	7	37	43	48			
W							
Y							
V	1	13	41				

Tablica 3.1: Tablica indeksa za dio niza iz jednog retka

Njemu pridružujemo vektor h_i duljine n_i čiji elementi su nenegativni cijeli brojevi. Broj na j-tom mjestu vektora h_i označava broj simbola koji se poklapaju sa simbolima na istoj poziciji u motivu kada je motiv poravnat s podnizom retka x_i počevši od mjesta j.

Pogledajmo na primjeru kako bi to izgledalo. Neka je x kao i u prethodnom odjeljku, a za motiv uzmimo niz y = VVFGDSLSDA, duljine $m = 10$. Definiramo vektor $h = (0, 0, \ldots, 0)$ duljine 50. U motivu y na nultoj poziciji nalazi se simbol V, pa pogledamo u tablici (2.1) na kojim indeksima se on pojavljuje u x. Vidimo da su to pozicije $\{1, 13, 41\}$ pa na mjestima h_1, h_{13} i h_{41} dodajemo 1. Na mjestu y_1 opet je simbol V, pa 1 dodajemo na pozicije $1-1 = 0$, $13 - 1 = 12$ i $41 - 1 = 40$, odnosno na h_0, h_{12} i h_{40}. Sljedeći simbol je F, on se pojavljuje samo na mjestu x_{33}, pa 1 dodajemo samo na h_{31}. Pogledajmo još za 3. simbol motiva y. Tamo se nalazi simbol D koji se javlja na mjestima x_2 i x_3. Kako je $2 - 3 < 0$, 1 dodajmo samo na mjesto h_0. U konačnici dobijemo vektor:

$$h = (2, 1, 0, 0, 0, 0, 0, 2, 0, 1, 0, 0, 1, 2, 3, 0, 1, 1, 0, 0, 0, 1, 1, 0, 1,$$
$$0, 1, 0, 0, 2, 0, 2, 0, 1, 0, 2, 1, 1, 0, 1, 2, 0, 1, 0, 0, 0, 0, 0, 0, 0)$$

Na kraju, uzmemo maksimum po svim elementima h_i gjde je $i = 0, 1, \ldots, n - m$, jer nas zanimaju poravnanja samo s podnizovima duljine jednake duljini motiva. U ovom slučaju maksimum se postiže na poziciji h_{14} s vrijednosti 3, pa bi optimalno lokalno poravnanje bilo:

```
MVGGKKKTKICDKVSHEEDRISQLPEPLISEILFHLSTKDSVRTSALSTK
--------------VVFGDSLSDA--------------------------
```

Pogledajmo kako bi postupak izgledao kada se ulazni motiv sastoji od više nizova. Odnosno, imamo niz $x =$ `MASKIRKVTNQNMRINSSLSLS` i motiv

```
VVFGDSLSDA
ISFGDSIADT
LILGDSKSAG
```

Opet definiramo vektor $h = (0, 0, \ldots, 0)$ iste duljine kao i niz x. Pogledamo koja sve slova se mogu pojaviti u i-tom u danom motivu, dohvatimo njihove indekse u nizu x te na ta mjesta umanjena za i u vektor h dodajemo 1. Primijetimo kada se neko slovo u stupcu pojavljuje više puta, tada se svako njegovo pojavljivanje boduje sa 1. Vidimo da u nultom stupcu moguća slova su {`V`,`L`,`I`}. `V` se pojavljuje na mjestu x_7 pa dodajemo 1 na h_7, `L` na mjestu x_{18} pa u h_{18} dodajemo 1, a `I` na mjestima x_4 i x_{14}, što dovodi do uvećanja h_4 i h_{14} za jedan. Izračunajmo još za peti stupac, u kojem se pojavljuje samo slovo `S`. Njega nalazimo na pozicijama x_2, x_{16}, x_{17} i x_{19}, pa vektor h na mjestima h_{11}, h_{12} i h_{14} uvećamo za 3. Na kraju dobivamo:

$$h = (0, 1, 0, 1, 1, 0, 1, 1, 1, 2, 2, 3, 7, 1, 7, 3, 5, 4, 4, 0, 4, 1)$$

Dakle, najdulje poravnanje postiže se na poziciji h_{12} te samo nju pamtimo. Jasno je da pozicija h_{14} ne dolazi u obzir jer se maksimum uzima samo po indeksima manjim ili jednakim $22 - 10 = 12$.

Jasno je da vektor h, kada se uspoređivanje radi s motivom sastavljenim od jednog niza, predstavlja broj poklapanja simbola na odgovarajućim pozicijama između tih dvaju nizova. S druge strane, kad se motiv sastoji od 2 ili više nizova, vektorom h opisujemo sumu svih broja poklapanja simbola na odgovarajućim pozicijama svakog od nizova motiva s nizom x.

Poglavlje 4

Testiranje metode i rezultati

Metodu opisanu u prethodnom poglavlju testirali smo pretražujući proteom biljke *Arabidopsis thaliana* za motivom proteina koji pripadaju GDSL familiji enzima.

GDSL familija uključuje hidrolitičke enzime s multifunkcionalnim svojstvima i velikim potencijalom za primjenu u prehrambenoj i farmaceutskoj industriji. Broj proteina koji su okarakterizirani kao moguće GDSL lipaze u posljednjih je nekoliko godina naglo porastao, posebice u biljnom svijetu, što ukazuje na to da bi biljke mogle biti dobar izvor novih GDSL enzima. Tipičan niz koji pripada ovoj familiji karakteriziran je s 5 motiva, obično se gledaju manje varijabilni blokovi I, III i V, a mi radimo samo s I blokom.

Arabidopsis thaliana ili *Talijin uročnjak* mala je cvjetnica porijeklom iz Euroazije. Često je pionir kamenih, pješčanih i karbonatnih tla, te se zbog svoje rasprostranjenosti na raznim narušenim staništima smatra korovom. Ta jednogodišnja biljka relativno kratkog životnog vijeka od samo 6 tjedana je popularan modelni organizam u genetici i botanici. Pogodnom za razna istraživanja čini je i relativno kratak genom, zbog čega je bila prva biljka kojoj je sekvencioniran genom.

Proteom na kojem ispitujemo metodu ima 33410 nizova čije duljine su dane histogramom (2.2).

Ispitat ćemo osjetljivost i specifičnost testa te ih usporediti s obzirom na odabir parametra *skala*, posebno kad je ulazni motiv samo jedan redak i kad se sastoji od više redaka. Listu dobivenih pozitivnih pogodaka uspoređivat ćemo s listom 127 eksperimentalno identificiranih redaka proteoma kao GDSL enzima.

Pogledajmo prvo rezultate motiva

$$x = \begin{matrix} \texttt{VVFGDSLSDA} \\ \texttt{ISFGDSIADT} \\ \texttt{LILGDSKSAG} \end{matrix}$$

		Testiranje		
		pozitivno ocijenjeni	negativno ocijenjeni	
GDSL enzimi	jesu	TP = 120	FN = 7	osjetljivost = 0.9449
	nisu	FP = 160	TN = 33123	specifičnost = 0.9952
		PPV = 0.4286	NPV = 0.9998	

Tablica 4.1: Veličine uspješnosti testa za *skala* = 5

		Testiranje		
		pozitivno ocijenjeni	negativno ocijenjeni	
GDSL enzimi	jesu	TP = 120	FN = 7	osjetljivost = 0.9449
	nisu	FP = 41	TN = 33242	specifičnost = 0.9988
		PPV = 0.7453	NPV = 0.9998	

Tablica 4.2: Veličine uspješnosti testa za *skala* = 6

		Testiranje		
		pozitivno ocijenjeni	negativno ocijenjeni	
GDSL enzimi	jesu	TP = 116	FN = 11	osjetljivost = 0.9134
	nisu	FP = 7	TN = 33276	specifičnost = 0.9998
		PPV = 0.9431	NPV = 0.9997	

Tablica 4.3: Veličine uspješnosti testa za *skala* = 7

		Testiranje		
		pozitivno ocijenjeni	negativno ocijenjeni	
GDSL enzimi	jesu	TP = 110	FN = 17	osjetljivost = 0.8661
	nisu	FP = 0	TN = 33283	specifičnost = 1
		PPV = 1	NPV = 0.9995	

Tablica 4.4: Veličine uspješnosti testa za *skala* = 8

Evo i rezultata za motiv

$$y = \texttt{VVFGDSLSDA}$$

		Testiranje		
		pozitivno ocijenjeni	negativno ocijenjeni	
GDSL enzimi	jesu	TP = 125	FN = 2	osjetljivost = 0.9842
	nisu	FP = 174	TN = 33109	specifičnost = 0.9948
		PPV = 0.4181	NPV = 0.9999	

Tablica 4.5: Veličine uspješnosti testa za *skala* = 5

		Testiranje		
		pozitivno ocijenjeni	negativno ocijenjeni	
GDSL enzimi	jesu	TP = 122	FN = 5	osjetljivost = 0.9606
	nisu	FP = 54	TN = 33226	specifičnost = 0.9983
		PPV = 0.6816	NPV = 0.9998	

Tablica 4.6: Veličine uspješnosti testa za *skala* = 6

		Testiranje		
		pozitivno ocijenjeni	negativno ocijenjeni	
GDSL enzimi	jesu	TP = 115	FN = 12	osjetljivost = 0.9055
	nisu	FP = 12	TN = 33271	specifičnost = 0.9996
		PPV = 0.9055	NPV = 0.9996	

Tablica 4.7: Veličine uspješnosti testa za *skala* = 7

		Testiranje		
		pozitivno ocijenjeni	negativno ocijenjeni	
GDSL enzimi	jesu	TP = 91	FN = 36	osjetljivost = 0.7165
	nisu	FP = 1	TN = 33282	specifičnost = 0.9999
		PPV = 0.9891	NPV = 0.9989	

Tablica 4.8: Veličine uspješnosti testa za *skala* = 8

Rezultati pokazuju da ubrzana metoda vrlo dobro prepoznaje retke proteoma identificirane kao GDSL enzime. Možemo primijetiti da odabir parametra *skala* ponajviše ovisi o tome kakav test želimo imati. U koliko nam je bitno da je test osjetljiviji, tada smanjimo *skalu*, dok za veću pozitivnu prediktivnu vrijednost, povećamo *skalu*. Kako je broj nizova u protemu koji se mogu okarakterizirati kao GDSL enzimi vrlo mali, nema osjetne razlike u specifičnosti i negativnoj prediktivnoj vrijednosti testa kod promjene *skale*.

Bibliografija

[1] M. Cigula, *Iterativna optimizacija modela i pretraživanje proteoma*, Diplomski rad, Sveučilište u Zagrebu, Prirodoslovno-matematički fakultet, Matematički odsjek, 2016.

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

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

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

Sažetak

Cilj ovog rada bio je ubrzati proces iterativnog traženja mogućih varijanti motiva nekog proteina u danom proteomu. U tu svrhu je opisana funkcija sličnosti koju koristimo za uspoređivanje dvaju nizova te je dana distribucija najsličnijih podnizova po svakom retku proteoma. Procedura je poboljšana indeksiranjem proteoma i uvođenjem vektora koji bilježi broj mjesta poklapanja simbola na odgovarajućim pozicijama između dvaju nizova. Na kraju, vidjeli smo da metoda daje vrlo dobre rezultate testirajući je na stvarnom proteomu biljke *Arabidopsis thaliana*.

Summary

In this paper, we are concerned with optimization of iterative motif scanning. To be specific, we want to reduce processing time of finding the closest match to a given motif in each sequence of given proteome. First, we describe the scoring model and comment distribution of best score in each sequence. Then we improve the model by indexing the proteome and generating a vector that represents the number of positions at which the corresponding symbols are equal between two sequences. Finally, we have seen that this model gives very good results when applied on real-life plant proteome.

Životopis

Rođena sam 6. travnja 1992. godine u Čakovcu. Svoje obrazovanje započela sam 1999. u Osnovnoj školi "Ivana Gorana Kovačića" u mjestu Sveti Juraj na Bregu. Nakon toga 2007. godine upisujem opću gimnaziju u rodnom gradu.

Po završetku srednjoškolskog obrazovanja, 2011. godine upisujem Preddiplomski sveučilišni studij Matematika na Matematičkom odsjeku Prirodoslovno-matematičkog fakulteta u Zagrebu. Završetkom preddiplomskog studija 2014. godine stječem akademski naziv sveučilišne prvostupnice te iste godine upisujem Diplomski sveučilišni studij Matematička statistika na istom fakultetu, kojeg upravo završavam.