Neki aspekti iterativnog pretraživanja proteoma

Kobovac, Mihaela

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Mihaela Kobovac

NEKI ASPEKTI ITERATIVNOG PRETRAŽIVANJA PROTEOMA

Diplomski rad

Voditelj rada:
doc. dr. sc. Pavle Goldstein

Zagreb, veljača, 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 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 posljednjih par desetljeća došlo je do ogromnog napredovanja u polju molekularne biologije, genetičkog inženjerstva i biotehnologije. To je dovelo do naglog porasta potrebe za analitičkom obradom bioloških podataka dobivenih znanstvenim istraživanjima iz tih područja. Istovremeno, došlo je do velikog tehnološkog napretka, a samim time i razvoja računalnih znanosti. Tako se razvila i bioinformatika koja predstavlja integraciju matematičkih, statističkih i kompjuterskih metoda u svrhu analize molekularno-bioloških, biokemijskih i biofizičkih podataka. Računala omogućuju stvaranje velikih baza podataka te organiziranje i brzu analizu tih podataka raznim alatima.

Jedno od pitanja bioinformatike jest kako u proteomu (skupu proteina) nekog organizma pronaći proteine iz određene proteinske familije. Time ćemo se baviti u ovom radu. Pokušati ćemo naći blokove aminokiselina najsličnije zadanom motivu, koji je karakterističan za određenu familiju, i potom procijeniti je li ta sličnost "dovoljno velika".

Ovaj rad podijeljen je u četiri poglavlja. U prvom poglavlju objašnjeni su pojmovi iz vjerojatnosti i statistike koji su fundamentalni za razumijevanje ovog rada. U drugom poglavlju detaljnije se objašnjava problem kojim ćemo se baviti. Opisuje se način traženja blokova aminokiselina u proteomu sličnih zadanom motivu, kao i princip ocjenjivanja te sličnosti. Potom je objašnjeno kako zaključujemo pripada li protein zadanoj proteinskoj familiji, te kako iterativno ponavljamo postupak ne bismo li dobili što točnije rezultate. U trećem poglavlju objašnjena je modifikacija metode građenja profila motiva te pokušaj ubrzanja pretraživanja proteoma. Naposljetku, objašnjena metoda je u zadnjem poglavlju primijenjena na proteom biljke Arabidopsis thaliana, te iznosimo detaljnu analizu dobivenih rezultata.

Poglavlje 1

Pojmovi iz vjerojatnosti i statistike

1.1 Vjerojatnost

Definicija 1.1.1. *Familija $\mathcal{F}$ podskupova od Ω ($\mathcal{F} \subset \mathcal{P}(\Omega)$) jest **$\sigma$-algebra skupova** (na Ω) ako je:*

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

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

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

Definicija 1.1.2. *Neka je Ω proizvoljan neprazan skup, a $\mathcal{F}$ σ-algebra na skupu Ω. Uređeni par $(\Omega, \mathcal{F})$ zove se **izmjeriv prostor**. Funkcija $\mathbb{P}\colon \mathcal{F} \to \mathbb{R}$ je **vjerojatnost** na $\mathcal{F}$ ako vrijedi:*

(i) $\mathbb{P}(A) \geq 0$, $A \in \mathcal{F}$

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

(iii) $A_i \in \mathcal{F}, i \in \mathbb{N}$ i $A_i \cap A_j = \emptyset$ za $i \neq j \Rightarrow \mathbb{P}(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$

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

Definicija 1.1.3. *Neka je $(\Omega, \mathcal{F}, \mathbb{P})$ proizvoljan vjerojatnosni prostor i $A \in \mathcal{F}$ takav da je $\mathbb{P}(A) > 0$. Definiramo funkciju $P_A(B)\colon \mathcal{F} \to [0, 1]$ na sljedeći način:*

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

Lako je provjeriti da je P_A vjerojatnost na $\mathcal{F}$ i nju zovemo ***uvjetna vjerojatnost uz uvjet A****, a $P(B|A)$ zovemo vjerojatnost od B uz uvjet A.*

Definicija 1.1.4. *Neka je $\mathcal{B}$ σ-algebra generirana familijom svih otvorenih skupova na $\mathbb{R}$. $\mathcal{B}$ zovemo* ***Borelova σ-algebra*** *skupova na $\mathbb{R}$, a elemente te σ-algebre zovemo* ***Borelovi skupovi****.*

Definicija 1.1.5. *Neka je $(\Omega, \mathcal{F}, \mathbb{P})$ vjerojatnosni prostor. Funkcija $X\colon \Omega \to \mathbb{R}$ 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 1.1.6. *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:*

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

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 slučajne varijable*** *ako za proizvoljne $B_i \in \mathcal{B}, i = 1, 2, \ldots, n$ vrijedi:*

$$\mathbb{P}\Big(\bigcap_{i=1}^{n}(X_i \in B_i)\Big) = \prod_{i=1}^{n} \mathbb{P}(X_i \in B_i). \tag{1.3}$$

1.2 Funkcija distribucije

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

$$\mathbb{P}_X(B) = \mathbb{P}(X^{-1}(B)) = \mathbb{P}(\omega \in \Omega\colon X(\omega) \in B) = \mathbb{P}(X \in B). \tag{1.4}$$

Time je definirana funkcija $\mathbb{P}_X\colon \mathcal{B} \to [0, 1]$, koja je vjerojatnosna mjera na $\mathcal{B}$ i zovemo ju vjerojatnost inducirana slučajnom varijablom X. Svakoj slučajnoj varijabli X je preko relacije (1.4) na prirodan način pridružen vjerojatnosni prostor $(\mathbb{R}, \mathcal{B}, \mathbb{P}_X)$ **induciran slučajnom varijablom X**. Problemi vezani uz slučajnu varijablu X rješavaju se u okviru tog vjerojatnosnog prostora.

Definicija 1.2.1. ***Funkcija distribucije*** *slučajne varijable X jest funkcija $F_X = F\colon \mathbb{R} \to [0, 1]$ definirana s:*

$$F(x) = \mathbb{P}_X(< -\infty, x]) = \mathbb{P}\{X \leq x\} = \mathbb{P}\{\omega\colon X(\omega) \leq x\}, x \in \mathbb{R}.$$

Postoje dvije glavne vrste slučajnih varijabli: diskretne i neprekidne.

Definicija 1.2.2. *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 1.2.3. *Kažemo da je slučajna varijabla X **apsolutno neprekidna** ili, kraće, **neprekidna slučajna varijabla** ako postoji nenegativna Borelova funkcija f na $\mathbb{R}$ takva da je*

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

Za funkciju distribucije oblika (1.5), odnosno za funkciju distribucije F_X slučajne varijable X kažemo da je apsolutno neprekidna funkcija distribucije. U tom slučaju se funkcija f iz (1.5) zove funkcija gustoće vjerojatnosti od X.

Uvedimo pojam matematičkog očekivanja slučajne varijable X.

Definicija 1.2.4. *Neka je X diskretna slučajna varijabla i neka je skup D iz definicije diskretne slučajne varijable, $D = \{x_1, x_2, \ldots\}$ i neka za svako k vrijedi $\mathbb{P}(\{x_k\}) = p_k$. Tada je **očekivanje diskretne slučajne varijable** X dano s:*

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

Definicija 1.2.5. *Neka je X neprekidna slučajna varijabla s funkcijom distribucije F_x. **Očekivanje neprekidne slučajne varijable** X dano je sljedećom relacijom:*

$$F_x(x) = \int_\Omega X d\mathbb{P} = \int_{\mathbb{R}} x dF_x(x).$$

Definicija 1.2.6. *Neka je $g\colon \mathbb{R} \to \mathbb{R}$ Borelova funkcija. Vrijedi*

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

Kako bi uveli pojam varijance slučajne varijable X, za $r > 0$ definiramo r-ti centralni moment od X.

Definicija 1.2.7. *Neka $\mathbb{E}(X)$ postoji. Tada$\mathbb{E}[(X - \mathbb{E}(X))^r]$ zovemo r-ti centralni moment od X.*

Definicija 1.2.8. *Drugi centralni moment od X zovemo **Varijanca** od X, i označavamo je s VarX ili σ_x^2.*

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

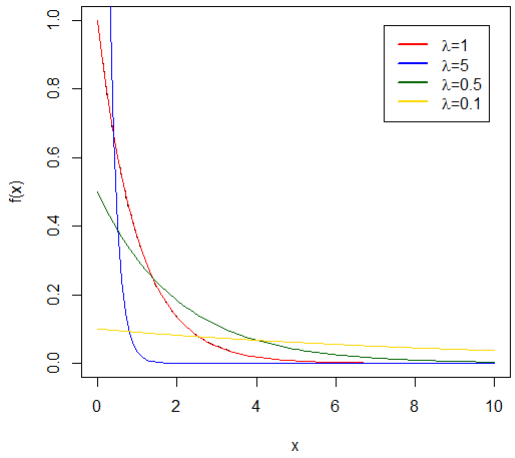
*Pozitivan drugi korijen iz varijance zovemo **standardna devijacija** od X i označavamo s σ_x*

1.3 Primjeri slučajnih varijabli

Eksponencijalna distribucija

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

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



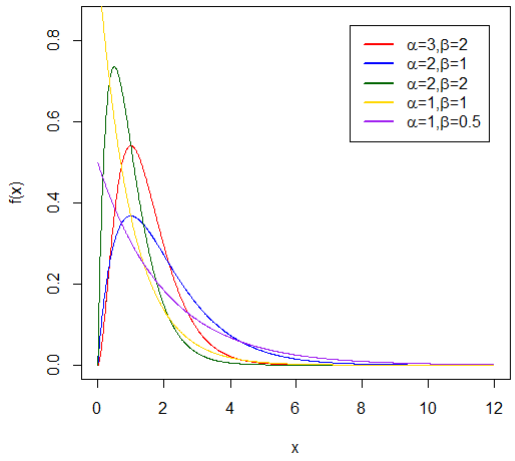
Slika 1.1: Graf funkcije gustoće eksponencijalne distribucije za razne vrijednosti λ

Gama 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 varijabla X ima **gama distribuciju** s parametrima α i β ako joj je funkcija gustoće f dana s:

$$f(x) = \begin{cases} \frac{1}{\gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{\frac{-x}{\beta}}, & \text{if } x > 0 \\ 0, & x \leq 0. \end{cases}$$

Ako je $\alpha = 1$ i $\beta = \frac{1}{\lambda}$,uočimo da tada X ima **eksponencijalnu distribuciju** s parametrom λ.



Slika 1.2: Graf funkcije gustoće gama distribucije za razne vrijednosti α i β

Logistička distribucija

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

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

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

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

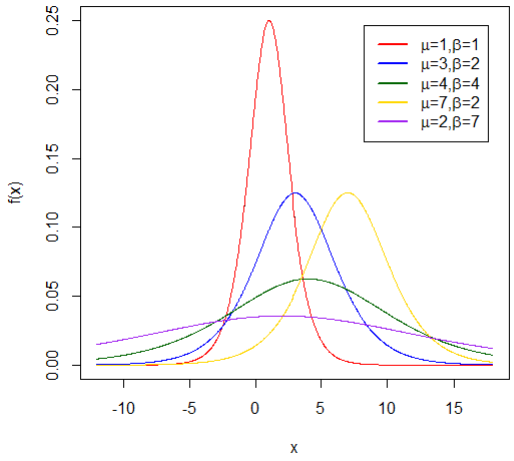
Pojmovi iz vjerojatnosti i statistike većinom su preuzeti iz [3] te se tamo mogu naći detaljnija objašnjanja i primjeri.

1.4 Teorija ekstremnih vrijednosti

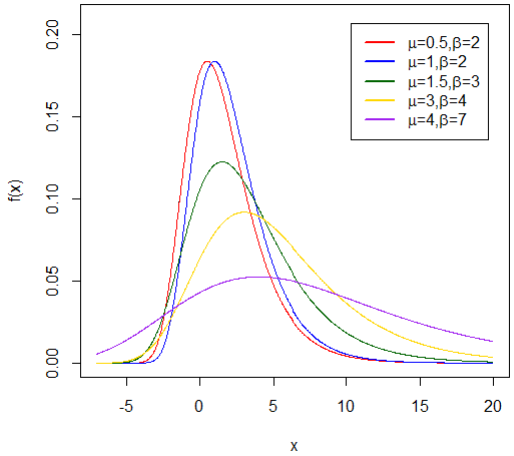
Gumbel distribucija

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

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



Slika 1.3: Graf funkcije gustoće logističke distribucije za razne vrijednosti μ i β



Slika 1.4: Graf funkcije gustoće Gumbel distribucije za razne vrijednosti μ i β

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

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

Korolar 1.4.1. *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:

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

Weibullova distribucija

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

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

Teorem 1.4.2. *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 $b_n \in \mathbb{R}$ takvi da je $\lim_{n\to\infty} \mathbb{P}\left(\frac{M_n - b_n}{a_n} \leq x\right) = F(x)$, gdje je F nedegenerirana distribucija, tada granična distribucija F pripada Gumbel, Fréchet ili Weibull distribuciji.*

1.5 Specifičnost i osjetljivost

Specifičnost i osjetljivost testa glavne su statističke mjere koje mjere uspješnost provedenog testa. **Specifičnost** (vjerojatnost detekcije negativnih) mjeri koji postotak negativnih je testom prepoznat kao negativan.

$$\begin{aligned} \text{specifičnost} &= \frac{\text{broj stvarno negativnih}}{\text{broj stvarno negativnih+broj lažno pozitivnih}} \\ &= \frac{\text{TN}}{\text{N}} = \frac{\text{TN}}{\text{TN+FP}} \end{aligned}$$

Osjetljivost (vjerojatnost detekcije pozitivnih) mjeri koji postotak pozitivnih je zaista prepoznat kao pozitivan.

$$\begin{aligned} \text{osjetljivost} &= \frac{\text{broj stvarno pozitivnih}}{\text{broj stvarno pozitivnih+broj lažno negativnih}} \\ &= \frac{\text{TP}}{\text{P}} = \frac{\text{TP}}{\text{TP+FN}} \end{aligned}$$

		predviđeno stanje		
	ukupna populacija	ocjenjeni pozitivno	ocjenjeni negativno	
stvarno stanje	pozitivno stanje	TP (stvarno pozitivni)	FN (lažno negativni)	osjetljivost
	negativno stanje	FP (lažno pozitivni)	TN (stvarno negativni)	specifičnost
		PPV (pozitivna prediktivna vrijednost)	NPV (negativna prediktivna vrijednost)	

Tablica 1.1: Tablica točnosti

Vrlo bitne mjere u ocjenjivanju testa također su pozitivna prediktivna vrijednost te negativna prediktivna vrijednost. **Pozitivna prediktivna vrijednost (PPV)** otkriva koliki postotak onih koji su proglašeni pozitivnim su zaista pozitivni dok **negativna prediktivna vrijednost (NPV)** otkriva koliki postotak onih koji su proglašeni negativnim su zaista negativni.

$$\begin{aligned} \text{PPV} &= \frac{\text{broj stvarno pozitivnih}}{\text{broj stvarno pozitivnih+broj lažno pozitivnih}} \\ &= \frac{\text{TP}}{\text{TP+FP}} \\ \text{NPV} &= \frac{\text{broj stvarno negativnih}}{\text{broj stvarno negativnih+broj lažno negativnih}} \\ &= \frac{\text{TN}}{\text{TN+FN}} \end{aligned}$$

Poglavlje 2

Problem

2.1 Uvod u biološke pojmove

Proteini ili bjelančevine fundamentalni su sastavni dijelovi svakog organizma. Od proteina su građeni mišići, krv, koža, kosa, nokti pa i unutarnji organi. Oni su sastavni dijelovi svake stanice što ih čini osnovom života na Zemlji. Skup svih proteina nekog organizma (ili stanice) koji nastaju kao posljedica ekspresije gena u određenom trenutku u vremenu pod određenim uvjetima zovemo proteom. Za razliku od genoma, proteom nije statičan, već varira s obzirom na vrstu stanice, stupanj razvoja organizma, kao i utjecaj okoliša.

Poteini su izgrađeni od 20 različitih standardnih aminokiselina međusobno povezanih poput karika u lancu. Građa proteina određuje specifične osobine svakog proteina. U tablici 2.1 dane su standardne aminokiseline.

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

Tablica 2.1: Standardne aminokiseline

Motiv proteinskog niza je kratak niz aminokiselina, u pravilu 5 do 20, koji je ostao sačuvam selekcijskim pročišćivanjem i ima neko biološko značenje.

2.2 Opis problema

Za protein od interesa postoji popis nekih njegovih dijelova koji su bolje očuvani, tzv. motiva, specifičnih za njega, koji se koristi u detektiranju homolognih proteina u nekom novom organizmu. Sličnost nizova aminokiselina u proteomima upućuje na zajedničko podrijetlo, a samim time i na sličnost biološke funkcije, zbog čega je od velikog interesa unaprijediti računalne tehnike pronalaženja sličnih nizova u proteomima novih organizama, kako bi se lakše odredila njihova funkcija. Kako bismo proširili popis proteina specifičnih za određenu proteinsku familiju pretražujemo proteom nekog organizma i pokušavamo te proteine prepoznati. Budući da se genetički materijal živih bića kroz generacije mijenja, tj. dolazi do mutacija, za zadane nizove aminokiselina pokušat ćemo u proteomu nekog organizma naći najsličnije.

Kada nađemo niz aminokiselina sličan zadanome motivu, zanimat će nas da li je ta sličnost značajna, tj. da li je ta sličnost dovoljna da bismo zaključili da je protein u kojem smo ga našli iz proteinske familije od interesa. Ako ustanovimo da sličnost nije slučajna, dodat ćemo novi niz u popis varijanti nizova specifičnih za danu proteinsku familiju. Takvim pristupom mogu se izbjeći skupi i dugotrajni eksperimenti u kojima bi se izravno utvrđivala funkcija pojedinog proteina iz proteoma novog organizma koji je predmet proučavanja.

2.3 Pretraživanje proteoma

Kao što je opisano u prethodnom poglavlju, neka nam je dan niz slučajnih varijabli $x = x_1x_2 \dots x_n$, tzv. ulazni motiv. Želimo naći njemu najsličniji niz aminokiselina duljine n u svakom proteinu, tj. u svakom retku danog proteoma. Tražiti ćemo samo nizove jednakih duljina kao motiv, tj. nećemo dozvoljavati inserciju i deleciju. Stoga u nizovima koje ćemo uspoređivati sa zadanim motivom neće biti praznina '-' koje bi nastale kada oni ne bi bili jednakih duljina kao zadani motiv.

Opisat ćemo postupak traženja niza najsličnijeg motivu za jedan protein $y = y_1y_2 \dots y_m$. Analogno postupamo za svaki redak danog proteoma. Metodom klizećeg prozora (engl. sliding window) usporedit ćemo naš motiv x sa svakim od blokova duljine n u nizu y počevši od 1. pozicije do $(m - n + 1)$-ve pozicije. Radi lakšeg shvaćanja, grafički ćemo prikazati princip uspoređivanja.

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

Uočimo da je nužno da bude $m \geq n$.

2.4 Ocjena sličnosti

U sljedećem odlomku objasnit ćemo princip uspoređivanja niza aminokiselina duljine kao zadani motiv s tim motivom, te način ocjenjivanja njegove sličnosti motivu. Pretpostavimo da nam je dan motiv $x = x_1x_2 \dots x_n$ duljine n, kao gore. Želimo s njim usporediti niz $z = z_1z_2 \dots z_n$ jednake duljine. Neka je npr. motiv zadan s 3 niza aminokiselina: "FVFGDSLVDN" , "AVLGDSLLFF" i "LFLGDFNVDK". Neka je z="VVFGDSVDDF". Dakle, za niz aminokiselina z želimo izračunati njegovu sličnost zadanom motivu x. Tu ocjenu sličnosti označit ćemo sa S (engl. score). Što je ocjena sličnosti veća, to je niz sličniji danom motivu i veća je vjerojatnost da protein u kojem smo ga našli pripada zadanoj proteinskoj familiji.

Najprije ćemo objasniti slučajni model R (engl. random) koji pretpostavlja da dva niza nisu povezana, tj. nezavisnost događaja x i z. Napomenimo da u oba modela koja ćemo opisati pretpostavljamo da je pojavljivanje bilo koje aminokiseline u proteinu, tj. karike u lancu, neovisno o prethodnoj aminokiselini u proteinu. Štoviše, model R pretpostavlja da se slučajni događaj z_1 dogodi neovisno o tome što je u motivu na prvoj poziciji, i tako analogno dalje. Vjerojatnost bilo koje od standardnih aminokiselina bit će određena neovisno o poziciji sljedećom distribucijom:

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

Neka je $q = (0.078, 0.051, 0.043, 0.053, 0.019, 0.043, 0.063, 0.072, 0.023, 0.053, 0.091, 0.059, 0.022, 0.039, 0.052, 0.068, 0.059, 0.014, 0.032, 0.066)$.

Taj vjerojatnosni vektor dobiven je računanjem relativnih frekvencija aminokiselina u proteomima većeg skupa organizama. Označimo s q_a vjerojatnost da se aminokiselina a pojavi u nizu. Slijedi da je:

$$\mathbb{P}(z|R) = \prod_{i=1}^{n} q_{z_j}. \tag{2.1}$$

Opišimo sada drugi model M (engl. match) koji pretpostavlja zavisnost aminokiselina na istim pozicijama. Dakle u tom modelu računamo vjerojatnost da varijabla Z proizvede događaj z koji na i-toj poziciji ima z_i uz uvjet x, odnosno x_i. U tom slučaju vjerojatnost događaja z dana je s :

$$\mathbb{P}(z|M_x) = \prod_{i=1}^{n} \mathbb{P}_{x_i}(z_i). \tag{2.2}$$

Objasnit ćemo na koji se način računaju vjerojatnosti u (2.2). Krenimo od ulaznog motiva, tj. zadanog niza ili nizova aminokiselina jednake duljine. Za svaki stupac danog motiva računamo relativne frekvencije aminokiselina iz $\mathcal{A}$. Dakle, za i-ti stupac motiva dobijemo empirijsku distribuciju $f_i = (f_{i_1}, f_{i_2}, \ldots, f_{i_{20}})$. Vjerojatnosti zapisujemo redoslijedom kojim su dane u q. Kako bismo ispravili mogući nedostatak nezavisnosti u ulaznom motivu na svaki od $f_i, i = 1, 2, \ldots, n$ primjenjujemo blagu težinsku shemu. To ima smisla samo u slučaju da se ulazni motiv sastoji od više nizova aminokiselina na istoj poziciji. Ukratko, ukoliko se ulazni motiv sastoji od dva ista niza aminokiselina i jednog različitog, kod zbrajanja apsolutnih frekvencija nećemo reći da su frekvencije aminokiselina koje su se pojavile na istoj poziciji 1 pri svakom pojavljivanju, nego npr. 0.8. Na taj način će se uništiti ova naznaka zavisnosti pri pojavljivanju istih aminokiselina. Označimo s *nos* (engl. number of sequences) broj nizova aminokiselina koji čine motiv. Budući da se ulazni motiv najčešće sastoji od svega nekoliko nizova aminokiselina, vjerojatnosti u f_i će za mnoge aminokiseline biti 0. Kako bismo to ispravili dodajemo mali pseudozbroj (koji ovisi o broju nizova) i definiramo novu funkciju distribucije $g_i = (g_{i_1}, g_{i_2}, \ldots, g_{i_{20}}), \forall i \in 1, 2, \ldots, n$ na sljedeći način:

$$g_{ij} = \frac{f_{ij} + \frac{0.01}{nos}}{1 + \frac{0.2}{nos}}. \tag{2.3}$$

Označimo s A tranzicijsku matricu PAM koja sadrži vjerojatnosti da pojedina aminokiselina mutira u drugu aminokiselinu za aminokiseline iz $\mathcal{A}$. Dakle, to je stohastička matrica, tj. vrijedi $\sum_{j=1}^{20} a_{i_j} = 1, \forall i \in \{1, 2, \ldots, n\}$. Neka je sada $B = (b_{i_j}) = A^k$, gdje je k dosta velik (mi uzimamo $k = 120$). Uočimo da vektor redak $b_i = (b_{i_1}, b_{i_2}, \ldots, b_{i_{20}})$ predstavlja očekivani vektor mutacije i-te aminokiselinke nakon k milijuna godina evolucije. Konačno, vektor $(p_{i_1}, p_{i_2}, \ldots, p_{i_{20}})$ definiramo kao linearnu kombinaciju vektora $b_1, b_2, \ldots, b_{20}$ s koeficijentima $g_{i_1}, g_{i_2}, \ldots, g_{i_{20}}$ za $i = 1, 2, \ldots, n$:

$$p_i = \sum_{l=1}^{20} g_{i_l} b_l. \tag{2.4}$$

Dakle, dobili smo n funkcija distribucije $p_1, p_2, \ldots, p_n$ takvih da je

$$p_{i_j} = \sum_{k=1}^{20} g_{i_k} b_{k_j}, i = 1, 2, \ldots, n$$

vjerojatnost da se u i-tom stupcu dogodi j-ta aminokiselina. (Vjerojatnosti u p_i možemo shvatiti na sljedeći način: p_{i_j} je korigirana empirijska vjerojatnost da padne aminokiselina A $\times$ vjerojatnost da A mutira u j-tu aminokiselinu + vjerojatnost da padne aminokiselina C $\times$ vjerojatnost da C mutira u j-tu aminokiselinu itd.) Razmatrajući gornji opis računanja distribucije za model M, dolazimo do zaključka da je model M zapravo niz onoliko modela koliki je broj stupaca u ulaznom motivu, tj. $M = (M_1, M_2, \ldots, M_{20})$, jer se za svaki stupac posebno računa funkcija distribucije. Dakle, formula (2.2) preciznije se može izraziti na sljedeći način:

$$\mathbb{P}(z|M_x) = \prod_{i=1}^{n} \mathbb{P}(z_i|M_{x_i}) = \prod_{i=1}^{n} p_{i_{z_i}} \tag{2.5}$$

Vraćamo se na računanje ocjene sličnosti. Cilj je dakle nizu aminokiselina dodijeliti ocjenu kojom iskazujemo u kojoj mjeri je on povezan s motivom u odnosu na to koliko nije. Sada kada smo objasnili modele u oba slučaja, promotrimo njihov omjer. Omjer formula (2.2) i (2.1) zovemo omjer šansi (engl. odds ratio) i on izgleda ovako:

$$\frac{\mathbb{P}(z|M)}{\mathbb{P}(Z|R)} = \frac{\prod_{i=1}^{n} p_{i_{z_i}}}{\prod_{i=1}^{n} q_{z_i}} = \prod_{i=1}^{n} \frac{p_{i_{z_i}}}{q_{z_i}}.$$

Budući da baratamo s vrlo malim brojevima, gornji omjer još logaritmiramo pa dobivamo:

$$S = \sum_{i=1}^{n} s(x_i, z_i), \text{ gdje je } s(x_i, z_i) = \log \frac{p_{i_{z_i}}}{q_{z_i}}. \tag{2.6}$$

Ocjene $s(x_i, z_i)$ zapisujemo u tzv. PSSM (engl. position specific scoring matrix), tj. matricu ocjena. Nakon što je objašnjen princip ocjenjivanja sličnosti ulaznog motiva s nizom aminokiselina jednake duljine vraćamo se na problem nalaženja niza u proteinu najsličnijeg zadanom motivu. Formulu (2.6) evaluiramo za nizove $y_k y_{k+1} \ldots y_{k+n-1}$ za $k = 1, 2, \ldots, m - n + 1$, tj. računamo:

$$s_k = \sum_{i=0}^{n-1} \log \frac{\mathbb{P}(y_{k+1}|M_i)}{\mathbb{P}(y_{k+1}|R)}, k = 1, 2, \ldots, m - n + 1 \tag{2.7}$$

Budući da tražimo najsličniji niz zadanome motivu, želimo pronaći niz s maksimalnom ocjenom, tj. tražimo

$$S = \max_{k=1,2,\ldots,m-n+1} s_k$$

za svaki redak proteoma.

2.5 Ubrzanje pretraživanja

Bitno je napomenuti da proteomi sadrže jako velik broj proteina, a svaki protein je opet sačinjen od velikog broja aminokiselina. Primjerice proteom biljke Arabidopsis thaliana (Talijin uročnjak) sastoji se od oko 33000 proteina prosječne duljine 400. Metodom klizećeg prozora u svakom retku proteina, ako pretpostavimo da je motiv duljine 10, usporediti ćemo motiv s blokom aminokiselina na 391 poziciji. To će značiti da ćemo otprilike 13 milijuna puta ($\approx 33000\times391$) računati formulu (2.7), što će, bez obzira na tehnološki napredak, dugo trajati. Kako bi se izvođenje programa značajno ubrzalo "diskretizirat" ćemo model. Preciznije, smanjit ćemo broj pozicija na kojima ćemo računati formulu (2.7), čime ćemo značajno ubrzati traženje sličnih varijanti zadanog motiva, ali pronalazak optimalnog rješenja nije zajamčen. Uočimo da je naš problem traženja blokova aminokiselina u proteomu najsličnijih zadanome motivu sličan traženju riječi u tekstu, ako uzmemo u obzir da je moglo doći do grešaka prilikom pisanja ili je riječ u tekstu u obliku drugačijem od zadanog. Ako je dan ulazni string $x = x_1x_2 \dots x_n$ duljine n i tekst $y = y_1y_2 \dots y_m$, $(m \geq n)$ definirajmo udaljenost od x do podstringa od y duljine n kao broj pozicija na kojima se taj podstring razlikuje od zadane riječi x. Označimo tu udaljenost s

$$d(y_{k+1}y_{k+2} \dots y_{k+n}, x), \quad k = 0, 1, \dots m - n.$$

Zovemo je Hammingova udaljenost. Što je udaljenost manja, to su stringovi sličniji. Najčešće je cilj naći sve podstringove koji zadovoljavaju uvjet da je udaljenost manja od neke konstante $l \in \mathbb{Z}$, tj.

$$d(y_{k+1}y_{k+2} \dots y_{k+n}, x) < l.$$

Vratimo se na problem pretraživanja niza aminokiselina (retka proteoma) kako bismo našli blok najsličniji ulaznom motivu. Definiramo funkciju sličnosti sličnu Hammingovoj udaljenosti. Razlika je u tome što vrijednost funkcije sličnosti definiramo kao broj pozicija na kojima su odgovarajući simboli jednaki, a ne različiti kao kod Hammingove udaljenosti. Dakle, definirali smo funkciju koja nizu aminokiselina duljine kao ulazni motiv dodijeli nenegativan cijeli broj. Za svako podudaranje u simbolima na istoj poziciji poveća se vrijednost funkcije za 1. Dakle, što je vrijednost funkcije sličnosti veća, to je niz aminokiselina iz proteina sličniji ulaznom motivu. Primjerice, funkcija sličnosti niza aminokiselina z="VVFGDSVDDF" s motivom x="FVFGDSLVDN" iznosit će 6, jer je na 6 pozicija isti znak.

motiv	F	V	F	G	D	S	L	V	D	N
niz	V	V	F	G	D	S	V	D	D	F

Ako je motiv zadan s više nizova aminokiselina, npr. motiv x se sastoji od nizova "FVFGDSLVDN","AVLGDSLLFF" i "LFLGDFNVDK", sličnost niza z="VVFGDSVDDF" s tim motivom iznosti će 13.

motiv	F	𝕍	𝔽	𝔾	𝔻	𝕊	L	V	𝔻	N
motiv	A	𝕍	L	𝔾	𝔻	𝕊	L	F	L	𝔽
motiv	L	F	L	𝔾	𝔻	F	N	V	𝔻	K
niz	V	V	F	G	D	S	V	D	D	F

Za svaki redak proteoma i za svaki blok duljine n ćemo na ovaj način izračunati njegovu vrijednost funkcije sličnosti. U svakom retku će nas najviše zanimati blokovi s najvećom vrijednosti funkcije sličnosti. Ukoliko je najveća vrijednost funkcije sličnosti u danom retku, koju označavamo s d_{max}, veća ili jednaka $\frac{n}{2}$ pamtimo sve pozicije na kojima je vrijednost funkcije sličnosti veća ili jednaka $\frac{n}{2}$. Ako je pak

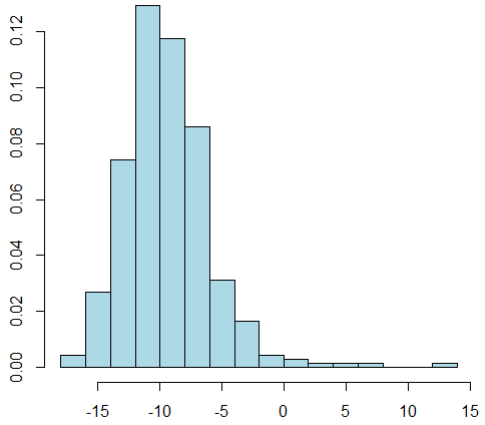
$$d_{max} < \frac{n}{2},$$

pamtimo sve pozicije na kojima se postiže ta maksimalna vrijednost funkcije sličnosti. Ocjenu sličnosti niza s motivom potom ćemo računati samo na zapamćenim pozicijama. To će značiti da ćemo na puno manje blokova aminokiselina računati (2.7) i znatno ubrzati proces traženja. Više o tome u [2].

2.6 Značajnost ocjene

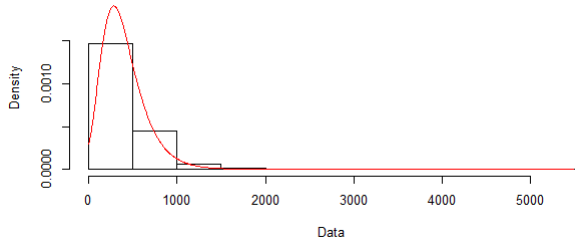
Iterativno ćemo graditi profil motiva. Preciznije, za dobivene nizove aminokiselina s najvećom ocjenom u svakom retku proteoma zanimati će nas koji od tih nizova su dovoljno slični motivu. Onim nizovima za koje zaključimo da jesu dovoljno slični nadograditi ćemo profil motiva te ponoviti postupak traženja i ocjenjivanja na temelju novo dobivenog motiva.

Postavlja se pitanje kako odrediti da li je sličnost niza aminokiselina zadanom motivu dovoljno velika da bismo zaključili da je pripadajući protein iz proteinske familije od interesa. Zapravo, ono što trebamo odrediti jest koje su od maksimalnih ocjena sličnosti dovoljno značajne. Jednostavnije rečeno, dobili smo oko 33000 brojeva i zanima nas koji od njih su dovoljno veliki. Kako bismo to mogli, najprije moramo odrediti zakon distribucije tih brojeva. Uočimo da zapravo želimo proučiti desni rep funkcije distribucije ocjena poravnanja po svim pozicijama za svaki pojedini redak proteoma posebno budući da će se maksimalne ocjene nalaziti upravo u desnom repu. Promotrimo histogram ocjena sličnosti koje su zapravo logaritmi omjera šansi po svim mogućim pozicijama u jednom retku proteoma. Uočimo iz histograma 2.1 da te ocjene imaju pomaknutu gama distribuciju s eksponencijalnim repom, a rep je upravo ono što nas zanima budući da su tu maksimalne ocjene.



Slika 2.1: Histogram ocjena sličnosti za jedan protein

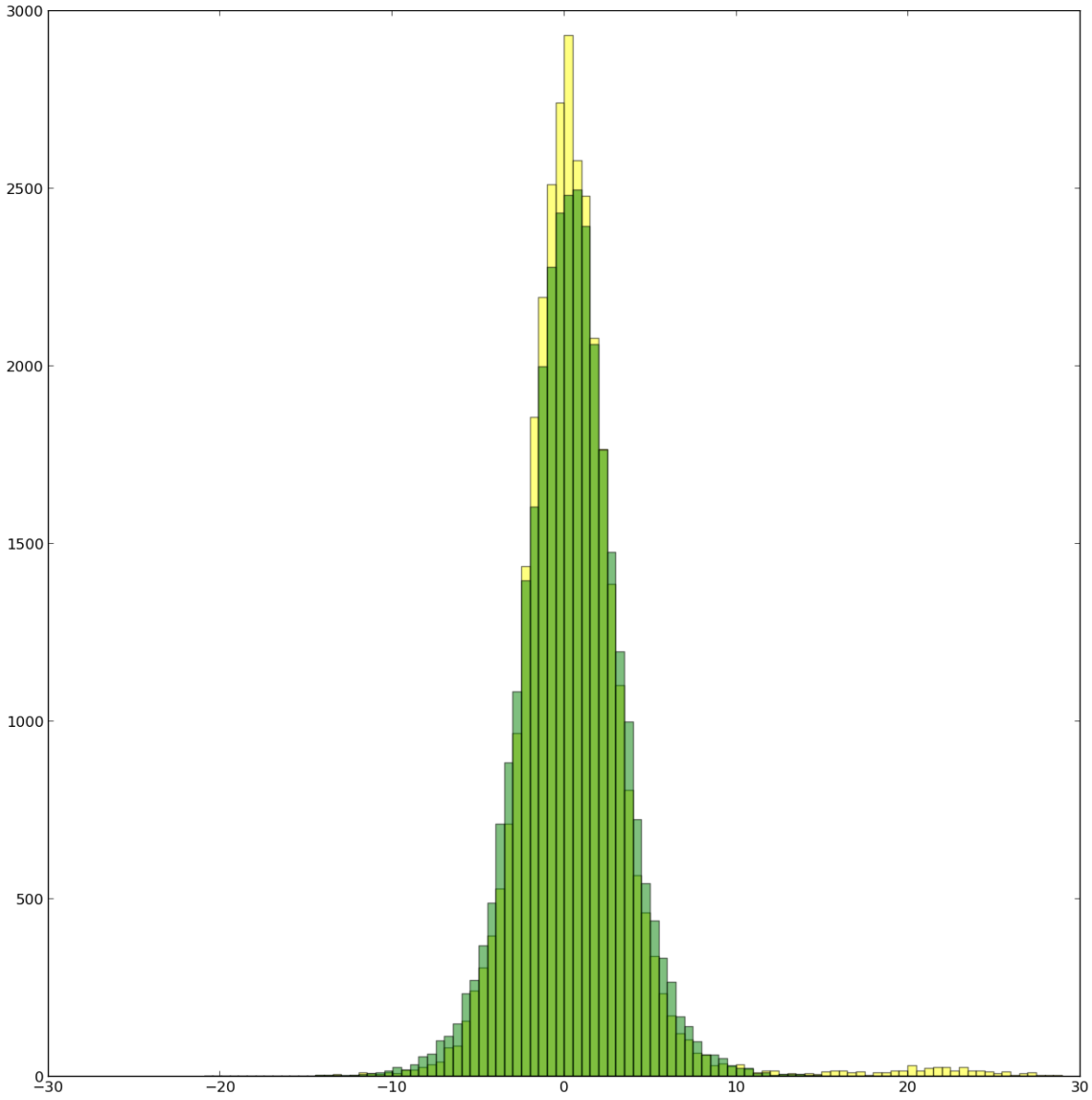
Dobro je poznato da maksimumi nezavisnih jednako distribuiranih eksponencijalnih slučajnih varijabli imaju Gumbel distribuciju. To se može lako provjeriti simulacijama. Više o tome u [4]. No, budući da nizovi aminokiselina u proteomima nisu jednakih duljina, ne možemo pretpostaviti da su eksponencijalne slučajne varijable jednako distribuirane. Očito je da, što je redak proteoma dulji, to je veća vjerojatnost da ćemo pronaći sličniji niz, tj. onaj s većom ocjenom sličnosti. Isto tako, što je motiv kraći, veće su šanse da ćemo naći sličniji niz. Međutim, to nam ne predstavlja problem budući da u svakom retku tažimo najsličnijeg istom, unaprijed zadanom motivu. Zbog navedenog želimo doznati distribuciju duljina redaka proteoma.



Slika 2.2: Histogram duljina proteina u proteomu

Na slici 2.2 prikazan je histogram duljina proteina u proteomu nekog organizma zajedno s funkcijom gustoće Gumbel distribucije. Uočavamo da bi duljine nizova mogle

pratiti Gumbelovu distribuciju. Dakle, baratamo s dvije slučajne varijable s Gumbelovim distribucijama, Iz poglavlja 1.4 slijedi da razlika dviju takvih slučajnih varijabli slijedi logističku distribuciju.



Slika 2.3: Histogram maksimalnih ocjena i logističke distribucije

To se iz histograma 2.3 vrlo dobro vidi. Maksimalne ocjene za retke proteina u prote-

omu zaista slijedi histogram logističke distribucije. Više o distribuciji maksimalnih ocjena sličnosti u [1].

Utvrdili smo da su maksimalne ocjene po nizovima logistički distribuirane, te možemo utvrditi na koji način ćemo odlučiti koje su od tih ocjena statistički značajne. Zapravo ćemo testirati nul-hipotezu da je sličnost niza aminokiselina slučajna naspram alternativne hipoteze da nije slučajna, tj. da je odgovarajući protein zaista iz te proteinske familije. Testna statistika je ocjena poravnanja koju smo definirali u (2.7). P-vrijednost testa je vjerojatnost da, ako vrijedi nulta hipoteza, dobijemo broj veći ili jednak testnoj statistici, tj. da je ocjena poravnanja veća ili jednaka opaženoj. Ako dobijemo malu p-vrijednost odbacujemo nultu hipotezu o nepovezanosti nizova. Želimo odrediti prag od kojeg nadalje ćemo maksimalne ocjene smatrati dovoljno velikim da bismo zaključili da su nizovi zaista biološki povezani. Parametar β logističke distribucije možemo izraziti na sljedeći način:

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

tj. pomoću standardne devijacije.

Prag ćemo definirati na način da se za određeni broj $\frac{\sqrt{3}}{\pi}\sigma$ udaljimo od prosječne ocjene. Taj broj koliko puta ćemo se odmaknuti za $\frac{\sqrt{3}}{\pi}\sigma$ zovemo skala i određujemo je ovisno o tome koliko želimo da test bude specifičan i senzitivan. Dakle, prag smo definirali na sljedeći način: $prag = \mu + skala \times \beta$. Nama su se najboljima pokazale skale 6,7 i 8. To naravno ovisi o veličini proteoma kojeg pretražujemo.

Funkcija distribucije slučajne varijeble X sa standardnom logističkom distribucijom dana je s:

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

Kod standardne logističke distribucije je $\mu = 0$ i $\beta = 1$, tj. $\sigma = \frac{\pi}{\sqrt{3}}$. Slijedi da je $prag = skala$ pa ukoliko uzmemo skalu 8

$$1 - F(8) = \frac{e^8}{1 + e^8} = 0.0003354$$

će se $0,03\%$ maksimalnih ocjena sličnosti smatrati značajnim.

2.7 Iteriranje

Rekli smo da ćemo iterativno graditi profil motiva. Opišimo malo detaljnije na koji način. Najprije zadajemo ulazni motiv koji se sastoji od jednog ili više nizova aminokiselina jednakih duljina, te skalu. Potom, na temelju tog motiva, izgradimo profil motiva kako je

opisano u poglavlju 2.4. Pomoću njega računamo maksimalne ocjene u svakom retku proteoma, a potom određujemo ovisno o skali koje od ocjena su značajne, tj. za koje od proteina ćemo zaključiti da su iz tražene proteinske familije. Nizove aminokiselina s dovoljno velikim ocjenama potom dodajemo u listu nizova koji čine motiv. Zatim ponavljamo postupak, na temelju novog motiva izgradimo profil motiva i opet tražimo nizove aminokiselina s dovoljno velikim ocjenama. Iteriranje staje kada se lista nizova, tzv. pozitivaca koji čine motiv ne promijeni, odnosno kada se postigne zadani maksimalni broj iteracija. Na kraju postupka imat ćemo listu pozitivaca, tj. kraćih nizova aminokiselina sprecifičnih za proteinsku familiju, kao i listu proteina za koje smo zaključili da pripadaju toj familiji.

Poglavlje 3

Poboljšanje

3.1 Optimizacija

U poglavlju 2.5 objasnili smo na koji način ćemo smanjiti broj pozicija na kojima ćemo računati ocjenu sličnosti (logaritam omjera vjerojatnosti) pomoću funkcije sličnosti. Željeli bismo taj broj još smanjiti. Logično je pretpostaviti da nizovi koji se s motivom poklapaju tek na jednoj ili dvije pozicije neće biti značajni. U proteinu u kojem maksimum funkcije sličnosti iznosi tek 1 ili 2, tj. kada je

$$d_{max} \in \{1, 2\},$$

najčešće ćemo naći velik broj nizova s takvom vrijednosti funkcije sličnosti budući da je vrlo vjerojatno da će se niz podudarati s motivom na barem jednoj poziciji. To može značiti da ćemo u tom retku proteoma jako puno puta evaluirati formulu (2.7),nekad čak i preko 100 puta, a očekujemo da na kraju niti jedan od nizova iz tog proteina nećemo proglasiti dovoljno sličnim. Stoga ćemo, kako bismo ubrzali cijeli postupak, u proteinima za koje je $d_{max} = 1$ ili 2 ocjenu sličnosti izračunati samo za prvi niz s maksimalnom vrijednosti funkcije sličnosti u tom proteinu. Taj score ćemo smatrati maksimalnim za taj protein. Na taj način ćemo značajno ubrzati traženje, ali uočimo da nećemo naći pravi maksimum za neke retke proteina kao što bismo trebali. Pravi maksimum naći ćemo samo kada baš taj prvi niz s maksimalnom vrijenosti funkcije sličnosti ima i maksimalan score.

3.2 Promjena modela M

Razmotrimo ponovo postupak građenja profila motiva i ocjenjivanja. Uočimo da, ukoliko se ulazni motiv sastoji od jednog ili svega nekoliko nizova aminokiselina, ima smisla da nam u građenju profila motiva teorijska distribucija igra veliku ulogu. Promotrimo npr.

prvi stupac motiva. Naime, pri određivanju funkcije distribucije za prvu poziciju, očito će najveću vjerojatnost imati upravo ona aminokiselina koja se pojavila u motivu na toj poziciji. Međutim, na temelju samo jedne aminokiseline, bez teorijeske distribucije, ne možemo znati što će umjesto te aminokiseline doći s većom, a što s manjom vjerojatnosti. Naprotiv, ako se motiv sastoji od velikog broja nizova, željeli bismo da empirijska distribucija ima veću ulogu od teorijske. Naime, ukoliko su se na prvom mjestu puno puta pojavile određene aminokiseline želimo da njihove vjerojatnosti budu velike bez obzira na teorijsku distribuciju.

Zbog navdenog promijenit ćemo model M koji pretpostavlja ovisnost promatranog niza aminokiselina o motivu, tj. u kojem se računa vjerojatnost da varijabla Z proizvede događaj z koji na i-toj poziciji ima z_i uz uvjet x, odnosno x_i, gdje je x motiv. Dakle promijenit ćemo princip računanja vjerojatnosti (2.5) koji je objašnjen u poglavlju 2.4.

Vjerojatnosna distribucija $p_i = (p_{i_1}, p_{i_2}, \ldots, p_{i_{20}})$ dobivena je na temelju i-tog stupca ulaznog motiva i PAM matrice koja sadrži vjerojatnosti da pojedina aminokiselina mutira u drugu. S $f_i = (f_i, f_{i_2}, \ldots, f_{i_{20}})$ označili smo empirijsku funkciju distribucije za i-ti stupac motiva. Konačnu funkciju distribucije k_i računat ćemo kao konveksnu kombinaciju teorijske funkcije distribucije p_i i empirijske funkcije distribucije f_i, $\forall i \in \{1, 1, \ldots, n\}$ pri čemu će koeficijenti konveksne kombinacije varirati.

Označimo s *nos* (engl. numer of sequences) broj nizova od kojih se sastoji motiv. Želimo da, kada je $nos = 1$, teorijska i empirijska distribucija imaju jednak utjecaj, tj. svaka 50%. Kada se motiv sastoji od otprilike 50 nizova želimo da utjecaj teorijske distribucije bude manji, oko 25%. Kada pak *nos* bude jako velik, oko 200 želimo da utjecaj teorijske distribucije padne na tek oko 10%.

Definirat ćemo funkciju α na $[1, +\infty >$ za koju želimo da određuje koeficijent konveksne kombinacije ovisno o broju nizova od kojih se sastoji motiv. Definiramo ju na sljedeći način:

$$\alpha(x) = \begin{cases} 0.505514 - 0.00551429x, & \text{za } 1 \leq x < 36 \\ -\frac{9.05}{x^2} + \frac{9.5}{x} + 0.05, & \text{za } x \geq 36 \end{cases} \tag{3.1}$$

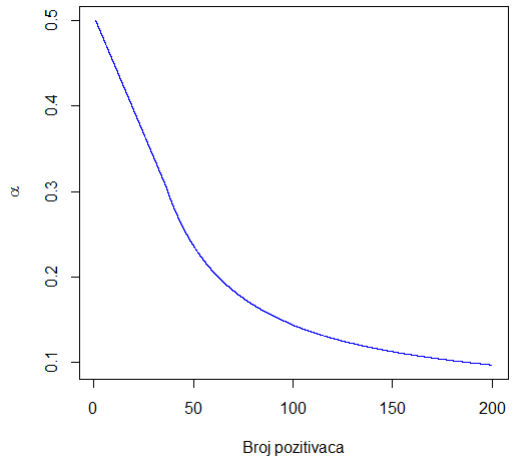
Naime kada bismo funkciju alfa računali samo kao:

$$\alpha'(x) = -\frac{9.05}{x^2} + \frac{9.5}{x} + 0.05$$

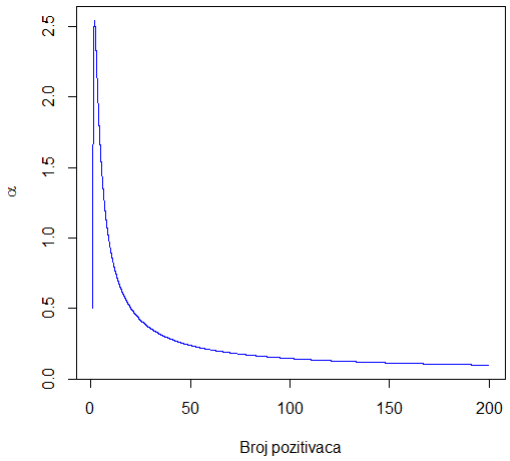
njome ne bismo mogli računati koeficijent konveksne kombinacije, kao što se vidi iz na slici 3.2

Distribuciju $k_i = (k_{i_1}, k_{i_2}, \ldots, k_{i_{20}})$ ćemo računati na sljedeći način:

$$k_i = \alpha(nos) \cdot p_i + (1 - \alpha(nos)) \cdot f_i \tag{3.2}$$



Slika 3.1: Graf funkcije α



Slika 3.2: Graf funkcije α'

Budući da je α padajuća funkcija, očito je da se s povećanjem broja nizova u motivu smanjuje utjecaj teorijske distribucije kao što smo željeli. Takav način računanja vjerojatnosti ima više smisla ako uzmzmo u obzir prirodu ovog zadatka. Označimo novo dobiveni model s M'. Vrijedi:

$$\mathbb{P}(z|M'_x) = \prod_{i=1}^{n} \mathbb{P}(z_i|M'_{x_i}) = \prod_{i=1}^{n} k_{i_{z_i}}, \tag{3.3}$$

pa ćemo u skladu s time ocjenu poravnanja računati kao:

$$s'_k = \sum_{i=0}^{n-1} \log \frac{\mathbb{P}(y_{k+1}|M'_i)}{\mathbb{P}(y_{k+1}|R)} = \sum_{i=0}^{n-1} \log \frac{k_{i_{k+1}}}{q_{y_{k+1}}}, k = 1, 2, \ldots, m - n + 1 \tag{3.4}$$

Poglavlje 4

Analiza metode na proteomu biljke talijin uročnjak

Talijin uročnjak (lat. Arabidopsis thaliana) je mala biljka s cvijetom koja pripada familiji Brassicaceae, u koju pripadaju i neke kultivirane biljke poput kupusa i rotkvice. To je jednogodišnja biljka relativno kratkog životnog vijeka. U znanstvenim krugovima se često označava kao biljna vinska mušica. Naime, Arabidopsis je prva biljka s potpuno sekvencioniranim genomom što je čini pogodnom za danja istraživanja u genetici i molekularnoj biologiji. Služi za razumijevanje mnogih biljnih osobina.

Slika 4.1: Arabidopsis thaliana

Dugo se mislilo da ima najkraći genom od svih cvjetnica, ali se ispostavilo da nije tako. Njen genom sastoji se od 24498 gena koji kodiraju proteine iz 11 000 proteinskih familija.

Proteom joj se sastoji od 35176 proteina. Za navedenu biljku eksperimentalno, ali i raznim durgim metodama utvrđeno je koji proteini iz proteoma pripadaju kojoj proteinskoj familiji. Na temelju tih podataka mi ćemo testirati uspješnost naših metoda na primjeru traženja proteina iz familije GDSL enzima.

GDSL enzimi relativno su novo otkrivena podklasa lipolitičkih enzima koji su vrlo bitni i atraktivni predmeti proučavanja zbog svojih višefunkcionalnih svojstva. Stoga, imaju velik potencijal 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. Proteini iz te familije pokazuju malu međusobnu sličnost u građi. U preoteomu biljke Arabidopsis tražit ćemo proteine iz te familije pomoću motiva koji se sastoji od samo jednog bloka, koji je karakterističan za tu familiju. Općenito je tipičan protein koji pripada ovoj familiji okarakteriziran s 5 motiva.

Dakle pretražujemo proteom od 33410 nizova kako bismo našli proteine koje pripadaju familiji GDSL enzima. Za ulazni motiv uzet ćemo x="FVFGDSLVDN". Kao što je već rečeno, mi te podatke imamo, tj. unaprijed znamo da je 127 proteina iz te familije i znamo točno koji su. Najprije ćemo tražiti "dovoljno slične" nizove aminokiselina našem motivu na način opisan u poglavlju 2.4 za različite vrijednost skale. Dobiveni rezultati prikazani su u tablicama 4.1 i 4.2. Stanje pozitivno nam označava da je protein zaista iz te familije, a negativno da nije. Br. pozitivaca označava broj proteina koji su proglašeni da pripadaju traženoj familiji.

skala	**br. pozitivaca**	**TP**	**FP**	**TN**	**FN**
4	747	125	622	32661	2
5	306	101	205	33078	26
6	171	96	75	33208	31
7	106	93	13	33270	34
8	80	79	1	33282	48
9	51	51	0	33283	76
10	32	32	0	33283	95
11	29	29	0	33283	98

Tablica 4.1: uz model *M*

Uočimo da je negativna prediktivna vrijednost uvijek blizu 1. To nije neočekivano budući da je od 33410 proteina samo 127 onih u pozitivnom stanju, pa će uvijek biti velik broj TN. Specifičnost je također dosta velika za sve vrijednosti skale, ali se malo povećava promjenom skale. Vidimo da za male vrijednosti skale ima puno FP, zbog čega je pozitivna prediktivna vrijednost premala. S povećanjem skale ona raste. Međutim, s povećanjem

skala	osjetljivost	specifičnost	PPV	NPV
4	0.9843	0.9814	0.1673	0.9999
5	0.7953	0.9938	0.3301	0.9992
6	0.7559	0.9977	0.5614	0.9991
7	0.7323	0.9996	0.8774	0.9989
8	0.6220	0.99996	0.9875	0.9986
9	0.4016	1	1	0.9977
10	0.2520	1	1	0.9972
11	0.2283	1	1	0.9972

Tablica 4.2: uz model M

skale drastično se smanjuje osjetljivost testa jer mnogi proteini iz GDSL familije nisu prepoznati. Skale 7 i 8 se stoga čine kao optimalan odabir.

Nadalje, razmotrit ćemo rezultate dobivene na način da nismo u svakom retku tražili pravi maksimum, nego smo u recima s niskim vrijednostima ocjene poravnanja uzele prvi blok na koji smo naišli kao najbolji kako bi ubrzali proces traženja. Detaljnije je objašnjeno u poglavlju 3.1. U tablicama 4.3 i 4.4 prikazani su dobiveni rezultati. Usporedimo sada tablicu te tablice s tablicama 4.1 i 4.2. Fokusirajmo se najprije na broj TP. Uočimo da je on za gotovo sve vrijednosti skale manji u odnosu na iste vrijednosti skale dobivene kada smo tražili pravi maksimum. Uočimo da je osjetljivost testa također manja za sve vrijednosti skale, osim 4 kada je jednaka. Za skalu 7 ovom metodom nađeno je 13 proteina manje nego početnom. Osjetljivost testa pala je s 0.7323 na 0.6299. Pretpostavljamo da je naš postupak poremetio distribuciju maksimalnih score-ova budući da smo dio maksimalnih ocjena zamijenili s manjim ocjenama. To je moglo dovesti do gubitka simetričnosti. Također, očekivanje se promijenilo. Zbog svega navedenog, dobit ćemo nešto lošije rezultate. Međutim, PPV je za sve vrijedosti skale čak veća ili jednaka, što je zaista dobro ako želimo biti sigurniji da su pozitivci zapravo TP. Međutimo, kako bismo uhvatili sve više TP zaključujemo da je u ovom slučaju ipak bolje ne ubrzavati traženje na ovaj način.

Naposlijetku, pretraživali smo proteom uz pomoć izmijenjenog modela M'. Dobiveni rezultati dani su u tablicama 4.5 i 4.6. Želimo proučiti da li je bolji od početnog modela. Uočimo da je vrijednost TP porasla za vrijednosti skale 4, 5, 8, 9, 10 i 11, za skalu 6 je ostala ista, dok je za skalu 7 nađen jedan manje TP. Za skalu 8 prepoznato su čak 32 TP više, što su značajno bolji rezultati. Osjetljivost testa također je veća za gotovo sve vrijednosti skale. Jedino se za skalu 7 smanjila, dok je za skalu 8 čak za 0.252 veća. NPV i specifičnost testa velike su za sve vrijednosti skale iz već objašnjenog razloga da negativnih ima puno više od pozitivnih. PPV drastično raste povećanjem skale. Uočimo da se za skalu 7 neočekivano dobiva manja osjetljivost i manji broj TP od očekivanog. To možemo objasniti time da je u početnoj iteraciji previše negativaca prepoznato kao pozitivci te se na temenju "loših"

skala	br. pozitivaca	TP	FP	TN	FN
4	612	125	487	32796	2
5	240	99	141	33192	28
6	141	94	47	33236	33
7	86	80	6	33277	47
8	77	77	0	33283	50
9	36	36	0	33283	91
10	30	30	0	33283	97
11	25	25	0	33283	102

Tablica 4.3: uz model M, ne uzima pravi maksimum

skala	osjetljivost	specifičnost	PPV	NPV
4	0.9843	0.9854	0.2042	0.9999
5	0.7795	0.9958	0.4125	0.9992
6	0.7402	0.9986	0.6667	0.9990
7	0.6299	0.9998	0.9302	0.9986
8	0.6063	1	1	0.9985
9	0.2835	1	1	0.9973
10	0.2362	1	1	0.9971
11	0.1969	1	1	0.9970

Tablica 4.4: uz model M, ne uzima pravi maksimum

blokova izgradio profil motiva. Sve u svemu, vidimo da uz model M' dobivamo znatno bolje rezultate u ovom slučaju. Dakle, naposlijetku možemo zaključiti da se najtočniji rezultati dobivaju uz model M'.

skala	br. pozitivaca	TP	FP	TN	FN
4	787	126	661	32622	1
5	336	119	217	33066	8
6	185	96	89	33194	31
7	111	92	19	33264	35
8	119	111	8	33275	16
9	90	90	0	33283	37
10	40	40	0	33283	87
11	30	30	0	33283	97

Tablica 4.5: uz model M'

skala	**osjetljivost**	**specifičnost**	**PPV**	**NPV**
4	0.9921	0.9801	0.1601	0.99997
5	0.9370	0.9935	0.3542	0.9998
6	0.7559	0.9973	0.5189	0.9964
7	0.7244	0.9994	0.8288	0.9989
8	0.8740	0.9998	0.9328	0.9995
9	0.7087	1	1	0.9989
10	0.3150	1	1	0.9974
11	0.2362	1	1	0.9971

Tablica 4.6: uz model M'

Bibliografija

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

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

Sažetak

Opisali smo postupak traženja motiva iz neke proteinske familije u proteomu nekog organizma. Definirali smo ocjenu sličnosti bloka aminokiselina sa zadanim motivom te dobili da su maksimalne ocjene sličnosti, uz korekciju zbog nejenakih duljina nizova, logistički distribuirane. Pokušali smo ubrzati pretraživanje te smo poboljšali metodu alteracijom iterativnog građenja profila motiva. Analizom rezultata dobivenih traženjem GDSL enzima u proteomu biljke Arabidopsis thaliana pokazalo se da ova metoda daje vrlo dobre rezultate.

Summary

We have described a procedure for iterative searching of protein motifs in a large set of protein sequences. We have described a similarity score in terms of a log-odds ratio, and established that the scores are following a logistic distribution. Furthermore, we have sped up the search procedure, and adjusted the profile building method. Analyses of results given by searching for GDSL enzymes in a protein of a plant Arabidopsis thaliana showed that our method gives very good results.

Životopis

Rođena sam 1. svibnja 1993. godine u Zagrebu. Svoje školovanje započela sam u osnovnoj školi "Bogumil Toni" u Samoboru te ga nastavila 2007. godine u općoj gimnaziji "Antun Gustav Matoš" u Samoboru. Nakon završenog srednjoškolskog obrazovanja, 2011. godine upisujem Preddiplomski sveučilišni studij Matematike 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 Matematičke statistike, kojeg upravo završavam.