SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Mihaela Stojković

VIŠESTRUKO PORAVNAVANJE I HMM

Diplomski rad

Voditelj rada:
doc. dr. sc. Pavle Goldstein

Zagreb, srpanj, 2015.

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. _____

Zadovoljstvo mi je zahvaliti se mentoru doc. dr. sc. Pavlu Goldsteinu na posvećenom vremenu i pruženoj pomoći u izradi ovog diplomskog rada. Zahvaljujem se i svojoj obitelji koja je vjerovala u mene i pružala mi potporu tijekom studiranja.

Sadržaj

Uvod

Ovaj rad se bavi proučavanjem proteinskih familija, na primjeru AT domena iz poliketidnih sintetaza (AT označava aciltransferazu). Bioinformatika se bavi proučavanjem bioloških karakteristika živih bića, kao što je analiza genoma, predviđanje strukture, time i funkcije proteina uz pomoć matematike, statistike i programiranja. U ovom radu ću opisati dio u koji ulazi višestruko poravnanje nizova, te kakav matematički model za opis proteinskih familija možemo primijeniti. Problem kojim se bavimo je analiza višestrukog poravnanja. Pristupiti ćemo problemu na način da za svaki niz odredimo "score" što će biti mjera slaganja sa matematičkim modelom u pozadini promatrane familije proteina.

Rad je podijeljen u pet dijelova. U prvom dijelu govori o osnovnim pojmovima iz teorije vjerojatnosti i statističkim metodama. Detaljnije razrađeni pojmovi navedeni u prvom dijelu mogu se pronaći u [6]. Taj dio služi kao potkrepa matematičkom aparatu koji se koristi u drugim djelovima rada. Drugi dio rada govori o osnovnim pojmovima vezanim uz proteine. To je kratki uvod u biološka i kemijska svojstva proteina. Važno je obratiti pažnju na tercijarnu strukturu proteina, te na činjenicu da je ona u potpunosti određena primarnom, budući da je ona faktor koji određuje funkciju proteina u organizmu. Detaljan opis bioloških i kemijskih svojstava proteina može se naći u [1]. Nadalje, u trećem poglavlju uvodimo matematički aparat koji se koristi za opis primarne strukture proteina. U tom dijelu definiramo skriveni Markovljev model, zatim ga povezujemo s analizom proteina, te dajemo algoritam kojim ćemo računati spomenute "score"-ove. Zatim dolazi još jedan rezultat iz matematike i statistike pomoću kojeg ćemo analizirati izračunate "score"-ove. Četvrti dio se bavi vezom između biološkog i matematičkog značenja višestrukog poravnanja, pa je ono razdvojeno na proučavanje biološkog značenja poravnanja i poravnanja u smislu skrivenog Markovljevog modela. Na kraju rada ćemo pogledati rezultate, vidjet ćemo kako smo procijenili parametre skrivenog Markovljevog modela i koliko dobro smo ih procijenili simulacijom novih nizova. Odgovorit ćemo na pitanje kako se ponašaju "score"-ovi i pokušati poboljšati dobiveni model.

Poglavlje 1

Teorija vjerojatnosti i statistika

1.1 Osnovni pojmovi

Vjerojatnosni prostor

Slučajni pokus je pokus čiji ishodi nisu jednoznačno određeni uvjetima u kojima izvodimo pokus. Ishode slučajnih pokusa zovemo **događaji**. Neka je A proizvoljan događaj vezan uz neki slučajni pokus. Svojstvo **statističke stabilnosti** relativnih frekvencija [1] je u tome da se prilikom ponavljanja pokusa veliki broj puta relativne frekvencije događaja A grupiraju oko nekog fiksnog broja.

Definicija 1.1.1. *Ako slučajni pokus zadovoljava uvjet statističke stabilnosti relativnih frekvencija, tada se* ***vjerojatnost a posteriori*** *proizvoljnog događaja A vezanog uz taj pokus definira kao realan broj* $\mathbb{P}(A), 0 \leq \mathbb{P}(A) \leq 1$*, oko kojeg se grupiraju, odnosno kojemu teže relativne frekvencije tog događaja.*

Definicija 1.1.2. *Neka je* Ω *neprazan proizvoljan skup. Familija* $\mathcal{F}$ *podskupova od* Ω *je* σ*-algebra skupova na* Ω *ako vrijedi:*

1. $\emptyset \in \mathcal{F}$
2. $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$
3. $A_i \in \mathcal{F}, i \in \mathbb{N} \Rightarrow \bigcup_{i \in \mathbb{N}} A_i \in \mathcal{F}$

Definicija 1.1.3. *Neka je* Ω *proizvoljan neprazan skup, a* $\mathcal{F}$ σ*-algebra na tom skupu. Uređen par* $(\Omega, \mathcal{F})$ *zove se izmjeriv prostor. Nadalje, funkcija* $\mathbb{P} : \mathcal{F} \to \mathbb{R}$ *je* ***vjerojatnost*** *na* $\mathcal{F}$ *ako vrijedi:*

[1] Relativna frekvencija događaja A je broj pojavljivanja događaja A podijeljen sa ukupnim brojem ponavljanja pokusa.

1. $\mathbb{P}(A) \geq 0, A \in \mathcal{F}; \mathbb{P}(\Omega) = 1$

2. $A_i \in \mathcal{F}, i \in \mathbb{N} \quad i \quad A_i \cap A_j = \emptyset \quad za \quad i \neq j \Rightarrow \mathbb{P}\left(\bigcup_{i\in\mathbb{N}} A_i\right) = \sum_{i\in\mathbb{N}} P(A_i)$

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**.

Definicija 1.1.4. *Neka je $(\Omega, \mathcal{F}, \mathbb{P})$ vjerojatnosni prostor i $A \in \mathcal{F}$ takav da je $\mathbb{P}(A) > 0$. Definiramo funkciju $\mathbb{P}_A : \mathcal{F} \to [0, 1]$ tako da je:*

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

*$\mathbb{P}_A$ je vjerojatnost na $\mathcal{F}$ i zovemo je **uvjetna vjerojatnost uz uvjet** A*

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

Slučajne varijable

Definicija 1.1.5. *Funkcija $X : \Omega \to \mathbb{R}$ je **slučajna varijabla** ako je $X^{-1}(B) \in \mathcal{F}$ za proizvoljan skup $B \in \mathcal{B}$.*

Propozicija 1.1.6. *Neka je X slučajna varijabla na Ω i $g : \mathbb{R} \to \mathbb{R}$ Borelova funkcija. Tada je $g(X) = g \circ X$ također slučajna varijabla na Ω.*

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

$$\mathbb{P}\left(X_1 \in B_1, X_2 \in B_2, \ldots, X_n \in B_n\right) = \mathbb{P}\left(\bigcap_{i=1}^{n} X_i \in B_i\right) = \prod_{i=1}^{n} \mathbb{P}\left(X_i \in B_i\right) \tag{1.2}$$

Neka je $(\Omega, \mathcal{F}, \mathbb{P})$ vjerojatnosni prostor i X slučajna varijabla definirana na njemu. Definiramo **vjerojatnosnu mjeru induciranu slučajnom varijablom** X kao:

$$\mathbb{P}_X(B) = \mathbb{P}(X^{-1}(B)) = \mathbb{P}\{\omega \in \Omega; X(\omega) \in B\}, \quad B \in \mathcal{B} \tag{1.3}$$

Relacijom 1.3 definirana je vjerojatnost $\mathbb{P}_X : \mathcal{B} \to [0, 1]$ koju još zovemo **zakon razdiobe** od X. Pripadni vjerojatnosni prostor $(\mathbb{R}, \mathcal{B}, \mathbb{P}_X)$ je vjerojatnosni prostor induciran slučajnom varijablom X.

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

$$F_X(x) = \mathbb{P}_X(\langle -\infty, x]) = \mathbb{P}(X^{-1}(\langle -\infty, x]) = \mathbb{P}\{\omega \in \Omega; X(\omega) \leq x\} = \mathbb{P}\{X \leq x\}, x \in \mathbb{R} \tag{1.4}$$

Obično nas zanimaju vjerojatnosti događaja vezanih uz neku slučajnu varijablu, odnosno $\mathbb{P}(\omega \in \Omega : X(\omega) \in B) = \mathbb{P}(X \in B)$, za proizvoljno $B \in \mathcal{B}$. Način na koji to računamo ovisi o tipu slučajne varijable, pa tako imamo dva glavna tipa: diskretne i neprekidne.

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

Definicija 1.1.10. *Slučajna varijabla X je apsolutno neprekidna ili **neprekidna** slučajna varijabla ako postoji nenegativna Borelova funkcija f na $\mathbb{R}$ takva da vrijedi:*

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

Funkciju distribucije oblika F_X iz relacije 1.5 zovemo neprekidna funkcija distribucije, a funkciju f iz iste relacije tada zovemo funkcijom gustoće vjerojatnosti od X.

Matematičko očekivanje i varijanca

Definicija 1.1.11. *Neka je X diskretna slučajna varijabla, i neka je skup D iz definicije 1.1.9, $D = \{x_1, x_2, \ldots\}$. Definiramo **matematičko očekivanje** diskretne slučajne varijable X kao:*

$$\mathbb{E}X = \sum_k x_k \mathbb{P}_X(\{x_k\}) \tag{1.6}$$

Definicija 1.1.12. *Neka je X neprekidna slučajna varijabla s funkcijom distribucije F_X, tada je očekivanje slučajne varijable X*

$$\mathbb{E}X = \int_\Omega X d\mathbb{P} = \int_{\mathbb{R}} x dF_X(x) \tag{1.7}$$

Definicija 1.1.13. *Neka je X neprekidna slučajna varijabla i $g : \mathbb{R} \to \mathbb{R}$ Borelova funkcija, tada vrijedi:*

$$\mathbb{E}[g(X)] = \int_\Omega g(X) d\mathbb{P} = \int_{\mathbb{R}} g(x) dF_X(x) \tag{1.8}$$

Neka je X slučajna varijabla na vjerojatnosnom prostoru $(\Omega, \mathcal{F}, \mathbb{P})$ i $r > 0$. $\mathbb{E}[X^r]$ zovemo r-ti moment od X, $\mathbb{E}[|X|^r]$ je r-ti apsolutni moment od X. Neka je X takva slučajna varijabla da $\mathbb{E}X$ postoji, tada je $\mathbb{E}[(X - \mathbb{E}X)^r]$ r-ti centralni moment od X.

Definicija 1.1.14. *Varijanca od X, u oznaci $VarX$ ili σ_X^2 je drugi centralni moment od X,*

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

Pozitivan drugi korijen iz varijance, σ_X zovemo standardna derivacija od X.

Definicija 1.1.15. *Kažemo da niz slučajnih varijabli $(X_n)_{n\in\mathbb{N}}$ definiranih na vjerojatnosnom prostoru $(\Omega, \mathcal{F}, \mathbb{P})$* ***konvergira gotovo sigurno****, u oznaci (g.s.) prema slučajnoj varijabli X definiranoj na istom vjerojatnosnom prostoru ako vrijedi:*

$$\mathbb{P}\{\omega \in \Omega : X(\omega) = \lim_{n\to\infty} X_n(\omega)\} = 1. \tag{1.9}$$

Definicija 1.1.16. *Kažemo da niz slučajnih varijabli $(X_n)_{n\in\mathbb{N}}$ definiranih na vjerojatnosnom prostoru $(\Omega, \mathcal{F}, \mathbb{P})$* ***konvergira po distribuciji*** *prema slučajnoj varijabli X definiranoj na istom vjerojatnosnom prostoru ako vrijedi:*

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x), \forall x \in C(F_X), \tag{1.10}$$

gdje je $C(F_X)$ skup svih točaka neprekidnosti funkcije F_X.

Primjeri slučajnih varijabli

Počnimo od najjednostavnije slučajne varijable. Neka je

$$\epsilon(x) = \begin{cases} 0, & x < 0 \\ 1, & x \geq 0. \end{cases} \tag{1.11}$$

Za proizvoljan $c \in \mathbb{R}$ neka je $F(x) = \epsilon(x - c), x \in \mathbb{R}$. Slučajna varijabla X za koju je F funkcija distribucije je slučajna varijabla degenerirana u c.

Neprekidna slučajna varijabla X s funkcijom gustoće

$$\begin{cases} \frac{1}{b-a}, & a \leq x < b \\ 0, & \text{inače} \end{cases} \tag{1.12}$$

za proizvoljne $a, b \in \mathbb{R}, a < b$ je uniformno distribuirana slučajna varijabla na segmentu $[a, b]$. Očekivanje i varijanca ovako distribuirane slučajne varijable su
$\mathbb{E}X = \frac{1}{b-a}\int_a^b x dx = \frac{a+b}{2}$ i $VarX = \frac{(a-b)^2}{12}$.

Normalno distribuirana slučajna varijabla s parametrima μ i σ^2, gdje je $\sigma^2 > 0$, a $\mu \in \mathbb{R}$ je neprekidna slučajna varijabla s funkcijom gustoće vjerojatnosti

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(x-\mu)^2}{2\sigma^2}}. \tag{1.13}$$

Očekivanje takve slučajne varijable je $\mathbb{E}X = \mu$ i varijanca je $VarX = \sigma^2$.

Definiramo Γ funkciju, $\Gamma(x) = \int_0^\infty e^{-t}t^{x-1}dt, \forall x > 0$. Neka je $\alpha > 0$ i $\beta > 0$, neprekidna slučajna varijabla ima gama distribuciju s parametrima α, β ako joj je gustoća vjerojatnosti zadana sa

$$f(x) = \begin{cases} \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-\frac{x}{\beta}}, & x > 0 \\ 0, & x \leq 0. \end{cases} \tag{1.14}$$

Očekivanje i varijanca su $\mathbb{E}X = \alpha\beta$ i $VarX = \alpha\beta^2$.

Neka su $\mu, \beta \in \mathbb{R}$, uz $\beta > 0$. Slučajna varijabla X ima logističku distribuciju s parametrima μ i β, ako joj je funkcija gustoće zadana sa

$$f(x) = \frac{e^{-\frac{x-\mu}{\beta}}}{\beta\left(1 + e^{-\frac{x-\mu}{\beta}}\right)}, \tag{1.15}$$

te je očekivanje te slučajne varijable $\mathbb{E}X = \mu$, a varijanca $VarX = \frac{\beta^2\pi^2}{3}$.

Procjena parametara

Neka je X slučajna varijabla koju promatramo.

Definicija 1.1.17. ***Slučajni uzorak*** *duljine n za X je niz od n nezavisnih i jednako distribuiranih slučajnih varijabli $X_1, X_2, \ldots, X_n$ koje imaju istu razdiobu kao i X.*

Definicija 1.1.18. *Neka je $(x_1, x_2, \ldots, x_n)$ opaženi uzorak za slučajnu varijablu X s gustoćom $f(x|\theta)$, gdje je $\theta = (\theta_1, \theta_2, \ldots, \theta_k) \in \Theta \subseteq \mathbb{R}^k$ nepoznati parametar.* ***Funkcija vjerodostojnosti*** *$L : \Theta \to \mathbb{R}$ definirana je sa*

$$L(\theta) = f(x_1|\theta)f(x_2|\theta) \cdot \ldots \cdot f(x_n|\theta), \theta \in \Theta. \tag{1.16}$$

Vrijednost $\hat{\theta} = \hat{\theta}(x_1, x_2, \ldots, x_n) \in \Theta$ za koju je $L(\hat{\theta}) = \max_{\theta\in\Theta} L(\theta)$ zovemo ***procjena metodom maksimalne vjerodostojnosti****, a $\hat{\theta}$* ***procjeniteljem maksimalne vjerodostojnosti****.*

Poglavlje 2

Proteini

Da bismo shvatili biološki smisao ovog rada potrebno je pojasniti neke osnovne pojmove, kao što su:

- **aminokiseline**
- **proteini**

Aminokiseline su molekule čija je glavna uloga izgradnja proteina. Najvažnija kemijska reakcija aminokiselina je formiranje peptidne veze koja omogućava povezivanje dviju aminokiselina i stvaranje lanca aminokiselina. **Proteini** su velike biološke molekule, makromolekule, sastavljene od niza aminokiselina i čine osnovu za izgradnju živih bića. Proteini u organizmu služe za izgradnju, jačanje i popravljanje tkiva, zatim za proizvodnju antitijela, koja služe imunološkom sistemu, stvaraju hormone koji su važni za prijenos informacija po organizmu. Neke vrste proteina važne su za kretanje, druge za transport kisika, neke čine enzime, itd. Postoji 20 standardnih aminokiselina koje tvore proteine.

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: Popis aminokiselina s oznakama

2.1 Struktura proteina

Primarna struktura

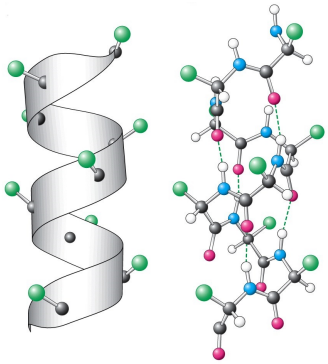
Niz aminokiselina se veže peptidnom vezom[1] tvoreći **polipeptidni lanac**. Peptidna veza je planarna, što znači da oko nje nema rotacije molekula. To svojstvo ima zbog djelomičnog obilježja dvostruke kovalentne veze. Pojedina aminokiselina u polipeptidnom lancu čini naše opažanje. Polipeptidni lanci u prirodi su obično sastavljeni od 50 do 2000 aminokiselina, te ih zovemo proteini. Svaki protein je zadan konačnim nizom u skupu aminokiselina, koji nazivamo **primarnom strukturom** proteina. Točan niz aminokiselina je važan jer određuje svojstva proteina, njihovu trodimenzionalnu strukturu i biološku zadaću.

Sekundarna struktura

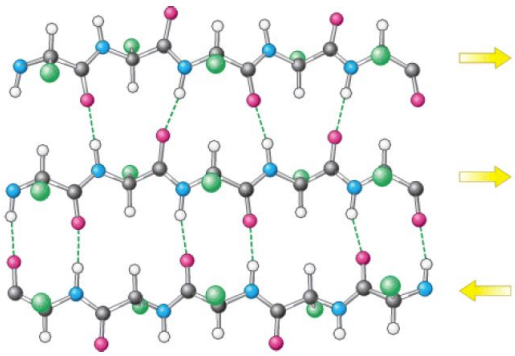
Polipeptidni lanac se može saviti u regularne strukture kao što su:

- α-zavojnice
- β-ploče
- β-okreti i Ω-petlje

α-zavojnica[2] određena je brojem aminokiselina u jednom okretu i rastom zavojnice, prikazana je na slici 2.1.



Slika 2.1: Prikaz α -zavojnice



Slika 2.2: Prikaz β- nabrane ploče

[1]Peptidna veza nastaje spajanjem dviju aminokiselina tako da se spoji ugljikov atom jedne s dušikovim atomom druge aminokiseline uz oslobađanje molekule vode. Generalić, Eni. "Peptidna veza." Englesko-hrvatski kemijski rječnik & glosar. 22 Feb. 2015. KTF-Split. 19 Mar. 2015.

[2]eng. α-helix

Na slici 2.2 je prikazana β nabrana ploča[3] koja nastaje stvaranjem vodikovih veza između atoma. β- ploča se sastoji od β-lanaca[4] koji su potpuno rastegnuti za razliku od α-zavojnica. Obično se tako veže oko 4-5 lanaca, ali može i više. Okreti i petlje su dijelovi koji nemaju periodičku strukturu kao α-zavojnice i β- ploče, zato su to vrlo reaktivni dijelovi slabe strukture.

Tercijarna struktura

Tercijarna struktura označava prostornu organizaciju proteina koja se događa zbog veza između aminokiselina koje u primarnoj strukturi nisu međusobno blizu. U različitim okolinama se proteini savijaju u kompaktne strukture i tek tada protein može obavljati svoju funkciju. Trodimenzionalna struktura je u potpunosti određena primarnom.

[3] eng. β-sheet
[4] eng. β-strand

Poglavlje 3

Skriveni Markovljevi modeli

3.1 HMM

Osnovno

Definicija 3.1.1. ***Slučajan proces*** *s diskretnim vremenom i prostorom stanja S je familija $X = (X_n : n \geq 0)$ slučajnih varijabli definiranih na nekom vjerojatnosnom prostoru $(\Omega, \mathcal{F}, \mathbb{P})$ s vrijednostima u S.*

Definicija 3.1.2. *Neka je S prebrojiv skup. Slučajni proces $X = (X_n : n \geq 0)$ definiran na vjerojatnosnom prostoru $(\Omega, \mathcal{F}, \mathbb{P})$ s vrijednostima u skupu S je* ***Markovljev lanac prvog reda*** *ako za svaki $n \geq 0$ i za sve $i_0, \ldots, i_{n-1}, i, j \in S$ vrijedi*

$$\mathbb{P}(X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) = \mathbb{P}(X_{n+1} = j | X_n = i) \tag{3.1}$$

uz pretpostavku da su obje uvjetne vjerojatnosti dobro definirane.

Svojstvo u relaciji 3.1 zovemo Markovljevim svojstvom, a skup S prostorom stanja Markovljevog lanca.

Definicija 3.1.3. *Neka je sa $p_{ij} = \mathbb{P}(X_{t+1} = j | X_t = i)$ označena vjerojatnost da slučajna varijabla X u trenutku $t + 1$ prijeđe u stanje j ako je u trenutku t bila u stanju i. Tada p_{ij} zovemo prijelazna, odnosno* ***tranzicijska vjerojatnost****.*

Markovljev model je Markovljev lanac zajedno sa zadanim prijelaznim vjerojatnostima.

Definicija

Kod skrivenog Markovljevog modela imamo poznati niz opaženih simbola i nepoznati, skriveni niz stanja. Kada je neko stanje posjećeno Markovljevim lancem, ono emitira simbol. Simbol predstavlja vidljivo svojstvo procesa koji promatramo. Emisija simbola ovisi o stanju iz kojeg je emitirano na način da za svako stanje postoji zasebna vjerojatnosna distribucija simbola.

Definicija 3.1.4. ***Skriveni Markovljev model*** [1] *je skup slučajnih varijabli* $E = E_1, E_2, \ldots, E_n$ *i* $S = S_1, S_2, \ldots, S_n$*, gdje S poprimaju diskretne vrijednosti, a E poprimaju diskretne ili kontinuirane vrijednosti, te za koje vrijedi:*

$$\mathbb{P}(S_t|S_{t-1}, S_{t-2}, \ldots, S_1) = \mathbb{P}(S_t|S_{t-1}), \forall t = 1, \ldots, n \tag{3.2}$$

$$\mathbb{P}(E_t|E_n, S_n, \ldots, E_{t+1}, S_{t+1}, S_t, E_{t-1}, S_{t-1} \ldots, E_1, S_1) = \mathbb{P}(E_t|S_t), \forall t = 1, \ldots, n \tag{3.3}$$

Napomenimo da u gornjoj definiciji E predstavljaju slučajne varijable koje opisuju opažene simbole, a S skup slučajnih varijabli koje opisuju stanja kroz koja su pripadajući simboli emitirani. Relacijom 3.2 ističemo da vjerojatnost prelaska iz prethodnog stanja u sljedeće ovisi samo o ta dva stanja, S_t i S_{t-1}, tj. ne ovisi i o svim ostalim stanjima kroz koja je proces prošao $S_{t-2}, \ldots, S_1$. Također, relacijom 3.3 ukazujemo da je vjerojatnost da je trenutni simbol E_t emitiran ovisi samo o trenutnom stanju S_t procesa, bez obzira na sva prethodna i buduća stanja i emisije. Možemo uvesti notaciju koja će dodatno pojasniti prethodnu definiciju. Neka je sa S označen skup od n stanja, $S = \{1, 2, \ldots, n\}$, a sa $B = \{b_1, b_2, \ldots, b_m\}$ skup mogućih simbola. Vjerojatnosti iz 3.2 i 3.3 označimo na sljedeći način:

$$a_{ij} = \mathbb{P}(S_t = j|S_{t-1} = i), i, j \in S \tag{3.4}$$

$$e_k(b_l) = \mathbb{P}(E_t = b_l|S_t = k), k \in S, b_l \in B \tag{3.5}$$

Definiramo matrice $A = \{a_{ij}\}, i, j \in S$ i $E = \{e_k(l)\}, k \in S, l \in \{1, 2, \ldots, m\}$. Opaženi niz simbola označimo sa $X = (x_1, x_2, \ldots, x_L)$. Matrica A sadrži **tranzicijske vjerojatnosti**, a matrica E **emisijske vjerojatnosti**

3.2 Viterbijev algoritam

Pretpostavimo sada da nam je zadan opaženi niz simbola $X = (X_1, X_2, \ldots, X_L)$. Budući da u skrivenom Markovljevom modelu nije iz opažanja vidljiv put $\pi = (\pi_1, \pi_2, \ldots, \pi_L)$

[1] eng. Hidden Markov model

Markovljevog lanca kroz stanja, postavlja se prirodno pitanje koja stanja su emitirala te simbole. Tu nastupa Viterbijev algoritam, to je jedan od načina otkrivanja (dekodiranja) puta kroz stanja modela. Obično postoji više putova kroz model pomoću kojih možemo emitirati upravo opaženi niz, htjeli bismo prepoznati za koji je od njih vjerojatnost da je opaženi niz emitiran kroz neki put stanja, s obzirom na parametre modela, najveća.

Označimo sa π^* put s najvećom vjerojatnosti, tada je

$$\pi^* = \underset{\pi}{\operatorname{argmax}} \mathbb{P}(X, \pi). \tag{3.6}$$

Najvjerojatniji put možemo izračunati rekurzivno . Kad bismo za svako stanje $k \in S$ znali vjerojatnost $V_k(i)$ da put završava u tom stanju s opažanjem i, tada bi se te vjerojatnosti mogle izračunati za sljedeće opažanje x_{i+1} kao

$$V_l(i+1) = e_l(x_{i+1}) \max_{k \in S} (V_k(i) a_{kl}), l \in S \tag{3.7}$$

Svaki niz možemo modelirati tako da ima početno stanje, nulto stanje (0), pa stavimo da je inicijalni uvjet $V_0(0) = 1$.

Algoritam

Inicijalizacija:

$$V_0(0) = 1, V_k(0) = 0, k > 0$$

Rekurzija:

$$V_l(i) = e_l(x_i) \max_k (V_k(i-1) a_{kl}), \quad i = 1, \dots L$$
$$pointer_i(l) = \underset{k}{\operatorname{argmax}} (V_k(i-1) a_{kl}), \quad i = 1, \dots L$$

Kraj:

$$\mathbb{P}(X, \pi^*) = \max_k (V_k(L) a_{k0}) \pi_L^* = \operatorname{argmax}_k (V_k(L) a_{k0})$$

Vraćanje puta:

$$\pi_{i-1}^* = pointer_i(\pi_i^*), \quad i = L, \dots 1$$

U praksi se ovakav algoritam pokazuje često nepraktičan jer u rekurziji dolazi do množenja vrlo malih brojeva i to više puta, ovisno o duljini opaženog niza, a i sveukupnog broja stanja. Zbog toga se često javlja *underflow* i tu pojavu izbjegavamo logaritmiranjem vjerojatnosti, tj. prelaskom u *log-space*. Dakle, koristimo $\log(V_l(i))$. Log-vjerojatnost iz završnog koraka Viterbijevog algoritma nazivamo "score".

3.3 Distribucija ekstremnih vrijednosti

Neka je X slučajna varijabla koja poprima vrijednosti log-vjerojatnosti nekog prolaska kroz model. U slučaju da računamo vjerojatnosti svih mogućih prolaza kroz model vidjeli bismo da su "score"-ovi distribuirani s eksponencijalnim repom. Pokazuje se da je maksimum takve distribucije Gumbelova distribucija, koja pripada klasi distribucija ekstremnih vrijednosti.

Uzmimo sada $n \in \mathbb{N}$ takvih slučajnih varijabli, $X_1, X_2, \ldots, X_n$. Tada $(X_i, i = 1, ..., n)$ odgovara slučaju da promatramo prolaske n nizova simbola, pa možemo na to gledati kao na niz jednako distribuiranih slučajnih varijabli. Ukoliko pretpostavimo njihovu nezavisnost vrijediti će u nastavku navedeni rezultati.

Neka je $X_1, X_2, \ldots, X_n$ niz nezavisnih jednako distribuiranih slučajnih varijabli zadanih na istom vjerojatnosnom prostoru $(\Omega, \mathcal{F}, \mathbb{P})$, sa funkcijom distribucije F. Definiramo $M_n = \max(X_1, X_2, \ldots, X_n)$, ona je zaista slučajna varijabla zbog 1.1.6. M_n predstavlja maksimum procesa tijekom n ponovljenih mjerenja.

Tada vrijedi:

$$\begin{aligned} \mathbb{P}(M_n \leq z) &= \mathbb{P}(X_1 \leq z, X_2 \leq z, \ldots, X_n \leq z) \\ &= \mathbb{P}\left(\bigcap_{i=1}^{n} \{X_i \leq z\}\right) \\ &= \prod_{i=1}^{n} \mathbb{P}\left(X_i \leq z\right) \\ &= F^n(z), \qquad \forall z \in \mathbb{R}, n \in \mathbb{N} \end{aligned} \tag{3.8}$$

gdje prva jednakost slijedi iz definicije maksimuma, treća prema 1.2 i zadnja zbog pretpostavke da su jednako distribuirane. U praksi ovo nije korisno jer je distribucija F nepoznata. Mogli bismo je procijeniti, ali bi zbog potenciranja u zadnjoj jednakosti u 3.8 greška u procjeni prebrzo rasla. Dakle, to ne bi bio dobar pristup.

Teorem 3.3.1. *Neka postoje nizovi $a_n > 0$ i $b_n \in \mathbb{R}$ takvi da*

$$\mathbb{P}\left(\frac{M_n - b_n}{a_n} \leq x\right) \to G(x), \quad \textit{kada} \quad n \to \infty \tag{3.9}$$

gdje je G nedegenerirana funkcija distribucije, tada G pripada jednoj od sljedećih familija

funkcija distribucije:

$$I : G(z) = \exp\left\{-\exp\left\{-\frac{z-b}{a}\right\}\right\}, \quad z \in \mathbb{R} \tag{3.10}$$

$$II : Fr(z) = \begin{cases} 0, & z \leq b \\ \exp\left\{-\left(\frac{z-b}{a}\right)^{-\alpha}\right\}, & z > b \end{cases} \tag{3.11}$$

$$III : W(z) = \begin{cases} \exp\left\{\left(\frac{z-b}{a}\right)^{\alpha}\right\}, & z < b \\ 1, & z \geq b \end{cases} \tag{3.12}$$

$$\tag{3.13}$$

gdje su $a > 0$, b parametri, a za familije II i III imamo i parametar α.

Domena atrakcije Gumbelove distribucije

Definicija 3.3.2. *Kažemo da funkcija distribucije F leži u **domeni atrakcije** funkcije distribucije ekstremnih vrijednosti H ako postoje $a_n > 0$ i $b_n \in \mathbb{R}$, $n \geq 1$ takvi da vrijedi:*

$$\lim_{n\to\infty} F^n(a_n x + b_n) = H(x), \quad \forall x \in \mathbb{R} \tag{3.14}$$

Kao što je spomenuto na početku nas zanima Gumbelova distribucija, a kako je familija funkcija distribucija tipa I iz teorema 3.3.1 upravo Gumbelova definirat ćemo kriterij da neka funkcija distribucije bude u domeni atrakcije Gumbelove funkcije distribucije.

Neka je $x_0 = \sup_y\{F(y) < 1\}$ desni kraj funkcije distribucije. Za F iz domene atrakcije funkcije distribucije Gumbelovog tipa, $F \in D(G)$ x_0 može biti konačan ili beskonačno. Primjerice, uzmimo funkciju distribucije eksponencijalne razdiobe. $F(x) = 1 - e^{-x}, x > 0$ desni rep te funkcije je $x_0 = \infty$, a za a_n i b_n iz definicije 3.3.2 stavimo $a_n = 1$ i $b_n = \ln n$. Za $x \in \mathbb{R}$ vrijedi da je $\lim_n F^n(x + \ln n) = \lim_n \left(1 - e^{-(x+\ln n)}\right)^n = \lim_n \left(1 - \frac{e^{-x}}{n}\right)^n = exp(-exp\{-x\}) = G(x)$.

Definicija 3.3.3. *Funkcija distribucije $F_\#$ s desnim krajem x_0 je von Misesova funkcija ako postoji $z_0 < x_0$ takav da za $z_0 < x < x_0$ i $c > 0$ vrijedi*

$$1 - F_\#(x) = c \cdot \exp\left\{-\int_{z_0}^{x_0} \frac{1}{f(u)} du\right\}$$

za $f(u) > 0$, $z_0 < u < x_0$ i neprekidna na (z_0, x_0) sa funkcijom gustoće $f'(u)$ i $\lim_{u\nearrow x_0} f'(u) = 0$. f zovemo pomoćna funkcija.

Definicija 3.3.4. *Neopadajuća funkcija U je Γ-varirajuća ($U \in \Gamma$) ako je U definirana na intervalu (x_l, x_0), $x_0, x_l \in \mathbb{R}$, $x_l < x_0$, $\lim_{x\to x_0} U(x) = 0$ i ako postoji pozitivna funkcija f definirana na (x_l, x_0) takva da za svaki x vrijedi:*

$$\lim_{t\to x_0} \frac{U(t + xf(t))}{U(t)} = e^x, \tag{3.15}$$

gdje se f naziva pomoćnom funkcijom.

Definicija 3.3.5. *Nenegativna, neopadajuća funkcija V, definirana na intervalu $(z, \infty), z \in \mathbb{R}$ je Π-varirajuća ($V \in \Pi$) ako postoji pozitivna funkcija a i funkcija b takva da za svaki $x > 0$ vrijedi:*

$$\lim_{t\to x_0} \frac{V(tx) - b(t)}{a(t)} = \log x. \tag{3.16}$$

Funkcija a se naziva pomoćnom funkcijom.

Propozicija 3.3.6. *1. Ako je $F_\#$ von Misesova funkcija definirana kao u definiciji 3.3.3, tada je $F_\# \in D(G)$. Nizove $(a_n)_{n\in\mathbb{N}}$ i $(b_n)_{n\in\mathbb{N}}$ možemo definirati ovako:*

$$a_n = (1/(1 - F_\#))^{-1}(n)$$
$$b_n = f(b_n)$$

za $1/(1 - F_\#) \in \Gamma$ sa pomoćnom funkcijom f

2. Pretpostavimo da je F apsolutno neprekidna i druga derivacija F'' je negativna za sve $x \in (z_0, x_0)$. Ako je

$$\lim_{x\to x_0} F''(x)(1 - F(x))/(F'(x))^2 = -1 \tag{3.17}$$

tada je F von Misesova funkcija i $F \in D(G)$. Možemo staviti $f = (1 - F)/F'$. Obratno, von Miseseova koja je dva puta derivabilna zadovoljava 3.17.

Iskažimo sada lemu i propoziciju, bez dokaza, koje će nam biti važne za rezultat koji slijedi nakon njih.

Lema 3.3.7. *Ako su $H_n, n \geq 0$ neopadajuće funkcije i $H_n \to H_0$, onda i $H_n^{-1} \to H_0^{-1}$.*

Propozicija 3.3.8. *Ako je $U \in \Gamma$ s pomoćnom funkcijom f, tada je $U^{-1} \in \Pi$ s pomoćnom funkcijom $a(t) = f(U^{-1}(t))$. Nadalje, ako je $V \in \Pi$ s pomoćnom funkcijom $a(t)$, tada je $V^{-1} \in \Gamma$ s pomoćnom funkcijom $f(t) = a(V^1(t))$.*

Propozicija 3.3.9. *Za funkciju distribucije F stavimo $U = 1/(1 - F)$ tako da je U^{-1} definirana na $(1, \infty)$. Tada je ekvivalentno:*

i) $F \in D(G)$

ii) $U \in \Gamma$

Dokaz. Pretpostavimo da vrijedi i), tj. $F^n(a_n x + b_n) \to G(x)$, ako tu relaciju logaritmiramo i iskoristimo činjenicu da $-logz \sim 1 - z$, $z \nearrow 1$ dobivamo

$$\begin{aligned} &n(1 - F(a_n x + b_n)) \to e^{-x}, x \in \mathbb{R} \\ &n^{-1} U(a_n x + b_n) \to e^{x}, x \in \mathbb{R} \end{aligned} \tag{3.18}$$

Zbog 3.3.7 vrijedi

$$\begin{aligned} &(U^{-1}(ny) - b_n)/a_n \to \log y, y > 0 \\ &(U^{-1}(ny) - U^{-1}(n))/a_n \to \log y. \end{aligned} \tag{3.19}$$

Za $\epsilon > 0$ i dovoljno velik t, uz $a(t) > 0$ vrijedi:

$$\begin{aligned} \frac{U^{-1}(ty) - U^{-1}(t)}{a(t)} - \frac{U^{-1}(t(1+\epsilon)) - U^{-1}(t)}{a(t)} &\leq (U^{-1}(ty) - U^{-1}(t+1))/a(t) \\ &\leq (U^{-1}(ty) - U^{-1}(t+1))/a(t) \\ &\leq (U^{-1}(ty) - U^{-1}(t))/a(t) \\ &\leq (U^{-1}((t+1)y) - U^{-1}(t))/a(t) \\ &\leq (U^{-1}(ty(1+\epsilon)) - U^{-1}(t))/a(t) \end{aligned}$$

ako sada pustimo $t \to \infty$, te iskoristimo 3.19, dobivamo

$$\log y - \log(1+\epsilon) \leq \liminf_t (U(ty) - U(t))/a(t) \leq \limsup_t (U(ty) - U(t))/a(t) \leq \log y + \log(1+\epsilon).$$

Zbog proizvoljnosti od ϵ U^{-1} zadovoljava definiciju 3.3.5. Zbog propozicije 3.3.8 vrijedi $\left(U^{-1}\right)^{-1} \in \Gamma$. Nadalje, $\left(U^{-1}\right)^{-1} = \lim_{t \nearrow x} U(t) = U^{-}(x)$. Tvrdimo da $U^{-} \in \Gamma \Rightarrow U \in \Gamma$. Neka je $x \in \mathbb{R}$ i $\lim_{t \nearrow x_0} U^{-}(t + xf(t))/U^{-}(t) = e^{x}$, za $\epsilon > 0$ imamo

$$\frac{U^{-}(t + xf(t))}{U^{-}(t)} \leq \frac{U(t + xf(t))}{U^{-}(t)} \leq \frac{U^{-}(t + (x+\epsilon)f(t))}{U^{-}(t)}.$$

Pustimo $t \nearrow x_0$ i imamo

$$e^{x} \leq \liminf_{t \nearrow x_0} \frac{U(t + xf(t))}{U^{-}(t)} \leq \limsup_{t \nearrow x_0} \frac{U(t + xf(t))}{U^{-}(t)} \leq e^{x+\epsilon},$$

odnosno,

$$\lim_{t \nearrow x_0} \frac{U(t + xf(t))}{U^-(t)} = e^x, x \in \mathbb{R}. \tag{3.20}$$

Ako sada u 3.20 stavimo $x = 0$, dobivamo $U(t) \sim U^-(t)$, pa je $U \in \Gamma$, tj vrijedi *ii*).

Pretpostavimo sada da vrijedi ii) i pokažimo i). Neka vrijedi $\frac{U(t+xf(t))}{U(t)} \to e^x$. Zbog propozicije 3.3.8 vrijedi $U(U^{-1}(t)) \sim t$, pa imamo

$$\frac{U(U^{-1}(n) + xf(U^{-1}(n)))}{n} \to e^x$$

uz $a_n = f(U^{-1}(n))$ i $b_n = U^{-1}(n)$ ovo postaje

$$n(1 - F(a_n x + b_n)) \to e^{-x}$$

što je ekvivalentno 3.18, tj vrijedi i). □

Svojstva funkcije distribucije Gumbelovog tipa i procjena parametara

Kao što je već navedeno teoremom 3.3.1, Gumbelova funkcija distribucije glasi:

$$G(x) = \exp\left\{-e^{-(x-b)/a}\right\}, \quad x \in \mathbb{R}, b \in \mathbb{R}, a > 0. \tag{3.21}$$

Za $b = 0$ i $a = 1$ definirana je standardna Gumbelova distribucija. Funkcija gustoće Gumbelove slučajne varijable je

$$g(x) = \frac{1}{a} \exp\left\{-e^{-\frac{x-b}{a}} - (x-b)/a\right\}, \quad x \in \mathbb{R}, b \in \mathbb{R}, a > 0. \tag{3.22}$$

Neka je $Z = e^{-(X-b)/a}$, $a, b \in \mathbb{R}, a > 0$, gdje je X Gumbelova slučajna varijabla. Tada je Z eksponencijalno distribuirana s parametrom $\lambda = 1$ i vrijedi

$$\mathbb{E}[e^{t(X-b)/a}] = \mathbb{E}[Z^{-t}] = \Gamma(1-t), t < 1$$

Ako stavimo $t \to ta$ slijedi

$$\mathbb{E}[e^{tX}] = e^{tb}\Gamma(1-at), a|t| < 1$$

Uz oznake $\Psi(t) = \log \mathbb{E}[e^{tX}]$ i $\psi(t) = \frac{\Gamma'(t)}{\Gamma(t)}$, vrijedi

$$\Psi(t) = bt + \log\Gamma(1-at) \quad \text{i} \quad \Psi'(0) = b - a\psi(1).$$

Slučajna varijabla X ima k-ti moment jednak k-toj derivaciji funkcije Ψ u $t = 0$, pa je očekivanje Gumbelove slučajne varijable $\mathbb{E}[X] = b + a\gamma$, gdje je γ Eulerova konstanta, a varijanca $VarX = \frac{1}{6}\pi^2 a^2$.

Neka su $X_1, X_2, \ldots, X_n$ nezavisne jednako distribuirane slučajne varijable s funkcijom distribucije kao u 3.21. Metodom maksimalne vjerojdostojnosti procijenimo parametre a i b. Funkcija log-vjerodostojnosti je

$$l(a,b) = -n\log(a) - \sum_{i=1}^{n}\frac{x_i - b}{a} - \sum_{i=1}^{n}\exp\left\{-\frac{x_i - b}{a}\right\}$$

Maksimum se postiže za

$$\begin{aligned}\bar{a} &= \bar{x} - \frac{\sum_{i=1}^{n}(x_i \exp -x_i/a)}{\sum_{i=1}^{n}(\exp -x_i/a)}\\ \bar{b} &= -a\log\frac{1}{n}\sum_{i=1}^{n}\exp -n/a.\end{aligned} \tag{3.23}$$

Poglavlje 4

Višestruko poravnanje

Višestruko poravnanje proteina[1] je poravnanje tri ili više nizova proteina, odnosno rekonstrukcija evolucijske povijesti proteinske familije. Poravnanjem aminokiselina u proteinima se bavimo u svrhu usporedbe dva ili više nizova ne bismo li saznali primjerice njihovu pripadnost pojedinim proteinskim familijama ili evolucijsko svojstvo, vezu predak-potomak.

4.1 Biološko značenje poravnanja

Kod prijenosa genetskog materijala kroz generacije organizama, dolazi do promjene u sastavu nizova koji čine DNK, pa onda i u proteinima. Takve promjene u zapisu genetske informacije nazivamo mutacijama, do njih može doći iz raznih razloga. Najjednostavnija forma mutacija su točkovne mutacije, a to su supstitucija (zamjena) jedne aminokiseline drugom, insercija (umetanje neke aminokiseline) i delecija (brisanje aminokiseline). Primjerice, u nizovima

```
a) VRPS
b) VRPA
c) VRP_
```

imamo supstituciju aminokiseline S iz a) aminokiselinom A u b), deleciju prelaskom iz a) u c) ili b) i inserciju prelaskom iz c) u b) ili a). U višestrukom poravnanju proteina homologne aminokiseline su smještene u istom stupcu. Idealno bi bilo kad bismo mogli prepoznati homologne aminokiseline u evolucijskom i strukturalnom smislu. Dakle, želimo da se stupac poravnanja u trodimenzionalnim strukturama proteina nalazi u sličnom položaju, ali i da sve aminokiseline iz tog stupca potječu od zajedničke. Takvo poravnanje

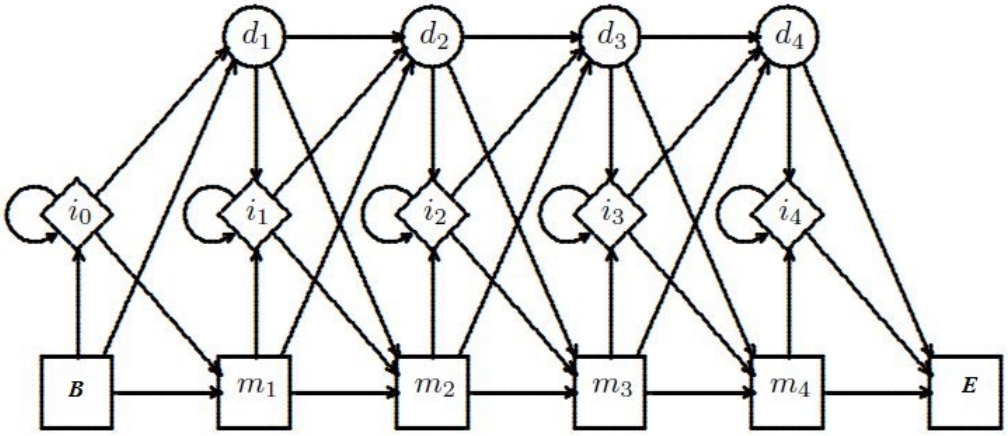
[1] eng. Multiple Sequence alignment

stručnjaci biolozi mogu ručno izvesti vrlo precizno. Zbog rubnih dijelova strukture proteina, kao što su okreti i petlje poravnanje će imati dijelove koji će jako varirati od proteina do proteina. Naime, β-ploče i α-zavojnice su dijelovi proteina koji su čvrsto vezani pa ostaju očuvani u poravnanju, te se mijenjaju uglavnom supstitucijom. Varijabilni dijelovi imaju puno više mutacija, pogotovo insercija i delecija. Očuvane regije želimo što bolje modelirati, a to je moguće upravo skrivenim Markovljevim modelom.

4.2 Poravnanje u terminima HMM-a

U prošlom poglavlju objašnjena je generalna ideja HMM-a, sada bismo htjeli u tim terminima objasniti poravnanje proteina i povezati to sa biološkim značenjem poravnanja. Pretpostavimo da je zadano točno poravnanje više nizova proteina. Za početak možemo postepeno graditi model, pa možemo pretpostaviti da je u pozadini poravnanja trivijalan HMM. Odnosno da je svako opažanje rezultat emisije iz zasebnog stanja koje nazivamo **match** stanjem. Tada imamo trivijalno poravnanje jer je vjerojatnost prelaska iz nekog stanja u sljedeće match stanje upravo 1. Nadalje, problem se javlja kod delecije jer imamo simbol '_' koji nije aminokiselina, pa ne može biti emitiran match stanjem, a analogno deleciji problem je i insercija, stoga dodajemo za svako match stanje pripadajuće **insert** i **delete** stanje. Također moramo modelirati početak i kraj niza. Stoga dodajemo **begin** B i **end** E stanje, ta stanja su "tiha" stanja, u smislu da ne emitiraju aminokiseline. Napomenimo još da su *m* (match) i *i* (insert) emitirajuća stanja, a *d* (delete) "tiho". Shema HMM-a s ovako modeliranim stanjima je prikazana na slici 4.1.

Uz ovu će shemu Viterbijev algoritam s najboljim putem kroz model davati optimalno poravnanje niza aminokiselina sa modelom.



Slika 4.1: Tranzicijska struktura HMM-a

Za potrebe daljnjeg razmatranja pretpostavimo da nam je poznat model, odnosno tranzicijske i emisijske vjerojatnosti. Zamislimo situaciju u kojoj možemo iz skrivenog Markovljevog modela ponovljeno gledati opažene nizove. Primjerice, imamo manji skup proteina reprezentativan za neku familiju. Želimo iskoristiti taj model kako bismo veći broj pripadnika te familije međusobno poravnali. Vidjeli smo kako poravnati svaki od tih nizova sa modelom, dakle, pronalaskom najvjerojatnijeg puta kroz model Viterbijevim algoritmom. Višestruko poravnanje u ovim uvjetima označava međusobno poravnanje stanja skrivenog Markovljevog modela. Poravnanje stanja se vrši tako da ukoliko smo na putu kroz model u stanju m_j onda u j-ti stupac poravnanja stavimo pripadajući niz aminokiselina, ako prolazimo kroz stanje d_j tada umjesto simbola za aminokiselinu stavimo znak '-', a kada prolazimo kroz neki insert i_j tada su u tom stupcu svi znakovi emitirani u stanju i_j. Za vizualizaciju ovog algoritma slijedi kratki primjer.

Primjer 4.2.1. *Neka su opaženi nizovi simbola* {*HPEW, PW, P, PEEW*}*. Neka je model takav da da su pripadajući Viterbijevi putevi bili* $\{(B, i_0, m_1, i_1, m_2, E)$*,* (B, m_1, m_2, E)*,* (B, m_1, d_2, E)*,* $(B, m_1, i_1, i_1, m_2, E)\}$*. Radi razlikovanja pojmova stupac poravnanja i pripadajuće stanje HMM-a označimo stupce poravnanja velikim slovima. Višestruko poravnanje koje dobijemo gore navedenim postupkom je tada:*

I_0	M_1	I_1	M_2
H	P	E	W
	P		W
	P		$-$
	P	EE	W

Insert stupce možemo poravnati proizvoljno, primjerice nalijevo. Stupce koji pripadaju pojedinom insert stanju nazivamo insert regijama. Insert regije predstavljaju dijelove proteina koji nisu tipični, nisu sačuvani te ih nema biološkog smisla poravnavati. Očekujemo da su to petlje i okreti jer su u homolognim proteinima ti dijelovi često strukturalno različiti.

Poglavlje 5

Rezultati

5.1 Procjena modela

Pretpostavimo da imamo k opaženih nizova $X_i, i = 1, \ldots k$. Kad bismo znali točan put stanja Π_i u pozadini svakog od nizova mogli bismo prebrojati frekvencije prelazaka iz stanja u stanje i frekvencije emisija pojedinih simbola, te iz njih procijeniti vjerojatnosti, tj. matrice A i E.

Kako bismo procijenili model pretpostavimo da je zadano početno poravnanje. Potrebno je odrediti stupce poravnanja koji su očuvani u $p\%$ proteina. Što znači da u $(100 - p)\%$ proteina dopuštamo deleciju. Takve stupce proglasimo match stupcima. Pokušati ćemo s različitim brojevima p, ali bilo bi poželjno da imamo što manje delecija, pa je donja granica $p > 50$.

Odavde počinjemo graditi model. Svi stupci poravnanja između dva susjedna match stupca dobiveni su iz točno određenog insert stanja. Pogledamo li shemu sa slike 4.1 vidjeti ćemo da ukoliko se nalazimo u stanju m_j jedino insert stanje koje možemo direktno posjetiti iz m_j je stanje i_j, ostala stanja u koja možemo prijeći su m_{j+1} i d_{j+1}, ali to su stanja koja se zajedno poravnavaju u sljedećem match stupcu poravnanja. Dakle, jedini izbor za insert regiju je i_j.

Na taj način odredimo za svaki simbol pripadajuće stanje u modelu. Tranzicijske vjerojatnosti ćemo odrediti računanjem frekvencija, odnosno relativnih frekvencija prelazaka iz stanja u sljedeće dopušteno stanje[1]. Dok ćemo emisijske vjerojatnosti pojedinog simbola, za match i insert stanja modela računati kao relativne frekvencije pojavljivanja pripadajućeg simbola u pojedinom stanju. Tako ćemo računati procjenjene emisijske vjerojatnosti, uz poznate frekvencije $f_{M_j}(a)$ da je u stanju j emitiran simbol a, preko procjenitelja maksimalne vjerodostojnosti:

[1] Dopušteno stanje u koje možemo prijeći iz nekog stanja HMM-a je ono za koje postoji pozitivna vjerojatnost prelaska u shemi HMM-a naznačena strelicom.

$$e_{m_j}(a) = \frac{f_{M_j(a)}}{\sum_{a'} f_{M_j(a')}} \tag{5.1}$$

Analogno, računamo i tranzicijske vjerojatnosti.

Radi što bolje procjene potrebno je dodati korekcije, pogotovo ako nemamo dovoljno podataka. Kod izračuna frekvencija prelazaka može se dogoditi da nemamo opažanja za neku tranziciju i tada koristimo jednostavan pseudozbroj. Naime, dodajemo neku konstantu na sve frekvencije čime vjerojatnosti ne mijenjamo, ali izbjegavamo dijeljenje s nulom. Ako je ta konstanta 1, takav pseudozbroj se zove Laplace-ovo pravilo. Uobičajeno je zapisivati tranzicijske vjerojatnosti u matricu oblika

A_j	m_{j+1}	i_j	d_{j+1}
m_j	$\mathbb{P}(S_{j+1} = m_{j+1} \mid S_j = m_j)$	$a_{m_j i_j}$	$a_{m_j d_{j+1}}$
i_j	$a_{i_j m_{j+1}}$	$a_{i_j i_j}$	$a_{i_j d_{j+1}}$
d_j	$a_{d_j m_{j+1}}$	$a_{d_j i_j}$	$a_{d_j d_{j+1}}$

za $j = 0, \ldots, k$ gdje je k duljina modela odnosno broj match stanja.

Navedeni postupak je izveden u Python kodu. Ulazni podaci bili su poravnanje AT-domena, priloženo pod nazivom at.aln. Izlazni podaci su parametri modela u obliku duljina modela k, A_j, E_j, $j = 0, \ldots, k$, gdje E_i čine dva vektora-retka emisijskih vjerojatnosti za pripadna stanja, redom m_j i i_j.

Slijedi isječak iz početnog poravnanja.

```
--VLFSGQGSQRPGMGRELHARYPVFAAAFDETVALLDARLGTS--------LRDIVWD---QDRTRLDDTRHTQ
--VLFSGQGSQRLGMGRELHARFPVFAEAFDEIAALLDRHTDRP--------LREVVWG---SDAELLNETGWTQ
--VLFSGQGSQRPGMGRELHARFPVFAAAFDEITALLDTHLDRP--------LREVVWG---TDADLLNDTGWAQ
---LFSGQGSQRLGMGRELYGRFPVFAEALDAVLAVLDGELEG--------SLREVMWG---EDAGLLNETGWTQ
---LFSGQGSQRLGMGRELYGRFPVFAEALDAVLAVLDGELDG--------SLREVMWG---EDAGLLNETGWTQ
---LFSGQGSQRLGMGRELYGRFPVFAEALDAVLAVLDGELGG--------SLREVMWG---EDAGLLNETGWTQ
---LFSGQGSQRLGMGRELYGRFPVFAEVLDAVLAVLDGELGG--------SLREVMWG---EDAGLLNETGWTQ
---LFSGQGSQRLGMGRELYGRVPVFTEALDAVLAVLDGELEG--------SLREVMWG---EDADLLNETGWTQ
--VLFSGQGSQRLGMGRELYERFPVFAEALDVVIDHLDAALPAQA------GLREVMWG---DDVELLNETGWTQ
--VLFTGQGAQRLGMGRELYGRFPVFAEALDVVVDHLDAALPAQA------GLREVMWG---DDVELLNETGWTQ
--VLFSGQGSQRLGMGRELYERFPVFAEALDVAIDHLDAALPAQA------SLREVMWG---DDVELLDETGWTQ
--VLFSGQGAQRLGMGRELYERFPVFAEALDVVVDHLDAALPAQA------GLREVMWG---DDAELLNETGWTQ
---LFSGQGSQRLGMGRVLYERFPAFAEALDTVLTALDAELGHP--------LRDIIWG---EDAQLVDRTGYTQ
--VLFSGQGSQQLGMGRDLYERFPAFAEAFDAVLARLDGHLGAS--------LRDVVWH---GDQETLNETGHTQ
--VLFSGQGSQRLGMGRELHERHPVFAEAFDSVLARLDDRLDTP--------LRDVVWG---TDEEALHATGNTQ
```

5.2 Simulacija i "score"-ovi

Nakon procjene modela, željeli bismo moći simulirati podatke iz procijenjenog modela. Dakle, želimo naći nove nizove koji će proširiti spektar "score"-ova početne familije. Na

način da iz distribucije prelazaka i emisija biramo na slučajan način sljedeće stanje i ukoliko je ono match ili insert također na slučajan način biramo iz distribucije emisija sljedeći simbol. Koristimo metodu inverznih transformacija.

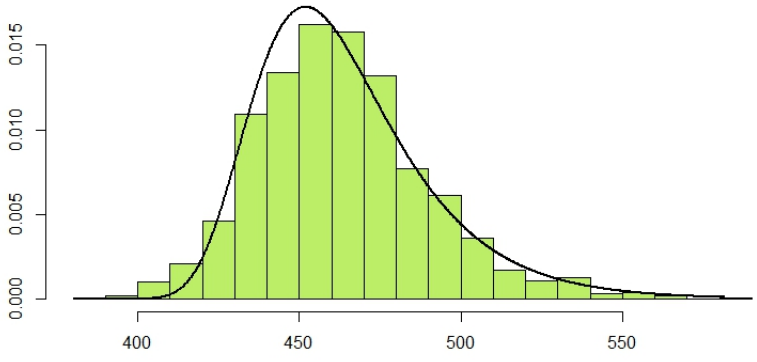
Teorem 5.2.1. *Neka je* $\{F(z), a \leq z \leq b\}$ *funkcija distribucije, s inverzom*

$$F^{-1}(u) = \inf\{z \in [a, b] : F(z) \geq u, 0 \leq u \leq 1\} \tag{5.2}$$

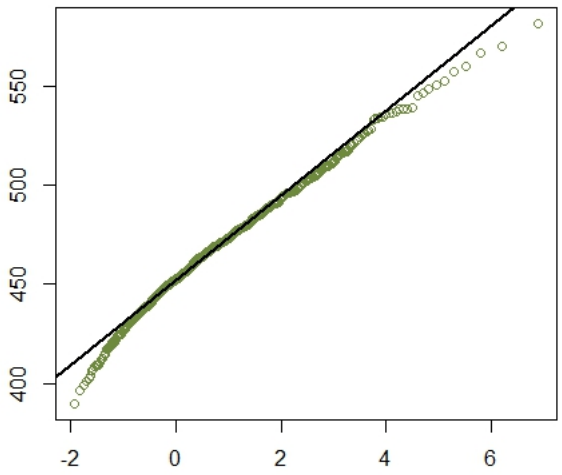
Neka je $U \sim \mathcal{U}(0, 1)$ *uniforma slučajna varijabla na segmentu* $[0, 1]$*, tada slučajna varijabla* $Z = F^{-1}(U)$ *ima funkciju distribucije* F*.*

Postupak procjene modela, opisan u 5.1, nije najbolji za potrebe simulacije novih nizova. Kada je nekoliko stupaca poravnanja potpuno očuvano, do na supstituciju, dolazi do spomenutog slučaja manjka informacija. Za simulaciju to predstavlja problem jer bi moglo doći do generiranja nizova kojima duljina vrlo jako varira, zbog čega gubimo biološki smisao. Zato je uvedena korekcija procijenjenog modela, na mjestima gdje je potrebno. Korekcija je provedena u smislu povećanja vjerojatnosti ostanka u insert stanju i prelaska iz insert stanja u match stanje, te prelaska iz delete stanja u match i delete stanje. Usporedno je smanjena vjerojatnost prelazaka iz insert u delete i obratno.

Na slici 5.2 prikazan je histogram "score"-ova 1000 simuliranih nizova i funkcija gustoće Gumbelove slučajne varijable s parametrima procijenjenima iz podataka, $\hat{a} = 21.33987$, $\hat{b} = 451.8948$. Iako su simulirani nizovi nejednake duljine, no s malim varijacijama, distribucija će biti približno Gumbelova. Usporedimo li kvantile Gumbelove distribucije i dane podatke dobivamo sljedeći Q - Q graf gdje su na x-osi prikazani kvantili podataka, a na y-osi kvantili teoretske distribucije.



Slika 5.1: Histogram simuliranih nizova iz procjenjenog modela duljine 255

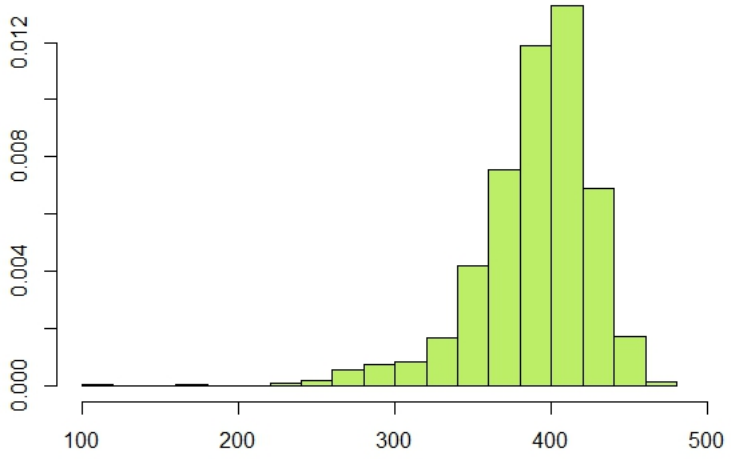


Slika 5.2: Q-Q graf

Vidimo da se kvantili dobro grupiraju oko pravca $y = \hat{a}x + \hat{b}$, stoga ne možemo odbaciti tvrdnju da maksimalni "score"-ovi slijede Gumbelovu distribuciju.

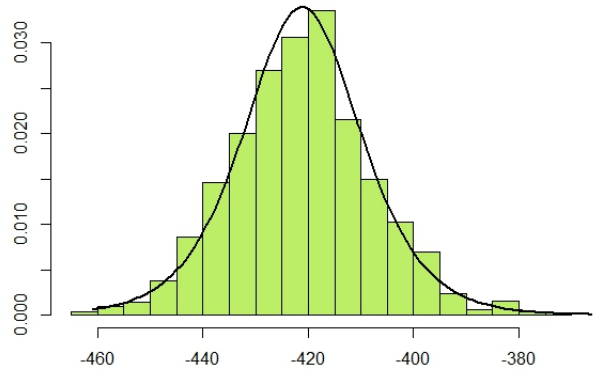
Ovakva simulacija, iako je najbolja, čini se složena i možda nepraktična, a mogla bi i predugo ostati u nekom insert stanju što nikako ne bismo željeli zbog duljine niza. Jednostavnija bi bila simulacija nizova jednake duljine, no pokazuje se da nemamo jasan alat pomoću kojeg bismo sačuvali nezavisnost unutar nizova. Jedna ideja je bila sljedeća. Ukoliko je tražena duljina nizova d, a trenutno smo emitirali d simbola, ali nismo u stanju *End*, tada simulacija tog niza završava, to je ekvivalentno prolasku kroz ostala delete stanja. Obratno, ukoliko smo emitirali manje od d simbola, a našli smo se u stanju *End*, nastavljamo s emisijom s parametrima zadnjeg insert stanja, to je ekvivalentno ostanku u zadnjem insert stanju sve dok naš niz ne napunimo do duljine d.

Takvom simulacijom dolazi do grafova koji gube smisao. Primjerice, istim modelom simuliramo 1000 nizova, ali sada fiksne duljine, za duljinu uzmemo otprilike prosjek duljina originalnih nizova. Ovdje je za duljinu određeno 260 simbola. Slijedi histogram "score"-ova tako simuliranih nizova. Vidimo sa slike 5.2 da iz histograma nije vidljiva nikakva očekivana distribucija.



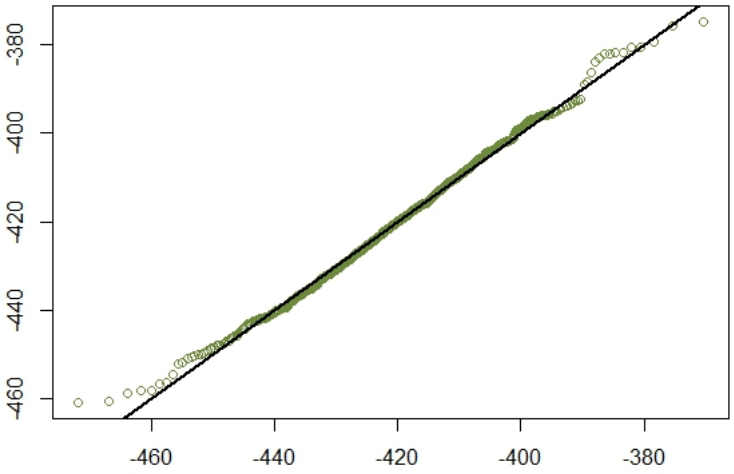
Slika 5.3: Histogram sa “score”-ovima nizova simuliranih iz modela s jednakim duljinama

Mogli bismo simulaciju nizova fiksne duljine izvesti iz neke zadane razdiobe, to je najjednostavniji način simulacije. Vrlo je intuitivno uzeti za razdiobu diskretnu distribuciju takvu da je vjerojatnost pojavljivanja svake aminokiseline jednaka, te iznosi $\frac{1}{20}$. Na taj način su dobiveni sljedeći rezultati.



Slika 5.4: Histogram sa "score"-ovima simuliranih nizova iz zadane razdiobe

Funkcija gustoće prikazana uz histogram pripada logističkoj distribuciji, s procijenjenim parametrima $\hat{\alpha} = \mu = -421.1418$ i $\hat{\beta} = \frac{\sigma\sqrt{3}}{\pi} = 7.368195$. U nastavku je i pripadajući Q-Q graf.



Slika 5.5: Q-Q graf

5.3 Iterativno poboljšanje modela

Htjeli bismo model pomoću kojega je u prethodnom poglavlju izvršena simulacija nizova upotrijebiti i u smislu poravnanja originalnih nizova. Na taj način ćemo vidjeti dobivamo li zaista traženo, da očuvani dijelovi proteina ostaju očuvani. Želimo rekonstruirati početno poravnanje prikazano na slici **??**. Ideja poravnanja nizova prema modelu opisana je Viterbijevim algoritmom, a višestruko poravnanje nizova prema modelu je poravnanje Viterbi prolazaka promatranih nizova kao što je opisano u primjeru 4.2.1 uz napomenu da insert stanja poravnavamo nalijevo, u tom slučaju bi poravnanje iz primjera 4.2.1 izgledalo kao u tablici 5.3.

H	P	E	–	W
–	P	–	–	W
–	P	–	–	–
–	P	E	E	W

Tablica 5.1: Primjer višestrukog poravnanja s poravnatim insert regijama

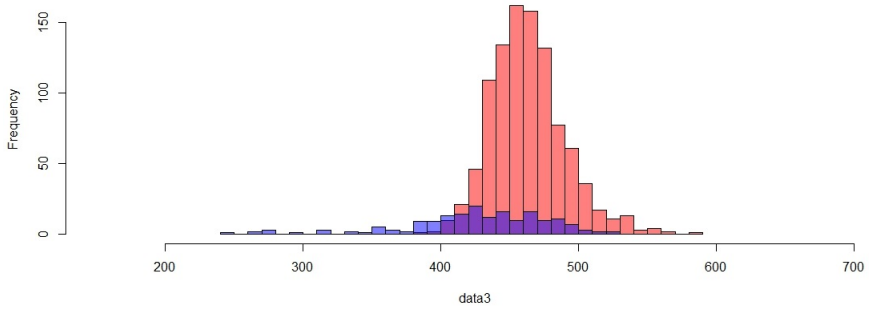
Ukoliko to napravimo direktno iz procjenjenog modela, dio poravnanja originalnih nizova je prikazan u nastavku.

```
V--LFSGQGSQRPGMGRELHARYPVFAAAFDETVALLDARLG-------TSLRDIVWDQDRTR-----------LDDTRHTQ
V--LFSGQGSQRLGMGRELHARFPVFAEAFDEIAALLDRHTD-------RPLREVVWGSDAEL-----------LNETGWTQ
V--LFSGQGSQRPGMGRELHARFPVFAAAFDEITALLDTHLD-------RPLREVVWGTDADL-----------LNDTGWAQ
---LFSGQGSQRLGMGRELYGRFPVFAEALDAVLAVLDGELE-------GSLREVMWGEDAGL-----------LNETGWTQ
---LFSGQGSQRLGMGRELYGRFPVFAEALDAVLAVLDGELD-------GSLREVMWGEDAGL-----------LNETGWTQ
---LFSGQGSQRLGMGRELYGRFPVFAEALDAVLAVLDGELG-------GSLREVMWGEDAGL-----------LNETGWTQ
---LFSGQGSQRLGMGRELYGRFPVFAEVLDAVLAVLDGELG-------GSLREVMWGEDAGL-----------LNETGWTQ
---LFSGQGSQRLGMGRELYGRVPVFTEALDAVLAVLDGELE-------GSLREVMWGEDADL-----------LNETGWTQ
V--LFSGQGSQRLGMGRELYERFPVFAEALDVVIDHLDAALP-------AQAGLREVMWGDDVEL---------LNETGWTQ
V--LFTGQGAQRLGMGRELYGRFPVFAEALDVVVDHLDAALP-------AQAGLREVMWGDDVEL---------LNETGWTQ
V--LFSGQGSQRLGMGRELYERFPVFAEALDVAIDHLDAALP-------AQASLREVMWGDDVEL---------LDETGWTQ
V--LFSGQGAQRLGMGRELYERFPVFAEALDVVVDHLDAALP-------AQAGLREVMWGDDAEL---------LNETGWTQ
---LFSGQGSQRLGMGRVLYERFPAFAEALDTVLTALDAELGHPLRDIIWGEDAQL------------------VDRTGYTQ
V--LFSGQGSQQLGMGRDLYERFPAFAEAFDAVLARLDGHLG-------ASLRDVVWHGDQET-----------LNETGHTQ
V--LFSGQGSQRLGMGRELHERHPVFAEAFDSVLARLDDRLD-------TPLRDVVWGTDEEA-----------LHATGNTQ
```

Poravnanje dobiveno ovakvom procjenom se ne razlikuje mnogo od početnog poravnanja, insert regije su poravnate drugačije, no već je rečeno da su to varijabilni dijelovi, a zato nisu ključni da bismo zaključili koliko je procijenjeni model dobar. Ukoliko pogledamo regije koje su ostale očuvane vidimo da se one nisu značajno promijenile. Dakle, model je dobro prilagođen našim podacima, odnosno nizovima.

Poboljšanje poravnanja se sastoji u tome da cijeli postupak procjene modela ponovimo, ali na način da izdvojimo nizove koji daju najveći "score" uz taj model. Izdvojimo nizove po kriteriju: ako je "score" niza bio veći od medijana "score"-ova simuliranih nizova iz modela, tada niz ulazi u odabir za novu procjenu.

Pogledajmo usporedne histograme "score"-ova simuliranih nizova i originalnih nizova. Samo mali dio najvećih "score"-ova ulazi u sljedeću procjenu.



Slika 5.6: Histogrami simuliranih i originalnih "score"-ova

Na taj način izdvojeno je 12 nizova. Prije procjene potrebno je međusobno poravnati samo njih 12. Drugim riječima, trebamo izbrisati višak crtica koje su poravnate same sa sobom. Sada možemo ući u ponovljenu procjenu. Iskoristimo algoritam opisan u poglavlju 5.1. Nakon procjene poravnamo izdvojene nizove. Novi model daje sljedeće poravnanje izdvojenih nizova.

```
VLFSGQGSQRPGMGRELAARFPVFADALDDALRALDRHLD--GPVREVMWGTDAALLDRTGWTQ
VLFSGQGSQRLGMGRELHARFPVFAEAFDEIAALLDRHTD--RPLREVVWGSDAELLNETGWTQ
VLFSGQGSQRPGMGRELHARFPVFAAAFDEITALLDTHLD--RPLREVVWGTDADLLNDTGWAQ
-LFSGQGSQRLGMGRELYGRFPVFAEALDAVLAVLDGELE--GSLREVMWGEDAGLLNETGWTQ
-LFSGQGSQRLGMGRELYGRFPVFAEALDAVLAVLDGELD--GSLREVMWGEDAGLLNETGWTQ
-LFSGQGSQRLGMGRELYGRFPVFAEALDAVLAVLDGELG--GSLREVMWGEDAGLLNETGWTQ
-LFSGQGSQRLGMGRELYGRFPVFAEVLDAVLAVLDGELG--GSLREVMWGEDAGLLNETGWTQ
-LFSGQGSQRLGMGRELYGRVPVFTEALDAVLAVLDGELE--GSLREVMWGEDADLLNETGWTQ
VLFTGQGAQRLGMGRELYGRFPVFAEALDVVVDHLDAALPAQAGLREVMWGDDVELLNETGWTQ
-LFSGQGAQRLGMGRELHARFPVFAEALDQVLDLLDEELD--ASLGDIIWGEEEAPLNETGFTQ
-LFSGQGAQRLGMGRELHARFPVFARALDTAVDLLDAELG--GTLREVIWGTDDAPLNETGFTQ
VLFSGQGSQRIGMGRELSGRYPVFAEAFDTVCAALDEHLD--RPLRDVVRGEDEELLNRTVYAQ
```

Novo poravnanje ima jako očuvane regije α-zavojnica i β-ploča. Ono što bismo mogli zaključiti iz ovog poravnanja je da vidimo zapravo kostur početne familije proteina.

Bibliografija

[1] J.M. Berg, J.L. Tymoczko, and L. Stryer. *Biochemistry, Fifth Edition*. W.H. Freeman, 2002.

[2] Laurens de Haan and Ana Ferreira. *Extreme Value Theory: An Introduction (Springer Series in Operations Research)*. Springer, 1 edition, 2006.

[3] Richard Durbin, Sean Eddy, Anders Krogh, and Graeme Mitchison. *Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge Univ*. Cambridge University Press, 1998.

[4] George Fishman. *Monte Carlo*. Springer Series in Operations Research and Financial Engineering. Springer, 1996.

[5] Sidney I. Resnick. *Extreme values, regular variation and point processes*. Springer, 2008.

[6] N. Sarapa. *Teorija vjerojatnosti*. Udžbenici Sveučiliš ta u Zagrebu. Školska knjiga, 1992.

[7] Silvija Vrbančić. *Lokalno poravnanje i prepoznavanje motiva, diplomski rad*. PMF-MO, Zagreb, 2014.

Sažetak

Višestruko poravnavanje je važan objekt u bioinformatici jer daje puno informacija proteinskim familijama. U ovom radu smo vidjeli kako napraviti višestruko poravnanje pomoću skrivenog Markovljevog modela. Pokazuje se da su rezultati vrlo osjetljivi i ovisni o uzorku. Provedena je analiza kojom je iz početnog poravnanja procijenjen model, zatim je na nekoliko načina provedena simulacija i pokazalo se da distribucija "score"-ova jest Gumbelova kako smo i očekivali. Naposljetku je napravljeno novo poravnanje. Provedena parametrizacija modela je vrlo osjetljiva, pa dobiveni model ne omogućava daljnju analizu. Zbog toga pokušavamo postepeno graditi model od najboljih poravnanja, koja su i najmanje varijabilna, a kako smo vidjeli to je dobar način da izbjegnemo neke od problema na koje smo naišli.

Summary

Multiple sequence alignment is an important object in bioinformatics for obtaining information about protein families. In this thesis we show how to build a multiple sequence alignment using hidden Markov models. We have observed that the results are very sensitive to the choice of various parameters and sample biased. Analysis carried out consists of model estimation from given alignment, simulation and realigning. Distribution of scores is approximately Gumbel, as expected. Since parametrisation of a family profile is a very sensitive procedure, we gradually build a model using less variable subsamples. This method provides a good solution to avoid some of the obstacles we encountered.

Životopis

Rođena sam 12.08.1990 godine u Karlovcu. Osnovno školovanje sam započela u osnovnoj školi ”Vladimira Nazora” u Dugoj Resi. Nastavljam sa srednjoškolskim obrazovanjem u ”Srednjoj školi Duga Resa” smjer opća gimnazija. Nakon toga, 2009. godine upisujem preddiplomski studij matematike, na Prirodoslovno-matematičkom fakultetu u Zagrebu. Završetkom preddiplomskog studija 2013. godine nastavljam školovanje upisivanjem diplomskog studija statistike na Prirodoslovno-matematičkom fakultetu.