SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Iva Sorić

STATISTIČKI ASPEKTI PRETRAŽIVANJA PROTEOMA

Diplomski rad

Voditelj rada:
Doc. dr. sc. Pavle Goldstein

Zagreb, rujan 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. _____

Mojim roditeljima.

Sadržaj

Uvod

Jedno od važnijih pitanja u bioinformatici je pitanje pripadnosti proteina nekoj proteinskoj familiji. U ovom radu bit će opisana jedna od metoda čija je svrha identifikacija nizova koji pripadaju familiji od interesa - motif scanning. Ova metoda koristi se kada su nizovi iz familije globalno vrlo varijabilni, a karakteriziraju ih specifični motivi.

Objasnit ćemo razvoj skrivenog Markovljevog modela dizajniranog za traženje motiva te implementirati Viterbi algoritam za računanje "score"-ova kao ocjene podudaranja niza s modelom. Provest će se statistička analiza "score"-ova na simuliranim podacima te konstruirati korekcija za log-odds ratio. Korigirani "score"-ovi trebali bi poslužiti kao dobar kriterij za diskriminaciju željenih nizova.

Metodu na kraju testiramo na stvarnim biljnim proteomima za traženje članova GDSL familije te analiziramo dobivene rezultate.

Poglavlje 1

Uvodni pojmovi iz vjerojatnosti

Definicija 1.0.1. *Neka je Ω proizvoljan neprazan skup i $\mathcal{F}$ σ-algebra na skupu Ω. Uređen par $(\Omega, \mathcal{F})$ zove se izmjeriv prostor.*

Definicija 1.0.2. *Neka je $(\Omega, \mathcal{F})$ izmjeriv prostor. Funkcija $\mathbb{P} : \mathcal{F} \to \mathbb{R}$ je vjerojatnost (na $\mathcal{F}$) ako vrijedi*
$(P1)$ $\mathbb{P}(A) \geqslant 0$, $A \in \mathcal{F}$; $\mathbb{P}(\Omega) = 1$
$(P2)$ $A_i \in \mathcal{F}$, $i \in \mathbb{N}$ i $A_i \cap A_j = 0$ *za* $i \neq j \Rightarrow \mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$.

Definicija 1.0.3. *Uređena trojka $(\Omega, \mathcal{F}, \mathbb{P})$, gdje je $\mathcal{F}$ σ-algebra na Ω i $\mathbb{P}$ vjerojatnost na $\mathcal{F}$, zove se vjerojatnosni prostor.*

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

Definicija 1.0.4. *Neka je $(\Omega, \mathcal{F}, \mathbb{P})$ vjerojatnosni prostor. Funkcija $X : \Omega \to \mathbb{R}$ je slučajna varijabla ako je $X^{-1}(B) \in \mathcal{F}$ za proizvoljno $B \in \mathcal{B}$ tj. $X^{-1}(\mathcal{B}) \subset \mathcal{F}$.*

Skup Ω na kojem je X definirana može biti sasvim općenit, ali ako nas zanima problem vezan za određenu slučajnu varijablu X, pogodnije je operirati s vjerojatnosnim prostorom koji je induciran s X.
Za $B \in \mathcal{B}$ stavimo

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

Relacijom (1.1) definirana je funkcija $\mathbb{P}_X : \mathcal{B} \to [0, 1]$ i to je vjerojatnosna mjera na $\mathcal{B}$. $\mathbb{P}_X$ zovemo vjerojatnosna mjera inducirana s X, a vjerojatnosni prostor $(\mathbb{R}, \mathcal{B}, \mathbb{P}_X)$ zovemo vjerojatnosni prostor induciran s X.

Prema tome, svakoj slučajnoj varijabli X se preko relacije (1.1) na prirodan način pridružuje vjerojatnosni prostor $(\mathbb{R}, \mathcal{B}, \mathbb{P}_X)$ i problemi vezani za slučajnu varijablu X rješavaju se u okviru toga vjerojatnosnog prostora. $\mathbb{P}_X$ često zovemo i zakon razdiobe od X.

Definicija 1.0.5. *Neka je X slučajna varijabla na Ω. Funkcija distribucije od X je funkcija $F_X : \mathbb{R} \to [0, 1]$ definirana s*

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

Često se stavlja $F_X = F$, ako je jasno o kojoj se slučajnoj varijabli radi.

Definicija 1.0.6. *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 1.0.7. *Slučajna varijabla X je apsolutno neprekidna ili, kraće, neprekidna slučajna varijabla ako postoji nenegativna realna Borelova funkcija f na $\mathbb{R}$ ($f : \mathbb{R} \to \mathbb{R}_+$) takva da je*

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

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

Matematičko očekivanje i varijanca

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

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

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

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

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

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

Definicija 1.0.8. *Neka $\mathbb{E}X$ postoji tj. konačno je. Tada $\mathbb{E}[(X - \mathbb{E}X)^r]$ zovemo r-ti centralni moment od X.*

Definicija 1.0.9. *Varijanca od X, u oznaci $VarX$ ili σ_X^2, je drugi centralni moment od X, tj.*

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

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

Uvjetna vjerojatnost i nezavisnost

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

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

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

Definicija 1.0.10. *Neka su $A, B \in \mathcal{F}$. Događaji A i B su nezavisni ako vrijedi*

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B).$$

1.1 Gumbelova distribucija

Funkciju distribucije ekstremnih vrijednosti tipa III zovemo Gumbelov tip distribucije. Ime je dobila prema Emilu Gumbelu (1891. - 1966.), njemačkom matematičaru koji se bavio modeliranjem ekstremnih vrijednosti u području strojarstva i meteorologije, a u teoriji vjerojatnosti i statistici koristi se za modeliranje maksimuma (ili minimuma) uzoraka različitih distribucija.

Gumbelova funkcija distribucije u općenitom obliku:

$$F(x) = \exp\{-e^{-(x-\mu)/\sigma}\}, \quad \mu, \sigma \in \mathbb{R}, \ \sigma > 0 \tag{1.3}$$

Funkcija gustoće f Gumbelove distribucije ima oblik

$$f(x) = \sigma^{-1} \exp\{-e^{-(x-\mu)/\sigma} - (x - \mu)/\sigma\}.$$

Uzimanjem $\mu = 0$ i $\sigma = 1$, dobivamo standardnu Gumbelovu funkciju distribucije

$$F(x) = \exp\{-e^{-x}\}, \quad x \in \mathbb{R}.$$

Funkcija gustoće tako definirane slučajne varijable glasi

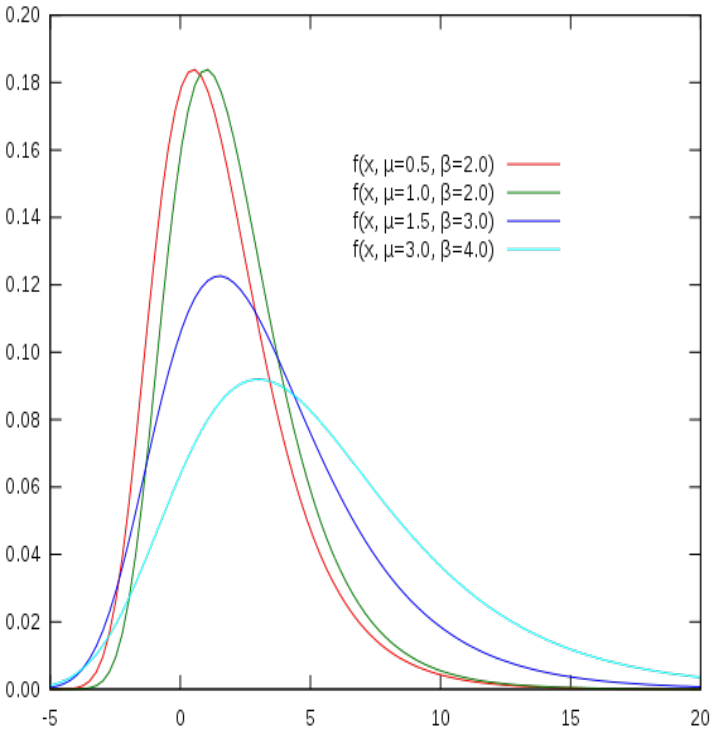
$$f(x) = \exp\{-x - e^{-x}\}, \quad x \in \mathbb{R}.$$

Očekivanje Gumbel distribuirane slučajne varijable X je

$$\mathbb{E}[X] = \mu + \sigma\gamma,$$

gdje je γ Eulerova konstanta, $\gamma \sim 0.5772$, a varijanca je jednaka

$$VarX = \frac{1}{6}\pi^2\sigma^2.$$



Slika 1.1: Funkcija gustoće

Procjena parametara distribucije

Metodom maksimalne vjerodostojnosti želimo procijeniti parametre distribucije, μ i σ. Dakle, tražimo maksimum log-vjerodostojnosti

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

Maksimum se postiže za

$$\begin{aligned} \hat{\mu} &= \sigma \log \frac{1}{n} \sum_{i=1}^{n} \exp\{-n/\sigma\} \\ \hat{\sigma} &= \bar{x} - \frac{\sum_{i=1}^{n} (x_i \exp\{-x_i/\sigma\})}{\sum_{i=1}^{n} \exp\{-x_i/\sigma\}} \end{aligned} \tag{1.4}$$

Poglavlje 2

Skriveni Markovljev model

2.1 Definicija

Osnovna razlika između klasičnog Markovljevog modela i skrivenog Markovljevog modela ili HMM-a (hidden Markov model) je u tome što u HMM-u ne postoji 1-1 korespondencija između stanja i opažanja. Poznat nam je samo niz opažanja, dok je niz stanja skriven i treba ga procijeniti. Niz stanja HMM-a modeliran je Markovljevim lancem 1. reda, a opažanja su međusobno nezavisna.

Definicija 2.1.1. *Skriveni Markovljev model (HMM) zadan je s dva niza slučajnih varijabli,* $Q = Q_1, Q_2, ..., Q_N$ *i* $X = X_1, X_2, ..., X_N$*, takva da vrijedi:*

$$\mathbb{P}(Q_t|Q_{t-1}, Q_{t-2}, ..., Q_1) = \mathbb{P}(Q_t|Q_{t-1}) \tag{2.1}$$

$$\mathbb{P}(X_t|Q_t, Q_{t-1}, X_{t-1}, ..., Q_1, X_1) = \mathbb{P}(X_t|Q_t). \tag{2.2}$$

Niz Q predstavlja niz stanja, a X niz opažanja (simbola). Simboli su međusobno nezavisni i ovise samo o stanju u kojem se emitiraju. Vjerojatnosti emitiranja simbola u određenom stanju zovu se *emisijske vjerojatnosti* i modelirane su relacijom (2.2). Vjerojatnosti prelaska iz jednog stanja u drugo zovu se *tranzicijske vjerojatnosti* i za njih vrijedi Markovljevo svojstvo (2.1); vjerojatnost određenog stanja ovisi samo o prethodnom stanju.

Skriveni Markovljev model ima sljedeće parametre:

- skup stanja $S = \{1, ..., N\}$, broj stanja N
- skup opažanja (simbola) $O = \{b_1, ..., b_M\}$, broj mogućih opažanja M

• matrica tranzicijskih vjerojatnosti $A = [a_{ij}]$,

$$a_{ij} = \mathbb{P}(Q_t = j | Q_{t-1} = i),\ 1 \leq i, j \leq N$$

• matrica emisijskih vjerojatnosti $E = [e_j(k)]$,

$$e_j(k) = \mathbb{P}(X_t = b_k | Q_t = j),\ 1 \leq k \leq M,\ 1 \leq j \leq N$$

2.2 Viterbi

Da bismo opaženom nizu simbola pridružili odgovarajući niz "skrivenih" stanja, koristimo dekodiranje Viterbijevim algoritmom. Neka je x opaženi niz. Niz stanja zovemo put i označavamo sa π. Optimalni put π^*, $\pi^* = \arg\max_{\pi} \mathbb{P}(x, \pi)$, računa se rekurzivno:

Neka je $v_k(i)$ vjerojatnost optimalnog puta koji završava u stanju k, a pri tome su emitirani simbol $x_1, ..., x_i$. Vrijedi

$$v_l(i+1) = e_l(x_{i+1}) \max(v_k(i)\, a_{kl})$$

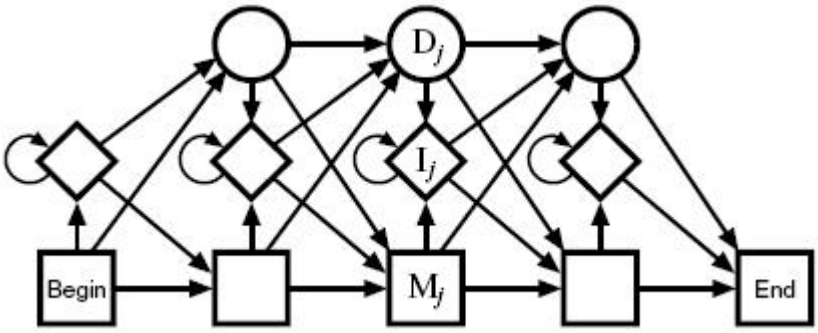
(ako je stanje l emitirajuće stanje) uz inicijalni uvjet $v_0(0) = 1$.
Viterbi algoritam računa π^* - optimalan prolaz niza kroz model i "score" - vjerojatnost tog puta.

Kod implementacije algoritma javlja se problem - tzv. underflow error. Množenjem velikog broja vjerojatnosti dobiva se izrazito mali broj zbog čega na računalu dolazi do pogreške. Iz tog razloga se Viterbijev algoritam računa u "log space"-u odnosno računa se $\log(v_k(i))$. Koristeći logaritam svih vjerojatnosti: $\tilde{a}_{kl} = \log a_{kl}$, te $\tilde{e}_l(x_i) = \log e_l(x_i)$, rekurzija u algoritmu postaje

$$V_l(i+1) = \tilde{e}_l(x_{i+1}) + \max_k(V_k(i) + \tilde{a}_{kl}).$$

2.3 Profile HMM

Model koji koristimo je specijalan tip skrivenog Markovljevog modela i njegova najpopularnija primjena u molekularnoj biologiji - profile HMM. Njegova osnovna svrha je detektiranje nizova koji pripadaju određenoj familiji od interesa. Model se gradi na temelju višestrukog poravnanja nekog uzorka nizova iz te familije i koristi za pronalazak novih nizova koji ostvaruju značajno podudaranje s modelom tj. visoki "score". Model ima ponavljajuću strukturu stanja, ali s različitim vjerojatnostima na različitim pozicijama.



Slika 2.1: Struktura općenitog HMM-a

Stanja u donjem redu zovu se *match* stanja, ona modeliraju stupce poravnanja. Rombovima su označena *insert* stanja - njih uvodimo radi dijelova niza koji se ne podudaraju s modelom, dok *delete* stanja, označena krugovima, predstavljaju dijelove višestrukog poravnanja koji se ne podudaraju s niti jednim dijelom niza. To su tzv. "tiha" stanja, ona ne emitiraju simbole. Početak i kraj modeliramo tako da uvodimo *begin* stanje s oznakom $\mathcal{B}$ i *end* stanje s oznakom $\mathcal{E}$. To su također "tiha", odnosno neemitirajuća stanja koja jednostavno služe kao početna i završna točka. Duljinu modela definiramo kao broj match stanja.

Nas zanima specifičan slučaj kada je globalna sličnost nizova iz iste familije vrlo mala, ali su prisutni karakteristični motivi. Kako su nizovi jako varijabilni, gotovo ih je nemoguće kvalitetno poravnati, odnosno kvaliteta njihovog višestrukog poravnanja bila bi neočekivano niska s obzirom na to da pripadaju istoj familiji. U takvim slučajevima, traženje proteina koji pripadaju familiji od interesa temelji se na traženju motiva koji ih karakteriziraju. Uzimajući u obzir da se proteini s vremenom mijenjaju odnosno sakupljaju mutacije, traženje motiva u nizovima puno je drugačije od jednostavnog traženja substringova, te zahtijeva sofisticiranije metode. Jedna od tih metoda je *profile HMM* prilagođen traženim motivima, i to je model kakav ćemo koristiti u ovom radu.

2.4 Parametrizacija modela

Nakon određivanja strukture modela, potrebno je procijeniti parametre - tranzicijske i emisijske vjerojatnosti te duljinu modela. Parametrizacija kreće od osnovnog skupa podataka - određenog broja nizova pripadnika familije od interesa. U njihovom višestrukom poravnanju uočavaju se karakteristični motivi. Te dijelove poravnanja predstavljaju match stanja u modelu, a time je određena i duljina modela kao broj match stanja. Stupci poravnanja

koriste se za određivanje emisijskih i tranzicijskih vjerojatnosti.

Za procjenu parametara modela, uz matematičke alate, potrebno je i stručno biološko znanje. Parametrizacija za model koji se koristi u ovom radu preuzeta je iz članka [3], ali je radi potpunosti ovdje opisujemo.

Model koji ćemo koristiti prilagođen je GDSL proteinskoj familiji. Osnovni skup podataka za parametrizaciju čine 23 niza eksperimentalno utvrđenih GDSL enzima. Njihovo višestruko poravnanje daje 5 karakterističnih motiva, a izabrani su očuvaniji blokovi - blokovi I, III i V duljina 10, 9, 9 respektivno. Za određivanje emisijskih i tranzicijskih vjerojatnosti korišteni su stupci poravnanja.

Emisijske vjerojatnosti

Neka je $P = [p_{ls}]$ matrica mutacijskih vjerojatnosti konstruirana iz BLOSUM 50 matrice. p_{ls} je vjerojatnost da aminokiselina l nakon određenog vremena mutira u aminokiselinu s. Neka je $F_R^j = [f_k^j]$ vektor relativnih frekvencija aminokiseline u j-tom stupcu poravnanja motiva, $j \in \{1, ..., n\}$, $k \in \{1, ..., 20\}$, gdje je n duljina motiva. Zbog male veličine osnovnog skupa podataka, relativnim frekvencijama se dodaje pseudo-zbroj 10^{-2}. Time dobivamo vektore $\hat{F}_R^j = [\hat{f}_k^j]$, gdje je $\hat{f}_k^j = \frac{f_k^j + 0.01}{1.2}$. Emisijska vjerojatnost aminokiseline k u j-tom stupcu motiva je

$$e^j(k) = \sum_{l=1}^{20} p_{lk} \cdot \hat{f}_l^j.$$

Kako j-ti stupac poravnanja odgovara HMM stanju M_j, to je također emisijska vjerojatnost $e_j(k)$. Svim insert stanjima dodijeljene su jednake emisijske vjerojatnosti - 0.05.

Tranzicijske vjerojatnosti

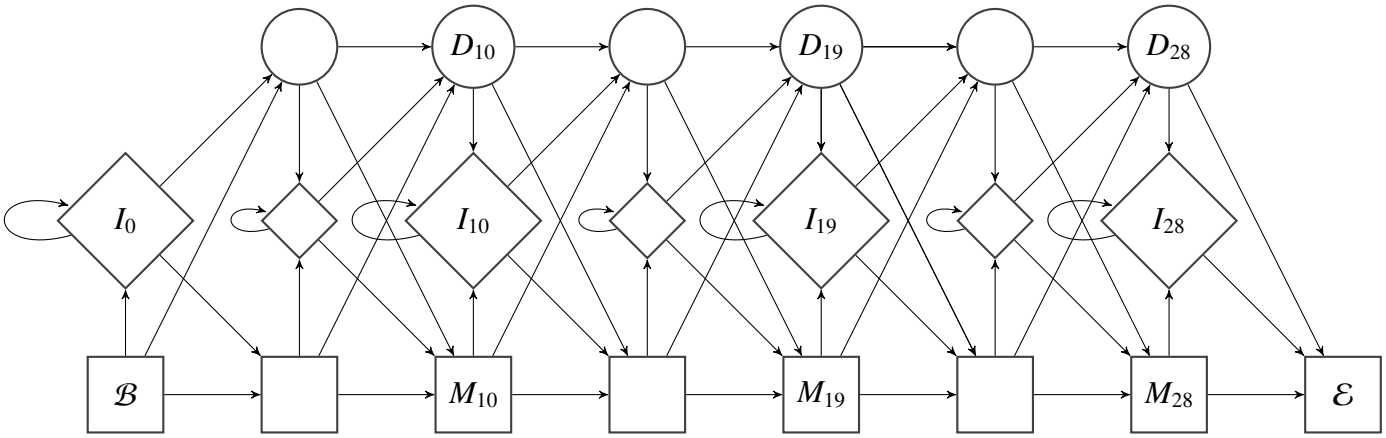
Testiranjem je utvrđeno da su najbolje tranzicijske vjerojatnosti sljedeće:

$$t_{M_iM_{i+1}} = 0.99 \quad t_{M_iI_i} = 0.005 \quad t_{M_iD_{i+1}} = 0.005$$

$$t_{I_iM_{i+1}} = 0.99 \quad t_{I_iI_i} = 0.01 \quad t_{I_iD_{i+1}} = 0$$

$$t_{D_iM_{i+1}} = 0.99 \quad t_{D_iI_i} = 0 \quad t_{D_iD_{i+1}} = 0.01$$

Tranzicijske vjerojatnosti iz match u insert/delete stanje su jako male jer insercije i delecije nisu očekivane u GDSL blokovima. Međutim, između motiva očekujemo veliki broj insercija, pa te tranzicije imaju velike vjerojatnosti: $t_{I_0 I_0} = 0.99$, $t_{I_{10} I_{10}} = t_{I_{19} I_{19}} = t_{I_{28} I_{28}} = 0.999$.



Slika 2.2: Struktura dobivenog HMM-a

Ovako konstruiran model označit ćemo s M.

Poglavlje 3

Korekcija

Proteini ili bjelančevine su makromolekule koje se sastoje od jednog ili više lanaca aminokiselina. U živim organizmima obavljaju mnogo različitih funkcija uključujući ubrzavanje metaboličkih reakcija, replikaciju DNK, odgovaranje na podražaje, transport molekula itd. Proteini su sastavni dijelovi svake stanice. U jednoj stanici neke vrste može biti oko tisuću različitih molekula proteina. Izgrađeni su od dvadeset različitih aminokiselina, međusobno povezanih poput karika u lancu.

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

Tablica 3.1: Aminokiseline i njihove kratice

Proteom je skup svih proteina koje organizam proizvodi tijekom života.

3.1 Simulacija proteoma

Neka je $\mathcal{A}$ skup aminokiselina

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

te q distribucija iz koje simuliramo

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

Vjerojatnost pojavljivanja i-te aminokiseline iz $\mathcal{A}$ je $q(i)$, $i = 1, 2, ..., 20$. Vektor q dobiven je kao vektor relativnih frekvencija u proteomima nekog većeg skupa organizama. Neka je p kumulativna distribucija; $p(1) = 0$, $p(i + 1) = \sum_{k=1}^{i} q(k)$, $i = 1, 2, ..., 20$.

$$p = (0,0.078, 0.129, 0.172, 0.225, 0.244, 0.287, 0.35, 0.422, 0.445,\\ 0.498, 0.589, 0.648, 0.67, 0.709, 0.761, 0.829, 0.888, 0.902, 0.934, 1)$$

Ako želimo simulirati niz duljine n odnosno protein koji sadrži n aminokiselina, tada n puta generiramo slučajan realan broj x iz $[0, 1)$. Ako je $p(i) \leq x < p(i + 1)$, tada nizu pridružujemo i-tu aminokiselinu.

U ovom radu razmatrat ćemo dvije simulacije proteoma:

• proteom ima 10000 proteina jednake duljine - svaki protein se sastoji od 1000 aminokiselina
• proteom ima 27000 proteina različitih duljina - duljine od 200 do 1500, s korakom 50

3.2 Log-odds ratio

Da bismo ocijenili koliko dobro niz odgovara modelu, koristimo Viterbijev algoritam koji nam daje optimalan put π^* zajedno s njegovom vjerojatnošću $\mathbb{P}(x, \pi^*|M)$. U praksi, rezultat koji želimo razmatrati kod ocjenjivanja podudaranja s modelom je ta vjerojatost u odnosu na općeniti model R tj. *log-odds ratio*

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

Kada niz dobro odgovara našem modelu M, "log-odds score" je visok, a kada odgovara nul-modelu bolje, "log-odds" je negativan.

Za općeniti model R uzimamo model iste duljine ($n = 28$) čije su emisije uniformno distribuirane - sve emisijske vjerojatnosti su $\frac{1}{20}$, a tranzicijske vjerojatnosti su sljedeće:

$$t_{M_iM_{i+1}} = 0.8 \quad t_{M_iI_i} = 0.1 \quad t_{M_iD_{i+1}} = 0.1$$

$$t_{I_iM_{i+1}} = 0.7 \quad t_{I_iI_i} = 0.2 \quad t_{I_iD_{i+1}} = 0.1$$

$$t_{D_iM_{i+1}} = 0.7 \quad t_{D_iI_i} = 0.1 \quad t_{D_iD_{i+1}} = 0.2$$

Vjerojatnosti tranzicija u match stanja su najveće jer se očekuje grupiranje match stanja u blokove.

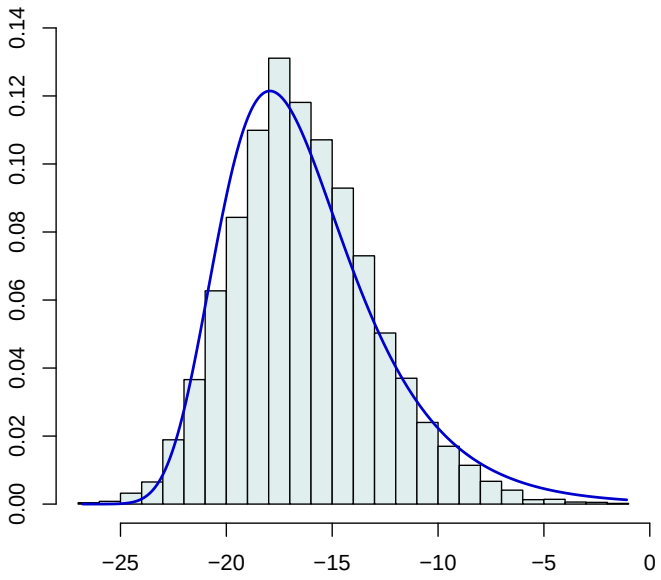
Proteom s jednakim duljinama

Za simulirani proteom razmatramo "log-score"-ove izračunate Viterbijevim algoritmom za svaki niz. Kako je korekcija za svaku emitiranu aminokiselinu u nizu $\frac{1}{20}$, korigirani "score"-ovi su

$$\log \mathbb{P}(x|M) - \log\left(\left(\frac{1}{20}\right)^{|x|}\right), \tag{3.1}$$

gdje je $|x|$ duljina niza. Pokaže se da oni slijede Gumbelovu distribuciju. Procijenimo parametre μ i σ Gumbelove distribucije; iz (1.4) slijedi $\hat{\mu} = -17.94067$, $\hat{\sigma} = 3.02852$.

Na slici (3.1) prikazani su "log-score"-ovi nakon korekcije (3.1) i funkcija gustoće Gumbelove distribucije.

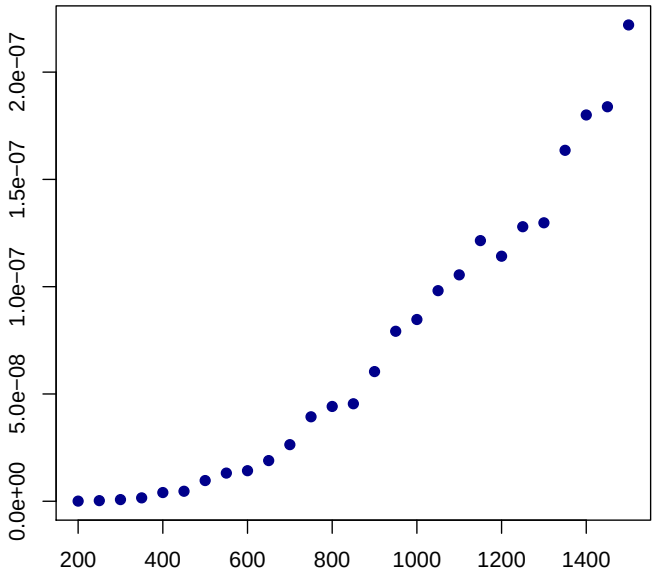


Slika 3.1: Histogram i Gumbelova razdioba

Međutim, za nizove različitih duljina ne dobivamo slične rezultate, pa možemo zaključiti da "score"-ovi ovise o duljini niza.

Proteom s različitim duljinama

Za svaku duljinu generiramo 1000 nizova. Kako rezultat ne bi ovisio o jednoj simulaciji, uzimamo prosječni "log-score" za svaku duljinu, te ih za bolji uvid eksponenciramo.

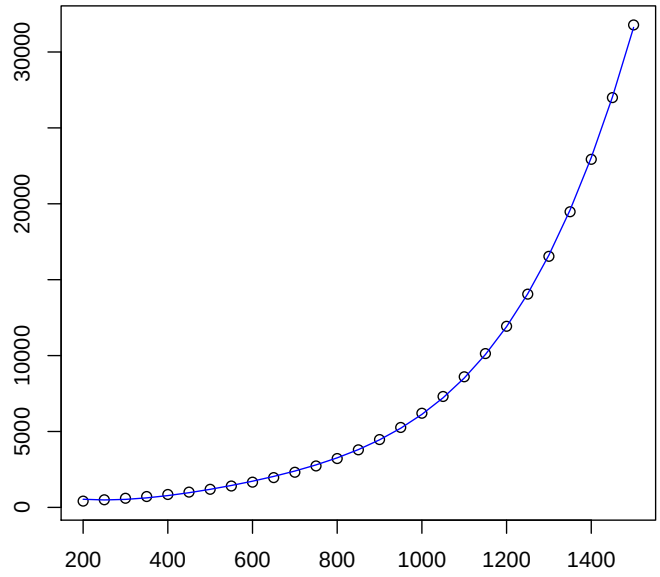


Slika 3.2: Prikaz prosječnih "score"-ova u ovisnosti o duljinama

Već iz slike je jasno da će se kod procjene metodom najmanjih kvadrata pojaviti heteroskedastičnost odnosno rast varijance s rastom duljine niza. Zbog toga "log-score"-ove dijelimo s duljinom niza te računamo korekciju za potpuni log-odds. "Score"-ovima oduzimamo i korekciju za tranzicijske vjerojatnosti. Kako su svi simulirani nizovi jako dugi u odnosu na model, to se odnosi na tranzicije $M \to M$, kakvih ima 25, te tranzicije $I \to I$ - takvih je $|x| - 26$. Dobivamo

$$\log \frac{\mathbb{P}(x|M)}{\mathbb{P}(x|R)} = \log \mathbb{P}(x|M) - \log\left(\left(\frac{1}{20}\right)^{|x|}\right) - \log\left(0.8^{25} p^{|x|-26}\right). \tag{3.2}$$

Pri tome je $p = 0.99675$ što je iznos uprosječenih vjerojatnosti tranzicija $I_0 \to I_0$, $I_{10} \to I_{10}$, $I_{19} \to I_{19}$ te $I_{28} \to I_{28}$ u modelu; time želimo dobiti jednaku ocjenu za kraće i duže nizove. Zatim metodom najmanjih kvadrata procijenimo polinom koji najbolje opisuje podatke. Dobiven je polinom 4. stupnja koji je zajedno s podacima prikazan na slici (3.3).



Slika 3.3: Procjena korigiranih “score”-ova polinomom

Koeficijent determinacije modela (R^2) iznosi 99%, što znači da je model izvrsno prilagođen podacima. Pri tome uzimamo u obzir da se radi o simuliranim, a ne stvarnim podacima, ali možemo zaključiti da je korekcija (3.2) korektna.

Dakle, iako smo na početku pretpostavili da duljina nizova igra važnu ulogu u distribuciji “score”-ova, ovaj rezultat pokazuje da korekcija za duljinu nije potrebna, već je dovoljno u potpunosti uzeti u obzir slučajni model.

Poglavlje 4

Proteomi biljaka

4.1 Motif scanning

Kada su nizovi koji pripadaju familiji od interesa vrlo varijabilni i pokazuju samo značajniju lokalnu sličnost, a karakterizirani su s konzerviranim motivima, metoda koja se pokazala dobra za njihovu diskriminaciju je *HMM motif scanning*. Metoda se temelji na skrivenom Markovljevom modelu, kakav je opisan u Poglavlju 2, za traženje karakterističnih motiva. Optimalan prolaz niza kroz model odnosno njegov "score" daje nam informaciju o tome koliko dobro niz odgovara modelu, dakle s kolikom sigurnošću možemo zaključiti da sadrži tražene motive. Metodu smo testirali tražeći proteine koji pripadaju GDSL familiji u proteomima četiri modelne biljke čiji su GDSL enzimi otprije poznati.

GDSL familija uključuje hidrolitičke enzime s multifunkcionalnim svojstvima i velikim potencijalom za primjenu u prehrambenoj i farmaceutskoj industriji. Ubrzan razvoj biotehnologije potaknuo je potragu za novim enzimima s korisnim svojstvima, a novootkriveno obilje GDSL enzima u biljnom svijetu indicira da bi biljke mogle biti dobar novi izvor tih enzima. Stoga je potraga za novim GDSL enzimima u biljnom carstvu od općeg interesa.

Tipičan niz koji pripada ovoj familiji karakteriziran je s 5 motiva, a izabrani su manje varijabilni blokovi I, III i V. Opisana metoda primijenjena je na četiri biljna proteoma s ciljem identifikacije nizova koji sadrže tražene motive. Pomoću Viterbijevog algoritma izgrađenog skrivenog Markovljevog modela, računali smo "score"-ove odnosno korigirane "score"-ove nizova i na taj način pokušali identificirati hidrolitičke enzime koristeći "score" kao kriterij.

4.2 Arabidopsis thaliana

Arabidopsis thaliana je mala biljka s cvjetovima. Autohtona je biljka Europe, Azije i sjeverozapadne Afrike, ali je naturalizirana i u mnogim drugim zemljama. Arabidopsis je godišnja biljka, obično naraste 20-25 cm u visinu. Iako nije od velike agronomske važnosti, dobro je istražena i popularna je za istraživanja u molekularnoj biologiji i genetici. Zbog toga što je diploid i ima genom male veličine, pogodna je za genetsko sekvenciranje.

Proteom ove biljke ima 35176 nizova. Koristeći Viterbi algoritam, odredili smo "score" za svaki od tih nizova, te napravili rangiranu listu nizova prema korigiranim "score"-ovima. Dobivene rezultate usporedili smo s listom otprije poznatih GDSL enzima iz ovog proteoma.

U svrhu analize našeg rangiranja definiramo pojmove TP (true positives) i TN (true negatives) kao GDSL nizove koje je ova metoda uspješno identificirala, odnosno nizove koji ne pripadaju GDSL familiji, a naša metoda ih točno svrstava među "negativce". Dodatno definiramo FP (false positives) - nizove koje motif scanning detektira kao "pozitivce", a ne pripadaju familiji lipaza, te FN (false negatives) - GDSL enzime koje ova metoda svrstava ispod praga. Za analizu učinkovitosti metode koriste se, između ostaloga, prediktivne vrijednosti. Pozitivna prediktivna vrijednost (PPV) otkriva koliki postotak pozitivnih nizova pripada GDSL familiji, dok negativna prediktivna vrijednost (NPV) otkriva koliki postotak nizova s negativnim rezultatom nema tražena svojstva.

Napomena 4.2.1. *Određivanje praga koji bi dijelio nizove čiji bi se "score" smatrao pozitivnim rezultatom i one negativne je teorija za sebe, pa analizu provodimo gledajući gdje se poznati GDSL enzimi nalaze na našoj listi tj. koliko je dobar ovaj način rangiranja. Kada bismo imali prag, mogli bismo odrediti točan broj TP, FP, TN, FN, odnosno sve mjere učinkovitosti metode. Međutim, za to bi bilo potrebno dodatno istraživanje.*

U proteomu biljke arabidopsis thaliana biolozi su pronašli 114 GDSL enzima, a na našem su popisu oni rangirani jako visoko. Točnije, tih 114 proteina nalazi se na prvih 118 mjesta. False positives se među njima nalaze na 23., 66., 76. i 88. mjestu. Ako gledamo prvih 50 nizova, naš model je uspješno identificirao 49 od 50 GDSL enzima, što znači da je PPV (positive predictive value) za ovaj slučaj 98%. Uzimajući u obzir prvih 100 nizova, taj iznos je 96/100 tj. 96%.

Kako je općenito broj pripadnih GDSL enzima izrazito mali u odnosu na veličinu biljnog proteoma, velika većina nizova bit će "negativci". Stoga NPV (negative predictive value) nije pogodan parametar za ocjenu učinkovitosti metode. U takvim slučajevima korisnije je pogledati FDR (false discovery rate), $FDR = \frac{FP}{TP+FP} = 1 - PPV$.

Kada bismo postavili prag na točno 118 nizova, PPV bi iznosila $114/118 = 96.61\%$, a FDR samo 3.39% što je jako dobar rezultat.

Neke standardne mjere točnosti, uz prag na 120 nizova:

$TP = 114$	$FP = 6$	$\longrightarrow$ $PPV = 95\%$
$FN = 0$	$TN = 35056$	$\longrightarrow$ $NPV = 100\%$
$\downarrow$	$\downarrow$	
$TPR = 100\%$	$TNR = 99.98\%$	

TPR (true positive rate) = TP/(TP+FN)
TNR (true negative rate) = TN/(TN+FP)
PPV (positive predictive value) = TP/(TP+FP)
NPV (negative predictive value) = TN/(TN+FN)

4.3 Oryza sativa

Oryza sativa je žitarica poznata kao azijska riža ili samo riža. Jednogodišnja je biljka, naraste do 1.8 m u visinu. *Oryza* je latinska riječ za rižu, a *sativa* znači kultivirana. Poznata je po tome da ju je lako genetski modificirati. U biologiji često služi kao modelni organizam.

Slično kao za arabidopsis, sada analiziramo rang listu rižinih proteina dobivenu pomoću korigiranih "score"-ova. Od 50 najviše rangiranih nizova, 43 su GDSL enzimi. Dakle, PPV za ovaj slučaj iznosi 86%. Računajući prvih 100 proteina, njih 91 pripada GDSL familiji, dok se u prvih 150 nalazi njih 125. Poznato je 126 GDSL enzima ovog proteoma, a na našoj listi se oni nalaze na prvih 163 mjesta. Ako pretpostavimo da je prag upravo 163, dobivamo da je PPV 77.3%, a FDR 22.7%.

4.4 Populus trichocarpa

Populus trichocarpa tj. kalifornijska topola je širokolisno listopadno stablo iz Sjeverne Amerike. Može narasti 30-50 m u visinu, a značajna je kao modelna biljka. Njen puni genom objavljen je 2006. godine. To je prva vrsta stabla čiji je genom sekvenciran.

Od 50 proteina populus trichocarpe s najvišim "score"-om, svih 50 su poznati GDSL enzimi. Dakle na vrhu rang liste nalaze se upravo traženi proteini. Ako pogledamo i sljedećih 50 mjesta, naša metoda uspješno je identificirala 90 enzima odnosno ostvaruje PPV 90%. Svi ranije utvrđeni GDSL enzimi populus trichocarpe, njih 97, nalaze se na prvih 126 mjesta na listi. Uzimajući za prag upravo 126, dobivamo PPV = 76.98%, te FDR = 23.02%.

4.5 Sorghum

Sorghum je rod brojnih biljnih vrsta iz porodice trava. Neke od njih se uzgajaju kao žitarice ili kao stočna hrana ili za proizvodnju sirupa i alkoholnih pića. Uglavnom se uzgaja radi žita u toplim klimama diljem svijeta.

U proteomu sorghuma nalaze se 104 otprije poznata GDSL enzima. Naša metoda na prvih 50 mjesta liste smješta samo tražene enzime, dok se unutar prvih 100 nalazi samo jedan FP. Najniže rangirani GDSL enzim nalazi se na 107. mjestu, što znači da se do tog mjesta nalaze samo 3 FP-a. Kada bismo prag postavili na 107, PPV bi iznosila $104/107 = 97.2\%$, dok gledajući 150 nizova s najvišim "score"-om dobivamo PPV = 69.33%.

U sva četiri slučaja traženi proteini bili su vrlo visoko rangirani na našoj listi. To pokazuje da ova metoda ima potencijal za korištenje u pretraživanju proteoma radi identifikacije enzima s hidrolitičkim svojstvima.

4.6 P - vrijednosti

Drugi način za rangiranje nizova je pomoću p-vrijednosti. Za svaki niz odnosno njegov score pogledamo simulaciju nizova odgovarajuće duljine te iz "fitane" distribucije odredimo p-vrijednost.

Simulirani su nizovi duljina od 50 do 1500 s korakom 50 - za svaku duljinu po 1000 nizova. Kao što je u Poglavlju 3 utvrđeno, korigirani "score"-ovi tih nizova slijede Gumbelovu distribuciju. Proteom biljke arabidopsis thaliana ima 35176 nizova različitih duljina - od samo 16 do 5393 aminokiselina. Za svaki niz iz proteoma pogledamo simulaciju najbliže odgovarajuće duljine odnosno "fitanu" distribuciju, te odredimo p-vrijednost, dakle $\mathbb{P}(score \geqslant x)$, gdje je x "score" danog niza. Koristeći Gumbelovu funkciju distribucije (1.3), to je

$$p = \mathbb{P}(score \geqslant x) = 1 - \exp\{-e^{-(x-\mu)/\sigma}\},$$

gdje su μ i σ parametri odgovarajuće distribucije. Tako određene p-vrijednosti daju ranglistu nizova od kojih bi oni s najmanjim p-vrijednostima trebali biti *positives* (P). Dobivena lista ne razlikuje se značajno od one dobivene samo sa "score"-ovima. Iako redoslijed jest nešto drugačiji, točnost ostaje otprilike jednaka; 114 poznatih GDSL enzima iz ovog proteoma nalazi se među prvih 119 nizova na listi. Dobivene p-vrijednosti su jako male. Najmanja (najviše rangirani niz) iznosi $1.03 \cdot 10^{-7}$, a zadnji TP ima p-vrijednost 0.00024.

U bioinformatici je uobičajeno uz p-vrijednost gledati i e-vrijednost, $E = p * D$, gdje je D veličina uzorka (broj nizova). Ona ponekad ima prednost pred p-vrijednosti jer je manje osjetljiva na duljinu niza koja može jako varirati, a interpretira se kao očekivani broj nizova sa "score"-om jednakim ili većim od opaženog, koji se može i samim slučajem pojaviti u pretraživanju uzorka dane veličine. Što je veći "score", to je manja e-vrijednost. Mala e-vrijednost znači da su male šanse slučajnog pojavljivanja niza sa sličnim "score"-om, a time je značajnost opaženog podudaranja veća. Za proteom arabidopsis thaliana e-vrijednosti su, za TP nizove, od $3.59 \cdot 10^{-3}$ za najviše rangiranog do 8.41 za najniže rangiranog TP.

	protein	score
1	11795	35.13
2	27336	34.02
3	17627	33.76
4	26000	33.25
5	1625	31.94
6	24616	31.36
7	3054	30.74
8	7195	30.67
9	3055	30.62
10	33101	30.25
⋮	⋮	⋮
118	31062	12.24
119	3449	12.10
120	4957	11.65
⋮	⋮	⋮

Tablica 4.1: Lista prema score-ovima

	protein	p-vrijednost	e-vrijednost
1	11795	$1.03 \cdot 10^{-7}$	$3.59 \cdot 10^{-3}$
2	27336	$1.47 \cdot 10^{-7}$	$5.13 \cdot 10^{-3}$
3	17627	$1.59 \cdot 10^{-7}$	$5.55 \cdot 10^{-3}$
4	26000	$1.87 \cdot 10^{-7}$	$6.25 \cdot 10^{-3}$
5	1625	$2.85 \cdot 10^{-7}$	$9.94 \cdot 10^{-3}$
6	24616	$3.43 \cdot 10^{-7}$	$1.2 \cdot 10^{-2}$
7	7195	$4.27 \cdot 10^{-7}$	$1.49 \cdot 10^{-2}$
8	3055	$4.33 \cdot 10^{-7}$	$1.51 \cdot 10^{-2}$
9	24584	$6.06 \cdot 10^{-7}$	$2.11 \cdot 10^{-2}$
10	17628	$6.58 \cdot 10^{-7}$	$2.3 \cdot 10^{-2}$
⋮	⋮	⋮	⋮
118	8703	0.00022	7.55
119	31062	0.00024	8.41
120	17743	0.00029	10.12
⋮	⋮	⋮	⋮

Tablica 4.2: Lista prema p-vrijednostima

Bibliografija

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

[2] N. Sarapa, *Teorija vjerojatnosti*, Školska knjiga, 1987.

[3] I. Vujaklija, A. Bielen, T. Paradžik, S. Biđin, P. Goldstein, D. Vujaklija, *Limitations of family profile search: a case study of GDSL enzyme annotation*, preprint, 2015.

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

Sažetak

Opisan je razvoj skrivenog Markovljevog modela za traženje motiva karakterističnih za članove određene proteinske familije. Nakon analize "score"-ova dobivenih na simuliranim proteomima konstruirana je korekcija tako da korigirani "score"-ovi služe kao kriterij za identifikaciju nizova koji sadrže tražene motive. Metoda je primijenjena na četiri biljna proteoma s pozitivnim rezultatima.

Summary

In this work we study one of the methods used for identifying proteins characterized by specific motifs. We have described a hidden Markov model designed for searching for motifs. After analyzing the scores obtained on simulated data, we have constructed a score correction that should yield a valid criterion for discriminating sequences with wanted motifs. In the end we apply this method to four plant proteomes with positive results.

Životopis

Rođena sam 02. lipnja 1991. godine u Zagrebu. Školovanje sam započela u Osnovnoj školi Savski Gaj, koju sam pohađala od 1998. do 2006. godine, i nastavila u V. gimnaziji u Zagrebu, gdje sam maturirala 2010. godine. Nakon toga upisala sam Preddiplomski studij Matematika na Prirodoslovno-matematičkom fakultetu u Zagrebu. Završetkom preddiplomskog studija 2013. godine upisala sam Diplomski studij Matematička statistika također na Prirodoslovno-matematičkom fakultetu u Zagrebu.