SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

SEMINARSKI RAD

IMPLEMENTACIJA SKRIVENIH MARKOVLJEVIH MODELA U DOMENSKOJ KLASIFIKACIJI PROTEINSKIH SEKVENCI

Domagoj Sviličić

Voditelj: doc. dr. sc. Krešimir Križanović

Zagreb, svibanj, 2025.

Implementacija skrivenih Markovljevih modela u domenskoj klasifikaciji proteinskih sekvenci

Domagoj Sviličić

Sažetak

U ovom radu istražena je primjena i implementacija skrivenih Markovljevih modela (HMM) u postupku klasifikacije nepoznatih proteinskih sekvenci odgovarajućim proteinskim domenama. HMM-ovi su statistički modeli koji nalaze široku primjenu u analizi sekvencionalnih podataka, a posebno su korisni u bioinformatici i biostatistici za prepoznavanje i klasifikaciju proteina. Iz perspektive strojnog učenja HMM-ovi pripadaju skupini algoritama probabilističkih grafičkih modela i karakterizira ih mogućnost efikasne ugradnje vanjskog znanja bez isključivog oslanjanja na podatke u skupu za učenje. U ovom radu implementirani su HMM-ovi za klasifikaciju različitih proteinskih domena koristeći skup podataka iz javno dostupne baze podataka Pfam. Analizirane su performanse modela i uspoređujemo ih s drugim metodama klasifikacije. Rezultati pokazuju da HMM može postići visoku točnost u klasifikaciji proteinskih domena, čime se potvrđuje njegova korisnost u bioinformatici.

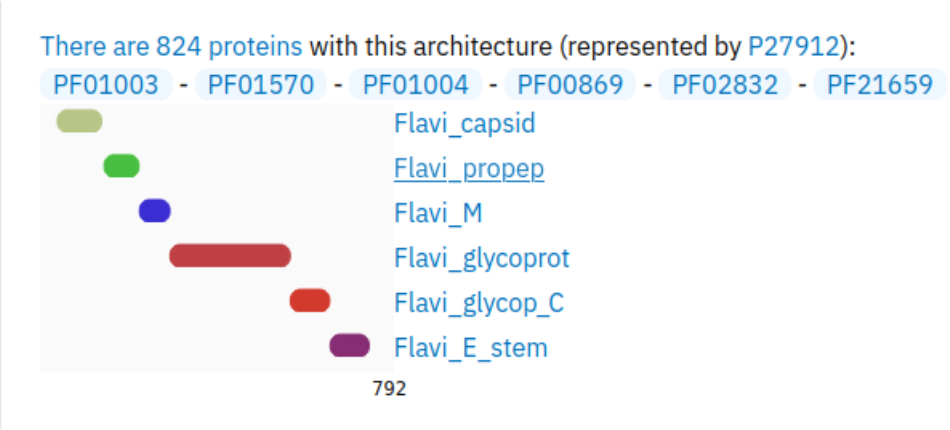
Ključne riječi: skriveni Markovljev model, klasifikacija, proteinske domene, bioinformatika, strojno učenje, biostatistika, Pfam

Sadržaj

1. Uvod

Određivanje kojoj proteinskoj skupini pripada nepoznata proteinska sekvenca predstavlja temeljni zadatak u bioinformatici s brojnim primjenama – od razumijevanja funkcije proteina do otkrivanja evolucijskih veza između organizama. Iako se na prvi pogled može činiti da je riječ o jednostavnom zadatku usporedbe, problem je u svojoj srži vrlo kompleksan. Protein je u svojoj funkcionalnoj stvarnosti trodimenzionalna i kompleksno savijena molekula, no u biološkom i bioinformatičkom kontekstu polazi se od njegove primarne strukture – linearne sekvence aminokiselina iz koje nastaje. Ta se sekvenca zapisuje kao niz znakova, gdje svaki znak označava jednu od dvadeset standardnih aminokiselina (npr. A za alanin, K za lizin, F za fenilalanin itd.). Primjer takve sekvence može izgledati ovako: MKTAYIAKQRQISFVKSHFSRQLEERLGLIEVQANLQEL.... Iako se time gubi trodimenzionalna informacija o stvarnom obliku proteina, ovaj niz sadrži ključne podatke koji određuju kako će se protein savijati i kakvu će funkciju imati, te se u mnogim analitičkim pristupima pokazuje kao dovoljan za donošenje zaključaka o pripadnosti obitelji, funkciji i evolucijskoj povezanosti. Proteinske sekvence mogu biti vrlo različite i varijabilne, čak i unutar iste funkcionalne skupine, zbog čega trivijalne metode pretraživanja poput direktnog uparivanja sekvenci često nisu dovoljne za pouzdanu klasifikaciju.

Domena je dio proteina (sekvenca aminokiselina) koji se stabilno savija u određeni 3D oblik i ima biološku funkciju, kao što je vezanje DNA, enzimska aktivnost, ili interakcija s drugim proteinima. Može postojati samostalno ili u kombinaciji s drugim domenama, što dodatno objašnjava zašto proteini mogu sadržavati više različitih domena. Na slici 1.1. prikazan je opis proteinske arhitekture iz baze Pfam.



Slika 1.1. Grafički prikaz proteinskih domena u jednom proteinu u bazi Pfam

Na primjer, humani protein BRCA1 sadrži više funkcionalnih domena, uključujući BRCT domenu (PF00533) i RING-finger domenu (PF00097). Svaka od tih domena se zasebno modelira kroz vlastita SEED poravnanja, čime se postiže preciznije prepoznavanje i anotacija funkcionalnih regija u proteinima.

U tom kontekstu, profilni skriveni Markovljevi modeli (Hidden Markov models - HMM-ovi) pokazuju iznimnu snagu. Oni kombiniraju statistički model učenja s biološki relevantnim predznanjem, koje se ugrađuje kroz algoritme višestrukog poravnanja sekvenci (MSA - Multiple Sequence Alignment). Ova sinergija omogućuje modelu da „nauči“ obrasce specifične za pojedine proteinske obitelji, ne samo na temelju sirovih sekvenci, već i na temelju njihove konzerviranosti, umetanja i delecija koje su evolucijski informativne.

Multiple sequence alignment predstavlja ključni korak u izgradnji profilnog HMM-a jer osigurava temelj za procjenu strukture modela – identificiraju se pozicije u sekvenci koje su konzervirane među članovima obitelji, što sugerira njihovu biološku važnost. U praksi, MSA algoritmi poravnavaju više srodnih sekvenci tako da se maksimizira njihova podudarnost kroz cijelu duljinu, pri čemu se uvodi i penaliziranje za umetke i delecije. Takav rezultat nije samo vizualno informativan, već i kvantitativno vrijedan jer služi za izračun parametara HMM-a: matrice emisije (koja opisuje koje se aminokiseline najvjerojatnije emitiraju na svakoj poziciji) i matrice prijelaza (koja definira vjerojatnosti prelaska iz jednog skrivenog stanja u drugo).

Time se u HMM zapravo ugrađuje vanjsko znanje – filogenetski i funkcionalni obrasci

detektirani višestrukim poravnanjem – što HMM čini znatno informiranijim i robusnijim modelom od klasičnih pristupa strojnom učenju koji počinju „od nule“. Svaka proteinska obitelj dobiva svoj profilni HMM, čime se modelira zajednički uzorak sekvenci unutar obitelji.

Kada se zatim analizira nova, nepoznata sekvenca, računa se njezina logaritamska vjerojatnost pod svakim profilnim HMM-om – koristeći forward algoritam. Korišten je i trivijalni slučajni pozadinski model koji predstavlja referentu točku u odnosu na koju se može reći koliko bolje konstruirani HMM opisuje ispitnu sekvencu u odnosu na slučajni model. Na taj način se može odrediti kojoj obitelji pripada nepoznata sekvenca.

Takav pristup omogućuje visoku preciznost klasifikacije i duboko razumijevanje filogenetskih odnosa među sekvencama, dok istovremeno koristi provjerene metode statističkog modeliranja i biološke interpretacije.

2. Priprema podataka

Polazna točka za izgradnju profilnog HMM-a za pojedinu proteinsku obitelj jest skup proteinskih sekvenci poravnatih pomoću algoritma višestrukog poravnanja sekvenci (MSA – Multiple Sequence Alignment). Ovaj postupak omogućuje identifikaciju konzerviranih regija unutar obitelji, koje su od presudne važnosti za konstrukciju modela jer upravo te regije nose evolucijski i funkcionalno najrelevantnije informacije.

U ovom radu korišten je skup podataka iz javno dostupne baze Pfam, koja sadrži velik broj proteinskih domena zajedno s reprezentativnim sekvencama za svaku od njih. Jedna od glavnih prednosti Pfam baze je dostupnost tzv. seed poravnanja – pažljivo odabranih i ručno kuriranih reprezentativnih sekvenci koje najbolje opisuju svoju obitelj, već poravnatih MSA algoritmom. Time se osigurava da su ulazni podaci visoke kvalitete, filogenetski raznoliki i biološki relevantni.

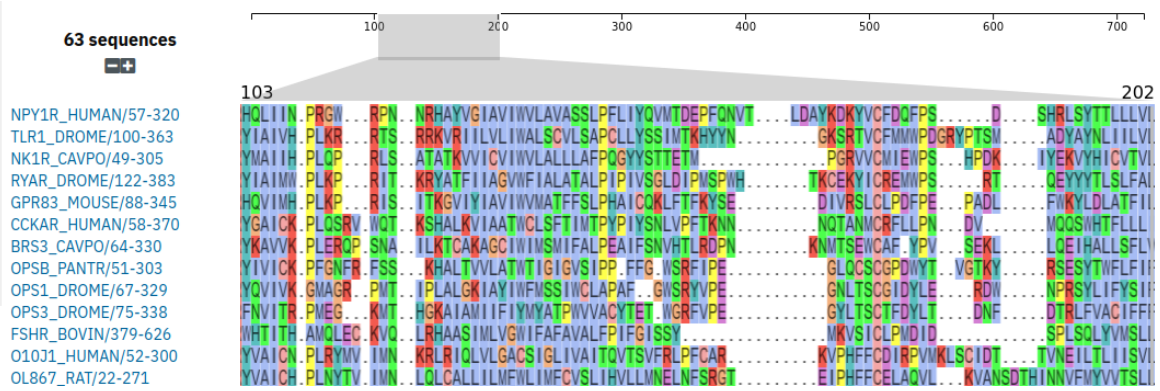
U ovoj implementaciji parametri HMM-a – emisijske i prijelazne vjerojatnosti – određuju se eksplicitno na temelju rezultata MSA poravnanja, računajući frekvencije pojedinih aminokiselina i prijelaza među stanjima izravno iz poravnatog skupa podataka. Ovakav pristup naziva se eksplicitnim treniranjem, jer model ne „nagađa“ parametre već ih izračunava direktno iz poznatih uzoraka.

Suprotno tome, kod implicitnog (iterativnog) treniranja, model se inicijalizira s početnim (često nasumičnim) parametrima, a zatim se postupno optimizira pomoću algoritama poput EM (Expectation-Maximization) ili Baum-Welch metode. Taj se pristup koristi kada nema dostupnog poravnanja ili kada želimo model trenirati na velikim skupovima sirovih, neporavnatih sekvenci. On je fleksibilniji, ali i računalno zahtjevniji te skloniji lokalnim minimumima.

Eksplicitni pristup, korišten u ovom radu, prikladniji je kada su dostupni visoko kva-

litetni i već poravnati podaci, kao što je slučaj s seed sekvencama iz Pfam baze. Takav pristup omogućuje potpunu kontrolu nad strukturom modela, veću transparentnost u načinu postavljanja parametara, i preciznu vezu između bioloških uzoraka i stohastičkog modela.

U nastavku je prikazan odsječak iz Pfam baze koji sadrži podatke o proteinskoj obitelji PF00001 u tzv. Stockholm formatu, koja se odnosi na domenu proteina poznatih kao "Rhodopsin-transmembrane receptors". Primjer sa slike 2.1.ilustrira kako su sekvence već poravnate i spremne za daljnju analizu.



Slika 2.1. Prikaz poravnatih seed sekvenci iz Pfam baze podataka za obitelj PF00001 ("Rhodopsin-transmembrane receptors")

U svrhu izrade ovog rada odabrano je deset proteinskih domena iz Pfam baze koje su filogenetski više i manje udaljene. Ove obitelji su odabrane kako bi se testirala sposobnost HMM-a da klasificira sekvence iz različitih evolucijskih skupina. Za svaku obitelj su iz poravnatih seed sekvenci izračunati parametri HMM-a što će biti detaljno opisano u nastavku.

3. Skriveni Markovljevi modeli (HMM-ovi)

Skriveni Markovljevi modeli (HMM) predstavljaju temeljnu klasu probabilističkih grafičkih modela koji omogućuju modeliranje sekvencijalnih podataka u kontekstu neizravnog promatranja. Ključna ideja HMM-a je da se opažena sekvenca generira iz niza latentnih (skrivenih) stanja koja slijede Markovljevski proces prvog reda, pri čemu svako stanje emitira simbole prema određenoj emisijskoj distribuciji [1]. Ta struktura čini HMM posebno pogodnim za slučajeve u kojima postoji podloga "skrivene" strukture (npr. strukturne regije proteina), dok su stvarna opažanja (npr. niz aminokiselina) izravno dostupna.

Formalno, HMM je definiran trojkom $(\mathbf{A}, \mathbf{B}, \pi)$, gdje $\mathbf{A}$ označava matricu prijelaznih vjerojatnosti između skrivenih stanja, $\mathbf{B}$ predstavlja skup emisijskih vjerojatnosti za opažene simbole u svakom stanju, dok je π inicijalna distribucija vjerojatnosti nad stanjima. Pretpostavka Markovljeve ovisnosti znači da je vjerojatnost prelaska u neko stanje ovisna isključivo o prethodnom stanju, dok su opažanja uvjetno neovisna kada se zna trenutno stanje.

Kao što je opisano u Durbinovoj literaturi [1], HMM-ovi su naročito korisni u bioinformatici jer omogućuju elegantno modeliranje bioloških nizova u kojima je sekvencijalna struktura rezultat skrivenih bioloških procesa, primjerice konzerviranih regija u proteinima, umetanja i delecija, ili strukturnih elemenata poput heliksa i petlji. U tim se slučajevima HMM-ovi koriste za opisivanje obiteljskih obrazaca putem profilnih modela ("profile HMMs"), gdje se sekvence poravnavaju kako bi se identificirale informativne regije koje zatim određuju arhitekturu modela.

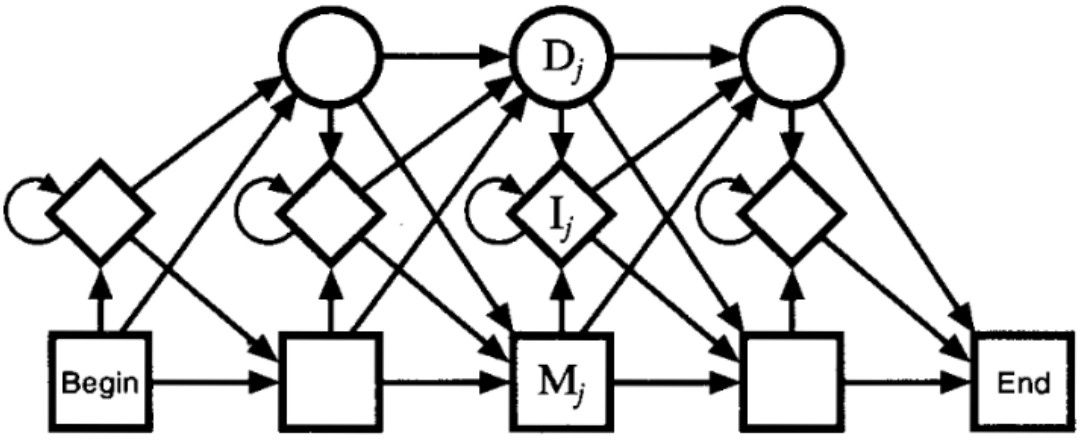
Za rad s HMM-ovima ključni su algoritmi koji omogućuju analizu i treniranje mo-

dela: Forward algoritam omogućuje računanje vjerojatnosti promatranog niza s obzirom na model; Viterbijev algoritam pronalazi najvjerojatniji slijed skrivenih stanja; dok Baum-Welch algoritam (poseban slučaj algoritma očekivanje-maksimizacija) služi za estimaciju parametara modela iz sirovih podataka kada nije dostupna referentna struktura.

HMM-ovi imaju široku primjenu izvan bioinformatike – od automatskog prepoznavanja govora i prirodnog jezika, do prepoznavanja obrazaca, detekcije anomalija i vremenskih serija u financijama. No upravo u analizi bioloških sekvenci njihova primjena ima duboko teorijsko utemeljenje, jer kombinira stohastičko modeliranje s evolucijskom i strukturnom biologijom.

4. Postupak konstrukcije profilnog HMM-a nad proteinskom obitelji

Konstrukcija profilnog HMM-a počinje se od poravnatih sekvenci iz Pfam baze koja nudi preuzimanje alignment.seed datoteka (2.1.) za svaku pojedinu proteinsku domenu, koje su već prošle kroz višestruko poravnanje. MSA algoritam osigurava da su sekvence usklađene na način koji maksimizira sličnost među njima, a sposoban je i razdvojiti podudarne regije uvođenjem praznina. Udio tih praznina u pojedinom stupcu poravnanja bit će ključan kriterij za utvrđivanje konzerviranosti dotičnog stupca. Iz poravnatih sekvenci moguće je definirati Insert, Delete i Match stanja u postupku konstrukcije profilnog HMM-a za dotičnu proteinsku domenu promatrane seed.alignment datoteke. U Durbinovoj literaturi [1] ponuđena je formalna topologija profilnog HMM-a za postupak klasifikacije sekvenci u proteinske obitelji, a navedena topologija prikazana je na slici 4.1.



Slika 4.1. Topologija profilnog HMM-a za klasifikaciju sekvenci u proteinske domene

Izvor: Durbin et al. [1]

Kvadratna polja označavaju Match stanja, krugovi Delete stanjam, a rombovi Insert stanja. U nastavku je na primjeru poravnanja 5 sekvenci opisan heursitički algoritam za konstrukciju profilnog HMM-a. Radi jednostavnosti opisa korištene su genomske sekvence koje se sastoje od samo 4 baze (A, C, G, T), a ne od 20 aminokiselina. U postupku klasifikacije proteina se dakako koristi abeceda od 20 znakova, ali je postupak konstrukcije identičan.

Slika 4.2. prikazuje poravnatih 5 sekvenci koje se koriste za konstrukciju profilnog HMM-a. Svaka sekvenca je predstavljena kao niz znakova, a praznine su označene s "-". U ovom slučaju, svaka sekvenca je dugačka 10 znakova, a ukupno imamo 5 sekvenci.

```
-GAT-T---
T---GTG--
TG-TG-GAA
--A-GC---
----GTC--
  **
```

Slika 4.2. Poravnanje 5 sekvenci korištenih za konstrukciju profilnog HMM-a

Neka zvjezdice označavaju pozicije konzerviranih stupaca. Postupak kreće redom prolaskom kroz svaku sekvencu u seed.alignment datoteci. Kada se u nekom redku dođe do znaka baze na poziciji stupca koji nije konzerviran u listi za dotičnu sekvencu bilježim prijelaz u stanje Insert. Ako se dođe do znaka A, T, C ili G u stupcu koji je označen kao konzerviran, bilježim prijelaz u stanje Match. Ako se dođe do praznine u stupcu koji je označen kao konzerviran, bilježim prijelaz u stanje Delete. U ovom slučaju, stupac 5 i 6 su konzervirani, dok su ostali stupci ne-konzervirani. Gledano za jedan redak poravnjanja ako iz Match stanja napravim prijelaz u Match stanje povećavam indeks stanja, dok kod prijelaza između dva susjedna Insert stanja ne povećavam indeks stanja. Indeks insert stanja se povećava ako se pojavi prijelaz iz susjednog Match ili Delete stanja. U našem primjeru će sukladno tome stanje I0 u sebi inkorporirati prvih 5 stupaca. Indeksiranje Delete stanja odvija se analogno Match stanjima.

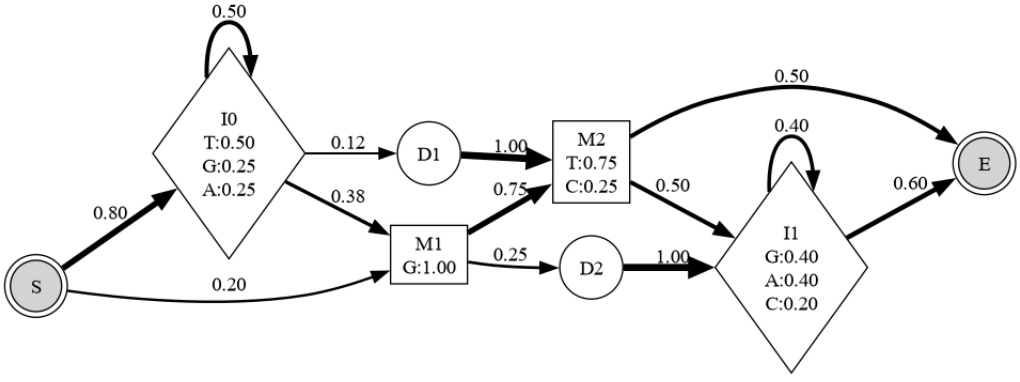
U procesu parsiranja svakog retka poravnanja generira se niz stanja kroz koja sek-

venca prolazi u profilnom HMM-u. Na primjer, za dani skup poravnanja dobivene su sljedeće liste prijelaza za pojedine sekvence: za prvi redak prijelazi su ['S', 'I0', 'I0', 'I0', 'D1', 'M2', 'E'], za drugi ['S', 'I0', 'M1', 'M2', 'I2', 'E'], za treći ['S', 'I0', 'I0', 'I0', 'M1', 'D2', 'I2', 'I2', 'I2','E'], za četvrti ['S', 'I0', 'M1', 'M2', 'E'], a za peti ['S', 'M1', 'M2', 'I2', 'E']. Ove liste prikazuju konkretne prijelaze između stanja (počevši iz početnog stanja SS, završavajući u završnom EE), te odražavaju kako različite sekvence koriste različite puteve kroz model – primjerice, koristeći Insert i Delete stanja ovisno o prisutnosti praznina u poravnanju.

Iz tako dobivenih listi prijelaza moguće je izračunati vjerojatnosti prijelaza između stanja. Na primjer, ako imamo 5 sekvenci i 10 stupaca, možemo izračunati koliko puta se dogodio prijelaz iz jednog stanja u drugo. Ako se prijelaz iz stanja S dogodio 5 puta (u svih pet listi prijelaza) od toga četiri puta u stanje I0 i jedan put u stanje M1, tada je vjerojatnost prijelaza iz S u I0 jednaka $4/5 = 0.8$, a vjerojatnost prijelaza iz S u M1 jednaka $1/5 = 0.2$. Ovaj postupak se ponavlja za sve moguće prijelaze između stanja.

Emisijske vjerojatnosti određuju se na temelju učestalosti pojavljivanja pojedinih baza u svakom od stanja. Ako I0 obuvaća prvih 5 stupaca u kojima se četiri puta javlja T, dva puta A i dva puta G onda su emisijske vjerojatnosti za stanje I0: $P(T|I0) = 4/8 = 0.5$, $P(A|I0) = 2/8 = 0.25$, $P(G|I0) = 2/8 = 0.25$. Ovaj postupak se ponavlja za svako stanje u modelu. Delete stanja ne emitiraju nikakve znakove.

Dobiveni grafički model u končnici je moguće vizualizirati pomoću biblioteke Graphviz što je prikazano na slici 4.3.



Slika 4.3. Poravnanje 5 sekvenci korištenih za konstrukciju profilnog HMM-a

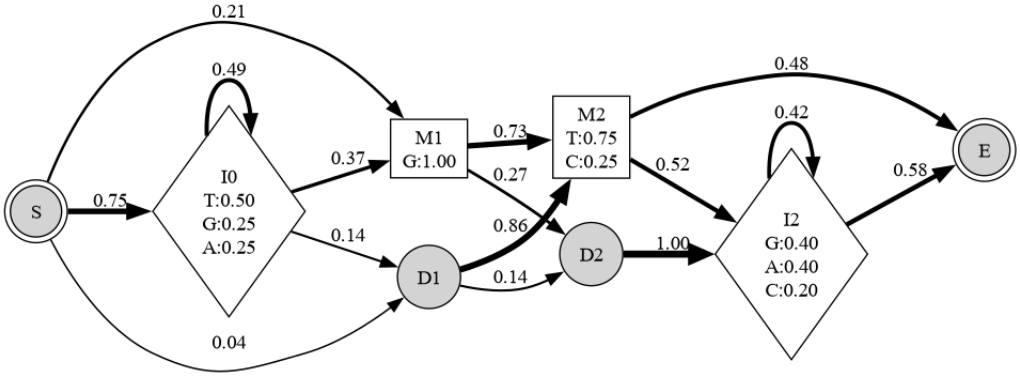
U sljedećem koraku izvršen je postupak Laplaceovog zaglađivanja koji se koristi za

izbjegavanje problema s nultim vjerojatnostima. Ovaj postupak dodaje malu konstantu (npr. 1) na brojnik i prilagođava nazivnik kako bi se osigurala normalizacija. Na taj način modelu se daje šansa da modelira i pojave koje nisu striktno prisutne u početnim podacima čime se omogućuje da model bude robusniji i otporniji na nepredviđene situacije. Formula za Laplaceovo zaglađivanje je sljedeća:

$$P(x|y) = \frac{N(x,y) + \alpha}{N(y) + \alpha \cdot |X|} \tag{4.1}$$

gdje je $N(x,y)$ broj pojavljivanja simbola x u stanju y (misleći pritom na zaglađivanje emisijskih vjerojatnosti), α je konstanta koja se dodaje (npr. 1), $N(y)$ je ukupan broj simbola u stanju y, a $|X|$ je broj različitih simbola u abecedi.

Navedeni postupak moguće je primjeniti i na prijelazne i na emisijske vjerojatnosti. Prethodni graf nakon Laplaceovog zaglađivanja (samo na prijelaznim vjerojatnostima) prikazan je na slici 4.4.

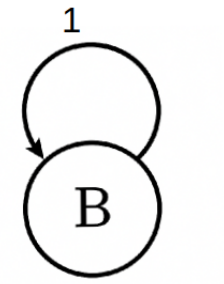


Slika 4.4. Poravnanje 5 sekvenci korištenih za konstrukciju profilnog HMM-a

5. Postupak klasifikacije i analiza rezultata

U prethodnom poglavlju opisan je postupak konstrukcije profilnog HMM-a čiji je konačni cilj definiranje skupa stanja modela te matria prijelaznih i emisijskih vjerojatnosti. U ovom poglavlju opisuje se postupak klasifikacije nepoznate sekvence pomoću već definiranih parametara modela. Sama brada seed.alignment datoteka i stvaranja parametarskih matrica obavljena je u programskom jeziku python, dok se naknadno korištenje modela vrši u programskom jeziku R. Klasifikacija se provodi korištenjem Forward algoritma koji omogućuje izračunavanje vjerojatnosti promatranog niza s obzirom na model. Ovaj algoritam koristi dinamičko programiranje kako bi izbjegao ponavljanje izračuna i optimizirao vrijeme izvođenja što je direktna posljedica činjeice da se ciljna topologija HMM-a temelji na Markovljevom procesu prvog reda što znači da je vjerojatnost trenutnog stanja ovisna samo o prethodnom stanju. Ne treba zavaravati činjenica da je u trenutno stanje moguće doći iz tri različita Match, Insert i Delete stanja jer sva tri stanja imaju istu vremensku točku koja iznosi t-1 ako je t trenutna vremenska točka.

Postupak klasfikacije nepoznatih proteinskih sekvenci prema konstriranim HMM-ovima temelji se na izračunu logaritamske vjerojatnosti promatranog niza pod svakim profilnim HMM-om. Potom se koristi i pozadinski trivijalni model čiji je graf prikazan na slici 5.1. koji služi kao referentna točka za usporedbu.



A	1/20
C	1/20
...	
Y	1/20

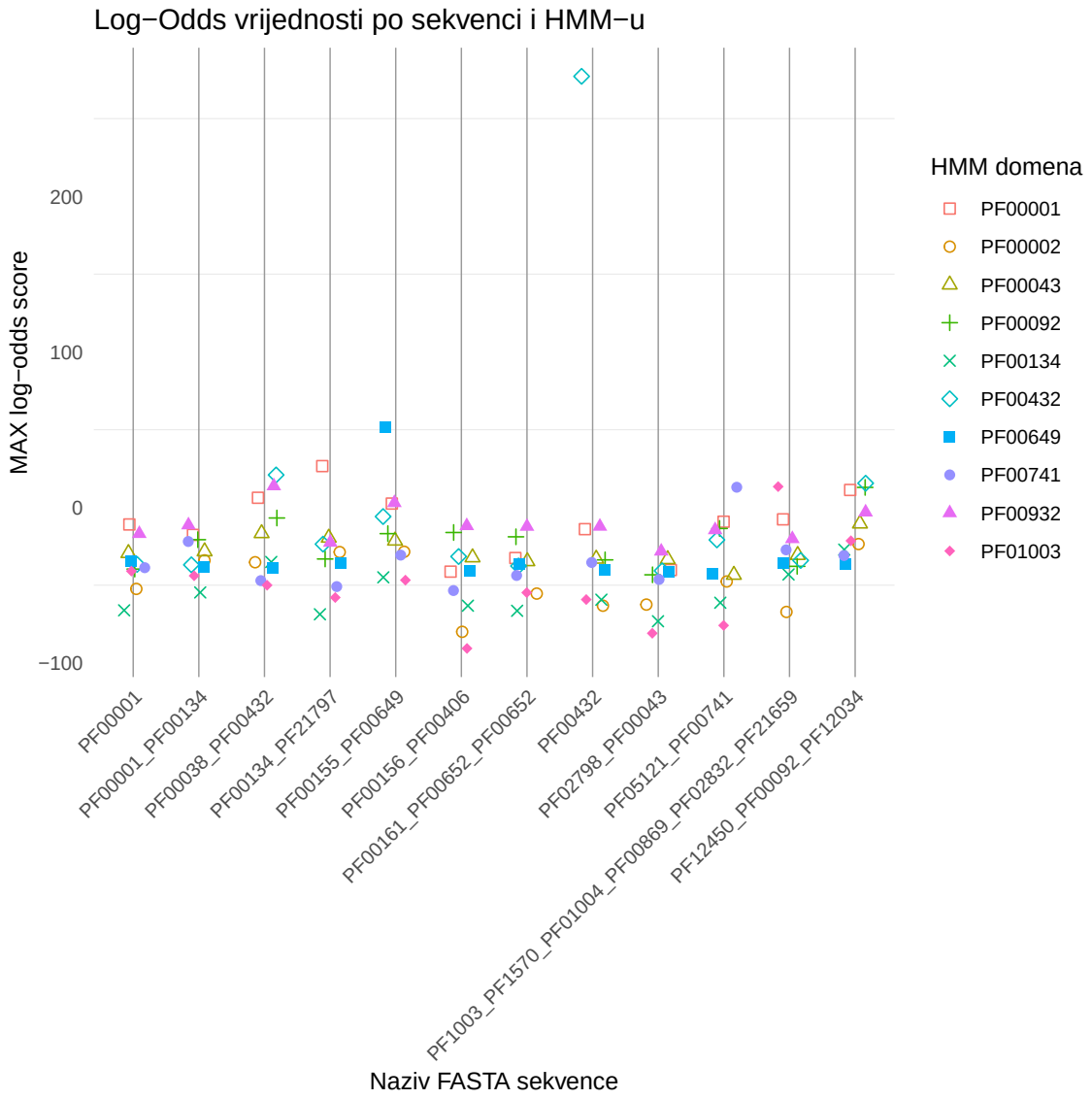
Slika 5.1. Grafički prikaz trivijalnog pozadinskog modela

Ovaj model se koristi kao kontrola kako bi se osiguralo da je klasifikacija nepoznate sekvence rezultat stvarnog preklapanja s profilnim HMM-om, a ne slučajnog poduda-ranja. Konačna vjerojatnost na temelju koje se sekvenca klasificira naziva se log-odds i izračunava se kao razlika između logaritma vjerojatnosti promatranog niza pod profil-nim HMM-om i logaritma vjerojatnosti promatranog niza pod pozadinskim modelom. Ova razlika omogućuje da se uzmu u obzir samo značajni rezultati koji su veći od onih koje bi se moglo očekivati slučajno. Formula za izračun log-odds vjerojatnosti je sljedeća:

$$\text{log-odds} = \log(P(X|HMM)) - \log(P(X|pozadinski\ model)) \tag{5.1}$$

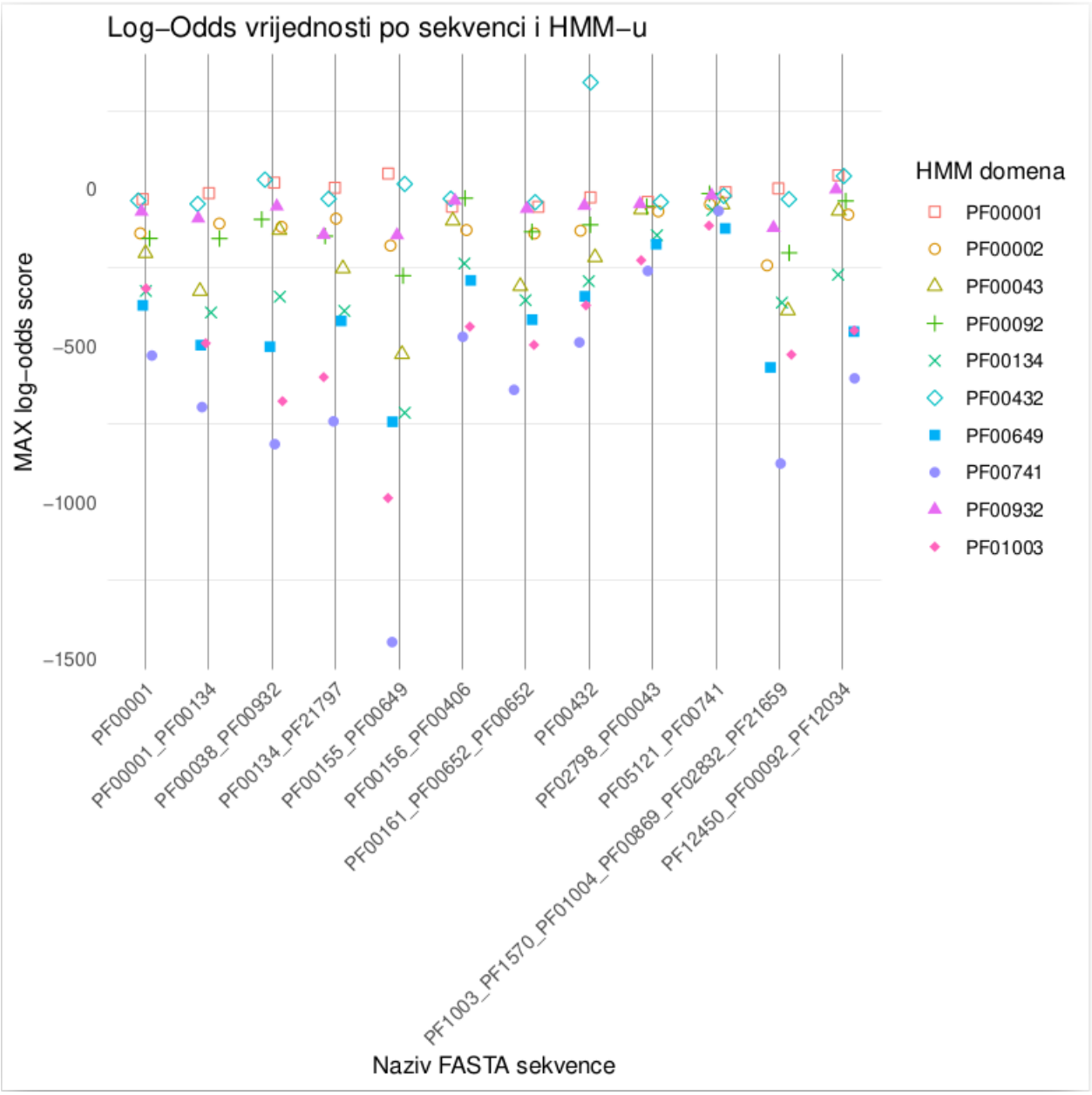
gdje je $P(X|HMM)$ vjerojatnost promatranog niza pod profilnim HMM-om, a $P(X|pozadinski\ model)$ vjerojatnost promatranog niza pod pozadinskim modelom. Obje vjerojatnosti se izračunavaju korištenjem forward algoritma. Koristi se logaritamski oblik vjerojatnosti kako bi se izbjegli problemi s malim brojevima i numeričkom stabilnošću. Na temelju izračunatih log-odds vjerojatnosti, sekvenca se klasificira kao član one obitelji čiji HMM daje najvišu log-odds vrijednost. Ovaj pristup omogućuje visoku preciznost klasifikacije i duboko razumijevanje filogenetskih odnosa među sekvencama, dok istovremeno koristi provjerene metode statističkog modeliranja i biološke interpretacije. Zbog činjenice da i sama duljina ispitne sekvence može značajno utjecati na rezultat dobiven forward algoritmom, potrebno je čitavu ispitnu sekvencu obići posmačnim prozorom one duljine koja najbolje odgovara topologiji HMM-a koji se trenutno koristi za klasifikaciju. Kada takav

prozor pogodi regiju ispitne sekvence koja sadrži domenu koja se poklapa s HMM-om očekivano će forward algoritam u tome trenutku dati najveću log vjerojatnost da sadržaj prozora pripada dotičnoj domeni. Duljina prozora za svaki HMM dobije se tako da se stotinu puta uzorkuje slučajna sekvenca iz dotičnog HMM-a te se računa prosječna duljina te sintetičke sekvence koju nam vraća konstrirani model. Posmak prozora iznosi 20% duljine samog prozora. Nužno je postojanje preklapanja prozora kako bi se izbjeglo preskakanje regija koje su konzervirane i koje se nalaze na rubovima prozora. Na taj način se osigurava da se ne propuste važne informacije koje bi mogle utjecati na klasifikaciju. Konačni rezultati postupka klasifikacije skupa nepoznatih sekvenci pomoću skupa konstruiranih profinih hmm-ova koristeći tehniku posmačnog prozora prikazani su na slici 5.2.



Slika 5.2. Grafički prikaz dodijeljenih log-odds vjerojatnosti za 10 različitih proteinskih domena

Svaka prikazana točka predstavlja log-odds vjerojatnost za određenu sekvencu i odgovarajući HMM. Na x-osi su prikazane različite proteinske domene, dok je na y-osi prikazana log-odds vjerojatnost. Svaka točka predstavlja rezultat klasifikacije za određenu sekvencu i HMM. Ovaj grafički prikaz omogućuje vizualizaciju rezultata klasifikacije i usporedbu između različitih domena. Pojedina proteinska sekvenca bit će klasificirana onome HMM-u koji joj je dodijelio najvišu log-odds vjerojatnost. Na temelju ovog grafičkog prikaza moguće je uočiti koje su sekvence najbliže određenoj domeni, a koje su najdalje. Također, može se primijetiti i koliko su različite sekvence unutar iste domene, što može ukazivati na evolucijske promjene unutar te domene. Sve navedeno moguće je testirati i odgovarajućim statističkim testovima. Druga, četvrta i deveta sekvenca su pogrešno klasificirane. Šesta i sedma sekvenca uopće ne posjeduju domene kojima odgovaraju konstrirani HMM-ovi. Vidljivi su slučajevi kada je najveća ostvarena log-odds vrijednost za pojedinu sekvencu negativna, no bez obzira na to i dalje je moguće ostvariti klasifikacijski kriterij prema načelu najmanjeg odstupanja od pozadinskog modela. Pritom je problematično što nije jednostavno razaznati proteine koji sadrže domene za koje nemamo konstruirane pripadne HMM-ove. Navedeno nas navodi na zaključak da bi trebalo ispitativati u odnosu na znatno veću bazu profilnih HMM-ova. Ovaj rezultat ukazuje na to da su te sekvence evolucijski udaljene od ostalih sekvenci u bazi podataka i da ne pripadaju nijednoj od definiranih domena. Jasno je vidljivo kako najveća pozdanost u klasifikaciju pripada sekvenci PF00432. Kao svojevrsna potvrda modularnosti i evolucijske konzerviranosti domenskih regija na proteinskim sekvencama postupak klasifikacije je ponovljen ali bez korištenja posamčnog prozora. U tom slučaju očekivano će prema definiciji log-odds vjerojatnosti, ta vrijednosti pasti za većinu opservacija ispod nule što objašnjava kako je pozadinski model u tom slučaju puno vjerojatniji tvorac ispitne sekvence u odnosu na bilo koji od konstriranih HMM-ova. Navedeno je prikazano na slici 5.3.

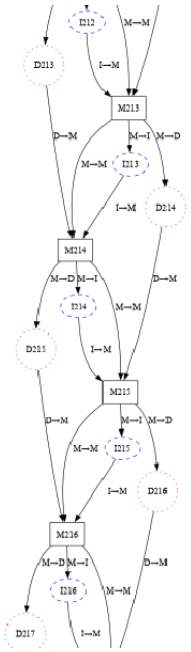


Slika 5.3. Grafički prikaz dodijeljenih log-odds vjerojatnosti za 10 različitih proteinskih domena bez korištenja posmačnog prozora

Čak i pri ovome krivome postupku sekvenca PF00432 je klasificirana ispravno zbog jednostvne proteinske arhitekture koja se sastoji od samo jedne domene. Rezultati za domenu PF00001 također se značajno ne razlikuju u odnosu na postupak sa korištenjem posmačnog prozora zbog istog razloga.

6. HMMER

U ovom poglavlju opisuje se HMMER, specijalizirani bioinformatički alat koji se koristi za analizu i pretraživanje sekvenci pomoću skrivenih Markovljevih modela. HMMER je razvijen kako bi omogućio efikasno modeliranje i prepoznavanje sekvenci u biološkim podacima, posebno u kontekstu proteina i nukleinskih kiselina. Ovaj alat koristi HMM-ove za identifikaciju i klasifikaciju sekvenci, kao i za predikciju funkcionalnih domena unutar proteina što je ujedno i tema ovog rada stoga je razumno dati osvrt i na alat koji se u praksi najčešće koristi u ove svrhe i usporediti ga s vlastitom implementacijom. Pomoću pomoćnih datoteka koje ovaj alat pri pokretanju generira moguće je dobiti i grafički prikaz HMM-a koji se koristi za pretraživanje sekvenci. Ovaj grafički prikaz omogućuje vizualizaciju strukture HMM-a, uključujući stanja, prijelaze i emisijske vjerojatnosti. Na slici 6.1. prikazan je primjer grafičkog prikaza HMM-a generiranog pomoću HMMER-a.



Slika 6.1. Grafički prikaz HMM-a za pretraživanje sekvenci u alatu HMMER

U odnosu na vlastitu implementaciju, kod strukture HMM-a uočljiv je manji broj

prijelaza među stanjima što navodi na to da je kod alata stavljen naglasak na determinitički pristup kako bi preciznije radio u realnim scenarijima izninmo velikog broja sekvenci i profilnih HMM-ova. Deterministočnost se postiže učinkovitim ugrađivanjem vanjskog domenskog znanja. Za razliku od jednostavnog Laplaceovog zaglađivanja, HMMER koristi Dirichletove mješavine kao prior distribucije za procjenu emisijskih vjerojatnosti u profil-HMM-ovima. Ove mješavine, naučene iz velikih skupova višestrukih poravnanja, predstavljaju biološki utemeljene obrasce raspodjele aminokiselina na različitim vrstama pozicija (npr. konzerviranim ili varijabilnim). Kombiniranjem ovih priora s opaženim frekvencijama u poravnanju, HMMER dobiva posteriorne vjerojatnosti koje su statistički stabilne i robusne čak i kada je broj uzoraka mali. Time se izbjegava overfitting i osigurava veća biološka relevantnost modela. Odsječak rezultatnog ispisa navedenog alata za klasifikaciju ispitne sekvence PF00001 prikazan je na slici 6.2.

```
#                                                                            --- full sequence --- -------------- this domain -------------
# target name        accession   tlen query name           accession   qlen   E-value  score  bias   #  of  c-Evalue  i-Evalue  score  bias
#------------------- ---------- ----- -------------------- ---------- ----- --------- ------ ----- --- --- --------- --------- ------ -----
7tm_1                PF00001.26   260 A0A821XNW9           -            782   1.4e-62  197.6  12.8   1   2   1.4e-62   1.4e-62  197.6  12.8
  receptor (rhodopsin family)
7tm_1                PF00001.26   260 A0A821XNW9           -            782   1.4e-62  197.6  12.8   2   2      0.18      0.18   -2.7   2.2
  receptor (rhodopsin family)
#
# Program:         hmmscan
# Version:         3.4 (Aug 2023)
# Pipeline mode:   SCAN
# Query file:      A0A821XNW9.fa
# Target file:     binaries/PF00001.hmm
# Option settings: hmmscan --domtblout rezultati.tbl binaries/PF00001.hmm A0A821XNW9.fa
# Current dir:     /home/domagoj/Desktop/Seminar2_Implementation_of_Hidden_Markov_Models_in_the_Phylogenetic_Classification_of_Protein_Sequer
# Date:            Fri May 16 12:37:25 2025
# [ok]


```

Slika 6.2. Grafički prikaz HMM-a za pretraživanje sekvenci u alatu HMMER

Rezultati prikazani iz alata HMMER (hmmscan) predstavljaju poravnanje ispitne proteinske sekvence s HMM-om domene `7tm_1` iz Pfam baze, koja pripada obitelji receptora s 7 transmembranskih domena (rhodopsin family). Uočene su dvije domene u sekvenci A0A821XNW9 koje odgovaraju modelu `PF00001.26`, pri čemu je prva detekcija statistički značajna, dok je druga marginalna.

E-value (ocjena očekivanja) označava broj rezultata s jednakim ili boljim poravnanjem koji bi se mogli očekivati slučajno u bazi iste veličine. Manja vrijednost znači veću statističku značajnost. Prva domena ima vrlo nisku E-vrijednost (1.4×10^{-62}), što upućuje na visokosignifikantno poravnanje. Druga domena ima E-vrijednost od 0.18, što je blizu granice statističke pouzdanosti.

Score je sirova HMMER-ova log-odds ocjena (izražena u bitovima), koja kvantificira

koliko je vjerojatnije da poravnanje potječe iz HMM modela nego iz slučajnog modela. Veći score označava bolje poravnanje. Prva domena ima score od 197.6, dok druga ima negativan score (-2.7), što dodatno ukazuje na njezinu nisku pouzdanost.

Bias označava koliko je score poravnanja mogao biti pojačan zbog ponavljajućih ili nisko-kompleksnih regija; viši bias može ukazivati na umjetno napuhane rezultate. U ovom slučaju, bias je 12.8 za prvu domenu, što se smatra umjerenim i očekivanim za transmembranske proteine.

Koordinate **hmm coord**, **ali coord** i **env coord** odnose se na pozicije poravnanja:

- `hmm coord`: pozicije unutar HMM modela (od pozicije 1 do 260 za prvu domenu),
- `ali coord`: stvarne pozicije u sekvenci koje su poravnate s modelom (56–380),
- `env coord`: prošireni interval koji uključuje cijeli poravnati kontekst u sekvenci (također 56–380), tj. područje koje najvjerojatnije sadrži cijelu domenu.

Parametar **acc** (accuracy) označava prosječnu posteriornu vjerojatnost ispravnog poravnanja po poziciji, s vrijednostima bližima 1 označavajući veće povjerenje. Uočena vrijednost 0.97 za prvu domenu potvrđuje vrlo visoku pouzdanost.

Ukratko, rezultati upućuju na postojanje jedne vrlo pouzdane transmembranske domene tipa 7tm_1 u sekvenci A0A821XNW9, dok je druga detekcija vjerojatno rezultat slabije konzerviranog ili nepotpunog poravnanja.

U HMMER-u, sirovi **score** predstavlja log-odds ocjenu, definiranu kao logaritam omjera vjerojatnosti da je sekvenca generirana HMM modelom u odnosu na pozadinski (slučajni) model: $S = \log_2\left(\frac{P(\text{sekvenca}|\text{HMM})}{P(\text{sekvenca}|\text{background})}\right)$ što je ekvivalentno kao u vlastitoj implementaciji. Score se izražava u bitovima. Veći score označava veću sličnost s modelom.

E-vrijednost (E-value) procjenjuje broj poravnanja s jednakim ili višim scoreom koji bi se mogli slučajno pojaviti prilikom pretraživanja baze određene veličine. Računa se pomoću ekstremne vrijednosne distribucije (Gumbelove distribucije), prema formuli $E = K \cdot N \cdot e^{-\lambda S}$, gdje su K i λ empirijski određeni parametri distribucije, N je efektivna veličina baze, a S je score. Ova mjera omogućuje statističku procjenu značajnosti poravnanja u kontekstu cijele baze podataka.

7. Zaključak

U ovom radu opisan je postupak konstrukcije profilnog HMM-a temeljenog na višestrukom poravnanju sekvenci. Ovaj pristup omogućuje modeliranje evolucijskih odnosa među sekvencama i identifikaciju konzerviranih regija unutar proteina. Kroz analizu rezultata klasifikacije nepoznatih sekvenci, pokazano je da je ovaj pristup učinkovit u prepoznavanju proteinskih domena i njihovih funkcija. Iako su rezultati klasifikacije pokazali solidnu preciznost, također su uočeni i slučajevi pogrešne klasifikacije, što ukazuje na potrebu za daljnjim istraživanjem i poboljšanjem modela. U budućim radovima, preporučuje se korištenje većih baza podataka i dodatnih metoda za optimizaciju modela kako bi se povećala točnost klasifikacije. Također, usporedba s postojećim alatima poput HMMER-a pokazuje da je vlastita implementacija konkurentna, ali postoje i mogućnosti za daljnje poboljšanje. Skriveni Markovljevi modeli omogućuju efikasno ugrađivanje vanjskog domenskog znanja što direktno utječe na preciznost modela i ističe ga u odnosu na druge tehnike strojnog učenja.

Literatura

[1] 1. Durbin R, Eddy SR, Krogh A, Mitchison G. Markov chains and hidden Markov models. In: Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press; 1998:47-80.