Univerzitet u Beogradu
Matematički fakultet

Master Rad

Predviđanje funkcija proteina metodama binarne klasifikacije

Autor:
Anja Bukurov

Mentor:
dr Jovana Kovačević

Članovi komsije:

dr Jovana Kovačević
prof. dr Gordana Pavlović-Lažetić
dr Mladen Nikolić

Beograd, 2019

UNIVERZITET U BEOGRADU

Sažetak

Matematički fakultet
Katedra za Računarstvo i informatiku

Informatičar

Predviđanje funkcija proteina metodama binarne klasifikacije

Anja Bukurov

Proteini su makromolekuli koji igraju važnu ulogu u organizmu svakog živog bića. To su biološki najaktivniji molekuli neophodni za izgradnju i funkcionisanje ćelija i imaju veliki broj esencijalnih funkcija. Struktura proteina zavisi od rasporeda aminokiselina i utiče na njegovu funkciju. Primarna struktura obuhvata sekvencu aminokiselina koje izgrađuju protein. Ona je osnovni izvor informacija o proteinu i njegovoj funkciji. Poznato je da proteini sa sličnim primarnim strukturama teže da obavljaju iste funkcije.

Svake godine sekvencira se veliki broj novih genoma čime raste broj novootkrivenih proteina. Funkcija proteina određuje se eksperimentalno što je skup i spor proces zbog čega se ulaže trud u razvoj računarskih metoda koje mogu da predvide funkciju proteina. Cilj ovog rada je razvoj alata za određivanje funkcije proteina na osnovu njegove primarne strukture pomoću metoda binarne klasifikacije.

Sadržaj

Glava 1

Uvod

Predviđanje funkcije proteina je jedan od najbitnijih zadataka bioinformatike koji može pomoći u velikom broju bioloških problema. Poznavanje funkcije proteina daje nam informacije o njegovim ulogama u organizmu. Metode za eksperimentalno određivanje funkcije proteina su spore u odnosu na brzinu sekvencioniranja genoma koje uvećava broj novih sekvenci. Mnoge metode predviđanja funkcije proteina zasnivaju se na poređenju sekvenci ili struktura proteina za koje je utvrđena funkcija sa onim proteinima za koje je funkcija nepoznata.

U ovom radu prikazan je razvoj alata za predviđanje funkcije proteina na osnovu njihove primarne strukture. Korišćene su metode binarne klasifikacije i to: metod potpornih vektora, logistička regresija i slučajne šume. Alat je razvijan u programskom jeziku Python.

U poglavlju 2 uvedeni su biološki pojmovi neophodni za razumevanje rada. U poglavlju 3 prikazan je način predstavljanja bioloških podataka u računaru, a onda su ukratko predstavljene metode binarne klasifikacije korišćene za razvoj prediktora. Zatim, u poglavlju 4, analizirani su podaci o proteinima i njihovim funkcijama, nakon čega je opisana implementacija alata za predviđanje funkcije proteina. Na kraju, u poglavlju 5 sumirani su rezultati koje su prediktori dali na izdvojenom skupu proteina za testiranje.

Glava 2

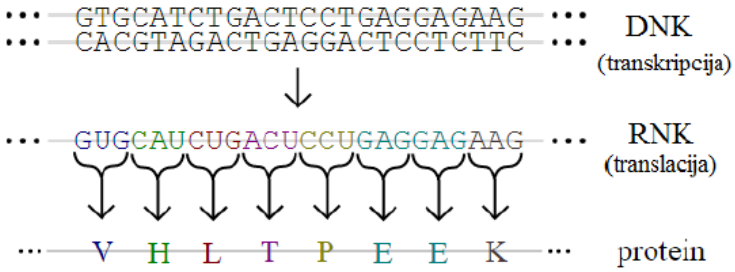
Proteini

Sva živa bića sastoje se iz ćelija. U ćelijama se neprestano odvijaju različiti procesi u kojima učestvuju nukleinske kiseline (dezoksiribonukleinka kiselina - DNK i ribonukleinska kiselina - RNK) i proteini. Unutar molekula DNK šifrovan je genetski materijal koji sadrži uputstva za sintezu proteina.

Proteini su makromolekuli koji igraju mnoge kritične uloge u organizmu. Sačinjavaju više od 50% suvog dela ćelije i važni su za njenu izgradnju i funkcionisanje. Kontrakcija mišića, strukturna podrška, ubrzavanje i usporavanje hemijskih reakcija, odbrana od virusa i bakterija samo su neke od mnogobrojnih uloga koje proteini obavljaju [1, 2].

2.1 Sinteza proteina

DNK sadrži informacije koje su neophodne ćeliji za izgradnju veoma važnog tipa molekula - proteina. Proteini se sintetišu prilikom genske ekspresije i to u dva koraka: transkripcija i translacija (slika 2.1). Prvi korak je dekodiranje genske poruke, prilikom čega se od DNK sekvence dobija RNK sekvenca. U sastav obe nukleinske kiseline ulazi 4 nukleotida i oni su prikazane u tabeli 2.1. S obzirom da su tri nukleotidne baze iste, proces transkripcije sastoji se iz zamene svakog molekula T molekulom U.



Slika 2.1: Prikaz procesa sinteze proteina [2].

DNK	adeinin (A)	guanin (G)	citozin (C)	timin (T)
RNK	adeinin (A)	guanin (G)	citozin (C)	uracil (U)

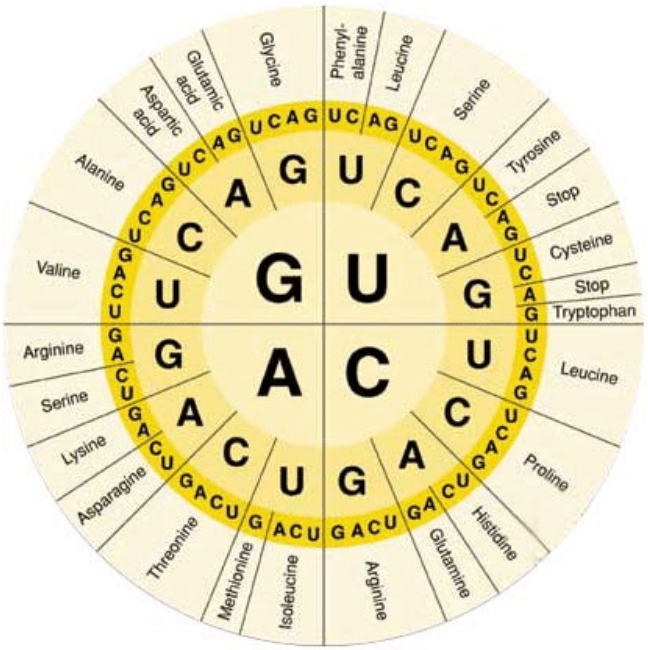
Tabela 2.1: Prikaz nukleinskih kiselina sa nukeotidima koji ih grade

Sledeći korak, proces translacije, jeste grupisanje aminokiselina kako bi se dobio protein. Genetski kod se čita u grupama od 3 nukleotida koje nazivamo *kodoni*. Svaki

Aminokiselina	Oznaka	Simbol	Aminokiselina	Oznaka	Simbol
Alanin	ALA	A	Arginin	ARG	R
Asparagin	ASN	N	Asparaginska kiselina	ASP	D
Cistein	CYS	C	Glutamin	GLN	Q
Glutaminska kiselina	GLU	E	Glicin	GLY	G
Histidin	HIS	H	Izoleucin	ILE	I
Leucin	LEU	L	Lisin	LYS	K
Metionin	MET	M	Fenilalanin	PHE	F
Prolin	PRO	P	Serin	SER	S
Treonin	THR	T	Triptofan	TRP	W
Tirosin	TYR	Y	Valin	VAL	V

TABELA 2.2: Prikaz standardnih aminokiselina sa oznakama i simbolima

kodon odgovara tačno jednoj aminokiselini ili služi da označi kraj sekvence (stop kodon). Na primer, kodon **GUA** kodira aminokiselinu valin, dok kodon **UAG** označava kraj sekvence. Na slici 2.2 je dat šematski prikaz svih kodona i odgovarajućih aminokiselina. Jedan po jedan, kodoni se prevode u odgovarajuće aminokiseline čime se dobija sekvenca aminokiselina koja čini protein [3, 4].



SLIKA 2.2: Prikaz kodona i odgovarajućih aminokiselina

2.2 Aminokiseline

Aminokiseline su organska jedinjenja koja se sastoje od karboksilne grupe ($COOH$), aminogrupe (NH_2) i bočnog lanca (R-grupa) koji je vezan za α-ugljenikov atom i karakterističan je za svaku aminokiselinu. Postoji 20 standardnih aminokiselina i one su prikazane u tabeli 2.2. Pod standardnim aminokiselinama podrazumevaju se one aminokiseline za koje postoji najmanje jedan specifičan kodon u genetskom kodu [5, 6, 7].

Aminokiseline možemo podeliti u nekoliko grupa prema osobinama bočnog lanca [5, 6, 8]:

1. **aminokiseline sa nepolarnim bočnim lancem**
   Bočni lanac ovih aminokiselina ne može da otpušta niti da vezuje protone, kao ni da učestvuje u vodoničnim ili jonskim vezama. Zbog svoje nepolarnosti, one su hidrofobne i obično popunjavaju praznine u unutrašnjosti proteina čime doprinose oblikovanju njegove strukture. U ovu grupu ubrajamo 7 standardnih aminokiselina: alanin, valin, leucin, izoleucin, metionin, fenilalanin, triptofan.

2. **aminokiseline sa nenaelektrisanim polarnim bočnim lancem**
   R-grupa aminokiselina iz ove grupe može da gradi vodonične veze sa molekulima vode što ih čini rastvorljivijim u odnosu na aminokiseline iz prethodne grupe. Zbog polarnosti, ove aminokiseline se obično nalaze na spoljašosti proteina. Ova grupa obuhvata 6 standardnih aminokiselina i to: serin, treonin, tirozin, asparagin, glutamin i cistein.

3. **aminokiseline sa naelektrisanim polarnim bočnim lancem**
   U ovu grupu spadaju veoma hidrofilne aminokiseline zbog čega se one nalaze na površini proteina. Dodatno ih možemo podeliti na kisele i bazne aminokiseline. Kisele imaju jednu karboksilnu grupu više i imaju negativno naelektrisanje, dok su bazne aminokiseline pozitivno naelektrisane. Asapraginska i glutaminska kiselina su kisele aminokiseline, a lizin, histidini i arginin spadaju u bazne aminokiseline.

4. **konformaciono važne aminkiseline**
   Preostale dve standardne aminokiseline, glicin i prolin se po svojoj strukuturi razlikuju od ostalih. Glicin nema bočni lanac i može da se prilagođava konformacijama koje su nedostupne drugim aminokiselinama. Prolin sadrži jedan heterociklički prsten i u svojoj strukuturi sadrži sekundarnu amino grupu.

Bilo koje dve aminokiseline mogu izgraditi veći molekul, dipeptid, formiranjem peptidne veze između njih. Peptidna veza se ostvaruje između atoma ugljenika iz karboksilne grupe i atoma azota iz amino grupe. Peptidne veze omogućavaju stvaranje lanaca aminokiselina, tzv. polipeptida. Peptidna veza nastaje reakcijom dve aminokiseline pri čemu se spajaju karboksilna grupa jedne sa amino grupom druge aminokiseline uz izdvajanje vode. Prilikom tog vezivanja pojavljuje se niz koji se zove kičma polipeptidnog lanca koji čine ugljenikov atom karboksilne grupe, atom azota aminogrupe i α-ugljenikov atom. To je osnovni niz i isti je za sve proteine, a oni se međusobno razlikuju po bočnim lancima aminokiselina [6, 8].

Peptide možemo podeliti prema broju aminokiselina koje sadrže i to na oligopeptide i polipeptide. Oligopeptidi su sačinjeni od najviše 10 aminokiselina, dok polipeptidi sadrže do 100 aminokiselina. Jedinjenja sa više od 100 aminokiselina u lancu spadaju u proteine [6].

2.3 Struktura proteina

U sastav proteina ulazi 20 standardnih aminokiselina. Sekvenca aminokiselina, koja se formira peptidnim vezama, specifična je za svaki protein. Ona je primarni izvor informacija o proteinu i njegovoj funkciji. Složenost proteinske strukture najbolje se analizira kroz četiri nivoa: primarna, sekundarna, tercijerna i kvaterna struktura [5].

Primarna strktura Jedinstveni redosled aminokiselina koje su povezane peptidnom vezom kako bi formirale protein čini primarnu strukturu proteina. Proteini koji imaju slične sekvence često imaju i slične osobine i funkcije. Zbog toga je poređenje sekvenci prvi korak u izučavanju proteina. Razumevanje primarne sekvence je bitno zbog mnogih genetskih bolesti koje za posledicu imaju proteine sa neispravnim sekvencama što vodi do pogrešnog savijanja i nefunkcionalnog proteina [5, 6, 8].

Sekundarna struktura Polipeptidni lanac ne zauzima bilo kakav oblik u prostoru već ima opšti raspored aminokiselina koje se u lancu nalaze jedna blizu druge. Taj raspored označava sekundarnu strukturu proteina i podrazumeva savijanje ili uvijanje polipeptidnog lanca. Lanac može da uzme oblik α-heliksa (engl. α-helix), β-traka (engl. β-sheet) ili β-okreta (engl. β-turn). α-heliks je periodična struktura u kojoj se kičma proteina spiralno uvrće, a bočni lanci aminokiselina izviruju izvan nje. β-traka formiraju se kao parovi lanaca aminokiselina koji se uzdužno vezuju vodoničnim vezama. β-okret menja pravac polipeptidnog lanca čime mu pomaže dobije kompaktan, loptast oblik [5, 8, 9].

Tercijarna struktura Prostorna struktura čitavog molekula proteina predstavlja ternarnu strukturu. Hidrofobni bočni lanci nepolarnih aminokiselina teže da budu unutar molekula proteina zaštićeni od vode, dok se kisele i bazne aminokiseline obično nalaze na površini proteina pošto su hidrofilne. α-heliksi i β-listovi služe da obezbede maksimalan broj vodoničnih veza u unutrašnjosti molekula, čime sprečavaju da se molekuli vode vežu za hidrofilne grupe i time naruše integritet proteina [5, 6].

Kvaternarna struktura Mnogi proteini su formirani grupisanjem više savijenih polipeptidnih lanaca. Pojedinačnu komponentu nazivamo podjedinica. One mogu biti međusobno različite ili potpuno iste. Raspored ovih podjednica predstavlja kvaternarnu strukturu. U kvaternarnu strukturi podjedinice se međusobno drže zajedno nekovalentnim interakcijama i kovalentnim vezama [2, 5, 9].

2.4 Uloga proteina

Proteini su najbrojniji i funkcionalno najrazličitiji molekuli u živom svetu. Svaki od njh ima veoma važnu ulogu u organizmu. Na primer:

- Enzimi su proteini koji olakšavaju hemijske reakcije. Učestvuju u skoro svim reakcijama u ćelijama i pomažu u izgradnji novih molekula.
- Antitela su proteini koje proizvodi imuni sistem da bi pomogli u odstranjivanju stranih supstanci i kako bi se borile protiv infekcija. Oni se vezuju za nepoznate čestice, poput bakterija i virusa čime brane telo
- Kontrakcijski proteini učestvuju u kontrakcijama mišića i kretanju.
- Strukturni proteini su vlaknasti i obezbeđuju strukturu i podršku ćelijama. Učestvuju u izdgradnji kose, noktiju, kože, kostiju, itd.
- Transportni proteini prenose molekule kroz telo.
- Hormonski proteini prenose signale kako bi upravljali biološkim procesima među ćelijama, tkivima i organima.
- Skladišni proteini čuvaju aminokiseline za kasniju upotrebu [10, 11, 12].

Glava 3

Podaci i metode binarne klasifikacije

U ovom poglavlju biće opisani korišćeni podaci i način njihovog predstavljanja u računaru. Zatim će ukratko biti opisane metode binarne klasifikacije koje su korišćene za predviđanje funkcija proteina.

3.1 Podaci

Podaci o proteinima mogu se pronaći u biomedicinskim bazama podataka, a neke od njih prikazane su u tabeli 3.1.

Baza podataka	URL	Opis
UniProtKB	uniprot.org	Proteinske sekvence i funkcije proteina
PFAM	pfam.xfam.org	Proteinske familije
PDB	wwpdb.org	Eksperimentalno utvrđene strukture
ModBase	modbase.compbio.ucsf.edu	Strukture utvrđene predviđanjem
I2D	ophid.utoronto.ca	Interakcije između proteina
GEO	www.ncbi.nlm.nih.gov/geo	Podaci o genskoj ekspresiji
PRIDE	www.ebi.ac.uk/pride	Podaci dobijeni masenom sprektometrijom

Tabela 3.1: Prikaz nekih javno dostupnih biomedicinskih baza podataka [1, 2].

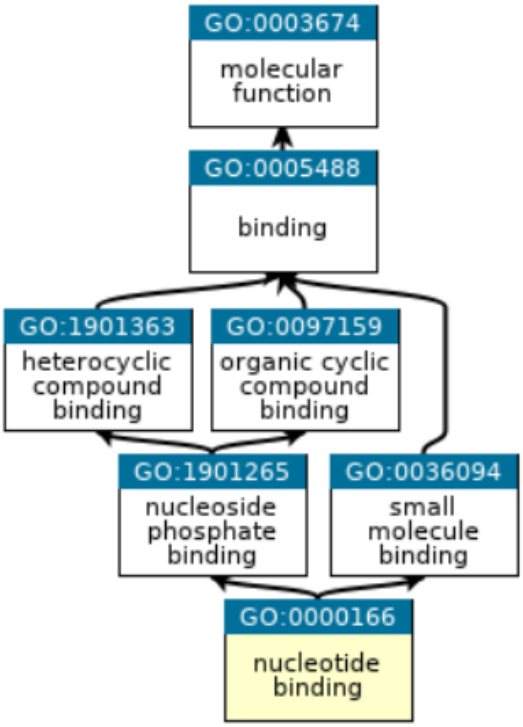
3.1.1 Predstavljanje proteina

Kao što je već rečeno, proteini su izgrađeni od 20 različitih aminokiselina, a svaka aminokiselina ima jedinstveni simbol (tabela 2.2). Najjednostavniji način za predstavljanje proteina u računaru jeste kao niska karaktera nad azbukom $\Sigma = \{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\}$. Nad ovako predstavljenim proteinima mogu su koristiti algoritmi za rad sa tekstom kao što je poravnanje sekvenci [1].

3.1.2 Predstavljanje funkcije proteina

Da bi predviđanje funkcija proteina bilo moguće neophodno je da postoje dobro definisani odnosi između funkcija. Sistem za predstavljanje funkcije proteina koji se trenutno najviše koristi je *Gene Ontology*. Ovaj sistem deli funkcije proteina na tri ontologije: biološki procesi (BPO), molekulske funkcije (MFO) i ćelijske komponente (CCO).

Svaka ontologija predstavljena je kao usmereni aciklički graf gde su čvorovima pridruženi nazivi funkcija, a grane koje ih povezuju definišu relaciju „is_a". Hijerarhijska organizacija obezbeđuje da svaki čvor ima specifičniju funkciju od roditeljskog čvora. U ovoj hijerarhiji jedan čvor može imati više roditeljskih čvorova što je prikazano na slici 3.1 [2, 13]. U korenu svake ontologije nalazi se funkcija sa nazivom te ontologije, a u listovima su najspecifičnije fukcije.



Slika 3.1: Prikaz svih predaka lista označenog funkcijom „nucleotid binding" u ontologiji molekulskih funkcija.

3.2 Binarni klasifikatori

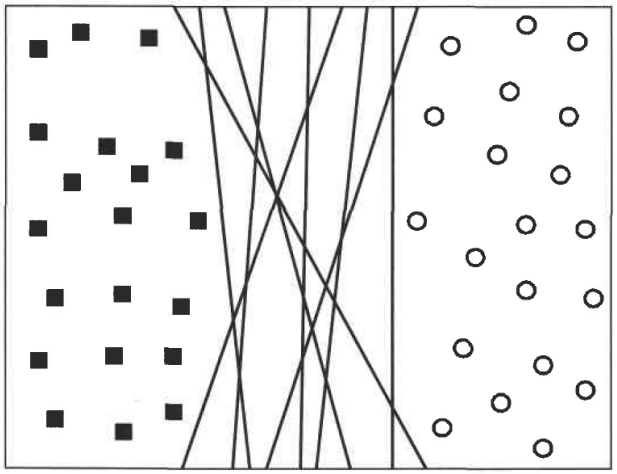
Klasifikacija, odnosno, zadatak dodeljivanja objekata jednoj od više predefinisanih kategorija, rasprostranjen je problem koji obuhvata mnoštvo različitih primena. Primeri uključuju otkrivanje spam poruka na osnovu zaglavlja poruke i njenog sadržaja, kategorisanje ćelija kao malignih ili beningnih na osnovu rezultata magnetne rezonance, klasifikaciju galaksija na osnovu njihovog oblika, itd. Binarna klasifikacija je slučaj klasifikacije u kojoj postoje tačno dve predefinisane kategorije u koje treba razvrstati date objekte. Obično se za jednu kategoriju kaže da je to pozitivna klasa, a za drugu da je negativna.

Svaki klasifikator upotrebljava algoritam za učenje kako bi odredio model koji najbolje odgovara vezi između skupa atributa i klase ulaznih podataka. Model koji algoritam generiše trebalo bi da odgovara ulaznim podacima kao i da tačno predviđa klasu slogova koje ranije nije video.

3.2.1 Metod potpornih vektora

Metod potpornih vektora (engl. *support vector machine*) je tehnika za klasifikaciju zasnovana na ideji vektorskih prostora. Ova metoda generiše klasifikacioni model koji predstavlja funkciju. Osnovni algoritam definisan je za binarnu klasifikaciju.

Osnovna ideja ove metode jeste pronalazak razdvajajuće hiperravni takve da su instance iste klase sa iste strane ravni. Sa tako postavljenim uslovom, razdvajajućih hiperravni može biti više od jedne što je prikazano na slici 3.2. Iako svaka od ravni razdvaja podatke bez greške, nema garancije da će se podjednako dobro ponašati sa novim podacima koje je potrebno klasifikovati.



Slika 3.2: Primeri razdvajajućih hiperravni [14]

Na slici 3.3 izdvojene su dve hiperravni B_1 i B_2 i za svaku su dodate dve pomoćne hiperavni b_{i1} i b_{i2}. Pomoćne ravni paralelne su glavnoj i pomerene u levu ili desnu stranu do najbliže instance jedne klase. Rastojanje između pomoćnih ravni odnosno rastojanje između najbližih instanci iz obe klase u odnosu na hiperravan naziva se *margina*, a instance oslonjene na hiperravni su *potporni vektori*. Cilj je pronaći hiperravan koja maksimizuje veličinu margine. Sa slike 3.3 jasno se vidi da je bolja hiperravan B_1. Jednačina optimalne hiperravni predstavlja klasifikacioni model. Korak klasifikacije nepoznate instance sastoji se iz izračunavanja njenog rastojanja od hiperravni na osnovu čega se određuje klasa kojoj instanca pripada [14, 15].

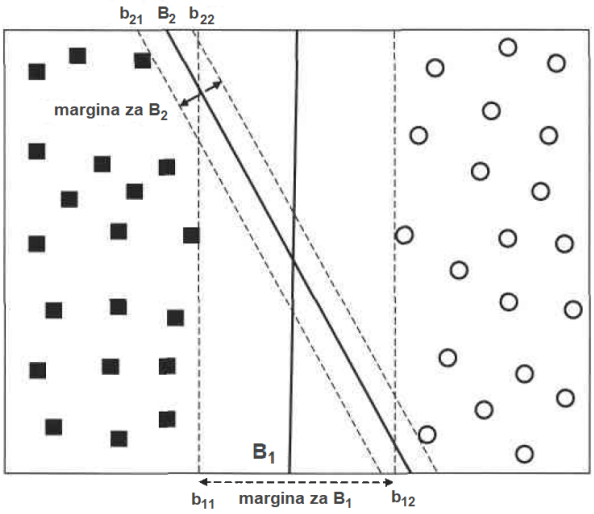
Jednačina hiperravni je

$$w \cdot x + w_0 = 0$$

gde je w_0 slobodan član. Na osnovu jednačine rastojanja tačke od hiperravni

$$\frac{|w \cdot x + w_0|}{||w||_2}$$

i činjenice da za svaku od tačaka sa ovih hiperravni važi $w \cdot x + w_0 = 1$, dobija se da je ukupno rastojanje između klasa, u pravcu normalnom u odnosu na optimalnu hiperravan $\frac{2}{|w|}$. Tako se optimalna hiperravan dobija pronalaženjem koeficijenata koji maksimizuju ovaj izraz pod uslovima da su sve tačke sa pravih strana te hiperravni odnosno da su podaci linearno razdvojivi. Optimizacioni problem može da se zapiše i kao problem minimizacije i glasi:



SLIKA 3.3: Margine razdvajajućih hiperravni [14]

$$\min_{w,w_0} \frac{||w||_2}{2}$$
$$y_i(w \cdot x_i + w_0) \geq 1 \qquad i = 1, \ldots, N$$

Dodatni uslov će obezbediti da sve tačke budu na većem rastojanju od hiperravni u odnosu na potporne vektore [16].

S obzirom da je čest slučaj da podaci nisu linearno razdvojivi potrebno je prihvatiti greške tj. dozvoliti da se neka instanca nađe sa pogrešne strane razdvajajuće hiperravni. U tu svrhu uvode se nove promenljive, ξ_i koje mere koliko je svaka pogrešno klasifikovana instanca udaljena od hiperravni. Taj metod nazivamo metod potpornih vektora sa *mekom marginom*. Optimizacioni problem se menja:

$$\min_{w,w_0} \frac{||w||_2}{2} + C\sum_{i=0}^{N} \xi_i$$
$$y_i(w \cdot x_i + w_0) \geq 1 - \xi_i \qquad i = 1, \ldots, N$$
$$\xi_i \geq 0 \qquad i = 1, \ldots, N$$

Metaparametar C kontroliše koliki značaj imaju greške. Ukoliko je vrednost jednaka nuli, onda greške ne igraju nikakvu ulogu, a ako je vrednost velika, onda su greške veoma važne, a pravac hiperravni i širina pojasa nisu bitne [16].

3.2.2 Logistička regresija

Logistička regresija (engl. *logistic regression*) je statistički zasnovan metod za analizu skupa podataka u kom jedna ili više nezavisnih promenljivih određuju ishod. Osnovna pretpostvaka je Bernulijeva rasporedela ciljne promenljive y pri datim vrednostima atributa x odnosno, za date vrednosti atributa x, postoji parametar $\mu \in [0, 1]$ tako da važi:

$$p(y = 1|x) = \mu$$

odakle je $p(y = 0|x)$ jednoznačno određeno. Zadatak je sličan kao u prethodnoj metodi, potrebno je pronaći razdvajajuću hiperravan koja deli podatke tako da sa jedne strane budu instance iste klase. Najjednostavniji je linearan model:

$$f(x) = w \cdot x$$

S obzirom da funkcija uzima vrednosti iz intervala $[-\infty, \infty]$, a parametar μ mora imati vrednost iz intervala $[0, 1]$ da bi verovatnoća bila ispravno definisana, ovakav model nije prihvatljiv. Zbog toga se vrednost linearnog modela transformiše monotonom funkcijom u interval $[0, 1]$. U te svrhe, najčešće se koristi sigmoidna funkcija:

$$\sigma(t) = \frac{1}{1 + exp(-t)}$$

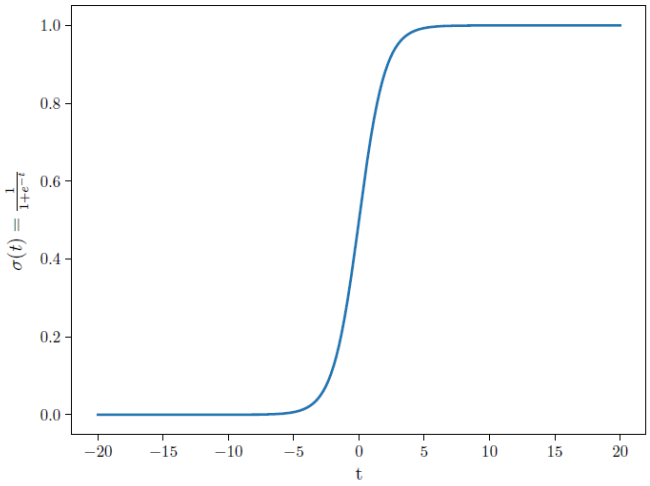
čiji je grafik prikazan na slici 3.4. Nakon transformacije vrednosti linearnog modela sigmoidnom funkcijom, model logističke regresije određen je relacijom:

$$p_w(y = 1|x) = \sigma(w \cdot x)$$

Iz toga se može odrediti i puna specifikacija problema:

$$p_w(y|x) = \sigma(w \cdot x)^y (1 - \sigma(w \cdot x))^{1-y}$$

Verovatnoća da instanca pripada jednoj klasi je veća što je instanca dalje od hiperravni sa odgovarajuće strane [16].



Slika 3.4: Grafik sigmoidne funkcije [16]

Ocena parametara ovog modela zasniva se na principu maksimalne verodostojnosti. Uz pretpostavku nezavisnosti instanci, funkcija verodostojnosti zadata je izrazom:

$$\mathcal{L}(w) = \prod_{i=1}^{N} p_w(y_i|x_i)$$

i potrebno je rešiti problem:

$$\max_w \mathcal{L}(w)$$

Ukoliko se pređe na negativan logaritam verodostojnosti optimizacioni problem postaje:

$$\min_{w} -\sum_{i=0}^{N}[y_i \log f_w(x_i) + (1-y_i)\log(1-f_w(x_i))]$$

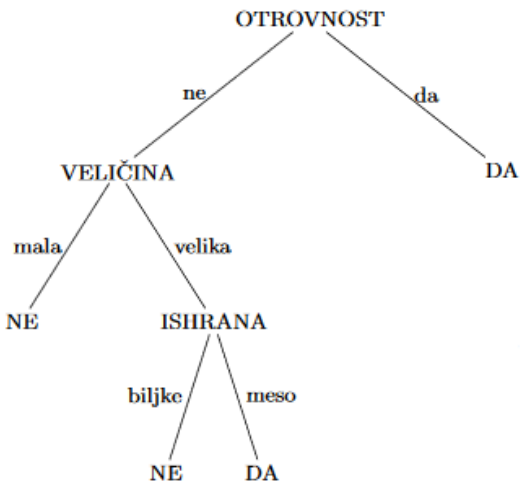
3.2.3 Slučajne šume

Metod slučajne šume (engl. *random forests*) spada u grupu metoda specijalno dizajnirane za stabla odlučivanja. Ona kombinuje predviđanja više različitih stabala, gde je svako stablo generisano na osnovu vrednosti nezavisnog skupa slučajno odabranih vektora.

Stabla odlučivanja

Problem klasifikacije rešava se postavljanjem pažljivo sastavljenih pitanja o atributima podataka. Pitanja se postavljaju sve dok nije moguće zaključiti klasu date instance. Niz pitanja i odgovora organizovan je u hijerarhijsku strukturu koja se sastoji iz čvorova i direktinih grana. Na slici 3.5 je prikazano stablo odlučivanja koje određuje da li je životinja opasna ili ne na osnovu podataka o otrovnosti, veličini i ishrani.

Svaki list u stablu ima dodeljenu klasu, a unutrašnji čvorovi i koren sadrže uslove koji razdvajaju instance sa različitim karakteristikama. Grane predstavljaju odgovor na pitanje čvora iz kog izlaze. Jednom kada je stablo konstruisano, klasifikacija je jednosmerna. Kreće se od korenog čvora i za konkretnu instancu odgovara se na pitanja koja se nalaze u čvorovima praćenjem odgovarajućih grana sve do listova koje sadrže konačan odgovor. Klasa koja je pridružena listu dodeljuje se instanci [14].



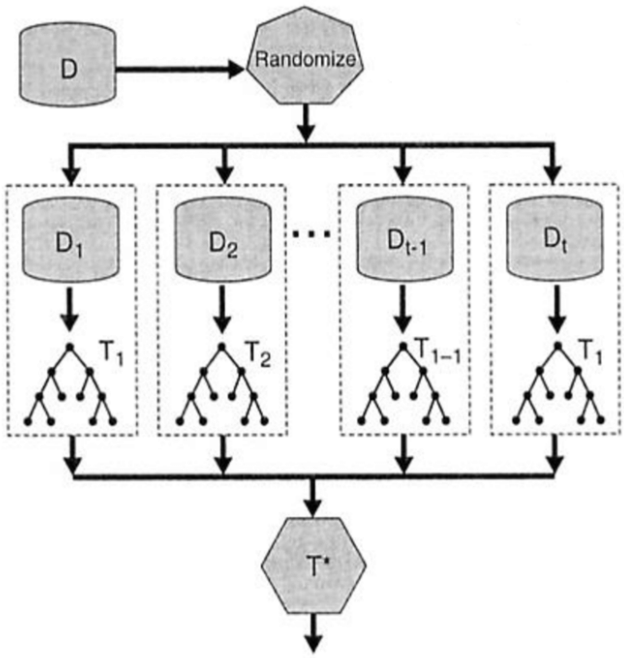
Slika 3.5: Primer stabla odlučivanja koje određuje da li je životinja opasna ili ne [17]

Predviđanje korišćenjem slučajnih šuma

Osnovna ideja je kombinovanje više stabala odlučivanja u jedan model. Stabla koja se kombinuju treniraju se nad nezavisnim, slučajno odabranim podskupovima

podataka, a može se koristiti i slučajno odabran podskup atributa. Obučavaju se nad različitim skupovima kako bi njihove greške bile što slabije korelisane [16].

Instanca se klasifikuje glasanjem. Svako od konstruisanih stabala klasifikuje instancu u jednu od dve klase, a zatim se svi odgovori broje i odgovor klasifikatora slučajne šume je ona klasa koja ima više glasova tj. ona klasa koju je više stabala predvidelo. Slika 3.6 ilustruje proces treniranja i klasifikacije.



Slika 3.6: Primer slučajne šume [14]. Prvo se iz početnog skupa slučajno odabiraju instance i formira se podskup za svako stablo. Zatim se nad odgovarajućim podskupovima treniraju stabla. Svako stablo klasifikuje nepoznatu instancu i odgovori se kombinuju u jedan, konačan, odgovor modela slučajnih šuma.

3.3 Evaluacija modela binarne klasifikacije

Metode binarne klasifikacije za test instancu daju jedan od dva moguća odgovora, na primer, ,,da" ili ,,ne". Pošto se jedna klasa obično posmatra kao pozitivna a druga kao negativna, neka u ovom primeru ,,da" bude pozitivna klasa, a ,,ne" neka bude negativna klasa. Prilikom ocenjivanja kvaliteta klasifikacionog modela od značaja su 4 veličine:

- tp - broj instanci za koju je predviđena pozitivna klasa i čija je stvarna klasa pozitivna
- tn - broj instanci za koju je predviđena negativna klasa i čija je stvarna klasa negativna
- fp - broj instanci za koju je predviđena pozitivna klasa, a čija je stvarna klasa negativna
- fn - broj instanci za koju je predviđena negativna klasa, a čija je stvarna klasa pozitivna.

Kada su definisane ove veličine, mogu se odrediti i neke mere kvaliteta kao što je, na primer, tačnost modela. Tačnost (*engl. accuracy*) određuje koliko je instanci tačno klasifikovano u odnosu na ukupan broj instanci i definiše se formulom:

$$\text{accuracy} = \frac{tp + tn}{tp + tn + fp + fn}$$

Ova metrika se često koristi u mašinskom učenju, međutim, ona ne daje uvek dobru ocenu metoda. U slučaju nebalansiranih klasa[1] model može da daje visoku vrednost za tačnost, a da ipak loše predviđa. Razlog je to što često loši modeli predviđaju skoro uvek samo jednu, dominantnu klasu, a pošto je instanci dominantne klase značajno više, većina instanci će biti ispravno klasifikovana. Međutim, cilj je napraviti model koji će biti uspešan u klasifikovanju obe klase, a ne samo jedne [18].

Zbog toga se definišu još neke mere kvaliteta modela. Prva je preciznost (*engl. precision*), a druga je odziv (*engl. recall*) i definišu se formulama:

$$\text{precision} = \frac{tp}{tp + fp} \qquad \text{recall} = \frac{tp}{tp + fn}$$

Preciznost određuje koliko je pozitivnih instanci ispravno klasifikovano u odnosu na ukupan broj instanci koje su klasifikovane kao pozitivne. Sa druge strane, odziv određuje udeo ispravno klasifikovanih pozitivnih instanci u odnosu na ukupan broj pozitivnih instanci u skupu.

Preciznost i odziv pojedinačno nisu korisne. Ukoliko su sve instance klasifikovane kao pozitivne, odziv će biti maksimalan, međutim preciznost će biti katastrofalna, dok sa druge strane, ukoliko su sve instance klasifikovane kao negativne model ne greši i preciznost je maksimalna, ali odziv je veoma loš. Stoga ima smisla posmatrati ih zajedno što se često radi određivanjem njihove harmonijske sredine koja je nazvana f_1-mera [16]:

$$f_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

[1] Pod nebalansiranim klasama podrazumeva se da u skupu podataka postoji mnogo više instanci koji pripadaju jednoj klasi u odnosu na broj instanci koje pripadaju drugoj klasi.

Glava 4

Implementacija predviđanja funkcija proteina

Predviđanje funkcije proteina vršeno je metodama binarne klasifikacije i to metodom potpornih vektora, slučajnim šumama i logističkom regresijom. Trenirani su binarni klasifikatori za pojedinačne funkcije ontologije molekulskih funkcija. Ulaz predstavljaju sekvence proteina za koje je određeno da li obavljaju ili ne obavljaju konkretnu funkciju. Odgovor koji svaki od klasifikatora daje je da li zadati protein izvršava odgovarajuću funkciju ili ne.

Kada su svi modeli istrenirani, testirani su nad 100 proteina različitih organizama. Za svaki protein traže se odgovori svih binarnih klasifikatora jedne metode, a njihovi odgovori se spajaju u konačan odgovor - podgraf ontologije koji predstavlja funkciju zadatog proteina.

4.1 Podaci

Podaci o proteinima korišćeni u ovom radu preuzeti su sa adrese https://biofunctionprediction.org/cafa-targets/CAFA3_training_data.tgz. Oni su podeljeni u dve datoteke:

1. **uniprot_sprot_exp.fasta** - proteini i njihove sekvence,
2. **uniprot_sprot_exp.txt** - proteini i eksperimentalno utvrđene funkcije koje obavljaju.

Dodatne informacije o organizmima iz kojih proteini potiču preuzete su sa `https://www.uniprot.org/` i to za organizme:

- čovek (human)
- miš (mouse)
- pacov (rat)
- ešerihija koli (ecoli)
- arabidopsis (arath).

Informacije o ontologijama preuzete su sa http://geneontology.org/docs/download-ontology/ u OBO formatu.

Preuzeti podaci nisu bili u pogodnom obliku za ulaz klasifikatora zbog čega je bilo neophodno njihovo parsiranje.

Ontologija Datoteka *go.obo* sadrži funkcije iz sve tri ontologije. Njenim parsiranjem izdvojena je ontologija molekulskih funkcija (MFO). Ona se sastoji iz približno 12000 čvorova, međutim, neki čvorovi su zastareli (*engl. obsolete*) zbog čega su izbačeni iz grafa. Pored toga, postoje čvorovi koji predstavljaju alternativni identifikator nekog drugog čvora te su takvi čvorovi ujedinjeni u jedan. Time je broj čvorova smanjen na 11078 molekulskih funkcija. Broj je dodatno umanjen zbog prirode podataka. Pre svega, približno 5500 funkcija se uopšte ne pojavljuje u trening skupu što znači da za njih nema pozitivnih instanci odnosno proteina koji ih izvršavaju pa su one izbačene. Zatim, oko 5000 funkcija se pojavljuje manje od 100 puta u trening skupu. Pokušaji treninga klasifikatora za takve funkcije su bili neuspešni te su i one izbačene iz skupa. Nakon svih redukcija ostalo je 399 funkcija sa 100 ili više pojavljivanja u trening skupu za koje su trenirani modeli.

Proteini i funkcije Parsiranjem *uniprot_sprot_exp.txt* izdvojeno je više informacija - proteini sa funkcijama koje obavljaju kao i funkcije sa proteinima za koje je utvrđeno da ih obavljaju. Prvi skup podataka je obogaćen podacima iz ontologije s obzirom da su zadati samo krajnji čvorovi, a ne i svi preci, kako bi se dobio ceo podgraf ontologije koji predstavlja funkciju proteina. Drugi skup poslužio je za prebrojavanje pojavljivanja funkcije u trening skupu kao i za kasnije formiranje skupa pozitivnih i negativnih instanci.

Sekvence proteina U datoteci *uniprot_sprot_exp.fasta* nalazi se 66817 proteina. Među njima se nalaze i proteini koji ne obavljaju neku od funkcija iz MFO. Pored toga, postoje proteini čije sekvence nisu validne u smislu aminokiselina koje sadrže. Pod validnim sekvencama podrazumevaju se samo one koje se sastoje isključivo iz 20 standardnih aminokiselina. Nakon eliminacije ovakvih proteina preostaje 34785 onih koji obavljaju bar jednu molekulsku funkciju. Nakon redukcije broja funkcija na 399 smanjio se i skup proteina. Naime, izbačeni su svi proteini za koje je utvrđeno da vrše neku od eliminisanih funkcija. Nakon svih redukcija, veličina trening skupa je 20960.

4.1.1 Predstavljanje proteina

Mnoge metode mašinskog učenja koriste vektore kao ulaz zbog čega je pogodno da se niska aminokiselina prepiše u niz. Jedan pogodan način za to jeste prebrojavanjem pojavljivanja svakog mogućeg trigrama nad azbukom 20 standardnih aminokiselina. Dimenzija jednog niza je samim tim 20^3, a jedan element sadrži broj pojavljivanja odgovarajućeg trigrama u niski aminokiselina. Ono što je neophodno jeste da za svaki trigram postoji jedinstveno određen redni broj u nizu. U te svrhe, prvo je potrebno odrediti brojeve pojedinačnih aminokiselina, a šema korišćena u ovoj implementaciji prikazana je u tabeli 4.1.

A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19

Tabela 4.1: Preslikavanje aminokiselina u broj

Sada, za svaki trigram može da se odredi njegov jedinstveni broj koji predstavlja poziciju u nizu i to formulom:

$$kmer\_index = aa_1 * 20^2 + aa_2 * 20 + aa_3$$

Sa ovakvim preslikavanjem trigrama u brojeve, jednostavnim prolaskom kroz nisku sa korakom od 3 karaktera dobija se odgovarajući niz.

4.2 Treniranje modela

Program je pisan u programskom jeziku Python i korišćene su implementacije metoda binarne klasifikacije iz Python-ove biblioteke *sklearn*. Trenirano je 399 modela za svaki metod pojedinačno. Početni skup proteina podeljen je na trening i test skup u razmeri 3:1. Tokom treniranja izvršen je i odabir modela na validacionom skupu koji je izdvojen iz trening skupa u istoj razmeri. Odabir najboljeg modela izvršen je na osnovu f1-mere.

Nakon što je obučavanje jednog modela završeno, on je sačuvan u posebnoj datoteci sa nazivom koji odgovara identifikatoru funkcije za koju je model treniran i to korišćenjem još jedne Python-ove biblioteke - *pickle*. Ova biblioteka omogućava čuvanje i kasnije čitanje modela mašinskog učenja u pogodnom obliku, tako da nema potrebe za obučavanjem ispočetka već su modeli odmah spremni za predviđanje.

Metod potpornih vektora Prilikom odabira modela birana je vrednost za parametar C i to iz skupa $\{0.01, 0.1, 1, 10\}$. Pored toga, odabiran je bolji od dva kernela, linearan i gausov. Zbog dugačkog treniranja jednog modela i velikog broja modela koje je trebalo obučiti, svim nizovima redukovana je dimenzionalnost na 1000. U te svrhe korišćena je Python-ova implementacija algoritma analize glavnih komponenti iz biblioteke *sklearn*.

Logistička regresija Prilikom odabira modela birana je vrednost za parametar C i to iz skupa $\{0.0001, 0.001, 0.01, 0.1, 1\}$. Neki modeli nisu uspevali da nauče ništa iz podataka i njihova f1-mera bila je jednaka 0. Za takve modele izvršen je dodatan trening na proširenom skupu podataka. Proširenje skupa se odnosi na generisanje sintetičkih instanci kako bi se ublažila nebalansiranost pozitivnih i negativnih instanci. Za proširivanje skupa korišćena je Python-ova biblioteka *imblearn*, a skup je obogaćen tako da odnos pozitivnih i negativnih instanci bude 1:2.

Slučajne šume Kod modela slučajnih šuma trenirani su modeli sa različitim brojem stabala iz skupa $\{100, 400, 700, 1000\}$. Slično kao kod logističke regresije, za modele čija je f1-mera bila 0 izvršen je dodatan trening sa dodatnim pozitivnim instancama.

4.3 Objedinjavanje modela

Nakon što su svi modeli za odabrani metod obučeni prelazi se na testiranje. Izdvojen je skup od 100 proteina nad kojim je testiran prediktor. Prediktor je formiran na osnovu 399 prethodno obučenih modela koji se na samom početku učitavaju u memoriju. Prilikom predviđanja funkcije jednog proteina, protein se prosleđuje kao ulaz svakom od binarnih klasifikatora koji daju vrednosti 0 ili 1. Ujedinjavanjem svih odgovora dobija se konačan odgovor. Sve funkcije za koje je odgovarajući klasifikator dao 1 kao odgovor predstavljaju čvor podgrafa.

4.4 Evaluacija modela

Kao mera kvaliteta pojedinačnih modela korišćena je f_1 mera. U okviru biblioteke *sklearn* implementirana je funkcija koja određuje ovu vrednost na osnovu pravih i predviđenih klasa instanci iz test skupa.

Ista mera korišćena je za evaluaciju konačnog prediktora koji ujedinjuje sve odgovore. S obzirom da prediktor daje strukturu kao odgovor (usmereni aciklički graf) treba preciznije definisati kako se ova mera određuje. Pretpostavimo da je datoj test instanci pridružen izlazni vektor $y = [0, 1, 1, 0, 1, 1]$, a da je prediktor dao odgovor $y' = [0, 0, 1, 1, 0, 1]$ za istu test instancu. Poređenjem dva vektora može se lako utvrditi koje su klase ispravno određene, a koje pogrešno odnosno mogu se odrediti veličine tp, tn, fp i fn opisane u sekciji 3.3:

$$y' = [\underset{\in tn}{0}, \underset{\in fn}{0}, \underset{\in tp}{1}, \underset{\in fp}{1}, \underset{\in fn}{0}, \underset{\in tp}{1}]$$

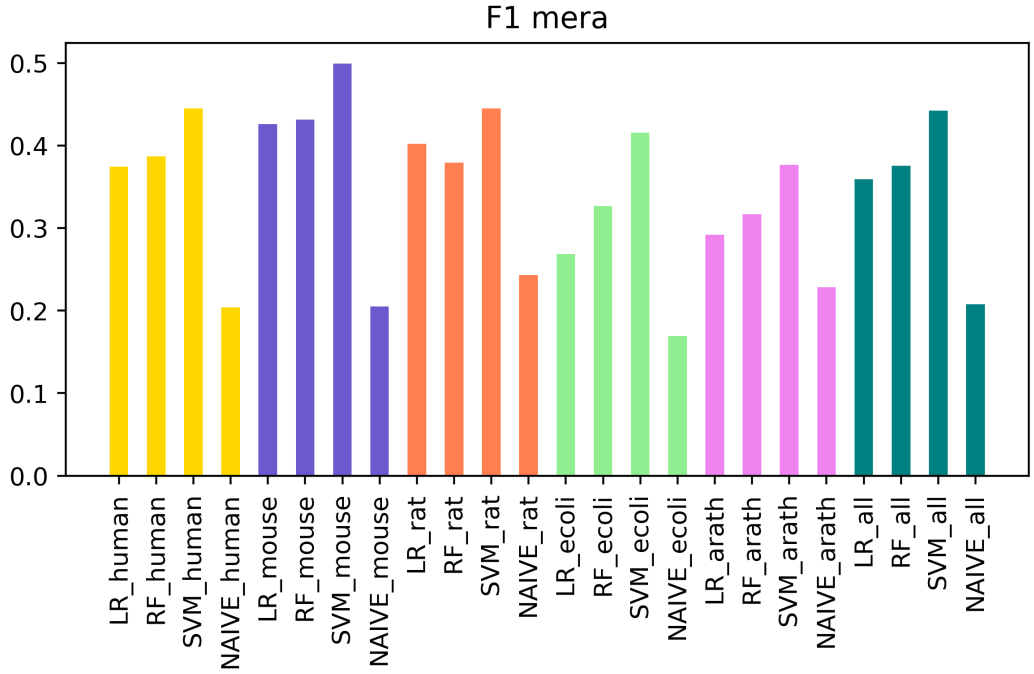
Na osnovu ovih veličina dalje se mogu odrediti preciznost i odziv, a onda i f_1 mera.

Glava 5

Rezultati

Za svaki od tri prethodno opisana metoda binarne klasifikacije trenirano je po 399 modela na celom trening skupu koji su kasnije ujedinjeni u 3 prediktora za predviđanje funkcije proteina. Pored toga, implementiran je još jedan jednostovaniji klasifikator koji je poslužio kao osnovna metoda za poređenje rezultata. U pitanju je naivni klasifikator koji svakom čvoru dodeljuje vrednost koja odgovara njegovoj frekvenciji pojavljivanja u trening skupu i tako formirani graf pridružuje svakom test primeru [2].

Naivni klasifikator je testiran na istom test skupu kao i 3 prediktora, a rezultat je prikazan na slici 5.1. Prilikom svakog testiranja računata je prosečna vrednost f_1 mere i to na nivou pojedinačnih organizama, kao i na nivou celog skupa. Sa slike se može videti da je svaki od tri prediktora bolji od naivnog klasifikatora. Pored toga, metod potpornih vektora daje najbolje rezultate kako za pojedinačne organizme, tako i za ceo skup.



Slika 5.1: Poređenje performansi prediktora i naivnog klasifikatora. Rezultati jednog organizma prikazani su istom bojom.

U tabelama 5.1, 5.2 i 5.3 prikazane su f_1 mere za 20 pojedinačnih klasifikatora za svaku upotrebljenu metodu binarne klasifikacije. Pored toga, prikazan je i broj pojavljivanja u trening skupu za svaku funkciju, kao i udeo broja pojavljivanja funkcije

u celom trening skupu. Vrednosti za sve klasifikatore prikazane su na adresi http://poincare.matf.bg.ac.rs/~anja_bukurov/master.

Funkcija	f1-mera	Broj pojavljivanja u skupu	Procenat pojavljivanja u skupu
GO:0005525	0.3	290	1.4%
GO:0008134	0.11	861	4.1%
GO:0019899	0.11	2381	11.4%
GO:0044325	0.13	144	0.7%
GO:0003723	0.12	1899	9.1%
GO:0016787	0.25	3202	15.3%
GO:0019900	0.12	5268	20.8%
GO:0019888	0.13	71	0.3%
GO:0004722	0.45	123	0.6%
GO:0004867	0.18	105	0.5%
GO:0005496	0.11	71	0.3%
GO:0016829	0.12	316	1.5%
GO:0046872	0.12	1824	8.7%
GO:0016758	0.11	172	0.8%
GO:0003729	0.13	5611	21.5%
GO:0004930	0.12	312	1.5%
GO:0030594	0.42	86	0.4%
GO:0004497	0.16	5710	21.7%
GO:0030246	0.18	180	0.9%
GO:0031406	0.12	5730	21.7%

Tabela 5.1: Prikaz f1-mere za pojedinačne klasifikatore metode slučajne šume

Iako su se trenirani prediktori pokazali bolje od naivnog klasifikatora, nemaju približnu moć predviđanja u poređenju sa aktivnim rezultatima prikazanim na poslednjem CAFA takmičenju, najrelevantnijem takmičenju u ovoj oblasti. Planovi za unapređenje prediktora uključuju treniranje posebnih modela za svaki od organizama na skupu proteina koji potiču isključivo iz odabranog organizma, zatim povećanje trening skupa i korišćenje raznovrsnijih metoda binarne klasifikacije.

Funkcija	f1-mera	Broj pojavljivanja u skupu	Procenat pojavljivanja u skupu
GO:0005525	0.46	290	1.4%
GO:0005524	0.2	671	3.2%
GO:0051117	0.11	101	0.5%
GO:0008134	0.26	861	4.1%
GO:0019899	0.31	2381	11.4%
GO:0019904	0.15	681	3.3%
GO:0044877	0.16	985	4.7%
GO:0003714	0.16	292	1.4%
GO:0046982	0.16	663	3.2%
GO:0044325	0.28	144	0.7%
GO:0003723	0.39	1899	9.1%
GO:0031625	0.15	303	1.5%
GO:0003779	0.28	402	1.9%
GO:0030234	0.27	1047	5.0%
GO:0019901	0.18	668	3.2%
GO:0042803	0.15	1218	5.8%
GO:0042802	0.25	2480	11.9%
GO:0005102	0.3	1423	6.8%
GO:0016787	0.49	3202	15.3%
GO:0004222	0.46	128	0.6%

Tabela 5.2: Prikaz f1-mere za pojedinačne klasifikatore metode logistička regresija

Funkcija	f1-mera	Broj pojavljivanja u skupu	Procenat pojavljivanja u skupu
GO:0005525	0.36	290	1.4%
GO:0005524	0.15	671	3.2%
GO:0008134	0.27	861	4.1%
GO:0019899	0.19	2381	11.4%
GO:0044877	0.11	985	4.7%
GO:0003714	0.15	292	1.4%
GO:0046982	0.16	663	3.2%
GO:0003723	0.33	1899	9.1%
GO:0003779	0.24	402	1.9%
GO:0030234	0.18	1047	5.0%
GO:0019901	0.14	668	3.2%
GO:0042802	0.11	2480	11.9%
GO:0005102	0.15	1423	6.8%
GO:0016787	0.43	3202	15.3%
GO:0004222	0.24	128	0.6%
GO:0019888	0.21	71	0.3%
GO:0004722	0.43	123	0.6%
GO:0004867	0.18	105	0.5%
GO:0030145	0.11	162	0.8%
GO:0051287	0.2	123	0.6%

Tabela 5.3: Prikaz f1-mere za pojedinačne klasifikatore metode potpornih vektora

Bibliografija

[1] Predrag Radivojac. *A (not so) Quick Introduction to Protein Function Prediction*. 2013.

[2] Jovana Kovačević. *Strukturna predikcija funkcije proteina i odnos funckionalnih kategorija i neuređenosti*. 2015.

[3] Rick Ricer Michael A. Lieberman. *Biochemistry, Molecular Biology, and Genetics*. New Science Press Ltd, 2004.

[4] OpenStax. "Anatomy and Phisiology". In: (2013).

[5] Denise R. Ferrier Richard A. Harvey. *Biochemistry Fifth Edition*. Lippincott Williams & Wilkins, a Wolters Kluwer business, 2011. ISBN: 978-1-60831-412-6.

[6] Vesna Spasojević-Kalimanovska Slavica Spasić Zorana Jelić-Ivanović. *Opšta biohemija*. 2002.

[7] Dubravka Cvorišćec Ivana Čepelak. *Štrausova medicinska biokemija*. Medicinska naklada, 2009.

[8] Marek Kimmel Andrzej Polanski. *Bioinformatics*. Springer-Verlag, 2007. ISBN: 978-3-540-24166-9.

[9] Dagmar Ringe Georgy A Pesko. *Protein Structure and Function*. New Science Press Ltd, 2004.

[10] Regina Bailey. "The Function and Structure of Proteins". In: (2019).

[11] "Role of proteins in the body". In: (2011).

[12] "What are proteins and what do they do?" In: (2019).

[13] *Gene Ontology*.

[14] Kumar Vipin Tan Pang-Ning Steinbach Michael. *Introduction to Data Mining*. Pearson Education, 2006.

[15] Jelena Graovac. *Prilog metodama klasifikacije teksta: matematički modeli i primene*. 2014.

[16] Anđelka Zečević Mladen Nikolić. *Mašinsko učenje*.

[17] Mladen Nikolić Predrag Janičić. *Veštačka inteligencija*.

[18] Ishaan Dey. "Evaluating Classification Models". In: (2019).