SVEUČILIŠTE U ZAGREBU

FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 1223

Predviđanje površine dostupne otapalu iz slijeda aminokiselinskih ostataka

Irena Popović

Zagreb, travanj 2011.

Sadržaj

1. Uvod

Proteini su odgovorni za sve stanične funkcije, stoga je poznavanje njihove strukture te razumijevanje interakcija proteina od velike važnosti. Neke od mogućih primjena vezane su za razvoj novih lijekova i cjepiva, analizu metaboličkih reakcija, te promatranje i praćenje razvoja organizma.

Postoje mnoge javne baze sa sad već velikim brojem poznatih 3-D struktura proteina dobivenih eksperimentalno pomoću spektroskopskih i kristalografskih tehnika, no sam broj mogućih proteinskih interakcija i struktura je toliki da je naspram njega broj pronađenih i definiranih neznatan. Kao zamjena skupim i vremenski ograničenim eksperimentalnim tehnikama udruženim snagama bioinformatičara i računaraca razvijene su brojne metode i algoritmi za računalno predviđanje strukture proteina, no postignuti rezultati još nisu dovoljno točni.

Svakodnevno se traže nova bolja rješenja i dodaju novi uvjeti već postojećim metodama, sve u svrhu povećanja točnosti.

Smatra se da slijed aminokiselina koje grade protein sadrži dovoljno informacija da se odredi njegova trodimenzionalna struktura. No to je moguće samo za proteine koji imaju veliku sličnost u slijedu s proteinom čija je struktura već poznata.

Metoda koja se često koristi pri predviđanju strukture proteina i mjesta proteinskih interakcija jest predviđanje dostupnosti otapalu (engl. *solvent accessibility*). Dostupnost otapalu pokazuje stupanj interakcije aminokiselinskog ostataka (engl. *amino acid residue*) s molekulama otapala i bitan je pokazatelj stanja smatanja proteina. Kako je većina mjesta interakcije proteina smještena na njihovoj površini, predviđanje izloženosti ostataka važno je za razumijevanje i predviđanje mjesta interakcije.

Ovaj rad se bavi jednom od metoda predviđanja površine dostupne otapalu.

2. Teorijski uvod

2.1. Proteini

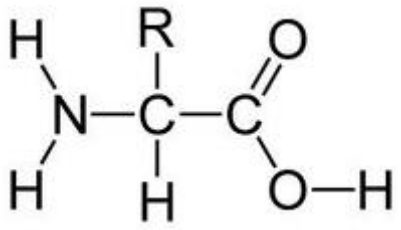
Proteini [1] [2] nastaju međusobnim povezivanjem aminokiselina peptidnom vezom u duge lance od stotinu do približno tisuću aminokiselinskih ostataka. Razumijevanje strukture proteina ključno je za razumijevanje problema predviđanja dostupnosti otapalu, te će ovdje biti dan kratak opis njihove strukture i formiranja. Ukratko će se opisati neke od strukturalnih karakteristika proteina s posebnim naglaskom na površinu dostupnu otapalu (engl. *accessible surface area*).

2.1.1. Građa proteina

Osnovne građevne jedinice proteina su aminokiseline[1], kao takve uvjetuju oblik proteina te njegova svojstva. Sadrže dvije karakteristične funkcionalne skupine: amino-skupinu NH_2 i karboksilnu skupinu COOH.

Kod aminokiselina koje nas prvenstveno zanimaju kao sastavni dio bjelančevina, amino-skupina je uvijek smještena u α-položaju [1] prema karboksilnoj grupi, odnosno amino-skupina je vezana na drugi C-atom u lancu.

Općenita formula aminokiseline je NH2-CHR-COOH.



Slika 2.1 Općenita struktura aminokiselina

Aminokiseline se međusobno razlikuju po R-skupini koja se još naziva i bočni ogranak (engl. *side chain*), a može biti od jednog atoma vodika (u slučaju glicina) do složenih molekula. Upravo bočni ogranci aminokiselina koje grade protein određuju njegova svojstva i funkciju.

Tablica 2.1 Najvažnije aminokiseline podijeljene obzirom na vrstu R-skupine [2]

Aminokiselina	**Simbol**	**FASTA zapis**
Nepolarna, alifatska R-skupina		
Glicin	Gly	G
Alanin	Ala	A
Prolin	Pro	P
Valin	Val	V
Leucin	Leu	L
Izoleucin	Ile	I
Metionin	Met	M
Aromatska R-skupina		
Fenilalanin	Phe	F
Tirozin	Tyr	Y
Triptofan	Trp	W
Polarna, nenabijena R-skupina		
Serin	Ser	S
Treonin	Thr	T
Cistein	Cys	C
Asparagin	Asn	N
Glutamin	Gln	Q
Pozitivno nabijena R-skupina		
Lizin	Lys	K
Histidin	His	H
Arginin	Arg	R
Negativno nabijena R-skupina		
Asparaginska kiselina	Asp	D
Glutaminska kiselina	Glu	E

U tablici 2.1 koja je preuzeta iz knjige autora Lehningera i suradnika [2] nalazi se dvadeset tipičnih aminokiselina podijeljenih u grupe obzirom na narav R-skupine od koje su izgrađene. Osim imena aminokiselina i kratica, u tablici se nalazi zapis aminokiselina u jednoslovačnom kodu [3](tzv. *FASTA zapis*).

Spajanjem dvije ili više aminokiselina u peptidni lanac, aminokiseline gube vodu, a preostali dio aminokiseline naziva se ostatak aminokiseline (engl. *amino acid residue*).

Jedno od bitnih svojstava aminokiselina je topljivost. Zbog svoje polarnosti, voda dobro otapa nabijene i polarne tvari. Tvari koje se dobro otapaju u vodi zovemo hidrofilnima, a one koje se u vodi loše otapaju zovemo hidrofobnima. Topljivost aminokiseline u otapalu (vodi) određena je polarnošću bočnog ogranka. Važnost svojstava bočnog ogranka dolazi zbog utjecaja koji ima na interakciju ostataka aminokiselina s drugim strukturama, bilo to unutar istog proteina ili između proteina. Raspodjela hidrofobnih i hidrofilnih aminokiselina ima veliku važnost pri promatranju proteinskih struktura i interakcija.

2.1.2. Strukture proteina

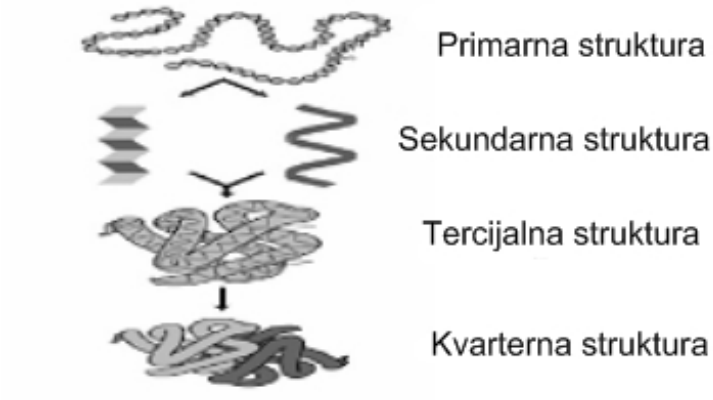
Primarna struktura [1] proteina je redoslijed tj.sekvenca aminokiselina koje izgrađuju taj protein i genetski je određena. Često se među različitim proteinima opaža velika podudarnost slijeda aminokiselina te se za njih kaže da su homologni proteini.

Sekundarna struktura [1] je prostorni raspored samog peptidnog lanca i rezultat je vodikovih veza >N-H…..O=, kojima se peptidni lanci mogu međusobno povezati. Vodik iz >N-H skupine jednog lanca naslanja se na nepodijeljeni par elektrona karboksilnog kisika iz drugog lanca ako se približe na udaljenost od 0.28nm. Energija vodikove veze iznosi oko 10% energije kovalentne veze C-C. No takvih veza između lanaca nastane mnogo pa se oslobađa znatna energija i postiže se stabilnija struktura.

Postoji nekoliko tipova sekundarne strukture koji su posebno stabilni te se pojavljuju u jako velikom broju proteina. Najčešće strukture su α–uzvojnica [1], koja je uzrokovana vodikovim vezama unutar polipeptidnog lanca, te β-ploča [1]. Za razliku od α–uzvojnica, β–ploča je uzrokovana vodikovimim vezama između susjednih polipeptidnih lanaca.

Cjelokupan trodimenzionalni raspored svih atoma nekog proteina zove se tercijarna struktura proteina. Za razliku od sekundarne strukture koja opisuje raspored atoma koji se u polipeptidnom lancu nalaze blizu jedan drugom, tercijarna struktura uzima u obzir i mnoge udaljenije dijelove lanca. Aminokiseline koje su u lancu jako udaljene, u tercijarnoj strukturi mogu biti jedna pored druge. Ona je uvjetovana i međusobnim djelovanjem bočnih ogranaka aminokiselina. Pojedini dijelovi peptidnog lanca u tercijarnoj strukturi međusobno se povezuju vodikovim, disulfidnim, ionskim i van der Waalsovim vezama.

Neki proteini sastoje se od više odvojenih polipeptidnih lanaca. Njihov trodimenzionalni raspored nazivamo kvartarna struktura.

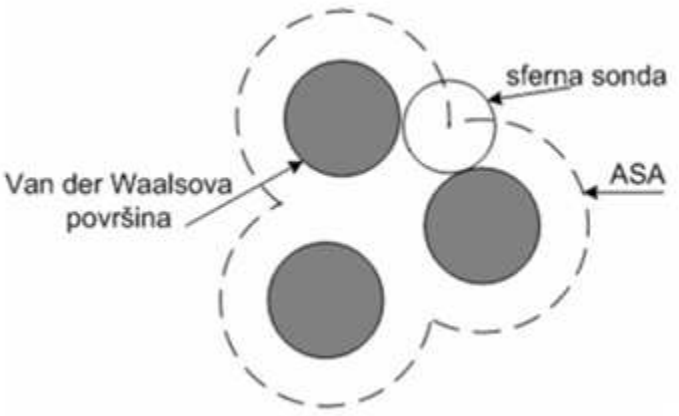


Slika 2.2 Strukture proteina

2.1.3. Otapalu dostupno područje površine

U vodenom okruženju hidrofobni bočni ogranci okreću se prema unutrašnjosti proteina. Jedan od načina kvantifikacije hidrofobnog ukopavanja aminokiselina je pomoću otapalu dostupnog područja površine. Time je opisano područje površine proteina na kojem je moguć kontakt s otapalom u kojem se protein nalazi (najčešće vode). Poznavanje topologije površine proteina nam je prvenstveno bitno zbog činjenice da obično dijelovi površine koji su izloženi direktno sudjeluju u interakcijama proteina. Topologija površine usko je povezana i sa samom funkcijom proteina.

Van der Waalsov radijus atoma [4] je radijus imaginarne sfere koji se koristi za razne reprezentacije atoma. Eksperimentalno je određen za atome mjerenjem prostora između para nevezanih atoma unutar kristala. Ako svakome atomu pridružimo njegov van der Waalsov radijus, odnosno svaki atom zamijenimo sferom van der Waalsova radijusa, dobiti ćemo van der Waalsovu površinu.



Slika 2.3 Otapalu dostupno područje površine atoma

Otapalu dostupno područje površine (engl. *Accessible Surface Area, ASA*) sada možemo definirati pomoću sferne sonde koja predstavlja model molekule otapala. Sferna sonda kotrlja se po van der Waalsovoj površini te definira otapalu dostupno područje površine kao što se vidi i na slici 2.3 preuzetoj iz rada [4]. Za radijus sonde se uzima obično iznos od 1.4Å koji predstavlja radijus molekule vode.

ASA vrijednost pojedine aminokiseline dobijemo tako da zbrojimo ASA vrijednosti svih atoma koji grade tu aminokiselinu. Osim ukupne ASA (engl. *Total*)[], postoje još četiri vrijednosti koje se često spominju pri proučavanju topologije površine aminokiseline:

a) ASA glavnog lanca (engl. *Backbone*) – suma ASA svih atoma koji grade glavni lanac aminokiseline

b) ASA bočnog lanca (engl. *Side-chain*) – suma ASA svih atoma koji grade bočni lanac aminokiseline

c) ASA polarnog dijela (engl. *Polar*) – suma ASA svih polarnih atoma (atomi kisika i dušika) koji grade aminokiselinu

d) ASA nepolarnog dijela (engl *Non-polar*) – suma ASA svih nepolarnih atoma (svi atomi osim kisika i dušika) koji grade aminokiselinu

Često se umjesto ASA vrijednosti koristi njezina relativna vrijednost (engl. *Relative Solvent Accessibility, RSA*) , tj. odnos ASA vrijednosti ostatka i maksimalne ASA vrijednosti toga ostatka dok ona nije sastavni dio proteina. Pošto se aminokiseline nikad ne nalaze same u prostoru, te vrijednosti su izračunate tako da se promatra aminokiselina okružena sa još dvije aminokiseline (npr. Ala-X-Ala ili Gly-X-Gly trojka).

Kao i kod apsolutnih vrijednosti, postoji pet relativnih vrijednosti (ukupna, glavnog lanca, bočnog lanca, polarna i nepolarna) koje se računaju omjerom apsolutne i standardne vrijednosti pomnožene sa 100. Relativne vrijednosti ASA aminokiselina opisuju kolikim dijelom svoje površine, izraženim u postotcima, je aminokiselina dostupna otapalu.

U ovom radu apsolutne i relativne ASA vrijednosti aminokiselinskih ostataka dobivene su korištenjem PSAIA alata [5]. Pri predviđanju se koristila samo ukupna relativna ASA vrijednost koja će se dalje u tekstu spominjati kao stvarna rASA.

2.1.4. Proteinske baze podataka

Dugi niz godina, u središtu biologije i biokemije nalazili su se geni te struktura i funkcija nukleinskih kiselina. Razvoj tehnologije donio je metode za masovno sekvencioniranje genoma, kao i metode za određivanje trodimenzionalne strukture proteina. Ubrzo se stvorila potreba za pohranjivanjem velikog broja informacija o sekvenci i strukturi podataka. Ti podaci danas su pohranjeni u bazama podataka od kojih su najvažnije PDB (*Protein Data Bank*) i UniProt.

PDB [6] je baza proteinskih struktura u čijim zapisima se nalaze podaci o prostornim koordinatama svih „teških“ atoma u proteinu. Pod teškim atomima se podrazumijevaju svi atomi osim vodika. Trenutno se u bazi nalazi više od 50.000 struktura dobivenih kristalografijom X-zrakama ili NMR-om (magnetskom rezonancijom).

2.2. Rad s pomičnim prozorima

Dužina proteinskih lanaca mjeri se u stotinama i tisućama aminokiselinskih ostataka. Kada bi se predviđanje vršilo uzimanjem svih aminokiselinskih ostatka lanca u obzir, stvorio bi se problem zbog ograničenih računalnih resursa. Poboljšanje točnosti klasifikacije time ne bi bilo zajamčeno, već bi samo postojala mogućnost da se unese šum zbog nedostatka podataka o dugim aminokiselinskim sekvencama.

Pri predviđanju se stoga često koriste pomični prozori fiksne duljine. Obično je riječ o prozorima s neparnim brojem aminokiselinskih ostataka, duljine između 3 do 21. Povećanjem duljine prozora povećava se broj aminokiselinskih ostataka na početku i kraju sekvence za koje se ne vrši predviđanje. Svaki od elemenata prozora sadrži više različitih svojstava na temelju kojih se vrši predviđanje.

Pomični prozor se pomiče od početka do kraja sekvence, dok se predviđanje vrši za središnji ostatak prozora .

GLU	ARG	TYR	GLU	ASN	LEU	PHE	ALA	GLN	LEU	ASN	ASP	ARG

Slika 2.4 Primjer pomičnog prozora duljine 9. Trenutni središnji ostatak je PHE, slijedeći središnji ostatak biti će ALA, dok je prethodni bio LEU.

2.3. Predviđanje dostupnosti otapalu regresijom

Postoje dva bitno različita načina predviđanja površine dostupne otapalu. Predviđanje klasifikacijom svrstava aminokiselinske ostatke u unaprijed određen broj kategorija (klasa), dok se predviđanjem regresijom pokušava odrediti točna ASA vrijednost aminokiselinskih ostataka. Često se umjesto površine dostupne otapalu predviđa njezina relativna vrijednost.

Jedan od problema klasifikacije je proizvoljan odabir pragova koji čine granice između klasa i predstavljeni su kao relativna ASA vrijednost da se izbjegne problem zbog velike razlike u ASA vrijednostima pojedinih aminokiselinskih ostataka. Proizvoljan izbor pragova ne uzima u obzir raspodjelu ASA vrijednosti te se gubi dio informacija, osim toga utječe na točnost predviđanja te ujedno predstavlja problem pri usporedbi metoda.

U ovome radu predviđanje će se vršiti regresijom, a predviđat će se relativna ASA vrijednost aminokiselinskih ostataka. Usporedba dosadašnjih rezultata i korištenih metoda bit će detaljno obrađena u poglavlju 5.

3. Podaci

3.1. Korišteni skupovi proteina

U istraživanjima se koriste razni skupovi proteina što otežava usporedbu točnosti između različitih metoda. Pojedini skupovi se češće upotrebljavaju te su stoga neki od njih odabrani i u ovom radu.

Prvi korišteni skup prvotno je predložen u radu Rosta i Sandera [7], sadrži 126 proteinskih lanaca pri čemu je sličnost između lanaca manja od 25%, što je naknadno opovrgnuto. Usprkos tome skup se još uvijek često koristi, a dalje u radu zvat ćemo ga Rost-Sander skup.

Drugi izabrani skup dolazi iz rada Manesha [8] te sadrži 215 nehomolognih proteina sa sličnosti između lanaca manjom od 25%, u daljnem tekstu će ga se nazivati Manesh skup.

Još jedan od skupova korištenih u radu, jest skup nadalje u tekstu spominjan kao skup Cuff-Barton, prema autorima koji su ga prvi koristili [9]. Sastavljen je od 502 nehomologna proteinska lanca s više od 83 000 ostataka i međusobnom sličnosti između proteina manjom od 25%.

Posljednji skup zvat ćemo Carugo skup koji se prvo koristio u radu autora Caruga [10]. Skup čine 338 nehomolognih monomernih proteinskih struktura sa sličnosti među proteinima u skupu manjom od 25%.

3.2. Svojstva na osnovu kojih će se vršiti predviđanje

Za predviđanje površine dostupne otapalu koristila su se sljedeća dva svojstva:

- sekvenca duljine 9 do 21 aminokiselinskih ostataka
- profili slijeda svakog od ostataka unutar prozora

Kako svaka aminokiselina ima svoj profil slijeda, ukupan broj svojstava po elementu pomočnog prozora iznosi 21. Pomični prozor se stoga može smatrati vektorom dimenzije 21×n, gdje je n duljina prozora.

3.2.1. Profili slijeda aminokiselinskih ostataka

Profili slijeda [11] (engl. *probability profiles*) koriste se kao mjera evolucijske očuvanosti, to su vjerojatnosti pronalaska bilo koje od dvadeset standardnih aminokiselina na onom mjestu u sekvenci na kojemu se nalazi aminokiselina čiji se profil određuje. Često se koriste za predviđanje različitih strukturalnih karakteristika.

Za određivanje profila slijeda prvo se izvuku imena aminokiselina koje grade određeni lanac te se zapišu u FASTA [3] formatu, pri čemu formiraju niz slova odnosno sekvencu. Zatim se takva sekvenca propušta kroz PSI-BLAST[12] algoritam koji kao rezultat daje PSSM matricu (engl. *Position-Specific Scoring Matrix*) odnosno profil. Profili su građeni u odnosu na SWIISS-PROT proteinsku bazu podataka. U poglavlju 4.1 bit će detaljno objašnjen način određivanja profila slijeda.

3.3. Struktura podataka

Podaci o strukturi proteina nalaze se u PDB [6] formatu. Iako se unutar PDB datoteka nalaze samo osnovni podaci, datoteka sadrži velik broj informacija kao što su način na koji je određena struktura proteina, od kojih se lanaca ostataka, molekula i atoma sastoji promatrani proteinski kompleks te prostorni raspored svakog pojedinog atoma.

PDB datoteke služe kao ulazni podaci iz kojih se određuju ASA vrijednosti ostatka primjenom PSAIA alata [5]. Za svaku od PDB datoteka alat generira po jednu izlaznu XML datoteku. Nakon što se obrade sve PDB datoteke skupa koji se koristi za predviđanje, pokreće se skripta napisana u Python programskom jeziku koja u izlazne XML datoteke dodaje profile slijeda. Druga skripta generira prozore različitih duljina te sve navedene podatke zapisuje u odgovarajuće ARFF (engl. *Attribute-Relation File Format*) datoteke.

3.3.1. Priprema podataka za regresiju

Za predviđanje relativne površine dostupne otapalu koristit će se regresija algoritmom slučajnih šuma [11][13]. Metoda slučajnih šuma detaljno će biti objašnjena u poglavlju 4.2. U radu se koriste alati za strojno učenje: Rattle biblioteke u statističkom paketu R, te vlastito napisane R skripte gdje se kao ulazni podaci koriste ARFF datoteke. ARFF format počinje imenom relacije, slijedi definicija atributa te na kraju dolazi blok s podacima. Definicija jednog atributa daje mu ime i jednu od mogućih vrijednosti: numeric (brojčane), string (tekstualne) ili skup kategorija u vitičastim zagradama.

Kako se koriste pomični prozori različitih duljina, za svaki skup proteina imat ćemo 7 ulaznih datoteka. Svaka datoteka sadržava podatke o ostacima unutar prozora, njihovim profilima, te relativne ASA vrijednost. U nastavku je prikazan jedan zapis unutar ARFF datoteke:

1A7J,A,247,PHE,PRO,TYR,LEU,THR,SER,MET,ILE,HIS,0,0,0,0,5,0,0,0,0,0,0,0,0,0,84,0,0,0,0,11,0,0,0,0,0,0,13,0,0,0,0,0,0,0,0,87,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,18,82,0,0,0,0,0,0,0,0,0,0,100,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,68,0,0,0,0,0,32,0,0,0,18,0,0,0,0,0,0,0,0,0,0,0,0,0,82,0,0,0,0,0,0,0,0,0,0,0,0,0,0,100,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,69,31,0,0,0,0,0,0,0,0,11,0,0,0,0,18,0,0,59,0,0,0,0,0,11,0,0,0,0,0,30

U primjeru se radi o prozoru duljine 9 aminokiselinskih ostataka, kojemu je središnji ostatak za kojega se vrši predviđanje treonin (THR). Sekvenca pripada lancu A proteina 1A7J, serijski broj središnjeg ostatka unutar lanca je 247. Nakon tih informacija slijede profili slijeda za svih 9 ostataka unutar prozora. Zadnji broj je relativna ASA vrijednost središnjeg ostatka, dakle treonina. Ti podaci služe za treniranje slučajnih šuma. Skupovi za treniranje i testiranje formirat će se iz orginalnog skupa krosvalidacijom.

Krosvalidacija je postupak dijeljenja orginalnog skupa podataka na *N* podskupova tako da jedan od tih podskupova čini skup za testiranje, a preostalih (*N-1*) skup za treniranje. Jednom kad je skup podijeljen, raspored trening i test podataka se rotira *N* puta te se s tim podacima radi *N* predviđanja. U ovom radu će se raditi 10-struka krosvalidacija, dakle radit će se 10 predviđanja za svaki skup proteina.

4. Metode

4.1. Određivanje profila slijeda. PSI-BLAST

PSI-BLAST [11][12] je poboljšani algoritam tehnike BLAST pomoću kojeg je moguće pronaći evolucijsku očuvanost aminokiselinskog ostatka. Koristi se metodom sekvencijalnog poravnanja (engl. *sequence alignment*) za prepoznavanje sličnosti dvaju proteina koji nisu blisko povezani, tj.nemaju bliskog zajedničkog homologa. Proteini mogu biti sekvencijalno ili strukturalno slični, pri čemu sličnosti nisu nužno povezane. Temeljna ideja algoritma je prepoznati zajedničku strukturu iz slabe sekvencijalne sličnosti. Toj se problematici pristupa metodom sekvencijalnog poravnavanja u sklopu tehnike BLAST.

4.1.1. Sekvencijalno poravnanje

Prvi korak u gradnji profila jest za neku proteinsku sekvencu pronaći da li ona pripada već poznatoj familiji proteina. Poravnavanjem primarnih sekvenci koje predstavljaju neku familiju proteina vrši se prepoznavanje sličnih elemenata što može upućivati na funkcionalnu, strukturnu ili evolucijsku povezanost između sekvenci. Poravnati aminokiselinski ostaci se prikazuju kao redovi unutar matrice. Ako dvije poravnate sekvence dijele zajedničkog pretka, nepravilnosti u sekvencama mogu se interpretirati kao točke mutacije ili pak praznine koje su posljedice delecije i insercije tokom evolucije, u odnosu na izvornu sekvencu, homologa. Dijelovi sekvenci za koje se ocjeni da su slični ili čak jednaki, nazivaju se motivi. Za motive se smatra da se tokom evolucije nisu mijenjali tj.da su konzervirani te da su strukturno ili funkcionalno važni.

Poravnate sekvence u odnosu na aminokiselinske ostatke prikazuju se grafički i tekstualno. U gotovo svim zapisima, sekvence su zapisane u recima tako da se slični ili jednaki aminokiselinski ostaci nalaze u istom stupcu.

U grafičkim prikazima na slici 4.1 pruzetoj iz rada [4] boje simboliziraju različite skupine aminokiselinskih ostataka. Tako su crvenom bojom prikazane male te hidrofobne (uključujući aromatske) aminokiseline.

```
AAB24882        TYHMCQFHCRYVNNHSGEKLYECNERSKAFSCPSHLQCHKRRQIGEKTHEHNQCGKAFPT 60
AAB24881        --------------------YECNQCGKAFAQHSSLKCHYRTHIGEKPYECNQCGKAFSK 40
                                    ****: .***:  * *:** * :****.:* ********..

AAB24882        PSHLQYHERTHTGEKPYECHQCGQAFKKCSLLQRHKRTHTGEKPYE-CNQCGKAFAQ- 116
AAB24881        HSHLQCHKRTHTGEKPYECNQCGKAFSQHGLLQRHKRTHTGEKPYMNVINMVKPLHNS 98
                 **** *:***********:***:**.: .***************    :  *.: :
```

Slika 4.1 Sekvencijalno poravnanje između dva proteina iz zinc finger porodice proteina. Prikaz je dobiven pomoću programa ClustalW pri čemu je korišten FASTA format prikaza aminokiselinskih ostataka. Identični dijelovi prikazani su simbolom '*' konzervirani s ':' te sa simbolom '.' djelomično konzervirani ostaci.

Sekvence se mogu poravnati na lokalnoj i globalnoj razini. Pri globalnom poravnanju svi se ostaci pojedinačno poravnaju te se čuva duljina sekvence. Lokalno poravnavanje ima veći učinak kada se sekvence razlikuju, ali se sumnja na eventualnu sličnost pojedinih dijelova. Također postoji i hibridno poravnavanje koje je kombinacija lokalnog i globalnog.

```
FTFTALILLAVAV
F--TAL-LLA-AV

FTFTALILL-AVAV
--FTAL-LLAAV--
```

Slika 4.2 Globalno (gore) i lokalno (dolje) poravnanje [4]

4.1.2. BLAST. BLOSUM supstitucijske matrice

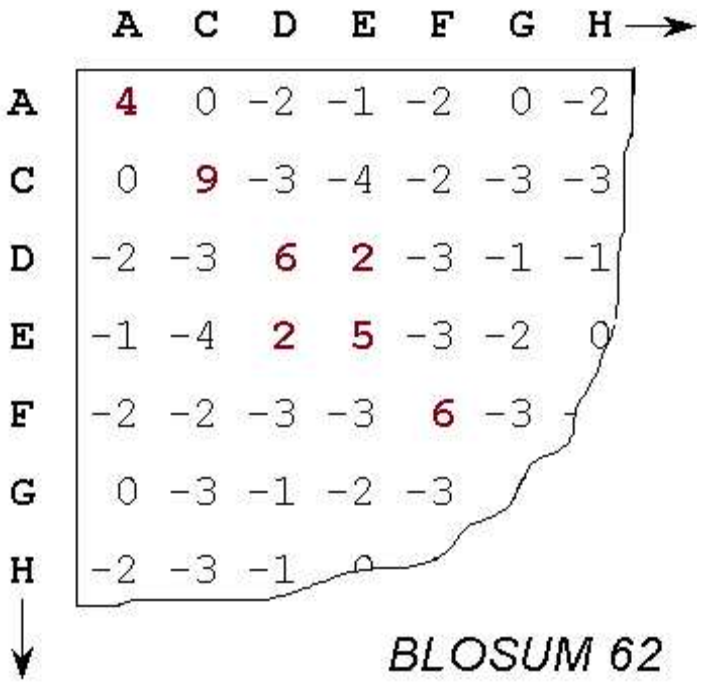
BLAST [11][12] (engl. *Basic Local Alignment Search Tool*), je algoritam za usporedbu sekvenci aminokiselinskih ostataka proteina ili nukleotida DNK lanca. BLAST omogućuje da se za neku sekvencu koja se ispituje, pretraži proteinska baza podataka.

Tako algoritam, pretražujući bazu, nalazi sekvence koje odgovaraju sekvenci koja se ispituje, pri tome zadovoljavajući određeni prag sličnosti koji u pravilu zada korisnik. Kada se radi sravnjenje sljedova aminokiselinskih ostataka, algoritam BLAST koristi supstitucijsku matricu za procjenu sličnosti sljedova. Postoji nekoliko takvih matrica, među kojima se najčešće koriste BLOSUM i PAM matrice.

BLOSUM (engl. *Blocks Substitution Matrix*) bazira se na lokalnom poravnavanju, a prvi put je predstavljena u radu Henikoff and Henikoff [14]. Nastala je empirijski na temelju poznatih i vrlo konzervativnih regija proteinskih familija u proteinskoj bazi podataka i računanja relativnih frekvencija pojavljivanja pojedinih aminokiselina. Za razliku od PAM matrica koje su dobivene uspoređivanjem poznatih i sličnih sekvenci tj.onih koje slabo divergiraju, BLOSUM matrice su nastale iz evolucijski divergentnih sekvenci.

Postoji više BLOSUM matrica ovisno o bazi podataka iz koje su nastale, a označuju se brojem koji upućuje na sličnost sljedova iz kojih su nastale. Primjerice, BLOSUM80 znači da se radi o sekvencama sličnosti iznad 80%. Takva će se matrica koristiti u slučaju manje evolucijski divergentnih sekvenci, dok će se BLOSUM45 koristiti u slučaju više divergentnih sekvenci.

Na slici 4.3 pruzetoj iz rada [4] prikazana je matrica BLOSUM62, kakva se koristila u ovome radu.



Slika 4.3 BLOSUM62 – broj 62 upućuje na sličnost od barem 62%

.

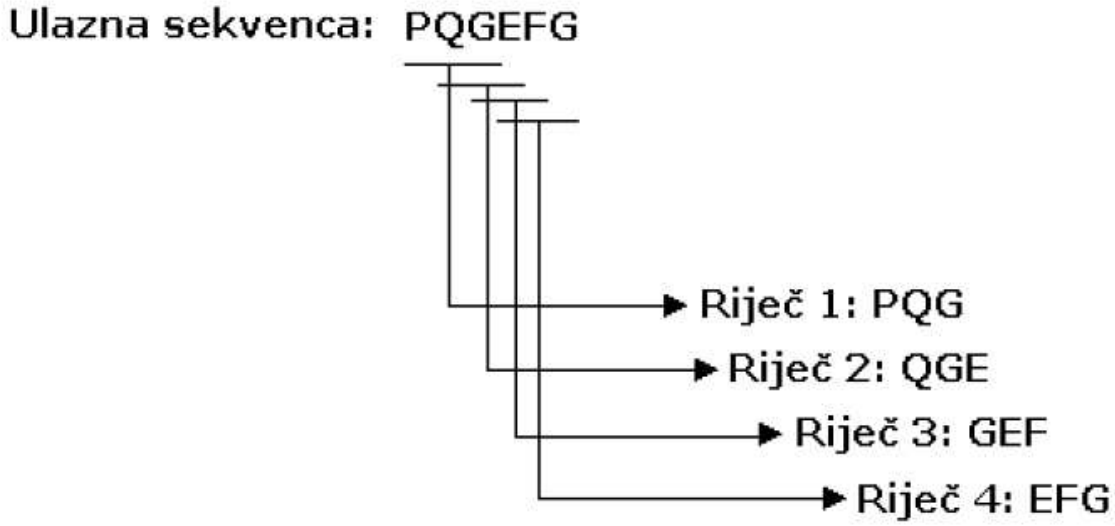
Vrijednosti matrice mogu se izračunati sljedećim izrazom:

$$S_{ij} = \left(\frac{1}{\lambda}\right)\log\left(\frac{p_{ij}}{q_i \cdot q_j}\right) \tag{4.1}$$

gdje je p_{ij} vjerojatnost da će aminokiseline *i* i *j* zamijeniti jedna drugu u homolognoj sekvenci, a q_i i q_j su vjerojatnosti slučajnog nalaženja aminokiselina *i* i *j* u sekvenci proteina. λ je skalirajući faktor.

4.1.3. Opis algoritma

Osnovna ideja algoritma jest da svako dobro ocjenjeno lokalno poravnanje dviju sekvenci gotovo uvijek sadrži dobro očuvanu jezgru. Za parove ostataka u slijedu, određuje se ocjena poravnanja i ako je ona iznad nekog zadanog praga, taj se par ostataka naziva dobro ocjenjeno lokalno poravnanje tj. HSP (engl. *High-scoring Segment Pairs*). BLAST pretražuje sekvence nalazeći dobro ocjenjena poravnanja između sekvence koja se ispituje i onih sekvenci u bazi sekvenci. Algoritam radi na način da ulaznu sekvencu koja se ispituje podijeli u trigrame, tj.u riječi od po 3 slova kao što je prikazano na slici 4.4 uzetoj iz rada [4].



Slika 4.4 Rastavljanje sekvence u riječi od po tri slova

Za svaku takvu riječ se pronalaze ciljne riječi odnosno svi trigrami na alfabetu aminokiselina koji imaju dovoljno veliku sličnost s početnim. Sličnost se iščitava iz supstitucijske matrice za svaki par aminokiselinskih ostataka u trigramu. Primjerice, uspoređujući riječ PQG s riječi PEG ocjena sličnosti je (iščitavajući BLOSUM62)15, dok je sličnost riječi PQG i primjerice, riječi PQA 12. Ako se zada prag sličnosti 13, tada je ciljna riječ PEG, te se kao takva zadržava na listi ciljnih riječi. Nakon što se pronađu ciljne riječi za svaku riječ od tri slova ulazne sekvence, slijedi traženje tih istih ciljnih riječi u sekvencama baze.

Kada se pronađe ciljna riječ u sekvenci baze, ona može upućivati da s odgovarajućom riječi ulazne sekvence čini jezgru. Da bi se to zaključilo vrši se proširivanje u oba smjera. Odnosno, gledaju se susjedni ostaci, te računa ocjena. Proširivanje poravnanja traje sve dok ocjena sličnosti (koja se čita iz BLOSUM matrice) ne počne padati (slika 4.5 pruzeta iz rada [4]).

Slika 4.5 Proširivanje ciljne riječi na susjedne dok ocjena sličnosti ne počne padati

U cilju bržeg rada algoritma osmišljena su poboljšanja. Jezgru produljenja poravnanja sada čine dva pogotka sličnih riječi takva da leže na istoj dijagonali. To znači da su dvije riječi jednako udaljene u obje sekvence. Pri tome se mora smanjiti prag sličnosti za ciljne riječi da bi se zadržala osjetljivost. Ujedno se i smanjuje broj produljenja. Produljenje se radi Smith- Waterman algoritmom koji vrši poravnavanje s razmacima (engl. *gapped alignment*). Ova se verzija BLAST algoritma prema tome naziva gapped BLAST.

4.1.4. Procjena značajnosti ocjene lokalnog poravnanja

Dobro ocjenjeno lokalno poravnanje ne mora nužno značiti da su odgovarajuće sekvence slične te da imaju zajedničkog homologa. Lokalno poravnanje može biti posljedica slučajnosti. Stoga se radi model slučajnih sekvenci u cilju uklanjanja takvih pojava.

Jednostavan model proteina sastoji se od slučajno odabranih aminokiselinskih ostataka na temelju njihovih specifičnih frekvencija pojavljivanja [11] (engl. *background probability*). Ocjena lokalnog poravnanja poprima negativnu vrijednost u slučaju da je poravnanje slučajno. Inače bi dugačka poravnanja imala visoku vrijednost ocjene poravnanja neovisno da li su evolucijski povezani. U dovoljno dugačkim sekvencama duljine *m* i *n*, značajnost ocjene lokalnog poravnanja karakteriziraju dva parametra *K* i *λ*. Očekivani broj dobro ocjenjenih lokalnih poravnanja, *E*-value, koji su posljedica slučajnosti, a vrijednost ocjene barem *S′* jest:

$$E = \left(\frac{N}{S'} \right) \tag{4.2}$$

gdje je $N = mn$, a S' normalizirana ocjena S (S je prag za dobro ocjenjeno lokalno poravnanje):

$$S' = \frac{\lambda S - \ln K}{\ln 2} \tag{4.3}$$

Iz izraza za *E*-value može se dobiti izraz za normaliziranu vrijednost praga koje mora zadovoljiti lokalno poravnanje da bi bilo dobro ocjenjeno. Navedeni izrazi se odnose na BLAST bez razmaka, ali se mogu primijeniti i na BLAST s razmacima. Međutim, statistički parametri *K* i *λ* se više ne određuju teorijski, već eksperimentalno.

U slučaju BLAST algoritma bez razmaka, parovi dobro ocjenjenih poravnatih riječi odnosno aminokiselinski ostaci koji čine riječ, pojavljuju se s frekvencijom:

$$q_{ij} = P_i P_j e^{\lambda_u s_{ij}} \tag{4.4}$$

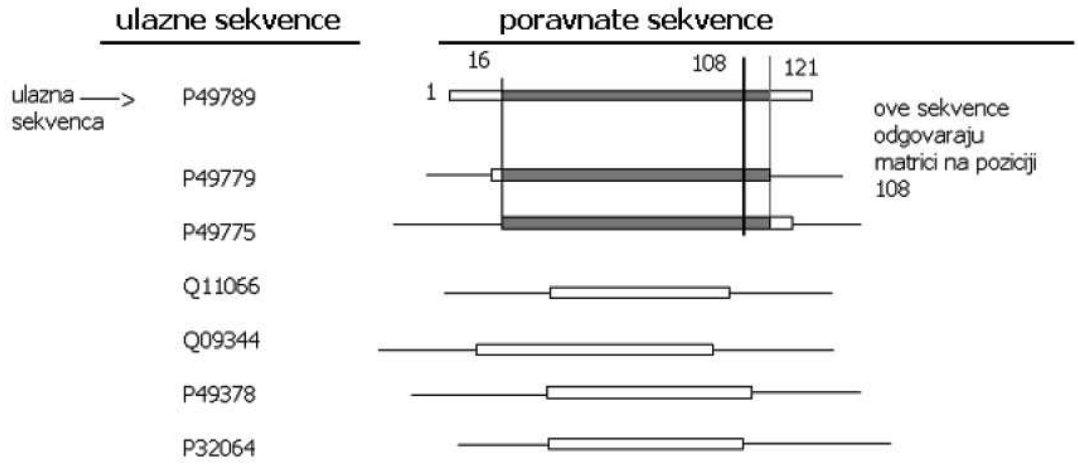
koja teži prema 1. Vrijednosti s_{ij} su elementi supstitucijske matrice:

$$s_{ij} = \left\lfloor \ln\left(q_{ij} / P_i P_j \right) \right\rfloor \lambda_u \tag{4.5}$$

4.1.5. PSI-BLAST

PSI-BLAST [12] punog naziva *Position Specific Iteration BLAST* inačica je BLAST algoritma u kojemu se profil, odnosno matrica vjerojatnosti pronalaženja svake od 20 aminokiselina na mjestu aminokiseline čiji se profil traži (engl. *Position Specific Scoring Matrix*, PSSM), gradi iz višestrukog sekvencijalnog poravnanja te najviše ocjenjenih lokalnih poravnanja koja se traže u inicijalnom BLAST algoritmu. Visoko konzervirane pozicije dobivaju visoke ocjene, a slabo konzervirane pozicije dobiju ocjenu oko nule. Profil izgrađen u prvoj iteraciji se koristi za drugu iteraciju i tako dalje, sve dok se proces ne izvrši zadani broj iteracija ili ne konvergira. Iterativni postupak poboljšavarezultat i povećava osjetljivost.

PSI-BLAST omogućava pronalazak udaljenih sekvenci odnosno onih manje sličnih. Profil izgrađen u prvoj iteraciji se temelji na sličnim sekvencama ulaznoj sekvenci te služi za sljedeću iteraciju u kojoj se pronalaze udaljene, a slične sekvence. Algoritam se sastoji od sljedećih elemenata. Korištenjem BLAST algoritma, u prvoj se iteraciji izgradi profil koji se zatim uspoređuje s proteinskom bazom podataka odnosno njihovih sekvenci. Početna točka u kreiranju profila jest grupa sekvenci koje su poravnate, a ujedno su i izlazni podatak BLAST algoritma. Taj se rezultat reducira u cilju određivanja vrijednosti profila. Za svaki stupac poravnatih sekvenci, u obzir se uzimaju i susjedni aminokiselinski ostaci.Tako se poravnati redci sekvenci reduciraju tj.uzimaju se samo oni redovi čiji su stupci postavljeni na način da svaki sadrži određeni ostatak ili prazninu, s time da su redovi iste duljine.



Slika 4.6 Za profil se uzimaju samo one sekvence odgovarajuće duljine čiji se stupci podudaraju ovisno o aminokiselinskim ostacima [4]

Sljedeći korak je računanje vrijednosti profila odnosno matrice. U računu se koriste vrijednosti BLOSUM matrice, a izraz po kojemu se dobivaju vrijednosti elemenata matrice profila je:

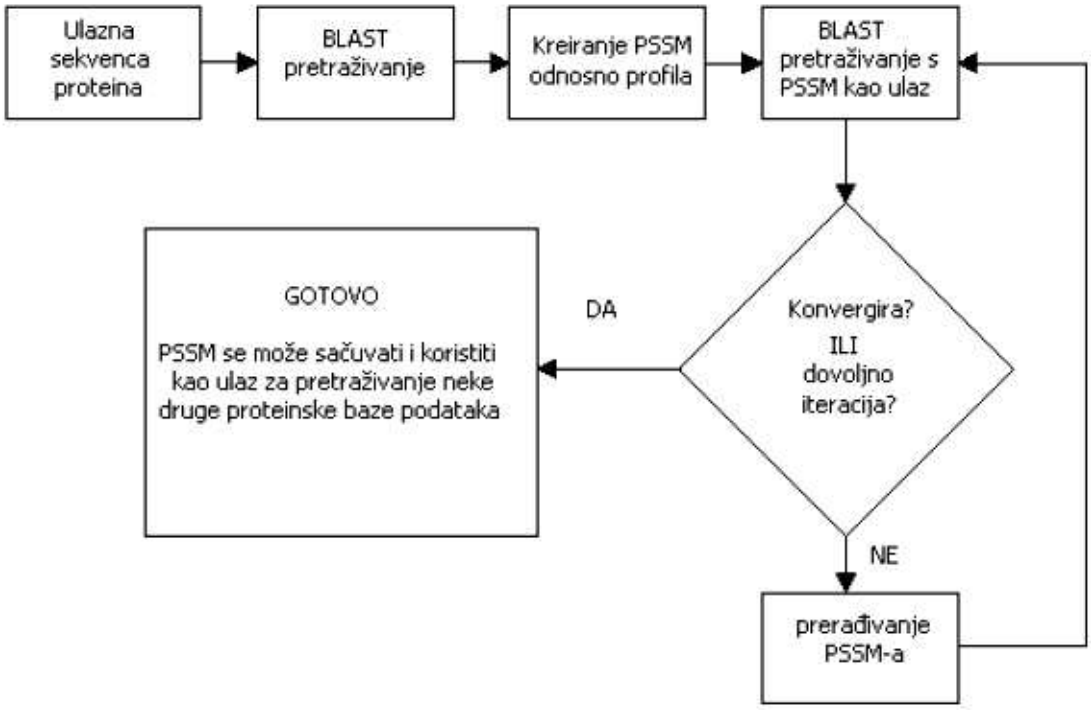
$$Pofile(r,c) = \sum_{d=1}^{20} \sum_{i=1}^{N} weight(i)\delta(A_{ir}, d) \times Comp(residue_d, residue_c) \quad (4.6)$$

gdje je *Profile(r,c)* vrijednost profila za redak *r* i stupac *c*; *r* može imati vrijednosti od 1 do N (duljina seta), *c* i *d* poprimaju vrijednosti od 1 do 20, prezentirajući aminokiseline; *i* je pozicija sekvence u setu; *N* ukupan broj sekvenci; $\delta(A_{ir}, d)$ ima vrijednost 1 ako ostatak na poziciji *r* u sekvenci *i* je aminokiselina *d*, inače je jednak nuli. $omp(residue_d, residue_c)$ je vrijednost u supstitucijskoj tablici. *Weight(i)* se odnosi na težinu sekvence *i*. Težina ekvence se može izračunati aproksimativno iterativnom metodom na sljedeći način:

1. skupiti aminokiseline koje se nalaze na pojedinoj poziciji poravnatog seta sekvenci.
2. inicijalna vrijednost svake sekvence je nula.

3. slučajnim dabirom odabrati sekvencu, birajući na svakoj poziciji (stupcu matrice) jednu aminokiselinu (praznine se tretiraju kao dodatna aminokiselina).
4. izračunati udaljenost slučajne sekvence od ostalih sekvenci.
5. dodati 1 težini najbliže sekvence. Ako je više takvih, njih K, težini svake od tih sekvenci se dodaje vrijednost 1/K.
6. ponoviti korake 3-5 dok težine ne konvergiraju. Kriterij konvergencije nalaže da je relativna promjena težine bliska nuli.
7. normalizacija težine da zbroj težina bude 1.

Na slici 4.7 [4] prikazan je princip rada PSI-BLAST algoritma.



Slika 4.7 Dijagram PSI-BLAST algoritma

4.2. Metoda slučajnih šuma

Slučajne šume[11][13] (engl. *Random Forest*) su u ovome radu odabrane kao metoda regresije zbog slijedećih svojstava:

- velika točnost prepoznavanja,
- relativno je otporna na outliere i šum

Slučajna šuma , u kasnijem tekstu RF, je općeniti naziv za skupinu metoda koje se koriste stablastim klasifikatorima {h(x,Θ_k), k=1,…,} gdje je {Θ_k} skup jednoliko distribuiranih, međusobno potpuno neovisnih vektora, a x ulazni vektorski uzorak. Prilikom treniranja, RF algoritam stvara veliki broj stabala, od kojih se svako trenira na određenom broju uzoraka originalnog trening seta odabranih bootstraping metodom. Za razliku od klasičnih stabala gdje se odabire najbolji atribut, za grananje RF koristi m slučajno odabranih varijabli ($m \ll M$, obično log2 $M+1$) i uzima one koja omogućavaju najbolje grananje. Vrijednost m se unaprijed određuje i konstantna je za cijelu šumu.

Za klasifikaciju svako stablo unutar RF daje glas jednoj od klasa unutar skupa x. Izlaz klasifikatora ovisi o broju glasova stabala svakoj pojedinoj klasi. Trening skup za pojedino stablo stvara se tako da se iz početnog skupa za treniranje, veličine N, uzme N instanci, slučajnim odabirom s ponavljanjem. Iz tako stvorenog skupa za treniranje stabala, vrijednosti koje nisu odabrane koriste se za procjenu pogreške. Ove instance se nazivaju oob instance (engl. *out of bag*) i ima ih oko 38 % ukupnog broja instanci N početnog skupa i koriste se za dobivanje nepristrane procjene greške klasifikacije. Također se koriste i za procjenu važnosti pojedinih varijabli ulaznih instanci.

Kod slučajne šume nema potrebe za krosvalidacijom ili korištenjem posebnog seta za testiranje kako bi se dobila nepristrana procjena greške.

Svako stablo se stvara tako da se koristi podskup iz početnih podataka za učenje koji se naziva bootstrap podskup. Svaki uzorak izostavljen pri stvaranju k-tog stabla, oob instance, treba pustiti niz k-to stablo da bi se dobila klasifikacija.

Nakon završene obrade definiramo *j* kao klasu koja je dobivala najviše glasova u slučaju kada je *n* bila *oob* instanca. Omjer broja izlaza kada *j* nije bila jednaka pravoj klasi instance *n* s obzirom na sve instance naziva se procjena pogreške *oob*-a.

Za svako se stablo u šumi uzimaju oob instance te zbroje glasovi koji su ispravno doneseni s obzirom na klasu. U sljedećem se koraku slučajno permutiraju vrijednosti varijable *m* u oob instancama, te ih se ponovo propusti kroz stablo. Nakon toga se oduzima broj glasova za ispravnu klasu oob instanci s permutiranom *m* varijablom od broja glasova za ispravnu klasu neupotrijebljenih oob instanci.

Srednja vrijednost dobivene razlike u svim tablima unutar šume naziva se važnost varijable *m*. Ukoliko su vrijednosti ove važnosti nezavisne od stabla do stabla, njezinim dijeljenjem sa standardnom pogreškom dobiva se z-skor.

Kod regresije algoritam stvara stabla ovisno o slučajnom vektoru Θ, takvom da $h(x,\Theta_k)$ poprimi numeričku vrijednost umjesto oznaku klase. Izlazi svih stabala se zbroje te tvore konačno predviđanje što je zapravo srednja vrijednost predviđanja svakog pojedinog stabla.

U ovom radu stvarna vrijednost relativne površine dostupne otapalu predviđala se regresijskom analizom metode slučajnih šuma. Za izradu i evaluaciju modela koristio se Rattle alat u statističkom paketu R te vlastoručno napisane skripte.

4.2.1. Postupak izgradnje stabala

Postupak izgradnje stabla odlučivanja je rekurzivan proces. Stablo se grana od početnog čvora po različitim značajkama i njihovim vrijednostima. Grananje je završeno u trenutku kada se određeni skup vrijednosti značajki poveže s klasom kojoj pripada.

Ulazni skup je vektor od *N* značajki, a izlaz je klasa *M* kojoj taj skup pripada. Prilikom izgradnje stabla koristi se skup od *n* uzoraka trening skupa, čiji je razred poznat.

Koraci izgradnje stabla odluke su sljedeći:

1. u korijenu stabla je čvor koji sadrži sve uzorke iz trening skupa
2. ako svi uzorci iz skupa promatranog čvora pripadaju istom razredu, vraća se odgovarajuća klasa te se grananje završava
3. inače, ako su sve ulazne vrijednosti jednake, vraća se klasa koje ima najviše te se grananje završava
4. inače se skup uzoraka u promatranom čvoru dijeli na podskupove određene vrijednostima značajke N_i . N_i je pri tome značajka koja nosi najveću količinu informacije.
5. razvija se k novih čvorova iz promatranog čvora gdje je k broj različitih vrijednosti značajke N_i koje se javljaju u čvoru roditelju. Svaki čvor dijete poprima jednu od k vrijednosti i nasljeđuje one uzorke iz roditeljskog skupa koji imaju odgovarajuću vrijednost značajke N_i .
6. koraci 2-5 se rekurzivno ponavljaju za svaki čvor

4.3. Mjerenje uspješnosti predviđanja

4.3.1. Pearsonov koeficijent korelacije

Mjera koja će se koristiti za mjerenje uspješnosti predviđanja je koeficijent korelacije. Koeficijent korelacije izražava mjeru povezanosti između dvije varijable neovisno o konkretnim jedinicama mjere u kojima su izražene vrijednosti varijabli. Varijable između kojih će se mjeriti povezanost su opažene relativne ASA vrijednosti aminokiselinskih ostataka dobivene PSAIA alatom te predviđene relativne ASA vrijednosti tih ostataka.

Korelacija će se izražavati preko Pearsonovog koeficijenta korelacije, čije se vrijednosti kreću od +1 (savršeno pozitivna korelacija) do -1 (savršeno negativna korelacija). Predznak koeficijenta nas upućuje na smjer korelacije – da li je pozitivna ili negativna, ali nas ne upućuje na snagu korelacije.

Pearsonov koeficijent korelacije [4] bazira se na usporedbi stvarnog utjecaja promatranih varijabli jedne na drugu u odnosu na maksimalni mogući utjecaj dviju varijabli. Označava se malim latiničkim slovom r. Za izračun koeficijenta korelacije potrebna su tri različite sume kvadrata (SS): suma kvadrata varijable X, suma kvadrata varijable Y i suma umnožaka varijabli X i Y.

Suma kvadrata varijable X jednaka je sumi kvadrata odstupanja vrijednosti varijable X od njezine prosječne vrijednosti:

$$SS_{XX} = \sum_{i=1}^{n} \left(X_i - \overline{X} \right)^2 \tag{4.7}$$

dok je prosječna vrijednost varijable X dana relacijom:

$$\overline{X} = \frac{1}{n} X_i \tag{4.8}$$

Nakon što definiramo sumu umnožaka varijabli X i Y kao sumu umnožaka odstupanja vrijednosti varijabli X i Y od njihovih prosjeka relacijom:

$$SS_{XY} = \sum \left(X_i - \overline{X} \right) \left(Y_i - \overline{Y} \right) \tag{4.9}$$

možemo definirati koeficijent korelacije:

$$r = \frac{SS_{XY}}{\sqrt{SS_{XX} \cdot SS_{YY}}} \tag{4.10}$$

5. Rezultati

Predviđanje relativne površine dostupne otapalu vršilo se regresijskom analizom metode slučajnih šuma. Koristila se šuma od 100 stabala, te 13 svojstava pri grananju. Ulazni podaci bili su pomični prozori duljine 9 do 21 aminokiselinskih ostataka te profili slijeda za svaki ostatak unutar prozora.

Kako se koristila 10.-struka krosvalidacija, svaki od korištenih skupova podijeljen je na deset podskupova. Svaki podskup sadrži jednak broj proteina iz orginalnog skupa dobivenih slučajnim odabirom. Radilo se deset predviđanja za svaki skup, na način da je svaki od podskupova u jednom krugu bio testni, dok je preostalih devet spojeno i korišteno za učenje šume. Prilikom učenja koristi se ukupna rASA dobivena PSAIA alatom.

Konačan rezultat u obliku koeficijenta korelacije dobiven je kao srednja vrijednost koeficijenata korelacije svih deset predviđanja na testnim skupovima pojedinog skupa. Upravo ti rezultati će i biti prikazani.

5.1. Ovisnost uspješnosti predviđanja o duljini prozora

U tablici 5.1 nalazi se ovisnost koeficijenta korelacije o duljini pomičnog prozora za svaki od korištenih skupova, također su istaknute maksimalne vrijednosti koeficijenta korelacije za svaki od skupova.

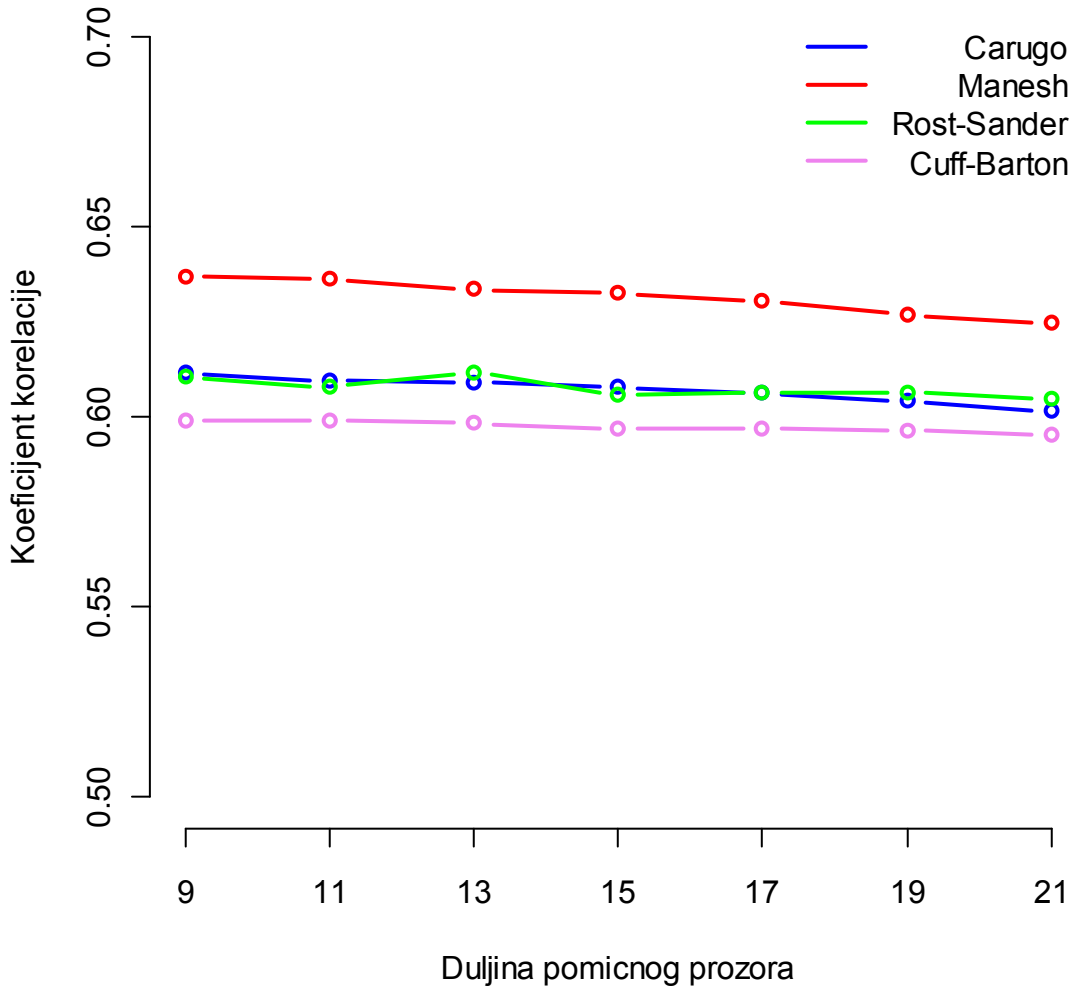
Tako je za skupove Manesh i Carugo korelacija najveća za prozor duljine 9, za skup Rost-Sander za prozor duljine 13 dok je kod skupa Cuff-Barton korelacija najveća u slučaju pomičnog prozora duljine 11 no tek je nešto veća od one za prozor duljine 9.

Tablica 5.1 Ovisnost koeficijenta korelacije o duljini prozora za korištene skupove.

	Duljina pomičnog prozora						
Korišteni skup	9	11	13	15	17	19	21
Carugo	**0.61175**	0.60949	0.60923	0.60797	0.60644	0.60434	0.60170
Manesh	**0.63697**	0.63648	0.63369	0.63259	0.63074	0.62692	0.62494
Rost-Sander	0.61078	0.60790	**0.61179**	0.60594	0.60659	0.60669	0.60489
Cuff-Barton	0.59906	**0.59918**	0.59852	0.59700	0.59714	0.59650	0.59525

Ponašanje koeficijenta korelacije odnosno uspješnosti predviđanja u ovisnosti o duljini pomičnog prozora može se vidjeti i na slici 5.1. Varijacije vrijednosti koeficijenta korelacije ovisno o duljini prozora nisu velike, no kod tri od četiri korištena skupa može se uočiti blagi pad koeficijenta korelacije za svako povećanje duljine prozora, što ne vrijedi za skup Rost-Sander.

Također se vidi da su najbolji rezultati postignuti za skup Manesh te najlošiji za skup Cuff-Barton.



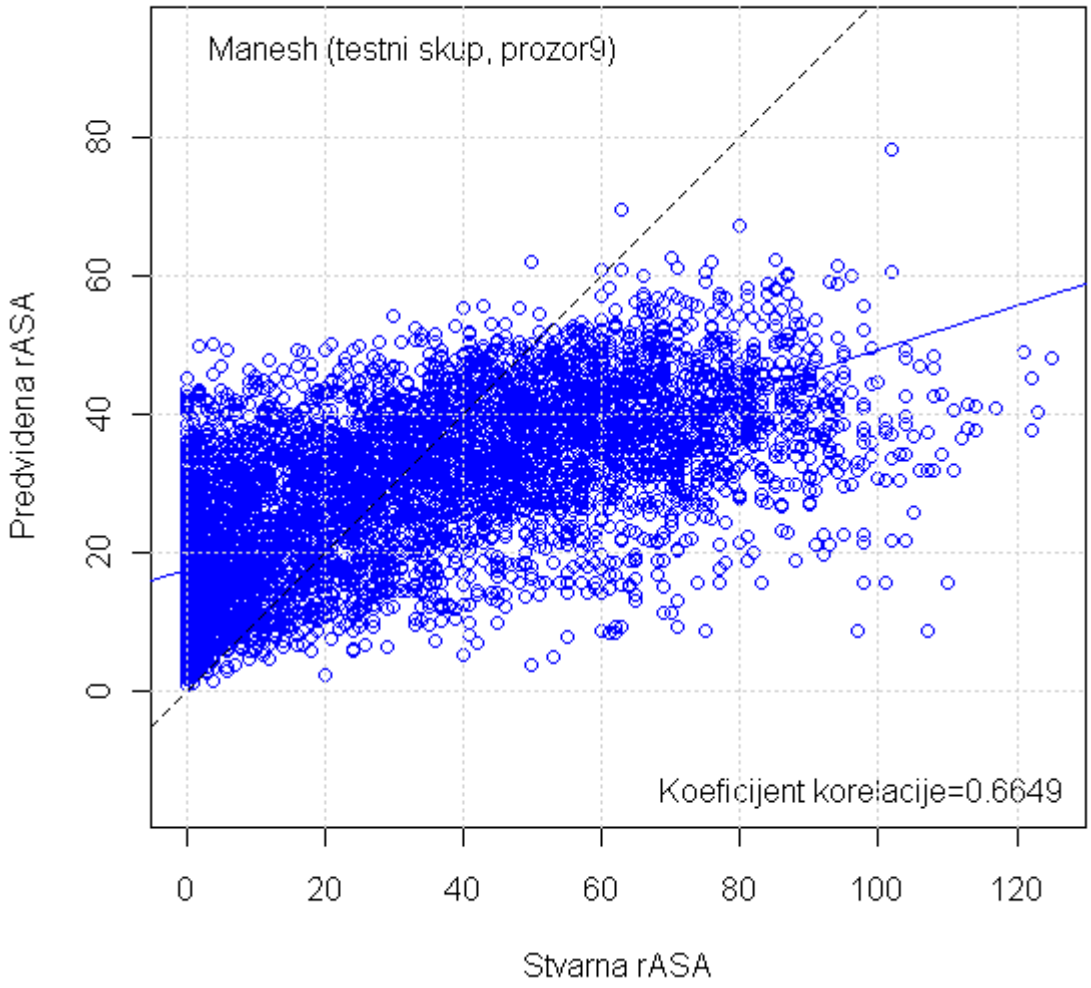
Slika 5.1 Ovisnost koeficijenta korelacije o duljini prozora za različite skupove

5.2. Prikaz rezultata

Budući da se predviđanje površine dostupne otapalu vršilo regresijom a ne klasifikacijom, osim koeficijenta korelacije može se još jedino prikazati odnos predviđenih rASA vrijednosti i opaženih odnosno stvarnih rASA vrijednosti aminokiselinskih ostataka za pojedine skupove.

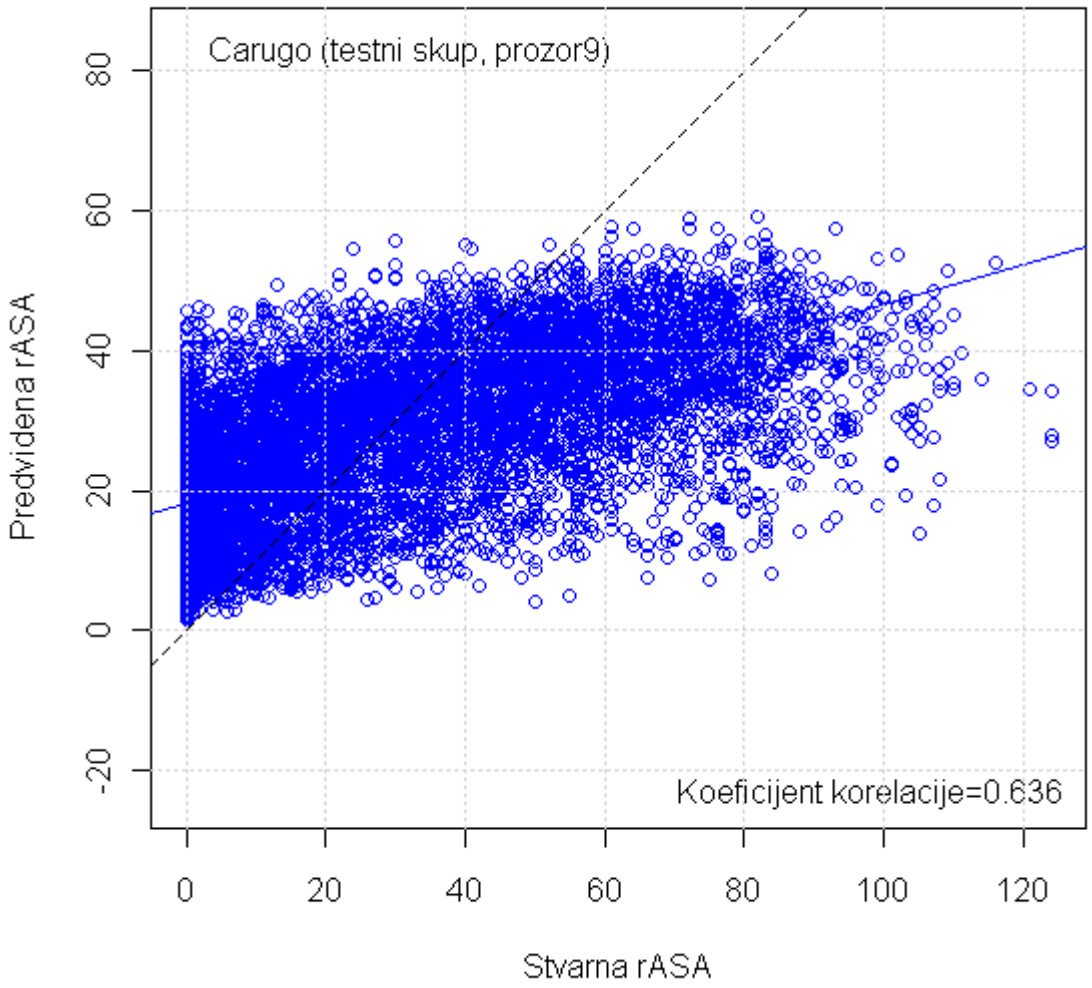
Prikazivat će se samo odnosi za duljine pomičnih prozora koji su dali najbolje ukupne rezultate korelacije za pojedini skup.

Na slici 5.2 nalazi se odnos predviđenih rASA vrijednosti i stvarnih rASA vrijednosti za jedan od testnih skupova Manesh skupa i pomični prozor duljine 9.



Slika 5.2 Odnos predviđenih naspram stvarnih rASA vrijednosti aminokiselinskih ostataka za jedan od testnih skupova skupa Manesh, te pomični prozor duljine 9

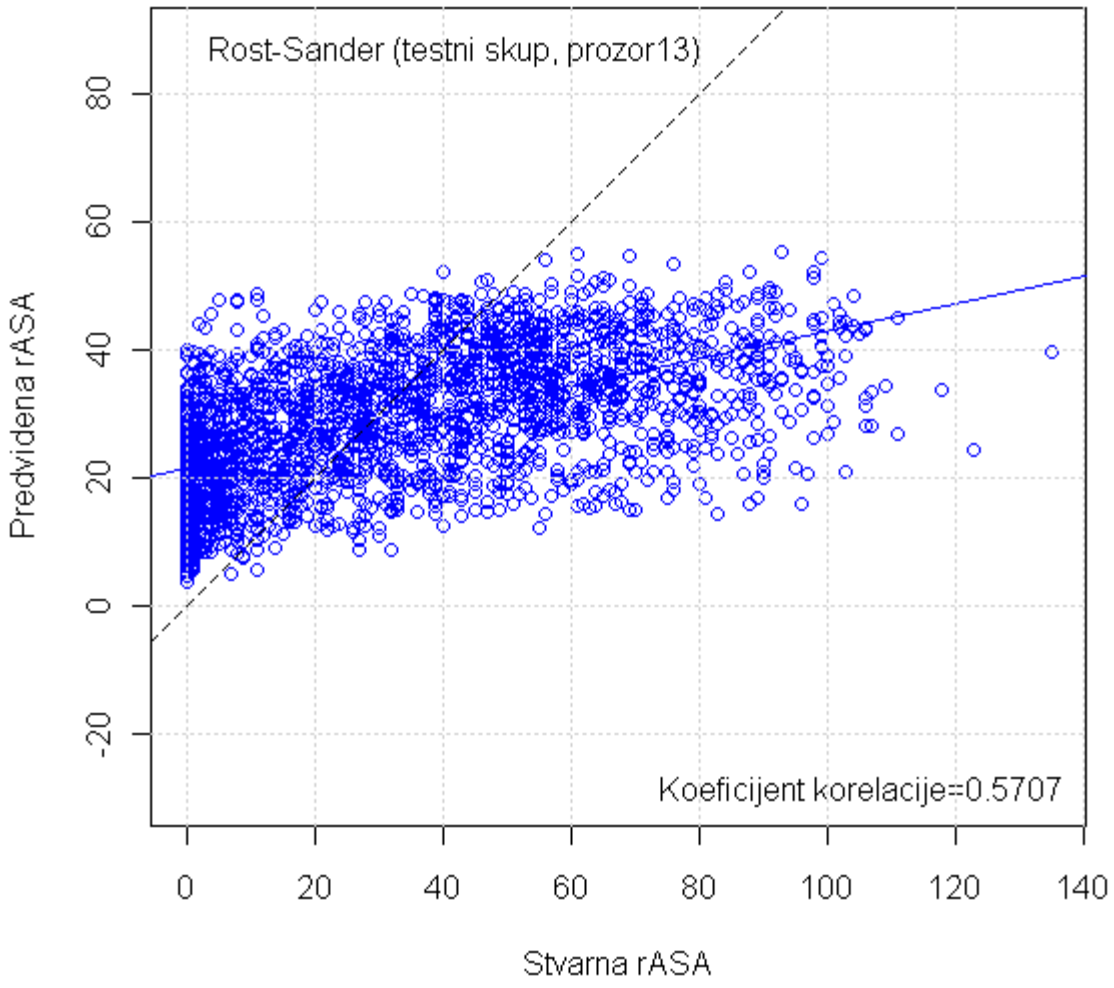
Za Carugo skup gledat ćemo također prozor duljine 9, a odnos predviđenih i stvarnih rASA vrijednosti jednog od testnih skupova nalazi se na slici 5.3 .



Slika 5.3 Odnos predviđenih naspram stvarnih rASA vrijednosti aminokiselinskih ostataka za jedan od testnih skupova skupa Carugo, te za pomični prozor duljine 9

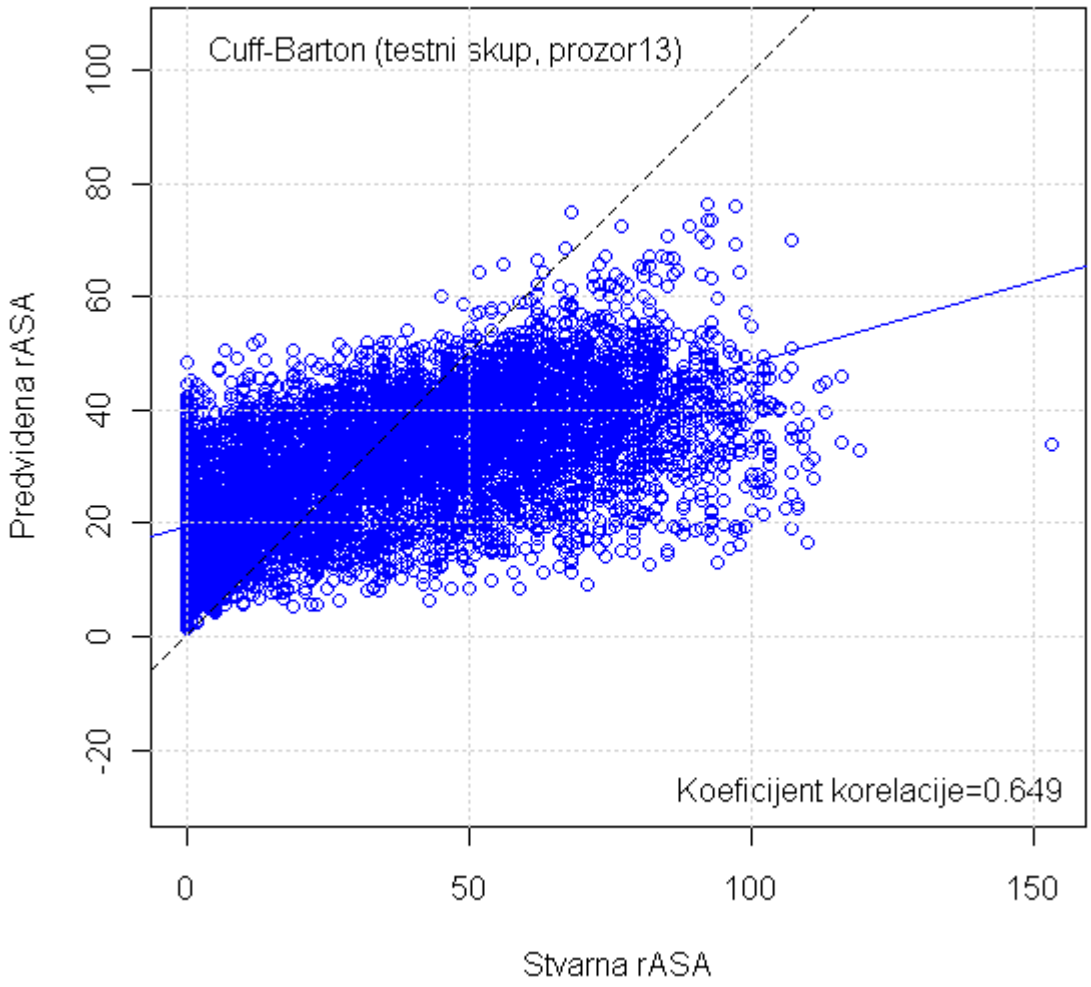
Na slici 5.4 imamo prikaz odnosa predviđene i stvarne rASA vrijednosti za jedan od testnih skupova skupa Rost-Sander pri pomičnom prozoru duljine 13.

Za skup Rost-Sander napravljeno je i predviđanje prilikom kojeg se koristila slučajna šuma od 300 stabala umjesto 100. Opet su najbolji rezultati postignuti za prozor duljine 13 te koeficijent korelacije iznosi 0.62231, što je oko 1% bolje. Očekivano šuma s većim brojem stabala daje i bolje rezultate, no za veće skupove zbog problema alociranja memorije nije bilo moguće graditi šume veće od 100 stabala. Ipak na osnovu rezultata može se zaključiti da i šuma od 100 stabala daje prilično dobre rezultate.



Slika 5.4 Odnos predviđenih naspram stvarnih rASA vrijednosti aminokiselinskih ostataka za jedan od testnih skupova skupa Rost-Sander, te za pomični prozor duljine 13

Odnos predviđene rASA vrijednosti naspram stvarne rASA vrijednosti, za pomični prozor duljine 11 ostataka, jednog od testnih skupova skupa Cuff-Barton nalazi se na slici 5.5.



Slika 5.5 Odnos predviđenih naspram stvarnih rASA vrijednosti aminokiselinskih ostataka za jedan od testnih skupova skupa Rost-Sander, te za pomični prozor duljine 13

5.3. Prikaz rezultata drugih autora

U nastavku će se prikazati rezultati nekih od metoda koje su također predviđale stvarnu ASA vrijednost i pritom su autori koristili barem jedan od skupova korišten u ovom radu, na taj način će biti moguća usporedba.

Prva metoda predstavljena od autora S.Ahmada i suradnika [15], koristi standardnu neuronsku mrežu i pomični prozor duljine 7 ostataka. Korišteni su skupovi Carugo, Barton i Rost-Sander. Rađena je 3-struka krosvalidacija, no za razliku od načina u ovom radu oni jedan podskup koriste za treniranje, drugi za testiranje i treći za validaciju.

Za vrijeme učenja na skupu za treniranje, konstantno se promatrala pogreška skupa za testiranje. Težine su sačuvane pri svakom smanjenju greške predviđanja. Na taj način konačne trenirane težine predstavljaju mrežu u trenutku kada je imala minimalnu srednju apsolutnu pogrešku.

Treći skup, onaj za validaciju se drži izvan procesa treniranja da bi se osigurala točnost predviđanja za te podatke koja predstavlja predvidljivost mreže i slobodna je od pristranosti uzrokovane skupom za treniranje/testiranje. Svaki od tri skupa će se koristiti za treniranje/testiranje/validaciju, te postoji šest mogućih kombinacija. Konačni rezultat je srednja vrijednost grešaka odgovarajućih skupova dobivenih u šest krugova tereniranja i predviđanja.

U tablici 5.2 nalaze se vrijednosti korelacije koje su dobili za validacijski skup, najbolji rezultati, iznosa korelacije 0.4870, dobiveni su za skup Carugo.

Tablica 5.2 Rezultati predviđanja metode [15] koristeći neuronske mreže

Korišteni skup	Rost-Sander	Carugo	Cuff-Barton
Korelacija	0.4718	0.4870	0.4800

Njihovi rezultati su dosta lošiji od onih postignutih u ovom radu, ipak direktna usporedba rezultata nije moguća budući da oni koriste pomični prozor duljine 7 koji u ovom radu nije uzet u obzir. No budući da je u radu pokazano da su rezultati bolji za prozore manje duljine, u našem slučaju za prozore duljine 9 do 13, za pretpostaviti je da bi i za prozor duljine 7 rezultati bili jednako dobri kao oni za prozor duljine 9 ili neznatno manji što je i dalje mnogo bolji rezultat.

Da se otprilike vidi o kolikoj bi se razlici u uspješnosti radilo evo naših rezultata dobivenih za iste skupove, no za prozor duljine 9: Rost-Sander: 0.61078, Carugo: 0.61175, te Cuff-Barton: 0.59906.

Razlog ovako boljem rezultatu leži u tome što autori ove metode kao ulazne podatke koriste samo slijed aminokiselinskih ostataka bez profila slijeda. Poznato je da profili slijeda značajno doprinose poboljšanju rezultata, stoga ne čudi da se zadnjih par godina intezivno koriste u predviđanjima, što dostupnosti otapalu, što mjestu proteinskih interakcija i drugog.

Druga metoda [16] autora Nguyena i Rajapaksea za predviđanje koristi SVR-regresiju potpornim vektorima (engl. *Support Vector Regression*) u dva stupnja u kombinaciji s evolucijskom informacijom odnosno profilima slijeda.

Odabir trening i test skupova uzeli su kao i u prethodnoj metodi radi mogućnosti objektivne usporedbe. Dakle primjenjena je 3-struka krosvalidacija te se validacijski skup držao izvan procesa treniranja

Cilj korištenja regresije potpornim vektorima u dvije faze jest uzeti u obzir utjecaj ASA vrijednosti susjednih ostataka na ASA vrijednost ostatka za kojeg vršimo predviđanje. Za prvu fazu koristio se pomični prozor veličine 13 aminokiselinskih ostataka dok se za drugu fazu koristio prozor veličine 21. Ulaz u drugi SVR prediktor su predviđanja iz prve faze.

Koeficijenti korelacije dobiveni ovom metodom za različite skupove nalaze se u tablici 5.3 . Najbolje rezultate dobili su za Manesh skup, a najlošije za Cuff-Barton skup, što odgovara i rezultatima ovog rada. Njihovi rezultati su nešto bolji, budući da su naši za prozor duljine 13, te skupove Manesh, Carugo i Cuff-Barton redom 0.63369, 0.60923, 0.59852.

Tablica 5.3 Rezultati predviđanja metode [16] koristeći SVR u 2 faze i profile slijeda

Korišteni skup	Manesh	Carugo	Cuff-Barton
Korelacija	0.6800	0.6700	0.6600

No opet je teško napraviti usporedbu budući da oni koriste SVR u dva stupnja, te svaki stupanj koristi različitu duljinu prozora. Veća uspješnost se vjerojatno krije u tome što se osim profila slijeda koriste i ASA vrijednosti susjednih ostataka, budući da ostatak djeluje ne samo svojom interakcijom već i svojom dostupnošću otapalu.

Treća metoda autora Wanga i suradnika [17] koristi višestruku linearnu regresiju za predviđanje stvarne ASA vrijednosti. Rezultati predviđanja su evaluirani 5-strukom krosvalidacijom(jedan testni, ostali za treniranje), a korišten je pomični prozor veličine 13 ostataka.

Predviđanje je vršeno samo za sekvencu aminokiseline, zatim koristeći profile slijeda poboljšanje s dva svojstva te koristeći profile slijeda uz dodatni 21-dimenzionalni vektor (20 elemenata za sastav aminokiseline i jedan za dužinu sekvence).

U tablici 5.4 nalaze se rezultati predviđanja dobiveni za Cuff-Barton skup i različite ulazne podatake. Korištenjem samo informaciji o slijedu aminokiselina koeficijent korelacije je najlošiji i iznosi 0.5200. Korištenje profila slijeda (PSSM) sa dva dodatna svojstva(insercijom/delecijom i entropijom) poboljšava korelaciju na 0.6300. Dok korištenje PSSM uz dodatni 21-dimenzionalni vektor (20 elemenata za sastav aminokiseline i jedan za dužinu sekvence) daje najbolje rezultate.

Cuff-Barton skup ima najlošije dobivene rezultate u ovom radu, za prozor duljine 13 ostataka, koeficijent korelacije iznosi 0.59852 što je bolje od njihova rezultata kad se koristi samo sekvenca aminokiseline. Za preostale slučajeve kad koriste poboljšane profile slijeda njihovi rezultati su bolji. No da je bilo mogućnosti koristiti šumu s većim brojem stabala od 100, rezultati bi možda bili jednaki ili čak bolji od njihovih.

Tablica 5.4 Rezultati predviđanja metode [17] koristeći višestruku linearnu regresiju

Korišteni skup	samo sekvenca	PSSM+2svojstva	PSSM+kompozicija+duljina slijeda
Cuff-Barton	0.5200	0.6300	0.6400

6. Zaključak

Zadatak ovog rada bio je istražiti primjenu RF regresije za predviđanje površine dostupne otapalu na osnovu podataka iz slijeda aminokiselinskih ostataka. Kao ulazni podaci koristili su se pomični prozori duljine 9 do 21 aminokiselinskih ostataka te profili slijeda za svaki ostatak unutar prozora. Skupovi za učenje i testiranje dobiveni su 10-strukom krosvalidacijom iz orginalnih skupova.

Pokazano je da su rezultati bolji za manje duljine prozora, one od 9 do 13, nego li za veće. Tako pomični prozor duljine 9 daje najbolje rezultate za dva korištena skupa (Manesh i Carugo) s tim da je i za preostala dva skupa drugi najbolji prozor. U slučaju Rost-Sander skupa najuspješniji je prozora duljine 13 dok je to prozor duljine 11 u slučaju Cuff-Barton skupa.

Najbolji rezultati predviđanja postignuti su za skup Manesh gdje je za prozor duljine 9 dobiven koeficijent korelacije 0.63697 dok su najlošiji dobiveni za skup Cuff-Barton gdje je za prozor duljine 11 dobiven koeficijent korelacije 0.59918.

Za skupove Rost-Sander, Carugo i Cuff-Barton postignuti su bolji rezultati od onih autora S.Ahmada i suradnika [15] za iste skupove, jedan od glavnih razloga je korištenje profila slijeda za razliku od njih koji su koristili samo podatke iz slijeda aminokiselinskih ostataka. Za skup Cuff-Barton te prozor duljine 13 dobiveni su bolji rezultati i od onih autora Wanga i suradnika [17] za slučaj korištenja samo slijeda aminokiselinskih ostataka. Za preostale slučajeve kad koriste poboljšane profile slijeda njihovi rezultati su nešto bolji. No da je bilo mogućnosti koristiti šumu s većim brojem stabala od 100, rezultati bi možda bili jednaki ili čak bolji od njihovih.

7. Literatura

[1] V. Hankonyi, Organska kemija za studente medicine, interna skripta., Medicinski fakultet Sveučilišta u Zagrebu

[2] Albert L. Lehninger, David Lee Nelson, Michael M. Cox, Principles of biochemistry, W. H. Freeman, 2005

[3] http://www.ncbi.nlm.nih.gov/blast/db/fasta.html

[4] T. Puđa, "Predviđanje površine dostupne otapalu iz slijeda aminokiselinskih ostataka", diplomski rad, FER, 2008.

[5] J. Mihel, M. Sikic, S. Tomic, B. Jeren, and K. Vlahovicek, "PSAIA – Protein Structure and Interaction Analyzer", University of Zagreb, 2008.

[6] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The Protein Data Bank," Nucleic Acids Res, vol. 28, pp. 235-42, Jan 1 2000.

[7] B. Rost and C. Sander, "Improved prediction of protein secondary structure by using sequence profiles and neural networks", Proc. Natl. Acad. Sci. USA, vol. 90, pp. 7558-7562, August 1993.

[8] HN. Manesh, M. Sadeghi, S. Arab, AM. Movahedi, "Prediction of protein surface accessibility with information theory", Proteins, vol. 42, pp. 452-459, 2001.

[9] Cuff JA, Barton GJ: Application of multiple sequence alignment profiles to improve protein secondary structure prediction.Proteins vol.40,pp.502-511.,2000

[10] Carugo O: Predicting residue solvent accessibility from protein sequence by considering the sequence environment.Protein Eng, 13(9):607-609.16.,2000

[11] V. Dragosavljević, "Predviđanje mjesta proteinskih interakcija iz profila slijeda aminokiselinskih ostataka", diplomski rad, 2008.

[12] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," Nucleic Acids Res, vol. 25, pp. 3389-402, Sep 1 1997.

[13] Leo Breiman: „Random Forests,“ Machine Learning, 45, 5–32, 2001

[14] S. Henikoff and J. G. Henikoff, "Amino acid substitution matrices from protein blocks", Proc. Natl. Acad. Sci. USA, vol. 89, pp. 10915-19, November 1992

[15] Ahmad S, Gromiha MM, Sarai A: Real value prediction of solvent accessibility from amino acid sequence. Proteins,50(4):629-635.,2003

[16] Nguyen MN, Rajapakse JC: Prediction of protein relative solvent accessibility with a two-stage SVM approach. Proteins ,59(1):30-37.,2005

[17] Wang JY, Lee HM, Ahmad S: Prediction and evolutionary information analysis of protein solvent accessibility using multiplelinear regression. Proteins , 61(3):481-491., 2005

Sažetak

U okviru ovog rada opisana je metoda predviđanja površine dostupne otapalu iz slijeda aminokiselinskih ostataka. Za predviđanje se koristila regresijska analiza metodom slučajnih šuma. Kao ulazni podaci koriste se pomični prozori duljine od 9 do 21 aminokiselinskih ostataka, te profili slijeda za svaki od ostataka unutar prozora. Tako imamo ulazni vektor dimenzija 21 x n, gdje je n broj ostataka unutar prozora. Predviđa se relativna površina dostupna otapalu, a predviđanje se vrši za središnji ostatak prozora. Rezultati se evaluiraju koristeći 10-struku krosvalidaciju, gdje se svaki od korištenih skupova podijeli na 10 podjednakih slučajno odabranih podskupova, jedan se koristi za testiranje a preostalih devet se spoji i služi za treniranje šume. Postupak se ponavlja sve dok svaki podskup ne bude tretiran kao testni.

Najbolji rezultati predviđanja postignuti su za skup Manesh gdje je za prozor duljine 9 dobiven koeficijent korelacije 0.63697. Dok su najlošiji dobiveni za skup Cuff-Barton gdje je za prozor duljine 11 dobiven koeficijent korelacije 0.59918. Također je pokazano da su rezultati bolji za prozore manje duljine nego veće i to u našem slučaju za one duljine 9 do 13.