SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 1153

PREDVIĐANJE MJESTA PROTEINSKIH INTERAKCIJA IZ PROFILA SLIJEDA AMINOKISELINSKIH OSTATAKA

Viktorija Dragosavljević

Zagreb, listopad 2008.

Hvala roditeljima na podršci tokom cijelog studija,
te hvala Mili Šikiću na svevremenskoj pomoći u izradi rada!

Sadržaj

1 Uvod

Udruživanjem računalne znanosti, matematike, informatike, statistike i biokemije, došlo je do nastanka i razvoja relativno nove znanosti, bioinformatike. Cilj bioinformatike jest pronaći odgovore na biološke upite pretežno na molekularnoj razini. Jedan od takvih upita je i pronalazak mjesta proteinskih interakcija te kako takva mjesta prepoznati odnosno predvidjeti. Ovaj problem je vrlo značajan prvenstveno u razvoju novih lijekova, cjepiva, analizi metaboličkih reakcija te u raznim drugim područjima. Broj eksperimentalno pronađenih i definiranih proteinskih kompleksa je relativno malen, stoga računalne metode za predviđanje proteinskih interakcija postaju značajne, prvenstveno radi svoje brzine pošto su eksperimentalne metode obično dugotrajne.

Zadatak ovog rada jest da se na osnovu podataka iz skupa za učenja, koristeći metodu slučajnih šuma, predvide mjesta proteinskih interakcija u testnom skupu podataka.

Za zadani skup podataka potrebno je korištenjem PSAIA alata i vlastito izrađenih programa izvući podatke o aminokiselinskim ostacima koji su u interakciji kao i informacije o njihovoj evolucijskoj očuvanosti (profili slijeda). Aminokiselinski ostaci u interakciji definiraju se kao oni čiji je bilo koji atom udaljen za manje od 6 Angstrema od najbližeg atoma susjednog lanca. Za ulazni vektor atributa koristi se vektor od 9 ostataka u nizu, kao i profili za svaki pojedini ostatak (20 po ostatku) što čini ulazni vektor od 189 atributa. Klasa za svaki pojedini vektor se označava pozitivnom ukoliko je barem središnji ostatak mjesto interakcije. Mjere za procjenu kvalitete rezultata su *preciznost, odziv*, *F-mjera*, *AUC* i *preciznost-odziv* krivulja. Pri radu se koriste alati za strojno učenje: Rattle biblioteke u statičkom paketu *R*, te paralelna implementacija algoritma slučajnih šuma PARF.

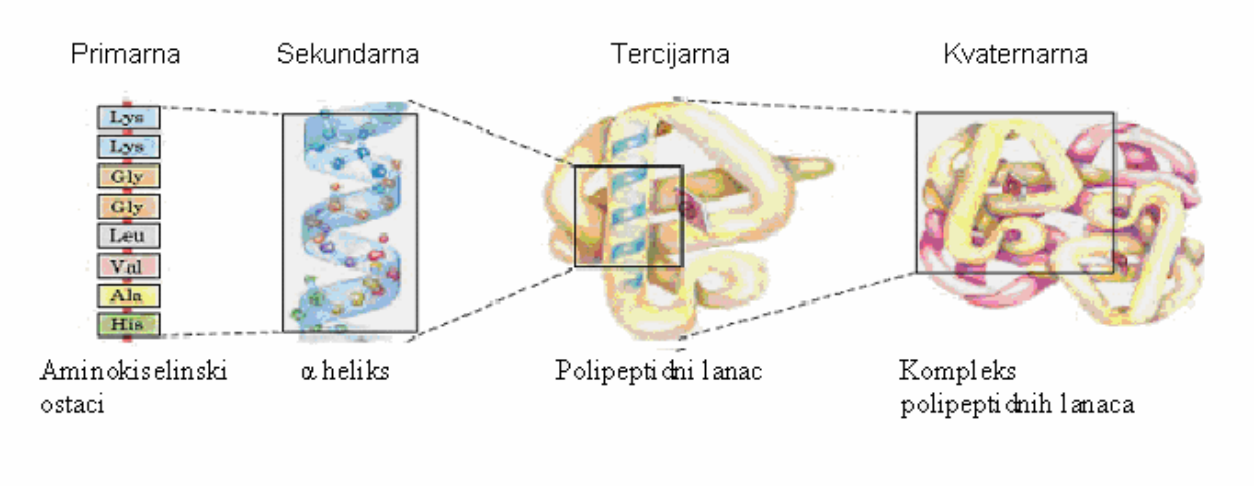
Rad je podijeljen na više poglavlja. U drugome je poglavlju teoretski uvod u područje. Opisana je građa proteina i aminokiselina te su navedene proteinske interakcije. Također su dani dosadašnji rezultati u predviđanju proteinskih interakcija. U trećem su poglavlju opisani podaci, njihovi zapisi i struktura te kako su podaci odabrani. Četvrto poglavlje sadrži opise korištenih metoda. Metoda za izvlačenje profila slijeda aminokiselinskih ostataka *PSI-BLAST*, metoda slučajnih šuma (engl. *Random Forest*) te mjere uspješnosti predviđanja i *OR* metoda. U petom poglavlju su izneseni rezultati predviđanja koristeći informacije iz sekvenci i profila sekvenci kao i slučaj kada se koriste uz navedena svojstva i strukturna svojstva. Zatim slijedi diskusija o dobivenim rezultatima te zaključak rada.

2 Teorijski uvod

2.1 Građa proteina

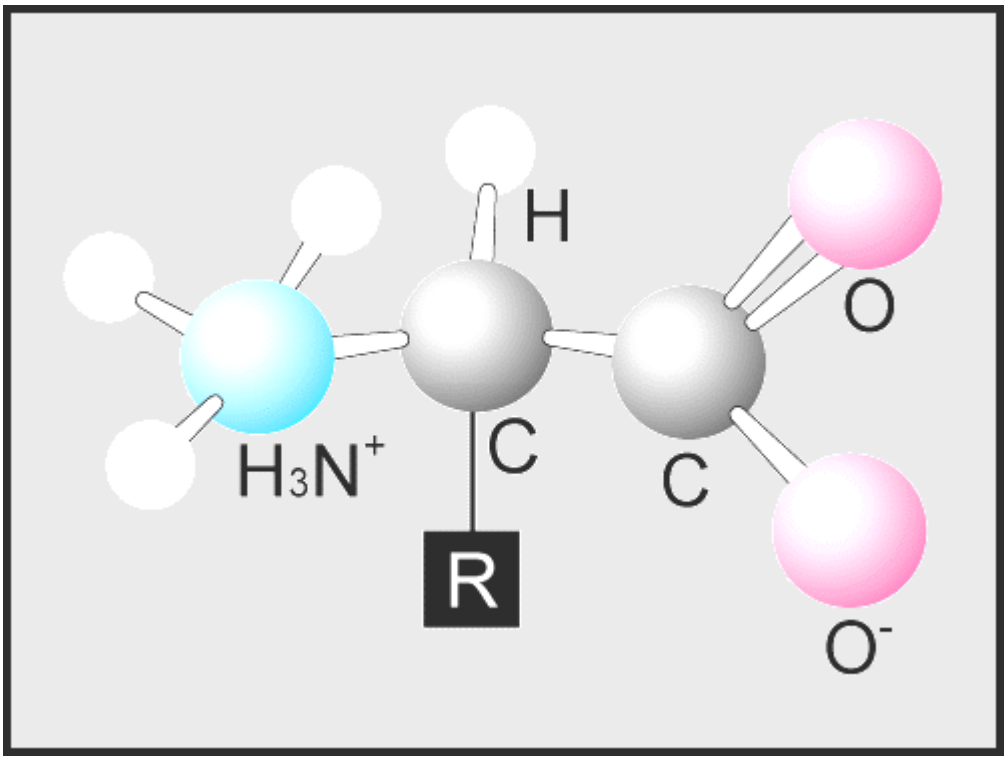
Proteini pripadaju skupini velikih organskih molekula. Sudjeluju u svim esencijalnim procesima unutar stanice. Protein svoju funkciju obavlja ovisno o prostornoj strukturi u koju se smata. Prostorna struktura proteina također određuje s kojim će drugim proteinom ili molekulama protein moći eventualno interagirati.

Protein je građen od niza aminokiselina koje su međusobno povezane peptidnom vezom u lance polipeptide. Veza nastaje između karboksilne i amino skupine aminokiseline pri čemu se izlučuje molekula vode i nastaje aminokiselinski ostatak. Funkcija proteina ovisi o aminokiselinskim ostacima koji čine lanac proteina pošto aminokiselinski ostaci tj sekvenca ostataka, određuju prostornu strukturu, a samim time i funkciju proteina. Prostornih struktura proteina ima nekoliko: primarna, sekundarna, tercijarna i kvaternarna struktura.



Slika 2.1 Strukture proteina

U prirodi postoji svega 20 različitih standardnih aminokiselina od kojih su građeni svi proteini. Razlikuju se po bočnim ograncima dok su im ostale skupine iste: karboksilna skupina i amino skupina vezane na isti atom ugljika. Slika 2.2 prikazuje općenitu strukturu aminokiseline.



Slika 2.2 Općenita struktura aminokiselina. Bočni ogranak je označen slovom R

Proteine se promatra kao niz 20 standardnih aminokiselina odnosno kao sekvencu koju čini 20 različitih slova engleske abecede. Naime, svakoj od 20 aminokiselina je pridruženo jedno slovo abecede osim slova B, J, O, U, X i Z. Primjer jedne takve sekvence je ...TGHARGGARTTVGRRDV...

Tablica 2.1 Popis svih 20 aminokiselina

alanin	ALA	A
arginin	ARG	R
asparaginska kiselina	ASP	D
asparagin	ASN	N
cistein	CYS	C
glutaminska kiselina	GLU	E
glutamin	GLN	Q
glicin	GLY	G
histidin	HIS	H

izoleucin	ILE	I
leucin	LEU	L
lizin	LYS	K
metionin	MET	M
fenilalanin	PHE	F
prolin	PRO	P
serin	SER	S
treonin	THR	T
triptofan	TRP	W
tirozin	TYR	Y
valin	VAL	V

2.2 Proteinske interakcije

Protein se u svom prirodnom okruženju nalazi u otapalu. Najčešće se radi o vodi i kao takvog ga se promatra. Voda svojim svojstvima, kao što je polarnost, utječe na veze među proteinima. Tipična veza između dijelova proteinskih molekula je vodikova veza koju između ostalog tvori molekula vode. Vodikovu vezu stvaraju atomi kisika i vodika koji se nalaze i u strukturama aminokiselina proteina. One su specijalni oblik polarnih interakcija pri kojima dva elektro-negativna atoma međusobno dijele jedan elektro-pozitivan atom vodika. Za razliku od tipičnih elektrostatskih veza, vodikove su veze usmjerene te su vrlo jake kada su sva tri atoma, koja sudjeluju u interakciji, na istome pravcu. Uz vodikove veze postoje nešto slabije veze između proteinskih molekula, Van der Waalsove. Ovdje se radi o privlačenju dva dipola koji oko sebe stvaraju elektronske oblake. Nastaju pri fluktuacijama elektronskog oblaka nepolarnog atoma stvarajući dipol. Takav će dipol inducirati stvaranje suprotno polariziranog dipola u susjednom atomu i nastati će slaba privlačnost među tim atomima. Uz dvije navedene vrste veza postoje i hidrofobne interakcije. Hidrofobne interakcije su odgovorne za međusobno vezanje proteina kao i za njihovo smatanje. Takve se interakcije pojavljuju u slučaju kontakta nepolarnih i polarnih molekula. Polarne molekule kao što je voda, ne mogu interagirati s nepolarnima. Proteini su građeni i od polarnih i nepolarnih molekula tako da se u dodiru s vodom pri formiranju strukture

maksimizira interakcija hidrofilnih dijelova s vodom te minimizira kontakt hidrofobnih dijelova proteina s vodom.

Osim nabrojenih veza postoje još ionske veze i kovalentne veze. Ionske veze nastaju između suprotno nabijenih atoma zbog elektrostatskih privlačnih sila. Ako se ovakva veza nalazi izvan vode, ona može biti prilično jaka, međutim, u vodi, zbog polarnih molekula vode nakupljenih oko nabijenih iona, međusobna privlačnost pada.

Svaka veza za sebe je relativno slaba, međutim njihovo skupno djelovanje stabilizira kompleks proteina smanjivanjem slobodne energije kompleksa.

2.2.1 Definicija mjesta proteinske interakcije

Postoji više definicija mjesta proteinske interakcije. Jedna od njih, koja se ujedno koristi u ovome radu, zasniva se na promatranju sekvence aminokiselinskih ostataka [1] i njihove evolucijske očuvanosti, profil slijeda [2]. Mjesto interakcije se definira kao pomični prozor od *n* uzastopnih ostataka pri čemu središnji kao i još barem *m* ostataka su u kontaktu sa susjednim lancem [1]. Vrijednosti parametra *n* se obično kreću od 9 do 13 ostataka u prozoru, a *m* između 1 i 6. U ovome se radu uzimalo 9 ostataka u prozoru, središnji kao mjesto kontakta te da su barem 4 ostatka u kontaktu sa susjednim lancem, a da pritom nisu udaljeni od središnjeg ostatka više od 3 ostataka. Za svaki se ostatak promatrao ujedno i njegov profil.

Aminokiselinski ostaci u kontaktu su oni čiji je bilo koji atom udaljen za manje od 6 Å (Angstrema) od najbližeg atoma susjednog proteinskog lanca.

Tako vektor atributa čine prozor od 9 ostataka i profil za svaki ostatak u prozoru te se definirani vektor atributa proglašava pozitivnim ukoliko je barem središnji ostatak mjesto kontakta.

Profil nekog aminokiselinskog ostatka jest vjerojatnost pronalaženja svake od 20 standardnih aminokiselina na mjestu aminokiselinskog ostatka čiji se profil ispituje. Profil se može promatrati kao mjera evolucijske očuvanosti aminokiselinskog ostatka.

2.3 Osvrt na dosadašnje rezultate

U dosadašnjim radovima autori su za predviđanje mjesta interakcije koristili ne samo informaciju iz sekvence već i strukturnu informaciju te njihovu kombinaciju. Uspješnost najboljih dosadašnjih rezultata predviđanja na temelju samo informacije iz sekvence je oko 70% preciznosti za 40% odziva [3]. Najuspješniji rezultat predviđanja je slučaj kada se koriste

informacije o prostornoj strukturi, a tada je preciznost preko 80% pri čemu se traži samo jedno mjesto interakcije na proteinu čime je otežano definirati odziv.

Ofran i Rost [1] u svome radu definiraju mjesto interakcije kao prozor od 9 ostataka pri čemu je prozor mjesto interakcije ukoliko je središnji ostatak u kontaktu s aminokiselinskim ostatkom drugog proteina te su barem još 4 ostatka u kontaktu, a da od središnjeg ostatka nisu udaljeni više od 3 ostataka. U kontaktu je onaj ostatak čiji je bilo koji teški atom udaljen 6 Å ili manje od teškog atoma proteina partnera. Predviđanje je rađeno koristeći neuronske mreže. Postigli su preciznost od 70% i odziv 0,5%.

Koike i drugi [4] koristili su metodu potpornih vektora (SVM) te prozor od 11 ostataka. Za preciznost od 40,2% dobili su odziv od 39,6% za ostatke koji se nalaze na površini. Za mjesto interakcije su definirali ostatak na površini koji se nalazi u središtu prozora od 11 ostataka i u kontaktu je s ostatkom proteina partnera. Kontakt čine oni ostaci čiji su teški atomi udaljeni manje od 5 Å. Profil sekvenci ostataka dobivene PSI-BLAST metodom su koristili kao svojstvo.

Koristeći SVM pri klasifikaciji Reš i ostali [2] uzeli su kombinaciju evolucijske informacije i sekvencu ostataka. Za mjesto kontakta definirali su one ostatke na površini čija se relativna ASA (površina dostupna otapalu) promijeni nakon interakcije. Postigli su preciznost od 26% te odziv od 59%. Pri tome su koristili prozor od 9 ostataka te prozore s N mjesta kontakta prozivali mjestom interakcije. Za $N = 6$ postigli su preciznost od 27,4% te odziv od 57,5%.

U novijem radu [5] Ofran i Rost postižu bolje rezultate tako da su predviđanju na osnovu sekvence dodali evolucijske profile, ASA-u i sekundarnu strukturu. U prvom su koraku najprije predvidjeli ASA-u i sekundarnu strukturu iz sekvence te su u drugome krugu dodali te informacije sekvenci. Kao klasifikator su kao i u ranijem radu koristili neuronske mreže. Postigli su preciznost između 60% i 70% pri odzivu iznad 10%. Mjestom interakcije su definirali onaj prozor čiji je središnji ostatak u kontaktu, te da na udaljenosti 5 ostataka od središnjeg, postoji još barem 6 ostataka koji su u kontaktu sa susjednim lancem.

Yan i ostali [6] u svome su radu koristili klasifikaciju u dva koraka koristeći informaciju da su ostaci koji sudjeluju u interakciji grupirani. U prvom su koraku koristeći SVM predvidjeli potencijalna mjesta kontakta, a u drugom su koraku koristili Bayesovu mrežu. Kao ulazna svojstva u Bayesovu mrežu koristili su 8 susjednih ostataka. Promatrali su samo ostatke na

površini proteina i među njima tražili one koji su u interakciji. Nešto su drugačije definirali mjesto kontakta u odnosu na ostale autore. Za mjesto kontakta su uzeli one ostatke kojima se nakon interakcije ASA promijeni za barem 1 Å. U prvome su koraku za preciznost od 44% dobili odziv od 43%, a u drugome za preciznost od 58% dobili su odziv od 39%.

Wang i ostali [7] su također promatrali isključivo ostatke na površini. Kao svojstva su koristili profile sekvenci i evolucijski omjer. Duljina prozora je 11 ostataka, a mjesto interakcije je prozor čiji je središnji ostatak u kontaktu sa susjednim lancem. Ostatak je u kontaktu ako je udaljenost njegova α atoma ugljika od α atoma ugljika bilo kojeg ostatka susjednog lanca manja od 12 Å. Koristeći SVM i kombinaciju rezultata dobivenih za pojedina svojstva postigli su preciznost od 49,7% uz odziv od 66,3%.

Najbolji rezultat ostvario je M. Šikić [3] koristeći informacije samo iz sekvence te uz preciznost od 60-70% postigao je odziv od oko 40%. Pri tome je kao mjesto interakcije definirao prozor čiji je središnji ostatak u kontaktu kao i još barem 4 ostatka koja su od središnjeg udaljeni ne više od 3 ostatka. Pri tome je ostatak u kontaktu, ako je udaljen manje od 6 Å od ostatka koji pripada proteinu partneru.
Isti kriteriji i svojstva uzimali su se i u ovome radu, osim što je dodana informacija o profilu sekvence kao jedno od svojstava predviđanja. Dok su se u drugome koraku, usporedbe radi, promatrala i strukturna svojstva.

3 Podaci

Za predviđanje mjesta proteinskih interakcija korišten je skup podataka od 1137 lanaca 333 različita proteinska kompleksa [3].

Zapisi o proteinima i njihovim eksperimentalno utvrđenim strukturama, nalaze se u bazi proteinskih struktura PDB (engl. *Protein Data Bank*) [8]. Osnovni podaci o strukturi proteina nalaze se u PDB formatu. PDB format je jednostavna tekstualna datoteka koja sadrži velik broj informacija o lancima proteina, prostornoj strukturi, koordinatama atoma i mnoge druge.

3.1 Priprema podataka

Za predviđanje mjesta interakcije koristila su se sljedeća svojstva:

- sekvenca od 9 ostataka
- profili slijeda aminokiselinskih ostataka

Za analizu i usporedbu koristila su se svojstva geometrijske strukture proteina [3] i to sljedeća svojstva:

- površina dostupna otapalu (engl. *accessible surface area, ASA*)
- ukopanost ostatka u proteinu
- izbočenost ostatka u proteinu
- hidrofobnost
- sekundarna struktura.
- elektro-ionski interakcijski potencijal (EIIP)
- elektrostatski potencijal

Navedena svojstva geometrijske strukture za analizu, određena su korištenjem alata PSAIA [9], čiji je izlaz u XML-u , te alata za dodavanje sekundarne strukture [3]. Svojstvo elektrostatski potencijal kao i profili dodani su u XML vlastito izrađenim alatima.

3.1.1 Profil slijeda aminokiselinskih ostataka

Profil kao mjera evolucijske očuvanosti daje informaciju o vjerojatnosti pronalaženja svake od dvadeset standardnih aminokiselina na mjestu one aminokiseline čiji se profil određuje.
Vlastito izrađenom skriptom u XML datoteke dodane su informacije o profilu aminokiselina svakog lanca pojedinačno, pri tome se koristeći alatom PSI-BLAST [10] koji je detaljno opisan u poglavlju 4.1.3. Informacija o profilu svake aminokiseline lanca dodana je u XML. Uz informacije o strukturi aminokiseline, XML datoteka sada je sadržavala i informaciju o profilu svake aminokiseline u lancu.
Za svaki su se lanac izvukla imena aminokiselina koje ga čine, formirajući pri tome niz slova (svako slovo se odnosi na odgovarajuću aminokiselinu) odnosno sekvencu. Takva bi se sekvenca propustila kroz PSI-BLAST koji bi kao rezultat dao PSSM matricu odnosno profil. SWISS-PROT [11] je proteinska baza podataka tj. sekvenci proteina, korištena u radu u odnosu na koju se gradio profil [12]. Čitajući PSSM svakoj aminokiselini trenutnog lanca, koji se obrađuje, dodana je 20-dimenzionalna informacija, profil.

3.1.2 Elektrostatski potencijal

Pojedine grupe aminokiselina stvaraju elektrostatski potencijal u prostoru zbog postojanja naboja jer sadrže ionizirane skupine atoma. Ovisno o energiji potencijalno polje koje stvara jedan protein privlači ionizirane dijelove drugog proteina. Stoga je elektrostatski potencijal kao značajka pri vezanju makromolekula promatran kao svojstvo u predviđanju mjesta proteinskih interakcija.
Potencijal svakog atoma svake aminokiseline dobiven je korištenjem već gotovog alata DelPhi v.4 [13]. Potencijal atoma aminokiseline računao se za svaki lanac pojedinačno zajedno s pripadajućim ligandima i molekulama vode. Uzeto je da ligand ili molekula vode pripadaju nekom lancu, ako je bilo koji atom liganda ili vode udaljen manje od 5 Angstrema od bilo kojeg atoma aminokiseline koja pripada tom lancu. Ostali lanci, ligandi i molekule vode izbrisani su iz PDB zapisa proteina o čijem se lancu trenutno radi. Tako pročišćenom PDB zapisu dodani su atomi vodika korištenjem alata PDB2PQR [14] i polja privlačenja Amber [15]. Kao što sam naziv govori, alat pretvara PDB datoteku u PQR. Na izlazu ovog programa dobije se PQR datoteka koja je tekstualna, a sadrži informacije zapisane u PDB datoteci uz dodane atome vodika.

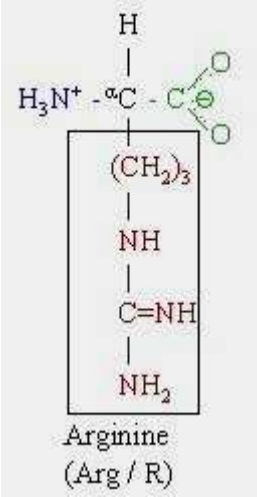
Nakon dodavanja atoma vodika lancu koji se ispituje, slijedi računanje potencijala svakog atoma u lancu pomoću DelPhi programa. Za izračun potencijala korištene su datoteke dobivene u sklopu paketa DelPhi:

- datoteka prot_amb.crg koja sadrži informaciju o naboju svakog pojedinog atoma aminokiseline.
- datoteka parseres.siz koja sadrži informaciju o radijusu svakog pojedinog atoma aminokiseline.

Ulazna datoteka programa jest PQR datoteka koja sadrži dodane vodike u prethodnom koraku, a izlazna FRC datoteka sadrži popis potencijala svakog atoma u aminokiselini koja pripada lancu koji se obrađuje.

Svi parametri [16] potrebni za izračun potencijala zapisani su u parametarskoj datoteci PRM koja se navodi prilikom poziva DelPhi programa.

Za svaku se aminokiselinu ispituje potencijal atoma isključivo bočnog ogranka i traži najveći po apsolutnoj vrijednosti te se dodaje aminokiselini u XML datoteci. Atomi koji nisu u bočnom ogranku su O, OXT, N, CA, H, HN, HA, HA1, HA2, HA3 osim u slučaju glicina kod kojega HA, HA1, HA2 i HA3 jesu u bočnom ogranku.



Slika 3.1 Bočni ogranak arginina

3.2 Čišćenje podataka

Nakon dodavanja profila i potencijala XML datotekama, informacije o lancima skupa proteina za klasifikaciju prevedene su u ARFF (engl. *Attribut Relation File Format*) format. ARFF format [17] koristi velik broj aplikacija za klasifikaciju. Sastoji se od zaglavlja i dijela s podacima. Zaglavlje sadrži informacije o skupu na što se odnosi i tipovima atributa. Glavni elementi

formata su @RELATION, @ATTRIBUTE i @DATA. Prva dva se odnose na zaglavlje, dok treći element predstavlja oznaku u formatu da slijede podaci.

ARFF datoteke u ovome radu sadržavaju informacije o 170 001 instanci, vrijednostima pripadajućih atributa te da li je neka instanca u ovisnosti o postavljenim uvjetima mjesto interakcije. Primjer zapisa jedne instance:

```
1A9X,A,109,GLN,GLY,VAL,LEU,GLU,GLU,PHE,GLY,VAL,14,3,3,2,1,31,5,2,3,1,14,4,3,2
,2,5,2,0,1,2,8,3,1,7,0,2,4,50,1,1,3,4,0,1,2,5,1,0,1,3,11,2,1,2,0,2,7,1,1,11,2
5,3,4,5,1,2,3,1,4,14,3,0,1,1,7,0,1,1,0,3,56,1,1,17,1,1,1,0,1,3,4,6,4,10,2,5,5
0,1,0,1,2,7,0,0,1,2,2,1,0,2,8,5,3,3,1,6,11,1,0,4,13,16,6,3,1,5,5,2,3,4,15,3,4
,1,1,2,7,2,1,3,20,2,11,5,1,4,3,0,12,3,2,3,12,6,1,3,3,57,2,1,1,4,0,0,2,2,1,0,1
,1,1,1,0,1,3,1,1,1,1,34,13,4,2,3,0,2,3,0,3,27,49.506,34.124,20.157,43.775,29.
349,21.064,22.600,34.519,26.906,33.312,0.643,3.016,0.000,0.725,2.473,0.000,0.
411,H,0.02167,0
```

U primjeru se radi o proteinu 1A9X, lancu A te aminokiselini GLU 109-oj u lancu koja je ujedno i središnja kiselina u prozoru od devet aminokiselina. Zatim slijede profili za svaku aminokiselinu u prozoru te vrijednosti svojstava zapisanih u XML datoteci kao što su površina dostupna otapalu, ukopanost ostatka, izbočenost ostatka, EIIP [18], sekundarna struktura, potencijal, te zadnja vrijednost u retku govori da li instanca je ili nije mjesto interakcije.

Vrijednosti nekih svojstava kreću se iznad granica koje se smatraju povoljnima. Stoga su se podaci morali dodatno pročistiti tj. izbrisati one instance čije se vrijednosti atributa ne nalaze u željenim okvirima.

Iz početnog skupa instanci maknute su one sa sljedećim vrijednostima atributa:

- otapalu dostupna površina (ASA) iznad vrijednosti 100 (odnosi se na sve relativne vrijednosti atributa vezanih za ASA-u)
- ukopanost ostatka ispod vrijednosti 0
- izbočenost ostatka ispod vrijednosti 0 i iznad vrijednosti 15
- sve negativne vrijednosti ostalih atributa osim hidrofobnosti i potencijala.

3.3 *Principal Component Analysis* (*PCA*)

S obzirom da je broj strukturnih informacija prilično velik potrebno ga je reducirati, odnosno odabrati dominantne strukturne informacije. Jedan od načina kako odabrati značajnije informacije u velikom skupu informacija je *PCA* metoda.

Veliki broj atributa instanci predstavlja n-dimenzionalni prostor u kojemu broj dimenzija odgovara broju atributa. Stoga je tako velike dimenzije potrebno reducirati na one koje 'pokrivaju' veći broj atributa. Drugim riječima, traže se statistički najvažnija svojstva s najvećim varijancama te se u njihovu smjeru zakreću koordinate koje se sada sastoje od linearne kombinacije značajnijih svojstava tj. atributa.

Ukratko rečeno ono što metoda radi jest da se na temelju podataka i njegovih koordinata, koje su u ovom slučaju svojstva instanci, računa matrica kovarijanci iz koje se vade svojstvene vrijednosti i svojstveni vektori. Oni svojstveni vektori s najvećom svojstvenom vrijednosti su upravo *principal components*. Jednostavnije rečeno svaki taj vektor je linearna kombinacija atributa u prostoru, dok oni s većom svojstvenom vrijednosti linearna su kombinacija dominantnijih atributa koje se i želi izdvojiti.

Ova metoda na jednostavan način ukazuje na sličnosti ili pak različitosti među atributima u prostoru velikih dimenzija kada nije moguće uočiti ikakvu vezu među istima.

3.3.1 Algoritam

U narednom tekstu matematički će biti opisana *PCA* metoda. Radi jednostavnosti neka se radi o dvodimenzionalnom prostoru s koordinatama *x* i *y* i neka se u takvome prostoru nalazi neki broj podataka. Svaki podatak u prostoru određen je vrijednostima *x* i *y* koordinata.

U prvome koraku svi podaci se normiraju. Odnosno računa se srednja vrijednost *x* i *y* koordinata svih podataka u prostoru izrazom:

$$X = \frac{\sum_{i=1}^{n} x_i}{n}, \qquad Y = \frac{\sum_{i=1}^{n} y_i}{n} \tag{3.1}$$

Zatim se računa kovarijanca između svake dvije dimenzije te slaže matrica kovarijanci. Kovarijanca između dvije dimenzije X i Y računa se prema sljedećem izrazu:

$$\text{cov}(X, Y) = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{n - 1}, \tag{3.2}$$

a općeniti izraz za matricu kovarijanci je:

$$C^{n \times n} = \left(c_{ij}, c_{ij} = \left(Dim_i, Dim_j\right)\right) \tag{3.3}$$

gdje je C^{nxn} matrica dimenzije $n \times n$, n broj dimenzija, c_{ij} vrijednost kovarijance između dimenzija i i j.

Matrica kovarijance za dvije dimenzije x i y računa se na način:

$$C^{2 \times 2} = \begin{pmatrix} \text{cov}(x, x) & \text{cov}(x, y) \\ \text{cov}(y, x) & \text{cov}(y, y) \end{pmatrix} \tag{3.4}$$

Nakon što se izračuna matrica kovarijanci slijedi izračun svojstvenih vrijednosti i svojstvenih vektora ove matrice. Postupak dobivanja vektora i svojstvene vrijednosti ovdje neće biti detaljno opisan. Bitno je napomenuti da nakon što se nađu vektori i njihove svojstvene vrijednosti, biraju se oni s najvećom svojstvenom vrijednosti. Takvi vektori linearna su kombinacija najznačajnijih koordinata u višedimenzionalnom prostoru i međusobno su okomiti, te se sada koordinatni sustav zakreće u njihovu smjeru. Na taj se način izdvajaju one koordinate od njih *n,* koje su dominantnije. U ovome se radu na taj način traže one značajke tj. atributi koji čine svojstveni vektor s najvećom svojstvenom vrijednosti. U konačnici se s *n*-dimenzionalnog prostora prelazi u prostor manjih dimenzija od *n* i to u prostor značajnijih koordinata odnosno atributa.

4 Metode

4.1 PSI-BLAST

PSI-BLAST [10] je poboljšani algoritam tehnike BLAST [10] pomoću kojeg je moguće pronaći evolucijsku očuvanost aminokiselinskog ostatka tj. profil. Koristi se metodom sekvencijalnog poravnavanja (engl. s*equence alignment*) za prepoznavanje sličnosti dvaju proteina koji nisu blisko povezani tj. nemaju bliskog zajedničkog homologa [19]. Proteini mogu biti ili sekvencijalno ili strukturno slični iz čega se može zaključiti njihovo zajedničko evolucijsko porijeklo. Sekvencijalna i strukturna sličnost međutim nisu nužno povezane – proteini mogu biti strukturno slični, ali ne i sekvencijalno, i obrnuto. Ideja jest predvidjeti iz slabe sekvencijalne sličnosti zajedničku strukturu. Toj se problematici pristupa metodom sekvencijalnog poravnavanja u sklopu tehnike BLAST.

4.1.1 Sekvencijalno poravnavanje

Prvi korak u gradnji profila jest za neku proteinsku sekvencu pronaći da li ona pripada već poznatoj familiji proteina. Usklađivanjem odnosno 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 i 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 (slika 4.1) koriste se boje radi označavanja onih dijelova sekvenci koji su identični, konzervirani ili pak djelomično konzervirani.

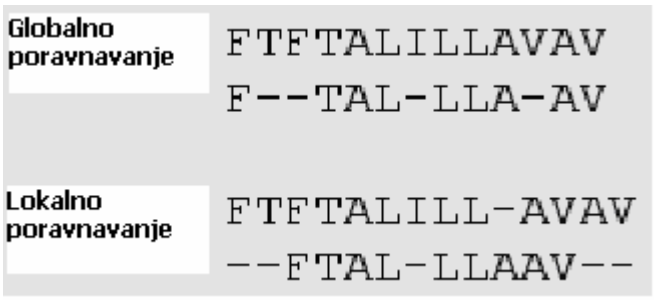
```
AAB24882        TYHMCQFHCRYVNNHSGEKLYECNERSKAFSCPSHLQCHKRRQIGEKTHEHNQCGKAFPT 60
AAB24881        --------------------YECNQCGKAFAQHSSLKCHYRTHIGEKPYECNQCGKAFSK 40
                                    ****: .***:  * *:** * :****.:* ********..

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

Slika 4.1 Identični i konzervativni ostaci su označeni simbolima: '*' - identični ostaci; ':' konzervirani ostaci; '.' djelomično konzervirani ostaci

Jedan od tekstualnih formata prikaza sekvenci je FASTA format [20] koji je ujedno i ulazna jedinica alata BLAST. Radi se o nizu slova u kojemu se svako slovo odnosi na neku od aminokiselina, a započinje oznakom '>'.

Sekvence (dvije sekvence ili više njih) mogu se uskladiti ili poravnati na lokalnoj i globalnoj razini. U globalnom se svi 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 hibridno poravnavanje koje je kombinacija lokalnog i globalnog.



Slika 4.2. Globalno i lokalno poravnavanje

4.1.2 BLAST

BLAST [10], punog naziva *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. Proteinska baza podataka [11] sadrži informacije o poznatim proteinskim sekvencama. Tako algoritam, pretražujući bazu, nalazi sekvence koje odgovaraju sekvenci koja se ispituje, pri tome zadovoljavajući određeni prag sličnosti (engl. *threshold*) koji u pravilu zada korisnik.

Kada se radi sravnjenje sljedova aminokiselinskih ostataka, algoritam BLAST koristi supstitucijsku matricu [21] za procjenu sličnosti sljedova. Postoji nekoliko takvih matrica, BLOSUM i PAM.

4.1.2.1 BLOSUM

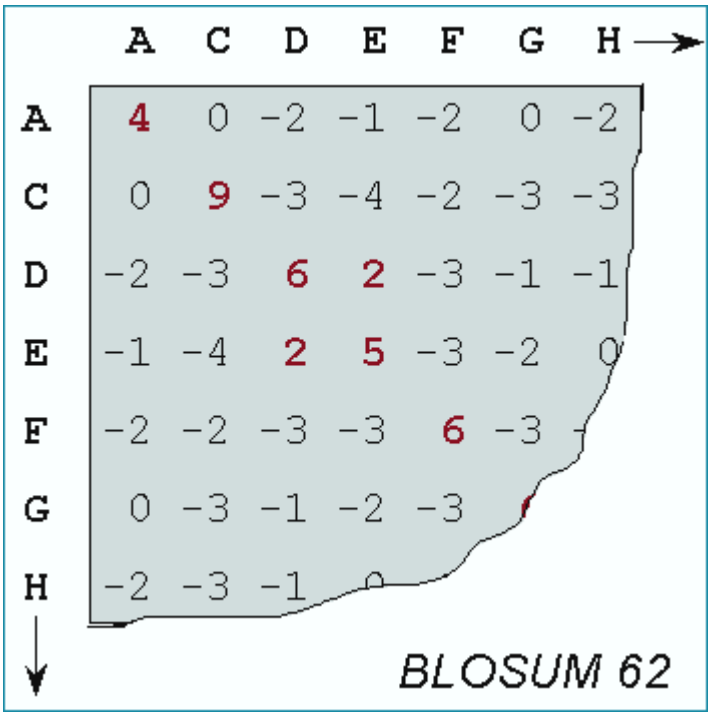
BLOSUM ili *Blocks Substitution Matrix* bazira se na lokalnom poravnavanju, a prvi put je predstavljena u radu Henikoff and Henikoff, 1992 [21]. Nastala je empirijski na temelju poznatih i vrlo konzervativnih regija proteinskih familija u proteinskoj bazi podataka i računanja relativnih frekvencija pojavljivanja 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. Slika 4.3 prikazuje jednu takvu matricu u kojoj je sličnost najmanje 62%. Takva se matrica koristila u ovome radu.

Vrijednosti matrice mogu se izračunati izrazom

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

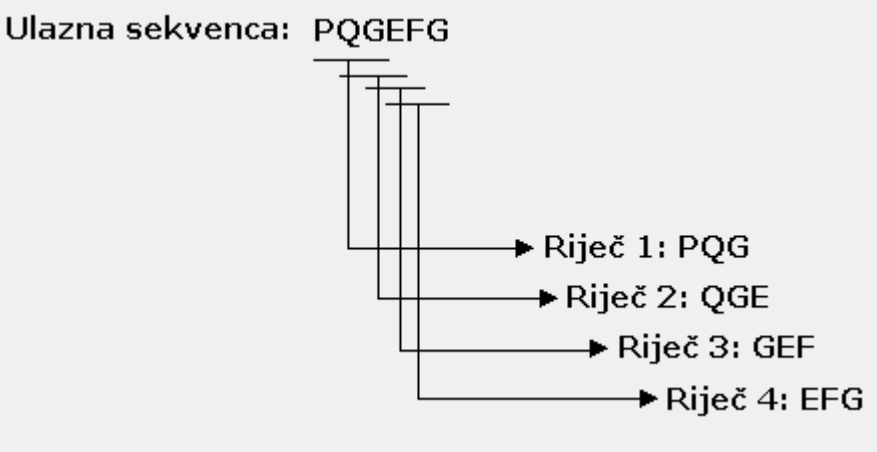
gdje je p_{ij} vjerojatnost da će aminokiseline i i j zamijeniti jedna drugu u homolognoj sekvenci, a q_i i q_j su vjerojatnosti slučajnog nalaženja aminokiselina i i j u sekvenci proteina; λ je skalirajući faktor.



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

4.1.2.2 Algoritam

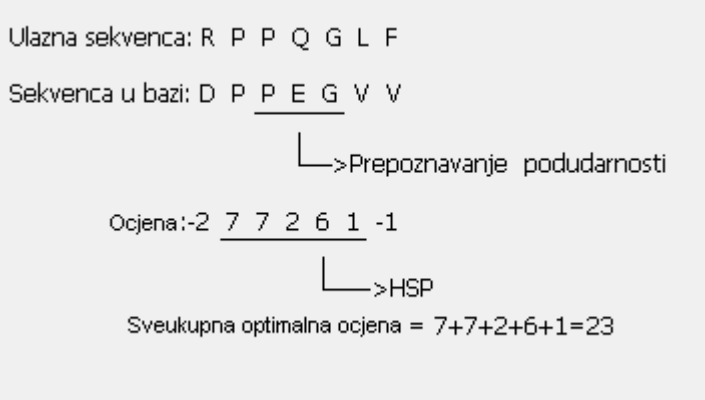
Osnovna ideja algoritma jest da svako, dobro ocjenjeno, lokalno poravnanje dviju sekvenci, gotovo uvijek sadrži dobro očuvanu jezgru. Za parove ostataka u slijedu, određuje se ocjena poravnanja i ako je ona iznad nekog zadanog praga, taj se par ostataka naziva dobro ocjenjeno lokalno poravnanje tj. HSP (engl. *High-scoring Segment Pairs*) [10]. 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. Slika 4.4 prikazuje metodu.



Slika 4.4 Rastavljanje sekvence u riječi od po tri slova

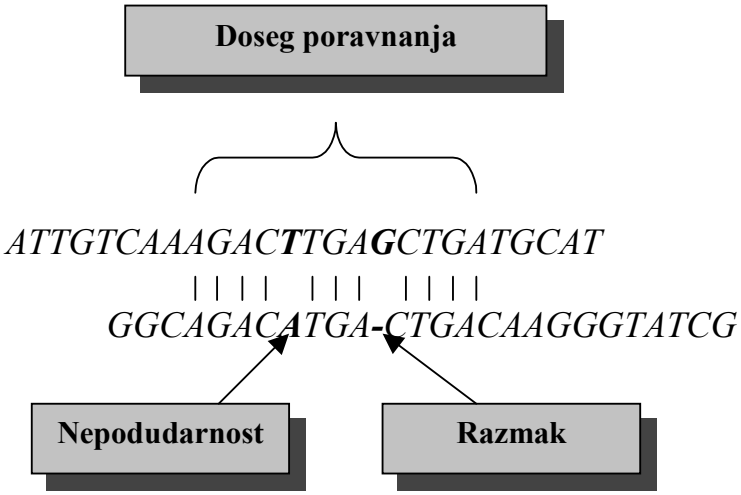
Za svaku takvu riječ se pronalaze ciljne riječi odnosno svi trigrami na alfabetu aminokiselina koji imaju dovoljno veliku sličnost s početnim. Sličnost se iščitava iz supstitucijske matrice za svaki par aminokiselinskih ostataka u trigramu. Primjerice, uspoređujući riječ PQG s riječi PEG ocjena sličnosti 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 prikazuje pojednostavljeni primjer:



Slika 4.5 Proširivanje ciljne riječi na susjedne dok ocjena poravnanja 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 [22] koji vrši poravnavanje s razmacima (engl. *gapped alignment*). Ova se verzija BLAST algoritma prema tome naziva *gapped BLAST* [10].



$$S = \sum(\text{ podudarnost, nepodudarnost }) - \sum(\text{bodovna kazna zbog razmaka})$$

$$\text{Ocjena poravnanja} = \text{Max}(S)$$

Slika 4.6 Poravnanje s razmacima i računanje ocjene poravnanja

4.1.2.3 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 (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:

$$E = N / S' \tag{4.2}$$

gdje je $N = mn$, a *S′* normalizirana vrijednost ocjene *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 $S' = \log_2(N/E)$ 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:

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

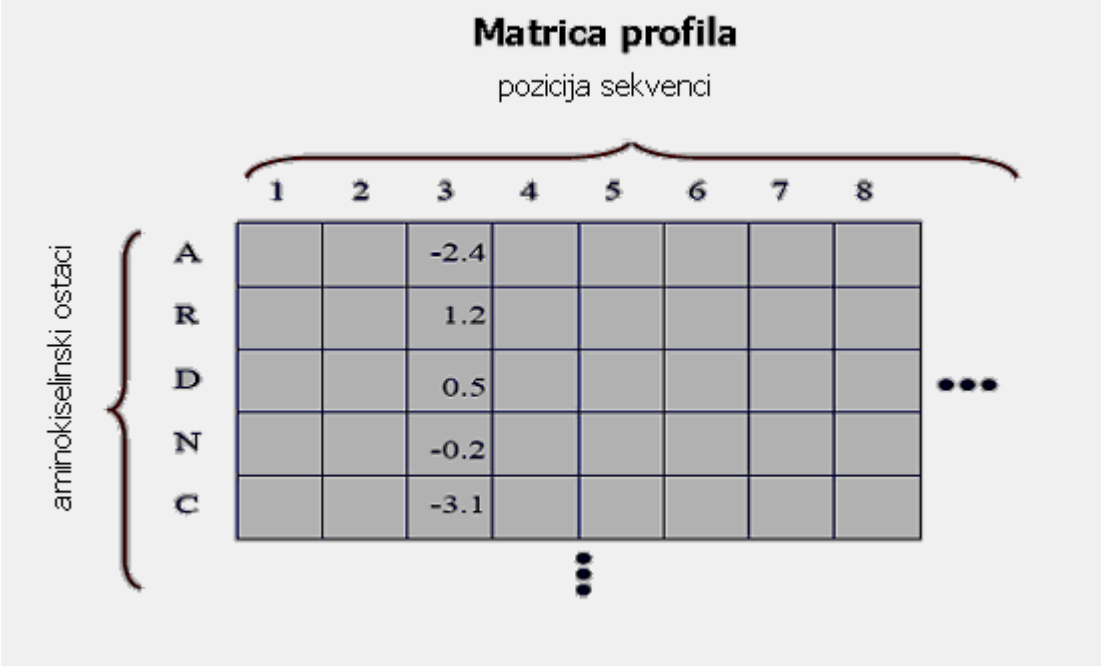
koja teži prema 1. Vrijednosti s_{ij} su elementi supstitucijske matrice:

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

4.1.3 PSI-BLAST

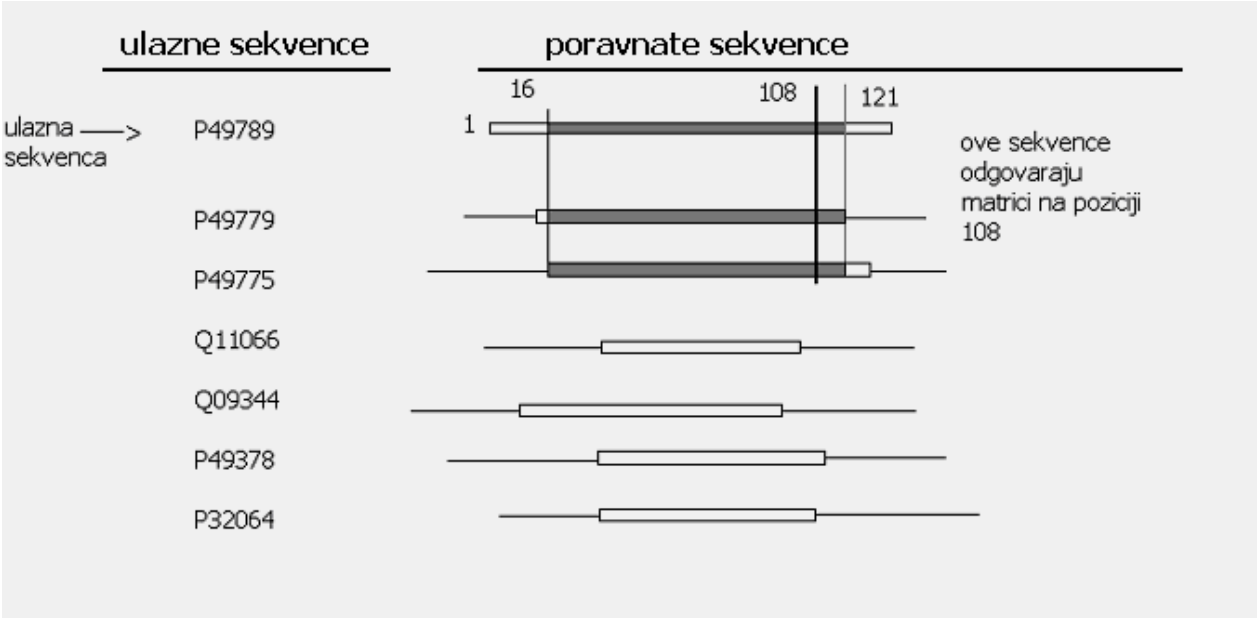
PSI-BLAST [10] punog naziva *Position Specific Iteration BLAST* inačica je BLAST algoritma u kojemu se profil, odnosno matrica vjerojatnosti pronalaženja svake od 20 aminokiselina na mjestu aminokiseline čiji se profil traži (engl. *Position Specific Scoring Matrix, PSSM*), gradi iz višestrukog sekvencijalnog 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 poboljšava rezultat 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. Time je PSI-BLAST osjetljiviji na evolucijski udaljene sekvence proteina.



Slika 4.7 Matrica profila

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 retci 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. Na slici 4.8 je prikaz takve redukcije.



Slika 4.8 Za profil se uzimaju samo one sekvence odgovarajuće duljine čiji se stupci podudaraju ovisno o aminokiselinskim ostacima

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:

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

gdje je

Profile(r,c) vrijednost profila za redak *r* i stupac *c*; *r* može imati vrijednost 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 $\delta(A_{ir},d)$ 0. Comp($residue_d$, $residue_c$) je vrijednost u supstitucijskoj tablici.

Weight(*i*) se odnosi na težinu sekvence *i.* Težina sekvence [23] 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 odabirom 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.

Zbroj težina sekvenci jest 1. Sličnije ili iste sekvence imaju manju težinu od onih *izoliranih*.

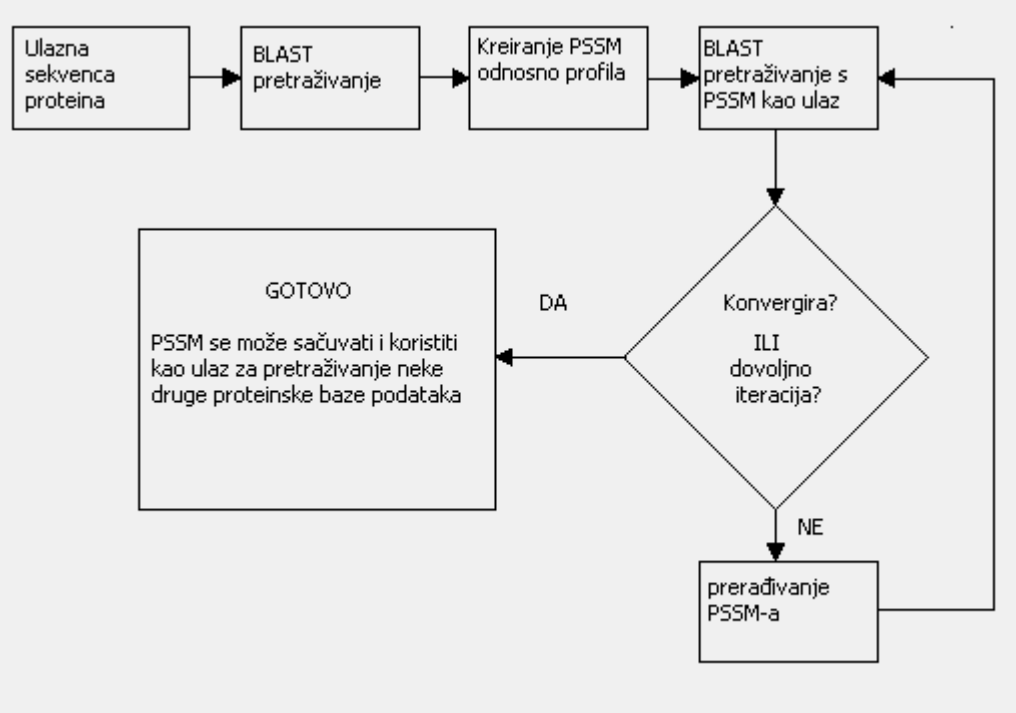
Vizualizaciju opisane metode prikazuje primjer:

sekvenca 1 AAAAAAAAAA težina = 0.25

sekvenca 2 AAAAAAAAAA težina = 0.25

sekvenca 3 CCCCCCCCCCC težina = 0.5

Slika 4.9 prikazuje načelo algoritma PSI-BLAST



Slika 4.9 Dijagram PSI-BLAST algoritma

4.2 Metoda slučajnih šuma

Slučajne šume (engl. *Random Forest*) kao metoda klasifikacije imaju sljedeća svojstva:

- velika *točnost* prepoznavanja
- relativno je otporna na *outliere* i šum
- brža je do *bagginga* i *boostinga*
- daje korisne interne procjene pogreške bez potrebe za korelacijom
- daje procjenu o važnosti pojedinih značajki za klasifikaciju
- čuva točnost ukoliko je skup podataka nepotpun i neuravnotežen
- čuva izgrađene šume i računa prototipove koji se mogu koristiti u uspostavljanju odnosa između značajki i klasifikacije
- računa udaljenosti između svaka dva uzorka koji se mogu koristiti u nenadgledanom učenju (engl. *clustering*) i omogućuje eksperimentalno utvrđivanje interakcija pojedinih značajki.

Slučajna šuma (*RF*), je općeniti naziv za skupinu metoda koje se koriste stablastim klasifikatorima $\{h(\mathbf{x}, \mathbf{\Theta_k}), k = 1,...,\}$ gdje je $\{\Theta_k\}$ skup jednoliko distribuiranih, međusobno potpuno neovisnih vektora, a **x** ulazni vektorski uzorak. Prilikom treniranja, *RF* algoritam stvara velik broj stabala, od kojih se svako trenira na određenom broju uzoraka originalnog trening seta odabranih *bootstraping* metodom. *RF* koristi *m* slučajno odabranih varijabli ($m << M$, obično $log_2 M + 1$) i uzima one varijable koje 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 stabla, vrijednosti koje nisu odabrane se koriste 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. Koriste se 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 uzoraka za testiranje. 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 definira se j kao klasa 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 stablima 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*.

Druga mjera koja se koristi za procjenu važnosti pojedinih varijabli je *Gini kriterij*. Općenito se kod stabala odluke za odabir varijable grananja koristi tzv. funkcija „nečistoća“ (engl. *impurity function*) $\varphi(p)$ koja se temelji na omjeru p uzoraka za učenje koje su u pojedinim klasama. Funkcija nečistoća treba biti takva da je maksimalna kada pojedini čvor sadrži jednak broj svake od mogućih klasa. (Ako npr. postoji isti broj slučaja koji pripadaju klasi ω_1 i klasi ω_2, tada je nemoguće pridružiti taj čvor jednoj od klasa, i u stvari, nemoguće je favorizirati bilo koju klasu pred drugima). Ako je omjer koji pripada ω_1 u roditeljskom čvoru p, a p_l i p_r omjeri iste klase u čvorovima djeci (lijevom i desnom) smanjenje nečistoće nastalo djeljenjem se modelira kao:

$$\varphi(p) - s\varphi(p_l) - (1-s)\varphi(p_r) \tag{4.7}$$

gdje s predstavlja omjer svih slučaja u lijevom čvoru djetetu. Kao jedna od najčešćih funkcija koja se koristi je *Gini kriterij* raznolikosti koji se definira kao:

$$\varphi(p) = 2p(1-p) \tag{4.8}$$

Svaki puta kada se napravi grananje na varijabli *m Gini* kriterij za dva nasljedna čvora je manji nego za čvor roditelj. Dodavajući *Gini* smanjenje za svaku individualnu varijablu preko svih

stabala u šumi dobiva se brza procjena važnosti varijable koja je često u skladu s mjerom permutirane važnosti.

U slučaju velikog broja varijabli, prvo se izgradi šuma sa svim varijablama i na taj se način odrede one najvažnije koje se koriste u sljedećem krugu. Ako je broj varijabli jako velik, znači svaka instanca se sastoji od većeg broja varijabli, moguće je obaviti klasifikaciju sa svim varijablama, pa opet ponoviti postupak samo s najvažnijim varijablama.

4.2.1 Postupak izgradnje stabla

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 pritom 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. Korake 2-5 rekurzivno ponavljati za svaki novi čvor.

4.2.2 Konvergencija slučajnih šuma

Uz skup klasifikatora $h_1(x), h_2(x), \ldots, h_K(x)$ te skupom za učenje stvorenim od nasumce izabranih instanci iz distribucije slučajnih vektora Y, X (X - vektor atributa, Y – klasa), funkcija margine glasi:

$$mg(X,Y) = av_k I(h_k(X) = Y) - \max_{y \neq Y} av_k I(h_k(X) = j) \tag{4.9}$$

gdje $I()$ je indikatorska funkcija. Margina daje podatak koliko prosječan broja glasova za ispravnu klasu nadmašuje prosječan broj glasova za bilo koju drugu klasu. Što je veća margina to je klasifikator točniji. Greška prilikom generalizacije zapisuje se kao:

$$PE^* = P_{X,Y}\left(mg\left(X,Y\right) < 0\right) \tag{4.10}$$

Kod slučajnih šuma $h_k\left(X\right) = h\left(X, \Theta_k\right)$. Kako se broj stabala povećava za gotovo sve sekvence $\Theta_1, \ldots$. PE^* konvergira prema

$$P_{X,Y}(P_\Theta(h(X,(\Theta) = Y) - \max_{y \neq Y} P_\Theta(h(X,\Theta) = j) < 0) \tag{4.11}$$

Ovaj rezultat potvrđuje da povećanje broja stabala ne dovodi do preprilagođenja podacima.

4.2.2.1 Snaga i korelacija

Funkcija margine za slučajnu šumu glasi:

$$mr(X,Y) = P_\Theta(h(X,\Theta) = Y) - \max_{y \neq Y} P_\Theta(h(X,\Theta) = j) \tag{4.12}$$

a snaga skupa klasifikatora $\{h(x,\Theta)\}$ se definira kao:

$$s = E_{X,Y} mr(X,Y) \tag{4.13}$$

Uz pretpostavku da je $s \geq 0$, Čebiševljeva korištenjem nejednakost se dobije:

$$PE^* \leq \frac{\mathrm{var}(mr)}{s^2} \tag{4.14}$$

Izraz koji bolje opisuje $var(mr)$ može se izvesti iz sljedećeg. Neka je

$$\hat{j}(X,Y) = \arg\max_{y \neq Y} P_\Theta(h(X,\Theta) = j) \tag{4.15}$$

dalje se može pisati:

$$\begin{aligned} mr(X,Y) &= (P_\Theta(h(X,(\Theta) = Y) - P_\Theta(h(X,\Theta) = \hat{j}(X,Y)) = \\ &= E_\Theta\left[I(h(X,(\Theta) = Y) - I(h(X,\Theta)) = \hat{j}(X,Y)\right] \end{aligned} \tag{4.16}$$

Funkcija neobrađene margine glasi:

$$rmg(\Theta, X, Y) = I(h(X, \Theta) = Y) - I(h(X, \Theta) = \hat{j}(X, Y)) \tag{4.17}$$

Potrebno je primijetiti da je *mr(X,Y)* očekivanje *rmg(Θ,X,Y)* s obzirom na *Θ*. Za svaku funkciju *f* identitet se računa kao:

$$\left[E_{\Theta} f(\Theta)\right]^2 = E_{\Theta,\Theta'} f(\Theta) f(\Theta') \tag{4.18}$$

Postoji ako *Θ* i *Θ′* su međusobno neovisne ali sa istom distribucijom. Dalje slijedi

$$mr(X, Y)^2 = E_{\Theta,\Theta'} rmg(\Theta, X, Y) rmg(\Theta', X, Y) \tag{4.19}$$

Koristeći gornji izraz može se pisati:

$$\begin{aligned} \text{var}(mr) &= E_{\Theta,\Theta'}(\text{cov}_{X,Y}\, rmg(\Theta, X, Y) rmg(\Theta', X, Y)) = \\ &= E_{\Theta,\Theta'}(\rho(\Theta, \Theta') sd(\Theta) sd(\Theta')) \end{aligned} \tag{4.20}$$

Gdje *ρ(Θ, Θ′)* predstavlja korelaciju između *rmg(Θ, X, Y)* i *rmg(Θ′, X, Y)* uz fiksne *Θ* i *Θ′*, a *sd(Θ)* standardna devijacija *rmg(Θ, X, Y)* uz fiksni *Θ*. Dalje je

$$\begin{aligned} \text{var}\left(mr\right) &= \bar{\rho}\left(E_{\Theta} sd\left(\Theta\right)\right)^2 \\ &\le \bar{\rho} E_{\Theta}\, \text{var}\left(\Theta\right) \end{aligned} \tag{4.21}$$

Gdje $\bar{\rho}$ predstavlja srednju vrijednost korelacije, i to

$$\bar{\rho} = \frac{E_{\Theta,\Theta'}\left(\rho\left(\Theta, \Theta'\right) sd\left(\Theta\right) sd\left(\Theta'\right)\right)}{E_{\Theta,\Theta'}\left(sd\left(\Theta\right) sd\left(\Theta'\right)\right)} \tag{4.22}$$

Gornja granica generalizacijske pogreške glasi:

$$PE^* \le \frac{\bar{\rho}\left(1 - s^2\right)}{s^2} \tag{4.23}$$

Ovaj izraz pokazuje kako dva osnovna elementa pogreške generalizacije slučajnih šuma su snaga pojedinog klasifikatora unutar šume i korelacija među njima u izrazu neobrađene funkcije margine. Izraz c/s^2 predstavlja omjer korelacije i kvadrata snage. Da bi se shvatio način na koji slučajne šume funkcioniraju ovaj omjer je jako koristan. Što je on manji to bolje.

Omjer c/s^2 za slučajnu šumu definiran je kao:

$$\frac{c}{s^2} = \frac{\bar{\rho}}{s^2} \tag{4.24}$$

U slučaju da ulazni podaci imaju samo dvije klase dolazimo do pojednostavljenja. Funkcija margine glasila bi:

$$mr\left(X, Y\right) = 2P_{\Theta}\left(h\left(X, \Theta\right) = Y\right) - 1 \tag{4.25}$$

Neobrađena margina izgleda $2I\left(h\left(X,\Theta\right)=Y\right)-1$, a korelacija $\bar{\rho}$ je između $I\left(h\left(X,\Theta\right)=Y\right)$ i $I\left(h\left(X,\Theta'\right)=Y\right)$.

U slučaju da su vrijednosti Y +1 i -1 slijedi

$$\bar{\rho}=E_{\Theta,\Theta'}\left[\rho\left(h\left(\cdot,\Theta\right),h\left(\cdot,\Theta\right)\right)\right] \tag{4.26}$$

4.2.3 Kreiranje slučajnih šuma

Najjednostavnije slučajne šume se kreiraju slučajnim odabirom grupe atributa u svakom čvoru te odabirom onog koji najbolje grana stablo. Stablo raste koristeći CART metodologiju do maksimalne veličine. Ne koristi se potkresivanje stabla (engl. *prune*). Ova metoda se naziva *Forest-RI* (engl. *random input*). Vrijednost grupe atributa koja se odabire je konstanta. Kao vrijednosti konstante F se obično koriste „1“ i $int(log_2(M)+1)$, gdje je M ukupan broj atributa.

Ukoliko postoji samo nekoliko ulaznih atributa, ako uzmemo da je F značajna vrijednost od M, može se dobiti povećanje u snazi klasifikatora, no istodobno i u korelaciji. Veći broj svojstava se može definirati na način da se uzme linearna kombinacija određenog broja, L, ulaznih atributa. U svakom čvoru L atributa se slučajno odabere i zbroji zajedno s koeficijentima koji su uniformno slučajni brojevi iz skupa [-1, 1]. Kreira se F linearnih kombinacija i između njih se odabire ona koja najbolje grana stablo. Ova metoda se naziva *Forest-RC* (engl. *random combination*).

U slučaju korištenja nominalnih varijabli, potrebno ih je pretvoriti tako da se mogu kombinirati s numeričkim. Kod korištenja nominalnih varijabli kod svakog grananja, u čvoru se odabere slučajan podskup kategorija i definira se zamjenska varijabla koja poprima vrijednost 1 kada je kategorijska vrijednost varijable unutar podskupa, a nula ako je izvan. Nominalne varijable s I vrijednosti se zamjenjuju sa I -1 0-1 varijabli. Na taj način, vjerojatnost da će pojedina nominalna varijabla biti izabrana u odnosu na numeričku je I-1 puta veća. U slučaju kada se koristi mnogo nominalnih varijabli vrijednost konstante F mora biti dva do tri puta veća od $int(log_2M+1)$ da bi se dobilo dovoljno snage za postizanje optimalne točnosti skupa za učenje.

4.3. *OR* metoda

Metoda za poboljšanje *F-mjere* koristeći kombinaciju klasifikatora [3]. Metoda se primjenjuje u slučaju nebalansiranih podataka. Nebalansirani podaci su oni u kojima je jedna klasa većinska, a druga manjinska. Manjinska se klasa označi pozitivnom oznakom, a većinska negativnom oznakom odnosno nulom. Kombinacija klasifikatora se provodi na način da ukoliko je barem jedan od klasifikatora proglasio neku instancu ili događaj kao pozitivnu klasu, tada je izlaz pozitivna klasa. U slučaju da svi klasifikatori neki događaj klasificiraju većinskom klasom, izlazna klasa je većinska.

Osnovna ideja *OR* metode je poboljšanje odziva uz istodobno poboljšanje *F-mjere* na račun mogućeg smanjenja preciznosti. Kako su obično vrijednosti preciznosti veće od odziva, te kako je *F-mjera* harmonijska sredina preciznosti i odziva, dobitak u odzivu nadmašuje gubitke u preciznosti, a često se i preciznost zna povećati.

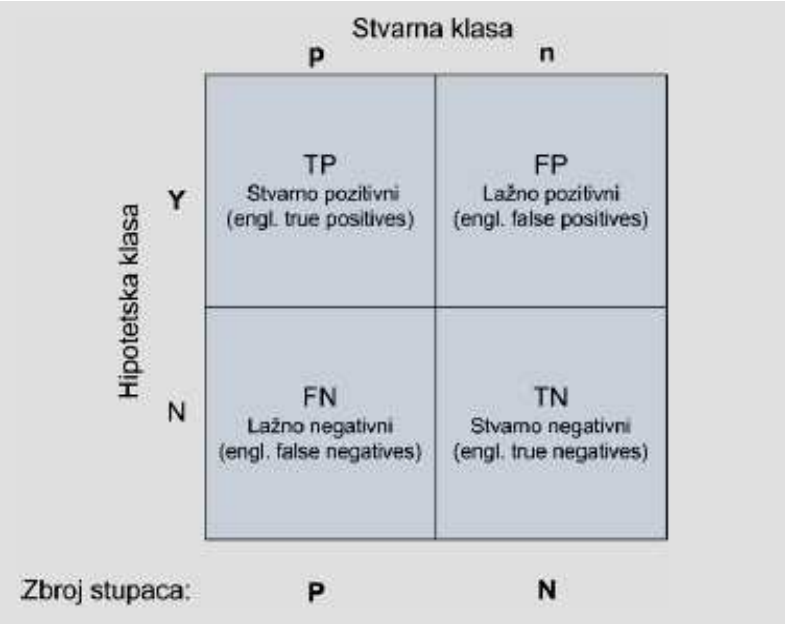
OR metoda se može primijeniti na nekoliko načina. Jedan način je da se isti klasifikatori primjenjuju na različite skupove ulaznih podataka dobivene izvodom iz istog početnog skupa, npr. *bootstraping* metodom. Drugi način je odabir različitih skupina svojstava i provođenje kombinacije tako dobivenih rezultata. Treći način je korištenje različitih klasifikatora na istom skupu podataka. Četvrti način je korištenje činjenice da se često unutar pojedine klase mogu definirati podklase te klase. Odabire se jedna od dobivenih podklasa i sve njoj pripadne instance označe se pozitivnom vrijednošću, a instance svih ostalih podklasa manjinske klase označe se negativnom vrijednošću i pridruže instancama većinske klase. Postupak odabira podklase odnosno podskupa skupa pozitivno označenih instanci se može ponavljati. Kombiniraju se rezultati klasifikatora na prvotno definiranom skupu podataka te kasnije definiranim skupovima s podklasama.

U slučaju kada su poznate manjinske klase, tada se primjenjuje *OR1* metoda, dok se u *OR2* metodi radi podjela manjinske klase pri čemu u metodi nije potrebno odabirati podklase i ponovo ih pridjeljivati instancama. U *OR2* se treniraju i ocjenjuju svi klasifikatori, te se poredaju po veličini tako da je prvi onaj klasifikator s najmanjom preciznošću.

Za detaljan opis *OR* metode pogledati doktorsku disertaciju M. Šikića [3].

4.4 Mjere uspješnosti predviđanja

Da bi se za neki klasifikator vidjelo da li dobro klasificira instance, potrebno je uvesti nekoliko mjera. U ovom slučaju klasifikator odlučuje da li neka instanca pripada jednoj od dvije klase. Za njega se može reći da je binarni klasifikator, a klase se označavaju kao pozitivna i negativna. Klasifikator uči na nekom odabranom skupu za učenje za kojeg se znaju polazne vjerojatnosti, nakon čega se evaluacija rezultata provodi na skupu za testiranje. Rezultat je matrica greške. Uzorci testnog objekta koje je klasifikator označio pozitivnima, raspoređuju se u dvije kategorije: stvarno pozitivni *TP* (engl. *true positives*) i lažno pozitivni *FP* (engl. *false positives*). Na isti se način raspoređuju oni uzorci proglašeni negativnima: stvarno negativni *TF* (engl. *true negatives*) i lažno negativni *FN* (engl. *false negatives*).



Slika 4.10 Matrica greške

Na temelju tablice računaju se dodatne mjere koje opisuju karakteristike klasifikatora.

Odziv se definira kao $odziv = \frac{TP}{TP + FN}$, odnosno kao omjer ispravno klasificiranih pozitivnih instanci i stvarnog broja pozitivnih instanci.

Preciznost se definira kao $preciznost = \frac{TP}{TP + FP}$, odnosno kao omjer ispravno klasificiranih pozitivnih instanci i instanci koje je klasifikator proglasio pozitivnima.

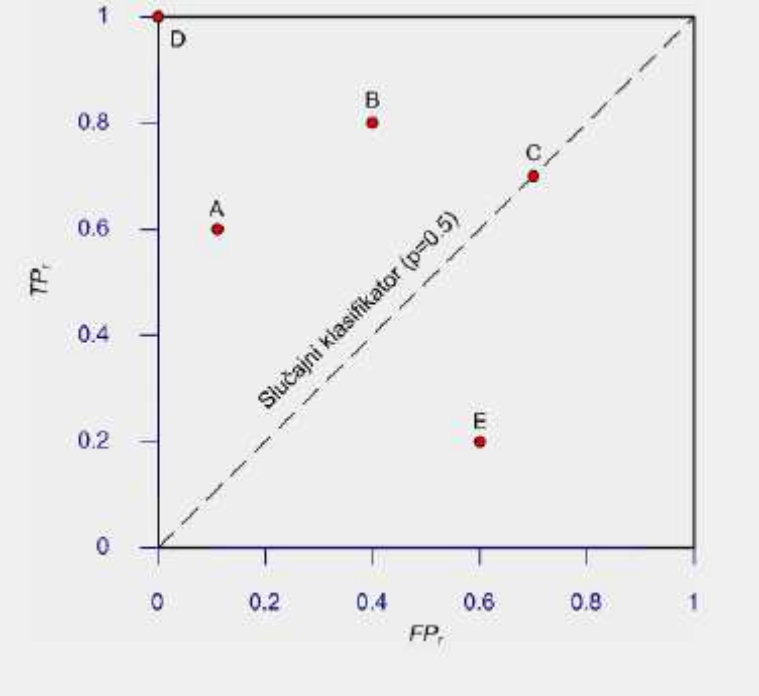
F-mjera je težinska harmonijska srednja vrijednost preciznosti i odziva. Definira se kao

$$F - mjera = \frac{2 \cdot (preciznost \cdot odziv)}{preciznost + odziv}.$$

Točnost je određena izrazom $točnost = \frac{TP + TN}{P + N}$, odnosno kao omjer ispravno klasificiranih i pozitivnih i negativnih instanci i ukupnog broja instanci.

4.4.1 Analiza ROC grafa i krivulje, površine ispod ROC krivulje i grafa preciznost-odziv

ROC grafovi su dvodimenzionalni grafovi u kojima je *TP* na Y-osi i *FP* na X-osi te prikazuju odnos između koristi (*TP*) i cijene (*FP*)
Na slici 4.11 je prikazan *ROC* graf s pet klasifikatora označenih od A do E.



Slika 4.11 *ROC* graf s 5 klasifikatora

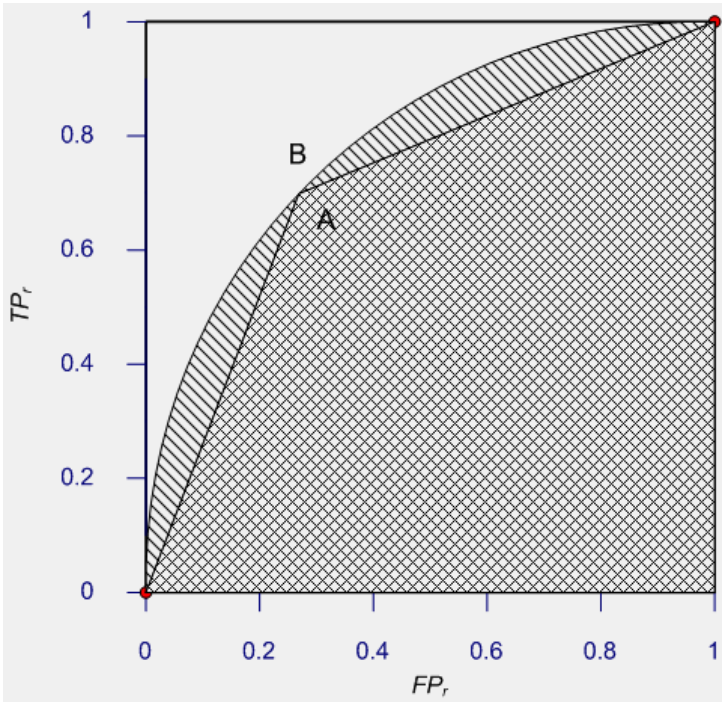
Svi klasifikatori koje prikazuje slika 4.11 su diskretni. Diskretni klasifikator na izlazu daje diskretni par (*TP, FP*) i on predstavlja jednu točku klasifikatora. Oni klasifikatori čija se točka nalazi bliže gornjem desnom kutu su bolji (na slici 4.11 klasifikator oznake D) dok su oni klasifikatori čije su točke bliže donjem lijevom kutu, lošijih performansi.
Dijagonalna linija predstavlja slučaj u kojem se oznaka klase bira slučajno.

Osim diskretnih klasifikatora postoje i vjerojatnosni klasifikatori. Takvi klasifikatori donose numeričku vrijednost koja predstavlja stupanj pripadnosti pojedine instance nekoj klasi. Ako ih se koristi s odgovarajućim pragom dobiju se diskretni klasifikatori; ako je izlaz iznad praga, izlaz klasifikatora je *T*, inače *N*. Svaka vrijednost praga daje različitu vrijednost u *ROC* prostoru.

Diskretni klasifikatori na izlaz daju oznaku klase za svaku testnu instancu. No kada se žele saznati rezultati klasifikatora tada se radi *ROC* krivulja. U tom se slučaju diskretni klasifikatori pretvaraju u one koji daju rezultate na način da se gleda statistika pojedinih instanci koje sadrže.

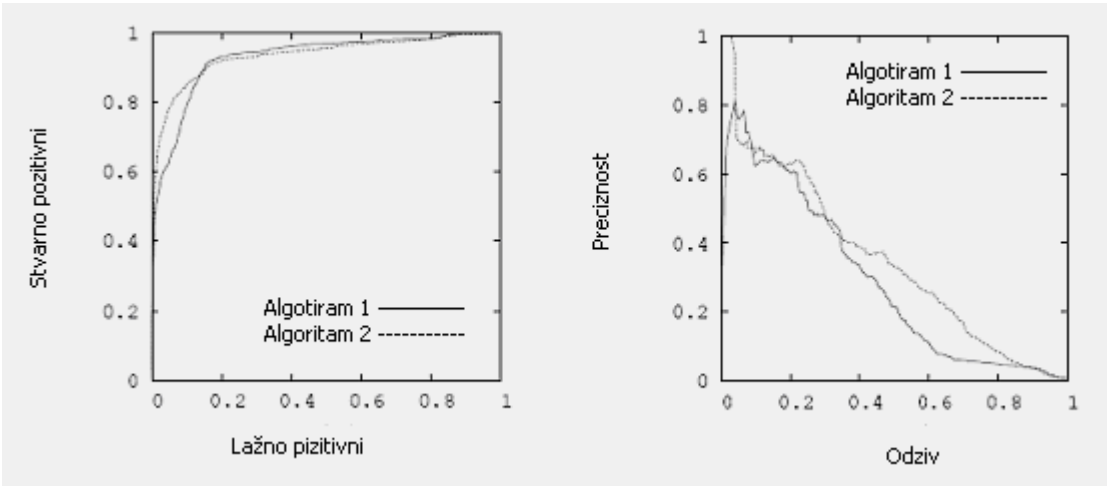
ROC krivulja je dvodimenzionalni prikaz performansi klasifikatora. Uobičajeno je izračunati površinu ispod krivulje koja se naziva *AUC* (engl. *Area under ROC curve*). Ta je površina uvijek manja od 1,0 pošto je dio površine jediničnog kvadrata s time da niti jedan realni klasifikator ne bi trebao imati *AUC* manju od 0,5.
Statističko svojstvo *AUC*-a je to da je jednak vjerojatnosti da će klasifikator rangirati slučajno odabranu pozitivnu instancu više od slučajno odabrane negativne instance.

Slika 4.12 prikazuje površine ispod dvije *ROC* krivulje, A i B. Klasifikator B ima veću površinu i stoga bolju srednju vrijednost performansi. Klasifikator predstavlja performanse klasifikatora B kada je B korišten s jednim, fiksnim pragom. Iako su performanse ova dva klasifikatora jednake u fiksnoj točci (A prag), performanse A su lošije od onih B dalje od te točke.



Slika 4.12 Graf prikazuje površinu ispod krivulje (*AUC*) za diskretni (A) i vjerojatnosni (B) klasifikator

Graf preciznost-odziv (*PR*) i njemu pripadajuća krivulja se često koriste kao alternativa *ROC* krivuljama u slučaju asimetričnosti (engl. *skew*) u raspodjelama klasa [24]. Važna razlika između *ROC* prostora i *PR* prostora je vizualna reprezentacija krivulja. Gledajući *PR* krivulje moguće je vidjeti razliku među algoritmima koja nije očita u *ROC* prostoru. Na slici 4.13 prikazane su *ROC* krivulja i *PR* krivulja za neka dva algoritma.



Slika 4.13 Usporedba algoritama u *ROC* i *PR* prostoru

Lijeva slika se odnosi na *ROC* krivulju, a desna na *PR* krivulju. U ROC prostoru rezultati su bolji ako krivulja teži prema gornjem desnom kutu, dok u PR prostoru su rezultati bolji ako krivulja teži prema gornjem lijevom kutu.

Prema desnoj slici koja prikazuje ROC prostor, moglo bi se zaključiti da su performanse oba algoritma približno jednake. Međutim, u PR prostoru se vidi mala, ali očita prednost jednog od algoritama. Prema tome, za evaluaciju performansi klasifikatora u ovome su se radu rezultati prikazivali u PR prostoru tj. *PR* krivuljama.

5 Rezultati

Za učenje i klasifikaciju odnosno predviđanje mjesta interakcije korištena je paralelna realizacija slučajnih šuma, PARF [25]. Inačica izvornog algoritma slučajnih šuma omogućava bržu implementaciju i korištenje metode slučajnih šuma jer se algoritam paralelno vrši na više računala. Za crtanje *preciznost-odziv* krivulja korišten je alat *R*.

5.2 Rezultati predviđanja koristeći informacije iz sekvence i profila sekvence

Definiciju mjesta interakcije korištenu u ovome radu predložili su Ofran i Rost [1]. Pri predviđanju mjesta interakcije korišten je pomični prozor od 9 ostataka s jednim do osam mjesta kontakta kao pragom za definiranje mjesta interakcije. Pritom je potrebno spomenuti da slučaj s vrijednošću praga 9 nije bio moguć zbog greške u radu *clustera.*

U sklopu metode slučajnih šuma korišteno je 200 stabala, te 14 svojstava pri grananju, a ulazni set za učenje činilo je 170 001 instanci. Za evaluaciju rezultata koristio se *oob* (engl. *out of bag error*) objašnjen u poglavlju 4.2. Za definiranje pozitivnih i negativnih klasa kod matrice greške uzimao se prag od 0,5.

Tablica 5.2 prikazuje rezultate za pojedine vrijednosti praga u prozoru. Promatrajući rezultate za prag 1, odnosno 1 ostatak u prozoru od 9 ostataka je mjesto kontakta, postiže se *preciznost* od 86,80% uz *odziv* od 35,21%*,* dok je *preciznost* maksimalna u slučaju praga 3 i iznosi 86,98%. Primjenom *OR1* metode može se uočiti porast *F-mjere* na 53,65% pri čemu je došlo do pada *preciznosti* u odnosu na prag 1 na 84,87% što je i za očekivati, ali je porastao i *odziv* na 39,23%. Također je i *AUC* najveći kada se primijeni *OR1* metoda.

Matrica greške je vrlo zoran prikaz performansi klasifikatora. Redak matrice predstavlja stvarnu klasu instanci dok stupac predstavlja odluku klasifikatora slučajne šume. Jedna takva matrica koja se odnosi na vrijednost praga 1 prikazana je tablicom 5.1. Iz tablice se može pročitati da je 46 019 instanci mjesto interakcije.

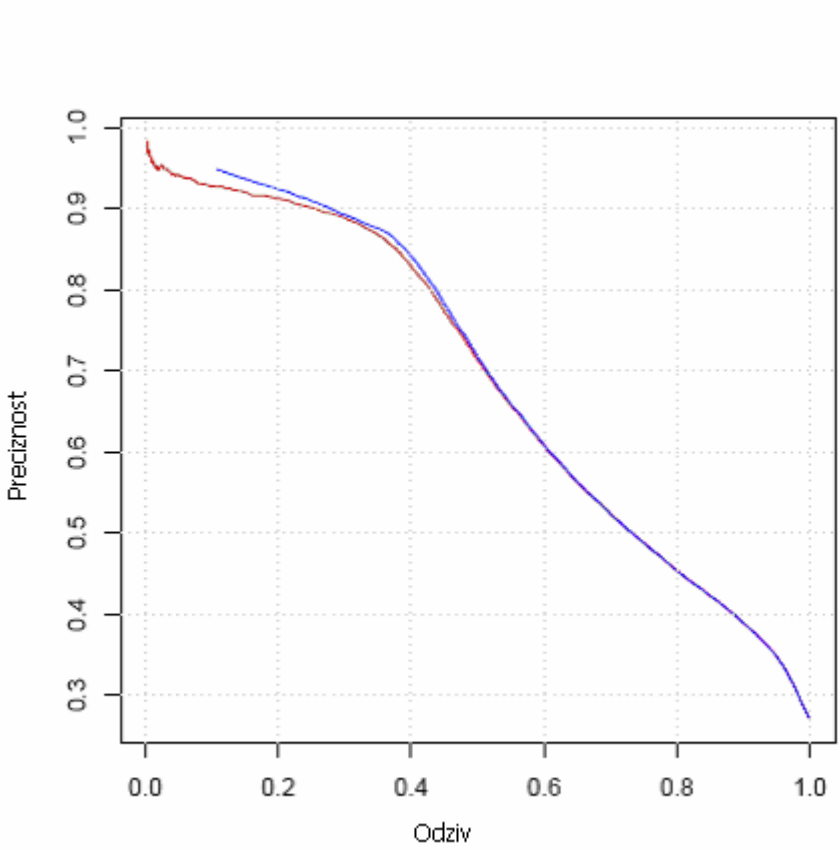
Tablica 5.1 Matrica greške za prag 1

	0	1
0	121 518	2 464
1	29 817	16 202

Tablica 5.2 Evaluacija rezultata za predviđanje mjesta interakcije iz sekvenci i profila sekvenci za različite vrijednosti praga te rezultat dobiven *OR1* metodom. U tablici su prikazani *preciznost*, *odziv*, *F-mjera, točnost* i *AUC*.

Prag	Preciznost	Odziv	F-mjera	Točnost	AUC
1	86,80	35,21	50,10	81,01	81,93
2	86,62	34,88	49,73	81,19	81,60
3	86,98	34,51	49,41	81,71	81,14
4	86,67	33,26	48,07	82,55	79,64
5	85,76	31,33	45,89	83,96	77,14
6	84,95	28,81	43,02	86,55	72,87
7	84,31	26,20	39,97	89,73	68,32
8	84,05	24,44	37,86	93,15	68,39
OR	84,87	39,23	53,65	81,65	82,00

Slika 5.1. prikazuje usporedbu krivulja *preciznost-odziv*. Može se uočiti poboljšanje primjenom *OR1* metode nad izlazima klasifikatora za različite vrijednosti praga.



Slika 5.1 Krivlja *preciznost-odziv* za prag 1 (crvena) te prag 1 i primjenu *OR1* metode (plava)

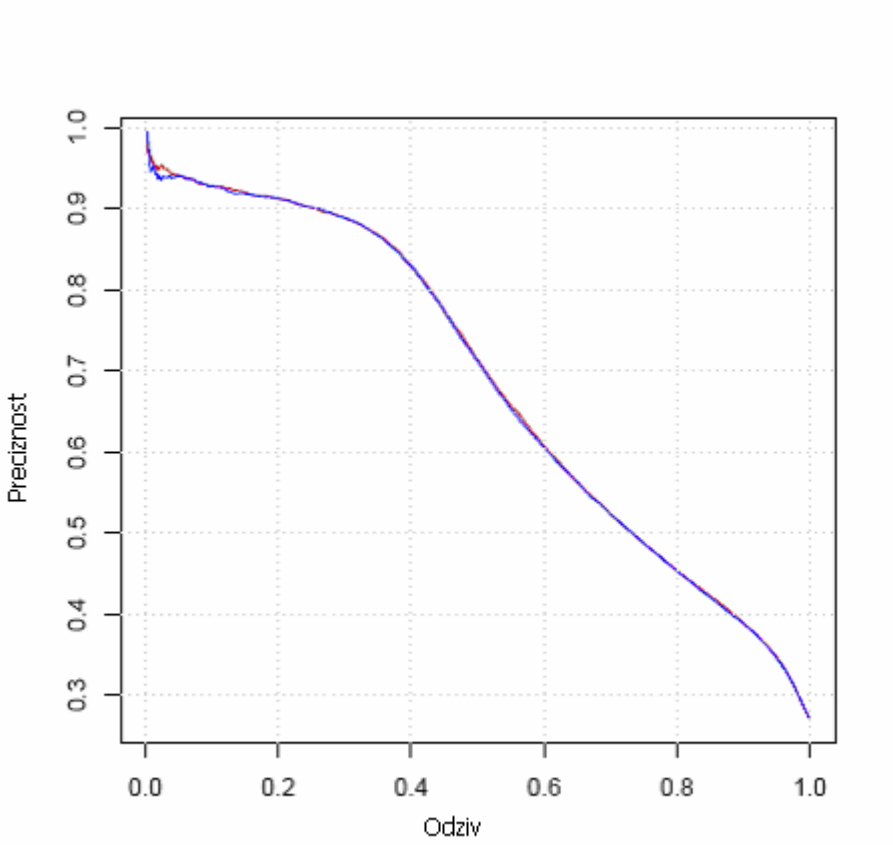
Prikazani rezultati odnose se na slučaj kada su težine pozitivne i negativne klase jednake odnosno 1:1. Ako se klasama pridruže različite težine dobiju se i različiti rezultati predikcije. Pri tome se pozitivnoj klasi pridruži veća težina već negativnoj klasi. Ispitivanjem je uspostavljeno da je optimalan izbor težine 3 za pozitivnu klasu i 2 za negativnu. Kriterij po kojem se bira koji omjer težina je dobar, je *AUC* odnosno površina ispod *ROC* krivulje. Ako se pozitivnoj klasi pridruži težina 3, a negativnoj klasi pridruži težina 2, dobiju se rezultati prikazani u tablici 5.3. Za uočiti je da je vrijednost *F-mjere* za prag 1 porasla u odnosu na slučaj jednakih težina te iznosi 56,42%. Također su porasle vrijednosti *odziva* za sve navedene vrijednosti praga nego u slučaju jednakih težina. Ispitivanja su pokazala da u slučaju različitih težina klasifikatori s vrijednošću praga većim od 5 ulaze u zasićenje tako da se nisu vršile predikcije za vrijednost praga iznad 5.

Primjenom *OR1* metode nad klasifikatorima do vrijednosti praga 5, postižu se još bolji rezultati za *odziv, F-mjeru* i *AUC*. Tako *odziv* raste s 44,01% na 51,54%, a *F-mjera* s 56,42% na 60,01%.

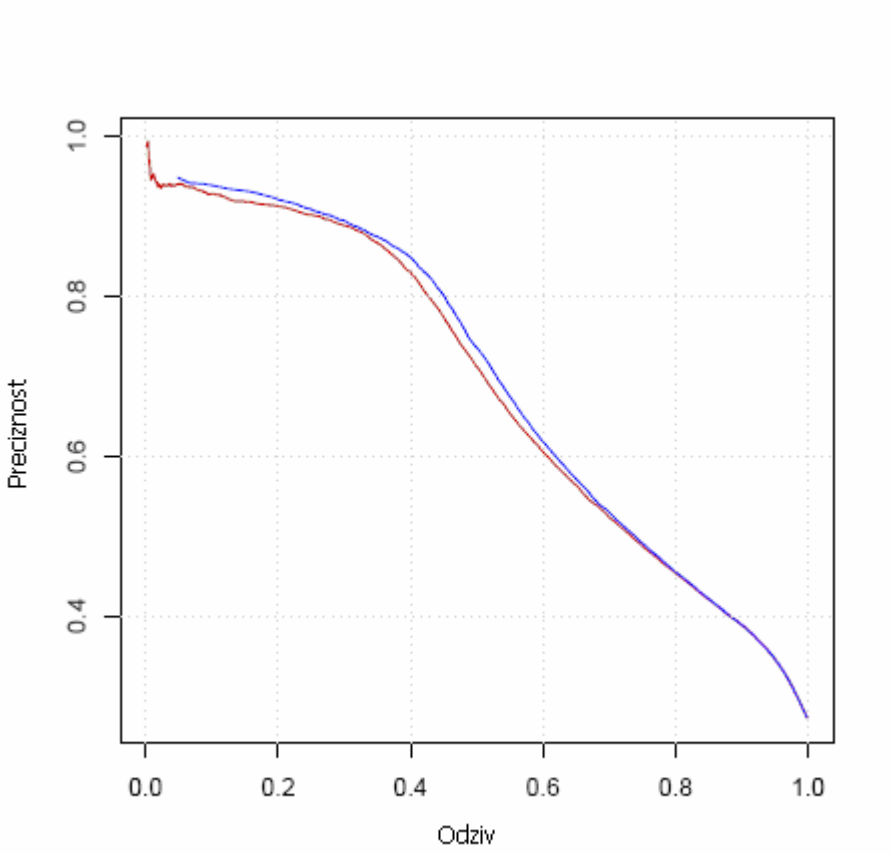
Tablica 5.3 Evaluacija rezultata za vrijednosti praga od 1 do 5 mjesta kontakta unutar 9 ostataka za omjer težine (težina pozitivne klase: težina negativne klase) 3:2

Prag	Preciznost	Odziv	F-mjera	Točnost	AUC
1	78,58	44,01	56,42	81,60	81,85
2	79,10	43,36	56,01	81,83	81,63
3	79,94	42,43	55,43	82,34	81,09
4	80,89	40,31	53,81	83,20	79,61
5	81,83	37,65	51,58	84,65	77,32
OR	71,82	51,54	60,01	81,40	82,16

Na slici 5.2 su prikazane *PR* krivulje u slučaju kad su težine jednake i kada je omjer težine pozitivne klase i negativne 3:2. Prema krivuljama je za uočiti da pridružene različite težine ne pridonose poboljšanju. Međutim, slika 5.3 pokazuje poboljšanje rezultata upotrebom *OR1* metode nad izlazima klasifikatora za pragove 1 do 5. Krivulja *preciznost-odziv* koja predstavlja primjenu *OR1* metode je iznad krivulje za prag 1.



Slika 5.2 Krivulja *preciznost-odziv* za jednaki omjer težina za prag 1 (crvena) i prag 1 s omjerom težina 3:2, pozitivna klasa : negativna klasa (plava)



Slika 5.3 Krivlja *preciznost-odziv* za prag 1 (crvena), prag 1 i primjenu *OR1* metode (plava) pri odnosu težina pozitivne i negativne klase 3:2

5.2 Rezultati predviđanja koristeći informacije iz sekvence, profila sekvence i strukture

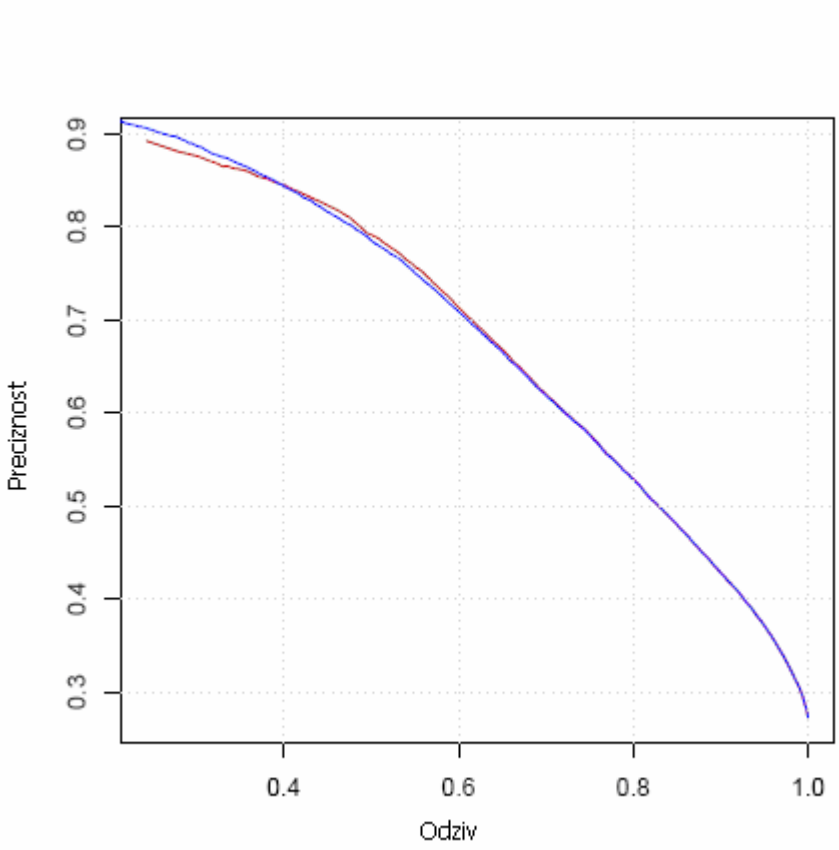
Informacijama iz sekvence i profila sekvence dodane su strukturne informacije. Strukturne informacije korištene u radu su ASA, relativna ASA, ASA okosnice, relativna ASA okosnice, ASA bočnog ogranka, relativna ASA bočnog lanca, nepolarna ASA, relativna nepolarna ASA, polarna ASA, relativna polarna ASA, srednja, maksimalna i minimalna vrijednost ukopanosti, srednja, maksimalna i minimalna vrijednost izbočenosti, srednja vrijednost hidrofobnosti te elektro - ionski interakcijski potencijal (EIIP). Kao klasifikator korištena je metoda slučajnih šuma tj. njezina inačica PARF s 200 stabala i 14 varijabli za grananje. Za evaluaciju rezultata korištena je uobičajena *oob.* Kao i u prethodnom slučaju za definiranje pozitivne i negativne klase kod matrice greške uziman je prag od 0,5.

Tablica 5.4 prikazuje rezultate za različite vrijednosti praga, te rezultat *OR1* metode koja kombinira rezultate klasifikatora za vrijednosti praga od 1 do 8.

Gledajući rezultate u tablici za uočiti je da je za vrijednost praga 1 *preciznost* 82,20% uz *odziv* 44,11% te predstavljaju najbolje vrijednosti u odnosu na ostale vrijednosti praga. Međutim za prag 4 se vidi najveća vrijednost *F-mjere* u odnosu na ostale vrijednosti praga. Ako se promatra samo prag 1 i primjena *OR1* metode, vrijednost *F-mjere* je s 57,42% porasla na 65,05%. Također je primjena *OR1* metode pridonijela porastu *odziva* s 44,11% na 51,38% dok *preciznost* pada s 82,20% na 78,34%.

Tablica 5.4 Evaluacija rezultata za predviđanje mjesta interakcije iz sekvenci, profila sekvenci i strukture za različite vrijednosti praga te rezultat dobiven *OR1* metodom. U tablici su prikazani *preciznost*, *odziv*, *F-mjera, točnost* i *AUC*.

Prag	Preciznost	Odziv	F-mjera	Točnost	AUC
1	82,20	44,11	57,42	82,29	85,52
2	82,11	43,92	57,23	82,48	85,23
3	82,18	43,73	57,09	82,98	84,74
4	81,86	42,82	63,23	83,81	83,34
5	81,56	41,36	54,89	85,24	81,38
6	81,08	36,69	52,39	87,60	78,15
7	80,21	35,52	49,24	90,44	74,58
8	81,69	28,30	42,04	93,34	71,16
OR	78,34	51,38	62,05	82,99	85,50



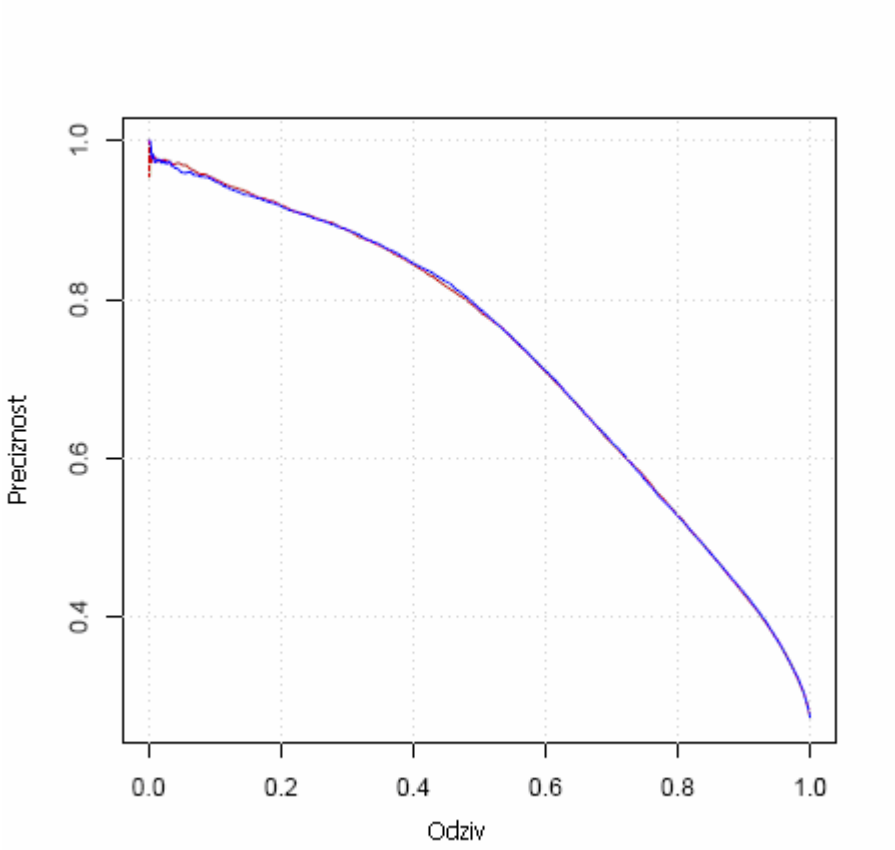
Slika 5.4 Krivlja *preciznost-odziv* za prag 1 (plava), prag 1 i primjenu *OR1* metode (crvena)

Tokom ispitivanja s različitim težinama pridruženih pozitivnoj i negativnoj klasi, ispostavilo se prema *AUC* kriteriju da je optimalan izbor težina 3 za pozitivnu klasu i 2 za negativnu klasu. Također se nije išlo dalje od vrijednosti praga 5 za klasifikatore s klasama različitih težina. U tablici 5.5 su prikazani rezultati za svaki klasifikator s pragom 1 do 5 te primjena *OR1* metode. *OR1*metoda je ponovno poboljšala vrijednosti *odziva, F-mjere* i *AUC*-a. Primjenom metode *odziv* je 64,38% uz *preciznost* od 68,05%, a *F-mjera* 66,16%. U slučaju klasifikatora s vrijednosti praga 1, i *odziv* i *F-mjera* imaju manje vrijednosti. Poboljšanje se također može uočiti na krivuljama *preciznost-odziv* na slici 5.6.

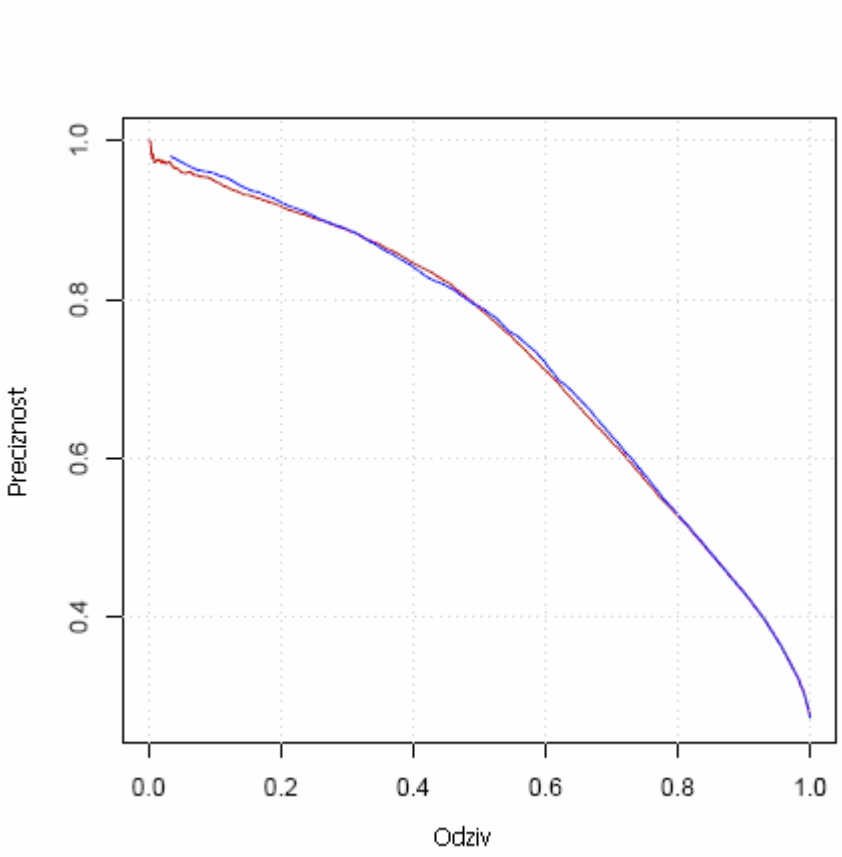
Tablica 5.5 Evaluacija rezultata za vrijednosti praga od 1 do 5 mjesta kontakta unutar 9 ostataka za omjer težine (težina pozitivne klase: težina negativne klase) 3:2

Prag	Preciznost	Odziv	F-mjera	Točnost	AUC
1	73,99	56,47	64,05	82,84	85,56
2	73,91	56,31	63,92	83,04	85,28
3	74,60	54,60	63,05	84,46	84,58
4	74,08	55,70	63,59	83,49	83,32
5	74,41	52,29	61,42	85,74	81,37
OR	68,05	64,38	66,16	82,17	85,68

Na slici 5.5 su prikazane *preciznost-odziv* krivulje u slučaju kad su težine jednake i kada je omjer težine pozitivne klase i negativne 3:2. Prema krivuljama je za uočiti da pridružene različite težine ne pridonose bitnom poboljšanju, međutim primjena *OR1* metode ponovno poboljšava rezultate, slika 5.6.

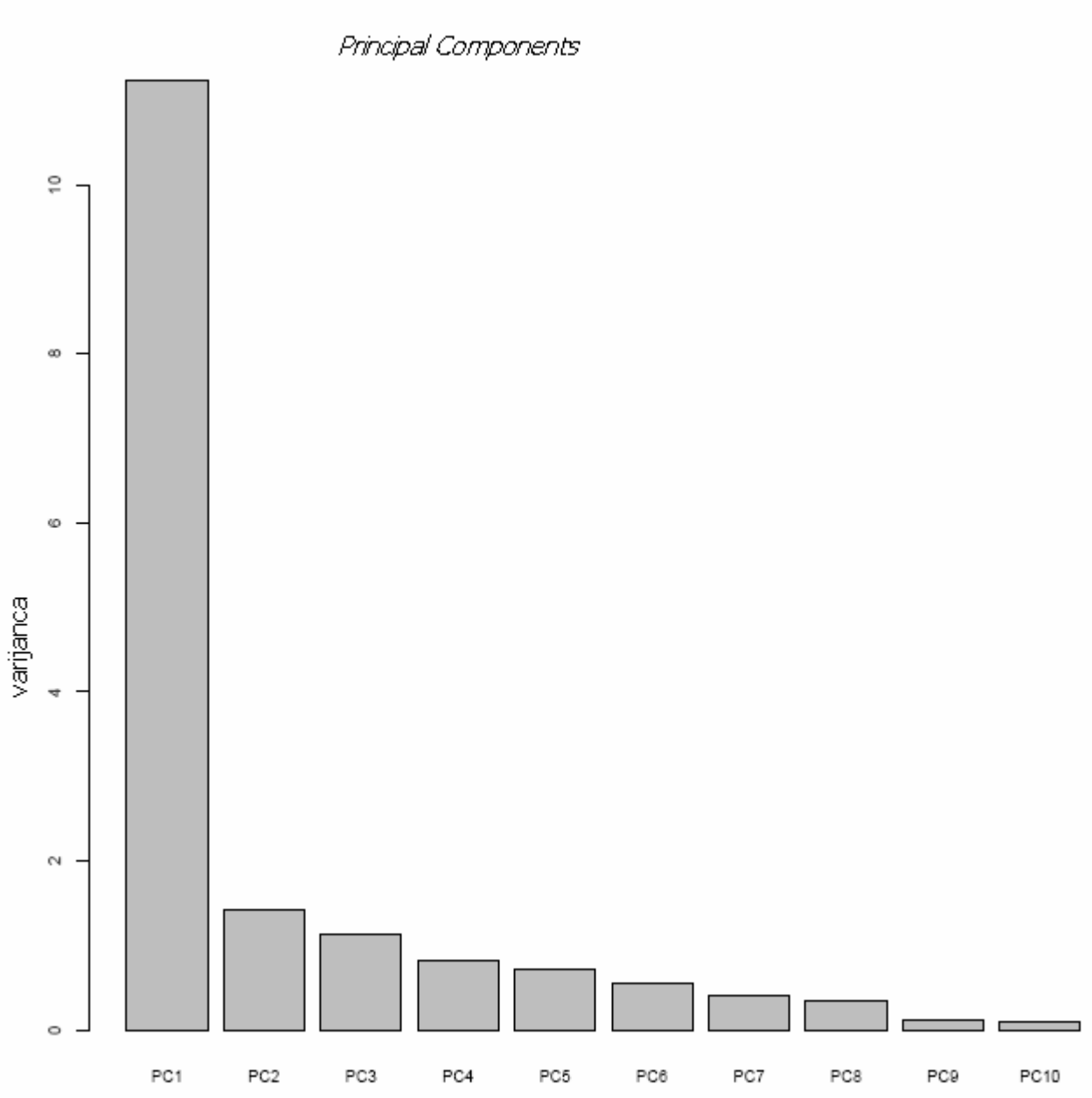


Slika 5.5 Krivulja *preciznost-odziv*: prag 1 s težinama 1:1 (crvena) i prag 1 s težinama 3:2 (plava)



Slika 5.6 Krivlja *preciznost-odziv*: prag 1 (crvena), prag 1 i primjenu *OR1* metode (plava) pri odnosu težina pozitivne i negativne klase 3:2

S obzirom na veliki broj strukturnih svojstava moguće je reducirati svojstva na ona značajnija. Svojstva su birana na dva načina. Prvi način je *z-skor* važnosti svojstava u sklopu metode slučajne šume. Tako su slučajne šume u *R* alatu kao dominantne strukturne atribute odredile srednju vrijednost ukopanosti, ASA-u, relativnu ASA-u, ASA-u bočnog ogranka te nepolarnu ASA-u. Drugi način određivanja dominantnih strukturnih atributa jest metoda *PCA* opisana u poglavlju 3.3. Kao rezultat, metoda daje komponente tj. vektore s najvećom svojstvenom vrijednosti, a one su linearna kombinacija svih atributa u prostoru. Značajnost pojedinog atributa može se vidjeti promatrajući njihove koeficijente u svakoj od komponenata. Tako se u tablici 5.6 mogu pogledati koeficijenti svakog atributa u prostoru koji čine *prvu* komponentu tj. onu s najvećom varijancom u histogramu na slici 5.7. Ova komponenta je najznačajnija i najjasnije pokazuje koji su atributi dominanti. Ostale komponente su ovdje zanemarene s obzirom da su koeficijenti atributa koji u biti grade svaku komponentu, relativno maleni s obzirom na koeficijetne atributa u *prvoj* komponenti, a samim time su i njihove varijance manje u odnosu na varijancu prve komponente.



Slika 5.7 Histogram 10 značajnijih komponenta

U tablici 5.6 navedeni su koeficijenti atributa *prve* komponente. Gotovo svi atributi se mogu smatrati približno značajnim s obzirom na vrijednosti koeficijenata. Međutim, potrebno je izdvojiti nekoliko značajnijih. U odabiru svojstava odnosno atributa, nije se išlo po sistemu odabira prvih pet s najvećim koeficijentima. U nekoliko se ispitivanja i kombinacijom svojstava pokazalo da nije nužno izabrati prvih pet s najvećim koeficijentima. Stoga su se s obzirom na rezultate tj. promatranjem *PR* krivulja, kao dominantni strukturni atributi uzeli sljedeći: relativna ASA, relativna nepolarna ASA, nepolarna ASA, ASA te srednja vrijednost ukopanosti.

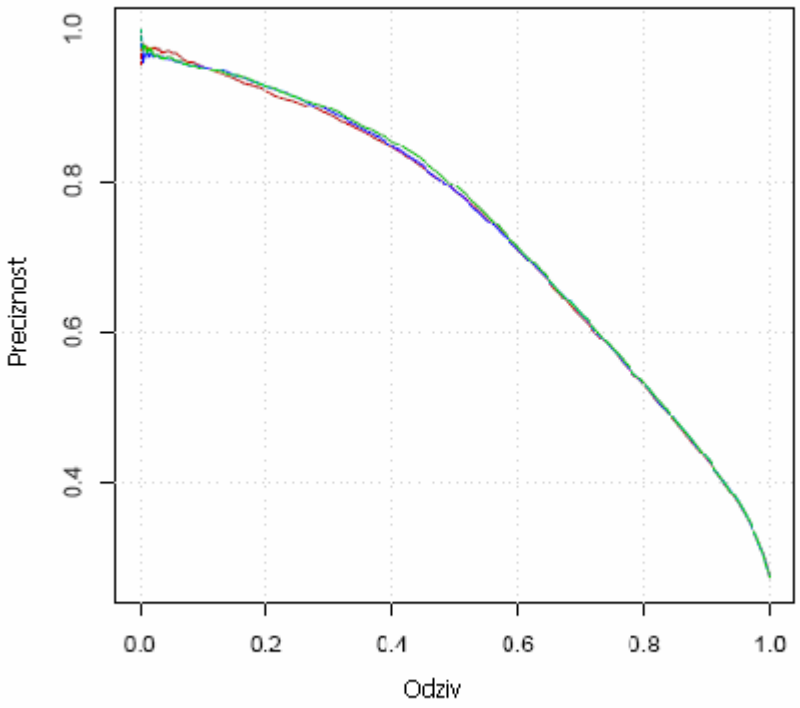
Tablica 5.6 Koeficijenti atributa koji čine linearnu kombinaciju gradeći *prvu* komponentu poredani po apsolutnoj vrijednosti koreliranosti atributa

Strukturna svojstva	Koeficijenti u komponenti
relativna ASA	0,2938971
površina dostupna otapalu (ASA)	0,2909842
srednja izbočenost	0,2801067
relativna nepolarna ASA	0,2783763
relativna ASA bočnog ogranka	0,2748286
ASA bočnog ogranka	0,2690842
relativna polarna ASA	0,2657842
nepolarna ASA	0,2652657
srednja ukopanost	-0,2472864
relativna ASA okosnice	0,2470808
polarna ASA	0,2461295
ASA okosnice	0,2340430
maksimalna izbočenost	0,2229732
maksimalna ukopanost	-0,2130318
minimalna izbočenost	0,1915919
hidrofobnost	-0,1212144

Promatrajući rezultate u tablici 5.7 vidi se da je došlo do poboljšanja u *odzivu*, *F-mjeri*, *točnosti* i u vrijednosti *AUC*-a u slučaju kada su se uzimala samo ona strukturna svojstva koja su odabrana putem metode *PCA*. Stoga bi se moglo reći da nije potrebno uzimati sva strukturna svojstva za predikciju već ona odabrana metodom *PCA* pošto je takva kombinacija svojstva dala bolje rezultate u odnosu na *RF* metodu *z-skor*. Međutim, *OR1* metoda za omjer težina 3:2 i kada se koriste sva strukturna svojstva, ipak daje bolje rezultate.

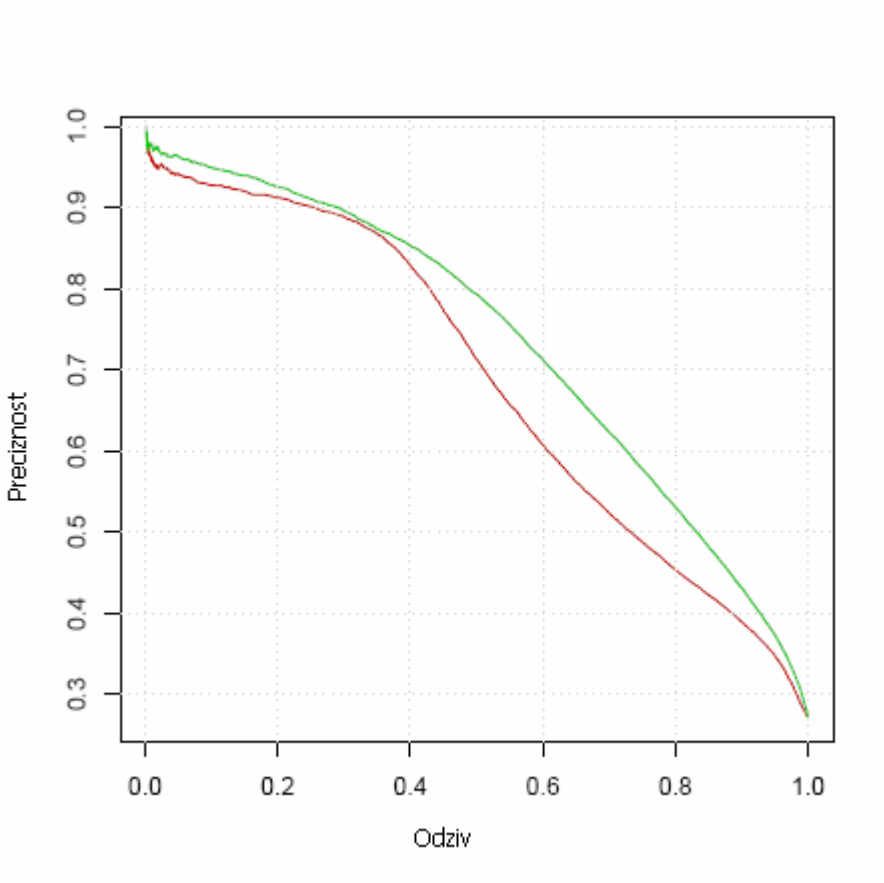
Tablica 5.7 Evaluacija rezultata za predviđanje mjesta interakcije iz sekvenci, profila sekvenci i strukture[1] te rezultati dobiveni iz informacija o sekvenci, profilu sekvenci i reduciranih strukturnih atributa putem *RF* metode[2] i iz informacija o sekvenci, profila sekvenci i reduciranih strukturnih atributa putem *PCA* metode[3] . U tablici su prikazani *preciznost*, *odziv*, *F-mjera, točnost* i *AUC*.

Prag = 1	Preciznost	Odziv	F-mjera	Točnost	AUC
svi atributi[1]	82,20	44,11	57,42	82,29	85,52
RF[2]	81,29	45,79	58,58	82,47	85,56
PCA[3]	81,97	46,01	58,94	82,65	85,62



Slika 5.8 Krivulja *preciznost-odziv*: informacije iz sekvence, profila sekvence i svi strukturni atributi (crvena); informacije iz sekvence, profila sekvence i reducirane informacije iz strukture metodom *RF* (plava); informacije iz sekvence, profila sekvence i reducirane informacije iz strukture metodom *PCA* (zelena).

Na slici 5.9 prikazan je odnos *preciznost-odziv* krivulja za slučajeve kada su svojstvima sekvence i profila dodane reducirane informacije o strukturi i onda kada nisu, odnosno predikcija se vrši samo na temelju sekvence i profila. Na slici je prikazan rezultat za vrijednost praga 1. Tako se vidi da je dodavanje strukturnih informacija bitno pridonijelo poboljšanju izgleda *PR* krivulje.

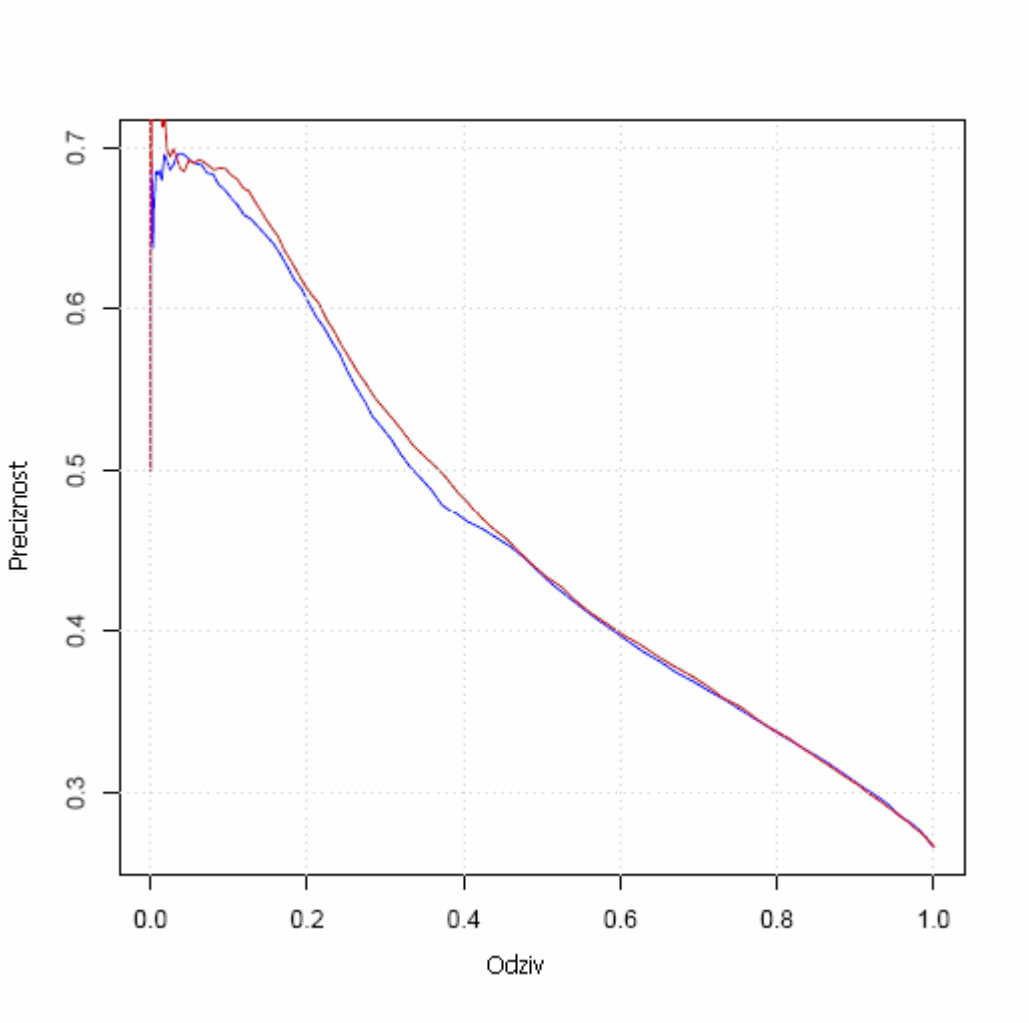


Slika 5.9 Krivulja *preciznost-odziv* za vrijednost praga 1: predikcija na osnovu sekvence i profila (crvena) i predikcija na osnovu sekvence, profila i reduciranih strukturnih informacija metodom *PCA* (zelena)

5.3 Utjecaj strukturne informacije – elektrostatički potencijal

Skupini od 124 864 instanci dodana je osim spomenutih strukturnih svojstava i svojstvo elektrostatski potencijal opisan u poglavlju 3.2.1. Cilj je bio vidjeti da li ovo svojstvo pomaže u predikciji odnosno da li na neki način utječe na prepoznavanje mjesta interakcije.
Za predikciju se koristila metoda slučajnih šuma u sklopu alata *R.*. Pri klasifikaciji je korišteno 200 stabala i 14 varijabli grananja.
Treba napomenuti da se u svim do sada navedenim predikcijama koristila 170 001 instanca, dok se u predikciji s dodanim elektrostatskim potencijalom kao svojstvo predviđanja, koristilo njih 124 864. Razlog tomu je što se u procesu računanja elektrostatičkog potencijala, neke ulazne PDB datoteke zbog svoje nekonzistentnosti nisu mogle obraditi, a samim time niti izračunati potencijal za svaki ostatak u lancu.

Slika 5.10 prikazuje odnos *preciznost-odziv* krivulja. Prema slici se može vidjeti da prisustvo atributa potencijal i nije pridonio predikciji, dapače. *AUC* vrijednost također pokazuje da dodavanjem potencijala nije postignuta bolja predikcija. Tako je *AUC* u slučaju kada se u predikciji koristi i svojstvo potencijal 68,65% dok je u predikciji bez potencijala 69,04%.



Slika 5.10 Krivulje *preciznost-odziv* za prag 1: bez potencijala (crvena) i s potencijalom (plava)

6 Diskusija i zaključak

Cilj rada bio je pokušati dodati neka nova svojstva u predikciji mjesta interakcije proteina te vidjeti da li su se rezultati predikcije poboljšali u odnosu na neke ranije radove.

Na osnovu sekvence od 9 ostataka i profila za svaki ostatak u prozoru vršila su se predviđanja metodom slučajnih šuma, odnosno njezine paralelne implementacije *PARF* te *OR1* metodom. Kasnije su se setu instanci s navedenim svojstvima dodale strukturne informacije i izdvojila ona strukturna svojstva koja su značajnija. Između ostalog dodao se i potencijal kao svojstvo predikcije, međutim time nije postignuto nikakvo poboljšanje.

Korištenjem samo sekvence od 9 ostataka i profila za svaki ostatak pri tome čineći vektor od 189 atributa te uz prag 1, odnosno uvjet da je središnji ostatak u kontaktu, postigla se *preciznost* od 86,80% uz *odziv* od 35,21%. Primjenom *OR1* metode poboljšala se vrijednost *F-mjere* s 50,10% na 53,65%, te se povećao *odziv* s 35,21% na 39,23% uz *preciznost* od 84,87%.

Ako se klasama pridijele različite težine dobiju se drugačiji rezultati predikcije. Tokom ispitivanja rezultata za različite težine klasa, kriterij za odabir najpovoljnijeg omjera težina jest *AUC.* Tako se uspostavilo da je klasifikator u slučaju kada se pozitivnoj klasi pridružila težina 3, a negativnoj težina 2, dao najbolji rezultat za *AUC* u odnosu na neke druge odnose težina. Stoga se odabrao odnos težina 3:2 za pragove klasifikacije 1 do 5. Omjer težina 3:2 nije bitno utjecao na poboljšanje *preciznosti* osim što se povećao *odziv,* ali i *F-mjeru.* Međutim, pridruživanje težine 3 pozitivnoj klasi i težine 2 negativnoj klasi daje bolje rezultate u odnosu na jednaki omjer težina tek kada se primjeni *OR1* metoda. Stoga, ako se promatra isključivo primjena *OR1* metode i kada je omjer težina pozitivne i negativne klase 3:2 postiže se *preciznost* od 71,82% uz *odziv* 51,54%, a *F-mjera* 60,01%. Dobiveni rezultati su bolji od slučaja jednakih težina i primjene *OR1* metode nad klasifikatorima s jednakim težinama klasa. Ovim rezultatima *OR1* metoda je potvrdila svoj doprinos porastu *odziva* i *F-mjere,* dok je različit omjer težina pozitivne i negativne klase postigao poboljšanje vrijednosti *odziva* i *F-mjere.*

U slučaju kada se za predikciju koriste i nereducirane strukturne informacije postignuta je *preciznost* 82,20% uz *odziv* od 44,11%. Primjena *OR1* metode pridonijela je povećanju *F-mjere* s 57,42% na 62,05% i povećanje *odziva* s 44,11% na 51,38%. Ako se uspoređuje slučaj za prag vrijednosti 1, kada se koriste sve strukturne informacije s onim slučajem u kojemu se predikcija vrši samo s dominantnim strukturnim svojstvima odabranim *PCA* metodom, došlo je do laganog poboljšanja. Prema rezultatima, reduciranje strukturnih svojstava je poboljšalo *odziv* i *F-mjeru*,

ali *OR1* metoda i dalje daje bolje vrijednosti za *odziv* i *F-mjeru* u slučaju korištenja svih strukturnih informacija.

I u ovom slučaju predikcije na osnovu sekvence, profila sekvence i svih strukturnih svojstava ponovno je odabran omjer pozitivne i negativne klase 3:2. Slika 5.5 pokazuje da i nije došlo do poboljšanja primjene različitih težina pošto se krivulje gotovo poklapaju osim u nekim dijelovima gdje je krivulja u slučaju omjera 3:2 nešto iznad krivulje za slučaj 1:1.

Međutim *OR1* metodom vrijednosti *odziva* i *F-mjere* ponovno rastu. Za *OR1* metodu i omjer težina 3:2 postignuta je *preciznost* 68,05% uz *odziv* od 64,38% i *F-mjera* od 66,16%. Te su vrijednosti veće od rezultata u slučaju omjera težina 1:1 i 3:2 za prag 1, kao i od rezultata kada se primjeni *OR1* za omjer jednakih težina.

Općenito *OR1* metoda poboljšava *odziv* i *F-mjeru*, a smanjuje *preciznost.* S time da je dobitak u porastu *odziva* i *F-mjere* veći od pada *preciznosti.* U slučaju pridjeljivanja težine 3 pozitivnoj klasi i težine 2 negativnoj klasi rezultati su još bolji.

7 Literatura

[1] Y. Ofran and B. Rost, "Predicted protein-protein interaction sites from local sequence information," *FEBS Lett*, vol. 544, pp. 236-9, Jun 5 2003.

[2] I. Res, I. Mihalek, and O. Lichtarge, "An evolution based classifier for prediction of protein interfaces without using protein structures," *Bioinformatics*, vol. 21, pp. 2496-501, May 15 2005.

[3] Šikić M., Računalna metoda za predviđanje mjesta proteinskih interakcija, doktorska disertacija, 2008

[4] A.Koike and T. Takagi, "Prediction od protein-protein interaction sites using support vector machines," *Protein Eng Des Sel,* vol.17, pp. 165-73, Feb 2004.

[5] Y. Ofran and B. Rost, "ISIS: interaction sites identigied from sequence," *Bioinformatics,* vol. 23, pp. el3-6, Jan 15 2007.

[6] C. Yan, D. Dobbs, and V. Honavar, "A two stage classifier for identification od protein-protein interface residues," *Bioinformatics,* vol.20 Suppl 1,pp.i371-8, Aug 4 2004.

[7] B. Wang, P. Chen, D.S. Huang, J.J.Li, T. M. Lok, and M. R. Lyu, "Predicting protein interaction sites from residue spatial sequence profile and evolution rate," *FEBS Lett,* vol.580, pp.380-4, Jan 23 2006.

[8] 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.

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

[10] 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.

[11] "The universal protein resource (UniProt)," *Nucleic Acids Res*, vol. 36, pp. D190-5, Jan 2008.

[12] ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA

[13] P.P. Vaidyanathan, "Genomics and proteomics: a signal processor's tour," *Circuits and Systems Magazine*, IEEE, vol. 4, pp. 6-29, 2004

[14] Todd J. Dolinsky, Jens E. Nielsen, J. Andree McCommon and Nathan A. Baker, "PDB2PQR: an automated pipeline for the setup od Poisson-Boltzmann electrostatics calculationc," *Nucleic Acids Reasearch*, vol. 32, Web Server issue © Oxford University Press 2004.

[15] Weiner,S.J., Kollman,P.A., Case,D.A., Singh,U.C., Ghio,C., Alagona,G., Profeta,S.Jr. andWeiner,P. (1984) Anewforce field for molecular mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc.*, 106, 765–784.

[16] J.R. Bradford and D.R. Westhead, "Improved predisction of protein-protein binding sites using support vector machines approach," *Bioinformatics*, vol. 21 pp 1487-94, 8 2005.

[17] I. H. Witten and E. Frank, Data Mining: "Practical machine learning tools and techniques," 2nd. ed. San Francisco: Morgan Kaufmann, 2005.

[18] I. Cosic, "Macromolecular bioactivity: is it resonant interaction between macromolecules? Theory and applications," *IEEE Trans Biomed Eng*, vol. 41, pp. 1101-14, Dec 1994.

[19] Michael Gribskov, Andrew D. McLanchlan and David Eisenberg, "Profile analysis: Detection of distantly related proteins," *Proc. Natl. Acad. Sci. USA*, vol. 84, pp 4355-4358, July 1987.

[20] http://www.ncbi.nlm.nih.gov/blast/fasta.shtml

[21] 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

[22] Smith, T.F. and Waterman, M.S. (1981) j.Mol.Biol., 147, 195-197

[23] R.Luthy, I. Xenarios nd P. Bucher "Improving the sensitivity of the sequence profile method," *Protein Science*, vol 3, Issue 1 139-146, 1994

[24] J. Davis, M. Goadrich: "The relationship between Precision-Recall and ROC curves," *ACM International Conference Proceeding Series*; vol 148, pp 233 -240, 2006.

[25] www.irb.hr/hr/research/projects/it/2004/2004-111/

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

Sažetak

U okviru ovog rada vršila se predikacija mjesta proteinske interakcije na osnovu sekvence, profila sekvence i strukturnih svojstava. Mjesto interakcije se definira kao prozor od 9 ostataka u kojemu je barem središnji aminokiselinski ostatak mjesto kontakta te su još 4 ostatka u kontaktu, a da su udaljeni od središnjeg ne više od 3 ostatka. Za ostatak se smatra da je u kontaktu ako je njegov atom udaljen od najbližeg atoma susjednog lanca manje od 6 Angstrema. Na skupu od 1137 lanaca vršila se predikacija alatom PARF, paralelnom implementacijom slučajnih šuma. Provjera rezultata izvršena je u sklopu alata PARF i *OR* metodom. Za predviđanje mjesta interakcije na temelju sekvence, profila sekvence i strukture korištena su sljedeća strukturna svojstva: površina dostupna otapalu (ASA) , relativna ASA, ASA okosnice, relativna ASA okosnice, ASA bočnog ogranka, relativna ASA bočnog ogranka, nepolarna ASA, relativna nepolarna ASA, polarna ASA, relativna polarna ASA, srednja, maksimalna i minimalna vrijednost ukopanosti, srednja, maksimalna i minimalna vrijednost izbočenosti, srednja vrijednost hidrofobnosti te elektro-ionski interakcijski potencijal (EIIP). Metodom *PCA* odabrana su sljedeća dominantna strukturna svojstva: relativna ASA, relativna nepolarna ASA, nepolarna ASA, ASA te srednja vrijednost ukopanosti. Za dodatno strukturno svojstvo elektrostatski potencijal, pokazano je da nema utjecaja na rezultate predikcije.

Za predikciju na osnovu sekvenci i profila sekvenci dobili su se bolji rezultati u odnosu na predikciju na osnovu samo sekvenci. Dodatak profila kao svojstvo pokazalo se uspješnim u predviđanju mjesta interakcije. Rezultati predikcije na temelju sekvence, profila sekvence i strukturnih informacija također su se pokazali boljima u odnosu na predviđanje na temelju sekvence i strukture. I u ovome slučaju profil kao svojstvo ostataka je poboljšao rezultate u odnosu na radove u kojima se profili ne koriste kao svojstvo.

Dodatno poboljšanje rezultata postignuto je pridruživanjem težine 3 pozitivnoj klasi i težine 2 negativnoj klasi te primjena *OR1* metode nad klasifikatorom s vrijednošću praga 1 i njegovim klasifikatorima podskupovima s vrijednostima praga 2 do 5.