SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 1173

PREDVIĐANJE MJESTA PROTEINSKIH INTERAKCIJA IZ SLIJEDA AMINOKISELINSKIH OSTATAKA I RELATIVNE POVRŠINE DOSTUPNE OTAPALU

Ana Pejaković

Zagreb, ožujak 2009.

Hvala mojim Pejakovićima, splitskim i zagrebačkim prijateljima i Neni,
jer mi čine život lijepim!
Hvala Mili Šikiću koji mi je pružio zadovoljstvo stvaranja ovog rada.

Sadržaj

1 Uvod

Veliki broj staničnih funkcija obavljan je posredstvom proteinskih interakcija. Stoga je razumijevanje takvih interakcija vrlo značajno pitanje prvenstveno u razvoju novih lijekova, cjepiva, analizi metaboličkih reakcija, promatranju i praćenju razvoja organizma te u mnogim drugim područjima.

Broj mogućih proteinskih interakcija je toliko velik da su zabilježeni primjeri pronađenih i definiranih proteinskih kompleksa neznatni naspram tog broja te je razumijevanje nepreglednog opsega mogućih proteinskih interakcija i uočavanje skupova pravila i uzoraka ponašanja proteinskih struktura vrlo zahtjevan problem. Ideja je na osnovu slijeda aminokiselina koje grade protein odrediti trodimenzionalnu strukturu proteina, zatim njegova svojstva, ponašanje u različitim medijima, i konačno prepoznati, odnosno predvidjeti, mjesta proteinskih interakcija i moguće interakcijske parove.

Eksperimentalne metode su se pokazale presporima i nedovoljno točnima, pa su se znanstveni timovi bioinformatičara i računaraca, u potrazi za boljim rješenjem, posvetili razvijanju brojnih metoda predviđanja mjesta interakcija na proteinima. Te su metode još u fazi razvoja i postignuti rezultati nisu dovoljno točni, ali neizbježnim širenjem baza podataka koje sadrže informacije o utvrđenim interakcijama i zabilježenim biofizičkim karakteristikama proteinskih kompleksa povećat će se i točnost rezultata postojećih i nadolazećih metoda.

Jedan od mogućih pristupa predviđanja mjesta interakcija proteina temelji se na poznavanju slijeda aminokiselinskih ostataka.

Cilj je uočiti koja sve svojstva koja je moguće predvidjeti ili izračunati iz slijeda aminokiselinskih ostataka imaju utjecaja na činjenicu da je neko mjesto na proteinu mjesto interakcije. Pokazalo se da profili slijeda aminokiselinskih ostataka daju jako dobre rezultate u predviđanju mjesta interakcije. Također se pokazalo se da je iz profila slijeda aminokiselinskih ostataka moguće jako dobro predvidjeti površinu dostupnu otapalu. Budući da se mjesta interakcije proteina

(engl. *active sites*) gotovo uvijek nalaze na njihovoj površini, za pretpostaviti je da je poznavanje izloženosti ostatka važna karika u prepoznavanju mjesta interakcije.

Ovaj rad se bavi jednom od metoda predviđanja mjesta proteinskih interakcija korištenjem profila slijeda aminokiselinskih ostataka te predviđene relativne površine dostupne otapalu. Naglasak rada je na sljedećem pitanju: je li moguće poboljšati rezultate predviđanja mjesta interakcije dobivene korištenjem profila slijeda aminokiselinskih ostataka uvođenjem nekog dodatnog atributa koji opisuje relativnu površinu dostupnu otapalu, a predviđen je korištenjem tih istih profila slijeda aminokiselinskih ostataka?

2 Teorijski uvod

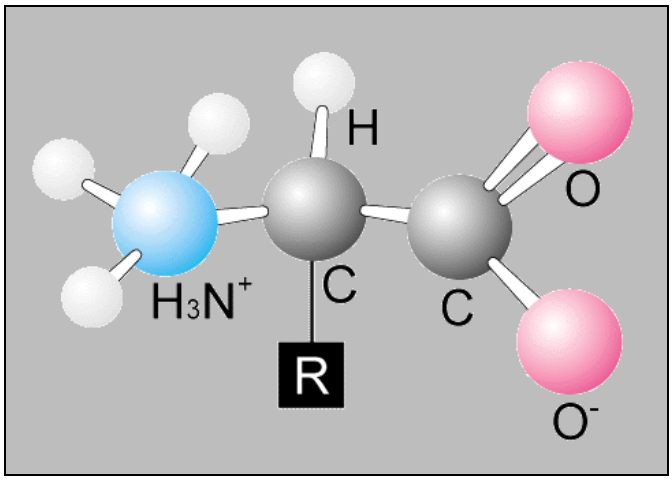
2.1 Proteini

Proteini su najzastupljenije biološke makromolekule u živom svijetu. Kemijski gledano proteini su polimeri sastavljeni od aminokiselina povezanih peptidnim vezama u strukture od dvije do nekoliko stotina tisuća aminokiselinskih ostataka.

Kako je razumijevanje strukture proteina ključno za razumijevanje problema predviđanja mjesta interakcije, ovdje će biti dan kratak opis njihove strukture i formiranja.

2.1.1 Građa proteina

Aminokiseline su osnovne građevne jedinice proteina. Kemijski su građene od dvije karakteristične funkcionalne skupine: amino skupine i karboksilne skupine te R skupine po kojoj se međusobno razlikuju. Općenita formula aminokiseline je *NH2-CHR-COOH*. R skupine ili tzv. bočni ogranci aminokiselina (engl. *side chain*) određuju svojstva i funkciju proteina.



Slika 2.1 Općenita struktura aminokiselina

Dvadeset je poznatih aminokiselina koje se pojavljuju u proteinima. Njihov pregled dan je u tablici 2.1. Osim imena aminokiselina, u tablici se nalazi zapis

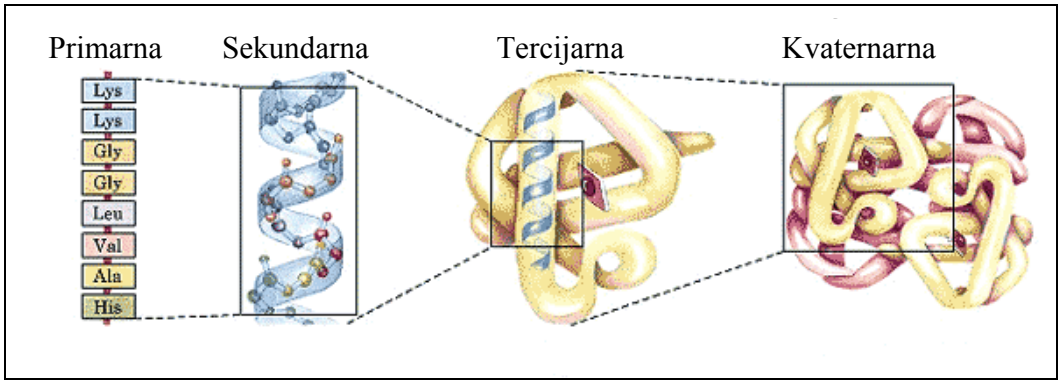
aminokiselina u *FASTA* formatu gdje svakoj od aminokiselina odgovara jedno slovo.

1	A	Ala	Alanin
2	C	Cys	Cistein
3	D	Asp	asparaginska kiselina
4	E	Glu	glutaminska kiselina
5	F	Phe	Fenilalanin
6	G	Gly	Glicin
7	H	His	Histidin
8	I	Ile	Isoleucin
9	K	Lys	Lizin
10	L	Leu	Leucim
11	M	Met	Metionin
12	N	Ans	Asparagin
13	P	Pro	Prolin
14	Q	Gln	Glutamin
15	R	Arg	Arginin
16	S	Ser	Serin
17	T	Thr	Treonin
18	V	Val	Valin
19	W	Trp	Triptofan
20	Y	Tyr	Tirozin

Tablica 2.1 Pregled aminokiselina

U različitom broju i omjerima su navedene aminokiseline ulančano povezane u lance – polipeptide. Veza nastaje između karboksilne i amino skupine aminokiseline pri čemu se izlučuje molekula vode i nastaje aminokiselinski ostatak. Spomenuti lanci aminokiselina zauzimaju određene oblike u prostoru pa se međusobno povezuju, preslaguju i tvore različite trodimenzionalne oblike. Budući da redoslijed aminokiselinskih ostataka određuje prostornu strukturu proteina, samim time određuje i njegovu funkciju.

Kako bi se lakše shvatila, struktura proteina može se promatrati u četiri stupnja: primarna, sekundarna, tercijarna i kvaternarna struktura (slika 2.2). U ovom radu lanci proteina će se promatrati kroz prizmu njihove primarne strukture.



Slika 2.2 Strukture proteina

2.1.2 Proteinske interakcije

Proteini i kompleksi proteina su trodimenzionalne makromolekule koje zauzimaju određene položaje u prostoru. Interakcijom dvaju proteina neki će slijedovi međudjelovati, a neki neće i dobiveni spoj će poprimiti neku nužno stabilnu formu. Kakvog će oblika taj produkt interakcije biti određuju sile koje postavljaju lance u neke poznate forme.

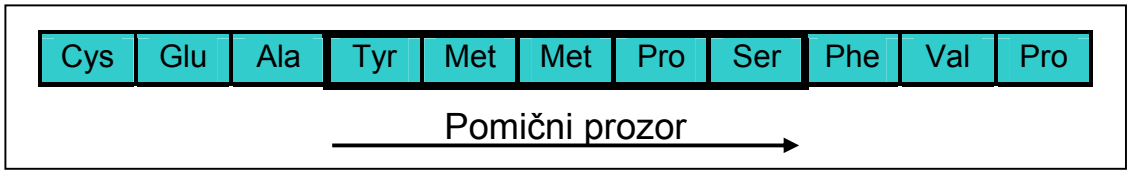
Veze koje najvećim dijelom određuju ove odnose su slabe nekovalentne veze između aminokiselinskih ostataka: hidrofobne, vodikove, ionske i Van der Waalsove.

Iako mnogo puta slabije od kovalentnih veza, ove su veze presudne u formiranju proteinskih interakcija jer djeluju istodobno u velikom broju, pa dobiveni protein čine dovoljno stabilnim.

Postoji više definicija mjesta proteinske interakcije. Jedna od njih, koja se ujedno koristi u ovom radu, zasniva se na promatranju slijed aminokiselinskih ostataka [6] i njihove evolucijske očuvanosti, profila slijeda [7].

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 [6]. Vrijednosti parametra n se obično kreću od 9 do 13 ostataka u prozoru, a m između 1 i 6.

U ovom se radu koristio prozor od 9 ostataka. Prozor predstavlja mjesto interakcije ukoliko je barem središnji ostatak mjesto kontakta. Za svaki se ostatak promatrao njegov profil i neki od atributa koji opisuju klasu relativne površine dostupne otapalu (engl. *Relative Accessible Surface Area, RASA*).



Slika 2.3 Pomični prozor duljine 5 aminokiselinskih ostataka

Aminokiselinski ostatak je, po definiciji korištenoj u ovome radu, u kontaktu ako je barem jedan atom tog ostatka udaljen manje od 6 Å (Angstrema) od najbližeg atoma susjednog proteinskog lanca.

Vektor ulaznih atributa čine prozor od 9 ostataka s pridruženim profilom slijeda i atributom klase RASA-e za svaki ostatak u prozoru. Definirani vektor atributa se proglašava pozitivnim ukoliko je barem središnji ostatak mjesto kontakta.

Profil promatranog 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.

Atributi klase RASA-e su veličine koje pomoću predviđenih klasa RASA-a za svaki aminokiselinski ostatak opisuju klasu RASA-e pomičnog prozora.

2.1.3 Proteinske baze podataka

U središtu zanimanja biologije i biokemije dugi niz godina 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 slijedu i strukturi podataka. Ti podaci danas su pohranjeni u bazama podataka od kojih su najvažnije PDB (*Protein Data Bank*) [12] i UniProt [15].

PDB 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 Dosadašnji rezultati

U dosadašnjim radovima autori su za predviđanje mjesta interakcije koristili ne samo informaciju iz slijeda već i strukturnu informaciju te njihovu kombinaciju.

Uspješnost najboljih dosadašnjih rezultata predviđanja na temelju samo informacije iz slijeda je preko 80% preciznosti za 40% odziva [1]. Donedavno je daleko najuspješniji rezultat predviđanja bio 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 [6] 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 kojemu je bilo koji teški atom (svi osim vodika) 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 [8] 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 kojima su teški atomi udaljeni manje od 5 Å. Profili slijeda ostataka dobiveni PSI-BLAST metodom su se koristili kao svojstvo.

Koristeći SVM pri klasifikaciji Reš i ostali [7] uzeli su kombinaciju evolucijske informacije i informaciju iz slijeda ostataka. Za mjesto kontakta definirali su one ostatke na površini čija se RASA (relativna površina dostupna otapalu) promijeni

nakon interakcije. Postigli su preciznost od 26% te odziv od 59%. Pritom su koristili prozor duljine 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 Ofran i Rost [9] postižu bolje rezultate tako da su predviđanju na osnovu slijeda dodali evolucijske profile, ASA-u i sekundarnu strukturu. U prvom koraku su predvidjeli ASA-u i sekundarnu strukturu iz slijeda, a u drugome su dodali te informacije slijedu. Kao klasifikator su, kao i u ranijem radu, koristili neuronske mreže. Postigli su preciznost između 60% i 70% uz odziv iznad 10%. Mjestom interakcije su definirali prozor kojemu 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 su u svom radu [10] 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 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 prvom su koraku za preciznost od 44% dobili odziv od 43%, a u drugom za preciznost od 58% su dobili odziv od 39%.

Wang i ostali su također promatrali isključivo ostatke na površini [11]. Kao svojstva su koristili profile slijeda i evolucijski omjer. Duljina prozora je 11 ostataka, a mjesto interakcije je prozor kojemu je središnji ostatak u kontaktu sa susjednim lancem. Ostatak je u kontaktu ako je udaljenost od njegovog α atoma ugljika do 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%.

M. Šikić je koristeći samo informacije iz slijeda uz preciznost od 60–70% postigao odziv od približno 40% [3]. Pritom je kao mjesto interakcije definirao

prozor kojemu je središnji ostatak u kontaktu kao i još barem 4 ostatka koja su od središnjeg udaljeni ne više od 3 ostatka. Ostatak je u kontaktu, ako je udaljen manje od 6 Å od ostatka koji pripada proteinu partneru.

V. Dragosavljević [1] je pod vodstvom M. Šikića radila na poboljšanju njegovih rezultata. S istim kriterijima i svojstvima, uz dodanu informaciju o profilu slijeda ostvarila je najbolje rezultate predviđanja dosad. Uz preciznost od 84,87% postigla je odziv od 39,23%.

S ciljem daljnjeg poboljšanja rezultata, i u ovom radu su se uzimali isti kriteriji i svojstva, osim što je dodana informacija o relativnoj površini dostupnoj otapalu kao jedno od svojstava predviđanja.

3 Podaci

3.1 Korišteni skupovi podataka

Za predviđanje mjesta proteinskih interakcija korišten je skup iz rada Ofrana i Rosta [6]. Skup se sastoji od 1137 lanaca 333 različita proteinska kompleksa. Nakon izbacivanja homolognih lanaca unutar pojedinih proteina, broj lanaca u skupu iznosi 833 te se kao takav koristio za prikaz različitih ocjena kvalitete predviđanja.

Osnovni podaci o strukturi proteina nalaze se u *PDB* formatu, tekstualnoj datoteci koja sadrži informacije o lancima proteina, prostornoj strukturi, koordinatama atoma i mnoge druge.

Spomenute PDB datoteke služe kao ulazni podaci iz kojih se određuju ASA vrijednosti ostataka primjenom *PSAIA* alata [13]. 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, pokreću se vlastoručno izrađene ili već gotove skripte napisane u Perl i Python programskim jezicima kojima se u izlazne XML datoteke dodaju profili slijeda te oznake klasa, a potom se podaci zapisuju u odgovarajuće *ARFF* datoteke.

3.2 Odabir svojstava za predviđanje

Za predviđanje RASA atributa koristila su se sljedeća svojstva:

- slijed od 9 aminokiselinskih ostataka
- profili slijeda svakog od ostataka unutar prozora

Kako svaka aminokiselina ima svoj profil slijeda, ukupan broj svojstava po elementu pomičnog prozora iznosi 21. Ukupan broj svojstava je stoga vektor dimenzije 21×9.

Za predviđanje mjesta interakcije koristila su se sljedeća svojstva:

- slijed od 9 aminokiselinskih ostataka

- profili slijeda svakog od ostataka unutar prozora
- predviđena klasa relativne površine dostupne otapalu središnjeg aminokiselinskog ostatka prozora
- predviđena klasa srednje vrijednosti RASA-a prozora
- izračunata klasa RASA prozora

3.2.1 Profil slijeda aminokiselinskih ostataka

Kao mjera evolucijske očuvanosti koriste se profili slijeda (engl. *probability profiles*). Profili slijeda su vjerojatnosti pronalaska bilo koje od dvadeset standardnih aminokiselina na onome mjestu u slijedu na kojemu se nalazi aminokiselina kojoj se profil određuje.

Uz informacije o strukturi svake aminokiseline, u XML datoteke dodane su informacije o profilu, pritom se koristeći alatom *PSI-BLAST* [14] koji je detaljno opisan u poglavlju 4.1.

Za svaki su se lanac izvukla imena aminokiselina koje ga čine, formirajući pritom niz slova (svako slovo se odnosi na odgovarajuću aminokiselinu) odnosno slijed. Takav bi se slijed propustio kroz PSI-BLAST koji bi kao rezultat dao PSSM matricu odnosno profil.

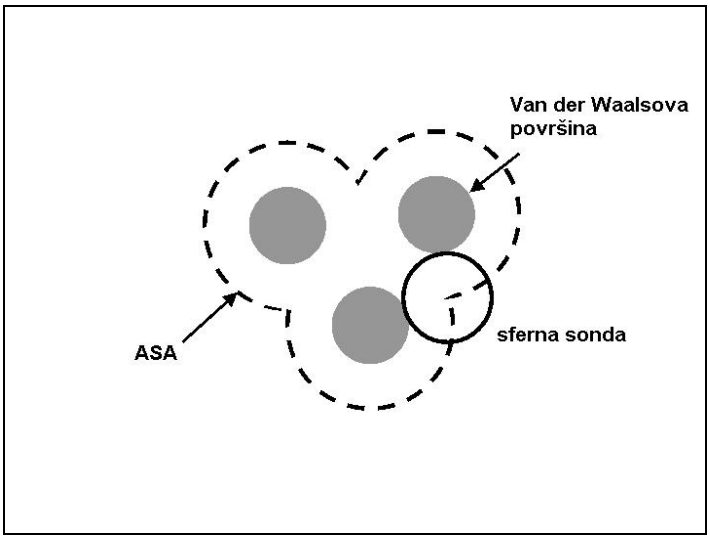
SWISS-PROT [15] je baza slijedova proteina u odnosu na koju se gradio profil [16]. Čitajući PSSM svakoj aminokiselini trenutnog lanca, koji se obrađuje, dodana je 20-dimenzionalna informacija, profil.

3.2.2 Relativna površina dostupna otapalu

Van der Waalsov radijus atoma je radijus imaginarne sfere koji se koristi za razne reprezentacije atoma. Ako svakome atomu pridružimo njegov van der Waalsov radijus, odnosno svaki atom zamijenimo sferom van der Waalsova radijusa, dobit ćemo van der Waalsovu površinu.

Otapalu dostupno područje površine (engl. *Accessible Surface Area, ASA*) može se definirati pomoću sferne sonde koja predstavlja model molekule otapala [19].

Sferna sonda kotrlja se po van der Waalsovoj površini te definira otapalu dostupno područje površine. Za radijus sonde se uzima obično iznos od 1,4 Å koji predstavlja radijus molekule vode.



Slika 3.1 Otapalu dostupno područje površine atoma

Za pojedinu aminokiselinu ASA vrijednost se dobije tako da se zbroje ASA vrijednosti svih atoma koji grade tu aminokiselinu. Osim ukupne ASA-e (engl. *Total*), postoje još četiri vrijednosti koje se često spominju pri proučavanju topologije površine aminokiseline:

a) ASA okosnice (engl. *Backbone*) – suma ASA-a svih atoma koji grade okosnicu aminokiseline

b) ASA bočnog ogranka (engl. *Side-chain*) – suma ASA-a svih atoma koji grade bočni ogranak aminokiseline

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

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

Često se umjesto ASA vrijednosti koristi njezina relativna vrijednost, RASA (engl. *Relative accessible surface area*), tj. odnos ASA vrijednosti ostatka i maksimalne ASA vrijednosti toga ostatka dok on nije sastavni dio proteina.

Pošto se aminokiseline nikad ne nalaze same u prostoru te vrijednosti su izračunate tako da se promatra aminokiselina okružena (po slijedu) s 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 postocima, je aminokiselina dostupna otapalu.

Predviđanje relativne površine dostupne otapalu može se obaviti na dva bitno različita načina. Kod predviđanja *regresijom* pokušava se odrediti točna RASA vrijednost za aminokiselinske ostatke, dok se *klasifikacijom* ostaci svrstavaju u predodređeni broj kategorija (klasa).

U ovom radu za predviđanje je korištena metoda klasifikacije pri čemu se broj kategorija kretao između dvije i pet.

Jedan od problema pri klasificiranju aminokiselinskih ostataka je izbor granica kategorija. Većina autora koristila je proizvoljne pragove, tj. granice kategorija po kojima su se vrijednosti raspoređivale po klasama.

Kao alternativa proizvoljnom izboru pragova u poglavlju 4.2 bit će objašnjeno kako se primjenom optimalnog kvantizatora mogu odrediti granice kategorija, a u poglavlju 5.3 će detaljno biti prikazana usporedba rezultata dobivenih koristeći pojedine metode odabira pragova i broja klasa.

Najveću pažnju pri analizi rezultata imat će klasifikacija u tri kategorije koristeći klasifikacijske pragove koje je odabrao Manesh za svoj skup [17].

3.2.2.1 Svojstva izvedena iz relativne površine dostupne otapalu

Svakoj XML datoteci su vlastito urađenim skriptama u Pythonu osim profila i iznosa RASA-e dodana i tri nova atributa koja na različite načine opisuju relativnu površinu dostupnu otapalu:

- *class_r_asa* – klasa RASA-e aminokiselinskog ostatka

- *class_mean_r_asa* – vrijednost dodijeljena klasi, određena kao srednja vrijednost svih vrijednosti koje se pojavljuju unutar neke klase
- *class_median_r_asa* – vrijednost dodijeljena klasi, određena kao središnja (medijan) vrijednost svih vrijednosti koje se pojavljuju unutar neke klase

Broj mogućih klasa je od 2 do 5, s dvije kombinacije određivanja pragova (za 5 klasa je samo jedan način određivanja pragova, jer Manesh nije radio s 5 klasa pa nema za njih navedene pragove), što sve skupa daje 7 mogućih kombinacija klasificiranja RASA-e.

Promatranjem podataka kao niza pomičnih prozora, iz navedena tri atributa su izračunati novi atributi koji opisuju RASA-u čitavog prozora. Novih atributa je 14 i to su redom:

- određeni iz vrijednosti RASA-e:
  - *r_asa_middle* – RASA središnjeg aminokiselinskog ostatka prozora
  - *r_asa_win_mean* – srednja vrijednost RASA-a svih aminokiselinskih ostataka prozora
  - *r_asa_win_median* – središnja (medijan) vrijednost RASA-a svih aminokiselinskih ostataka prozora
  - *r_asa_win_mean_class* – srednja vrijednost RASA-a svih aminokiselinskih ostataka prozora prikazana kao pripadnost pojedinoj klasi
  - *r_asa_win_median_class* – središnja (medijan) vrijednost RASA-a svih aminokiselinskih ostataka prozora prikazana kao pripadnost pojedinoj klasi
- određeni iz vrijednosti *class_r_asa*-e:
  - *class_r_asa_middle* – *class_r_asa* središnjeg aminokiselinskog ostatka prozora
  - *class_r_asa_win_mean* – srednja vrijednost *class_r_asa*-a svih aminokiselinskih ostataka prozora zaokružena na cijeli broj

- *class_r_asa_win_median* – središnja (medijan) vrijednost *class_r_asa*-a svih aminokiselinskih ostataka prozora

- određeni iz vrijednosti *class_mean_r_asa*-e:
  - *class_mean_r_asa_middle* – *class_mean_r_asa* središnjeg aminokiselinskog ostatka prozora
  - *class_mean_r_asa_win_mean* – srednja vrijednost *class_mean_r_asa*-a svih aminokiselinskih ostataka prozora
  - *class_mean_r_asa_win_median* – središnja (medijan) vrijednost *class_mean_r_asa*-a svih aminokiselinskih ostataka prozora
- određeni iz vrijednosti *class_median_r_asa*-e:
  - *class_median_r_asa_middle* – *class_median_r_asa* središnjeg aminokiselinskog ostatka prozora
  - *class_median_r_asa_win_mean* – srednja vrijednost *class_median_r_asa*-a svih aminokiselinskih ostataka prozora
  - *class_median_r_asa_win_median* – središnja (medijan) vrijednost *class_median_r_asa*-a svih aminokiselinskih ostataka prozora

RASA je realni broj koji se mora diskretizirati da bi se mogao predvidjeti klasifikacijom. Cilj uvođenja ovako velikog broja sličnih atributa je pokušati opisati RASA čitavog prozora na više načina da bi se kasnije promatranjem rezultata utvrdilo koja diskretna veličina ju najbolje opisuje.

3.3 Priprema podataka za klasifikaciju

Nakon dodavanja profila i klasa RASA-a XML datotekama, informacije o lancima skupa proteina za klasifikaciju prevedene su u ARFF (engl. *Attribut Relation File Format*) format. ARFF format [4] koristi velik broj aplikacija za klasifikaciju.

ARFF datoteke u ovome radu sadržavaju informacije o 170192 instanci, vrijednostima pripadajućih atributa te da li je neka instanca u ovisnosti o postavljenim uvjetima mjesto interakcije.

Primjer zapisa jedne instance:

1HCN,A,25,PRO,GLY,ALA,PRO,ILE,LEU,GLN,CYS,MET,2,1,1,1,0,1,2,2,1,1,35,1,
7,1,37,2,1,0,1,2,0,0,0,23,0,0,0,63,0,0,1,0,0,0,0,9,0,0,0,0,48,20,4,4,0,
0,0,6,0,0,1,0,0,0,0,3,4,0,0,7,0,6,0,0,0,0,0,0,0,0,1,0,0,0,89,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,70,6,0,0,0,0,0,0,0,0,24,0,0,0,0,0,0,0,0,0,0,4,0,6,12,
0,0,0,0,79,0,0,0,0,0,0,0,94,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,100,0,0,0
,0,0,0,0,0,0,0,0,0,0,0,0,0,8,0,0,0,0,0,0,0,0,0,0,4,56,0,0,6,18,0,0,9,4,47
,45,2,2,1,2,2,2,36,45,1,34,43,1

U gore navedenom primjeru radi se o proteinu 1HCN, lancu A te aminokiselini ILE, 25-oj aminokiselini u tom lancu, središnjoj u prozoru od devet popisanih aminokiselina. Zatim slijede profili za svaku aminokiselinu u prozoru (20×9) te vrijednosti 14 svojstava dobivenih iz RASA-a pojedine aminokiseline zapisanih u XML datoteci.

Ovih 14 vrijednosti je radi uštede memorije zaokruživano na cjelobrojne vrijednosti (ispostavilo se da je gubitak informacije biološki zanemariv u usporedbi s uštedom memorije). Posljednja vrijednost u retku je klasa koja označava je li promatrani prozor mjesto interakcije.

Za predviđanja u ovom radu najčešće su korišteni reducirani podaci gdje većina gore navedenih atributa nije ulazila u predviđanja, što je detaljnije objašnjeno u petom poglavlju.

4 Metode

U nastavku će biti opisane metode korištene prilikom stvaranja ovog rada i to redom: metoda za izvlačenje profila slijeda aminokiselinskih ostataka PSI-BLAST, metoda optimiranja Lloyd-Max kvantizator, metoda slučajnih šuma (engl. *Random Forest*) te mjere uspješnosti predviđanja.

4.1 PSI - BLAST

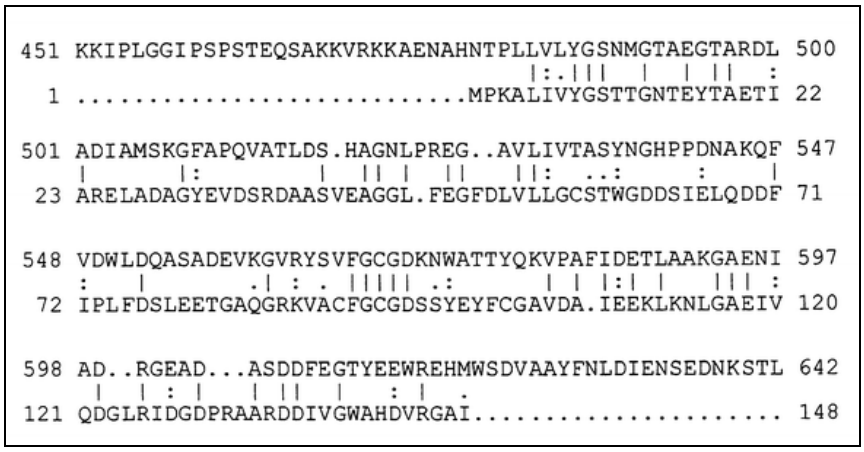
PSI-BLAST [14] je poboljšani algoritam tehnike BLAST [14] pomoću kojega je moguće pronaći evolucijsku očuvanost aminokiselinskog ostatka, tj. profil. Koristi se metodom poravnanja slijeda (engl. *sequence alignment*) za prepoznavanje sličnosti dvaju proteina koji nisu blisko povezani, tj. nemaju bliskog zajedničkog homologa. Proteini mogu biti strukturalno ili slijedno slični, pri čemu te dvije vrste sličnosti nisu nužno međusobno povezane. Temeljna ideja algoritma je prepoznati zajedničku strukturu iz slabe slijedne sličnosti. Toj se problematici pristupa metodom poravnavanja slijedova u sklopu tehnike BLAST.

4.1.1 Poravnavanje sljedova

Prvi korak u gradnji profila jest za neki proteinski slijed pronaći pripada li ona već nekoj poznatoj familiji proteina. Poravnavanjem primarnih sljedova koji predstavljaju neku familiju proteina vrši se prepoznavanje sličnih elemenata što može upućivati na funkcionalnu, strukturnu ili evolucijsku povezanost između sljedova.

Poravnati aminokiselinski ostaci se prikazuju kao redovi unutar matrice. Ako dva poravnata slijeda dijele zajedničkog pretka, nepravilnosti u slijedovima mogu se interpretirati kao točke mutacije ili pak praznine koje su posljedice delecije i insercije tokom evolucije u odnosu na izvorni slijed. Dijelovi slijeda 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.

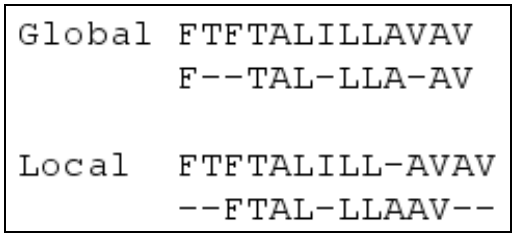
Poravnati sljedovi u odnosu na aminokiselinske ostatke prikazuju se grafički i tekstualno. U gotovo svim zapisima sljedovi su zapisani u recima tako da se slični ili jednaki aminokiselinski ostaci nalaze u istom stupcu. U grafičkim prikazima (slika 4.1) koriste se boje radi označavanja onih dijelova slijeda koji su identični, konzervirani ili pak djelomično konzervirani.



Slika 4.1 Slijedno poravnanje između dva proteina:
| – identični ostaci
: – konzervirani ostaci
. – djelomično konzervirani ostaci

Jedan od tekstualnih formata prikaza slijeda je FASTA format [16] 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 '>'.

Sljedovi (dva slijeda ili više njih) se mogu poravnati na lokalnoj i globalnoj razini. Pri globalnom poravnanju svi se ostaci pojedinačno poravnaju uz očuvanje duljine slijeda. Lokalno poravnanje ima veći učinak kada se slijedovi razlikuju, ali se sumnja na eventualnu sličnost pojedinih dijelova. Postoji još i hibridno poravnanje koje je kombinacija lokalnog i globalnog.



Slika 4.2 Globalno i lokalno poravnanje

4.1.2 BLAST i BLOSUM

BLAST [14] (engl. *Basic Local Alignment Search Tool*) je algoritam za usporedbu sljedova aminokiselinskih ostataka proteina ili nukleotida DNK lanca.

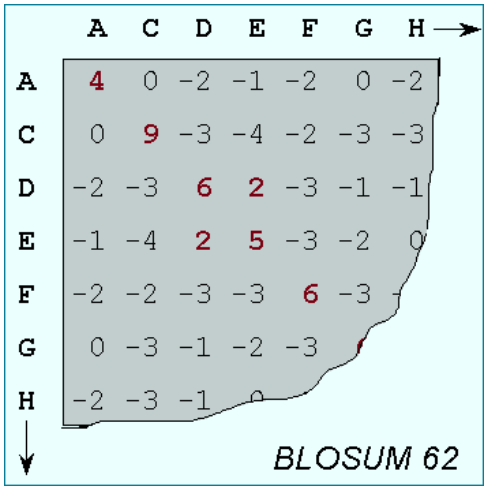
BLAST omogućuje da se za neki slijed koji se ispituje, pretraži proteinska baza podataka. Proteinska baza podataka [15] sadrži informacije o poznatim proteinskim slijedovima. Tako algoritam pretražujući bazu nalazi slijedove koje odgovaraju slijedu koji se ispituje, pritom zadovoljavajući određeni prag sličnosti (engl. *threshold*) koji u pravilu zada korisnik.

Kada se radi poravnavanje sljedova aminokiselinskih ostataka, algoritam BLAST koristi supstitucijsku matricu [24] za procjenu sličnosti sljedova. Postoji nekoliko takvih matrica, a najpoznatije među njima su BLOSUM i PAM.

BLOSUM (engl. *Blocks Substitution Matrix*) se bazira na lokalnom poravnavanju, a prvi put je predstavljena u radu Henikoffa i Henikoffa [24]. 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 slijedova tj. onih koje slabo divergiraju, BLOSUM matrice su nastale iz evolucijski divergentnih slijedova.

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

Na slici 4.3 prikazana je matrica BLOSUM62, kakva se koristila u ovom radu.



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

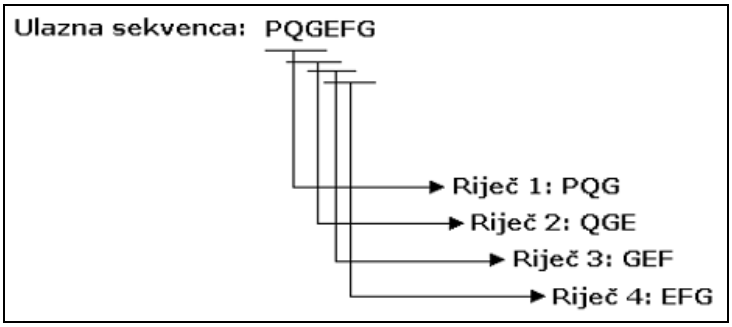
Vrijednosti matrice mogu se izračunati sljedećim izrazom:

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

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

4.1.3 Algoritam

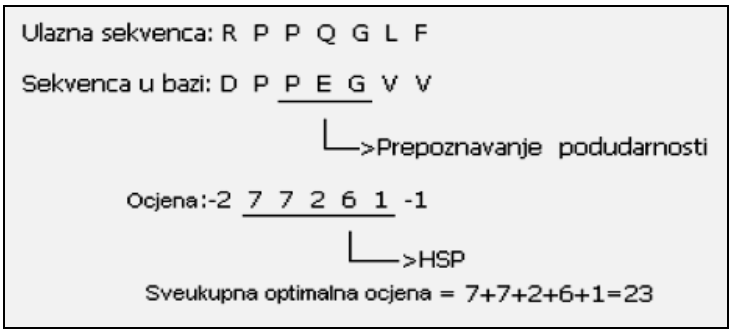
Osnovna ideja algoritma jest da svako dobro ocijenjeno lokalno poravnanje dvaju sljedova gotovo uvijek sadrži dobro očuvanu jezgru. Za parove ostataka u slijedu određuje se ocjena poravnanja i ako je ona iznad nekog zadanog praga, taj se par ostataka naziva dobro ocijenjenim lokalnim poravnanjem (engl. *High-scoring Segment Pairs, HSP*) [14]. BLAST *pretražuje* sljedove nalazeći dobro ocijenjena poravnanja između slijeda koji se ispituje i onih sljedova u bazi sljedova. Algoritam ulazni slijed koji se ispituje podijeli u trigrame, tj. u riječi od po 3 slova. Slika 4.4 prikazuje metodu.



Slika 4.4 Rastavljanje slijeda 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 ocjena sličnosti riječi PQG i 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 ulaznog slijeda, slijedi traženje tih istih ciljnih riječi u sljedovima baze. Kad se pronađe ciljna riječ u slijedu baze, ona može upućivati da s odgovarajućom riječi ulaznog slijeda čini jezgru. Da bi se to zaključilo vrši se proširivanje u oba smjera, odnosno gledaju se susjedni ostaci te se računa ocjena. Proširivanje poravnanja traje sve dok ocjena sličnosti (koja se čita iz BLOSUM matrice) ne počne padati (Slika 4.5).



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 oba slijeda. Pritom se mora smanjiti

prag sličnosti za ciljne riječi kako bi se zadržala osjetljivost. Ujedno se smanjuje i broj produljenja. Produljenje se radi Smith-Waterman algoritmom [25] koji vrši poravnanje s razmacima (engl. *gapped alignment*). Ova se verzija BLAST algoritma prema tome naziva *gapped BLAST* [14].

4.1.4 Procjena značajnosti ocjene lokalnog poravnanja

Dobro ocijenjeno lokalno poravnanje ne mora nužno značiti da su odgovarajući sljedovi slični te da imaju zajedničkog homologa. Lokalno poravnanje može biti posljedica slučajnosti. S ciljem uklanjanja takvih pojava, radi se model slučajnih sljedova. 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 o evolucijskoj povezanosti.

U dovoljno dugačkim sljedovima duljine m i n, značajnost ocjene lokalnog poravnanja karakteriziraju dva parametra: K i λ. Očekivani broj dobro ocijenjenih lokalnih poravnanja, *E-value*, koji su posljedica slučajnosti jest:

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

gdje je $N = mn$, a S' normalizirana ocjena S (S je prag za dobro ocijenjeno 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 kako bi bilo dobro ocijenjeno. 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 ocijenjenih 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:

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

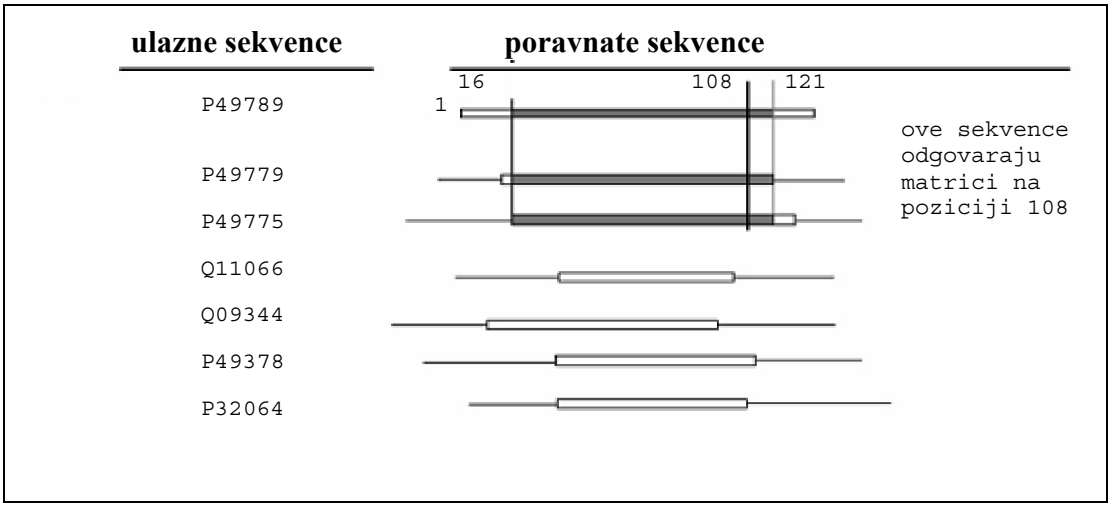
4.1.5 PSI-BLAST

PSI-BLAST (engl. *Position Specific Iteration BLAST*) [14] inačica je BLAST algoritma u kojemu se profil, odnosno matrica vjerojatnosti pronalaženja svake od 20 aminokiselina na mjestu aminokiseline kojoj se profil traži (engl. *Position Specific Scoring Matrix, PSSM*), gradi iz višestrukog slijednog poravnanja te najviše ocijenjenih 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 omogućuje pronalazak udaljenih sljedova odnosno onih manje sličnih. Profil izgrađen u prvoj iteraciji temelji se na sljedovima sličnim ulaznom slijedu te služi za sljedeću iteraciju u kojoj se pronalaze udaljeni, a slični sljedovi.

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 sljedova. Početna točka u stvaranju profila jest grupa sljedova koji su poravnati, a ujedno su i izlazni podatak BLAST algoritma. Taj se rezultat reducira u cilju određivanja vrijednosti profila. Za svaki stupac poravnatih sljedova, u obzir se uzimaju i susjedni aminokiselinski ostaci. Tako se poravnati reci sljedova reduciraju, tj. uzimaju se samo oni redovi kojima su stupci

postavljeni tako da svaki sadrži određeni ostatak ili prazninu, s time da su redovi iste duljine (slika 4.6).



Slika 4.6 Za profil se uzimaju samo oni sljedovi odgovarajuće duljine kojima 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 kojem se dobivaju vrijednosti elemenata matrice profila je:

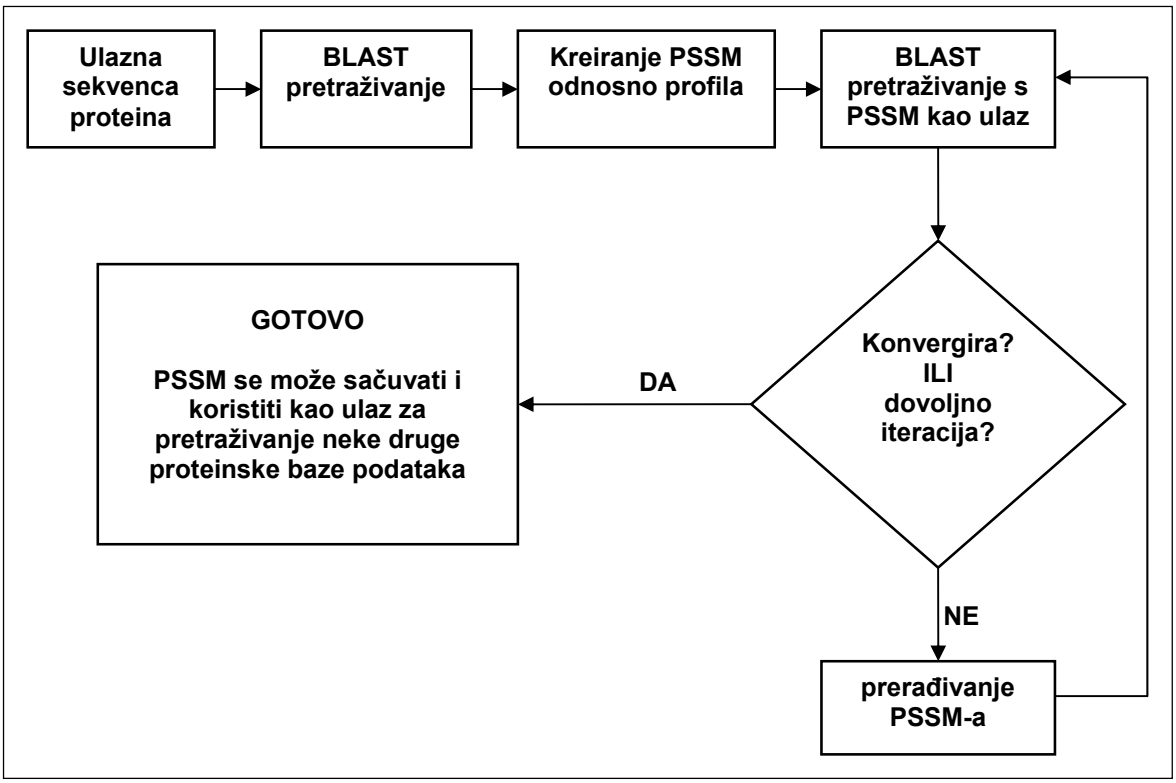
$$Profile(r,c) = \sum_{d=1}^{20} \sum_{i=1}^{N} weight(i)\delta(A_{ir}, d) \times Comp(residue_d, residue_c) \qquad (4.6)$$

gdje je *profile(r,c)* vrijednost profila za redak *r* i stupac *c*; *r* može imati vrijednosti od 1 do *N* (duljina skupa), *c* i *d* poprimaju vrijednosti od 1 do 20, predstavljajući aminokiseline; *i* je pozicija slijeda u skupu; *N* ukupan broj sljedova; $\delta(A_{ir},d)$ ima vrijednost 1 ako je ostatak na poziciji *r* u slijedu *i* aminokiselina *d*, inače je jednak nuli. $Comp(residue_d, residue_c)$ je vrijednost u supstitucijskoj tablici. *Weight(i)* se odnosi na težinu slijeda *i*. Težina slijeda [26] se može približno izračunati iterativnom metodom na sljedeći način:

1. Skupiti aminokiseline koje se nalaze na pojedinoj poziciji poravnatog skupa sljedova.
2. Inicijalna vrijednost svakog slijeda je nula.

3. Slučajnim odabirom odabrati slijed, birajući na svakoj poziciji (stupcu matrice) jednu aminokiselinu (praznine se tretiraju kao dodatna aminokiselina).
4. Izračunati udaljenost slučajnog slijeda od ostalih sljedova.
5. Dodati 1 težini najbližeg slijeda. Ako je više takvih, njih *K*, težini svakog od tih sljedova 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 prikazan je princip rada PSI-BLAST algoritma.



Slika 4.7 Dijagram PSI-BLAST algoritma

4.2 Lloyd-Max kvantizator

U dijelu skupova za predviđanje vrijednosti RASA-e su klasificirane u kategorije čije su granice preuzete od nekih drugih autora. Granice kategorija RASA-a su se kod preostalih skupova računale kvantizacijom.

Za kvantizaciju je odabran Lloyd-Max kvantizator, skalarni kvantizator optimalan s obzirom na srednju kvadratnu pogrešku. Kvantizator dijeli skup realnih brojeva u N podskupova R_1 do R_N. Svako od tih kvantizacijskih područja R_i predstavlja jednu kvantizacijsku razinu a_i, koja je i sama realni broj. Realni broj x koji pripada podskupu R_i tada se kvantizira na vrijednost a_i.

Osnovni problem je kako odabrati kvantizacijska područja i kvantizacijske razine da bi srednja kvadratna pogreška kvantizacije bila minimalna. Uz te uvjete određujemo granice intervala $b_1, b_2, \ldots, b_{N-1}$.

Navedeni problem sastoji se od dva dijela:

- kako za određeni skup kvantizacijskih razina odrediti kvantizacijska područja
- kako za određeni skup kvantizacijskih područja odrediti kvantizacijske razine

Odgovor na prvo pitanje jest: kako bi se minimizirala pogreška kvantizacije, granica b_i između intervala R_i i R_{i+1} mora ležati na polovini puta između a_i i a_{i+1}. Ovo ne ovisi o vjerojatnosti pojavljivanja pojedine vrijednosti x, dok odgovor na drugo pitanje ovisi.

Neka funkcija $Q(x)$ preslikava sve vrijednosti $x \in R_i$ u a_i i neka je $f_x(x)$ funkcija gustoće vjerojatnosti od x. Tada srednja kvadratna pogreška iznosi:

$$E\left[\left(x-Q(x)\right)^2\right] = \int_{-\infty}^{+\infty} \left(x-Q(x)\right)^2 f_x(x)\,dx = \sum_{i=1}^{N} \int_{R_i} f_x(x)\left(x-a_i^{\,2}\right)dx \qquad (4.7)$$

Za pojedini interval R_i dobije se da je srednja kvadratna pogreška minimalna za:

$$a_i = \overline{x_i} \qquad (4.8)$$

Lloyd-Max algoritam alternira ova dva uvjeta, optimizirajući prvo granice intervala b_i za određene kvantizacijske razine a_i, a potom određujući nove razine a_i za

dobivene granice b_i, sve dok se srednja kvadratna pogreška ne smanji na neku određenu vrijednost. Točnije, algoritam se sastoji od sljedećih koraka:

1. odabere se proizvoljan skup N razina $a_1 < a_2 < a_N$,
2. za $1 \leq i \leq N$ odrede se $b_i = 0{,}5(a_{i+1} + a_i)$,
3. za $1 \leq i \leq N$ odrede se a_i kao uvjetne srednje vrijednosti x, pri čemu je $x \in (b_{i-1}, b_i]$,
4. koraci se ponavljaju dok srednja kvadratna pogreška ne postane zanemarivo mala.

Potrebno je naglasiti da se u ovom radu za predstavljanje klasa neće koristiti kvantizacijske razine a_i, već redni broj klase.

Ako se za određivanje kvantizacijskih granica b_i uzme medijan vrijednost dviju susjednih kvantizacijskih razina a_i+1 i a_i, novi kvantizator bit će manje osjetljiv na vrijednosti koje značajno odstupaju od ostalih (engl. *outlier*).

4.3 Metoda slučajnih šuma

Kao metoda klasifikacije su u ovom radu odabrane Slučajne šume (engl. *Random Forest, RF*) [4], i to zbog sljedećih svojstava:

- velika točnost prepoznavanja,
- relativno je otporna na outliere i šum
- daje korisne interne procjene pogreške bez potrebe za korelacijom
- daje procjenu o važnosti pojedinih značajki za klasifikaciju
- algoritam je jednostavan za paralelizaciju te postoji paralelna inačica PARF [5] izrađena na IRB-u koja će se koristiti u ovom radu.

Slučajna šuma, u daljnjem tekstu *RF*, je općeniti naziv za skupinu metoda koje se koriste stablastim klasifikatorima $\{h(x,\theta_k), k=1,\ldots,\}$ gdje je $\{\theta_k\}$ 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 skupa odabranih *bootstrapping* metodom. Za razliku od klasičnih stabala gdje se odabire najbolji atribut, *RF* za grananje koristi *m* slučajno odabranih varijabli ($m<<M$, obično log_2M+1) i uzima one 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 danih 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.

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 kako 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 se zbroje glasovi koji su ispravno dani 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 neupotrebljenih *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*-vrijednost.

Za najveći dio predviđanja skupovi za treniranje i testiranje će se iz originalnog skupa formirati *krosvalidacijom*. Krosvalidacija je postupak dijeljenja originalnog 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 s tim podacima radi *N* predviđanja. Kao i kod *oob*-a zbroje se glasovi klasifikacije i promatra se uspješnost predviđanja. U ovom radu će se krosvalidacija raditi s deset podskupova, dakle 10 predviđanja.

4.3.1 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 kojima 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.

4.4 Mjere određivanja ovisnosti među kategorijama dvaju svojstava

Može se dogoditi kod predviđanja da težina informacije o vrijednosti nekog od atributa koji sudjeluju u predviđanju direktno ovisi o iznosu te vrijednosti. Primjerice, ako se radi o diskretnom atributu koji predstavlja pripadnost jednoj od *N* kategorija (klasa), moguće je da je pripadnost tog atributa kategoriji *a* od *N* značajnija informacija od primjerice pripadnosti kategoriji *b* od *N*. Zbog toga je nekad poželjno pomaknuti granice klasificiranja tih atributa. Ovo se radi postavljanjem *težina* u prvom krugu predviđanja, kad se klasificiraju takvi atributi. Prilikom klasificiranja se zadaju težinske vrijednosti koje utječu na kriterij, odnosno prag klasificiranja pojedinih klasa. Na taj način se postiže izmijenjena raspodjela klasificiranih vrijednosti u korist željene kategorije, što će u konačnici utjecati i na raspodjelu klasificiranja mjesta interakcije u drugom krugu predviđanja.

Da bi se ovako nešto uradilo, potrebno je dobiti informaciju o odnosu pojedinih klasa dvaju atributa. Cilj je uočiti koji su atributi statistički značajni za predviđanje mjesta kontakta i u kakvom su međusobnom odnosu različite klase istog atributa. Metode koje će se koristiti u ovom radu za određivanje spomenutih relacija su χ^2-test i omjer vjerojatnosti [27].

4.4.1 χ^2-test

Proučavanjem kvalitativnih podataka, može se pretpostaviti da neka teorijska raspodjela dobro opisuje opaženu raspodjelu frekvencija. Da bi se ta pretpostavka (*nulta hipoteza*) provjerila, primjenjuje se χ^2-test (*hi-kvadrat test*) – test nezavisnosti dviju varijabli.

χ^2-test (poznat i kao *Pearsonov test*) [27] je neparametarski test pomoću kojega se testira nulta hipoteza kojom se tvrdi da obilježje *X* ima određenu (teorijsku) razdiobu nasuprot alternativne hipoteze da nema tu razdiobu. χ^2-test je vrlo

praktičan kad se želi utvrditi odstupaju li neke dobivene (opažene) frekvencije od frekvencija koje bi se očekivale pod određenom hipotezom (4.9). Ovim testom se također pokazuje vjerojatnost povezanosti između dvije varijable.

$$\chi^2 = \sum \frac{(f_0 - f_t)^2}{f_t} \tag{4.9}$$

pri čemu f_0 označava opažene frekvencije, a f_t očekivane (teoretske) frekvencije, tj. frekvencije koje bismo očekivali pod nekom određenom hipotezom.

Tablicom niže dan je jednostavan primjer jedne takve raspodjele frekvencija za dvije varijable s po dvije kategorije, a izračun izrazom (4.10). Nulta hipoteza u promatranom slučaju kaže da su raspodjele po kategorijama dvaju tipova podataka jednake, tj. da su tipovi i kategorije međusobno neovisni.

	Kategorija 1	Kategorija 2	Ukupno
Tip 1	a	b	a+b
Tip 2	c	d	c+d
Ukupno	a+c	b+d	a+b+c+d

Tablica 4.1 Raspodjela frekvencija za dva tipa podataka raspoređena u dvije kategorije

$$\chi^2 = \sum \frac{(ad - bc)^2 (a+b+c+d)}{(a+b)(c+d)(b+d)(a+c)} \tag{4.10}$$

Stupanj slobode promatranog skupa podataka računa se kao (broj promatranih tipova podataka - 1) × (broj mogućih kategorija - 1), odnosno (broj stupaca - 1) × (broj redova - 1).

Jednom kad je određen broj stupnjeva slobode te iznos veličine χ^2, može se iz tablice raspodjele ili računalnim putem odrediti i nivo vjerojatnosti kojim se tvrdi istinitost nulte hipoteze. Općenito se uzima za one vrijednosti za koje je nivo vjerojatnosti manji od 5% da su statistički značajne, tj. da pobijaju nultu hipotezu. Drugim riječima, da bi se nulta hipoteza proglasila istinitom, mora se potvrditi sa sigurnošću većom od 5%.

4.4.2 Omjer vjerojatnosti, z-test

Omjer vjerojatnosti [27] je mjera određivanja odnosa između različitih klasa istog atributa.

U primjeru (Tablica 4.1) mogu se definirati sljedeće relacije:

- vjerojatnost da podatak koji je *Tip 1* poprimi vrijednost iz *Kategorije 1* iznosi:

$$P_{11} = \frac{a}{a+b}, \tag{4.11}$$

- vjerojatnost da podatak koji je *Tip 1* poprimi vrijednost iz *Kategorije 2* iznosi:

$$P_{12} = \frac{b}{a+b}, \tag{4.12}$$

te analogno za ostale kombinacije raspodjele podataka.

Iz (4.11) i (4.12) slijedi da je za podatak koji je *Tip 1* vjerojatnost u korist poprimanja vrijednosti iz *Kategorije 1* naspram *Kategorije 2* jednaka:

$$P_{1,1/2} = \frac{\textbf{Tip 1 u Kategoriji 1}}{\textbf{Tip 1 u Kategoriji 2}} = \frac{\left(\frac{a}{a+b}\right)}{\left(\frac{b}{a+b}\right)} = \frac{a}{b} \tag{4.13}$$

Analogno, vjerojatnost pojavljivanja podatka koji je *Tip 2* u *Kategoriji 2* naspram vjerojatnosti pojavljivanja u *Kategoriji 2* je:

$$P_{2,1/2} = \frac{\textbf{Tip 2 u Kategoriji 1}}{\textbf{Tip 2 u Kategoriji 2}} = \frac{\left(\frac{c}{c+d}\right)}{\left(\frac{d}{c+d}\right)} = \frac{c}{d} \tag{4.14}$$

Sada se može definirati omjer vjerojatnosti $\hat{\theta}$, koji će za gornji primjer iznositi

$$\hat{\theta}_{1,2} = \frac{a/b}{c/d} = \frac{ad}{bc}, \tag{4.15}$$

a predstavlja omjer vjerojatnosti da podatak u *Kategoriji 1* pripada *Tipu 1* naspram vjerojatnosti da taj podatak pripada *Tipu 2*.

Veličina $\hat{\theta}$ može poprimiti vrijednosti od 0 do ∞ što nije spretno za određivanje koliko je jak odnos između klasa promatranog atributa. Zbog toga se uvodi veličina *z*:

$$z = \frac{\ln\left(\hat{\theta}\right) - 0}{SE\left[\ln\left(\hat{\theta}\right)\right]}, \tag{4.16}$$

gdje je $SE\left[\ln\left(\hat{\theta}\right)\right]$ standardna pogreška za $\ln\left(\hat{\theta}\right)$ i za gornji primjer iznosi:

$$SE\left[\ln\left(\hat{\theta}\right)\right] = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}} \tag{4.17}$$

Vrijednost veličine *z* govori je li promatran odnos među klasama pojedinog atributa, sadržan u veličini $\hat{\theta}$, statistički značajan ili ne. Općenito se za statistički značajne rezultate uzimaju oni omjeri vjerojatnosti za koje je $z > 2$.

4.5 Mjere uspješnosti predviđanja

4.5.1 Točnost, preciznost, odziv, F-mjera

Jedna od mjera koje će se koristiti za ocjenu uspješnosti predviđanja je *točnost* (engl. *accuracy*). Podaci o točnosti izvlačit će se iz matrice greške koju generira PARF. Za definiranje točnosti koristit će se primjer matrice greške za dvije klase. U slučaju modela s dvije klase često se definira pozitivna i negativna klasa. Nazivi klasa nemaju praktično značenje, već predstavljaju dvije različite kategorije u koje klasifikator raspodjeljuje objekte nekog skupa. Tako objekti koje je klasifikator označio kao pozitivne mogu biti stvarno pozitivni (engl. *true positives, TP*) i lažno pozitivni (engl. *false positives, FP*). Analogno tome, uzorci podataka označeni kao negativni mogu biti stvarno negativni (engl. *true negatives, TN*) i lažno negativni (engl. *false negatives, FN*).

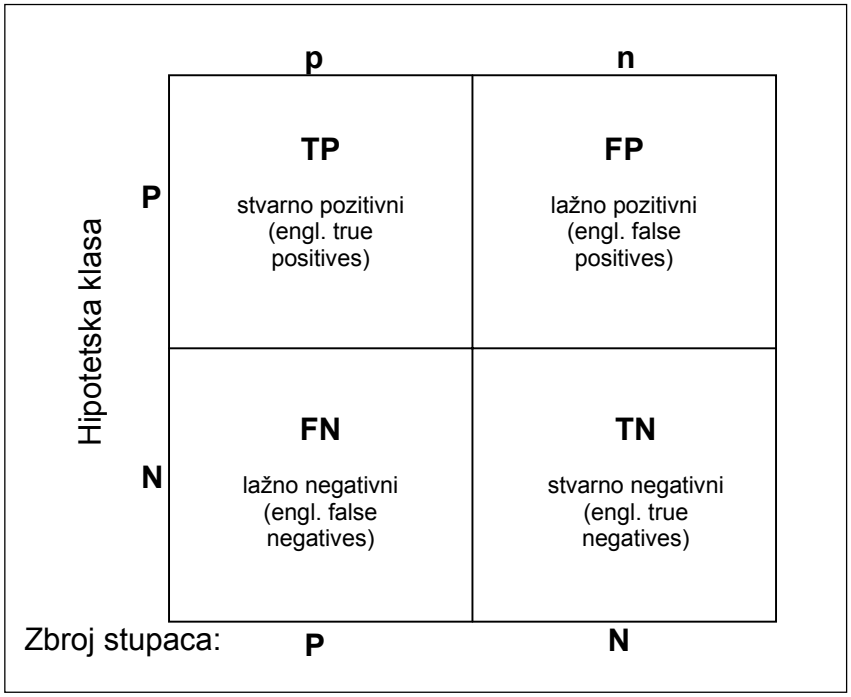
Točnost klasifikacije sada se može definirati kao omjer točno klasificiranih uzoraka te ukupnog broja uzoraka sljedećim izrazom:

$$točnost = \frac{TP + TN}{N} = \frac{TP + TN}{TP + FN + TN + FP} \tag{4.9}$$

Pogledom na matricu greške (slika 4.8) može se uočiti da brojnik odgovara dijagonali matrice, dok je nazivnik jednak sumi svih elemenata. Zbroj stupaca matrice odgovara broju objekata u pojedinoj klasi iz čega proizlazi da suma svih elemenata matrice odgovara broju objekata skupa.

Time se definira točnost za *N* klasa sljedećim izrazom:

$$točnost = \frac{\textbf{suma elemenata na dijagonali matrice}}{\textbf{suma svih elemenata matrice}} \tag{4.10}$$



Slika 4.8 Matrica greške za klasifikaciju u dvije kategorije

Pomoću gore definiranih veličina računaju se dodatne mjere koje opisuju karakteristike klasifikatora.

Odziv (engl. *recall*) se definira kao omjer ispravno klasificiranih pozitivnih instanci i stvarnog broja pozitivnih instanci:

$$odziv = \frac{TP}{TP + FN} \tag{4.11}$$

Preciznost (engl. *precision*) se definira kao omjer ispravno klasificiranih pozitivnih instanci i instanci koje je klasifikator proglasio pozitivnima:

$$preciznost = \frac{TP}{TP + FP} \tag{4.12}$$

F-mjera je težinska harmonijska srednja vrijednost preciznosti i odziva. Definirana je izrazom:

$$F - mjera = \frac{2 \cdot (preciznost \cdot odziv)}{preciznost + odziv} \tag{4.13}$$

4.5.2 Analiza ROC i PR krivulje

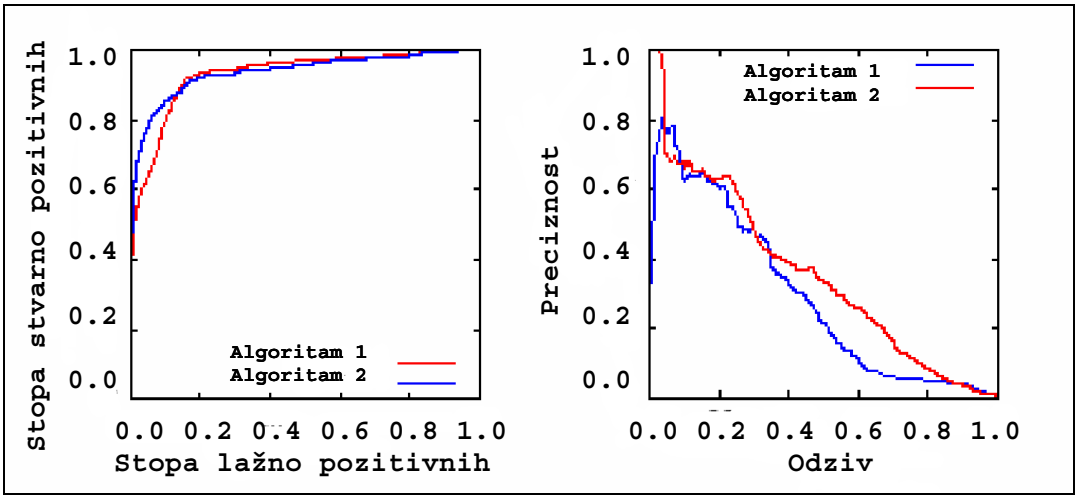
ROC (engl. *Receiver Operating Characteristic*) krivulja je dvodimenzionalni prikaz performansi klasifikatora u kojima je *TP* na Y-osi i *FP* na X-osi te prikazuju odnos između koristi (*TP*) i cijene (*FP*).

Uobičajeno je izračunati površinu ispod ROC 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.

Graf preciznost-odziv (engl. *Precision-Recall,* PR) i njemu pripadajuća krivulja se često koriste kao alternativa ROC krivuljama u slučaju asimetričnosti (engl. *skew*) u raspodjelama klasa [28]. Važna razlika između PR prostora i ROC prostora je vizualna reprezentacija krivulja. Gledajući PR krivulje moguće je vidjeti razliku među algoritmima koja nije očita u ROC prostoru.

Na slici 4.9 prikazan je primjer razlika PR i ROC krivulja za dva različita algoritma.



Slika 4.9 Usporedba algoritama u ROC i PR prostoru

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

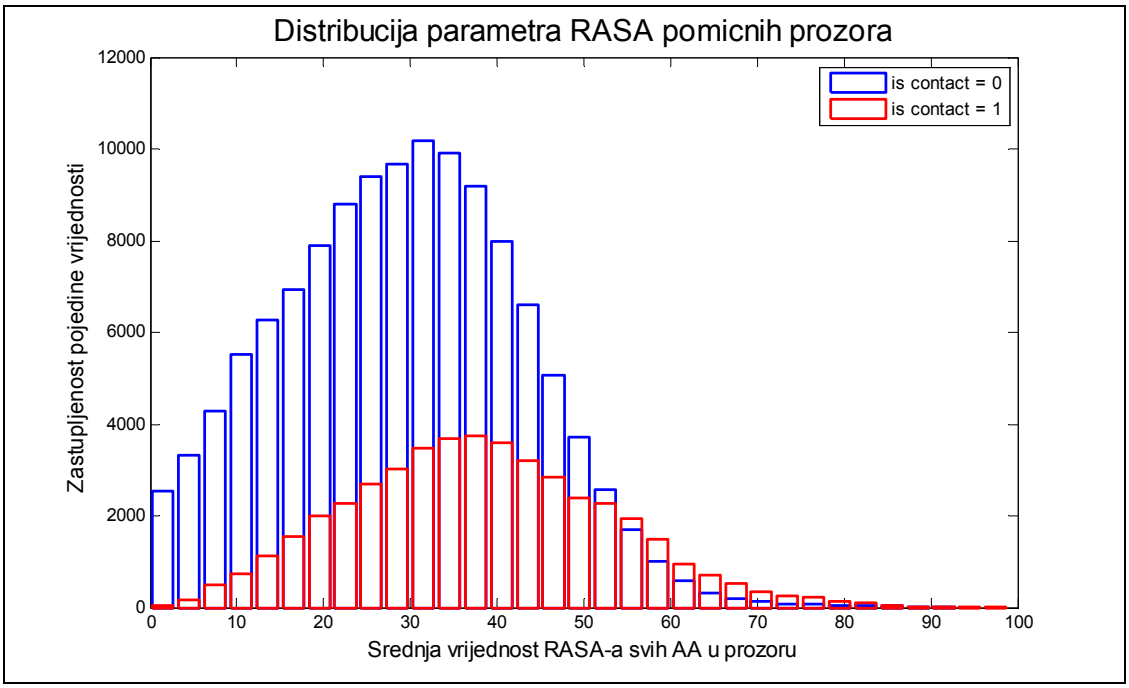
Prema desnom grafu (slika 4.9) koji prikazuje PR prostor, moglo bi se zaključiti da su performanse oba algoritma približno jednake. Međutim, u PR prostoru se vidi mala, ali očita prednost jednog od algoritama. Upravo zbog toga, rezultati predviđanja su se u ovom radu prikazivali PR krivuljama.

Za crtanje PR i AUC krivulja korišten je alat R [29].

5 Rezultati

5.1 Utjecaj raspodjele RASA vrijednosti na mjesto kontakta

Na slici 5.1 prikazan je utjecaj raspodjele RASA vrijednosti pomičnih prozora na mjesto kontakta. Promatra se atribut *r_asa_win_mean*, dakle srednja vrijednost RASA-a 9 aminokiselinskih ostataka koji čine prozor. Plavom bojom je na grafu označena raspodjela za prozore koji nisu mjesta kontakta, a crvenom bojom je označena raspodjela za one prozore koji su proglašeni mjestom kontakta.



Slika 5.1 Raspodjela RASA vrijednosti pomičnih prozora

Vidi se iz grafa da je cijela raspodjela za mjesta kontakta pomaknuta blago u desnu stranu, što je bilo i za očekivati; gotovo da nema mjesta kontakta s RASA vrijednošću oko nule, a ona mjesta koja imaju iznimno visoku vrijednost RASA su gotovo sigurno mjesta kontakta. Raspodjele su slične normalnoj raspodjeli, pogotovo ova za mjesta kontakta.

Očigledno je da će prozori s jako malom RASA-om i oni s jako velikom biti važno mjerilo predviđanja mjesta kontakta. RASA prozora blizu nule gotovo isključuje

vjerojatnost pojavljivanja mjesta kontakta, za one s RASA-om preko 50 već je vjerojatnije da su mjesto kontakta nego da to nisu, dok se za one s RASA-om preko 80, iščitavajući graf (slika 5.1), može gotovo sigurno zaključiti da su mjesto kontakta.

Međutim, cijelo područje između tih ekstremnih vrijednosti ima jako sličnu razdiobu za mjesta koja jesu mjesta interakcije i za ona koja to nisu. Uz to, predviđanje će se raditi klasifikacijom pa među korištenim atributima nipošto neće biti srednja RASA prozora, već će biti neki od atributa koji tu vrijednost približno opisuje.

5.2 Odabir ulaznih RASA atributa za predviđanje

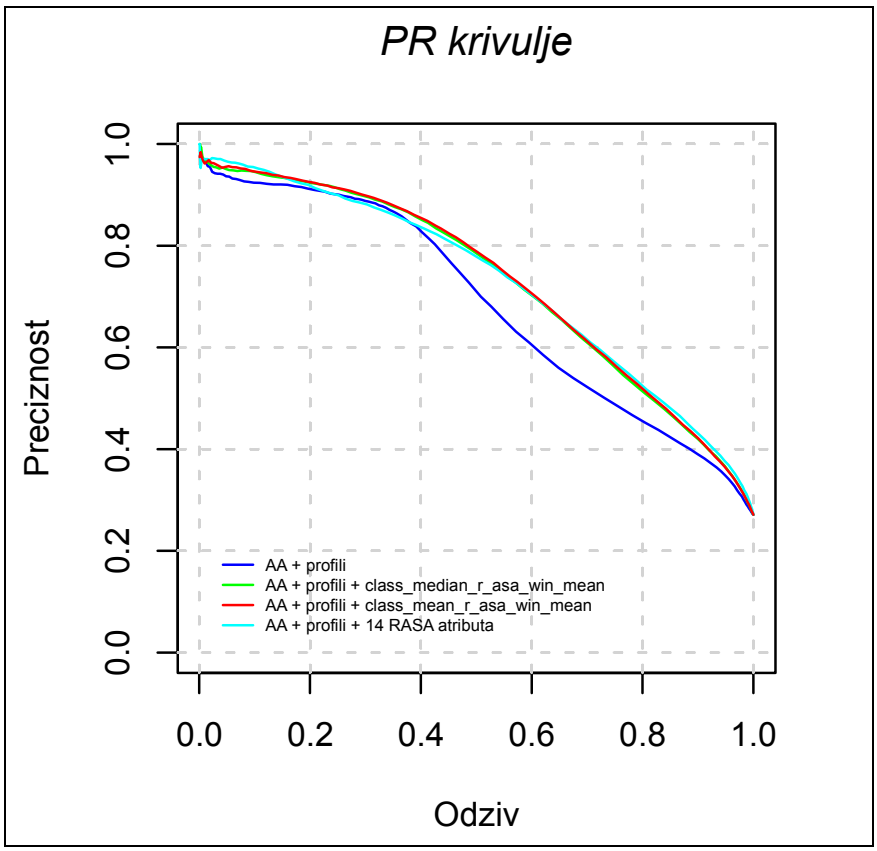
Predviđanje mjesta interakcije odvija se u dva koraka. U prvom koraku se za predviđanje koristi slijed od 9 aminokiselinskih ostataka u nizu te profili za svaki pojedini ostatak (20 po ostatku) što čini ulazni vektor od 189 atributa. Među 14 atributa koji na različite načine opisuju RASA-u (poglavlje 3.2.2.1), njih 5 je klasificirano u 2-5 kategorija, pa ih je moguće predvidjeti u prvom krugu predviđanja. Preostalih 9 su numeričke vrijednosti od kojih je njih 6 moguće izračunati poznavajući onih 5 koji se mogu predvidjeti. To sve skupa čini 11 atributa koji opisuju RASA-u prozora (ili središnjeg ostatka prozora), a koje je moguće predvidjeti, odnosno izračunati bez poznavanja prave vrijednosti RASA-e. Ideja je među spomenutih 11 atributa uočiti koji atribut ili kombinacija atributa najbolje pridonose predviđanju mjesta interakcije.

U drugom se krugu predviđanja – koristeći slijed od 9 aminokiselinskih ostataka u nizu, profile za svaki pojedini ostatak te atribute predviđene odnosno izračunate u prvom krugu – predviđaju mjesta interakcije.

Za neki skup podataka je učinjeno predviđanje mjesta kontakta *out-of-bag* metodom koristeći za predviđanje slijed od 9 aminokiselinskih ostataka u nizu te profile za svaki pojedini ostatak (20 po ostatku) i dodatno 14 atributa koji različitim pristupima opisuju RASA-u prozora.

U jednoj od datoteka dobivenoj ovim predviđanjem zapisane su značajnosti atributa. Atributi su poredani od najznačajnijeg do najmanje značajnog; već na prvi pogled se vidi da se dobar dio RASA atributa pokazuje značajnim. Najistaknutiji među njima su oni atributi koji direktno govore vrijednost RASA (*r_asa_win_mean*, *r_asa_middle*, *r_asa_win_median*). Nadalje, među atributima koji se mogu klasificirati po značajnosti najviše odudaraju *class_median_r_asa_win_mean* i *class_mean_r_asa_win_mean*.

Učinjeno je još nekoliko predviđanja mjesta kontakta *out-of-bag* metodom, ovaj put koristeći za isti skup nekoliko kombinacija sa spomenutim atributima. Slika 5.2 prikazuje usporedbu dobivenih PR krivulja.



Slika 5.2 PR krivulje predviđanja mjesta interakcije *out-of-bag* metodom koristeći samo slijed aminokiselinskih ostataka i njihove profile te kombinaciju sa a) *class_median_r_asa_win_mean*, b) *class_mean_r_asa_win_mean*, c) svih 14 RASA atributa

Iz gornjeg grafa je vidljivo da su svi rezultati bolji od dosadašnjih najboljih rezultata [1]. Detaljnijim proučavanjem vrijednosti veličina koje se koriste kao mjere uspješnosti predviđanja pokazuje se da se novim predviđanjima (bez obzira na korištene kombinacije atributa) za preciznost od oko 82% postiže odziv i do 45%.

Obzirom da su gore prikazana predviđanja mjesta kontakta urađena s poznatim vrijednostima RASA-e, za očekivati je da će uz predviđene vrijednosti atributa koji opisuju RASA-u, ovisno o uspješnosti tog predviđanja, konačni rezultati predviđanja mjesta kontakta biti nešto lošiji od gore prikazanih. Cilj je što bolje predvidjeti spomenute atribute i onda s njima krenuti u predviđanje mjesta kontakta. U najboljem slučaju očekuju se rezultati koji će po svim mjerama uspješnosti biti negdje između onih koje je postigla V. Dragosavljević [1] i gore prikazanih rezultata predviđanja.

Iz grafa (slika 5.2) je očito da je *class_mean_r_asa_win_mean* ipak nešto bolji za predviđanje mjesta kontakta od *class_median_r_asa_win_mean*. Iako su ova dva atributa (uz *class_median_r_asa_win_median* i *class_mean_r_asa_win_median*) međusobno vrlo slična, ubuduće je korišten *class_mean_r_asa_win_mean* jer se on ipak pokazao nešto boljim za sve ispitane skupove podataka. Rezultati prikazani slikom 5.2 dobiveni su za skup podataka *3a* (za taj skup RASA atributi su klasificirani u tri kategorije po pragovima dobivenim optimizacijom po Lloyd-Maxu), ali gotovo jednake relacije među atributima vrijede za predviđanja urađena na svim skupovima podataka korištenim u ovom radu.

Vrijednost atributa *class_mean_r_asa_win_mean* (poglavlje 3.2.2.1) predstavlja srednju vrijednost *class_mean_r_asa*-a svih aminokiselinskih ostataka prozora (s time da je *class_mean_r_asa* vrijednost dodijeljena klasi određena kao srednja vrijednost svih RASA vrijednosti koje se pojavljuju unutar neke klase). Atribut *class_mean_r_asa_win_mean* je realni broj i spada u one atribute koje nije moguće predvidjeti nego izračunati. Za računanje njegove vrijednosti treba poznavati vrijednosti *class_mean_r_asa*-a svih aminokiselinskih ostataka prozora. Ovaj problem može se svesti na određivanje atributa

class_r_asa_middle (kojeg je moguće predvidjeti!). Naime, klasu je lako pretvoriti u odgovarajući realni broj koji predstavlja srednju vrijednost svih vrijednosti koje su se pojavile unutar te klase. Taj realni broj je upravo *class_mean_r_asa* središnjeg ostatka u prozoru, a vrijednosti *class_mean_r_asa*-a preostalih ostataka promatranog prozora moguće je pronaći također kao *class_mean_r_asa* središnjeg ostatka u prozoru, gledajući nekoliko prethodnih odnosno sljedećih prozora. Primjerice, za prozor duljine 9 aminokiselinskih ostataka promatrat ćemo osim njega samoga i prethodna te sljedeća 4 prozora i njihove *class_r_asa_middle* (pretvorene u *class_mean_r_asa*) - tih 9 vrijednosti opisuje upravo 9 ostataka promatranog prozora i njihova srednja vrijednost predstavlja *class_mean_r_asa_win_mean* toga prozora.

Srednju vrijednost RASA-a svih aminokiselinskih ostataka prozora opisuje i jedan atribut kojeg je moguće predvidjeti – *r_asa_win_mean_class* (poglavlje 3.2.2.1). U datoteci značajnosti atributa, ovaj je atribut sljedeći po značaju nakon ranije spomenutih, pa će se u ovom radu osvrnuti i na predviđanja s korištenjem njegove vrijednosti.

Jednom kad su određeni atributi koji će biti korišteni za predviđanje mjesta kontakta, oni među njima koji se mogu predvidjeti predviđaju se u prvom krugu (to će biti *class_r_asa_middle*, odnosno *r_asa_win_mean_class*), potom se eventualno iz predviđenih izračunavaju preostali željeni atributi (*class_mean_r_asa_win_mean*) te se s predviđenim i izračunatim RASA atributima, uz informaciju iz slijeda i profila slijeda, predviđaju mjesta kontakta.

5.3 Odabir broja klasa i metode formiranja klasa

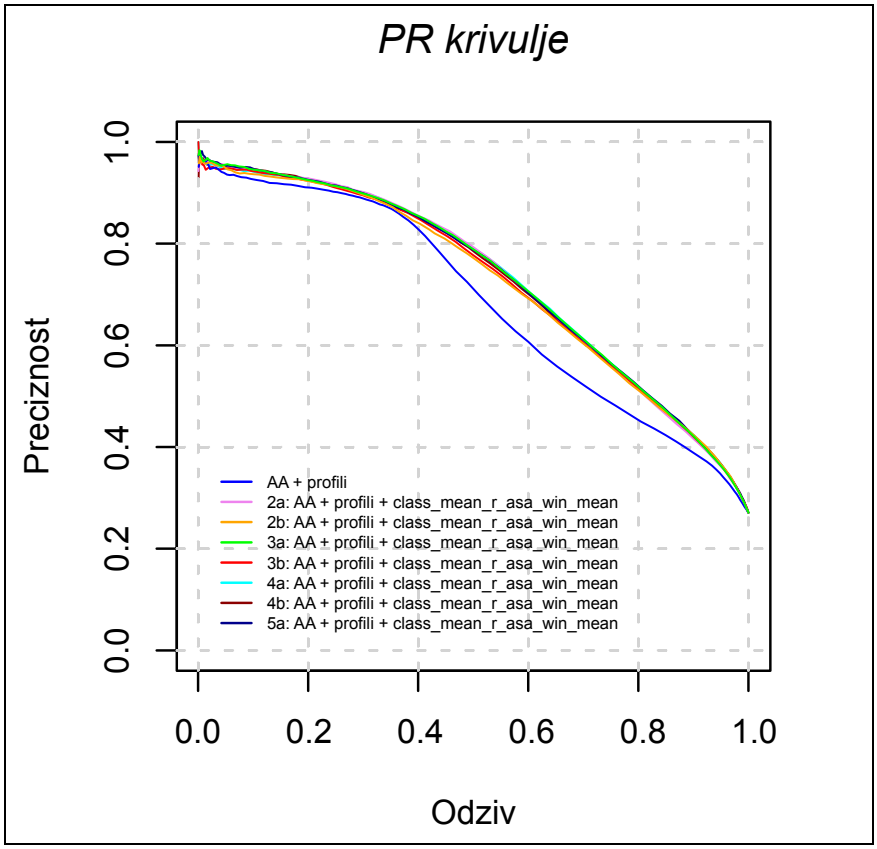
Obzirom na kombinacije broja klasa (2–5) te metoda formiranja pragova (Lloyd-Max kvantizator i pragovi preuzeti iz Maneshovog rada [17]) sedam je mogućih skupova koji će se koristiti za predviđanje, i to redom: *2a*, *2b*, *3a*, *3b*, *4a*, *4b* i *5a*. Kod oznake skupa prvi znak (broj) označava broj kategorija koji se koristio za klasifikaciju diskretnih atributa (2–5), a drugi znak je oznaka metode određivanja pragova za te kategorije; *a* - pragovi dobiveni Lloyd-Maxovim optimizacijskim

algoritmom; *b* – pragovi preuzeti iz Maneshovog rada [17]. Točni iznosi za oba slučaja prikazani su tablicom 5.1.

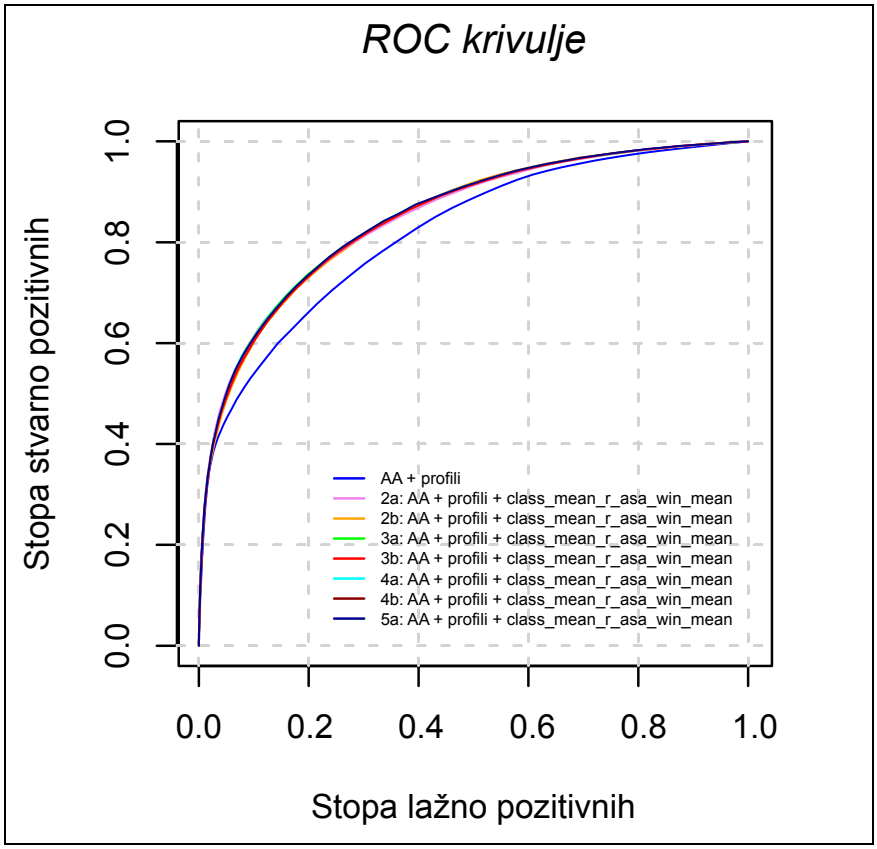
	Manesh	Lloyd-Max
2 klase	{ 9 }	{ 35,736 }
3 klase	{ 9, 64 }	{ 23,237, 55,742 }
4 klase	{ 4, 16, 49 }	{ 16,477, 40,357, 66,069 }
5 klasa	nije definirano	{ 12.81, 31,788, 51,209, 72.609}

Tablica 5.1 Pragovi za klasifikaciju RASA vrijednosti po kategorijama

Na slici 5.3 i 5.4 prikazani su rezultati predviđanja mjesta interakcije za ovih sedam skupova podataka kako bi se prepoznalo koji algoritam klasifikacije atributa bolje pridonosi predviđanju mjesta kontakta. Za potrebe uspoređivanja različitih skupova, predviđanje mjesta kontakta je urađeno *out-of-bag* metodom koristeći atribute koje je koristila V. Dragosavljević [1] te dodatno jedan od 14 atributa definiranih u ovom radu – prethodno izračunat *class_mean_r_a-sa_win_mean* za kojeg se pokazalo da je za predviđanje kontakta jedan od dva najznačajnija RASA atributa (poglavlje 5.3).



Slika 5.3 PR krivulje predviđanja mjesta interakcije *out-of-bag* metodom na 7 skupova podataka



Slika 5.4 ROC krivulje predviđanja mjesta interakcije *out-of-bag* metodom na 7 skupova podataka

	AUC	F-mjera	Točnost	Preciznost	Odziv
skup 2a	0, 8493	0, 5623	0, 8224	0, 845	0, 4213
skup 2b	0, 8489	0, 5575	0, 8227	0, 8275	0, 4204
skup 3a	0, 851	0, 5659	0, 7849	0, 8393	0, 4268
skup 3b	0, 8493	0, 5405	0, 8178	0, 8518	0, 3959
skup 4a	0, 8519	0, 5657	0, 8227	0, 8398	0, 4265
skup 4b	0, 8507	0, 5683	0, 8228	0, 8343	0, 4309
skup 5a	0, 8518	0, 5644	0, 822	0, 8368	0, 4258

Tablica 5.2 Mjere uspješnosti predviđanja mjesta interakcije *out-of-bag* metodom za 7 skupova podataka

Iako iz slika nisu vidljive velike razlike među skupovima, proučavanjem vrijednosti veličina koje se koriste kao mjere uspješnosti predviđanja (tablica 5.2)

vidi se da su generalno nešto bolje rezultate postigli skupovi s većim brojem klasa te pragovima dobivenim Lloyd-Max algoritmom.

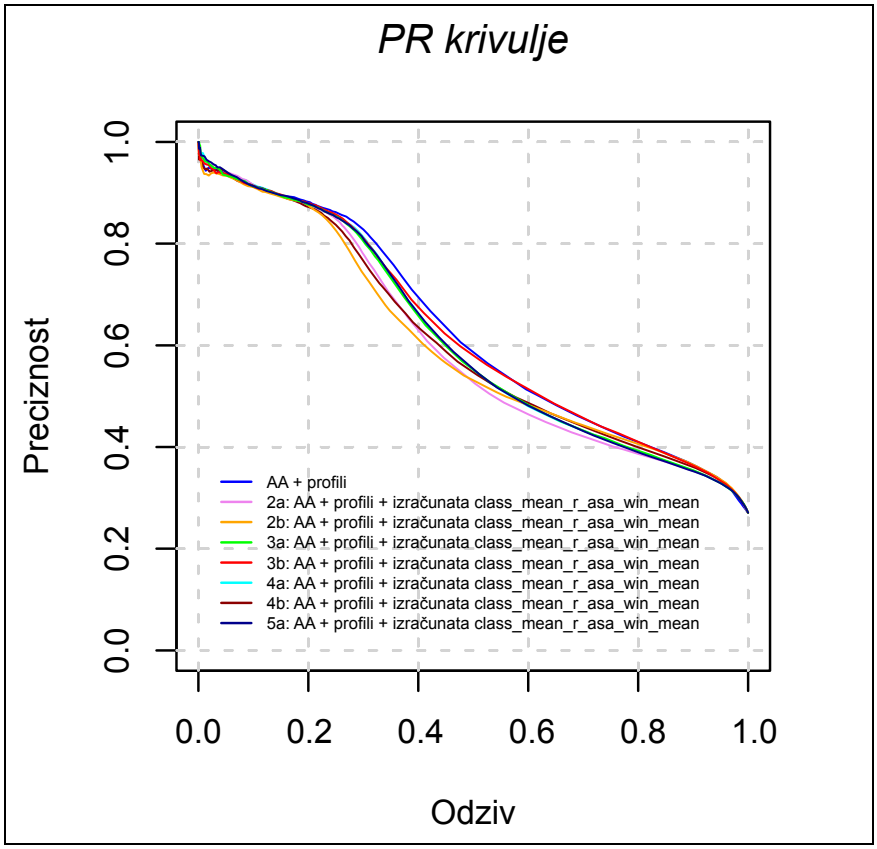
Gornji rezultati predviđanja su dobiveni uz poznate vrijednosti ulaznih atributa. Međutim, ukoliko svi korišteni atributi nisu poznati, nego se i oni sami predviđaju, uspješnost kojom se predvide direktno utječe na uspješnost predviđanja mjesta kontakta. U ovom radu će se RASA atributi predviđati iz slijeda aminokiselinskih ostataka i njihovih profila, pa je za predviđanje mjesta kontakta važno promotriti kakvo je predviđanje ovih atributa za pojedine skupove podataka. Tablicom 5.3 prikazan je odnos točnosti predviđanja jednog od RASA atributa – *class_r_asa_middle* (korištenog za izračunavanje *class_mean_r_asa_win_mean*) – za 9 različitih skupova.

	2a	2b	3a	3b	4a	4b	5a
Točnost	0,8186	0,8616	0,7111	0,7546	0,6473	0,6464	0,5962

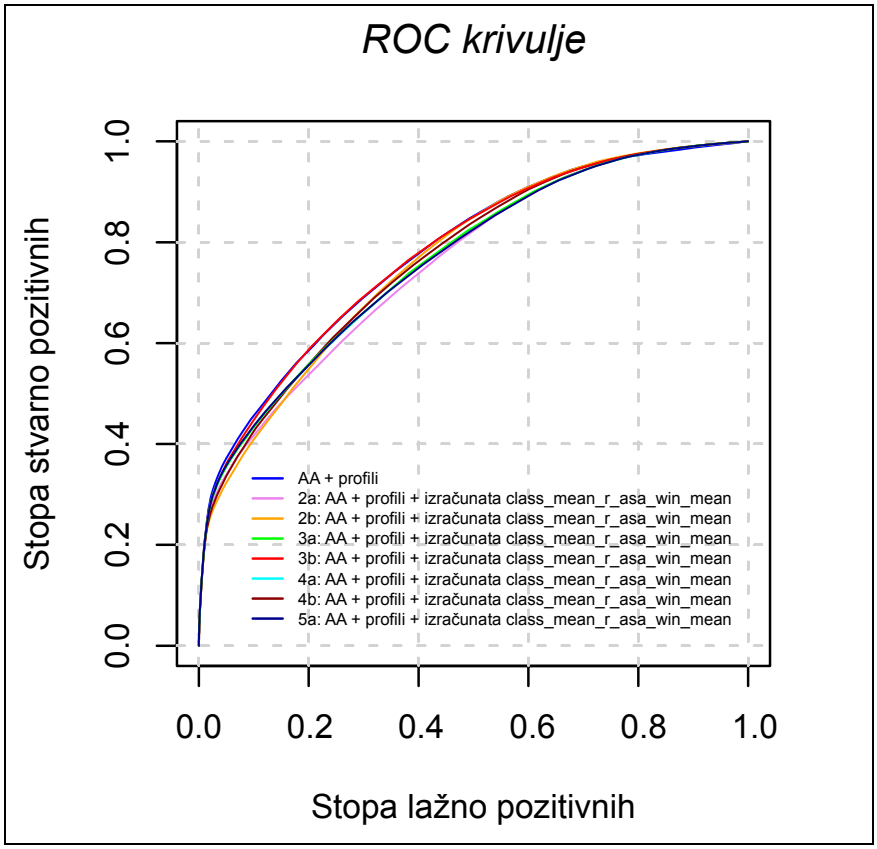
Tablica 5.3 Točnost klasificiranja atributa *class_r_asa_middle out-of-bag* metodom koristeći samo slijed aminokiselinskih ostataka i njihove profile

Vidi se iz tablice da se s manjim brojem klasa postiže veća točnost predviđanja (što je bilo za očekivati) te da su bolji rezultati postignuti sa skupovima s Maneshovim pragovima [17].

Očigledno je da su optimalni kriteriji klasificiranja atributa za predviđanja u prvom krugu upravo suprotni optimalnim kriterijima predviđanja mjesta kontakta u drugom krugu. Predviđanje RASA atributa je uspješnije što je broj kategorija manji, uz korištene Maneshove pragove. Predviđanje mjesta kontakta, s druge strane, je uspješnije što je broj kategorija veći (jer je time informacija sadržana u vrijednosti atributa značajnija), uz korištenje pragova dobivenih Lloyd-Max algoritmom. Optimalni skup za najbolje konačne rezultate predviđanja očigledno će biti kompromis između jednog i drugog. Da bi se odredio taj skup podataka urađeno je predviđanje atributa *class_r_asa_middle*, a potom i mjesta kontakta metodom krosvalidacije za svih sedam skupova podataka. Dobiveni rezultati prikazani su slikama 5.5 i 5.6 te tablicom 5.4.



Slika 5.5 PR krivulje predviđanja mjesta interakcije metodom krosvalidacije na 7 skupova podataka



Slika 5.6 ROC krivulje predviđanja mjesta interakcije metodom krosvalidacije na 7 skupova podataka

	AUC	F-mjera	Točnost	Preciznost	Odziv
skup 2a	0, 7581	0, 3988	0, 7862	0, 8354	0, 2619
skup 2b	0,769	0, 4532	0, 7778	0, 6791	0, 3401
skup 3a	0, 7659	0, 3839	0, 7849	0, 8547	0, 2476
skup 3b	0, 7802	0, 3885	0, 7859	0, 8569	0, 2512
skup 4a	0, 7654	0, 3863	0, 7853	0, 8546	0, 2495
skup 4b	0, 7689	0, 4155	0, 7858	0, 7955	0, 2812
skup 5a	0, 765	0, 386	0, 7852	0, 8533	0, 2494

Tablica 5.4 Mjere uspješnosti predviđanja mjesta interakcije metodom krosvalidacije za 7 skupova podataka

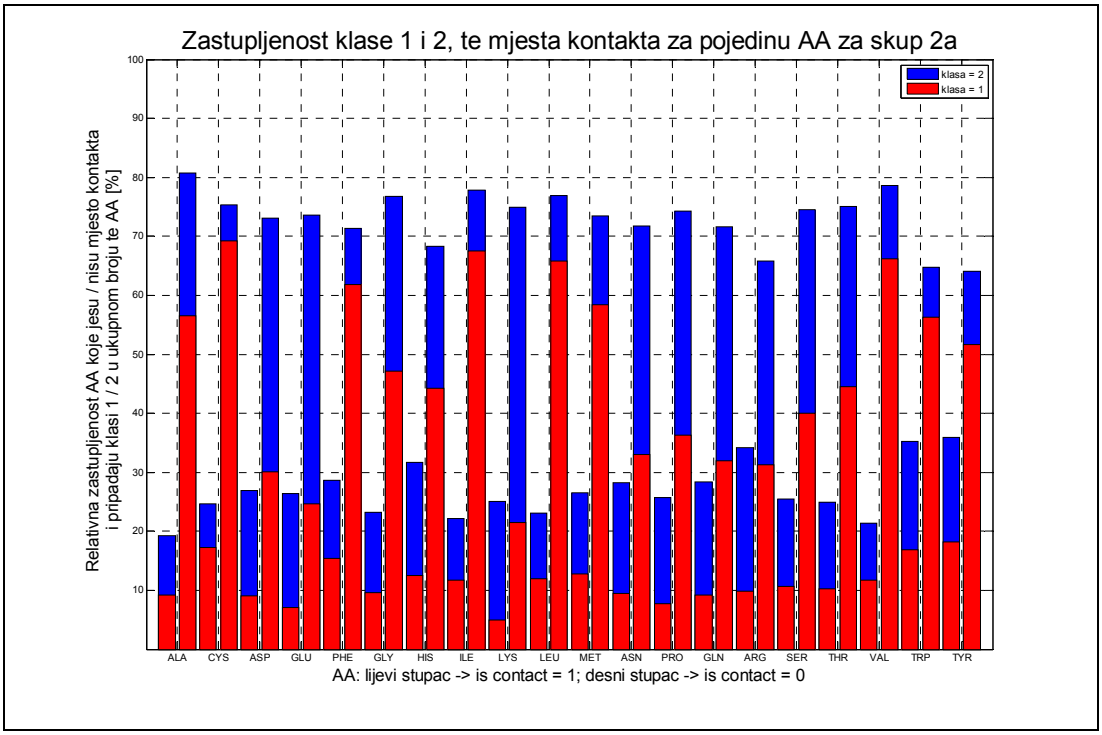
Rezultati predviđanja pokazuju da je ravnoteža između točnosti predviđanja RASA atributa i količine informacije sadržane u predviđenom atributu potrebne

za dobro predviđanje mjesta kontakta postignut za skup podataka iz grupe *3b*. Sljedeći najbolji rezultat bio je za skup *2b*, koji je ujedno imao daleko najbolju točnost predviđanja RASA atributa (tablica 5.3).

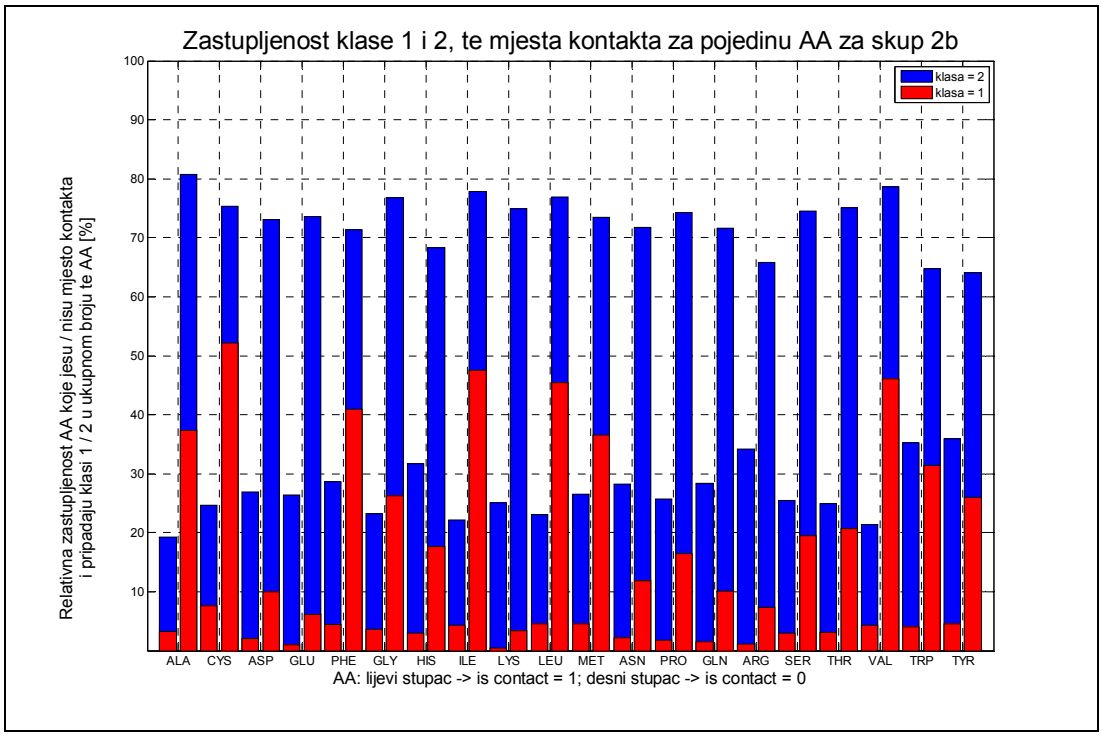
Uzevši gore navedeno u obzir, u nastavku ovog rada će se većina pažnje posvetiti upravo predviđanjima koristeći skupove iz grupa *2b* i *3b*.

Na slikama 5.7 i 5.8 promatra se za skupove s dvije klase utječe li atribut *class_r_asa* – klasa RASA-e aminokiselinskog ostatka na mjesto kontakta, tj. može li činjenica da pojedini ostatak pripada nekoj klasi uputiti na to da je u kontaktu. Plavom bojom prikazana je raspodjela onih aminokiselinskih ostataka čija je RASA klasa unutar klase 2, a crvenom bojom onih čija je RASA klasa unutar klase 1. Raspodjela je prikazana za svih 20 poznatih aminokiselina, za svaku lijevi stupac predstavlja raspodjele onih ostataka koji su u kontaktu, a desni stupac stoji za one koji nisu u kontaktu.

Iako iz slika nije bilo moguće donijeti nikakve konkretne zaključke, za neke aminokiselinske ostatke se može naslutiti da su klasa RASA-e i mjesto kontakta zavisne varijable. Ovo osobito vrijedi za klase koje su kategorizirane po Maneshovim pragovima (*2b*) – iz slike je općenito vidljivo da su u kontaktu češće ostaci čija je RASA u klasi 2 (koja je zastupljenija), a za neke aminokiselinske ostatke može se uočiti da ukoliko su u kontaktu su većinom klasa 2, naprotiv, ukoliko nisu u kontaktu su većinom klasa 1 (primjerice *CYS*, *PHE*, *ILE*, *LEU*, *VAL*).



Slika 5.7 Zastupljenost RASA klase 1 i 2 te mjesta kontakta za pojedini aminokiselinski ostatak u skupu *2a*



Slika 5.8 Zastupljenost RASA klase 1 i 2 te mjesta kontakta za pojedini aminokiselinski ostatak u skupu *2b*

5.4 Prikaz rezultata

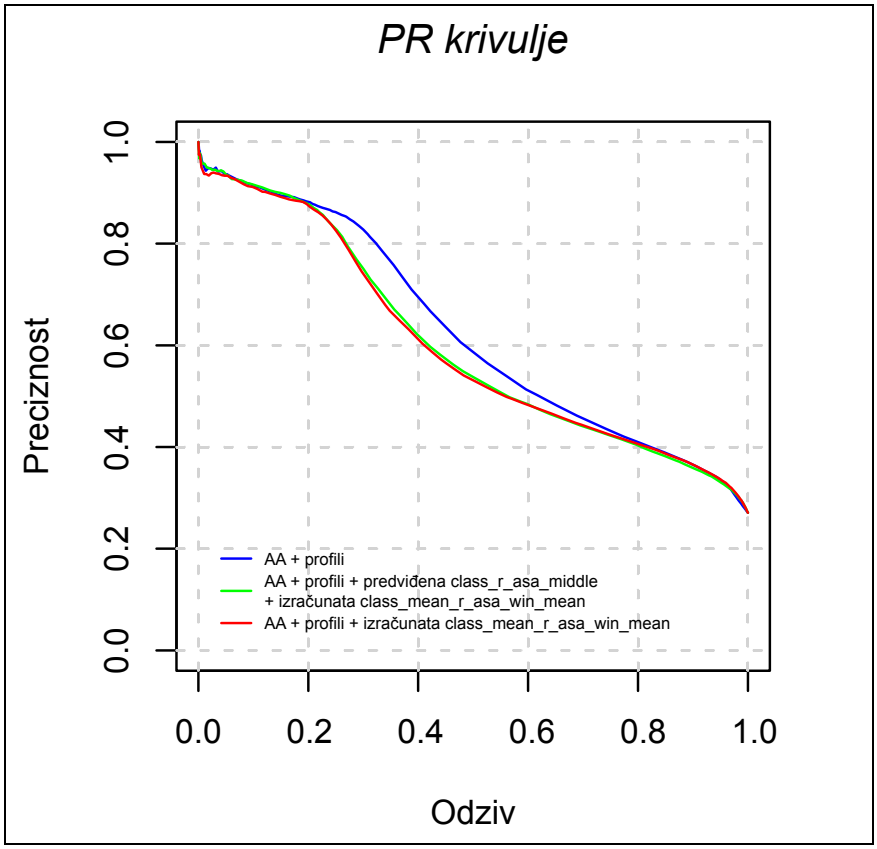
Zbog preglednosti rezultata na svim grafovima će najbolji dobiveni rezultat uvijek biti prikazan crvenom bojom. Također će na svim grafovima, radi usporedbe, u plavoj boji biti prikazana i krivulja dosadašnjeg najboljeg rezultata predviđanja (V. Dragosavljević [1]) čije poboljšanje je cilj ovog rada.

5.4.1 Rezultati predviđanja uz korištenje informacije iz slijeda, profila slijeda te izračunate srednje vrijednosti RASA-e prozora

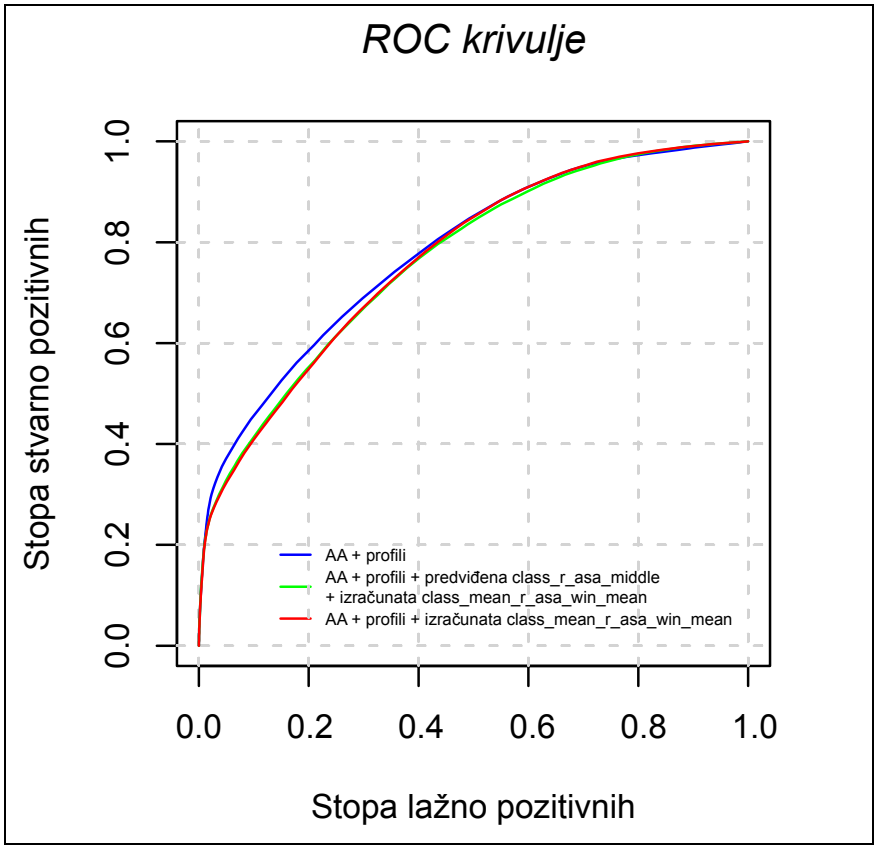
Obzirom da se atribut *class_mean_r_asa_win_mean* računa posredno, pomoću atributa *class_r_asa_middle* predviđenog za promatrani prozor te prethodna i naredna 4 prozora (kako je opisano u poglavlju 5.3), obratit će se pažnja na dvije varijante predviđanja:

- predviđanje mjesta kontakta iz slijeda aminokiselinskih ostataka, profila slijeda, predviđenih *class_r_asa_middle* i iz njih izračunate *class_mean_r_asa_win_mean*
- predviđanje mjesta kontakta iz slijeda aminokiselinskih ostataka, profila slijeda i izračunate *class_mean_r_asa_win_mean*.

Na slikama 5.9 i 5.10 su prikazane dobivene PR i ROC krivulje navedenih predviđanja urađenih na skupu *2b*, a na slikama 5.11 i 5.12 za skup *3b*.



Slika 5.9 PR krivulje predviđanja mjesta interakcije metodom krosvalidacije koristeći a) samo slijed aminokiselinskih ostataka i njihove profile te kombinaciju sa b) predviđenom *class_r_asa_middle* (dvije klase, Maneshovi pragovi) i izračunatom *class_mean_r_asa_win_mean*, c) izračunatom *class_mean_r_asa_win_mean*

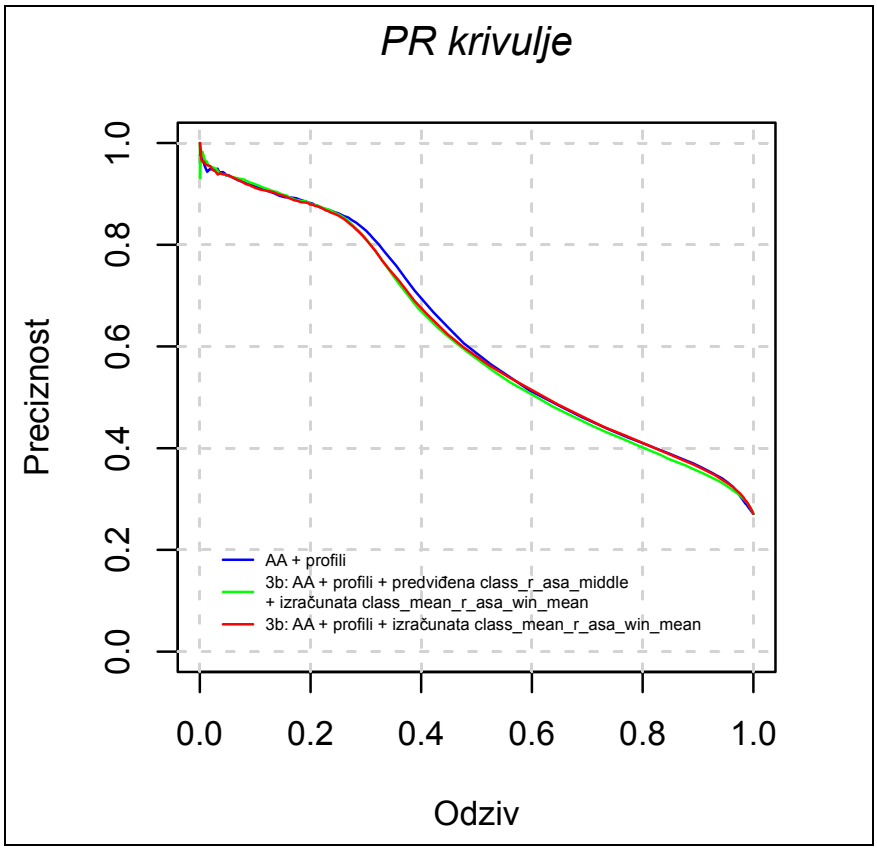


Slika 5.10 ROC krivulje predviđanja mjesta interakcije metodom krosvalidacije koristeći a) samo slijed aminokiselinskih ostataka i njihove profile te kombinaciju sa b) predviđenom *class_r_asa_middle* (dvije klase, Maneshovi pragovi) i izračunatom *class_mean_r_asa_win_mean*, c) izračunatom *class_mean_r_asa_win_mean*

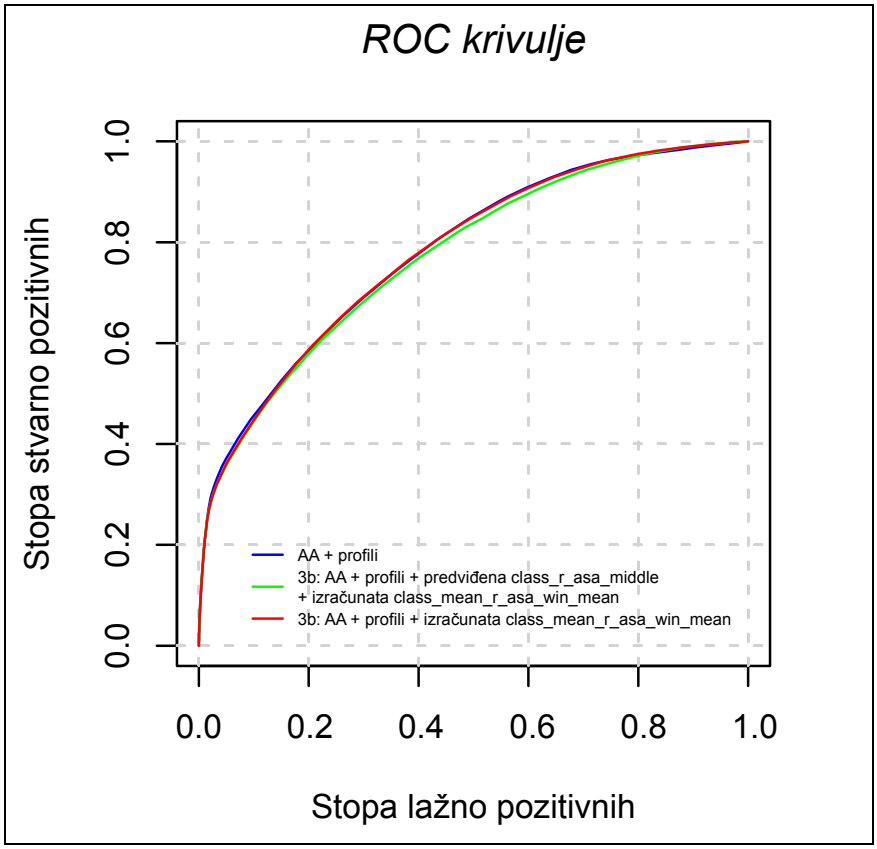
Iako su zelena i crvena krivulja vrlo slične, crvena – samo izračunata *class_mean_r_asa_win_mean* – ima bolju uspješnost predviđanja, što se vidi i iz tablice 5.5. Suprotno očekivanjima, oba nova predviđanja pokazala su se lošijima od originalnog kod kojeg se za predviđanje koristi samo informacija iz slijeda te profili slijeda.

	AUC	F-mjera	Točnost	Preciznost	Odziv
a)	0,781	0,4062	0,7891	0,8543	0,2664
b)	0, 7673	0, 4656	0, 7783	0, 6701	0, 3567
c)	0, 769	0, 4532	0, 7778	0, 6791	0, 3401

Tablica 5.5 Mjere uspješnosti predviđanja mjesta interakcije metodom krosvalidacije koristeći a) samo slijed aminokiselinskih ostataka i njihove profile te kombinaciju sa b) predviđenom *class_r_asa_middle* (dvije klase, Maneshovi pragovi) i izračunatom *class_mean_r_asa_win_mean*, c) izračunatom *class_mean_r_asa_win_mean*



Slika 5.11 PR krivulje predviđanja mjesta interakcije metodom krosvalidacije koristeći a) samo slijed aminokiselinskih ostataka i njihove profile te kombinaciju sa b) predviđenom *class_r_asa_middle* (tri klase, Maneshovi pragovi) i izračunatom *class_mean_r_asa_win_mean*, c) izračunatom *class_mean_r_asa_win_mean*



Slika 5.12 ROC krivulje predviđanja mjesta interakcije metodom krosvalidacije koristeći a) samo slijed aminokiselinskih ostataka i njihove profile te kombinaciju sa b) predviđenom *class_r_asa_middle* (tri klase, Maneshovi pragovi) i izračunatom *class_mean_r_asa_win_mean*, c) izračunatom *class_mean_r_asa_win_mean*

Crvena krivulja koja opisuje predviđanje kod kojeg se dodatno koristila samo izračunata *class_mean_r_asa_win_mean* opet prikazuje uspješnije predviđanje od zelene krivulje, što se vidi i iz tablice 5.6. Međutim, oba nova predviđanja, i s klasifikacijom atributa *class_r_asa_middle* u tri klase, su se pokazala lošijima od originalnog predviđanja. Ipak, predviđanje je očigledno bilo uspješnije nego u slučaju klasifikacije atributa *class_r_asa_middle* u dvije klase.

	AUC	F-mjera	Točnost	Preciznost	Odziv
a)	0,781	0,4062	0,7891	0,8543	0,2664
b)	0,7736	0, 402	0, 7881	0, 8513	0, 2631
c)	0, 7802	0, 3885	0, 7859	0, 8569	0, 2512

Tablica 5.6 Mjere uspješnosti predviđanja mjesta interakcije metodom krosvalidacije koristeći a) samo slijed aminokiselinskih ostataka i njihove profile te kombinaciju sa b) predviđenom *class_r_asa_middle* (tri klase, Maneshovi pragovi) i izračunatom *class_mean_r_asa_win_mean*, c) izračunatom *class_mean_r_asa_win_mean*

Kao što je ranije rečeno, uspješnost predviđanja kontakta direktno ovisi o tome koliko je bilo uspješno predviđanje atributa *class_r_asa_middle*. U prethodnom poglavlju se pokazalo da je uz poznavanje prave vrijednosti atributa *class_mean_r_asa_win_mean* predviđanje mjesta kontakta vidno uspješnije od originalnog predviđanja (slike 5.3 i 5.4). Može se stoga zaključiti da je razlog lošijim rezultatima predviđanja mjesta kontakta u neuspješnom predviđanju atributa *class_r_asa_middle* iz kojeg se potom izračunava atribut *class_mean_r_asa_win_mean*. U prethodnom poglavlju tablicom 5.3 prikazane su točnosti klasificiranja atributa *class_r_asa_middle* za pojedini skup *out-of-bag* metodom. Tablicom 5.7 prikazane su točnosti klasificiranja za klasifikaciju metodom krosvalidacije u dvije i tri kategorije po Maneshovim pragovima [17]. Iz podataka u tablici postaje očito da klasifikacija s točnošću od skoro 85% nije dovoljno dobra za uspješno predviđanje mjesta kontakta, naprotiv – predviđanje kontakta je lošije nego bez novih atributa.

Za klasifikaciju u tri kategorije točnost je osjetno lošija nego za klasifikaciju u dvije kategorije, ali klasifikacija u tri kategorije očigledno značajnije doprinosi predviđanju mjesta kontakta, pa su ukupni rezultati predviđanja bolji u slučaju kad je *class_r_asa_middle* klasificirana u tri kategorije.

	2b	3b
Točnost	0,8455	0,7319

Tablica 5.7 Točnost klasificiranja atributa *class_r_asa_middle* metodom krosvalidacije koristeći samo slijed aminokiselinskih ostataka i njihove profile

Logično je i da je predviđanje korištenjem samo *class_mean_r_asa_win_mean* za oba skupa podataka dalo bolje rezultate od predviđanja uz korištenje *class_r_asa_middle* – srednje vrijednosti devet realnih brojeva (klasa se pretvori u odgovarajući realni broj) ublažava pogrešku nastalu klasificiranjem tih devet vrijednosti.

Nakon predviđanja pokušalo se poboljšati dobivene rezultate postavljanjem težina prilikom klasifikacije, što pomiče prag klasifikacije u korist jedne klase. Primjerice, kod klasifikacije unutar dvije kategorije prag je po definiciji 50% glasova (ukoliko se eksplicitno ne zada drugačije) što znači da se svaka instanca koja dobije manje od pola glasova proglašava klasom 1, a sve ostale klasom 2. Težine takvih dviju kategorija su jednake. Zadavanje novih težina pomiče prag klasifikacije po kategorijama. Težine se navode za svaku kategoriju, veća težina odgovara većoj vjerojatnosti pojavljivanja neke instance u toj kategoriji.

U ovom radu težine su se zadavale pri klasifikaciji atributa *class_r_asa_middle* i to za skup *3b* kojim su postignuti najbolji rezultati. Za određivanje težina korišteni su statistički testovi χ^2-test (poglavlje 4.4.1) i omjer vjerojatnosti odnosno z-test (4.4.2). Rezultati testova prikazani su tablicama 5.8 i 5.9.

	klasa 1	klasa 2	klasa 3
is_contact = 0	44487	64344	15283
is_contact = 1	5826	30487	9765

Tablica 5.8 Raspored atributa *class_r_asa_middle* u odnosu na pojavu mjesta kontakta za skup *3b*

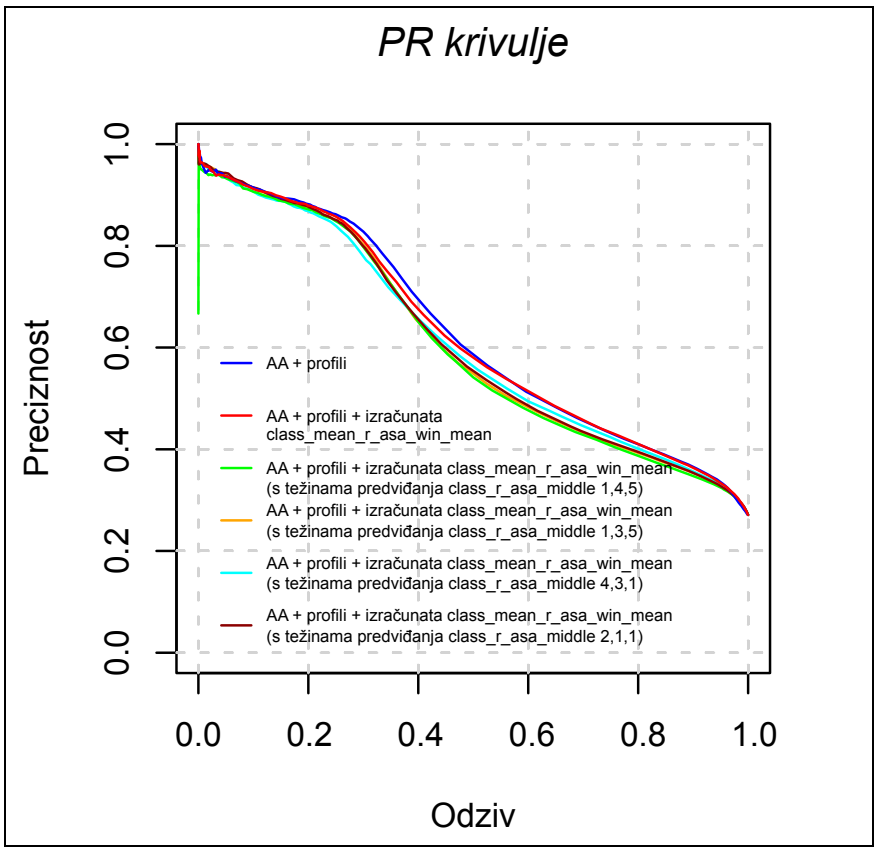
$\chi^2 = 9154{,}664$	**P < 5%**
$\hat{\theta}_{1/2} = 0{,}276$	z = -82.583
$\hat{\theta}_{1/3} = 0{,}205$	z = -83,306
$\hat{\theta}_{2/3} = \mathbf{0,}742$	z = -20,336

Tablica 5.9 Utjecaj atributa *class_r_asa_middle* na pojavu mjesta kontakta za skup *3b*

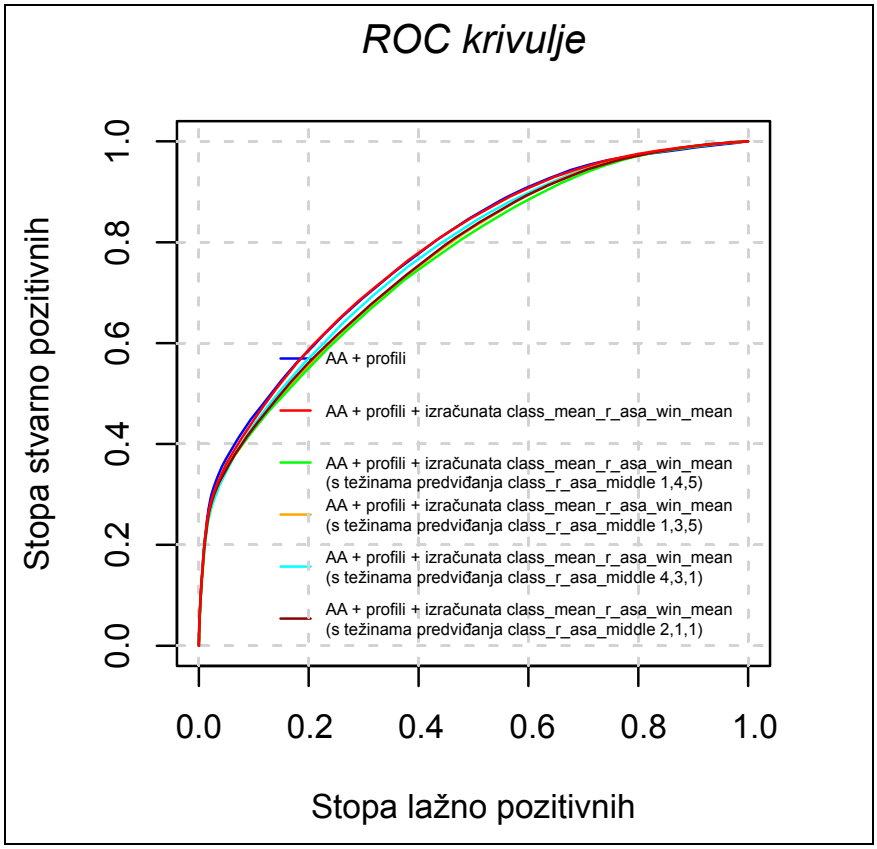
Iz tablica je vidljivo da su varijable *is_contact* i *class_r_asa_middle* međusobno zavisne. Obzirom na omjere vjerojatnosti među kategorijama atributa

class_r_asa_middle učinjeno je još nekoliko predviđanja mjesta kontakta, ali ovaj put su u prvom krugu predviđanja, pri predviđanju atributa *class_r_asa_middle*, postavljene težine. Kako su, suprotno očekivanjima, dobiveni nepoželjni rezultati, dodatno su učinjena i testiranja s potpuno drugačijim omjerima težina. Iako su se pokazali nešto boljima od prvih, ipak niti jedna kombinacija postavljenih težina nije u konačnici pridonijela predviđanju mjesta kontakta, stoga najbolja kombinacija ostaje ranije prikazano predviđanje uz *class_mean_r_asa_win_mean* izračunatu iz predviđene *class_r_asa_middle*, predviđenje klasično, bez eksplicitnih zadavanja težina (crvena krivulja na slikama 5.11–5.12).

Rezultati svih predviđanja sa zadanim težinama u prvom krugu predviđanja prikazani su slikama 5.13 i 5.14.



Slika 5.13 PR krivulje predviđanja mjesta interakcije metodom krosvalidacije koristeći slijed aminokiselinskih ostataka i njihove profile te *class_mean_r_asa_win_mean* izračunatu iz *class_r_asa_middle* predviđenu s postavljenim težinama

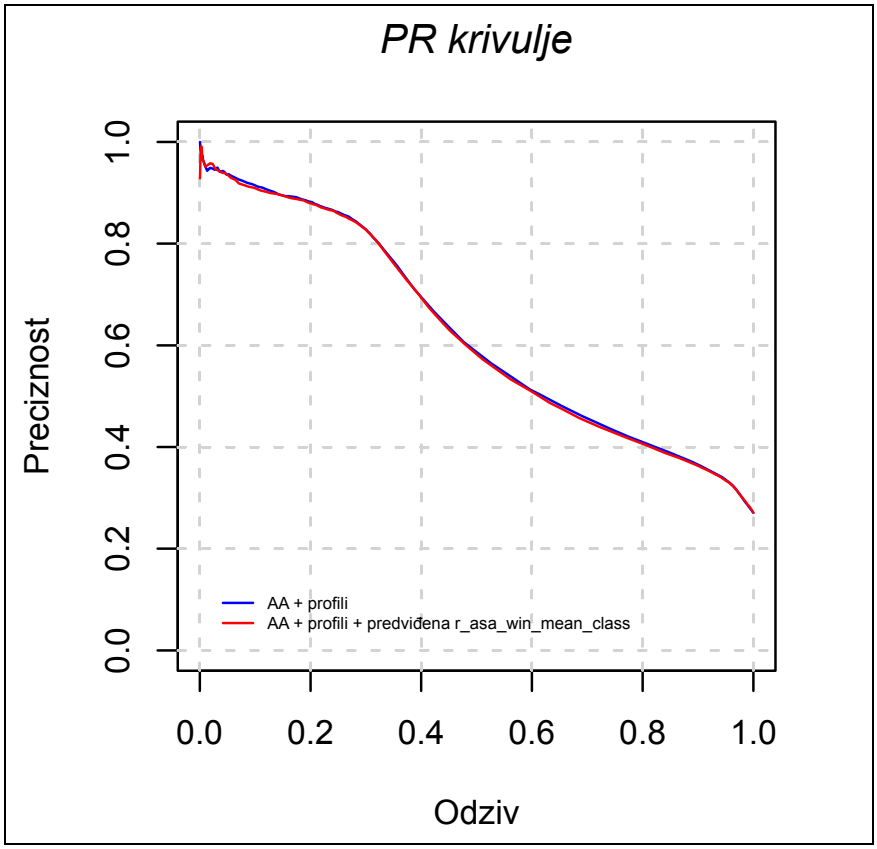


Slika 5.14 ROC krivulje predviđanja mjesta interakcije metodom krosvalidacije koristeći slijed aminokiselinskih ostataka i njihove profile te *class_mean_r_asa_win_mean* izračunatu iz *class_r_asa_middle* predviđenu s postavljenim težinama

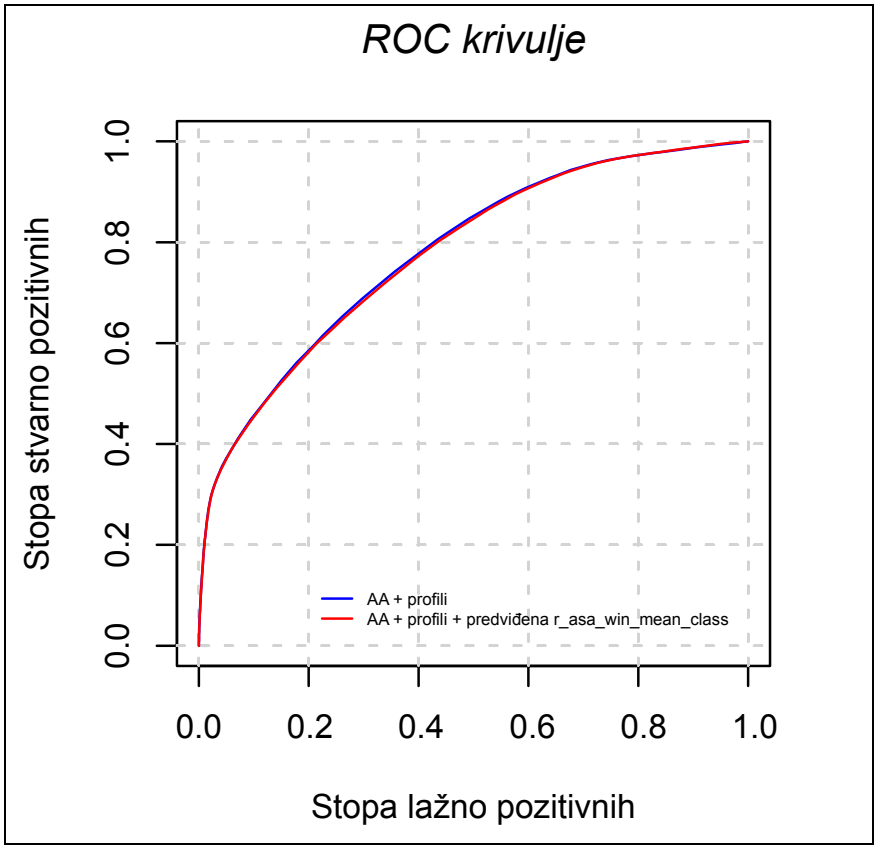
5.4.2 Rezultati predviđanja uz korištenje informacije iz slijeda, profila slijeda te predviđene klase srednje vrijednosti RASA-e prozora

Utjecaj predviđene klase srednje vrijednosti RASA-e prozora (*r_asa_win_mean_class*) na predviđanje kontakta također je proučen za skup *2b* i *3b*. U prvom krugu predviđanja klasificiran je *r_asa_win_mean_class* u jednu od dvije odnosno tri kategorije, a u drugom krugu predviđanja se za svaki pomični prozor uz informaciju iz slijeda i profil slijeda koristi i predviđena vrijednost novog atributa.

Rezultati predviđanja mjesta kontakta dani su slikama 5.15 i 5.16 te tablicom 5.10 za skup *2b* odnosno slikama 5.17 i 5.18 te tablicom 5.11 za skup *3b*.



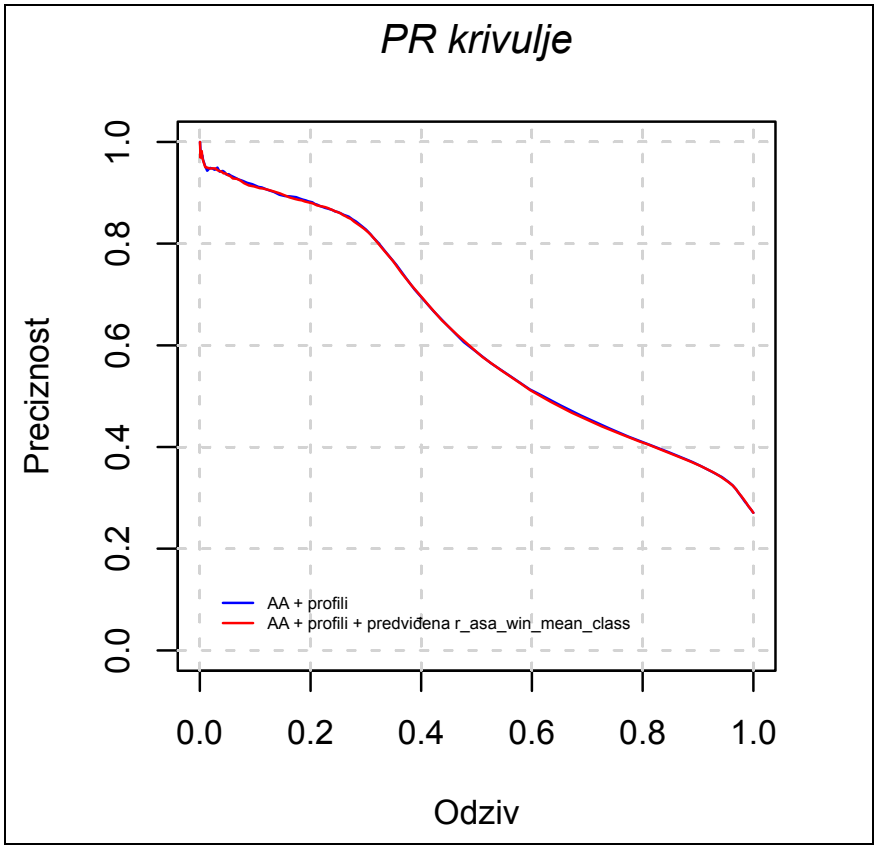
Slika 5.15 PR krivulje predviđanja mjesta interakcije metodom krosvalidacije koristeći a) samo slijed aminokiselinskih ostataka i njihove profile te kombinaciju sa b) predviđenom *r_asa_win_mean_class* (dvije klase, Maneshovi pragovi)



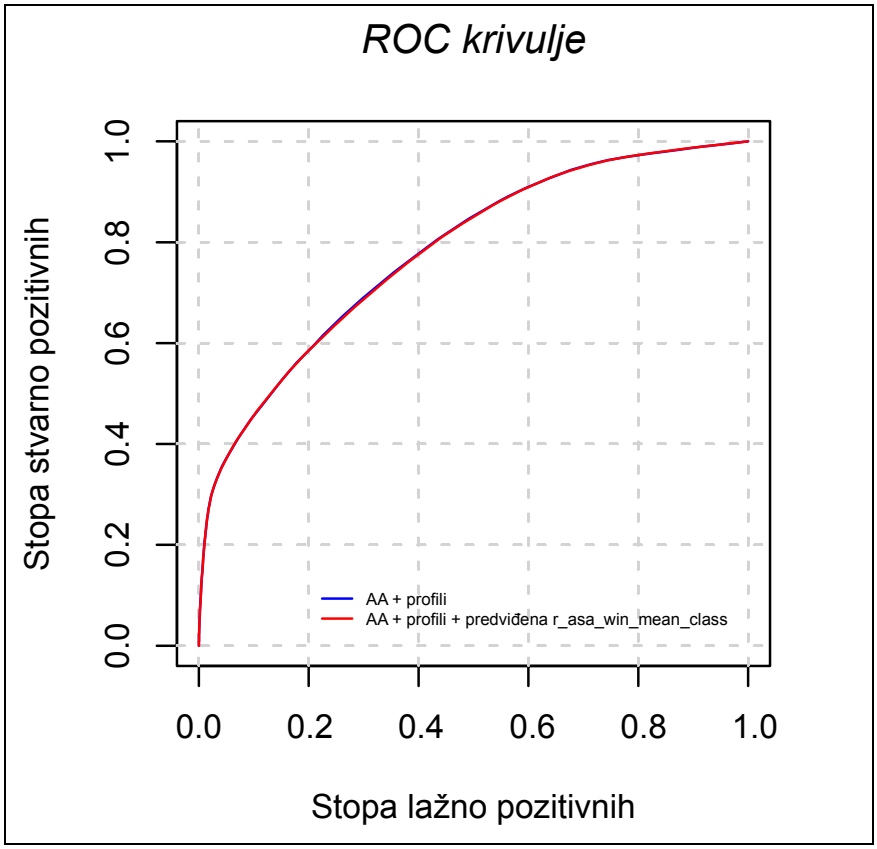
Slika 5.16 ROC krivulje predviđanja mjesta interakcije metodom krosvalidacije koristeći a) samo slijed aminokiselinskih ostataka i njihove profile te kombinaciju sa b) predviđenom *r_asa_win_mean_class* (dvije klase, Maneshovi pragovi)

	AUC	F-mjera	Točnost	Preciznost	Odziv
a)	0,781	0,4062	0,7891	0,8543	0,2664
b)	0, 7784	0, 4098	0, 7893	0, 8485	0, 2701

Tablica 5.10 Mjere uspješnosti predviđanja mjesta interakcije metodom krosvalidacije koristeći a) samo slijed aminokiselinskih ostataka i njihove profile te kombinaciju sa b) predviđenom *r_asa_win_mean_class* (dvije klase, Maneshovi pragovi)



Slika 5.17 PR krivulje predviđanja mjesta interakcije metodom krosvalidacije koristeći a) samo slijed aminokiselinskih ostataka i njihove profile te kombinaciju sa b) predviđenom *r_asa_win_mean_class* (tri klase, Maneshovi pragovi)



Slika 5.18 ROC krivulje predviđanja mjesta interakcije metodom krosvalidacije koristeći a) samo slijed aminokiselinskih ostataka i njihove profile te kombinaciju sa b) predviđenom *r_asa_win_mean_class* (tri klase, Maneshovi pragovi)

	AUC	F-mjera	Točnost	Preciznost	Odziv
a)	0,781	0,4062	0,7891	0,8543	0,2664
b)	0, 7802	0, 4038	0, 7885	0, 8532	0, 2645

Tablica 5.11 Mjere uspješnosti predviđanja mjesta interakcije metodom krosvalidacije koristeći a) samo slijed aminokiselinskih ostataka i njihove profile te kombinaciju sa b) predviđenom *r_asa_win_mean_class* (tri klase, Maneshovi pragovi)

Rezultati predviđanja su generalno nešto bolji nego ranije dobiveni (poglavlje 5.4.1). Kao što je ranije rečeno, uspješnost predviđanja mjesta kontakta ovisi o uspješnosti predviđanja RASA atributa u prvom krugu, ovaj put to se odnosi na točnost predviđanja atributa *r_asa_win_mean_class*. Radi usporedbe se promatra točnost predviđanja atributa *r_asa_win_mean_class* naspram točnosti predviđanja atributa *class_r_asa_middle* korištenog za predviđanja mjesta

kontakta u prethodnom poglavlju. Pretpostavlja se da je točnost predviđanja *r_asa_win_mean_class* veća, što bi objasnilo općenito bolje rezultate dobivene korištenjem tog atributa. Rezultati prikazani u tablici 5.12 potvrđuju očekivanja.

	class_r_asa_middle	*r_asa_win_mean_class*
Točnost (2b)	0,8455	0,9559
Točnost (3b)	0,7319	0,9393

Tablica 5.12 Točnost klasificiranja atributa *class_r_asa_middle* i *r_asa_win_mean_class* u dvije i tri kategorije po Maneshovim pragovima metodom krosvalidacije koristeći samo slijed aminokiselinskih ostataka i njihove profile

Opet su predviđanja na skupu podataka *3b* bolja od onih na skupu *2b*, bez obzira što je točnost predviđanja RASA atributa bolja za dvije klase. I za ovaj primjer je količina informacije sadržana u pojavljivanju neke vrijednosti unutar tri kategorije bila važnija od one sadržane u pojavljivanju iste te vrijednosti unutar dvije kategorije, čak i po cijenu točnosti klasificiranja koja opada s brojem kategorija.

Podešavanje težina najprije se radilo za skup *2b* iako su na njemu dobiveni nešto lošiji rezultati. Budući da eksperimenti podešavanja težina nisu dali zadovoljavajuće rezultate u prethodnom poglavlju (slike 5.13 i 5.14), ovaj put se pristupilo drugačijim metodama podešavanja težina, koje su (radi ograničenja korištenih aplikacija) bile moguće samo za binarne klasifikatore.

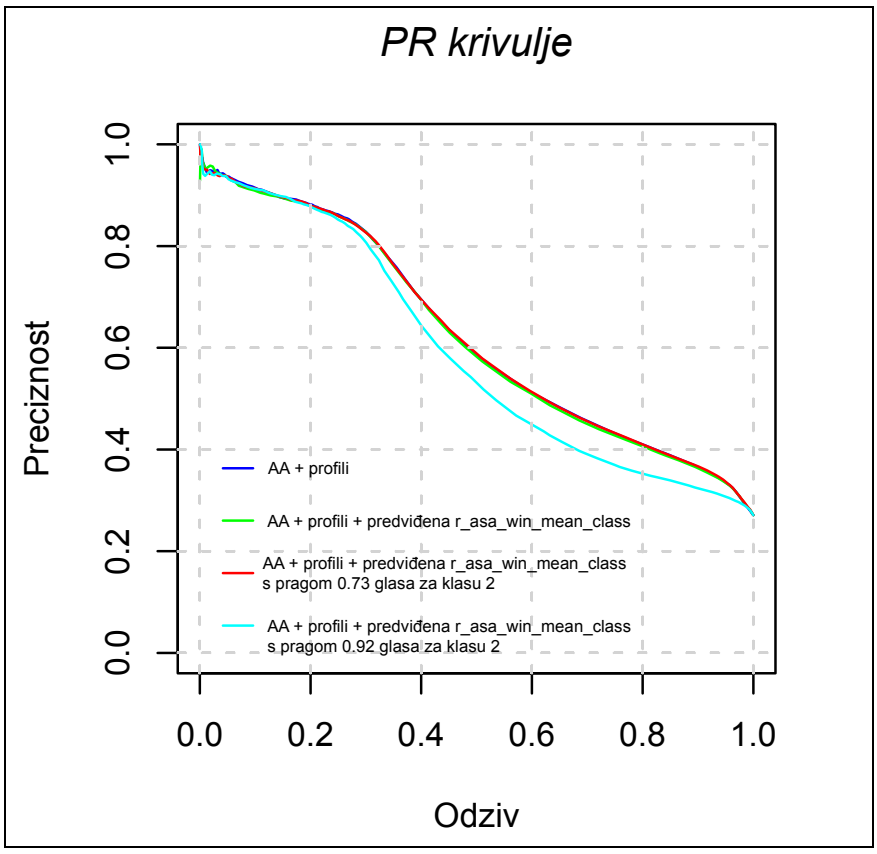
Proučavale su se vrijednosti koje opisuju točke ROC krivulje predviđanja atributa *r_asa_win_mean_class*. Ispitalo se kojim pragovima (pragovi idu od 0,00 do 1,00 s korakom od 0,01 – svakoj točki krivulje odgovara jedan prag) odgovaraju najbolje točke ROC krivulje. Ovaj postupak može se zamisliti kao "šetanje" po krivulji i traženje točke, odnosno pripadajućeg praga, u kojoj je ona najpovoljnija. Određene su dvije točke: točka za koju je točnost predviđanja najveća i točka koja je najbliža lijevom gornjem kutu ROC krivulje.

Matrica pogreške klasifikacije atributa *r_asa_win_mean_class* prikazana je tablicom 5.13. Prag dobiven za najveću točnost iznosi **0,73**, a prag dobiven iz točke ROC krivulje najbliže gornjem lijevom kutu iznosi **0,92** za skup podataka *2b*.

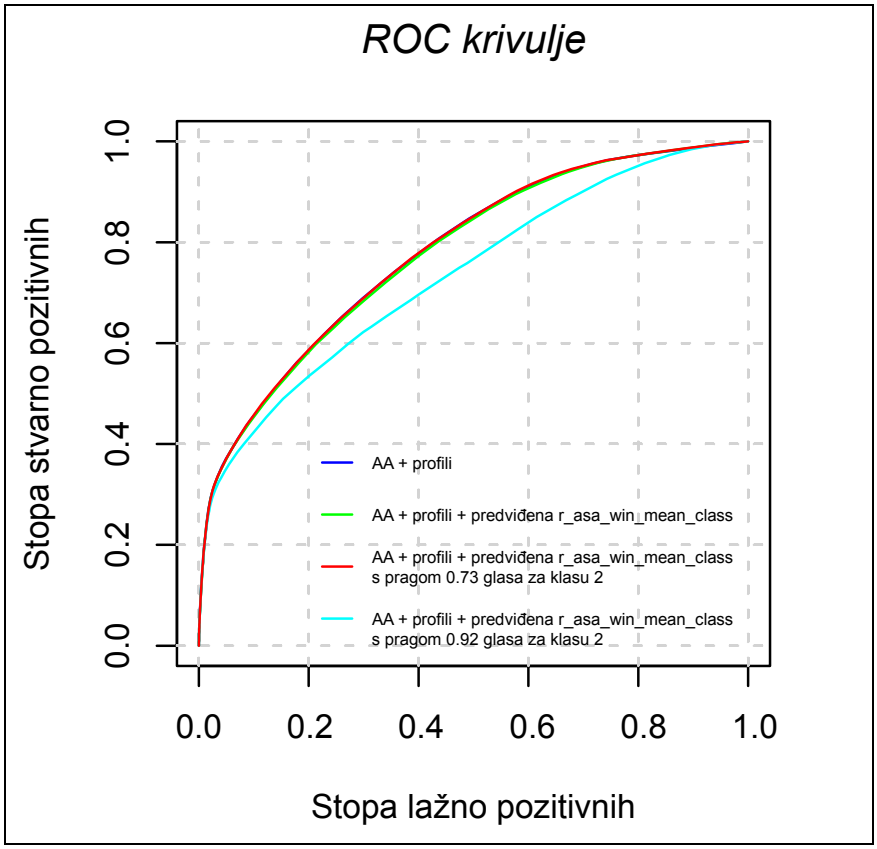
	Stvarna klasa 1	*Stvarna klasa 2*	**Total**
Predviđena klasa 1	2692	255	2947
Predviđena klasa 2	7255	159990	167245
Total	9947	160245	170192

Tablica 5.13 Matrica pogreške klasifikacije atributa *r_asa_win_mean_class* za skup *2b*

Glasovi iz prvog kruga predviđanja odredili su predviđeni RASA atribut, ali ovaj put je novi prag za klasu 2 iznosio jednom 73%, a drugi put 92% glasova. Predviđanja kontakta s tako dobivenim vrijednostima atributa *r_asa_win_mean_class* prikazana su slikama 5.19 i 5.20 te tablicom 5.14.



Slika 5.19 PR krivulje predviđanja mjesta interakcije metodom krosvalidacije koristeći a) samo slijed aminokiselinskih ostataka i njihove profile te kombinaciju sa predviđenom *r_asa_win_mean_class* (dvije klase, Maneshovi pragovi) uz prag klasificiranja RASA atributa postavljen na b) 0,5 (standardno), c) 0,73, d) 0,92



Slika 5.20 ROC krivulje predviđanja mjesta interakcije metodom krosvalidacije koristeći a) samo slijed aminokiselinskih ostataka i njihove profile te kombinaciju sa predviđenom *r_asa_win_mean_class* (dvije klase, Maneshovi pragovi) uz prag klasificiranja RASA atributa postavljen na b) 0,5 (standardno), c) 0,73, d) 0,92

	AUC	F-mjera	Točnost	Preciznost	Odziv
a)	0,781	0,4062	0,7891	0,8543	0,2664
b)	0, 7784	0, 4098	0, 7893	0, 8485	0, 2701
c)	0,7817	0,4102	0,7895	0,8496	0,2703
d)	0,7345	0,3917	0,786	0,8495	0,2545

Tablica 5.14 Mjere uspješnosti predviđanja mjesta interakcije metodom krosvalidacije koristeći a) samo slijed aminokiselinskih ostataka i njihove profile te kombinaciju s predviđenom *r_asa_win_mean_class* (dvije klase, Maneshovi pragovi) uz prag klasificiranja RASA atributa postavljen na b) 0,5 (standardno), c) 0,73, d) 0,92

Postavljanje vrijednosti praga za klasu 2 iznad uobičajene rezultiralo je boljim rezultatima predviđanja mjesta kontakta, iako jako visok prag nije dobro rješenje jer daje znatno lošije konačne rezultate. Tako su najbolji rezultati dobiveni za

predviđanje kod kojeg je u prvom krugu prag proglašavanja atributa *r_asa_win_mean_class* neke instance klasom 2 jednak 73% ukupnih glasova. To su ujedno i prvi rezultati koji su se pokazali boljima od onih koje je dobila V. Dragosavljević [1], iako je razlika uspješnosti statistički neznačajna. Obzirom da se korištenjem postojeće aplikacije analogni postupak određivanja pragova nije mogao provesti i za tri kategorije, rezultati dobiveni skupom *2b*, uz pragove glasova za klasu 2 postavljene na 73%, smatrat će se najboljima među predviđanjima mjesta kontakta korištenjem predviđene *r_asa_win_mean_class*.

5.4.3 Proširivanje vektora ulaznih atributa

Već iz prvih rezultata, dobivenih pri traženju atributa koji će se koristiti za predviđanje mjesta interakcije (poglavlje 5.2.), vidi se da su za predviđanje značajniji oni atributi koji opisuju RASA-u prozora od onih koji opisuju RASA-u pojedinačnog, središnjeg aminokiselinskog ostatka prozora. Nameće se pitanje je li moguće dobiti bolje rezultate predviđanja kontakta ako bi korišteni RASA atributi opisivali još šire područje od 9 aminokiselinskih ostataka koje pokriva pomični prozor korišten za predviđanje kontakta. Jedan način ispitivanja ove hipoteze je sljedeći: u prvom krugu predviđanja klase RASA atributa koristiti veći pomični prozor (npr. 13 aminokiselinskih ostataka), potom te RASA atribute pridijeliti manjem pomičnom prozoru (9 ostataka) koji se koristi u drugom krugu predviđanja. Međutim, predviđanja RASA atributa korištenjem većih pomičnih prozora traju iznimno dugo. Radi vremenskih ograničenja, u ovom radu korištena je drugačija metoda: ulazni RASA atribut proširen je na vektor od 9 takvih atributa koji se odnose redom na 9 aminokiselinskih ostataka promatranog prozora (analogno vektorima profila). Središnjem ostatku je pridružena klasa predviđena za taj prozor (dosad je ovo bio jedini atribut), a ostalim ostacima prozora su pridružene vrijednosti predviđene za prethodna, odnosno naredna 4 pomična prozora u kojima su ti ostaci bili središnji.

U procesu proširivanja ulaznih atributa, same podatke je bilo potrebno mijenjati, najčešće skraćivati. Primjerice, prvom pomičnom prozoru zapisanom u datoteci nije moguće pridružiti sve atribute, jer prva četiri aminokiselinska ostatka tog

prozora nijednom prozoru nisu bila središnja, pa za njih klase nisu bile ni predviđane. Ovakve i slične situacije rješavale su se nadomještanjem nepoznatih vrijednosti nekim procijenjenim vrijednostima ili naprosto izbacivanjem takvih prozora iz skupa. Svaki put kad su se ulazni podaci mijenjali, na dobivenom se skupu uradilo i jedno predviđanje kontakta samo korištenjem slijeda i profila. Na taj način se svaki novi rezultat promatra u odnosu na njemu usporedivi rezultat koji odgovara onom kojeg je dobila V. Dragosavljević [1].

5.4.3.1 Rezultati predviđanja uz korištenje informacije iz slijeda, profila slijeda te vektora predviđenih klasa RASA-a svih aminokiselinskih ostataka prozora

Prvo proširivanje ulaznih atributa je urađeno za atribut *class_r_asa_middle*. Za promatrani pomični prozor predviđena vrijednost atributa *class_r_asa_middle* postaje *class_r_asa5* (središnja vrijednost 9-dimenzionalnog vektora). Atributima *class_r_asa1*–*class_r_asa4* pridjeljuju se vrijednosti atributa *class_r_asa_middle* predviđenih za prethodna 4 pomična prozora, a *class_r_asa6*–*class_r_asa9* za naredna 4 pomična prozora. Na taj način je svaki aminokiselinski ostatak pomičnog prozora opisan svojim RASA atributom. Vektor ulaznih podataka sada je dimenzija 22×9.

Vlastito izrađenim programima stvaraju se novi skupovi podataka. Korišteni skup je iz grupe *3b* jer su za njega u poglavlju 5.4.1 dobiveni najbolji rezultati.

Osim klasom, novi atributi se u predviđanjima mogu nadomjestiti numeričkom vrijednošću koja se pridjeljuje ne samo ovisno o klasi, nego i ovisno o tipu aminokiselinskog ostatka.

Testirane su sljedeće tri kombinacije, svaka od njih za RASA atribute nezavisne o tipu ostatka te za RASA atribute zavisne o tipu ostatka:

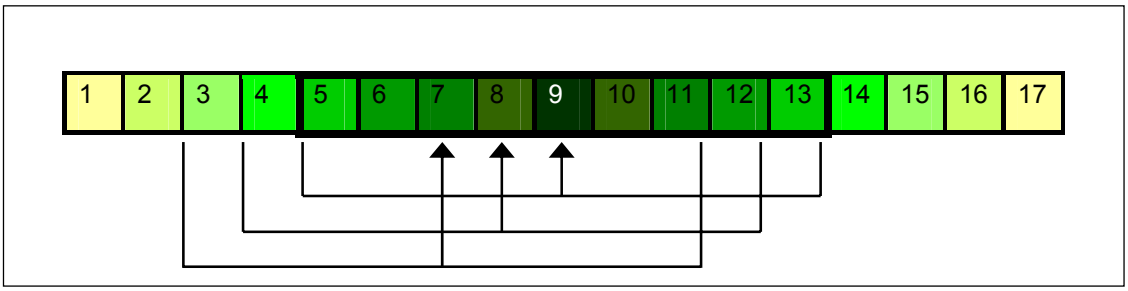
1. ulazni RASA atributi čine 9-dimenzionalni vektor vrijednosti predviđenih RASA klasa svih aminokiselinskih ostataka promatranog prozora (RASA klase zavisne o tipu ostatka predstavljene su numeričkom vrijednošću koja

opisuje srednju zastupljenost RASA vrijednosti promatranog ostatka unutar neke klase),

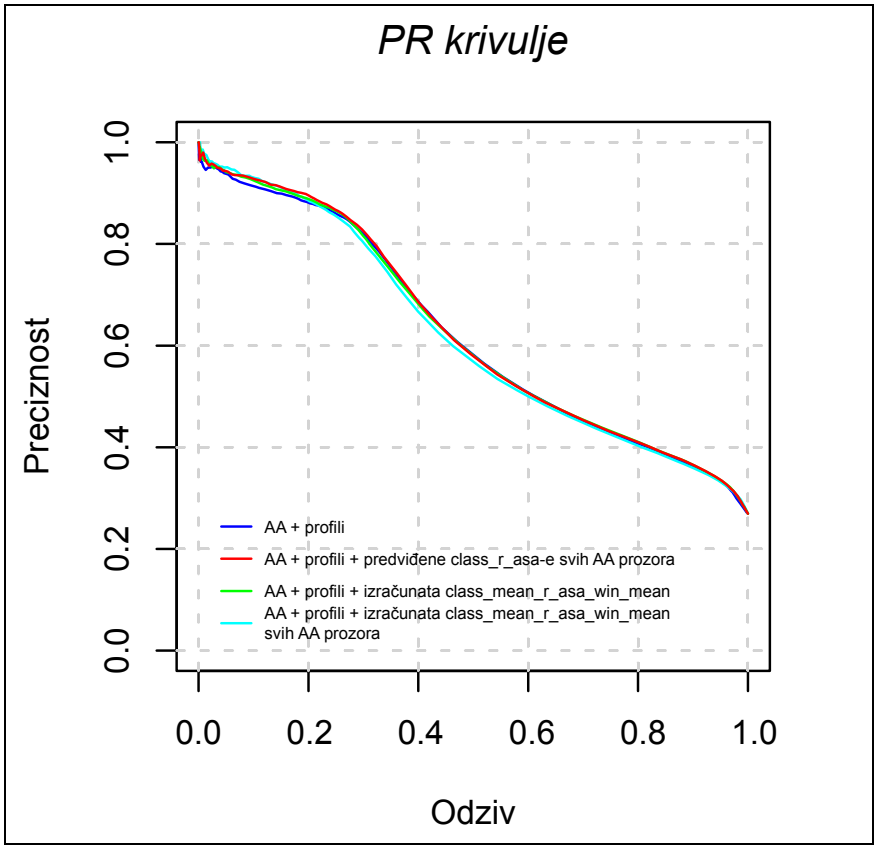
2. ulazni RASA atribut je jedna izračunata numerička vrijednost – aritmetička sredina gore opisanih devet vrijednosti: kod RASA atributa nezavisnih o tipu ostatka ta se vrijednost računa analogno računanju atributa *class_mean_r_asa_win_mean* (poglavlje 3.2.2.1) – broj koji predstavlja klasu prvo se pretvori u odgovarajuću numeričku vrijednost koja opisuje srednju zastupljenost RASA vrijednosti svih poznatih aminokiselina unutar neke klase,
3. ulazni RASA atributi čine 9-dimenzionalni vektor izračunatih vrijednosti – gore opisan atribut se proširuje u vektor, analogno proširivanju atributa *class_r_asa_middle*.

Atribut definiran točkom 3. je u pravom smislu proširivanje okoline koja utječe na predviđanje mjesta kontakta. Vrijednost koja opisuje cijeli pomični prozor pridijeljena je njegovom središnjem ostatku. Na taj način novih 9 atributa u sebi indirektno sadrže informaciju o RASA vrijednosti 17 aminokiselinskih ostataka. Za središnji među njima predviđa se je li mjesto kontakta. Pritom je utjecaj informacije središnjeg ostatka najjači (njegova predviđena RASA vrijednost utjecala je na svih 9 novih izračunatih!), što opada prema rubovima "velikog" prozora (slika 5.21). Informaciju o rubnim ostacima nosi tek po 1 od 9 ulaznih RASA atributa.

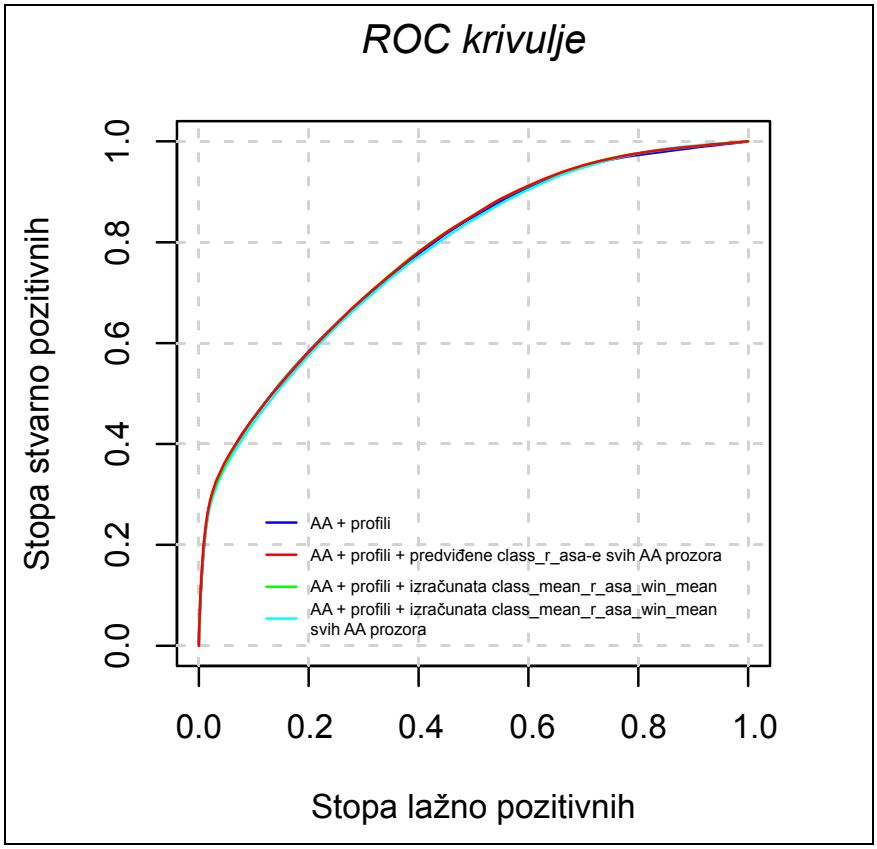
Rezultati novih predviđanja su zbog preglednosti prikazani u dva dijela: ovisno o tome tretira li se RASA atribut kao nezavisan (slike 5.22 i 5.23 te tablica 5.15) ili kao zavisan (slike 5.24 i 5.25 te tablica 5.16) o tipu aminokiselinskog ostatka.



Slika 5.21 Proširivanje ulaznih atributa – okolina koja svojim RASA vrijednostima utječe na iznose 9 atributa definiranih točkom 3: svjetlije su prikazani ostaci kojih je utjecaj manji, tamnije oni kojih je veći; promatrani prozor čine ostaci 5–13, za ostatak 9 se predviđa je li u interakciji



Slika 5.22 PR krivulje predviđanja mjesta interakcije metodom krosvalidacije koristeći a) samo slijed aminokiselinskih ostataka i njihove profile te kombinaciju sa b) predviđenim *class_r_asa*-ma svih ostataka prozora (neovisno o tipu ostatka), c) iz njih izračunatom *class_mean_r_asa_win_mean*, d) izračunatom *class_mean_r_asa_win_mean* svih ostataka prozora



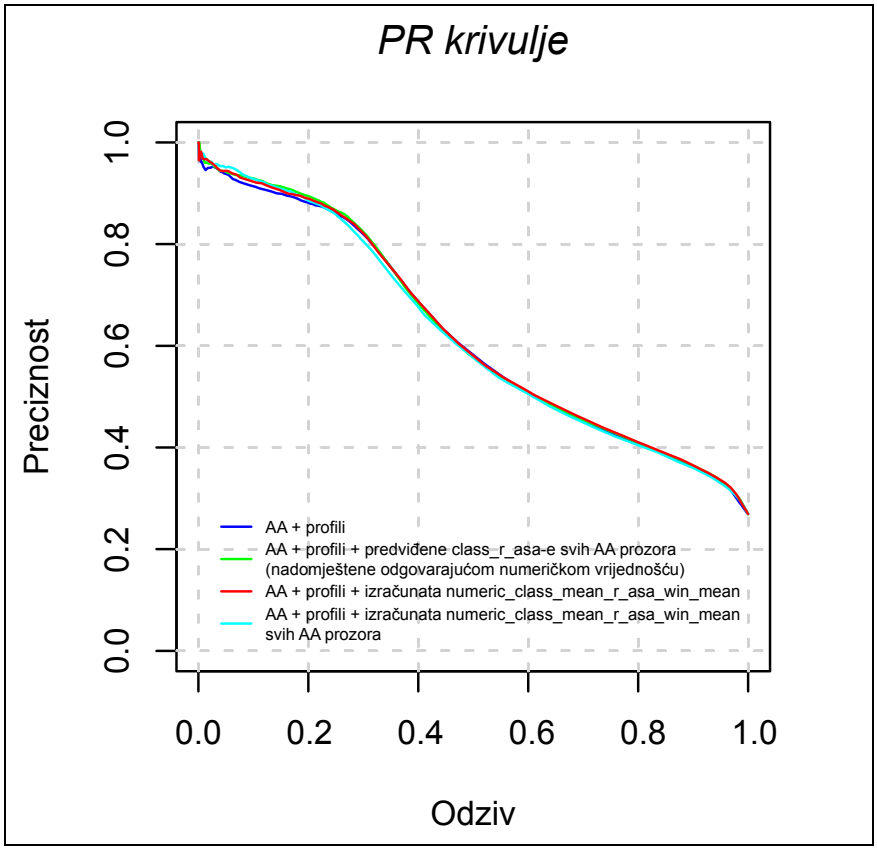
Slika 5.23 ROC krivulje predviđanja mjesta interakcije metodom krosvalidacije koristeći a) samo slijed aminokiselinskih ostataka i njihove profile te kombinaciju sa b) predviđenim *class_r_asa*-ma svih ostataka prozora (neovisno o tipu ostatka), c) iz njih izračunatom *class_mean_r_asa_win_mean*, d) izračunatom *class_mean_r_asa_win_mean* svih ostataka prozora

	AUC	F-mjera	Točnost	Preciznost	Odziv
a)	0, 7797	0, 3991	0, 7889	0, 8541	0, 2604
b)	0, 7818	0, 3994	0, 7895	0, 8608	0, 2601
c)	0, 7817	0, 3947	0, 7886	0, 8616	0, 256
d)	0,7763	0,3808	0,7858	0,8586	0,2447

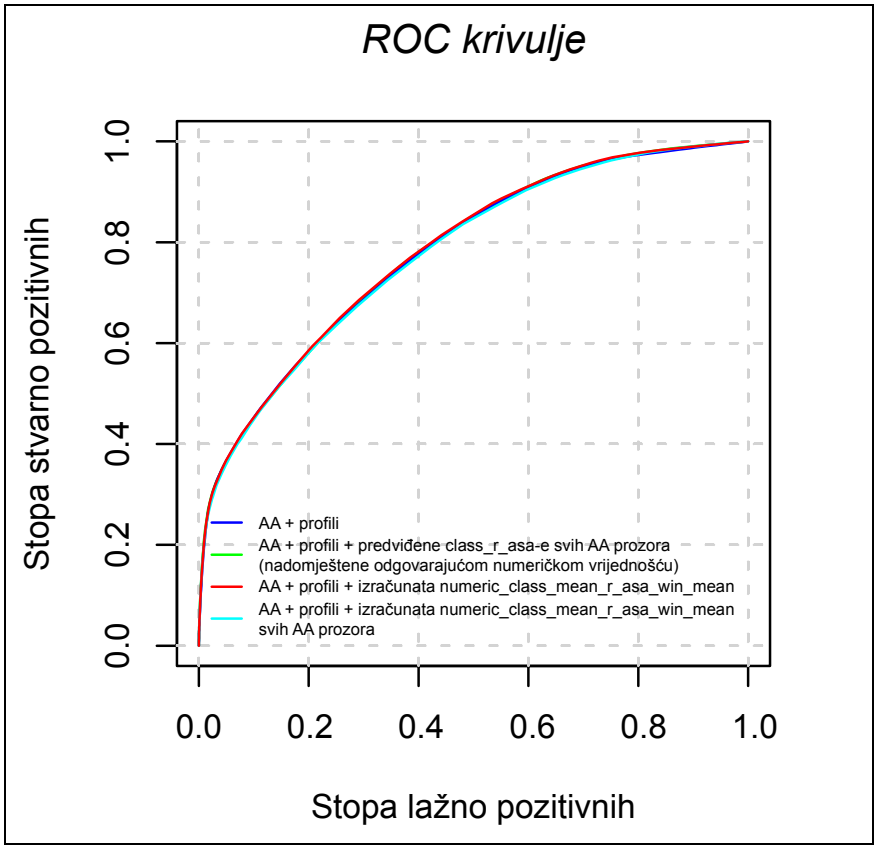
Tablica 5.15 Mjere uspješnosti predviđanja mjesta interakcije metodom krosvalidacije koristeći a) samo slijed aminokiselinskih ostataka i njihove profile te kombinaciju sa b) predviđenim *class_r_asa*-ma svih ostataka prozora (neovisno o tipu ostatka), c) iz njih izračunatom *class_mean_r_asa_win_mean*, d) izračunatom *class_mean_r_asa_win_mean* svih ostataka prozora

Među dobivenim rezultatima najboljim se pokazalo predviđanje urađeno korištenjem predviđenih *class_r_asa*-a svih ostataka prozora. Vrijednost AUC

pokazuje da je uspješnost predviđanja 0,21% bolja nego korištenjem samo slijeda aminokiselinskih ostataka i njihovih profila. Očigledno je da srednja vrijednost ovih 9 predviđenih vrijednosti ne nosi značajniju informaciju (stoga su rezultati c) bolji od b)), čak ni kad se ona proširi u 9-dimenzionalni vektor. Dapače, rezultati dobiveni točkom d) su najlošiji što znači da šira RASA okolina za predviđanje mjesta kontakta ili vodi do pogrešnih zaključaka, ili je metoda određivanja te RASA okoline pogrešna.



Slika 5.24 PR krivulje predviđanja mjesta interakcije metodom krosvalidacije koristeći a) samo slijed aminokiselinskih ostataka i njihove profile te kombinaciju sa b) predviđenim *class_r_asa*-ma svih ostataka prozora (zamijenjenima numeričkim vrijednostima ovisno o tipu ostatka), c) iz njih izračunatom *numeric_mean_r_asa_win_mean*, d) izračunate *numeric_mean_r_asa_win_mean* svih ostataka prozora



Slika 5.25 ROC krivulje predviđanja mjesta interakcije metodom krosvalidacije koristeći a) samo slijed aminokiselinskih ostataka i njihove profile te kombinaciju sa b) predviđenim *class_r_asa*-ma svih ostataka prozora (zamijenjenima numeričkim vrijednostima ovisno o tipu ostatka), c) iz njih izračunatom *numeric_mean_r_asa_win_mean*, d) izračunate *numeric_mean_r_asa_win_mean* svih ostataka prozora

	AUC	F-mjera	Točnost	Preciznost	Odziv
a)	0, 7797	0, 3991	0, 7889	0, 8541	0, 2604
b)	0, 7815	0, 4014	0, 7899	0, 8612	0, 2616
c)	0, 7827	0, 3956	0, 7886	0, 8592	0, 2569
d)	0,7773	0,3828	0,7864	0,8612	0,2461

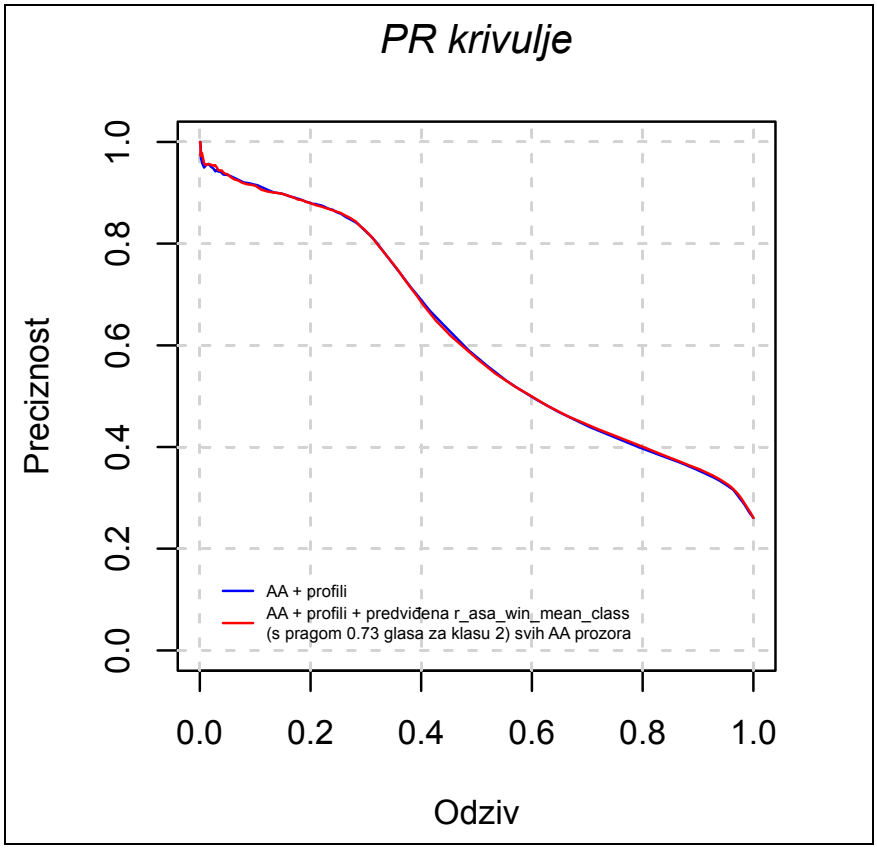
Tablica 5.16 Mjere uspješnosti predviđanja mjesta interakcije metodom krosvalidacije koristeći a) samo slijed aminokiselinskih ostataka i njihove profile te kombinaciju sa b) predviđenim *class_r_asa*-ma svih ostataka prozora (zamijenjenima numeričkim vrijednostima ovisno o tipu ostatka), c) iz njih izračunatom *numeric_mean_r_asa_win_mean*, d) izračunate *numeric_mean_r_asa_win_mean* svih ostataka prozora

Proučavanjem dobivenih rezultata vidi se da je predviđena class_r_asa svih ostataka prozora pridonijela predviđanju mjesta kontakta, ali nešto manje nego u slučaju kada se ona određivala kao diskretni broj iz skupa {1, 2} neovisan o tipu aminokiselinskog ostatka za koji je predviđen. Međutim, izračun srednje vrijednosti među tih 9 ima bolji utjecaj na predviđanje mjesta kontakta nego kod izračuna neovisnog o tipu ostaka. Štoviše, korištenjem ovako izračunate *numeric_mean_r_asa_win_mean* dobiveni su najbolji rezultati predviđanja (točka c)), prema AUC vrijednosti za 0,3% bolji od onih dobivenih samo korištenjem slijeda aminokiselinskih ostataka i njihovih profila. Međutim, čak i ovom slučaju, uključivanjem više tako dobivenih vrijednosti uspješnost predviđanja opada (točka d)) i daje najlošije predviđanje mjesta kontakta. Ovakav napredak (u oba slučaja) još uvijek je statistički zanemariv.

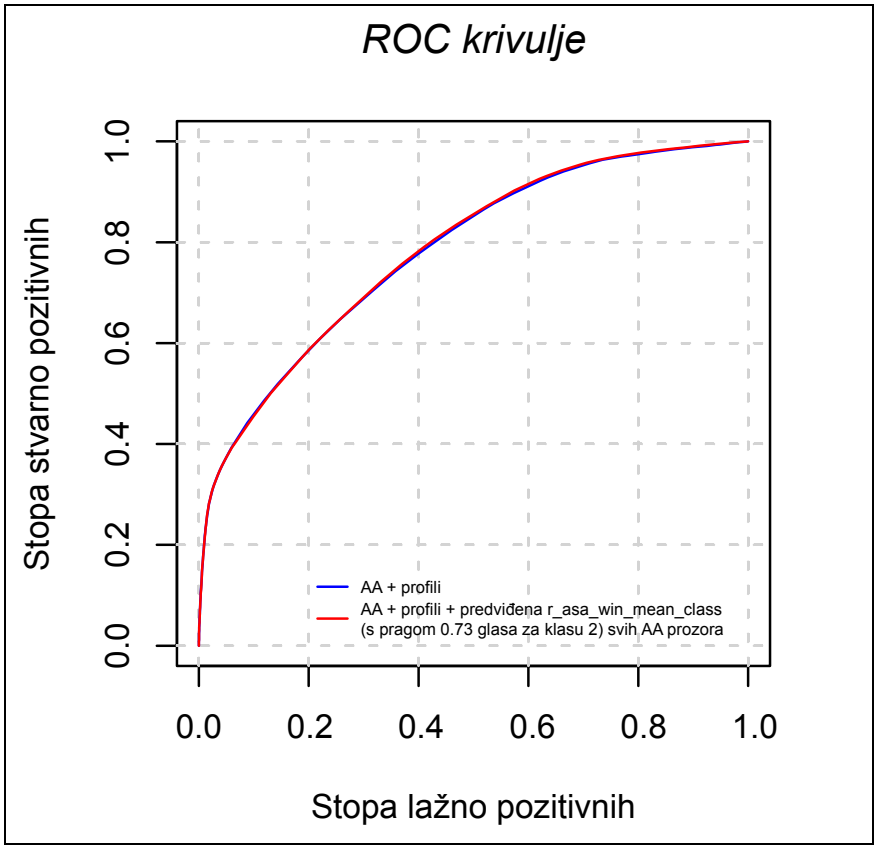
5.4.3.2 Rezultati predviđanja korištenjem informacije iz slijeda, profila slijeda i vektora predviđenih klasa srednjih vrijednosti RASA-a prozora

Analogno proširivanju atributa *class_r_asa_middle* u prethodnom poglavlju, za nekoliko sljedećih predviđanja proširuje se atribut *r_asa_win_mean_class*. Efekt korištenja vektora s 9 takvih predviđenih vrijednosti odgovara proširivanju atributa *numeric_mean_r_asa_win_mean* (definiran točkom 3. u prethodnom poglavlju) jer se na isti način u predviđanje uključuje informacija o RASA-i šireg područja. Ovaj put se umjesto vektora izračunatih srednjih vrijednosti prozora koristi vektor predviđenih srednjih vrijednosti prozora.

Kako su se u poglavlju 5.4.2 najbolji rezultati dobili za skup *2b* i to s pragom **0,73** glasa danih klasi 2 za proglašavanje instance klasom 2, skup podataka iz te grupe s proširenim ulaznim atributima koristio se i za sljedeća predviđanja. Koristile su se samo vrijednosti koje je bilo moguće predvidjeti, pa je skup znatno kraći od originalnoga. Rezultati predviđanja dani su slikama 5.26 i 5.27 te tablicom 5.17.



Slika 5.26 PR krivulje predviđanja mjesta interakcije metodom krosvalidacije koristeći a) samo slijed aminokiselinskih ostataka i njihove profile te kombinaciju sa b) predviđenim *r_asa_win_mean_class*-ama svih ostataka prozora (s pragom proglašavanja klasom 2 postavljenim na 0,73)



Slika 5.27 ROC krivulje predviđanja mjesta interakcije metodom krosvalidacije koristeći a) samo slijed aminokiselinskih ostataka i njihove profile te kombinaciju sa b) predviđenim *r_asa_win_mean_class*-ama svih ostataka prozora (s pragom proglašavanja klasom 2 postavljenim na 0,73)

	AUC	F-mjera	Točnost	Preciznost	Odziv
a)	0, 782	0, 3942	0, 7951	0, 8585	0, 2559
b)	0, 7839	0, 3938	0, 7951	0, 8599	0, 2554

Tablica 5.17 Mjere uspješnosti predviđanja mjesta interakcije metodom krosvalidacije koristeći a) samo slijed aminokiselinskih ostataka i njihove profile te kombinaciju sa b) predviđenim *r_asa_win_mean_class*-ama svih ostataka prozora (s pragom proglašavanja klasom 2 postavljenim na 0,73)

Iako su dobiveni rezultati nešto bolji nego kod predviđanja urađenih samo s jednom *r_asa_win_mean_class* po promatranom prozoru (poglavlje 5.4.2), razlika je neznatna. Vrijednost AUC je za 0,19% bolja nego samo s korištenjem slijeda i profila slijeda, dok je samo s jednim atributom bila 0,07% bolji. Iako ova razlika nije ni značajna, niti najveća dobivena dosad, treba primijetiti da je

koncept promatranja šire RASA okoline (uz skaliranje utjecaja kako je prikazano slikom 5.21) poboljšao predviđanje mjesta kontakta kad se radilo o vektoru **predviđenih** vrijednosti koje opisuju RASA-u prozora, dok su analogna predviđanja korištenjem **izračunatih** vrijednosti koje opisuju RASA-u prozora dala lošije rezultate od početnih.

6 Diskusija i zaključak

Cilj rada bio je pronaći odgovarajuće diskretne RASA atribute koji bi zadovoljili sljedeće uvjete:

- dovoljno velika točnost predviđanja tih atributa korištenjem slijeda aminokiselinskih ostataka i njihovih profila
- bolji rezultati predviđanja mjesta kontakta korištenjem predviđenih RASA atributa

Predviđanja su se pritom radila metodom slučajnih šuma, RASA atributi su se svrstavali u dvije do pet klasa, a za predviđanje mjesta interakcije koristio se pomični prozor duljine devet aminokiselinskih ostataka.

Obzirom da su se RASA atributi predviđali klasifikacijom, veću točnost su imala klasificiranja u manji broj klasa, i to tako definiranih da je broj pojavljivanja instanci unutar jedne klase daleko veći od broja pojavljivanja u svim ostalim klasama zajedno. Međutim, za predviđanje mjesta kontakta pokazalo se da više informacije u sebi sadrže RASA atributi klasificirani u što veći broj klasa tako definiranih da su zastupljenosti pojavljivanja unutar pojedine klase međusobno približno jednake. Ispitivanjem se ustanovilo da skupovi podataka s RASA atributima klasificiranim u tri klase definirane prema Maneshovim pragovima [17] predstavljaju najbolji kompromis između ta dva potpuno oprečna načina definiranja klasa RASA atributa.

Kombinacijama predviđanja u poglavljima 5.4.1–5.4.3 ispituje se utjecaj nekoliko varijacija RASA atributa na predviđanje mjesta kontakta. Pokušava se otkriti koliko je veliko okruženje nekog aminokiselinskog ostatka (u kontekstu promatranja RASA vrijednosti) koje utječe na činjenicu da je taj ostatak mjesto kontakta ili da to nije. Eksperimentira se s predviđanjima kod kojih je poznata predviđena klasa tog ostatka pa sve do vrijednosti koje skalirano opisuju RASA 16 aminokiselinskih ostataka oko promatranoga.

Nekoliko je rezultata pokazalo poboljšanje u odnosu na predviđanje mjesta kontakta koristeći samo slijed aminokiselinskih ostataka i njihove profile. Među njima je predviđanje korištenjem predviđene *r_asa_win_mean_class* za skup *2b*, uz postavljen prag proglašavanja instance klasom 2 na 73% glasova u prvom krugu predviđanja te predviđanje korištenjem 9-dimenzionalnog vektora s tim istim atributima proširenima na cijeli prozor. Najbolji rezultati dobiveni su uvođenjem izračunatih *numeric_mean_r_asa_win_mean* iz predviđenih klasa svih ostataka prozora pretvorenih u vrijednosti ovisne o tipu aminokiselinskog ostatka za skup iz grupe *3b* (slike 5.24 i 5.25). Prije proširivanja vektora ulaznih atributa boljima su se pokazali oni RASA atributi koji su opisivali područje cijelog prozora. Međutim, proširivanje na područje šire od promatranog pomičnog prozora dalo je poboljšane rezultate za predviđene atribute koji opisuju RASA-e prozora, ali ne i za izračunate atribute.

Nijedan od navedenih rezultata ipak nije pokazao dovoljno poboljšanje da isplati cijenu uvođenja dodatnog kruga predviđanja. Ukoliko bi se ipak nastavio ispitivati utjecaj RASA atributa na predviđanje mjesta kontakta, svakako bi trebalo proučiti mogu li se podešavanjem težina dobiti bolji rezultati na skupu 3b, koristeći predviđenu klasu RASA prozora te vektor od 9 predviđenih RASA klasa prozora. Radi ograničenja korištenih aplikacija spomenuta predviđanja nije bilo moguće uraditi, ali proučavanjem kretanja svih dobivenih rezultata vjerujem da bi se upravo za ova predviđanja dobili najbolji rezultati. Utjecaj RASA vrijednosti okoline na predviđanje mjesta kontakta mogao bi se ispitati i pomicanjem granica pomičnih prozora – prilikom predviđanja RASA atributa uzeti dulji pomični prozor nego prilikom predviđanja mjesta interakcije. Dosad se šira RASA okolina nije pokazala očekivano dobrim pokazateljem predviđanja mjesta kontakta, ali treba uzeti u obzir mogućnost da je razlog tomu odabrani pristup predviđanju, tj. izračunavanju tih vrijednosti te sam skup podataka koji se za te potrebe morao izmijeniti.

Ipak smatram da bi predviđanju mjesta kontakta najviše pridonijela konkretnija informacija o RASA vrijednosti, što nije moguće postići korištenjem kategorija (osobito kad je poznato da je klasifikacija RASA atributa tim lošija što je broj

kategorija veći). Realni broj dovoljno blizak stvarnoj RASA vrijednosti nekog aminokiselinskog ostatka će u svakom slučaju bolje upućivati na eventualnu pojavu mjesta kontakta od bilo kakve aproksimacije kategorijama. Zato bi trebalo pokušati uraditi predviđanje RASA vrijednosti regresijom – predvidjeti njihovu stvarnu vrijednost – i proučiti može li predviđanje regresijom biti dovoljno uspješno da više pridonese predviđanju mjesta kontakta od predviđanja klasifikacijom.

7 Literatura

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

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

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

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

[5] http://www.irb.hr/hr/research/projects/it/2004/2004-111/

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

[7] 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.

[8] 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.

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

[10] 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.

[11] 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.

[12] 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.

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

[14] 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.

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

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

[17] H. N. Manesh, M. Sadeghi, S. Arab, A. M. Movahedi, "Prediction of protein surface accessibility with information theory", Proteins, vol. 42, pp. 452–459, 2001.

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

[19] P. Norton, A. Samuel, D. Aitel, E. Foster-Johnson, L. Richardson, J. Diamond, A. Parker, M. Roberts, "Beginning Python", Wiley Publishing, Inc, 2005.

[20] A. L. Lehniger, M. M. Cox, D. L. Nelson, "Principles of Biochemistry", 4th edition, W. H. Freeman, 2004.

[21] A. Szilágy, V. Grimm, A. K. Arakaki, J. Skolnick, "Prediction of physical protein-protein interactions", Institute of Physics Publishing, Physical Biology 2, 2005.

[22] H. Kitano, "Computational systems biology", Nature Publishing Group, 2002.

[23] J. R. Bradford, C. J. Needham, A. J. Bulpitt, D. R. Westhead, "Insights into Protein-Protein Interfaces using a Bayesian Network Prediction Method", ScienceDirect, Elsevier, 2006.

[24] 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.

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

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

[27] R. L. Scheaffer, "Categorical Data Analysis", NCSSM Statistics Leadership Institute, 1999.

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

[29] T. Sing, O. Sander, N. Beerenwinkel, T. Lengauer, "ROCR: visualizing classifier performance in R", Bioinformatics 2005 21(20):3940-3941; doi:10.1093/bioinformatics/bti623

Sažetak

Nastojeći poboljšati dosada postignute rezultate predviđanja mjesta proteinskih interakcija na osnovu podataka iz slijeda aminokiselinskih ostataka i njihovih profila u ovome radu istražio se utjecaj klasificiranih atributa koji opisuju relativnu ASA-u promatranog ostatka i njegove uže okoline na pojavu mjesta interakcije.

Mjesto interakcije se definira kao prozor od 9 aminokiselinskih ostataka u kojemu je barem središnji ostatak u kontaktu. Za ostatak se smatra da je u kontaktu ako je barem jedan njegov atom udaljen od najbližeg atoma susjednog lanca manje od 6 Angstrema.

Predviđanje se radilo korištenjem alata PARF – paralelna inačica metode slučajnih šuma. Koristio se skup podataka od 1137 lanaca koji se prethodno obradio za potrebe predviđanja metodom krosvalidacije. Priprema podataka radila se vlastitim skriptama izrađenim u Pythonu i gotovim skriptama u Perlu, statističke analize pretežno u Matlabu, a obrada rezultata u programskom paketu R.

Atribut RASA kvantizirao se u dvije do pet razina korištenjem Lloyd-Max kvantizatora i pragova preuzetih iz Maneshovog rada [17]. Vektor ulaznih podataka kod predviđanja RASA atributa sastojao se od imena ostataka u prozoru i njihovih profila slijeda što ukupno čini 21×9 svojstava. Za predviđanje mjesta kontakta dodatno se koristilo 1–9 RASA atributa koji su opisivali promatrani ostatak, okolne ostatke, promatrani pomični prozor i područje šire od toga. Predviđanja i izračunavanja tih atributa radila su se ovisno i neovisno o tipu aminokiselinskog ostatka. Ne bi li se poboljšali rezultati klasificiranja RASA atributa, podešavale su se težine pojedinih klasa.

Iako su u nekoliko slučajeva dobiveni bolji rezultati predviđanja od dosad postignutih, ta su poboljšanja statistički zanemariva.