SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 1166

Predviđanje površine dostupne otapalu iz slijeda aminokiselinskih ostataka

Tomislav Puđa

Zagreb, prosinac 2008.

Sadržaj

1. Uvod

Poznavanje strukture proteina jedan je od ključnih čimbenika u razumijevanju mnogih važnih procesa u organizmu te zbog toga dobiva sve više i više pažnje od strane akademske zajednice. Proteinsku strukturu možemo eksperimentalno odrediti pomoću spektroskopskih i kristalografskih tehnika, no zbog visoke cijene postupka, potrebnog vremena za analizu te drugih čimbenika, eksperimentalne tehnike su obično ograničene na manje proteine. Alternativni pristup je da se skupe analize zamijene onima na računalima. Postojanjem javnih baza s tisućama riješenih 3-D struktura te uz veliku procesorsku moć današnjih računala omogućeno je pronalaženje novih metoda i algoritama za računalno određivanje proteinskih struktura.

Smatra se da slijed aminokiselina koje grade protein sadrži dovoljno informacija da se odredi njegova trodimenzionalna struktura. Ipak, specifični mehanizmi koji se odvijaju pri smatanju proteina (engl. *protein folding*) još nisu opisani. Stoga je precizna predikcija strukture moguća samo za one proteine koji imaju veliku sličnost u slijedu s proteinima već poznate strukture. Jedan od mogućih pristupa predviđanja strukture proteina temelji se na predviđanju strukturalnih karakteristika poput sekundarne strukture te dostupnosti otapalu (engl. *solvent accessibility*). Dostupnost otapalu određuje stupanj interakcije ostatka aminokiseline (engl. *amino acid residue*) s molekulama otapala te je važan pokazatelj stanja presavijanja proteina. Pošto se mjesta interakcije (engl. *active sites*) proteina gotovo uvijek nalaze na njihovoj površini, predviđanje izloženosti ostatka važna je karika za razumijevanje odnosa između strukture i funkcije. Poznavanje dostupnosti otapalu može poslužiti i kao potpora pri rješavanju problema smatanja te prianjanja proteina (engl. *protein docking*), temeljnim problemima ovoga područja.

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

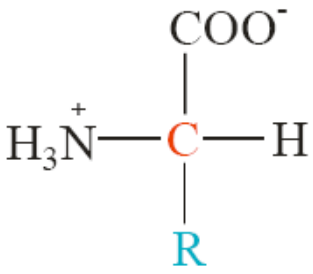
2. Teorijski uvod

2.1. Proteini

Proteine dobivamo međusobnim povezivanjem aminokiselina peptidnim vezama u duge lance od stotinu do približno tisuću aminokiselinskih ostataka. Kako je razumijevanje strukture proteina ključno za razumijevanje problema predviđanja dostupnosti otapalu, ovdje će biti dan kratak opis njihove strukture i formiranja. Ukratko će se opisati neke od strukturalnih karakteristika proteina s posebnim naglaskom na površinu dostupnu otapalu (engl. *accessible surface area*).

2.1.1. Aminokiseline. Ostatak aminokiselina

Aminokiseline su osnovne građevne jedinice proteina te kao takve uvjetuju oblik proteina te njegova svojstva. Jedan te isti skup aminokiselina zajednički je svim živim organizmima. Kao što ime kaže, aminokiseline posjeduju dvije karakteristične funkcionalne skupine: amino skupinu te karboksilnu skupinu. Kod aminokiselina, koje nas prvenstveno zanimaju kao sastavni dio proteina, amino skupina je uvijek smještena u α-položaju prema karboksilnoj grupi. Općenita formula aminokiseline je NH_2-*CHR*-*COOH*.



Slika 2.1 Općenita struktura aminokiselina. Zajednička je svim aminokiselinama osim jedne – prolin je prstenasta aminokiselina. R skupina je različita za svaku aminokiselinu. Ionizacijsko stanje je prikazano pri pH 7.0.

Aminokiseline se međusobno razlikuju po *R* skupini ili tzv. bočnom ogranku (engl. *side chain*) koji može biti od atoma do kompleksne molekule. Bočni ogranci aminokiselina koje tvore protein određuju njegova svojstva i funkciju. Pregled dvadeset aminokiselina koje redovito dolaze u proteinima dan je u tablici 1.1. Osim imena aminokiselina, u tablici se nalazi zapis aminokiselina u *FASTA* formatu gdje svakoj od aminokiselina odgovara jedno slovo. Koriste se sva slova engleskog alfabeta osim slova B, J, O, U, X i Z.

Tablica 2.1 Najvažnije aminokiseline

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	leucin
11	M	Met	metionin
12	N	Asn	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

Kada se dvije ili više aminokiseline spoje u lanac pri čemu tvore peptid, aminokiselina gubi vodu, a preostali dio aminokiseline naziva se ostatak aminokiseline (engl. *amino acid residue*). Ostatci aminokiselina mogu imati više različitih struktura. Prva mogućnost je da aminokiselini nedostaje vodik iz amino skupine. Moguće su i strukture u kojoj karboksilna skupina gubi -OH grupu te kombinacija nabrojanih; gubitak vodika iz amino skupine te -OH grupe iz karboksilne skupine.

Jedno od bitnih svojstava aminokiselina je njihova topljivost. Zbog svoje polarnosti, voda dobro otapa nabijene i polarne tvari. Tvari koje se dobro otapaju u vodi zovemo hidrofilnima, a one koje se u vodi loše otapaju

zovemo hidrofobnima. Topljivost aminokiseline u otapalu (vodi) određena je polarnošću bočnog ogranka. Važnost svojstava bočnog ogranka dolazi zbog utjecaja koji ima na interakciju ostataka aminokiselina s drugim strukturama, bilo to unutar istog proteina ili između proteina. Raspodjela hidrofobnih i hidrofilnih aminokiselina ima veliku važnost pri promatranju proteinskih struktura i interakcija.

2.1.2. Peptidi i proteini

Peptidi su polimeri aminokiselina. Kemijskom strukturom su amidi te hidrolizom daju aminokiseline. Aminokiseline se u peptidu vežu jedna uz drugu peptidnim (ili amidnim) vezama koje vežu amino skupinu jedne aminokiseline uz karboksilnu skupinu susjedne aminokiseline. Dvije aminokiseline izgrađuju dipeptid, tri čine tripeptid, osam oktapeptid itd. Ako nije međusobno povezano više od deset aminokiselina, govorimo o oligopeptidima, a ako ih je više o polipeptidima. Povećanjem lanaca polipeptida dolazimo do proteina koji su izgrađeni iz više od stotinu aminokiselina, koji se nazivaju i makropeptidima.

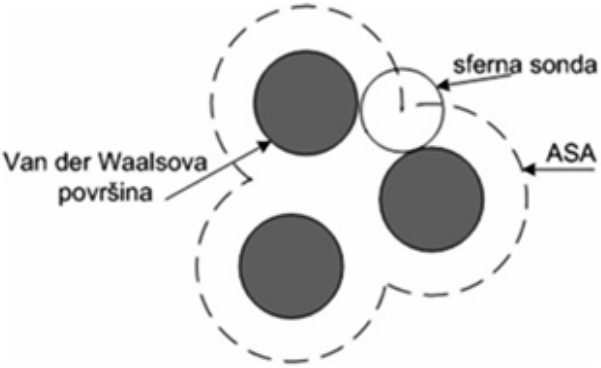
Kako bi se u potpunosti upoznao neki peptid, nije dovoljno poznavati samo vrstu i broj aminokiselina koje ga izgrađuju te koje se mogu osloboditi hidrolizom, već i redoslijed njihova povezivanja, tj. njihov slijed. To je razumljivo jer su npr. glicin-alanin i alanin-glicin potpuno različite supstancije.

2.1.3. Topologija površine proteina

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

2.1.4. Otapalu dostupno područje površine

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



Slika 2.2 Otapalu dostupno područje površine atoma

Otapalu dostupno područje površine sada možemo definirati pomoću sferne sonde koja predstavlja model molekule otapala. Sferna sonda kotrlja se po van der Waalsovoj površini te definira otapalu dostupno područje površine. Za radijus sonde se uzima obično iznos od 1.4 Å koji predstavlja radijus molekule vode.

2.1.5. Otapalu dostupno područje površine aminokiselina

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

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

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

U tablici 2.2 prikazane su standardne vrijednosti ASA koje su se koristile unutar ovoga rada. Vrijednosti su dobivene korištenjem programa PSAIA [1] te su izražene u $Å^2$.

Tablica 2.2 Standardne vrijednosti ASA za aminokiselinske ostatke

Vrsta ostatka	Ukupno	Glavni lanac	Bočni lanac	Nepolarno	Polarno
Ala	107,24	43,32	63,92	76,06	31,17
Arg	233,01	36,86	196,15	86,30	146,71
Asn	150,85	36,46	114,39	42,77	108,08
Asp	144,06	36,15	107,91	49,57	94,49
Cys	131,46	36,12	95,34	104,07	27,40
Gln	177,99	34,24	142,76	62,78	115,21
Glu	171,53	35,75	135,77	70,72	100,81
Gly	80,54	80,54	0,00	42,63	37,92
His	180,93	34,84	146,09	106,27	74,66
Ile	173,40	31,08	142,33	147,36	26,04
Leu	177,87	32,68	145,19	150,10	27,76
Lys	196,14	35,55	160,59	123,55	72,65
Met	186,80	34,06	152,74	159,15	27,65
Phe	200,93	33,67	167,26	174,16	26,77
Pro	133,78	32,90	100,88	112,82	20,96
Ser	115,30	40,86	74,44	52,93	62,37
Thr	136,59	34,14	102,45	80,39	56,20
Trp	240,12	32,51	207,61	186,22	53,90
Tyr	213,21	33,59	179,62	143,97	69,25
Val	149,34	31,33	118,01	123,43	25,91

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

Kao i kod apsolutnih vrijednosti, postoji pet relativnih vrijednosti (ukupna, glavnog lanca, bočnog lanca, polarna i nepolarna) koje se računaju omjerom

apsolutne i standardne vrijednosti pomnožene sa 100. Relativne vrijednosti ASA aminokiselina opisuju kolikim dijelom svoje površine, izraženim u postotcima, je aminokiselina dostupna otapalu.

2.2. Proteinske baze podataka

Dugi niz godina, u središtu biologije i biokemije nalazili su se geni te struktura i funkcija nukleinskih kiselina. Razvoj tehnologije donio je metode za masovno sekvencioniranje genoma, kao i metode za određivanje trodimenzionalne strukture proteina. Ubrzo se je stvorila potreba za pohranjivanjem velikog broja informacija o sekvenci i strukturi podataka. Ti podaci danas su pohranjeni u bazama podataka od kojih su najvažnije PDB [5] (*Protein Data Bank*) i UniProt.

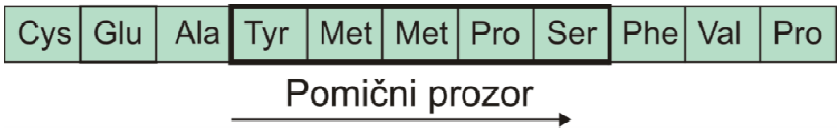
PDB je baza proteinskih struktura u čijim zapisima se nalaze podaci o prostornim koordinatama svih „teških“ atoma u proteinu. Pod teškim atomima se podrazumijevaju svi atomi osim vodika. Trenutno se u bazi nalazi više od 50.000 struktura dobivenih kristalografijom X-zrakama ili NMR-om (magnetskom rezonancijom).

2.3. Rad s pomičnim prozorima

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

Pri predviđanju se stoga često koriste pomični prozori fiksne duljine. Obično je riječ o prozorima s neparnim brojem aminokiselinskih ostataka, duljine između 3 do 21. Povećanjem duljine prozora povećava se broj aminokiselinskih ostataka na početku i kraju sekvence za koje se ne vrši

predviđanje. Svaki od elemenata prozora sadrži više različitih svojstava na temelju kojih se vrši predviđanje.



Slika 2.3 Kod predviđanja površine dostupne otapalu često se koristi pomični prozor fiksne duljine koji se pomiče od početka do kraja sekvence. Predviđanje se vrši za središnji ostatak, dok svaki od elementa prozora sadrži više različitih informacija.

2.4. Predviđanje dostupnosti otapalu klasifikacijom

Predviđanje površine dostupne otapalu može se obaviti na dva bitno različita načina. Kod predviđanje regresijom pokušava se odrediti točna ASA vrijednost za aminokiselinske ostatke, dok se klasifikacijom ostatci svrstavaju u predodređeni broj kategorija (klasa). Često se umjesto površine dostupne otapalu predviđa njezina relativna vrijednost. U ovome radu predviđanje će se vršiti klasifikacijom, pri čemu će se broj kategorija kretati između dva i pet. Najveću pažnju pri analizi rezultata imati će klasifikacija u dvije kategorije, prvenstveno zbog činjenice da je točnost pri klasifikaciju u veći broj kategorija još uvijek znatno manja.

Jedan od problema pri klasificiranju aminokiselinskih ostataka je izbor granica kategorija. Većina autora koristila je proizvoljne pragove, tj. granice kategorija u koje bi se nakon toga ostatci raspoređivali. Pragovi su se izražavali kao RSA vrijednosti te se tako izbjegao problem zbog velike razlike u ASA vrijednostima pojedinih aminokiselinskih ostataka. Proizvoljan izbor pragova ne uzima u obzir raspodjelu ASA vrijednosti te se time gubi dio korisnih informacija. Sam proizvoljni izbora pragova otežava i usporedba točnosti različitih metoda.

Kao alternativa proizvoljnom izboru pragova u poglavlju 4.1 biti će objašnjeno kako se primjenom optimalnog kvantizatora mogu odrediti granice kategorija, dok će usporedba rezultata metoda s proizvoljnim izborom praga te predložene metode u ovome radu biti detaljno obrađeno u poglavlju 5.

3. Podaci

3.1. Korišteni skupovi proteina

Usporedbu točnosti između različitih metoda često je teško provesti pošto autori obično ne koriste iste skupove. Ipak, postoji nekoliko skupova koji se češće upotrebljavaju te su stoga odabrani i u ovome radu.

Prvi izabrani skup potiče iz rada Ofrana i Rosta [2]. Skup se sastoji od 333 proteina s 1500 proteinskih lanca. Skup se često koristi i na Zavodu za elektroničke sustave i obradu informacija u sklopu diplomskih radova i doktorskih disertacija na temu predviđanja mjesta proteinskih interakcija. Nakon izbacivanja homolognih lanaca, broj lanaca u skupu iznosi 833 te se je kao takav koristio za prikaz različitih ocjena kvalitete predviđanja.

Slijeći odabrani skup, RS126, jedan je od najstarijih korištenih skupova za ocjenu kvalitete predviđanja sekundarnih struktura. Predložen je u radu Rosta i Sandera [3], sadrži 126 proteinskih lanaca pri čemu je sličnost između lanaca manja od 25%, što je naknadno opovrgnuto. Usprkos tome skup se još uvijek često koristi. Posljednji od skupova, MN215, predložen je u radu Manesha [4] te sadrži 215 nehomolognih proteina sa sličnosti između lanaca manjom od 25%.

3.2. Svojstva na osnovu kojih će se vršiti predviđanje

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

- sekvenca duljine 3 do 17 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. Pomični prozor se stoga može smatrati vektorom dimenzije $21 \times n$, gdje je n duljina prozora.

3.2.1. Profili 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 sekvenci na kojemu se nalazi aminokiselina čiji se profil određuje. Često se koriste za predviđanja različitih strukturalnih karakteristika.

Kako bi se odredili profili slijeda, prvo je potrebno izvući imena aminokiselina koje tvore određeni lanac. Imena aminokiselina se zatim zapisuju u *FASTA* formatu, formirajući pri tome niz slova (svako slovo se odnosi na odgovarajuću aminokiselinu), odnosno sekvencu. Takva se sekvenca propušta kroz PSI-BLAST algoritam koji daje rezultat kao *PSSM* (engl. *Position-specific scoring matrix*) matricu odnosno profil. Profili su građeni u odnosu na SWISS-PROT proteinsku bazu podataka. Način određivanja profila slijeda biti će detaljno objašnjen u poglavlju 4.1.

3.3. Struktura podataka

Podaci o strukturi proteina nalaze se u PDB [5] formatu. Iako se unutar PDB datoteka nalaze samo osnovni podaci, datoteka sadrži veliki broj informacija kao što su način na koji je određena struktura proteina, od kojih se lanaca ostataka, molekula i atoma sastoji promatrani proteinski kompleks te prostorni raspored svakog pojedinog atoma.

PDB datoteke služe kao ulazni podaci iz kojih se određuju ASA vrijednosti ostataka primjenom PSAIA alata [1]. Za svaku od PDB datoteka alat generira po jednu izlaznu XML datoteku. Nakon što se obrade sve PDB datoteke skupa koji se koristi za predviđanje, pokreće se skripta napisana u Python programskom jeziku. Nakon što se skripta izvrši, u izlazne XML datoteke dodani su profili slijeda te oznake klasa dobivene primjenom Lloyd-Max kvantizatora. Skripta u isto vrijeme generira prozore različitih duljina te sve navedene podatke zapisuje i u odgovarajuće *ARFF* (engl. *Attribute-Relation File Format*) datoteke.

3.3.1. Priprema podataka za klasifikaciju

Predviđanje površine dostupne otapalu vršiti će se primjenom paralelne inačice algoritma slučajnih šuma, *PARF* [6]. Algoritam je razvijen na Institutu Ruđer Bošković na temelju originalnog algoritma slučajnih šuma autora L. Breimana. *PARF* koristi izmijenjeni *ARFF* format za podatke, koji je većinom kompatibilan s originalnim formatom koji koristi Weka projekt. Ukratko, datoteka počinje imenom relacije, slijedi definicija atributa, te na kraju dolazi blok s podacima. Definicija jednog atributa daje mu ime i moguće vrijednosti: numeric (brojčane), string (tekstualne) ili skup kategorija u vitičastim zagradama.

Kako će se klasifikacija vršiti ovisno o vrsti aminokiselinskog ostatka te zbog korištenja prozora različitih duljina, broj ulaznih datoteka iznosi 160. Svaka datoteka sadržavati podatke o ostacima unutar prozora, njihovim profilima te klasama kojima pripadaju ostaci. U nastavku je prikazan jedan zapis unutar *ARFF* datoteke :

SER,CYS,ILE,ILE,SER,MET,VAL,VAL,GLY,GLN,LEU,1A2K,A,84,0,0,0,8,0,4,3,37,0,0,0,6,0,0,6,9,24,3,0,0,4,3,0,11,9,0,0,35,0,0,7,0,3,0,0,21,8,0,0,0,0,0,0,0,0,0,0,0,36,10,0,17,2,0,0,5,0,0,30,0,0,0,0,10,4,0,0,0,14,49,0,8,0,0,0,0,0,0,13,0,0,0,0,3,0,0,0,0,22,10,0,0,10,0,7,0,0,0,47,2,0,5,0,0,13,0,0,0,0,8,0,31,12,0,14,4,0,0,10,10,0,0,0,0,0,0,0,0,9,0,0,0,0,0,0,4,0,0,77,5,0,17,0,4,0,0,0,2,0,8,0,8,0,0,16,32,0,0,7,0,0,0,0,0,0,0,100,0,0,0,0,0,0,0,0,0,0,0,0,10,0,2,6,0,14,18,0,0,2,12,0,0,0,0,7,12,0,0,17,0,0,0,0,0,0,0,0,0,0,48,0,10,24,0,0,0,0,0,17,2,3,3,4

U primjeru se radi o prozoru duljine 11 aminokiselinskih ostataka, kojemu je središnji ostatak za kojega se vrši predviđanje metionin. Sekvenca pripada lancu A proteina 1A2K. Serijski broj središnjeg ostatka unutar lanca je 84. Nakon tih informacija slijede profili slijeda za svih 11 ostataka unutar prozora. Četiri zadnje znamenke predstavljaju klase kojima pripada središnji ostatka. Prvi broj predstavlja klasu kojoj pripada ostatak za klasifikaciju u dvije klase, drugi za klasifikaciju u tri klase itd. Ti podaci služe za treniranje slučajnih šuma, a dobiveni su primjenom Lloyd-Max kvantizatora.

4. Metode

4.1. Određivanje profila slijeda. PSI-BLAST

PSI-BLAST [7][8] je poboljšani algoritam tehnike BLAST pomoću kojeg je moguće pronaći evolucijsku očuvanost aminokiselinskog ostatka. Koristi se metodom sekvencijalnog poravnanja (engl. *sequence alignment*) za prepoznavanje sličnosti dvaju proteina koji nisu blisko povezani, tj. nemaju bliskog zajedničkog homologa. Proteini mogu biti sekvencijalno ili strukturalno slični, pri čemu sličnosti nisu nužno povezane. Temeljna ideja algoritma je prepoznati zajedničku strukturu iz slabe sekvencijalne sličnosti. Toj se problematici pristupa metodom sekvencijalnog poravnavanja u sklopu tehnike BLAST.

4.1.1. Sekvencijalno poravnanje

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

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

grafičkim prikazima na slici 4.1 boje simboliziraju različite skupine aminokiselinskih ostataka. Tako su crvenom bojom prikazane male te hidrofobne (uključujući aromatske) aminokiseline.

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

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

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

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

```
FTFTALILLAVAV      FTFTALILL-AVAV
F--TAL-LLA-AV      --FTAL-LLAAV--
```

a) globalno poravnanje b) lokalno poravnanje

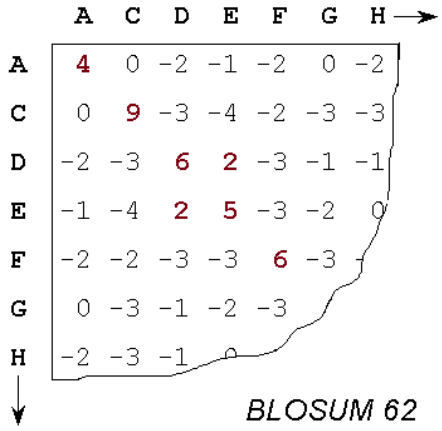
Slika 4.2 Globalno i lokalno poravnanje

4.1.2. BLAST. BLOSUM supstitucijske matrice

BLAST [7][8] (engl. *Basic Local Alignment Search Tool*), je algoritam za usporedbu sekvenci aminokiselinskih ostataka proteina ili nukleotida DNK lanca. BLAST omogućuje da se za neku sekvencu koja se ispituje, pretraži proteinska baza podataka. Tako algoritam, pretražujući bazu, nalazi sekvence koje odgovaraju sekvenci koja se ispituje, pri tome zadovoljavajući određeni prag sličnosti koji u pravilu zada korisnik. Kada se radi sravnjenje sljedova aminokiselinskih ostataka, algoritam BLAST koristi supstitucijsku matricu za procjenu sličnosti sljedova. Postoji nekoliko takvih matrica, među kojima se najčešće koriste BLOSUM i PAM matrice.

BLOSUM (*Blocks Substitution Matrix*) bazira se na lokalnom poravnavanju, a prvi put je predstavljena u radu Henikoff and Henikoff [10]. Nastala je empirijski na temelju poznatih i vrlo konzervativnih regija proteinskih familija u proteinskoj bazi podataka i računanja relativnih frekvencija pojavljivanja pojedinih aminokiselina. Za razliku od PAM matrica koje su dobivene uspoređivanjem poznatih i sličnih sekvenci tj. onih koje slabo divergiraju, BLOSUM matrice su nastale iz evolucijski divergentnih sekvenci.

Postoji više BLOSUM matrica ovisno o bazi podataka iz koje su nastale, a označuju se brojem koji upućuje na sličnost sljedova iz kojih su nastale. Primjerice, BLOSUM80 znači da se radi o sekvencama sličnosti iznad 80%. Takva će se matrica koristiti u slučaju manje evolucijski divergentnih sekvenci, dok će se BLOSUM45 koristiti u slučaju više divergentnih sekvenci. Na slici 4.3 prikazana je matrica BLOSUM62, kakva se je koristila u ovome radu.



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

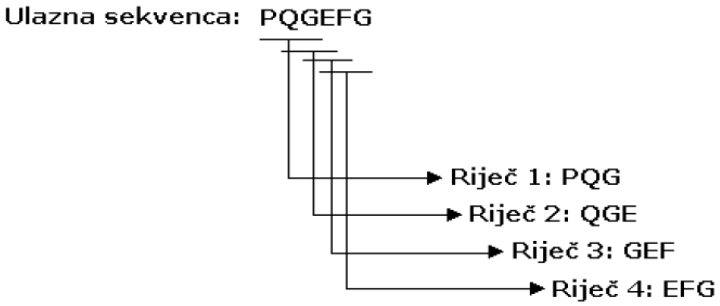
Vrijednosti matrice mogu se izračunati slijedećom izrazom:

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

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

4.1.3. Opis algoritma

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

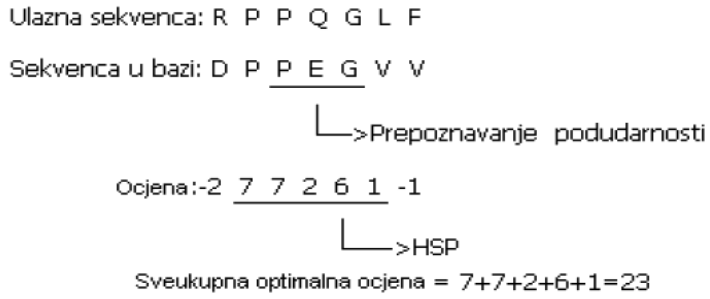


Slika 4.4 Rastavljanje sekvence u riječi od po tri slova

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

Nakon što se pronađu ciljne riječi za svaku riječ od tri slova ulazne sekvence, slijedi traženje tih istih ciljnih riječi u sekvencama baze. Kada se pronađe ciljna riječ u sekvenci baze, ona može upućivati da s odgovarajućom riječi ulazne sekvence čini jezgru. Da bi se to zaključilo vrši se proširivanje u oba smjera. Odnosno, gledaju se susjedni ostaci, te računa ocjena. Proširivanje

poravnanja traje sve dok ocjena sličnosti (koja se čita iz BLOSUM matrice) ne počne padati (slika 4.5).



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

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

4.1.4. Procjena značajnosti ocjene lokalnog poravnanja

Dobro ocjenjeno lokalno poravnanje ne mora nužno značiti da su odgovarajuće sekvence slične te da imaju zajedničkog homologa. Lokalno poravnanje može biti posljedica slučajnosti. Stoga se radi model slučajnih sekvenci u cilju uklanjanja takvih pojava. Jednostavan model proteina sastoji se od slučajno odabranih aminokiselinskih ostataka na temelju njihovih specifičnih frekvencija pojavljivanja (engl. background probability). Ocjena lokalnog poravnanja poprima negativnu vrijednost u slučaju da je poravnanje slučajno. Inače bi dugačka poravnanja imala visoku vrijednost ocjene poravnanja neovisno da li su evolucijski povezani.

U dovoljno dugačkim sekvencama duljine m i n, značajnost ocjene lokalnog poravnanja karakteriziraju dva parametra K i λ. Očekivani broj dobro

ocjenjenih lokalnih poravnanja, *E*-value, koji su posljedica slučajnosti, a vrijednost ocjene barem S' jest:

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

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

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

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

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

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

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

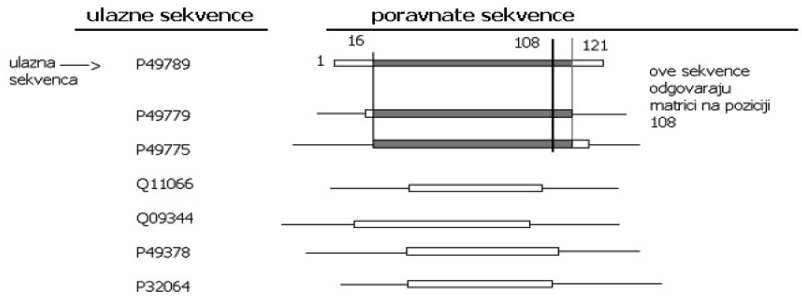
$$s_{ij} = \left[\ln(q_{ij} / P_i P_j \right] \lambda_u \tag{4.5}$$

4.1.5. PSI-BLAST

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

proces ne izvrši zadani broj iteracija ili ne konvergira. Iterativni postupak poboljšava rezultat i povećava osjetljivost. PSI-BLAST omogućava pronalazak udaljenih sekvenci odnosno onih manje sličnih. Profil izgrađen u prvoj iteraciji se temelji na sličnim sekvencama ulaznoj sekvenci te služi za sljedeću iteraciju u kojoj se pronalaze udaljene, a slične sekvence.

Algoritam se sastoji od sljedećih elemenata. Korištenjem BLAST algoritma, u prvoj se iteraciji izgradi profil koji se zatim uspoređuje s proteinskom bazom podataka odnosno njihovih sekvenci. Početna točka u kreiranju profila jest grupa sekvenci koje su poravnate, a ujedno su i izlazni podatak BLAST algoritma. Taj se rezultat reducira u cilju određivanja vrijednosti profila. Za svaki stupac poravnatih sekvenci, u obzir se uzimaju i susjedni aminokiselinski ostaci. Tako se poravnati redci sekvenci reduciraju, tj. uzimaju se samo oni redovi čiji su stupci postavljeni na način da svaki sadrži određeni ostatak ili prazninu, s time da su redovi iste duljine.



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

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

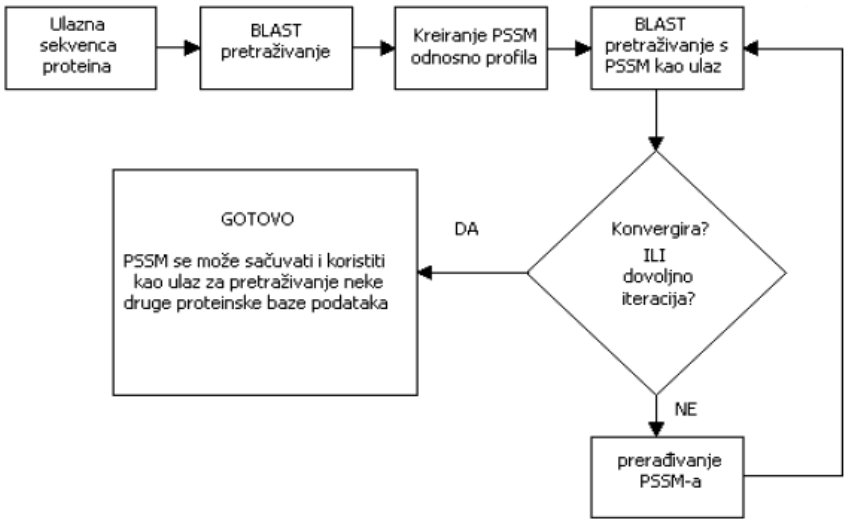
$$\text{Profile}(r,c) = \sum_{d=1}^{20} \sum_{i=1}^{N} weight(i)\delta(A_{ir}, d) \times Comp(residue_d, residue_c) \qquad (4.6)$$

gdje je Profile*(r,c)* vrijednost profila za redak *r* i stupac *c*; *r* može imati vrijednosti od 1 do *N* (duljina seta), *c* i *d* poprimaju vrijednosti od 1 do 20,

prezentirajući aminokiseline; *i* je pozicija sekvence u setu; *N* ukupan broj sekvenci; $\delta(A_{ir},d)$ ima vrijednost 1 ako ostatak na poziciji *r* u sekvenci *i* je aminokiselina *d*, inače je jednak nuli. $Comp(residue_d, residue_c)$ je vrijednost u supstitucijskoj tablici. *Weight(i)* se odnosi na težinu sekvence *i*. Težina sekvence se može izračunati aproksimativno iterativnom metodom na slijedeći način:

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

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



Slika 4.7 Dijagram PSI-BLAST algoritma

4.2. Lloyd – Max kvantizator

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 razina 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 zapravo se sastoji od dva dijela: kako za određeni skup kvantizacijskih razina odrediti kvantizacijska područja te kako za određeni skup kvantizacijskih područja odrediti kvantizacijske razine. Odgovor na prvo pitanje je jednostavan: da 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:

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

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 slijedećih 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 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 biti će manje osjetljiv na vrijednost koje značajno odstupaju od ostalih (engl. *outlier*).

4.3. Metoda slučajnih šuma

Slučajne šume (engl. *Random Forest*) su u ovome radu odabrane kao metoda klasifikacije zbog slijedećih 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* izrađena na IRB-u.

Slučajna šuma [9], u kasnijem tekstu *RF*, je općeniti naziv za skupinu metoda koje se koriste stablastim klasifikatorima $\{h(\boldsymbol{x}, \Theta_k), k=1,\ldots,\}$ gdje je $\{\Theta_k\}$ skup jednoliko distribuiranih, međusobno potpuno neovisnih vektora, a $\boldsymbol{x}$ ulazni vektorski uzorak. Prilikom treniranja, *RF* algoritam stvara veliki broj stabala, od kojih se svako trenira na određenom broju uzoraka originalnog trening seta odabranih *bootstraping* metodom. Za razliku od klasičnih stabala gdje se odabire najbolji atribut, za grananje *RF* koristi m slučajno odabranih varijabli ($m \ll$ M, obično $\log_2 M+1$) i uzima one koja omogućavaju najbolje grananje. Vrijednost m se unaprijed određuje i konstantna je za cijelu šumu. Za

klasifikaciju svako stablo unutar *RF* daje glas jednoj od klasa unutar skupa ***x***. Izlaz klasifikatora ovisi o broju glasova stabala svakoj pojedinoj klasi.

Trening skup za pojedino stablo stvara se tako da se iz početnog skupa za treniranje, veličine *N*, uzme *N* instanci, slučajnim odabirom s ponavljanjem. Iz tako stvorenog skupa za treniranje stabala, vrijednosti koje nisu odabrane koriste se za procjenu pogreške. Ove instance se nazivaju *oob* instance (engl. *out of bag*) i ima ih oko 38 % ukupnog broja instanci *N* početnog skupa i koriste se za dobivanje nepristrane procjene greške klasifikacije. Također se koriste i za procjenu važnosti pojedinih varijabli ulaznih instanci.

Kod slučajne šume nema potrebe za krosvalidacijom ili korištenjem posebnog seta za testiranje kako bi se dobila nepristrana procjena greške. Svako stablo se stvara tako da se koristi podskup iz početnih podataka za učenje koji se naziva *bootstrap* podskup. Svaki uzorak izostavljen pri stvaranju *k*-tog stabla, *oob* instance, treba pustiti niz *k*-to stablo da bi se dobila klasifikacija. Nakon završene obrade definiramo *j* kao klasu koja je dobivala najviše glasova u slučaju kada je *n* bila *oob* instanca. Omjer broja izlaza kada *j* nije bila jednaka pravoj klasi instance *n* s obzirom na sve instance naziva se procjena pogreške *oob*-a.

Za svako se stablo u šumi uzimaju oob instance, te zbroje glasovi koji su ispravno doneseni s obzirom na klasu. U sljedećem se koraku slučajno permutiraju vrijednosti varijable m u oob instancama, te ih se ponovo propusti kroz stablo. Nakon toga se oduzima broj glasova za ispravnu klasu oob instanci s permutiranom m varijablom od broja glasova za ispravnu klasu neupotrijebljenih oob instanci. Srednja vrijednost dobivene razlike u svim stablima unutar šume naziva se važnost varijable *m*. Ukoliko su vrijednosti ove važnosti nezavisne od stabla do stabla, njezinim dijeljenjem sa standardnom pogreškom dobiva se z-skor.

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, čiji je razred poznat.

Koraci izgradnje stabla odluke su sljedeći:

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

4.4. Mjerenje uspješnosti predviđanja

4.4.1. Točnost. Matrica greške

Jedna od mjera koja će se koristiti za ocjenu uspješnosti predviđanja je točnost. Podaci o točnosti izvlačiti će se iz matrice greške koju generira PARF. Za definiranje točnosti koristiti 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

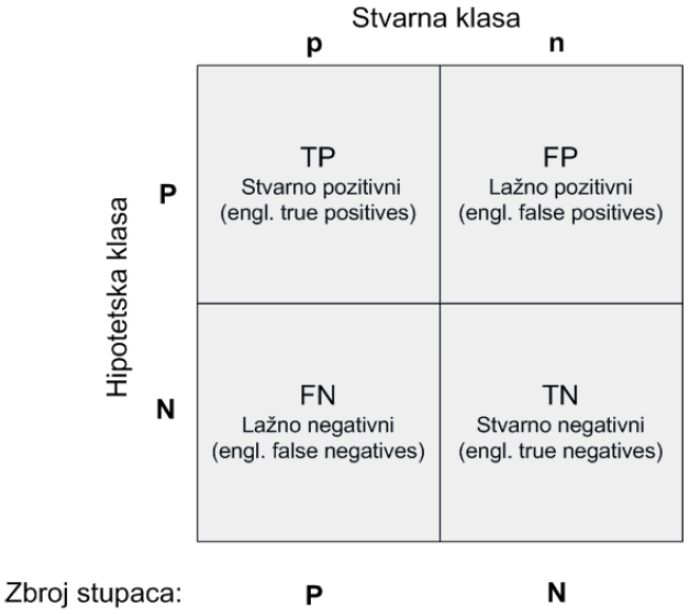
različite kategorije u koje klasifikator raspodjeljuje objekte nekog skupa. Tako objekti označeni kao pozitivni mogu biti klasificirani kao stvarno pozitivni TP (engl. *true positives*) i lažno pozitivni FP (engl. *false positives*). Uzorci podataka koje je klasifikator označio kao negativne također su raspoređeni u dvije kategorije, stvarno negativni TN (engl. *true negatives*) i lažno negativni FN (engl. *false negatives*).

Točnost klasifikacije sada možemo definirati kao omjer točno klasificiranih uzoraka te ukupnog broja slijedećim izrazom:

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

Pogledom na matricu greške prikazanu na slici 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 definiramo točnost za N klasa slijedećim izrazom:

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



Slika 4.8 Matrica greške za klasifikaciju u dvije kategorije

4.4.2. Pearsonov koeficijent korelacije

Slijedeća mjera koja će se koristiti za mjerenje uspješnosti predviđanja je koeficijent korelacije. Koeficijent korelacije izražava mjeru povezanosti između dvije varijable neovisno o konkretnim jedinicama mjere u kojima su izražene vrijednosti varijabli. Varijable između kojih će se mjeriti povezanosti su vektori hipotetskih te predviđenih klasa.

Korelacija će se izražavati preko Pearsonovog koeficijenta korelacije, čije se vrijednosti kreču od +1 (savršeno pozitivna korelacija) do -1 (savršeno negativna korelacija). Predznak koeficijenta nas upućuje na smjer korelacije – da li je pozitivna ili negativna, ali nas ne upućuje na snagu korelacije. Pearsonov koeficijent korelacije bazira se na usporedbi stvarnog utjecaja promatranih varijabli jedne na drugu u odnosu na maksimalni mogući utjecaj dviju varijabli. Označava se malim latiničkim slovom *r*. Za izračun koeficijenta korelacije potrebna su tri različite sume kvadrata (*SS*): suma kvadrata varijable *X*, suma kvadrata varijable *Y* i suma umnožaka varijabli *X* i *Y*.

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

$$SS_{XX} = \sum_{i=1}^{n} (X_i - \overline{X})^2 \tag{4.11}$$

dok je prosječna vrijednost varijable X dana relacijom:

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{4.12}$$

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

$$SS_{XY} = \sum (X_i - \overline{X})(Y_i - \overline{Y}) \tag{4.13}$$

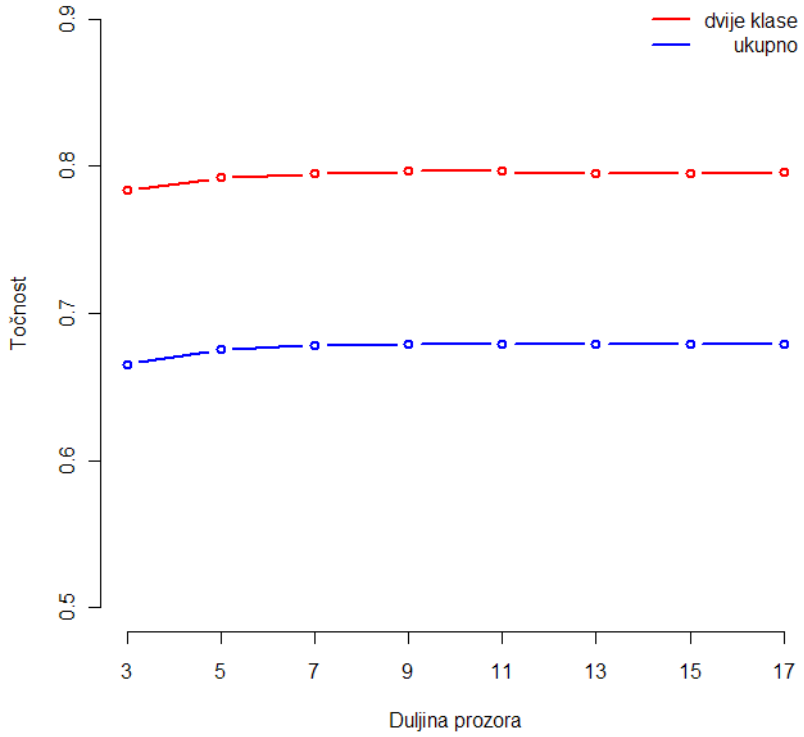
možemo definirati koeficijent korelacije:

$$r = \frac{SS_{XY}}{\sqrt{SS_{XX} \cdot SS_{YY}}} \tag{4.14}$$

5. Rezultati

5.1. Odabir duljine prozora.

Ovisnost točnosti o duljini prozora prikazana je na slici 5.1. Plava linija prikazuje kretanje srednje vrijednosti točnosti za klasifikaciju u dvije do pet kategorija (klasa). Iako je varijacija vrijednosti točnosti mala, lako se može uočiti povećanje točnosti za prozore s 5 ili više ostataka. Identična svojstva pokazuje klasifikacija u dvije kategorije (crvena linija).



Slika 5.1 Ovisnost točnosti o duljini prozora za skup Ofrana i Rosta

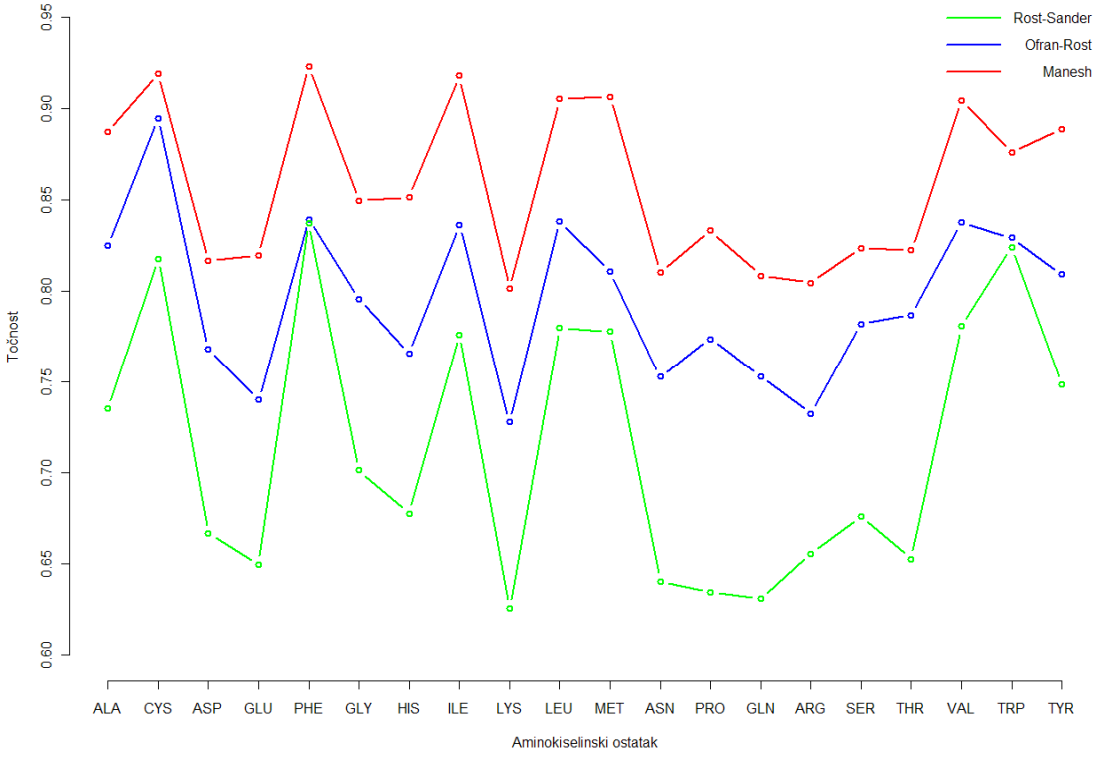
U tablici 5.1 dane su točnosti za klasifikaciju u dvije kategorije. Za dva od tri korištena skupa točnost klasifikacije je najveće za prozor duljine 9 ostataka. Zbog toga će se pri usporedbi točnosti koristiti upravo rezultati dobiveni za prozor duljine 9 aminokiselinskih ostataka, osim ako ne bude drugačije navedeno.

Tablica 5.1 Ovisnost točnosti o duljini prozora za različite skupove pri klasifikaciji u dvije kategorije. Istaknute su maksimalne vrijednosti točnosti za svaki od skupova.

	Duljina pomičnog prozora							
Korišteni skup	3	5	7	9	11	13	15	17
Ofran-Rost	78,36	79,20	79,49	**79,64**	79,63	79,50	79,52	79,55
Manesh	85,13	85,61	85,84	**85,89**	85,83	85,81	85,72	85,75
Rost-Sander	70,77	71,20	**71,43**	71,16	71,09	71,01	71,07	70,73

5.2. Ovisnost točnosti o aminokiselinskom ostatku

U poglavlju 3.3.1. rečeno je da će se predviđanje vršiti ovisno o vrsti aminokiselinskog ostatka. Razlog tome su razlike u razdiobi ASA vrijednosti te različit utjecaja susjednih ostataka na ASA vrijednost ostatka za kojeg se predviđanje vrši. Na slici 5.2 prikazana je ovisnost točnosti o vrsti aminokiselinskog ostatka za klasifikaciju u dvije kategorije.



Slika 5.2 Ovisnost točnosti klasifikacije o vrsti aminokiselinskog ostatka za različite skupove. Podaci su dani za klasifikaciju u dvije kategorije.

Sa slike su može uočiti velika sličnost u raspodjeli vrijednosti točnosti, pri čemu najveću razliku čini srednja vrijednost. U sva tri skupa najlošiji rezultati dobiveni su za lizin, dok je točnost klasifikacije bila najbolja za cistein i fenilalanin, ovisno o korištenom skupu. Sličnu pojavu prijavili su autori Z. Yuan i B. Huang u svome radu [11], iako su radili regresiju, a ne klasifikaciju. Najbolje rezultate dobili su za cistein, dok su za arginin, koji je i primjenom ove metode među najlošije klasificiranim ostacima, dobili najlošije rezultate.

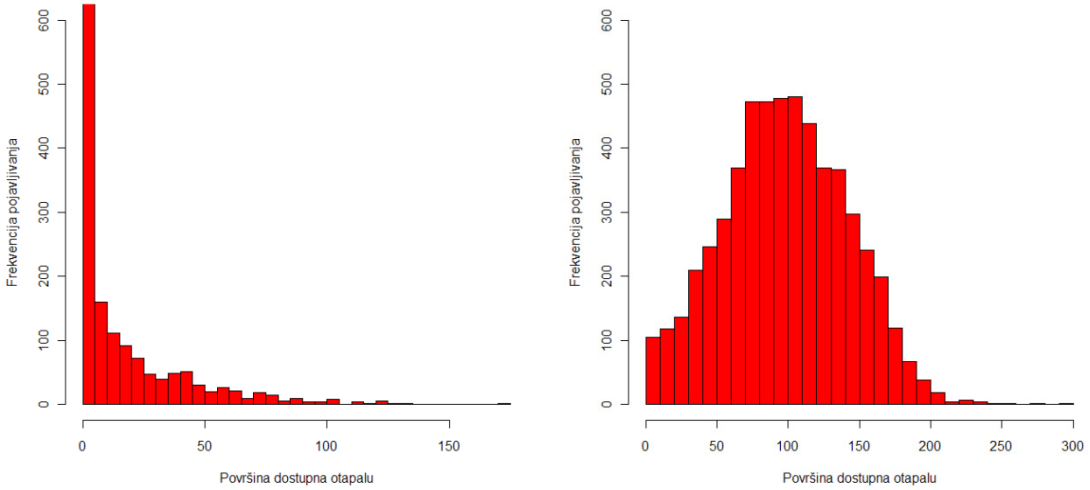
Tablica 5.2 Ovisnost točnosti predviđanja o vrsti aminokiselinskog ostatka. Istaknute su minimalne i maksimalne vrijednosti za svaki od skupova.

Vrsta aminokiselinskog ostatka	Ofran-Rost	Manesh	Rost-Sander
Ala	82,48	88,71	73,54
Cys	**89,46**	91,92	81,77
Asp	76,79	81,66	66,64
Glu	74,01	81,93	64,95
Phe	83,93	**92,31**	**83,74**
Gly	79,56	84,95	70,12
His	76,57	85,13	67,77
Ile	83,62	91,78	77,60
Lys	**72,79**	**80,12**	**62,55**
Leu	83,81	90,51	77,98
Met	81,07	90,51	77,77
Asn	75,27	81,00	64,04
Pro	77,37	83,31	63,44
Gln	75,31	80,85	63,04
Arg	73,25	80,45	65,54
Ser	78,20	82,36	67,58
Thr	78,68	82,26	65,26
Val	83,76	90,42	78,07
Trp	82,91	87,61	82,39
Tyr	80,91	88,88	74,83

Pošto je ponašanje spomenutih aminokiselinskih ostatka uočeno među raznim skupovima, koji su uostalom međusobno nezavisni, potrebno je pogledati razdiobe ASA vrijednosti kako bi se pokušao pronaći uzrok opažene pojave. Promatrati će se raspodjela ASA vrijednosti za lizin i cistein, pošto su za navedene aminokiselinske ostatke dobivene točnosti najveće, odnosno najmanje.

5.2.1. Utjecaj raspodjele ASA vrijednosti

Na slici 5.3 prikazane su raspodjele ASA vrijednosti za ostatke cistein i lizin. Podaci su uzeti iz skupa autora Manesha, no slične raspodjele uočene su i u ostalim skupovima. Raspodjele su prikazane kao histogrami, gdje y-os prikazuje frekvencije pojavljivanja ostatka unutar određenog intervala ASA vrijednosti.



Slika 5.3 Raspodjela ASA vrijednosti za cistein (lijevo) te lizin (desno)

Raspodjela ASA vrijednosti cisteina dosta se razlikuje od raspodjele vrijednosti za lizin. Cistein pripada skupini izrazito hidrofobnih aminokiselinskih ostataka, što objašnjava zašto velik broj ostataka ima ASA vrijednost blisku ili jednaku nuli. Od ukupno 1611 ostataka tipa lizin unutar skupa, PSAIA alat je za njih 802 odredio da im je površina dostupna otapalu manja od 5 $Å^2$. Fenilalanin također pripada grupi izrazito hidrofobnih ostataka te zbog toga ima razdiobu jako sličnu cisteinu. Sukladno tomu jedan je od najbolje predviđenih aminokiselinskih ostataka.

Pošto su prikazani rezultati za prozor duljine 9 ostataka, treba imati na umu da predviđanje nije vršeno za sve ostatke, već samo za one koji zadovoljavaju određenje uvjete. Spomenuti uvjeti su da nema prekida unutar prozora te da je ostatak za koji se vrši predviđanje središnji ostatak prozora.

Zbog toga je predviđanje vršeno za samo 1271 ostatak cisteina od ukupno 1611 što ih se nalazi u spomenutom skupu.

Osim raspodjele ASA vrijednosti, potrebno je pogledati vrijednosti pragova koje su korištene za raspored ostataka u hipotetske kategorije. Već je rečeno da je kvantizacija vršena pomoću Max-Lloyd kvantizatora s time da su korišteni sve vrijednosti aminokiselinskih ostataka koji su potencijalni elementi prozora. Dobiveni pragovi za spomenute aminokiselinske ostatke dani su u tablici 5.3, dok se u Dodatku nalaze svi pragovi korišteni u ovome radu. S pragom od 32,52 $Å^2$ te prikazanom razdiobom, većina ostataka cisteina, tj. njih 1063, dodijeljena je kategoriji 1.

Tablica 5.3 Pragovi za klasifikaciju u dvije kategorije

Vrsta ostatka	Prag [$Å^2$]
Cistein	32,52
Lizin	99,01

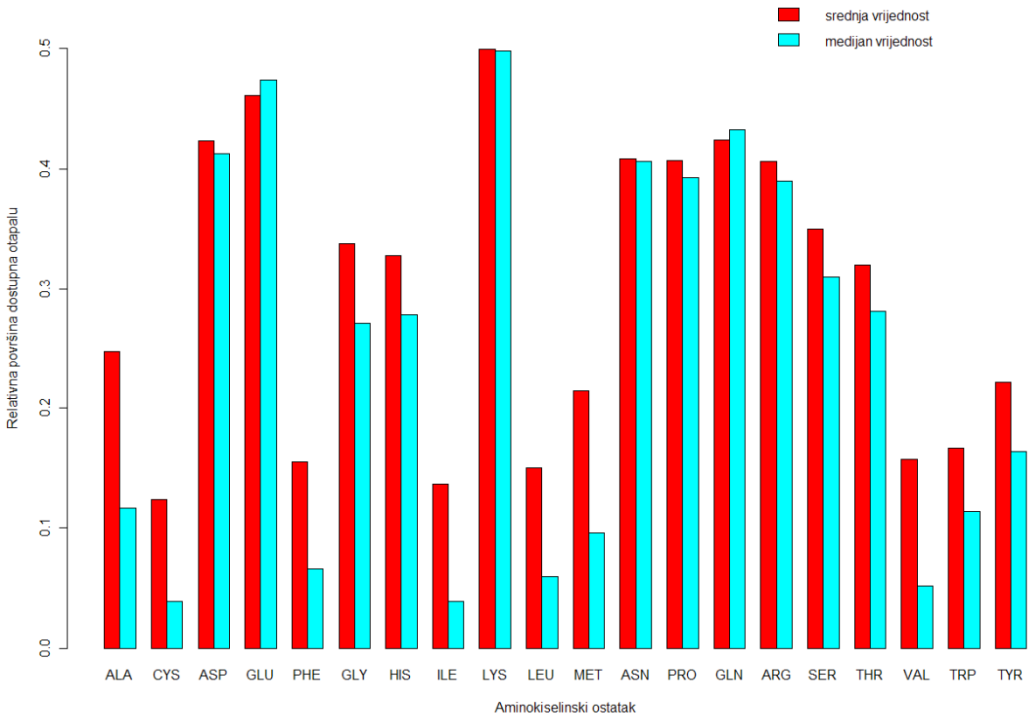
Zbog takve raspodjele klasa, očekivano je da će točnost biti visoka, jer samim favoriziranjem znatno brojnije klase postižu se dobri rezultati. Iz matrica greške koju generira PARF može se vidjeti da je od 1063 uzoraka ostatka koji pripadaju kategoriji 1, njih 1057 točno klasificirano, dok je od 208 uzoraka koji pripadaju kategoriji 2, samo njih 114 ispravno klasificirano. Pošto kategorija 1 broji višestruko više aminokiselinskih ostataka koji su klasificirani s jako visokom točnošću, ukupna točnost je također visoka.

Raspodjela vrijednosti za lizin nalikuje normalnoj razdiobi te se iz tablice 5.3 i pogledom na samu raspodjelu ASA vrijednosti može uočiti da prag dijeli skupinu ostataka na dvije kategorije s približno istim brojem elemenata. Zbog nepovoljnije situacije u kojoj nema većinske kategorije, točnosti klasifikacije po kategorijama su ujednačene, dok su ukupni rezultati predviđanja lošiji u odnosu na cistein.

Iz analize točnosti za cistein i lizin može se zaključiti da predstavljena metoda predviđanja ima znatno bolje rezultate za većinske kategorije. U slučaju aminokiselinskih ostataka čija raspodjela ASA vrijednosti nalikuje normalnoj razdiobi ne dolazi do stvaranja većinskih kategorija zbog ujednačenog broja

ostataka po kategorijama, što u konačnici rezultirala lošijim rezultatima predviđanja. Slične ponašanje može se uočiti u radu [4] M. Manesha i suradnika koji su predložili metodu baziranu na teoriji informacija. Autori su proizvoljno birali pragove koji su bili izraženi kao relativna površina dostupna otapalu, s time da su isti pragovi koristili neovisno o vrsti aminokiselinskog ostatka. Povećavajući vrijednost praga, stvarale su se većinske kategorije, prvenstveno kad je riječ o izrazito hidrofobnim ostacima, koji su zbog toga klasificirani s većom točnošću.

Zaključak ove analize je da je točnost predviđanja veća za izrazito hidrofobne aminokiselinske ostatke, što se u potpunosti slaže s rezultatima dobivenim za sva tri korištena skupa. U slučaju raspodjele ASA vrijednosti nalik normalnoj razdiobi, točnost predviđanja pada. Normalna razdioba je česta kod ostataka koji nemaju izraženu hidrofobnost, a takvi ostaci imaju i iznose površine dostupne otapalu veće od hidrofobnih ostatka.



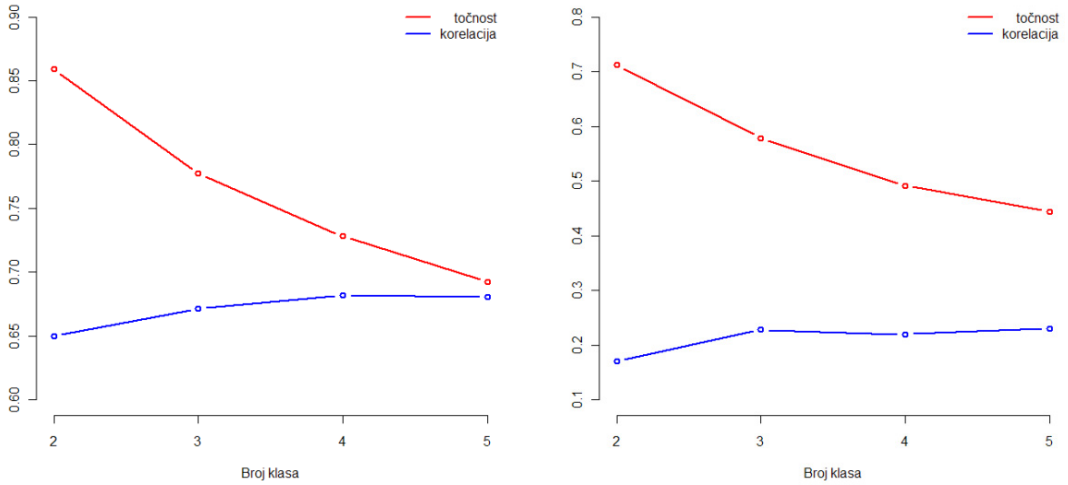
Slika 5.4 Raspodjela srednje i medijan RSA vrijednosti ovisno o aminokiselinskom ostatku. Ostaci kojima je srednja i medijan vrijednost bliska imaju raspodjelu vrijednosti nalik normalnoj razdiobi te je točnost predviđanja lošija.

5.3. Prikaz rezultata

Rezultati prikazani u nastavku dobiveni su za skupove Manesha te Rost-Sandera. Ti skupovi se često koriste pri provjeri točnosti metoda za predviđanju ASA vrijednosti te će se tako moći napraviti usporedba s drugim metodama. Za ocjenu kvaliteta predviđanja koristiti će se koeficijent korelacije te točnost. Prikazati će se i rezultati dobiveni zanemarivanjem profila slijeda, tj. korištenjem samo imena aminokiselina unutar pomičnog prozora. Na kraju će biti prikazani rezultati postignuti korištenjem pragova dobivenih pomoću medijan Lloyd-Max kvantizator opisanog u poglavlju 4.2.

5.3.1. Korištenje informacija iz sekvence i profila slijeda

Na slici 5.5 prikazana je ovisnost točnosti i korelacije o broju klasa. Rezultati su dani za pomični prozor duljine 9 aminokiselinskih ostatka. Isti rezultati prikazani su numerički u tablici 5.4.



Slika 5.5 Ovisnost točnosti i koeficijenta korelacije o broju klasa za skupove Manesha (lijevo) te Rost-Sandera (desno)

Iz slike se može uočiti da za oba skupa dolazi do naglog pada točnosti kako se povećava broj klasa u koje klasifikator raspodjeljuje ostatke. Zbog različitog mjerila, na prvi pogled možda nije vidljivo da je pad dosta izraženiji za skup Rost-Sandera. Korelacija između predviđenih i hipotetskih klasa pri tome neznatno rasta s povećanjem broja klasa kao posljedica ravnomjernijeg

rasporeda hipotetskih klasa. Dok je pad točnosti bio očekivan za veći broj klasa, postavlja se pitanje zašto postoji toliko velika razlika u kvaliteti predviđanja između skupova.

Tablica 5.4 Ovisnost točnosti i korelacije o broju klasa

	Manesh 215			
	2	3	4	5
Točnost	85,89	77,71	72,81	69,24
Korelacija	65,01	67,16	68,17	68,06
	Rost-Sander 126			
	2	3	4	5
Točnost	71,16	57,85	49,16	44,34
Korelacija	17,04	22,80	21,96	23,08

Skup Rost-Sander sadrži 126 relativno kratkih proteinskih lanca, od kojih 20 ima manje od 100 aminokiselinskih ostataka. Pri određivanju profila slijeda PSI-BLAST algoritam je u većini slučaja prekinut ograničenjem broja iterativnih pretraživanja, nego li konvergiranjem algoritma. Nedovoljno točno određeni profili slijeda tako smanjuju kvalitetu predviđanja te se postavlja pitanje mogu li se postići bolji rezultati ako se zanemare profili slijeda.

5.3.2. Predviđanje bez korištenja profila slijeda

Rezultati dobiveni korištenjem samo imena aminokiselinskih ostataka unutar prozora dani su u tablici 5.5. Rezultati su lošiji nakon što su zanemareni profili slijeda, pri čemu je pad koeficijenta korelacije znatno izraženiji od pada točnosti.

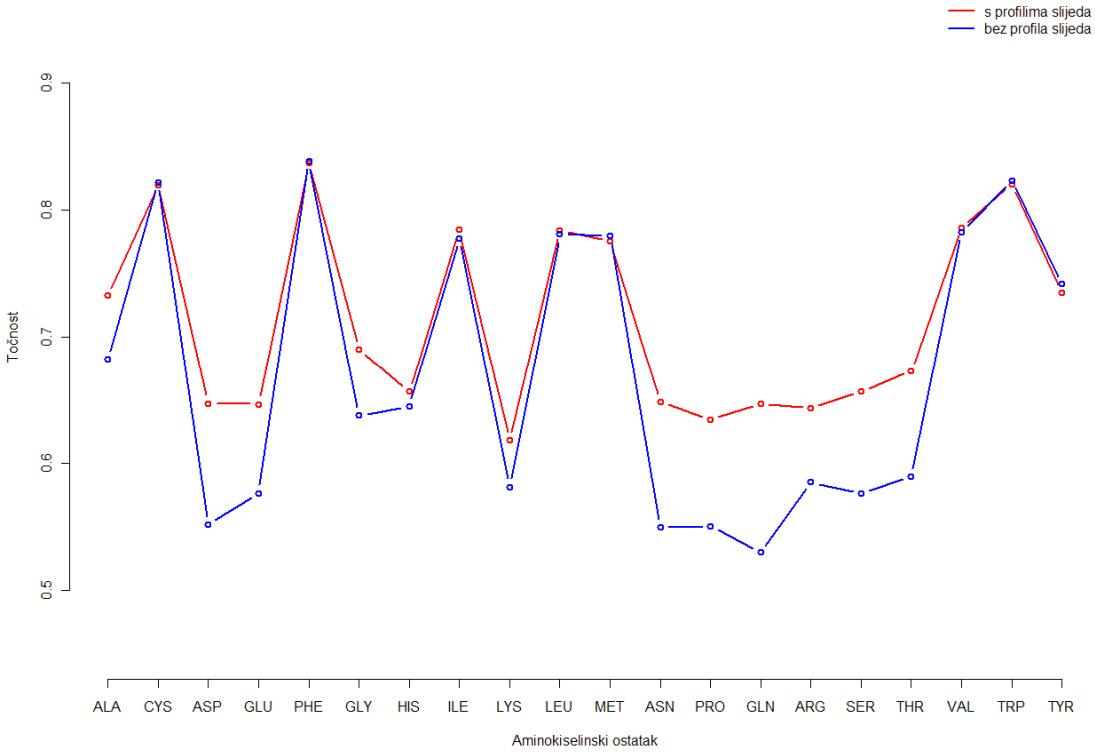
Tablica 5.5 Ovisnost točnosti i korelacije o broju klasa za skup Rost-Sander. Predviđanje je vršeno za prozor duljine 9 ostataka s zanemarenim profilima slijeda.

	2	3	4	5
Točnost	66,99	52,88	44,58	39,99
Korelacija	7,41	5,52	5,59	7,12

Može se zaključiti da imena aminokiselinskih ostataka nisu dostatna za postizanje zadovoljavajuće kvalitete predviđanja. Promatranjem podataka o važnosti varijabli može se uočiti da se za neke ostatke profili slijeda imaju

veću važnost, dok za druge imaju imena ostataka unutra prozora. Podatke o važnosti varijabli generira PARF, kao što je objašnjeno u poglavlju 4.3.

Sa slike 5.6 može se uočiti da za izrazito hidrofobne ostatke točnost ne pada ako se predviđanje vrši bez korištenja profila slijeda. Crvena linije prikazuje točnost dobivenu korištenjem profila slijeda, dok je plava linija dobivena zanemarivanjem profila slijeda te korištenjem samo imena ostataka unutar prozora.



Slika 5.6 Ovisnost točnosti o vrsti aminokiselinskog ostatka za skup Rost-Sander zanemarivanjem te korištenjem profila slijeda

5.3.3. Medijan Lloyd-Max kvantizator. Rezultati predviđanja

Rezultati dobiveni korištenjem Lloyd-Max kvantizatora koji za određivanje kvantizacijskih granica uzima medijan vrijednost susjednih kvantizacijskih razina dani su u nastavku. Korišteni kvantizator manje je osjetljiv na *outliere.* Na slici 5.7 prikazana je točnost za klasifikaciju u dvije kategorije koristeći dvije varijante Lloyd-Max kvantizatora. Točnosti su lošije za većinu

aminokiselinskih ostataka, pri čemu je naročito pogoršanje uočljivo za izrazito hidrofobne ostatke koji su primjenom klasičnog Lloyd-Max algoritma klasificirani s najvećom točnošću.

Dobiveni rezultati su se mogli i predvidjeti, pošto medijan Lloyd-Max algoritma ne iskorištava raspodjelu ostataka u većinske kategorije za slučaj izrazito hidrofobnih ostataka. Kako novi kvantizator daje identične pragove za slučaj normalne razdiobe, ostaci kojima raspodjela ASA vrijednosti nalikuje normalnoj razdiobi imaju pragove sličnog iznosa kao u slučaju klasičnog Lloyd-Max kvantizatora. Time je promjena točnosti znatno manja nego u slučaju ostataka s većinskim kategorijama.

Slika 5.7 Ovisnost točnosti o vrsti Lloyd-Max algoritma za kvantizaciju u dvije kategorije. Rezultati su dani za skup Manesh-215.

5.4. Prikaz rezultata koji su postigli drugi autori

U nastavku će se prikazati rezultati predviđanja za dvije različite metode. U oba rada, autori bi izbacili jedan od ukupno *k* lanaca skupa te bi se algoritam

trenirao na preostalim proteinskim lancima. Postupak se ponavlja *k* puta, tj. dok se ne izvrši predviđanje za svaki od izdvojenih lanca. Za granice klasa koristili su proizvoljno određene pragove, prikazane kao RSA vrijednosti.

Prva metoda [4] zasniva se na principima teorije informacija, a predložena je od autora H. Manesha i suradnika. Koristio se je pomični prozor duljine 17 aminokiselinskih ostataka, a prikazani slučaj dobiven je korištenjem informacija para (engl. *pair information*) središnjeg ostatka sa svakim od njegovih susjeda.

Slijedeća metoda [12] s kojom će se vršiti usporedba temelji se na, kao i u ovome radu, profilima slijeda, a predložena je od autora G. Gianesea i suradnika. Način određivanja profila te izbor pragova utjecali na razliku u kvaliteti predviđanja, što će se biti pokazano u nastavku. Autori koriste prozor duljine 17 ostatka, a način određivanje profila slijeda se razlikuje od metode opisane u ovome radu.

Tablica 5.6 Rezultati predviđanja dobiveni metodom [4] temeljenoj na teoriji informacija. Rezultati su prikazani za skup Manesh-215.

broj kategorija	pragovi	točnost	korelacija
dvije kategorije			
	4	75,1	0,49
	9	75,9	0,51
	16	75,5	0,50
	25	74,4	0,47
	36	74,1	0,41
	49	79,9	0,36
	64	97,2	0,46
	81	80,5	0,05
tri kategorije			
	4;25	49,3	0,39
	4;36	57,9	0,43
	9;16	62,3	0,42
	9;36	57,4	0,41
	9;64	74,1	0,47
	16;64	73,7	0,47
četiri kategorije			
	9;16;36	45,2	0,32
	9;36;49	41,2	0,25
	4;16;36	46,4	0,36
	4;16;49	51,8	0,37
	4;25;49	47,1	0,34

U tablici 5.6 prikazani su rezultati predviđanja dobiveni u radu [4] Manesha i suradnika za klasifikaciju u dvije do četiri kategorije. Korišten je skup od 215 proteina, koji je prvi put predstavljen od strane istoimenog autora.

Pri klasifikaciji u dvije kategorije točnost se kretala od 74,1% do 97,2%, s time da se s povećavanjem vrijednosti praga točnost povećavala. Povećanjem broja klasa točnost se smanjuje na vrijednosti oko 75%, što odgovara slučaju za klasifikaciju u 3 kategorije. Pri tome koeficijent korelacije neznatno pada. U slučaju klasifikacije u 4 kategorije, točnost nastavlja padati na iznos oko 45%, ovisno o pragu. Najveća točnost pri klasifikaciju u dvije kategorije za skup Rost-Sander iznosi 78,2%. Navedena vrijednost točnosti je postignuta za RSA prag od 9%.

Rezultati predviđanja pomoću metode [4] bazirane na profilima slijeda prikazani su u tablici 5.7. Prikazani su za skupove Rost-Sandera te Manesha za klasifikaciju u dvije i tri kategorije. Točnost klasifikacija je lošija za oba skupa u odnosu na metodu temeljenu na teoriji informacija.

Tablica 5.7 Rezultati predviđanja dobiveni metodom [12] temeljenoj na profilima slijeda. Prikazani su rezultati za klasifikaciju u dvije i tri kategorije

skup proteinskih lanaca	pragovi (%)	postotak točne klasifikacije	koeficijent korelacije
Rost-Sander-126			
	9	72,8	0,387
	16	71,5	0,421
	23	71,4	0,426
	6;64	61,7	0,408
	9;36	54,7	0,462
Manesh-215			
	4	75,5	0,374
	9	72,8	0,423
	16	72,2	0,445
	25	72,8	0,403
	36	79,5	0,282
	4;36	56,2	0,488
	9;16	66,4	0,457

6. Zaključak

Cilj ovoga rada bio je utvrđivanje kvalitete predviđanja površine dostupne otapalu na osnovu podataka iz slijeda aminokiselinskih ostatka. Predviđanje se je vršilo metodom slučajnih šuma, odnosno njenom paralelnom implementacijom *PARF*. Aminokiselinski ostaci su se svrstavali u dvije do pet kategorija prethodno određenih pomoću Lloyd-Max kvantizatora ovisno o vrsti ostatka. Kao ulazni podaci koristili su se pomični prozori duljine 3 do 17 aminokiselinski ostataka koji su osim imena ostataka, sadržavali i profile slijeda.

Najbolji rezultati predviđanja dobiveni su za duljinu prozora od 9 ostataka. Zbog karakteristika algoritma slučajnih šuma, razlika točnosti između prozora različite duljine je neznatna, iako prozor duljine 3 ostatka daje nešto lošije rezultate. Uočena je i ovisnost točnosti o vrsti aminokiselinskog ostatka, tako da su izrazito hidrofobni aminokiselinski ostaci poput cisteina i fenilalanina klasificirani s najvećom točnošću.

Najbolji rezultati su uočeni za skup Manesha, pri čemu je točnost za klasifikaciju u dvije kategorije iznosila 85,89%. Povećanjem broja kategorija točnost počinje padati sve do iznosa od 69,24%, koliko iznosi za klasifikaciju u pet kategorije. Korelacija se pri tome kreće između 65,01% i 68,06%.

Za skup Rost-Sandera, točnost pri klasifikaciji u dvije kategorije iznosi 71,16% Razlog pada točnosti leži u većem broju kratkih proteinskih lanca unutar skupa za koje PSI-BLAST algoritam nije dovoljno točno odredio profile slijeda. Korelacija za skup Rost-Sandera kreće se od 17,04% do 23,08%, koliko iznosi za klasifikaciju u dvije, odnosno pet kategorija.

7. Literatura

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

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

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

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

[5] **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.**

[6] **G. Topic and T. Smuc, "PARF - Parallel RF Algorithm," Zagreb: Institut Rudjer Boskovic, 2004.**

[7] **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.**

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

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

[10] **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**

[11] **Z. Yuan and B. Huang, "Prediction of Protein Accessible Surface Areas by Support Vector Regression", Proteins, vol. 57, pp. 558-564, 2004.**

[12] **G. Gianese, F. Bossa, S. Pascarella, "Improvement in prediction of solvent accessibility by probability profiles" , Protein Engineering, vol. 16 pp. 987-992, 2003**

Dodaci

Dodatak 1 Granice kategorija za skupove Manesh (MN215) i Rost-Sander (RS126)

MN 215	Broj kategorija									
	2	3		4			5			
	prag	1.prag	2.prag	1.prag	2.prag	3.prag	1.prag	2.prag	3.prag	4.prag
Ala	38,46	24,68	64,55	16,44	45,44	79,14	14,90	41,19	69,41	107,75
Cys	32,52	21,60	63,17	15,98	44,62	82,59	15,71	43,66	80,09	136,19
Asp	63,15	41,92	88,72	32,40	67,92	104,45	23,54	51,58	79,39	110,77
Glu	75,40	52,13	103,61	43,37	84,25	121,67	35,12	69,87	100,80	133,66
Phe	61,26	33,06	97,64	24,23	64,51	118,73	16,87	44,21	79,87	128,13
Gly	32,51	20,01	50,45	14,96	37,63	61,87	14,82	37,17	60,20	97,85
His	68,70	46,38	104,37	30,49	72,31	121,77	26,74	60,78	95,52	135,78
Ile	46,82	24,54	74,41	20,81	58,19	103,58	14,91	41,12	73,58	114,94
Lys	99,01	68,91	122,62	54,64	96,48	138,82	50,16	86,97	119,67	154,74
Leu	48,62	29,33	84,74	20,93	58,26	108,06	15,50	42,98	76,62	120,35
Met	64,33	35,70	102,18	26,96	69,99	122,30	28,80	65,44	106,99	151,85
Asn	62,36	44,31	95,56	34,02	71,89	110,14	23,49	53,56	83,25	116,63
Pro	57,05	36,44	81,99	27,95	62,81	96,70	18,22	44,25	71,72	101,22
Gln	72,23	50,91	103,54	39,08	80,04	119,68	30,64	64,37	95,62	128,81
Arg	101,07	66,66	132,06	54,16	102,89	156,23	46,43	88,33	128,72	174,50
Ser	45,57	29,81	70,59	21,82	51,13	82,37	18,09	43,04	68,66	96,71
Thr	48,49	33,35	77,95	23,17	56,54	93,62	19,26	46,59	72,49	102,70
Val	41,95	27,11	75,03	16,19	46,85	88,55	13,05	37,88	67,57	104,27
Trp	64,31	34,75	98,35	26,32	66,99	124,42	23,39	56,22	96,74	150,18
Tyr	71,38	39,40	104,38	29,45	74,00	133,76	25,01	59,20	100,96	155,31
RS 126	**Broj kategorija**									
	2	**3**		**4**			**5**			
	prag	**1.prag**	**2.prag**	**1.prag**	**2.prag**	**3.prag**	**1.prag**	**2.prag**	**3.prag**	**4.prag**
Ala	39,60	26,58	66,90	17,10	44,52	75,98	14,05	38,38	65,70	95,36
Cys	33,01	23,37	64,10	20,73	53,73	106,39	13,71	36,54	66,34	121,90
Asp	63,43	43,85	90,32	29,54	65,07	102,17	27,59	58,63	86,42	115,53
Glu	79,24	55,87	107,69	42,64	84,10	123,46	35,92	71,13	103,73	138,35
Phe	67,48	32,73	98,81	28,64	77,77	137,13	19,50	50,97	94,49	148,11
Gly	34,12	21,38	53,03	16,43	40,21	65,10	15,96	38,66	61,34	91,49
His	69,49	54,73	118,48	31,16	72,62	125,92	30,10	70,42	115,59	168,65
Ile	50,90	29,59	85,30	23,84	65,36	115,50	17,92	48,29	84,06	128,79
Lys	100,78	77,44	131,71	59,03	101,24	144,12	57,61	99,22	140,68	210,48
Leu	53,96	32,81	89,09	24,99	66,02	114,60	17,90	48,27	84,47	127,10
Met	63,91	44,51	115,82	25,10	69,49	125,80	25,09	69,11	119,64	173,02
Asn	61,65	44,41	93,59	31,63	68,96	107,09	27,18	58,70	87,80	118,40
Pro	58,39	38,88	84,71	31,58	67,05	100,53	23,40	52,55	80,69	110,19
Gln	74,92	52,68	107,37	45,26	89,43	132,96	32,66	68,73	101,74	138,34
Arg	105,40	69,17	137,88	54,05	107,54	162,37	47,59	90,57	128,56	170,93
Ser	47,83	33,23	76,35	24,16	55,81	88,74	19,43	46,51	74,41	105,29
Thr	54,60	38,00	86,23	26,61	59,99	96,92	19,05	45,77	73,10	104,84
Val	47,23	28,04	75,94	20,69	57,54	102,25	20,11	54,91	93,07	147,67
Trp	76,87	37,79	113,38	34,98	91,84	159,08	26,22	67,19	118,85	177,94
Tyr	74,55	46,27	108,58	28,58	71,20	125,42	27,96	69,83	122,51	191,55

Sažetak

U okviru ovoga rada opisana je metoda predviđanja površine dostupne otapalu na osnovu podataka iz slijeda aminokiselinskih ostatka. Kvantizacija površine dostupne otapalu vršena je primjenom Lloyd-Max kvantizator, pri čemu su vrijednosti kvantizirane u dvije do pet razina. Kvantizacijske razine zasebno su određene za svaku pojedinu vrstu ostatka. Za predviđanje se je koristila paralelna implementacija algoritma slučajnih šuma *PARF*. Previđanje se je vršilo za središnji ostatak pomičnog prozora, čija se je duljina kretala od 3 do 17 ostataka. Kao vektor ulaznih podataka koristila su se imena ostataka u prozoru te njihovi profili slijeda tako da je ukupan broj svojstava na temelju kojih se je obavljalo predviđanje iznosio $21 \times n$.

Pregledom rezultata uočena je najveća točnost za prozore duljine 9 ostataka. Uočena je i ovisnost točnosti o vrsti aminokiselinskog ostatka, tako da su izrazito hidrofobni aminokiselinski ostaci poput cisteina i fenilalanina klasificirani s najvećom točnošću. Ostaci čija je raspodjela vrijednosti površine dostupna otapalu nalik normalnoj razdiobi bile su klasificirane s najmanjom točnošću.

Za skup predložen od strane autora Rosta i Sandera dobiveni su najlošiji rezultati. Razlog leži u većem broju kratkih proteinskih lanca unutar skupa za koje PSI-BLAST algoritam nije dovoljno točno odredio profile slijeda. Pokušaj predviđanja korištenjem samo imena aminokiselina unutar prozora dao je nešto lošije rezultate u usporedbi s onima dobivenim korištenjem profila slijeda. Izrazito hidrofobni ostaci su pri tome klasificirani s istom točnošću neovisno o tome koriste li se informacije o profilima slijeda ili ne.