SVEUČILIŠTE U ZAGREBU

FAKLTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 66

PREDVIĐANJE MJESTA PROTEINSKIH INTERAKCIJA KORISTEĆI ALGORITAM SLUČAJNIH ŠUMA

Juraj Petrović

Zagreb, lipanj 2010.

SVEUČILIŠTE U ZAGREBU

FAKLTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 66

PREDVIĐANJE MJESTA PROTEINSKIH INTERAKCIJA KORISTEĆI ALGORITAM SLUČAJNIH ŠUMA

Juraj Petrović

Zagreb, lipanj 2010.

Sadržaj:

1. Uvod

Bioinformatika je ime relativno mladog područja znanosti usmjerenog na primjenu informacijskih tehnologija i računalnih znanosti u organizaciji, povezivanju, pohranjivanju, analizi, vizualizaciji složenih bioloških procesa i molekularnoj biologiji općenito. Prema definiciji američkog National Centera for Biotechnology Information [2], postoje tri značajne poddiscipline unutar bioinformatike i to su razvoj novih algoritama i statistike kojima se određuju veze između članova velikog skupa podataka, analiza i interpretacija različitih tipova podataka uključujući nukleotidne i aminokiselinske sljedove, proteinske domene i strukture proteina, te razvoj i implementacija alata koji omogućuju učinkovit pristup različitim tipovima podataka kao i njihovo učinkovito upravljanje. Primjena informatičkih metoda u ovim procesima rezultira brojnim prednostima među kojima sigurno valja istaknuti brzu obradu velikih količina podataka i pouzdane rezultate.

Jedno od mnogih područja kojima se bavi bioinformatika je i predviđanje proteinskih interakcija (engl. *protein-protein interactions*, PPI*s*), a to je ujedno i tema ovog diplomskog rada. Proteini su od ključne važnosti za gotovo sve procese u tijelu pa bi stoga uspješno predviđanje njihovih interakcija omogućilo bolji uvid u te procese, kao i mogućnost utjecanja na njih. U sklopu ovog diplomskog rada najprije je dan pregled dosadašnjih radova iz ovog područja, kao i važnosti i različitih definicija proteinskih interakcija u 2. poglavlju. Za predviđanje proteinskih interakcija korišteni su neredundantni skupovi proteinskih lanaca iz RSCB PDB baze, dostupni za preuzimanje putem Interneta. Informacije o tim lancima dobivene su iz odgovarajućih PDB datoteka, a pomoću PSAIA alata na njima su određena mjesta interakcija. Ovaj postupak opisan je u 3. poglavlju rada. Algoritmom slučajnih šuma, detaljno opisanim u 4. poglavlju, realiziran je klasifikator koji je ista mjesta trebao predvidjeti. Rezultati klasifikacije navedeni su u 5. poglavlju, a u 6. poglavlju je objašnjen njihov značaj i mogućnosti nadogradnje.

2. Proteinske interakcije

2.1 Proteini i proteinske interakcije

Proteini su makromolekule građene od aminokiselina, sa gotovo bezbrojnim funkcijama u ljudskom tijelu. Oni su odgovorni primjerice za izgradnju mišića i kože, probavu hrane, rast stanica i emocije. Provođenja signala (engl. *signal transduction*), koje u biologiji predstavlja mehanizam pretvaranja kemijskog ili mehaničkog podražaja stanice u njenu specifičnu reakciju, realizira se također posredovanjem proteina. Ljudsko tijelo proteine neprestano sintetizira, a greške u sintezi mogu dovesti do njihovih novih ili izmijenjenih oblika. Ova pojava najčešće ne utječe u većoj mjeri na tijelo, ali moguće je i da dovede do poboljšanja neke funkcije ili čak bolesti (Parkinsonova i Alzheimerova bolest)[3]. Proteini su rijetko aktivni samostalno i uglavnom djeluju uz prisutnost i interakciju s drugim proteinima, pa su stoga proteinske interakcije na gotovo najnižoj razini odgovorne za složene biološke procese.

2.2 Važnost predviđanja proteinskih interakcija

Važnost predviđanja proteinskih interakcija proizlazi iz potrebe i želje za objašnjenjem tih bioloških procesa i primjene stečenih spoznaja u drugim područjima. Praktična primjena takvih spoznaja moguća je primjerice u sintezi novih lijekova i cjepiva (engl. *drug design*), gdje bi one omogućile razumijevanje mehanizama bolesti i tako efikasnije liječenje i ispitivanje samog lijeka.

Mjesta na kojima dolazi do interakcije između dva proteinska lanca moguće je odrediti i eksperimentalno tehnikama od kojih su najpopularnije rentgenska kristalografija (engl. *X-ray crystallography*) i nuklearna magnetska rezonanca (engl. *nuclear magnetic resonance*). Ove tehnike mogu dati vrlo precizne rezultate čak i u relativno kratkom vremenu (do nekoliko dana)[5]. Eksperimentalno je određivanje strukture proteinskih kompleksa i interakcija unatoč tome vrlo složen proces, a broj tako određenih struktura je u porastu, no ipak ograničen, zbog čega računalne metode predviđanja istih dobivaju na važnosti [4].

Dosad su već razvijane brojne metode predviđanja proteinskih interakcija, no većina ih ima nedostatak da im je za relativno uspješno predviđanje potrebno vrlo mnogo informacija o proteinima čije interakcije se analiziraju [4]. Čak i uz te informacije rezultati klasifikatora pokazuju kako ima još mnogo prostora za poboljšanja. U nekim radovima, kao i u ovome,

autori mjesta proteinskih interakcija nastoje predvidjeti u najvećoj mjeri iz sekvence aminokiselinskih ostataka uz eventualno još neke informacije.

2.3 Definiranje mjesta proteinske interakcije

Kako bi uopće bilo moguće analizirati mjesta proteinskih interakcija potrebno je najprije definirati što uopće podrazumijevamo pod tim izrazom. Općenito, postoje različiti tipovi proteinskih interakcija [5], a i mjesto proteinske interakcije može se definirati na više načina.

U ovom radu korištene su dvije definicije mjesta proteinske interakcije: jedna, koja zadaje najveću dopuštenu udaljenosti između dva teška atoma proteina koji sudjeluju u interakciji, i druga, koja koristi PIADA algoritam PSAIA programskog alata.

Neke definicije na spomenutu najveću dopuštenu udaljenost dva teška atoma nadodaju da u pomičnom prozoru duljine n uzastopnih aminokiselinskih ostataka (n je obično neparan broj) navedeni uvjet udaljenosti teških atoma mora biti zadovoljen za središnji aminokiselinski ostatak, kao i za još m ostataka na udaljenosti najviše k od središnjeg ostatka. Vrijednosti parametra n obično se kreću između 9 i 13, a m između 1 i 6 [6]. U ovom radu korištene su vrijednosti n=9, m=1, k=4 i n=9, m=5, k=3.

2.4 Dosad ostvareni rezultati u ovom području

Rezultati dosad ostvareni u ovom području variraju i općenito ih nije jednostavno uspoređivati. Razlike proizlaze ne samo iz manje ili više uspješne realizacije klasifikatora odnosno modula za predviđanje, nego također i u ovisnosti o dva parametra: definiciji mjesta proteinske interakcije i skupa podataka korištenog za predviđanje. Neki od važnijih radova navedeni su u nastavku.

Ofran i Rost su u svom radu [7] za maksimalnu udaljenost teških atoma proteina koji sudjeluju u interakciji odredili vrijednost 6Å, a mjesto interakcije je bilo ono gdje je taj uvjet zadovoljen za središnji i još barem 4 aminokiselinska ostatka na udaljenosti ne više od 3 od središnjeg ostatka. Pomični prozor bio je veličine 9 aminokiselinskih ostataka. Uz klasifikator u obliku neuronske mreže treniran samo na podacima o sekvenci aminokiselinskih ostataka postigli su preciznost od 70% i odziv od 0.5%.

U kasnijem radu [8] isti autori postigli su bolji rezultat tako da su među informacije za predviđanje dodali evolucijske profile, površinu dostupnu otapalu (ASA) i sekundarnu strukturu. Mjesto interakcije redefinirali su tako da za istu maksimalnu udaljenost teških

atoma na udaljenosti najviše 5 ostataka od središnjeg postoji još barem 6 ostataka koji su u kontaktu sa susjednim lancem. Klasifikacijom neuronskom mrežom autori su ostvarili rezultate od 60-70% preciznosti i više od 10% odziva.

Wang i ostali [9] definirali su kao mjesto interakcije ono mjesto na proteinu kod kojega je udaljenost dva α atoma ugljika s dva susjedna lanca udaljeni za manje od 12Å, a u sklopu pomičnog prozora duljine 11 ostataka. Korištenjem SVM i kombinacijom rezultata dobivenih za pojedina svojstva postigli su preciznost od 49.7% i odziv od 66.3%.

M. Šikić [1] je koristeći informacije iz sekvence postigao preciznost od 60-70% i odziv od oko 40%. Mjesto interakcije definirao je kao mjesto gdje se na udaljenosti manjoj od 6Å nalaze dva teška atoma lanaca koji su u interakciji, a osim središnjeg ostatka u kontaktu moraju biti još barem 4 ostatka na udaljenosti ne većoj od 3 ostatka.

3. Podaci

3.1 Neredundantni skupovi proteina

Kao temeljni podaci za ovaj rad korišteni su neredundantni skupovi proteinskih lanaca: u prvom slučaju homo, u drugom hetero, a u trećem i homo i hetero lanaca. Skupovi su sadržavali redom 2760, 4280 i 6574 proteinskih lanaca, a uz njihov popis nužno je svakako imati pristup njihovim konkretnim opisima. Oni su pak dostupni za preuzimanje iz RSCB PDB baze podataka u obliku tekstualnih datoteka s PDB ili ENT ekstenzijom. Za sve lance sadržane u skupovima vrijedi da su dijelovi višelančanih proteina (multimera) koji sadrže barem po dva lanca duga barem 30 aminokiselinskih ostataka, da im je rezolucija manja ili jednaka 3.5 Å, te da je maksimalna sličnost između primarne strukture dva proteinska lanca u skupu 35%. Koji od proteina dostupnih u bazi su multimeri određeno je pomoću PISA Web servera dostupnog na adresi http://www.ebi.ac.uk/msd-srv/prot_int/cgi-bin/piserver. Zbog prirode eksperimentalnih postupaka kojima su dobiveni neki od postojećih opisa to nije moguće odrediti samo iz broja lanaca opisanih unutar pojedine ENT ili PDB datoteke.

Detaljan opis postupka izrade ovakvih skupova moguće je pronaći u radu [5], a same skupove na digitalnom mediju priloženom ovom diplomskom radu.

3.2 Priprema ulaznih podataka

3.2.1 Definicije mjesta proteinske interakcije

U radu su, kako je i prethodno spomenuto, korištene dvije definicije interakcije: prva, u kojoj se definira kako su u interakciji svaka dva teška atoma proteina koji sudjeluju u interakciji udaljena manje od 6 Å, i druga, koja koristi PIADA algoritam. Oba postupka implementirana su u sklopu PSAIA alata dostupnog za preuzimanje na adresi http://complex.zesoi.fer.hr/hr/PSAIA.html. Pomoću oba ova postupka najprije su određena pojedinačna mjesta kontakata na lancima proteina.

3.2.2 Definicija uzorka za potrebe predviđanja

Podaci koji su se koristili za predviđanje interakcija bili su informacije o sekvenci, to jest imena 9 uzastopnih aminokiselinskih ostataka iz pomičnog prozora koji se kretao po čitavom lancu i njima odgovarajući profili. Profili slijeda predstavljaju mjeru evolucijske očuvanosti i daju informaciju o vjerojatnosti pronalaska svake od 20 standardnih

aminokiselina na mjestu one aminokiseline čiji profil se određuje. Ovako definirani vektor podataka sastoji se od 189 elemenata: imena 9 aminokiselinskih ostataka i po 20 profila za svaki od njih.

Svakom vektoru podataka pridružuje se ciljna vrijednost klasifikacije 0 ili 1, u ovisnosti o zadovoljavanju uvjeta za proglašenjem mjesta interakcije. Za svaki od prethodnih načina određivanja pojedinačnih mjesta interakcije korištena su dva različita uvjeta: prvi koji definira kako je središnji od 9 aminokiselinskih ostataka mjesto interakcije ako je tako procijenjeno na temelju maksimalne udaljenosti teških atoma u prvom slučaju, odnosno PIADA algoritmom u drugom slučaju. Drugi uvjet definira kako je središnji od 9 aminokiselinskih ostataka mjesto interakcije, ako je tako procijenjeno na temelju maksimalne udaljenosti teških atoma u prvom slučaju, odnosno PIADA algoritmom u drugom slučaju, ali da to također vrijedi za još bar 4 aminokiselinska ostatka iz prozora na udaljenosti ne većoj od 3 mjesta od središnjeg ostatka. Ova definicija korištena je u radovima [1], [6] i [7].

3.2.3 **Generiranje uzoraka za potrebe predviđanja**

Uzorci za predviđanje generirani su iz izlaza PSAIA programskog alata. Kao ulaz za PSAIA alat poslužile su PDB datoteke čiji lanci su sadržani u spomenutim neredundantnim skupovima i na temelju njih su generirane XML izlazne datoteke sa podacima o kontaktima među proteinskim lancima, kao i datoteke s opisima samih lanaca. Navedene su datoteke, kako je već navedeno, generirane za dva slučaja: za definiranje mjesta interakcije pomoću maksimalne udaljenosti i PIADA algoritma, a željena od te dvije mogućnosti zadaje se u PSAIA alatu. Među dobivene podatke još je bilo potrebno dodati informacije o profilima, što je učinjeno pomoću skripte korištene u radu [6], te ciljnu vrijednost klasifikacije, pomoću skripte iz rada [1].

Budući da je za predviđanje proteinskih interakcija korišten paket *Rattle* programskog jezika *R*, za lakše je učitavanje i manipuliranje podacima ulazne vektore bilo potrebno iz XML formata povezati i organizirati u ARFF (engl. Attribut Relation File Format) datoteku. Ovaj format, koji se sastoji od zaglavlja i podataka, koristi velik broj aplikacija za klasifikaciju. Identifikatori zapisa u ARFF datoteci su @RELATION (javlja se na početku datoteke), @ATTRIBUTE (javlja se nakon @RELATION i označava postojanje atributa čije ime je navedeno iza ove ključne riječi) i @DATA (pojavljuje se samo jednom, nakon @ATTRIBUTE identifikatora i znači da slijedi blok podataka organiziranih u retke i

međusobno odvojene zarezima). Pretvorba u ARFF format obavljena je alatom iz rada [1], a primjer iz generirane ARFF datoteke nalazi se u nastavku:

```
1A0A,A,5,LYS,ARG,GLU,SER,HIS,LYS,HIS,ALA,GLU,5,40,0,0,0,15,3,0,0,0,0,35,0
,0,0,2,0,0,0,0,2,69,0,0,0,5,0,0,0,0,0,24,0,0,0,0,0,0,0,0,4,4,1,3,0,7,22,0
,2,2,15,10,6,0,0,8,12,0,0,3,15,5,22,0,0,3,1,0,0,0,2,8,2,2,1,22,12,0,0,4,7
,7,0,0,0,0,0,0,73,0,0,9,0,0,0,0,0,0,3,0,1,3,39,0,1,0,0,2,0,9,0,9,1,0,9,20
,2,0,0,5,17,0,0,1,0,5,12,0,2,18,21,2,3,2,0,6,0,0,0,10,28,7,0,0,0,0,0,0,0,
18,8,0,6,0,2,12,4,0,1,13,2,0,0,0,0,0,98,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
```

Prva dva elementa redom predstavljaju oznake PDB strukture i lanca kojemu promatrani uzorak pripada. Potom slijedi indeks središnjeg ostatka u cijelom lancu i nakon toga sekvenca od 9 aminokiselinskih ostataka. Nakon toga dolazi 9×20 profila, te na kraju pripadnost klasi, koja iznosi 0 ako središnji ostatak u navedenoj sekvenci nije mjesto interakcije, odnosno 1 ako on je mjesto interakcije.

Za potrebe rada generirano je ukupno dvanaest skupova podataka. Svaki od tri osnovna skupa (homo/hetero/homo i hetero) korišten je, naime, u četiri varijante nastale kombiniranjem svakog od dva načina određivanja pojedinačnih mjesta proteinskih interakcija u sklopu alata PSAIA (maksimalna udaljenost ili PIADA algoritam) sa svakim od dva kriterija za dodjeljivanje uzorku ciljne klasifikacije 0 ili 1 (središnji aminokiselinski ostatak u prozoru mora biti mjesto interakcije ili središnji i još bar 4 ostataka na udaljenosti ne većoj od 3 mjesta od središnjeg ostatka).

Broj uzoraka unutar skupa ne ovisi o načinu definiranja mjesta interakcije i kriteriju za dodjelu ciljne klasifikacije, nego samo o broju i duljini lanaca sa kojih se uzorci uzimaju. Skup koji se sastoji samo od hetero lanaca polučio je stoga 325170 uzoraka, skup homo lanaca polučio je 836415 uzoraka, a skup od homo i hetero lanaca polučio je 1064668 uzoraka.

4. Metode

4.1 Algoritam slučajnih šuma

4.1.1 Općenito o algoritmu slučajnih šuma

Algoritam slučajnih šuma (engl. *random forest*, *random forests*, *RF*) jedan je od algoritama iz područja strojnog učenja (engl. *machine learning*) koji se koristi za raspoznavanje uzoraka i regresijsku analizu. Autor ovog algoritma je profesor Leo Breiman sa američkog sveučilišta Berkely, a glavna značajka algoritma je da on pri raspoznavanju ne koristi samo jedan klasifikator nego kreira veći broj klasifikatora u obliku stabla odluke (engl. *decision trees*) od kojih svako sudjeluje u izglasavanju konačnog rezultata. Svako od spomenutih stabala trenira se na određenom broju uzoraka iz seta za treniranje dobivenih uzorkovanjem s ponavljanjem, pa je prisutna mogućnost višestrukog izbora istih uzoraka u skup za izgradnju pojedinog stabla [10].

4.1.2 Parametri i koraci algoritma slučajnih šuma

Algoritam slučajnih šuma obično koristi dva parametra: broj stabala koja će se generirati *n* i broj atributa iz kojeg će se svako stablo generirati *m*, a njihove optimalne vrijednosti ovise o broju uzoraka za učenje i broju atributa uzoraka. Često korištene okvirne vrijednosti parametra *m* su, ako broj atributa uzoraka označimo s M, definirane s $m = \log_2 M + 1$ ili $m = \sqrt{M}$, dok se za vrijednost parametra *n* preporuča vrijednost veća od 100. Korisno je pri tome imati na umu da povećanjem broja stabala *n* ne dolazi do pogoršanja performansi algoritma odnosno zasićenja, a u slučaju da je relativno velik broj vrijednosti atributa zašumljen, korisno je povećati broj atributa iz kojeg će se svako stablo raditi *m* [10].

Sam algoritam sastoji se od sljedećih koraka:

1. Neka su zadani broj stabala koji će se generirati *n*, broj atributa iz kojeg će se svako stablo raditi *m* i broj uzoraka u skupu za treniranje *N*.
2. Iz skupa za treniranje slučajnim odabirom s ponavljanjem (engl. *bootstrap sample*) izabrati *N* uzoraka koji će tvoriti skup za izgradnju stabla.
3. Izgradi stablo odluke na temelju skupa za izgradnju i to uzimajući pri izgradnji svakog čvora stabla u obzir *m* u tom koraku slučajno izabranih atributa

uzoraka. Od tih *m* slučajno izabranih atributa u čvoru će biti odabran onaj koji maksimizira informacijsku dobit.

4. Ponavljaj korake 2 i 3 dok nije izgenerirano *n* stabala.

4.1.3 Procjena pogreške i utjecanje na *oob* pogrešku

Budući da je kod odabira uzoraka od kojih ćemo generirati stablo u 2. koraku moguće ponavljanje, obično oko jedna trećina uzoraka originalnog skupa za treniranje *N* svaki put bude izostavljena iz skupa za generiranje stabla. Ti se uzorci nazivaju i *oob* uzorci (engl. *out-of-bag* ili skraćeno *oob*). Ova pojava rezultira jednim važnim svojstvom algoritma slučajnih šuma, a to je da nema potrebe za krosvalidacijom ili korištenjem posebnog skupa za testiranje kako bi se dobila procjena pogreške. Svaki uzorak iz originalnog skupa za treniranje *N* javlja se kao *oob* uzorak u otprilike jednoj trećini stabala šume i upravo na tim stablima zato ispitujemo njegovu klasifikaciju. Vjerojatnost pogrešne klasifikacije računa se kao broj stabala koja su pogrešno klasificirala *oob* uzorak podijeljen s ukupnim brojem stabala za koja je taj uzorak *oob*. Konačna procjena *oob* pogreške računa se kao prosječna vrijednost vjerojatnosti pogrešne klasifikacije između svih uzoraka u originalnom skupu za treniranje *N*. Pokazalo se kako je ovo zadovoljavajuća i neprostrana procjena pogreške šume [10].

Vrijednost *oob* pogreške dobivene šume ovisi o snazi odnosno uspješnosti svakog pojedinog stabla (što je veća snaga pojedinih stabala, to će pogreška biti manja) i o korelaciji između stabala šume (što je veća korelacija, to će i pogreška biti veća). Oba ova obilježja u izravnoj su vezi s parametrom *m*: povećanjem vrijednosti *m* raste i korelacija i snaga stabala i obrnuto. Optimalna vrijednost ovog parametra (ili češće relativno širok raspon optimalnih vrijednosti) može se utvrditi na temelju *oob* (engl. *out of bag*) pogreške [10].

4.1.4 Procjena važnosti atributa

Procjena važnosti atributa u algoritmu slučajnih šuma može se jednostavno odrediti, a vrlo je važna jer nam omogućava kvalitetan uvid u utjecaj atributa na klasifikaciju. Da bismo odredili važnost nekog atributa *A1*, najprije je potrebno svakim pojedinim stablom klasificirati njegove *oob* uzorke, a potom slučajno permutiramo vrijednosti atributa *A1* unutar *oob* uzoraka i ponovo klasificirati na taj način modificirane uzorke. Od zbroja točno klasificiranih originalnih *oob* uzoraka za svako stablo potom se oduzme zbroj točno klasificiranih *oob* uzoraka nakon permutiranja njihovih vrijednosti atributa *A1* i dobivena

razlika usrednjava brojem stabala u šumi. Dobivena vrijednost naziva se važnost atributa $A1$, a ako su neusrednjene vrijednosti važnosti nezavisne od stabla do stabla, tada se dijeljenjem sa standardnom pogreškom (koja se računa na uobičajen način) dobiva *z-skor*. Z-skor daje informaciju o udaljenosti neke slučajne varijable od njene srednje vrijednosti izraženo u jedinicama standardne devijacije [6].

Još jedna mjera koja se koristi za procjenu važnosti atributa je *Gini kriterij*. Općenito se kod stabala odluke koristi *funkcija nečistoća* (engl. *impurity function*) koja služi kao kriterij za odabir varijable grananja na temelju omjera p uzoraka za učenje koji su u pojedinim klasama. Jedna od funkcija koje se najčešće koriste je Gini kriterij raznolikosti koji se definira kao: $\varphi(p) = 2 \cdot p \cdot (1-p)$. Nakon svakog grananja po varijabli m, Gini kriterij za dva nasljedna čvora manji je nego za čvor roditelj. Dodajući Gini smanjenje za svaku individualnu varijablu preko svih stabla u šumi dobiva se brza procjena važnosti varijable koja je često u skladu s mjerom permutirane važnosti iz prethodnog odlomka [6].

4.1.5 **Nedostajuće vrijednosti u skupu za treniranje**

Algoritam slučajnih šuma predviđa dva moguća načina za nadoknađivanje nedostajućih vrijednosti. Prvi način nadoknađivanja je da ukoliko je atribut numerički, tada na mjesto nedostajuće vrijednosti zapišemo srednji vrijednost tog atributa svih uzoraka iz iste klase kao i on [10]. Ukoliko vrijednost atributa nije numerička, tada na mjesto nedostajuće vrijednosti zapišemo najčešću vrijednost tog atributa kod uzoraka koji su iz iste klase kao i taj uzorak. Ovakav način nadoknađivanja brz je i jednostavan.

Drugi način nadoknađivanja nedostajućih vrijednosti računarski je složeniji, ali zato daje bolje rezultate čak i za veliki broj nedostajućih vrijednosti [10]. Nedostajuće vrijednosti najprije se grubo aproksimiraju, a potom se računa matrica sličnosti (engl. *proximity matrix*). Matrica sličnosti dimenzije je $N \times N$ i na početku se inicijalizira na nulu, a potom korigira nakon svakog generiranog stabla tako da stablom klasificiramo cijeli skup za testiranje (i skup za generiranje stabla i *oob* uzorke) i za svaka dva uzorka koji unutar stabla završe u istom čvoru njihovu odgovarajuću vrijednost u matrici sličnosti povećamo za jedan. Primjerice, ako u k-tom stablu uzorci a i b nakon propuštanja kroz stablo završe u istom čvoru, tada će se element matrice s indeksima [a,b] povećati za 1. Nakon što smo izgenerirali sva stabla i matricu sličnosti, nju normaliziramo tako da sve njene elemente podijelimo s brojem stabala. Nedostajuća vrijednost potom se određuje tako da se vrijednosti odgovarajućeg atributa kod svih ostalih uzoraka kojima je njegova vrijednost

definirana pomnoži s odgovarajućim koeficijentom iz matrice sličnosti te izračuna prosjek. Primjerice, ako za uzorak *x* nije poznata vrijednost atributa *A1* i nju označimo kao *x.A1*, a atribut A1 ima poznatu vrijednost samo u uzorcima *a*, *b* i *c* i to su vrijednosti *a.A1*, *b.A1*, *c.A1*, tada ćemo te tri vrijednosti pomnožiti koeficijentima iz matrice sličnosti s indeksima redom [*a*,*x*], [*b*,*x*] i [*c*,*x*], te na kraju zbrojiti i usrednjiti dijeljenjem s tri.

4.2 Mjere uspješnosti klasifikacije

4.2.1 Tablica zabune i numeričko ocjenjivanje uspješnosti klasifikatora

Za donošenje utemeljenih i potpunih zaključaka o uspješnosti klasifikatora potrebno je raspolagati informacijama o nekoliko često korištenih mjera njegovih rezultata. Klasifikator konstruiran u ovom radu ima zadatak predviđanja kojoj od dvije moguće klase pripada promatrani uzorak: klasi 0 ako središnji ostatak u tom uzorku nije mjesto proteinske interakcije ili klasi 1 ako je. Ovakav klasifikator obično se naziva i binarni klasifikator. Uzorke klase 0 možemo nazvati i negativno klasificiranim uzorcima, a uzorke klase 1 pozitivno klasificiranim uzorcima. Nakon što je klasifikator konstruiran i naučen na nekom skupu, testira se na drugom skupu još neviđenih podataka. Rezultat klasifikacije testnog skupa može se prikazati u obliku tablice zabune (engl. *confusion matrix*), koja se ponekad naziva i matrica greške (engl. *error matrix*) ili kontingencijska matrica u statistici (engl. *contingency matrix*). Ova tablica prikazana je na slici 4.1.

Predviđena klasa \ Stvarna klasa	0	1
0	*TN*	*FN*
1	*FP*	*TP*

Slika 4.1 Tablica zabune

U matrici su uzorci kojima je klasifikator pridijelio klasu 0 raspodijeljeni u dvije skupine: prvu, označenu s *TN* (engl. *true negatives*) u kojoj su negativno klasificirani uzorci koji ujedno i pripadaju klasi 0, te drugu, označenu s *FN* (engl. *false negatives*) u kojoj su negativno klasificirani uzorci koji zapravo pripadaju klasi 1. Analogno tome podijeljeni su

i uzorci kojima je klasifikator pridijelio klasu 1: prvu skupinu takvih uzoraka, označenu s *FP*, čine pozitivno klasificirani uzorci koji zapravo pripadaju klasi 0, a drugu skupinu, označenu s *TP*, čine pozitivno klasificirani uzorci koji zaista pripadaju klasi 1. Očito je da bi kod dobrog klasifikatora vrijednosti na glavnoj dijagonali ove tablice, *TP* i *TN*, trebale biti što veće, a vrijednosti *FP* i *FN* što manje.

Za preciznije vrednovanje karakteristika klasifikatora uvode se sljedeće mjere koje se temelje na prethodno opisanim skupinama uzoraka:

1. Točnost (engl. *accuracy*), koja predstavlja udio točno klasificiranih primjera u svim primjerima.
   $tocnost = (TP + TN)/(TP + TN + FP + FN)$

2. Preciznost (engl. *precision* ili *positive predictive value*), koja predstavlja udio točno klasificiranih primjera u skupu svih pozitivno klasificiranih primjera.
   $preciznost = TP/(TP + FP)$

3. Odziv (engl. *recall*, *hit rate* ili *true positive rate*), koji predstavlja udio točno klasificiranih primjera u skupu svih pozitivnih primjera.
   $odziv = TP/(TP + FN)$

4. Specifičnost (engl. *specificity* ili *true negative rate*), koja predstavlja udio točno klasificiranih primjera u skupu svih negativnih primjera.
   $specificnost = TN/(TN + FP)$

5. *F-mjera*, koja predstavlja harmonijsku sredinu preciznosti i odziva.
   $$F - mjera = \frac{2 \cdot preciznost \cdot odziv}{preciznost + odziv}$$

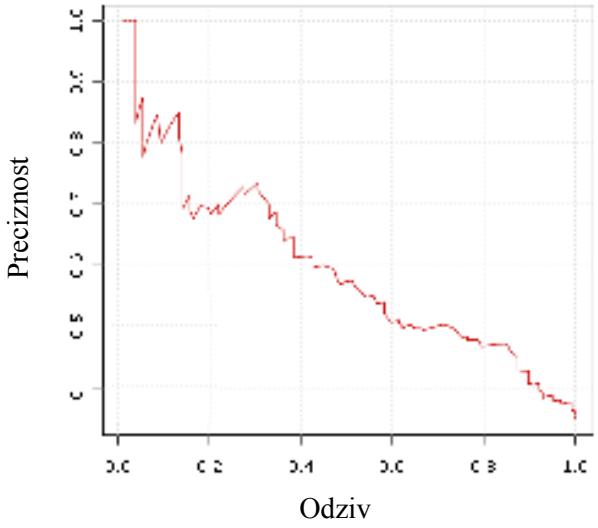
Niti jedna od navedenih mjera nije dovoljna sama za sebe nego se obično koriste kombinirano [13].

4.2.2 **Grafičko ocjenjivanje uspješnosti klasifikatora**

4.2.2.1 Krivulja preciznost-odziv

Često se uspješnost klasifikatora ocjenjuje i pomoću grafa *preciznost-odziv* (engl. *precision-recall curve*, *PR*), koji na *x*-osi sadrži vrijednost odziva, a na *y*-osi preciznosti. Na odnos ovih dvaju mjera moguće je utjecati pomoću vrijednosti *FN*, *FP* i *TP*. Naime,

ukoliko olakšamo klasifikaciju nepoznatog uzorka u klasu 1, iznos *FN* će se smanjiti, što će dovesti do povećanja odziva (ali često i do smanjenja preciznosti) [13]. Obrnuto, ukoliko olakšamo klasifikaciju nepoznatog uzorka u klasu 0, iznos *FP* će se smanjiti, odnosno povećati će se preciznost (često na uštrb odziva). Olakšavanje klasifikacije nepoznatog uzorka u klasu 1 ili 0 u ovom slučaju značilo bi korekciju praga ili modificiranje dovoljnog broja stabala šume koja trebaju dati svoj glas za određenu klasu na manje ili više od pola ukupnog broja stabla šume. Idealno, vrijednosti *FP* i *FN* trebale bi biti što manje. Primjer grafa preciznost-odziv generiranog u programskom jeziku R dan je na slici 4.2. Graf prikazuje kako povećanjem odziva dolazi do smanjenja preciznosti. U idealnom slučaju, graf *PR* krivulje bi težio gornjem desnom kutu, gdje su vrijednosti i odziva i preciznosti visoke [6].

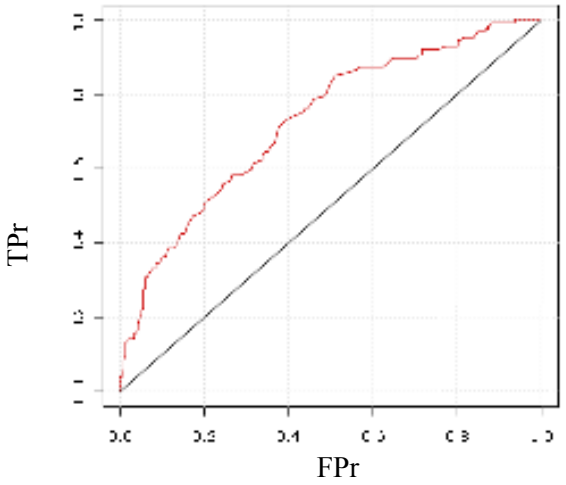


Slika 4.2 Graf preciznost-odziv

4.2.2.2 *ROC* krivulja

Ocjenu performansi binarnih klasifikatora pomoću ROC (engl. *receiver operator characteristic*) krivulje uveli su F. Provost i T. Fawcett 1997. Za iscrtavanje *ROC* krivulje potrebno je na *x*-osi naznačiti vrijednosti omjera broja *FP* uzoraka i zbroja *FP* + *FN* (engl. *false positive rate*), dok se na *y*-osi naznačuje omjer broja *TP* uzoraka i zbroja *TP* + *FN* (engl. *true positive rate*) za različite vrijednosti praga. U idealnom slučaju, graf *ROC* krivulje težio bi gornjem lijevom kutu [12], gdje je broj *FN* i *FP* uzoraka malen. S druge strane, ROC krivulja nekog klasifikatora trebala bi biti ograničena pravcem $y=x$ koji predstavlja performanse klasifikatora koji nepoznati uzorak klasificira u bilo koju od dvije klase s vjerojatnošću 0.5. Uobičajeno je također izračunati površinu ispod ROC krivulje i

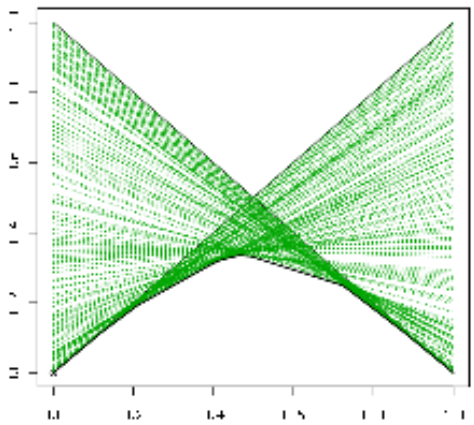
ona se naziva *AUC* (engl. *area under ROC curve*). S obzirom na to da se krivulja obično prikazuje u jediničnom kvadratu, AUC površina mora biti manja od 1 i također ne bi smjela biti manja od 0.5. Može se reći i da je klasifikator to bolji što je površina ispod *ROC* krivulje veća. Pri usporedbi dvije *ROC* krivulje, može se reći kako dominira onaj klasifikator čiji je graf viši, a ako im se *ROC* krivulje sijeku, tada je u nekim okolnostima bolji jedan, a u drugima drugi klasifikator. ROC graf za klasifikator iz prethodnog potpoglavlja prikazan je na slici 4.3.



Slika 4.3 *ROC* krivulja

4.2.2.3 Krivulja cijene

Krivulja cijene (engl. *cost curve*) jedan je od pokušaja nadomještanja nedostataka *ROC* krivulje koji su uveli C. Drummond i R. C. Holte. Matematički, krivulja cijene predstavlja *ROC* krivulju transformiranu linearnom transformacijom u novi prostor, pri čemu svakoj točki *ROC* krivulje odgovara jedan pravac krivulje cijene. Primjerice, uz pretpostavku kako je cijena pogreške u klasifikaciji pozitivnog i negativnog uzorka ista, točka (*TP*, *FP*) iz ROC prostora preslikava se u pravac koji prolazi kroz točke (0, *FP*) i (1, 1-*TP*), a zbog linearnog odnosa moguće je dakako i obrnuto preslikavanje. Krivulja cijene u osnovi se orijentira na cijenu odnosno trošak proizašao iz pogrešne klasifikacije uzorka. Kod krivulje cijene može se reći, suprotno nego kod *ROC* krivulje, da dominira onaj klasifikator, čija je krivulja cijene najniža [11]. Graf krivulje cijene za klasifikator iz prethodnog potpoglavlja prikazan je na slici 4.4.



Slika 4.4 Graf funkcije cijene vjerojatnosti

Cijena pogrešne klasifikacije u ovom radu nije od primarnog interesa pa se stoga za prezentaciju i analizu performansi klasifikatora neće koristiti krivulje cijene. Između *ROC* i krivulja preciznost-odziv za prezentaciju i analizu performansi klasifikatora odabrane su krivulje preciznost-odziv zbog lakšeg i potpunijeg vizualnog iščitavanja rezultata.

5. Rezultati

5.1 Analiza svojstava skupa

5.1.1 Kvantitativna i kvalitativna analiza skupa

U ovom radu korištena su u osnovi tri skupa proteinskih lanaca, od kojih su kasnije iz svakog izvedene četiri podvarijante, kako je opisano u poglavlju 3.2.3. Prvi od ovih temeljnih skupova sastoji se samo od lanaca ih hetero multimerskih proteina, drugi od homo multimerskih proteina, a treći sadrži lance i iz homo i iz hetero multimerskih struktura. Sastav ova tri skupa iskazan je u tablici 5.1.

Tablica 5.1 Sastav skupova

	Skup 1	Skup 2	Skup 3
Opis:	Samo lanci iz hetero multimera	Samo lanci iz homo multimera	Lanci iz homo i hetero multimera
Broj sadržanih lanaca:	2760	4280	6574
Broj obuhvaćenih PDB struktura:	1492	4277	1496

Niti jedan od navedena tri skupa ne sadrži redundanciju u vidu primarne strukture veću od 30%, a PDB strukture iz kojih sadržani lanci dolaze imaju bar dva lanca duljine barem 30 aminokiselinskih ostataka. Treći skup iz Tablice 5.1 kreiran je iz istih PDB struktura kao i prethodna dva skupa. Ipak, to ne znači da će svi njegovi lanci biti sadržani u nekom od prethodnih skupova. Hoće li neki lanac iz trećeg skupa biti sadržan u nekom od prva dva skupa ili ne, ovisi o načinu odabira lanca koji reprezentira svoj klaster. U radu [5] kojim su dobiveni ovi skupovi za lanac koji reprezentira klaster odabiran je prvi mogući iz liste lanaca koji čine taj klaster.

5.1.2 Statistička analiza interakcija u skupu

Prvo statističko ispitivanje interakcija provedeno nad dobivenim skupovima analizira učestalost pojavljivanja pojedinih tipova aminokiselinskih ostataka općenito u skupu, te u interakcijama. U tablicama 5.2, 5.3 i 5.4 prikazani su rezultati ovog ispitivanja za sva tri

skupa redom, uz korištenje prve definicije interakcije unutar PSAIA alata, odnosno maksimalne udaljenosti teških atoma sa lanaca u interakciji. Vrijednosti u stupcima prikazuju redom koliko se puta pojedini tip aminokiselinskog ostatka javlja u skupu, zatim koliko se puta javlja u svim interakcijama u kojima sudjeluje taj lanac, zatim omjer broja pojavljivanja u interakcijama i pojavljivanja u skupu i na kraju omjer broja pojavljivanja u interakcijama i ukupnog broja pojavljivanja svih tipova aminokiselinskih ostataka u interakcijama.

Tablica 5.2 Interakcije maksimalne udaljenosti u prvom (hetero) skupu

	Ukupno u skupu	Ukupno u interakcijama	Zastupljenost u skupu	Zastupljenost u interakcijama
ALA	49630	11986	0.0855	0.0742
ARG	30704	10692	0.0529	0.0662
ASN	24290	7094	0.0418	0.0439
ASP	31664	8799	0.0545	0.0545
CYS	9304	2342	0.0160	0.0145
GLN	22699	6923	0.0391	0.0429
GLU	40565	11424	0.0699	0.0708
GLY	39608	9882	0.0682	0.0612
HIS	12646	4025	0.0218	0.0249
ILE	33396	8551	0.0575	0.0530
LEU	55691	14873	0.0959	0.0921
LYS	37557	10260	0.0647	0.0636
MET	12138	3701	0.0209	0.0229
PHE	22897	6805	0.0394	0.0422
PRO	25662	7217	0.0442	0.0447
SER	34469	9589	0.0594	0.0594
THR	30800	8664	0.0531	0.0537
TRP	7538	2382	0.0130	0.0148
TYR	19506	6506	0.0336	0.0403
VAL	39761	9721	0.0685	0.0602

Tablica 5.3 Interakcije maksimalne udaljenosti u drugom (homo) skupu

	Ukupno u skupu	Ukupno u interakcijama	Zastupljenost u skupu	Zastupljenost u interakcijama
ALA	84924	17540	0.0842	0.0805
ARG	52753	14204	0.0523	0.0652
ASN	40741	9183	0.0404	0.0421
ASP	58423	12359	0.0579	0.0567
CYS	12159	2259	0.0121	0.0104
GLN	35914	8732	0.0356	0.0401
GLU	71932	15349	0.0713	0.0704
GLY	73593	14557	0.0730	0.0668
HIS	24370	6156	0.0242	0.0282
ILE	61133	11699	0.0606	0.0537
LEU	97989	20175	0.0972	0.0925
LYS	57130	12400	0.0566	0.0569
MET	16808	3969	0.0167	0.0182
PHE	41540	9464	0.0412	0.0434
PRO	45736	9920	0.0453	0.0455
SER	57887	12644	0.0574	0.0580
THR	53763	11711	0.0533	0.0537
TRP	13052	3167	0.0129	0.0145
TYR	34765	8845	0.0345	0.0406
VAL	73933	13677	0.0733	0.0627

Tablica 5.4 Interakcije maksimalne udaljenosti u trećem (homo i hetero) skupu

	Ukupno u skupu	Ukupno u interakcijama	Zastupljenost u skupu	Zastupljenost u interakcijama
ALA	119810	27937	0.0879	0.0833
ARG	77834	23501	0.0571	0.0701
ASN	60393	15318	0.0443	0.0457
ASP	83821	19897	0.0615	0.0593
CYS	19859	4309	0.0146	0.0128
GLN	54562	14731	0.0400	0.0439
GLU	104689	25282	0.0768	0.0754
GLY	103501	23000	0.0759	0.0686
HIS	34559	9641	0.0254	0.0288
ILE	87888	19107	0.0645	0.0570
LEU	143174	33008	0.1051	0.0984
LYS	85414	21256	0.0627	0.0634
MET	26643	7167	0.0196	0.0214
PHE	59993	15330	0.0440	0.0457
PRO	66448	16146	0.0488	0.0481
SER	85855	20840	0.0630	0.0621
THR	78636	19182	0.0577	0.0572
TRP	19276	5226	0.0141	0.0156
TYR	50446	14458	0.0370	0.0431
VAL	105695	22036	0.0776	0.0657

Iz navedenih podataka moguće je uočiti neka osnovna svojstva skupa. Među najzastupljenijim aminokiselinskim ostatcima u sva tri skupa nalaze se leucin, arginin, alanin i glicin, dok se u sva tri skupa najrjeđe javljaju metionin, cistein i triptofan. Isti aminokiselinski ostatci također su najzastupljeniji i u interakcijama. U nastavku su u tablicama 5.5, 5.6 i 5.7 prikazani isti rezultati, ali u slučaju da se kao definicija mjesta kontakta koristi PIADA algoritam PSAIA alata.

Tablica 5.5 PIADA interakcije u prvom (hetero) skupu

	Ukupno u skupu	Ukupno u interakcijama	Zastupljenost u skupu	Zastupljenost u interakcijama
ALA	43447	9727	0.0819	0.0737
ARG	30704	10573	0.0579	0.0801
ASN	24290	6088	0.0458	0.0461
ASP	31664	7630	0.0597	0.0578
CYS	9305	1895	0.0175	0.0144
GLN	22699	5783	0.0428	0.0438
GLU	40566	9656	0.0765	0.0732
GLY	37851	8317	0.0714	0.0630
HIS	12646	3687	0.0238	0.0279
ILE	33396	7394	0.0630	0.0560
LEU	55690	12857	0.1050	0.0974
LYS	35098	9577	0.0662	0.0726
MET	12138	3319	0.0229	0.0252
PHE	22897	6260	0.0432	0.0474
PRO	25662	5987	0.0484	0.0454
SER	34469	7789	0.0650	0.0590
THR	30800	7260	0.0581	0.0550
TRP	7538	2297	0.0142	0.0174
TYR	19506	5857	0.0368	0.0444
VAL	39761	8286	0.0750	0.0628

Tablica 5.6 PIADA interakcije u drugom (homo) skupu

	Ukupno u skupu	Ukupno u interakcijama	Zastupljenost u skupu	Zastupljenost u interakcijama
ALA	84924	13725	0.0909	0.0799
ARG	52753	13051	0.0564	0.0759
ASN	40741	7387	0.0436	0.0430
ASP	58423	10782	0.0625	0.0627
CYS	12159	1809	0.0130	0.0105
GLN	35914	7181	0.0384	0.0418
GLU	71932	13134	0.0770	0.0764
GLY	73593	11679	0.0787	0.0680
HIS	24370	5500	0.0261	0.0320
ILE	61133	9848	0.0654	0.0573
LEU	97989	17084	0.1048	0.0994
LYS	57130	10589	0.0611	0.0616
MET	16808	3426	0.0180	0.0199
PHE	41540	8427	0.0444	0.0490
PRO	45736	8252	0.0489	0.0480
SER	57887	9840	0.0619	0.0573
THR	53763	9392	0.0575	0.0546
TRP	13052	2922	0.0140	0.0170
TYR	34765	7838	0.0372	0.0456
VAL	73933	11320	0.0791	0.0659

Tablica 5.7 PIADA interakcije u trećem (homo i hetero) skupu

	Ukupno u skupu	Ukupno u interakcijama	Zastupljenost u skupu	Zastupljenost u interakcijama
ALA	119810	22173	0.0879	0.0775
ARG	77834	22314	0.0571	0.0780
ASN	60393	12666	0.0443	0.0443
ASP	83821	17285	0.0615	0.0605
CYS	19860	3464	0.0146	0.0121
GLN	54562	12180	0.0400	0.0426
GLU	104689	21531	0.0768	0.0753
GLY	103501	18756	0.0759	0.0656
HIS	34559	8685	0.0254	0.0304
ILE	87888	16299	0.0645	0.0570
LEU	143174	28241	0.1051	0.0988
LYS	85414	18844	0.0627	0.0659
MET	26643	6311	0.0196	0.0221
PHE	59993	13814	0.0440	0.0483
PRO	66448	13418	0.0488	0.0469
SER	85855	16495	0.0630	0.0577
THR	78636	15656	0.0577	0.0548
TRP	19276	4906	0.0141	0.0172
TYR	50446	12896	0.0370	0.0451
VAL	105695	18455	0.0776	0.0645

Iz tablica se vidi kako su najzastupljeniji aminokiselinski ostatci slični kako i u prethodnom slučaju. Radi preglednosti, u sljedećim tablicama 5.8 i 5.9 izdvojeni su podaci o relativnoj učestalosti pojavljivanja aminokiselinskih ostataka poredani po padajućem redoslijedu i to najprije za definiciju mjesta interakcije preko maksimalne udaljenosti, a zatim preko PIADA algoritma.

Tablica 5.8 Relativna učestalost pojavljivanja u interakcijama za maksimalnu udaljenost

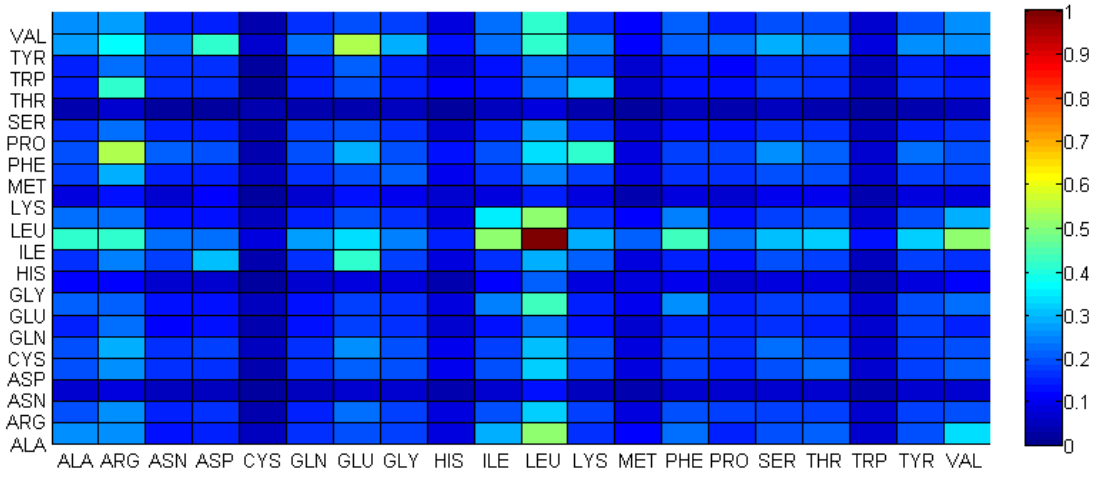
	Hetero		Homo		Homo i hetero
LEU	0.0921	LEU	0.0925	LEU	0.0984
ALA	0.0742	ALA	0.0805	ALA	0.0833
GLU	0.0708	GLU	0.0704	GLU	0.0754
ARG	.0662	GLY	0.0668	ARG	0.0701
LYS	0.0636	ARG	0.0652	GLY	0.0686
GLY	0.0612	VAL	0.0627	VAL	0.0657
VAL	0.0602	SER	0.0580	LYS	0.0634
SER	0.0594	LYS	0.0569	SER	0.0621
ASP	0.0545	ASP	0.0567	ASP	0.0593
THR	0.0537	ILE	0.0537	THR	0.0572
ILE	0.0530	THR	0.0537	ILE	0.0570
PRO	0.0447	PRO	0.0455	PRO	0.0481
ASN	0.0439	PHE	0.0434	ASN	0.0457
GLN	0.0429	ASN	0.0421	PHE	0.0457
PHE	0.0422	TYR	0.0406	GLN	0.0439
TYR	0.0403	GLN	0.0401	TYR	0.0431
HIS	0.0249	HIS	0.0282	HIS	0.0288
MET	0.0229	MET	0.0182	MET	0.0214
TRP	0.0148	TRP	0.0145	TRP	0.0156
CYS	0.0145	CYS	0.0104	CYS	0.0128

Tablica 5.9 Relativna učestalost pojavljivanja u interakcijama za PIADA algoritam

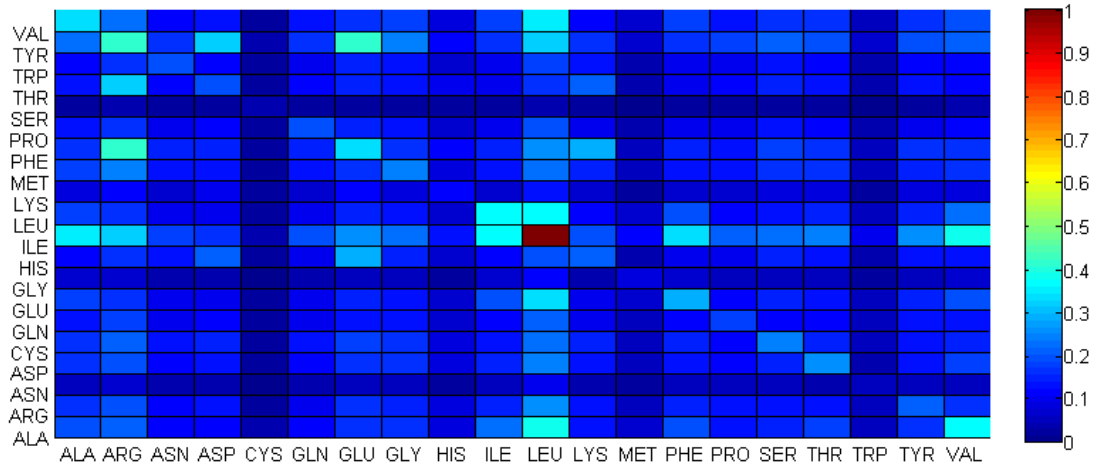
	Hetero		Homo		Homo i hetero
LEU	0.0974	LEU	0.0994	LEU	0.0988
ARG	0.0801	ALA	0.0799	ARG	0.0780
ALA	0.0737	GLU	0.0764	ALA	0.0775
GLU	0.0732	ARG	0.0759	GLU	0.0753
LYS	0.0726	GLY	0.0680	LYS	0.0659
GLY	0.0630	VAL	0.0659	GLY	0.0656
VAL	0.0628	ASP	0.0627	VAL	0.0645
SER	0.0590	LYS	0.0616	ASP	0.0605
ASP	0.0578	ILE	0.0573	SER	0.0577
ILE	0.0560	SER	0.0573	ILE	0.0570
THR	0.0550	THR	0.0546	THR	0.0548
PHE	0.0474	PHE	0.0490	PHE	0.0483
ASN	0.0461	PRO	0.0480	PRO	0.0469
PRO	0.0454	TYR	0.0456	TYR	0.0451
TYR	0.0444	ASN	0.0430	ASN	0.0443
GLN	0.0438	GLN	0.0418	GLN	0.0426
HIS	0.0279	HIS	0.0320	HIS	0.0304
MET	0.0252	MET	0.0199	MET	0.0221
TRP	0.0174	TRP	0.0170	TRP	0.0172
CYS	0.0144	CYS	0.0105	CYS	0.0121

Prethodne tablice ukazuju ne samo na veliku sličnost vrijednosti statistika za sva tri skupa uz istu definiciju mjesta interakcije i glede sastava skupa i interakcija u skupu, nego čak i uz različitu definiciju kriterija za mjesto interakcije. Ako je mjesto interakcije definirano preko maksimalne udaljenosti teških atoma, pozicije odgovarajućih aminokiselinskih ostataka razlikuju se najviše za tri mjesta i to samo u jednom slučaju (LYS). U slučaju PIADA algoritma ta razlika iznosi najviše dva. Moguće je još primijetiti kako su najzastupljeniji aminokiselinski ostatci u interakcijama (LEU) i četiri najmanje zastupljena (HIS, MET, TRP, CYS) aminokiselinska ostatka prisutni na istim rangiranim pozicijama unutar svih skupova i za obije definicije mjesta interakcije.

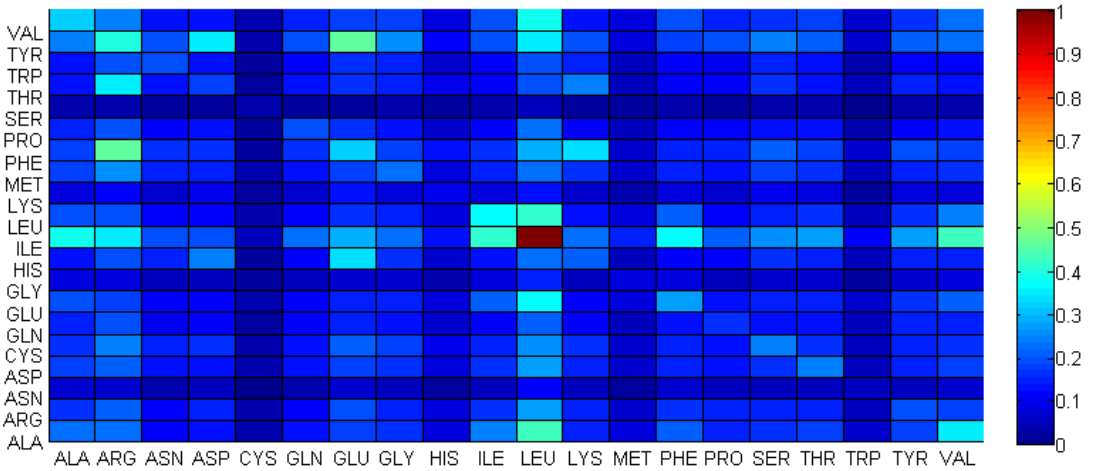
Sljedećih 6 slika (5.1-5.6) predstavlja zastupljenost interakcije pojedinog aminokiselinskog ostatka sa svakim od 20 standardnih. Ovi grafovi izraženi su ne numerički nego pomoću boja, kako bi usporedba i čitanje bili lakši. Vrijednosti grafova prethodno su radi efikasnije usporedbe skalirane na interval [0,1].



Slika 5.1 Interakcije za maksimalnu udaljenost i prvi (hetero) skup



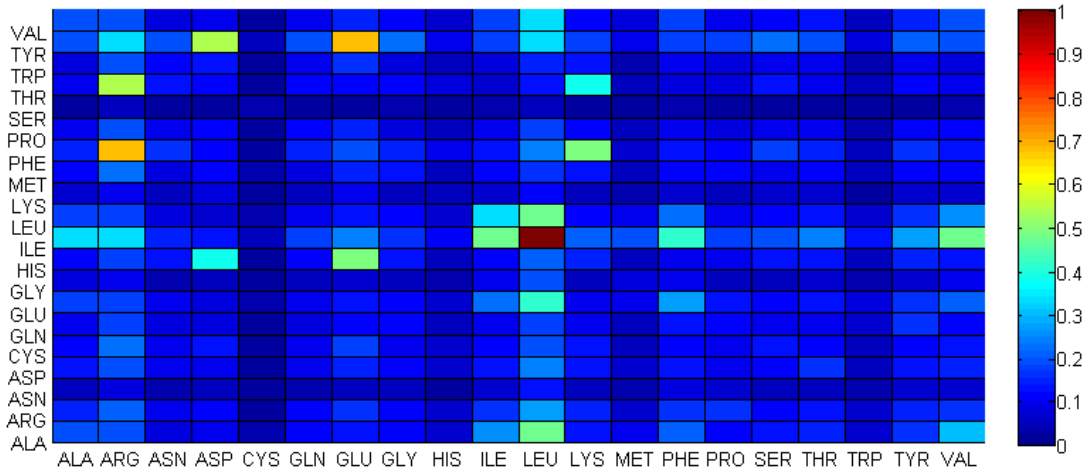
Slika 5.2 Interakcije za maksimalnu udaljenost i drugi (homo) skup



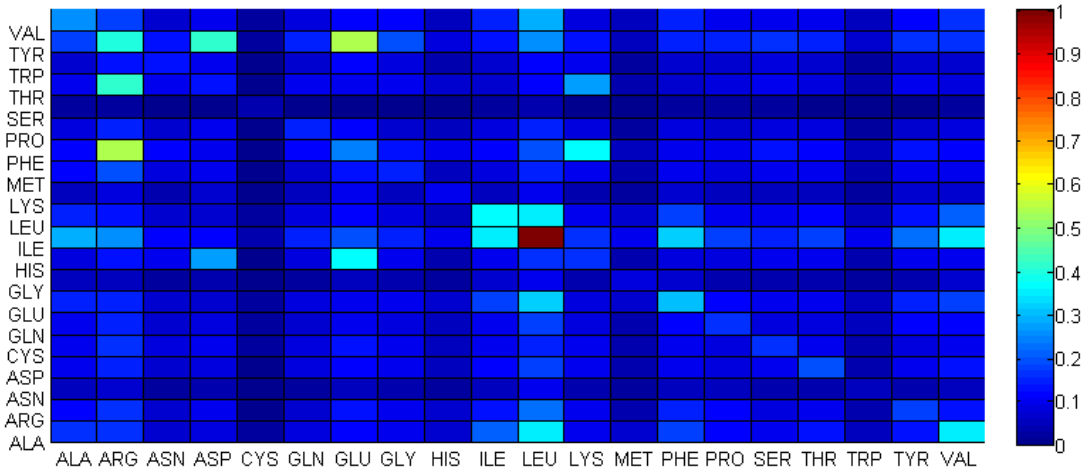
Slika 5.3: Interakcije za maksimalnu udaljenost i treći (homo i hetero) skup

Prikazana tri grafa za sva tri skupa u slučaju definiranja mjesta interakcije preko maksimalne udaljenosti imaju dosta sličnosti i na sva tri su prepoznatljive karakteristične linije, ali u različitim nijansama. Najviše razlika ima prvi graf koji predstavlja hetero skup,

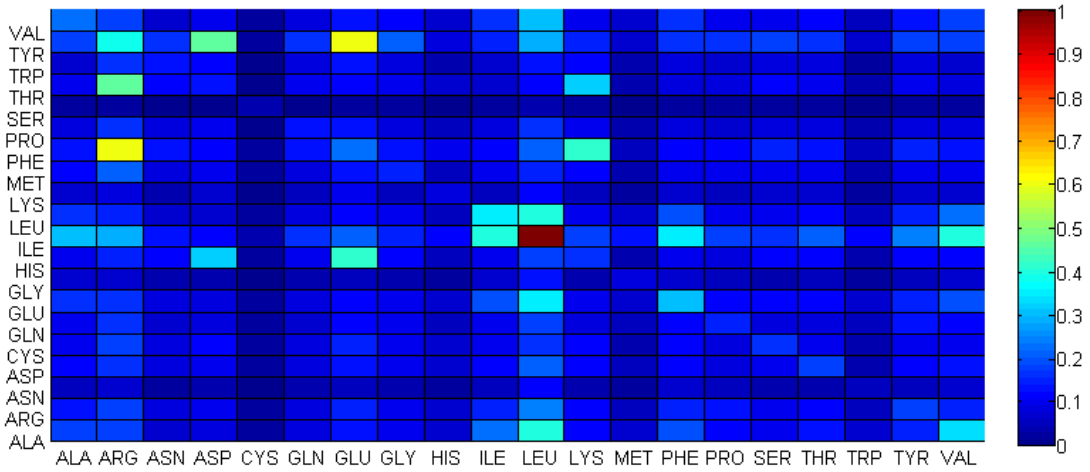
ali samo u vidu različitih nijansi, što bi moglo biti posljedica manjeg broja podataka nego kod ostalih skupova.



Slika 5.4 Interakcije za PIADA algoritam i prvi (hetero) skup



Slika 5.5 Interakcije za PIADA algoritam i drugi (homo) skup



Slika 5.6 Interakcije za PIADA algoritam i treći (homo i hetero) skup

Sa svih šest prethodnih slika postojeći trendovi u podatcima su sada nešto očitiji nego samo iz brojčanih vrijednosti. Isti glavni uzorci primjetni su na svih šest grafova: tamno plave horizontalne i vertikalne linije na mjestima prethodno navedenih najmanje aktivnih aminokiselinskih ostataka, svjetliji križ koji predstavlja dvije linije leucina i blago istaknuta dijagonala. Istaknuta polja na grafovima uglavnom su vrlo slična i razlikuju se tek u nijansama, što je još jedna potvrda ne samo sličnosti između samih skupova nego i definicija mjesta interakcije.

Iduća statistika predstavlja analizu sličnu prethodnoj, samo ne promatra pojedine interakcije nego dva susjedna aminokiselinska ostatka s jednog lanca koji su u interakciji s dva susjedna aminokiselinska ostatka s drugog lanca. Kombinacije s najviše ponavljanja prikazane su u tablici 5.10 za definiranje interakcije pomoću PIADA algoritma i u tablici 5.11 za definiranje interakcije maksimalnom udaljenošću.

Tablica 5.10 Dvojke aminokiselinskih ostataka u interakciji za PIADA algoritam

Hetero					Homo					Homo i hetero				
LEU	LYS	LYS	LEU	76	LEU	LEU	LEU	LEU	119	LEU	LEU	LEU	LEU	167
LEU	GLU	GLU	LEU	69	LEU	VAL	VAL	LEU	109	LEU	GLU	GLU	LEU	154
LEU	LEU	LEU	LEU	69	LEU	ALA	ALA	LEU	106	LEU	ALA	ALA	LEU	142
LEU	LEU	ARG	GLU	57	GLU	LEU	LEU	GLU	104	LEU	LYS	LYS	LEU	137
GLU	ARG	LEU	LEU	55	ARG	GLU	GLU	ARG	96	LEU	VAL	VAL	LEU	133
GLY	PRO	PRO	GLY	55	LEU	THR	THR	LEU	91	ARG	GLU	GLU	ARG	125
GLY	PRO	PRO	PRO	54	LEU	LEU	VAL	VAL	80	LEU	ARG	ARG	LEU	114
LEU	THR	LEU	LEU	52	LEU	LEU	GLU	ARG	74	ARG	GLU	LEU	LEU	111
LEU	ARG	ARG	LEU	48	LEU	ILE	ILE	LEU	73	LEU	LEU	GLU	ARG	110
LEU	ALA	ALA	LEU	47	LEU	SER	SER	LEU	73	LEU	LEU	VAL	VAL	110
LEU	LEU	LEU	VAL	47	LEU	ARG	ARG	LEU	72	LEU	THR	THR	LEU	108
GLN	LEU	LEU	GLN	42	LEU	LEU	THR	THR	72	LEU	LEU	ARG	GLU	100
ARG	GLU	GLU	ARG	39	LEU	LYS	LYS	LEU	72	GLU	ARG	LEU	LEU	99
LEU	VAL	ARG	GLU	38	ALA	GLY	GLY	ALA	69	LEU	SER	SER	LEU	99

Tablica 5.11 Dvojke aminokiselinskih ostataka u interakciji za maksimalnu udaljenost

Hetero					Homo					Homo i hetero				
GLY	PRO	PRO	PRO	136	LEU	LEU	LEU	LEU	235	LEU	LEU	LEU	LEU	311
PRO	PRO	GLY	GLY	113	ALA	LEU	LEU	ALA	228	ALA	LEU	LEU	ALA	298
GLY	PRO	PRO	GLY	110	LEU	GLY	GLY	LEU	218	LEU	VAL	VAL	LEU	257
LEU	LEU	LEU	LEU	105	GLY	ALA	ALA	GLY	215	LEU	SER	SER	LEU	256
GLU	LEU	LEU	GLU	97	LEU	VAL	VAL	LEU	209	ALA	GLY	GLY	ALA	248
LEU	ALA	ALA	LEU	92	LEU	THR	THR	LEU	202	GLY	LEU	LEU	GLY	247
LYS	LEU	LEU	LYS	92	LEU	SER	SER	LEU	200	LEU	THR	THR	LEU	236
LEU	THR	LEU	LEU	89	LEU	LEU	THR	THR	166	LEU	GLU	GLU	LEU	234
LEU	LEU	ARG	GLU	80	LEU	GLU	GLU	LEU	162	LEU	LYS	LYS	LEU	207
VAL	LEU	LEU	LEU	75	LEU	LEU	ALA	ALA	161	LEU	LEU	VAL	VAL	199
LEU	SER	SER	LEU	74	GLY	VAL	VAL	GLY	160	LEU	LEU	THR	THR	192
GLY	SER	GLY	GLY	72	LEU	LEU	VAL	VAL	150	LEU	LEU	ALA	ALA	191
LEU	LEU	LEU	VAL	72	ALA	GLY	ALA	ALA	147	ARG	GLU	LEU	LEU	188
LEU	VAL	VAL	LEU	69	PRO	LEU	LEU	PRO	145	GLY	SER	SER	GLY	187

U prethodnim tablicama kombinacije koje se ponavljaju unutar tablice označene su istom bojom. Pritom, zapis u obliku primjerice ALA ARG GLY SER označava kako su susjedni aminokiselinski ostatci s jednog lanca ALA i GLY redom u kontaktu s susjednim aminokiselinskim ostatcima ARG i SER. Kombinacije koje se ponavljaju u obije tablice također su označene istom bojom. Moguće je primijetiti kako je, pogotovo s obzirom na mogući broj kombinacija, nezanemariv broj istih četvorki prisutan u sva tri skupa. Ova situacija izraženija je u podacima čija su mjesta interakcija određena pomoću PIADA algoritma.

Posljednja provedena statistika je prebrojavanje različitih tipova interakcija između prozora koji su definirani kao neprekinuti slijed od devet aminokiselinskih ostataka dvaju lanaca, takvih da su središnji aminokiselinski ostaci u interakciji. Za takve prozore definira se *niže područje* kao četiri aminokiselinska ostatka koji prethode središnjem i *više područje* kao četiri aminokiselinska ostatka koji dolaze nakon središnjeg. Prebrojavanjem interakcija iz nižeg područja u više, iz višeg područja u više, iz višeg područja u niže i iz nižeg područja u niže obuhvaćenim dvama prozorima kojima su središnji aminokiselinski ostaci u interakciji dobiveni su rezultati prikazani u tablicama 5.13 i 5.14 za sva tri skupa, najprije za definiciju mjesta interakcije maksimalnom udaljenošću, a potom PIADA algoritmom. Navedeni tipovi interakcija označeni su redom s *NV*, *VV*, *VN*, *NN*. Također prebrojane su i sve interakcije u više područje (*V*), sve interakcije u niže područje (*N*) i sve interakcije prema središnjem aminokiselinskom ostatku (*S*).

Tablica 5.13 Statistika interakcija u prozorima za maksimalnu udaljenost

	NN	NV	VN	VV	N	V	S
Hetero	138566	226803	226183	1138145	453431	453439	523281
Homo	255953	406823	406235	256054	804329	805102	823967
Homo i hetero	358572	558463	558867	359284	1235378	1234582	1314825

Tablica 5.14 Statistika interakcija u prozorima za PIADA algoritam

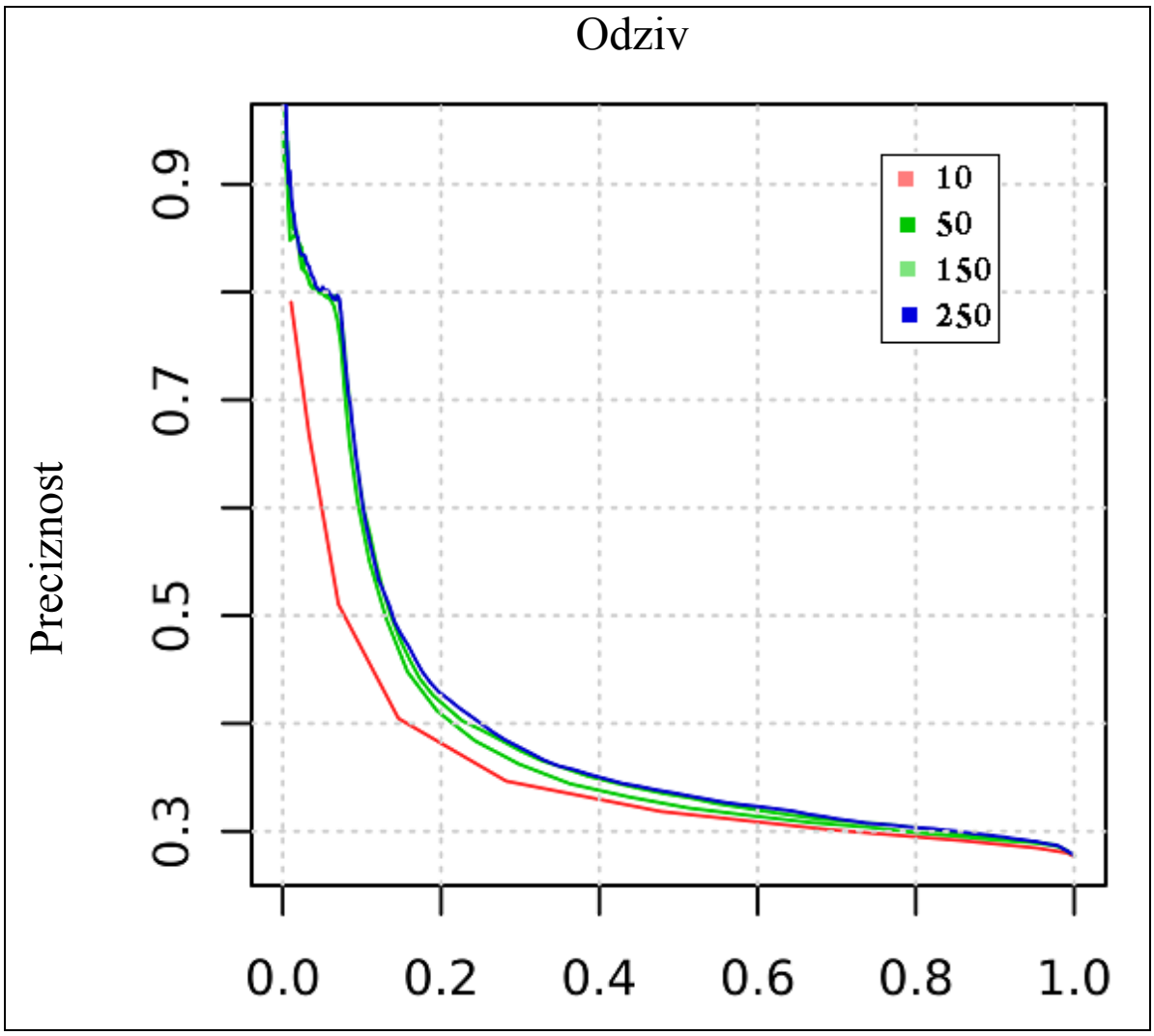
	NN	NV	VN	VV	N	V	S
Hetero	113165	203772	203940	113841	440479	4400084	512804
Homo	248225	365578	365334	249047	738328	738250	813276
Homo i hetero	344040	542472	542379	345513	1125610	1125225	1270191

5.2 Predviđanje interakcija

5.2.1 Implementacija i parametri klasifikatora

Klasifikatori kreirani za potrebe ovog rada temeljili su se na algoritmu slučajnih šuma i implementirani su u programskom jeziku R. Predikcija je vršena u prvom slučaju pomoću sljedova od po devet uzastopnih aminokiselinskih ostataka iz lanaca skupa, a u drugom slučaju su uz informaciju o slijedu odnosno primarnoj strukturi dodane i informacije o profilima slijeda i to za svaki od ukupno 12 korištenih skupova.

Broj stabala slučajne šume kojom bi se radila klasifikacija bilo bi uvijek jednak 250, a ta vrijednost određena je eksperimentalno. Broj atributa analiziranih za grananje u svakom čvoru stabla bio je jednak drugom korijenu iz ukupnog broja atributa, dakle tri kada se predikcija vršila samo na osnovu sekvence i 13 kada se predikcija vršila na osnovu sekvence i profila. Primjer utjecaja broja stabala šume na rezultate predikcije dan je na slici 5.7 u obliku grafa *preciznost-odziv*.



Slika 5.7 Utjecaj broja stabala slučajne šume klasifikaciju

Postupak kreiranja klasifikatora bio je identičan za svaki od njih ukupno 12. Svaki klasifikator bio je treniran na 70% slučajno odabranih uzoraka iz njemu dodijeljenog skupa, a preostalih 30% uzoraka bilo je korišteno za predviđanje odnosno procjenu pogreške klasifikatora. Također, u drugom pokušaju klasifikator bi bio treniran na svim dostupnim uzorcima iz skupa, a procjena performansi obavljala bi se temeljem *oob* pogreške.

Svi kreirani klasifikatori provjereni su krosvalidacijom pomoću tri skupa, iako to općenito, kao što je spomenuto ranije, nije potrebno provoditi. Originalni skup namijenjen nekom klasifikatoru slučajno bi se razdijelio na tri jednako velika podskupa, od kojih bi onda svaka dva bila korištena za treniranje, a treći za testiranje.

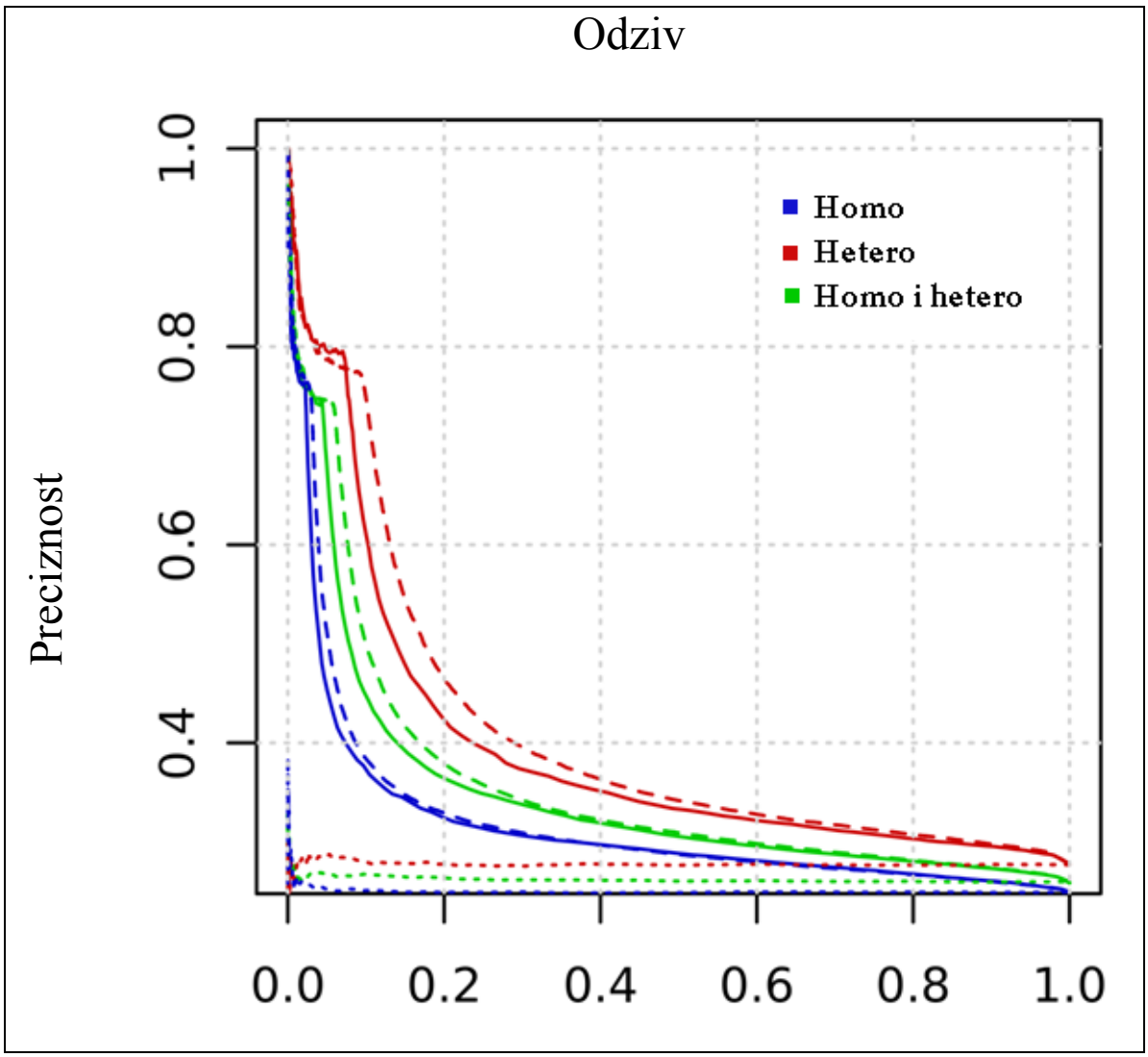
U nastavku su prikazani rezultati klasifikacije skupa za testiranje u obliku grafa *preciznost-odziv*, najprije za klasifikaciju temeljem samo sekvence aminokiselinskih ostataka u

podpoglavlju 5.2.2.1, a zatim za klasifikaciju temeljem i sekvence aminokiselinskih ostataka i profila slijeda u podpoglavlju 5.2.2.2.

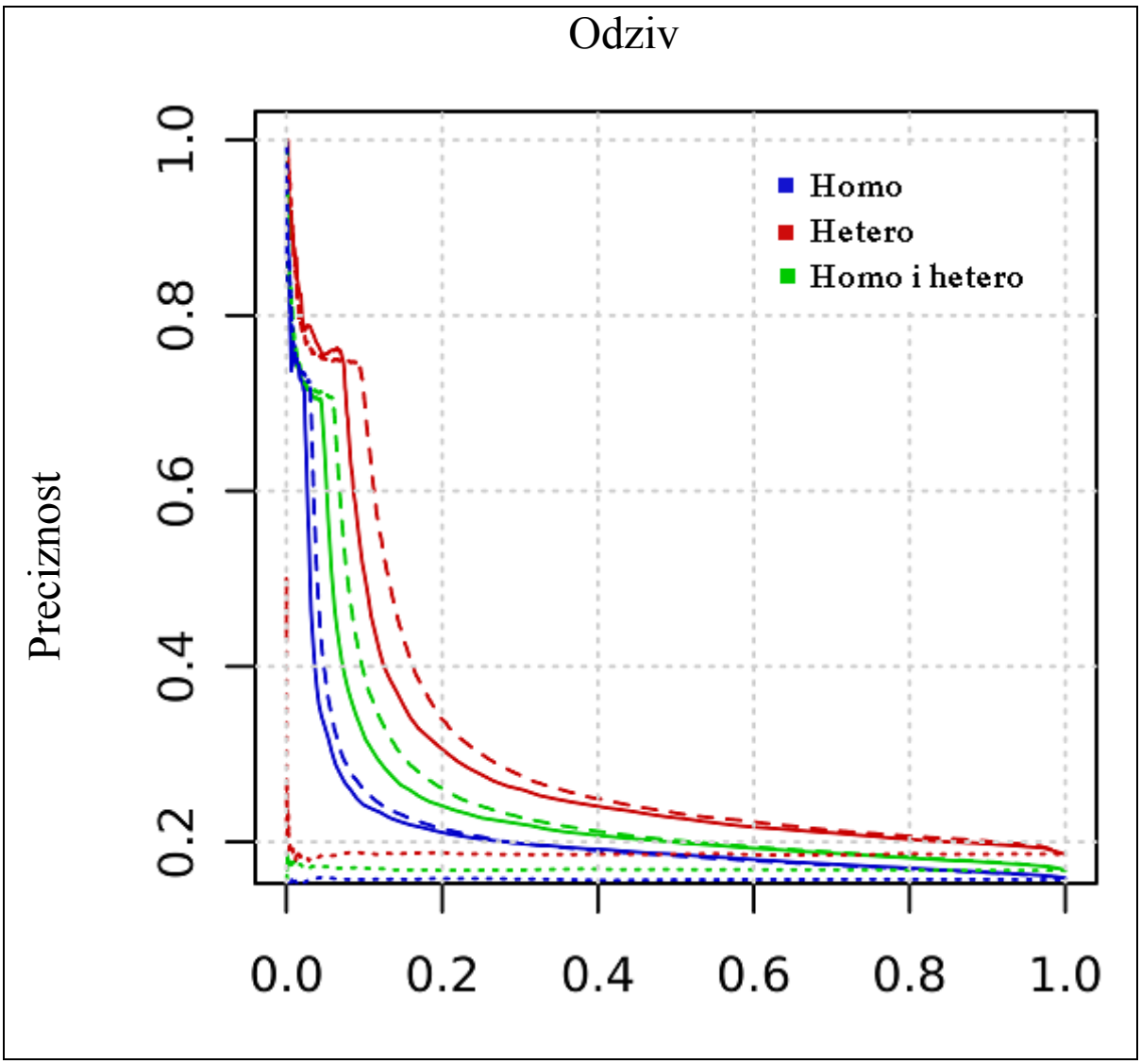
5.2.2 **Rezultati klasifikacije**

5.2.2.1 Rezultati klasifikacije temeljem sekvence aminokiselinskih ostataka

U ovom podpoglavlju najprije su prikazani rezultati ostvareni uz mjesta interakcije prvotno definirana preko maksimalne udaljenosti teških atoma lanaca u interakciji. Na slici 5.8 ako je za dodjeljivanje klase 1 uzorku dovoljno da je središnji aminokiselinski ostatak u kontaktu, a na slici 5.9 ukoliko je nužno da uz njega u kontaktu budu još barem 4 ostataka na udaljenosti ne više od 3 od središnjeg. Na obije slike, uz rezultate predikcije za testni skup (puna linija), nalaze se i rezultati predikcije kada su u testnom skupu ciljne vrijednosti klasa slučajno pomiješane (točkasta linija) i rezultati *oob* pogreške za treniranje klasifkatora cijelim skupom (crtkana linija). Ova tri podatka na grafu su za svaki skup označena istom bojom.



Slika 5.8 Graf *preciznost-odziv* za testne skupove (puna linija), slučajno testiranje (točkasta linija) i *oob* pogreška za treniranje na cijelom skupu (crtkana linija), ako je kriterij klasifikacije postojanje interakcije središnjeg aminokiselinskog ostatka u prozoru

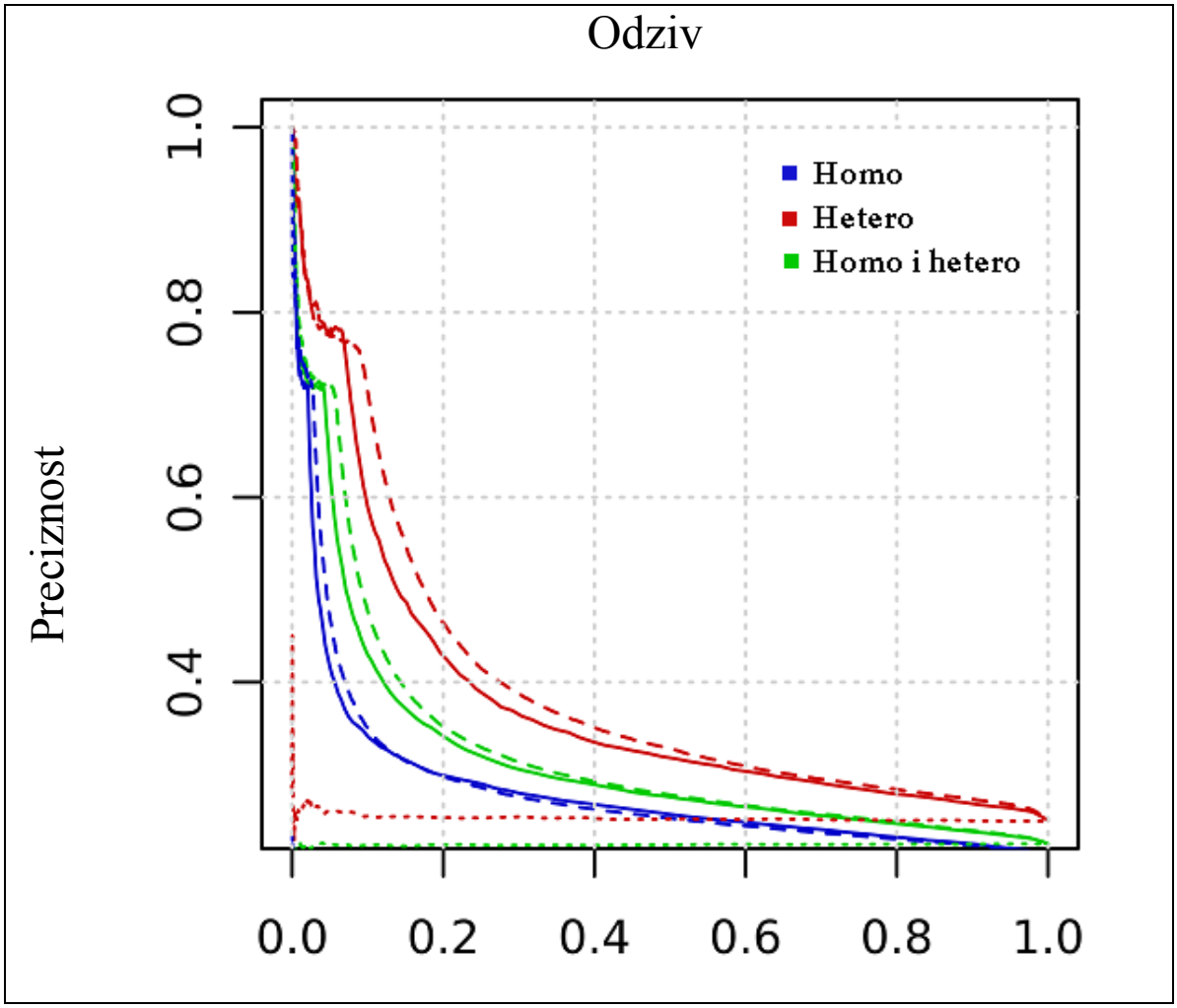


Slika 5.9 Graf *preciznost-odziv* za testne skupove (puna linija), slučajno testiranje (točkasta linija) i *oob* pogreška za treniranje na cijelom skupu (crtkana linija), ako je kriterij klasifikacije postojanje interakcije središnjeg aminokiselinskog ostatka u prozoru i još barem 4 interakcije na udaljenosti ne više od 3 od središnjeg aminokiselinskog ostatka

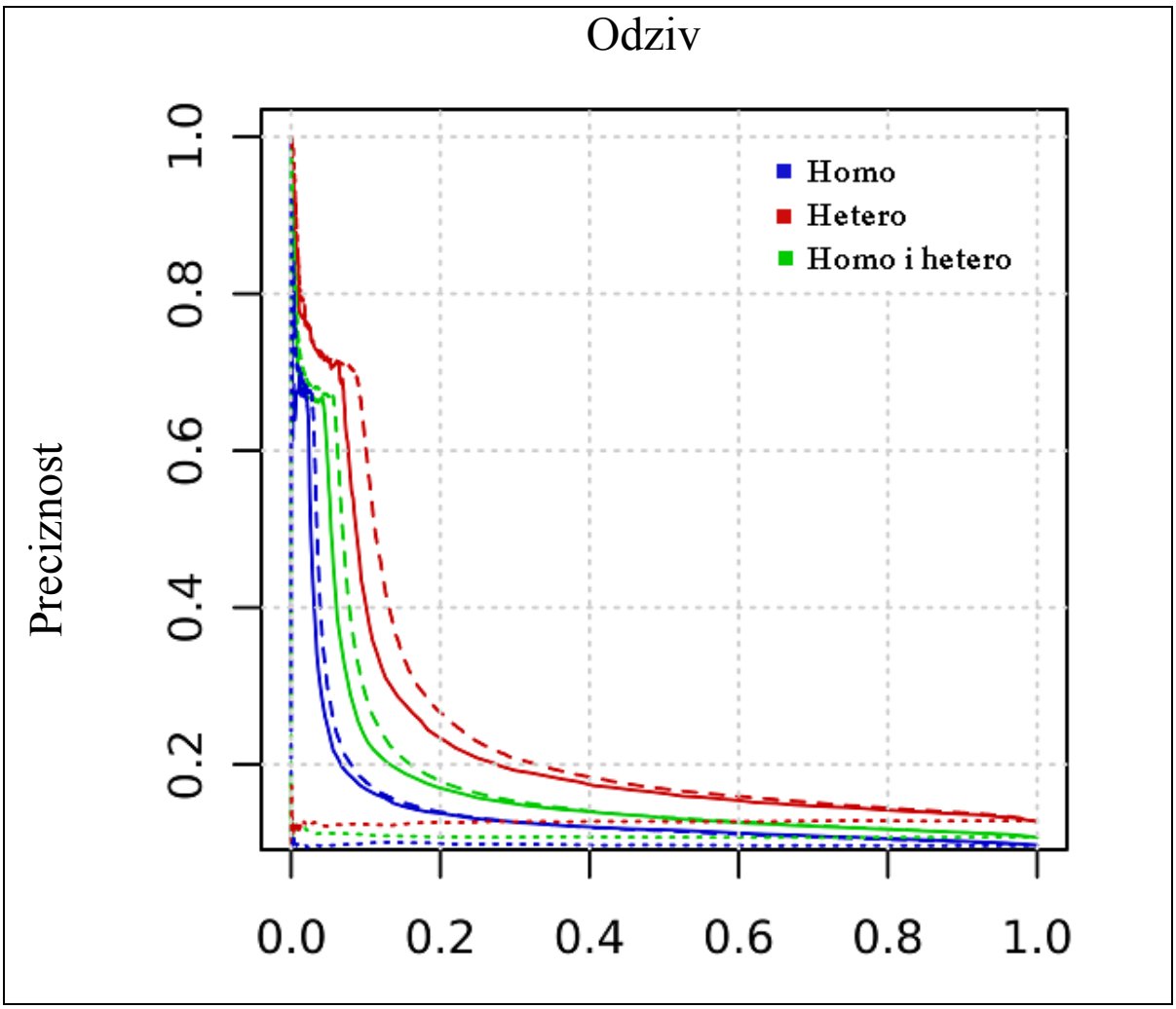
Sa navedena dva grafa možemo očitati mnoštvo informacija. Najprije, vidljiva je sličnost ponašanja klasifikatora odnosno njegovih sposobnosti za sve korištene skupove. Pune linije koje označuju odnos preciznosti i odziva sličnog su oblika i za homo, hetero i skup koji je njihova kombinacija, što ukazuje na sličnost zaključaka koje je klasifikator izveo iz uzoraka, odnosno sličnost obilježja samih uzoraka. Nadalje, promatrajući površinu između crtkane linije i odgovarajuće pune linije bilo kojeg skupa na oba prethodna grafa zapravo promatramo koliko je dobiveni klasifikator bolji od slučajnog klasifikatora. Klasifikator je stoga, kako se i vidi na grafu, uspio iz podataka o sekvenci aminokiselinskih ostataka u određenoj mjeri izvući informacije korisne pri klasifikaciji neviđenih uzoraka. Na kraju, iz odnosa pune i odgovarajuće crtkane linije za bilo koji skup na oba grafa zaključujemo kako

je korištenjem svih dostupnih uzoraka za učenje, umjesto 70% klasifikator uspio još malo poboljšati performanse.

U nastavku rada prikazani su isti rezultati, ali u slučaju kada su originalna mjesta interakcija određena pomoću PIADA algoritma. Na slici 5.10 u slučaju da je klasifikacija ovisila samo o kontaktu središnjeg aminokiselinskog ostatka u prozoru, a na slici 5.11 ukoliko je osim toga ovisila i o postojanju još bar 4 aminokiselinska ostatka na udaljenosti ne više od 3 od središnjeg.



Slika 5.10 Graf *preciznost-odziv* za testne skupove (puna linija), slučajno testiranje (točkasta linija) i *oob* pogreška za treniranje na cijelom skupu (crtkana linija), ako je kriterij klasifikacije postojanje interakcije središnjeg aminokiselinskog ostatka u prozoru



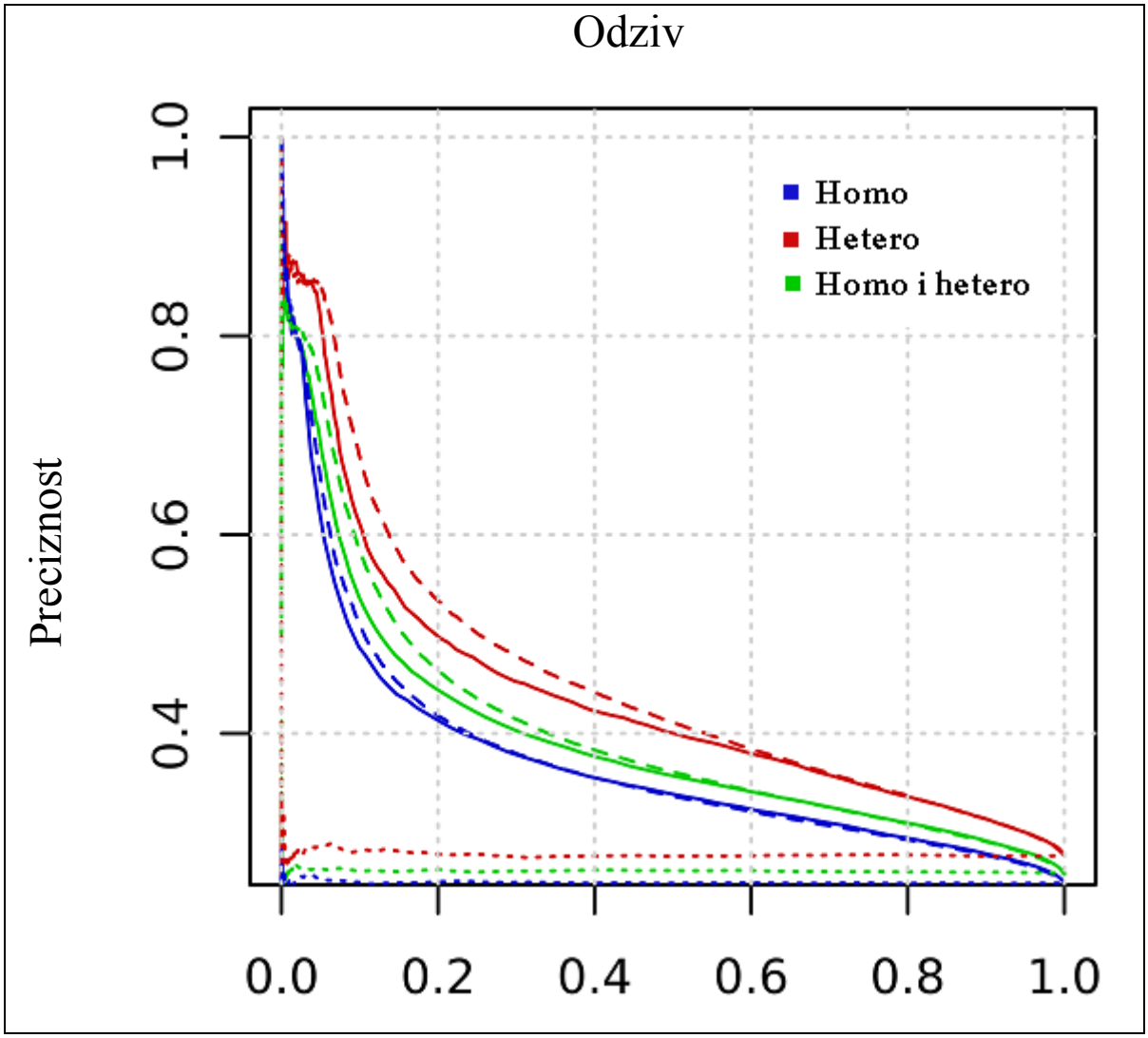
Slika 5.11 Graf *preciznost-odziv* za testne skupove (puna linija), slučajno testiranje (točkasta linija) i *oob* pogreška za treniranje na cijelom skupu (crtkana linija), ako je kriterij klasifikacije postojanje interakcije središnjeg aminokiselinskog ostatka u prozoru i još barem 4 interakcije na udaljenosti ne više od 3 od središnjeg aminokiselinskog ostatka

Rezultati pokazuju ista obilježja kao i rezultati sa prethodne dvije slike, ali su ipak malo slabiji. Ipak, prednost nad slučajnim klasifikatorom ostaje neupitna.

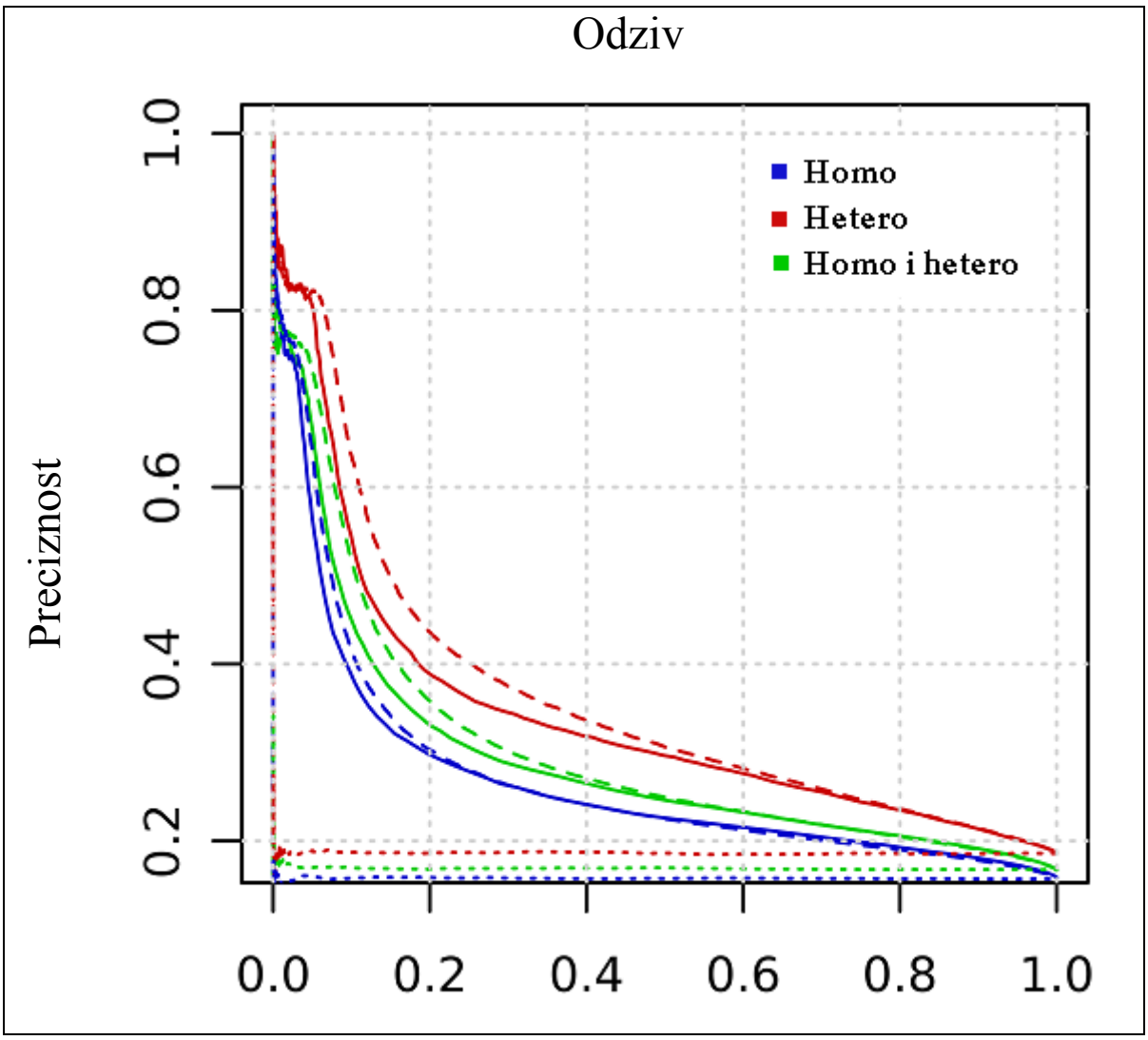
5.2.2.2 Rezultati klasifikacije temeljem sekvence aminokiselinskih ostataka i profila slijeda

U nastavku su prikazani rezultati ostvareni uz mjesta interakcije prvotno definirana preko maksimalne udaljenosti teških atoma lanaca u interakciji, ali za predviđanje obavljeno ne samo uz pomoć sekvence aminokiselinskih ostataka nego i profila slijeda. Na slici 5.12 ako je za dodjeljivanje klase 1 uzorku dovoljno da je središnji aminokiselinski ostatak u kontaktu, a na slici 5.13 ukoliko je nužno da uz njega u kontaktu budu još barem 4 ostatka na udaljenosti ne više od 3 od središnjeg. Isto kao i u prethodnom slučaju, na obije slike, uz rezultate predikcije za testni skup (puna linija), nalaze se i rezultati predikcije kada su u

testnom skupu ciljne vrijednosti klasa slučajno pomiješane (točkasta linija) i rezultati *oob* pogreške za treniranje klasifkatora cijelim skupom (crtkana linija).



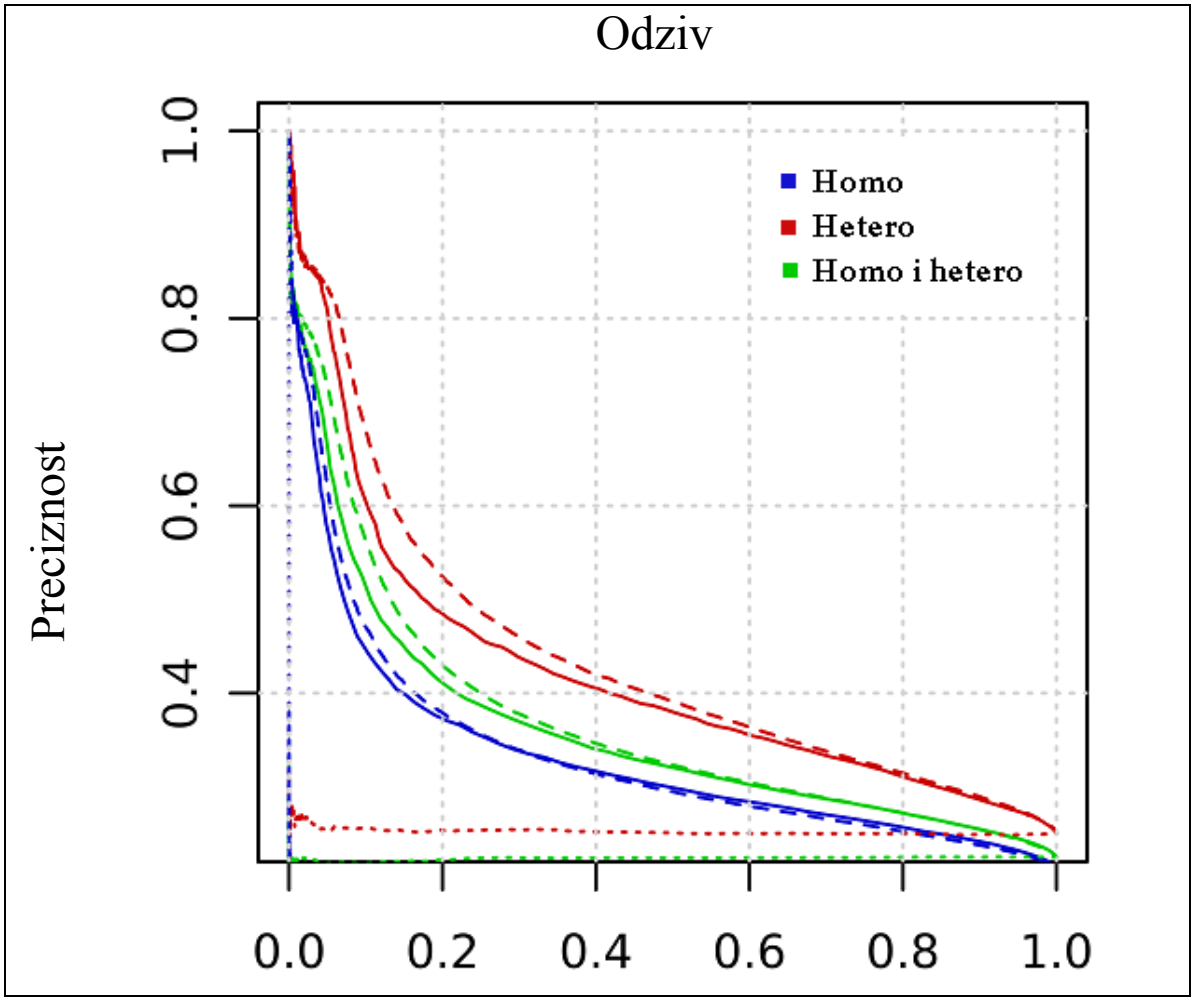
Slika 5.12 Graf *preciznost-odziv* za testne skupove (puna linija), slučajno testiranje (točkasta linija) i *oob* pogreška za treniranje na cijelom skupu (crtkana linija), ako je kriterij klasifikacije postojanje interakcije središnjeg aminokiselinskog ostatka u prozoru



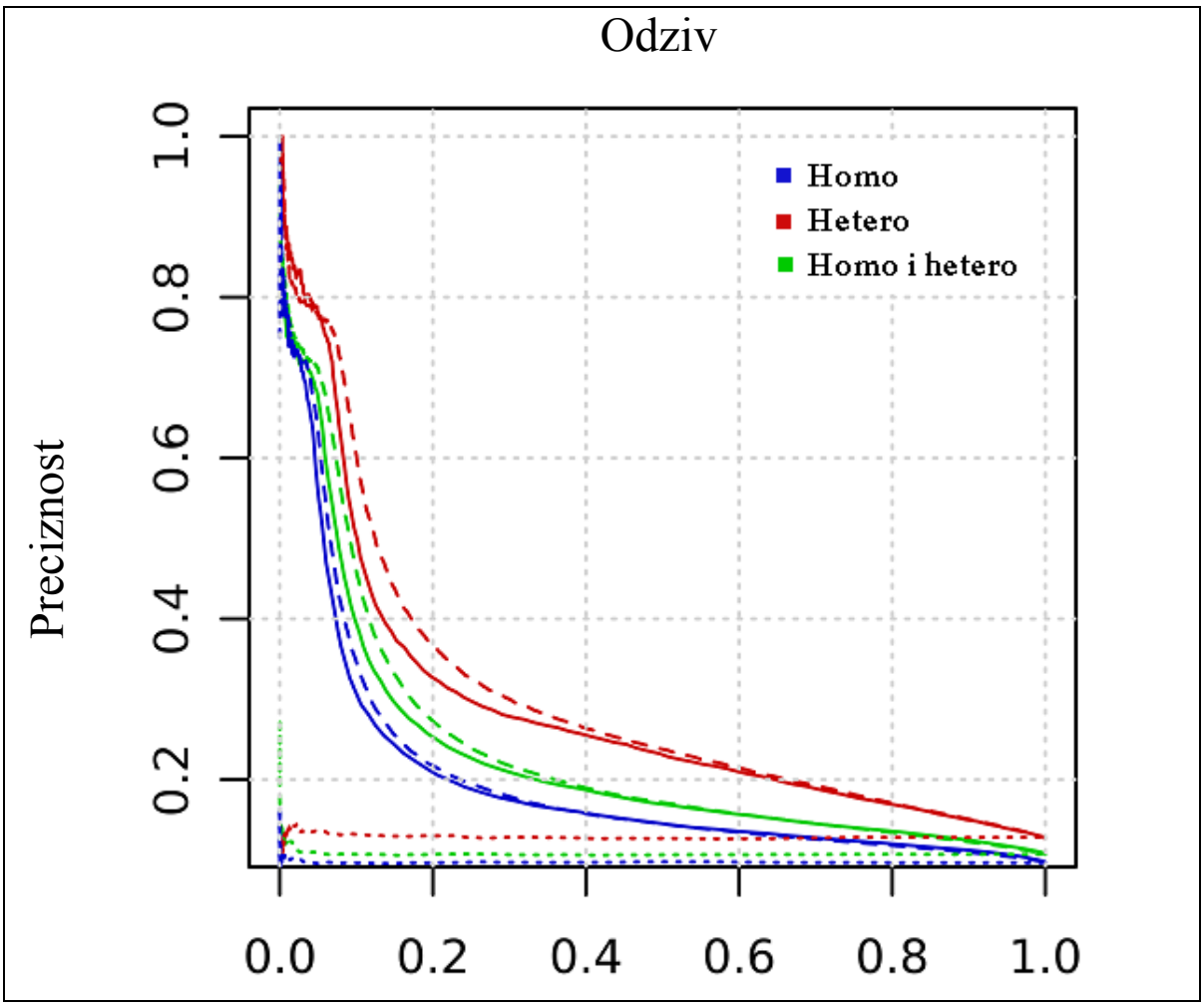
Slika 5.13 Graf *preciznost-odziv* za sve testne skupove (puna linija), slučajno testiranje (točkasta linija) i *oob* pogreška za treniranje na cijelom skupu (crtkana linija), ako je kriterij klasifikacije postojanje interakcije središnjeg aminokiselinskog ostatka u prozoru i još barem 4 interakcije na udaljenosti ne više od 3 od središnjeg aminokiselinskog ostatka

Na prethodnim grafovima izvodimo slične zaključke kao i ranije: definitivno je jasno vidljiva razlika između slučajnog i kreiranog klasifikatora, grafovi su i dalje sličnog oblika, a treniranjem na cijelom skupu rezultantna *oob* pogreška sugerira malo bolju predikciju. Ipak, valja primijetiti da su rezultati općenito bolji od prethodnih i pokazuje se kako je korištenjem informacija o profilima slijeda moguće poboljšati performanse klasifikatora. Također je moguće primijetiti kako se općenito uz definiranje uvjeta za dodjeljivanjem klase 1 uzorku u ovisnosti samo o interakciji središnjeg aminokiselinskog ostatka u prozoru postižu bolji rezultati nego uz dodatni uvjet kako isto mora vrijediti za još barem 4 aminokiselinska ostataka na udaljenosti ne većoj od 3 od središnjeg ostatka. Profili slijeda stoga zaista donose informaciju korisnu za predikciju.

U nastavku rada prikazani su još isti rezultati, ali u slučaju kada su originalna mjesta interakcija određena pomoću PIADA algoritma. Na slici 5.14 u slučaju da je klasifikacija ovisila samo o kontaktu središnjeg aminokiselinskog ostatka u prozoru, a na slici 5.15 ukoliko je osim toga ovisila i o postojanju još bar 4 aminokiselinska ostataka na udaljenosti ne više od 3 od središnjeg.



Slika 5.14 Graf *preciznost-odziv* za testne skupove (puna linija), slučajno testiranje (točkasta linija) i *oob* pogreška za treniranje na cijelom skupu (crtkana linija), ako je kriterij klasifikacije postojanje interakcije središnjeg aminokiselinskog ostatka u prozoru



Slika 5.15 Graf *preciznost-odziv* za testne skupove (puna linija), slučajno testiranje (točkasta linija) i *oob* pogreška za treniranje na cijelom skupu (crtkana linija), ako je kriterij klasifikacije postojanje interakcije središnjeg aminokiselinskog ostatka u prozoru i još barem 4 interakcije na udaljenosti ne više od 3 od središnjeg aminokiselinskog ostatka

I ovi rezultati ukazuju na slične zaključke kao i prethodni. Kriterij za mjesto interakcije u kojem je nužno samo da središnji aminokiselinski ostatak bude u interakciji daje bolje rezultate nego kriterij koji zahtijeva da osim njega u kontaktu budu bar još 4 ostatka na udaljenosti ne većoj od 3 od središnjeg. Također ako početna mjesta interakcija određena pomoću zadane maksimalne udaljenosti, rezultati predikcije su bolji nego pri korištenju PIADA algoritma.

Konkretne vrijednosti točaka sa krivulja klasifikatora prikazanih na prethodnim slikama nalaze se u tablicama 5.15 i 5.16. Kao što je prethodno pokazano, najbolji rezultati ostvareni su za sva tri skupa ukoliko su kontakti određeni pomoću maksimalne udaljenosti teških atoma, a klasa uzorka određena je samo na temelju kontakta središnjeg aminokiselinskog ostatka u prozoru.

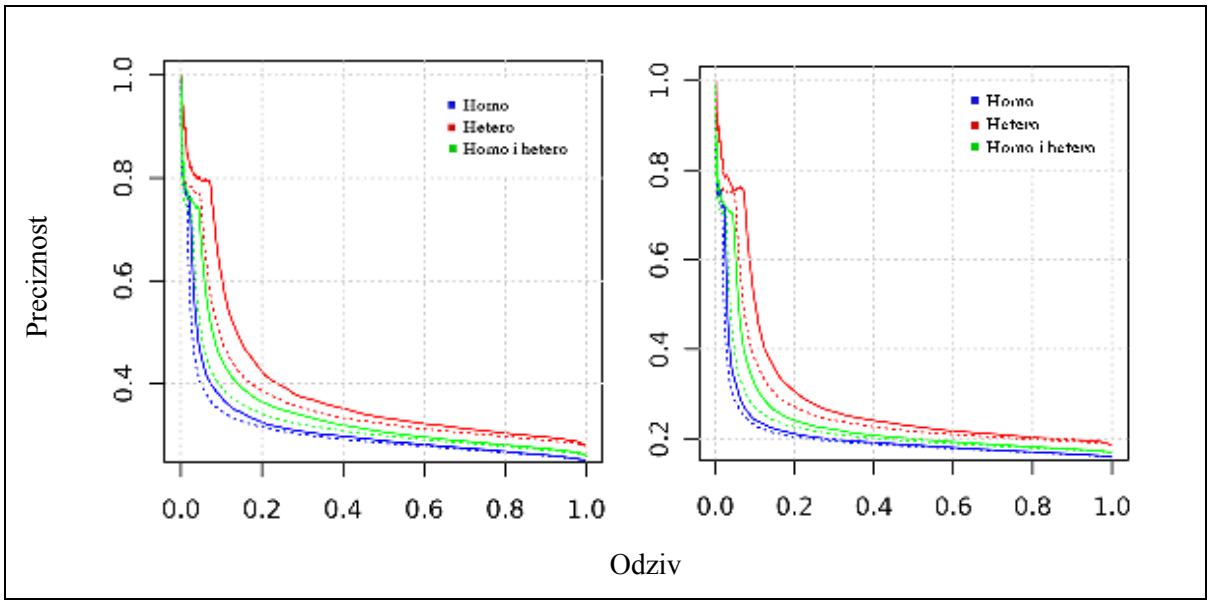
Tablica 5.15 Primjeri vrijednosti preciznosti i odziva izraženih u postotcima za klasifikaciju temeljem sekvence

	Središnji ostatak u prozoru je u interakciji						Središnji ostatak u prozoru je u interakciji i još barem 4 na udaljenosti najviše 3 mjesta od njega					
	Maksimalna udaljenost			PIADA			Maksimalna udaljenost			PIADA		
	Hetero	Homo	Homo i hetero	Hetero	Homo	Homo i hetero	Hetero	Homo	Homo i hetero	Hetero	Homo	Homo i hetero
Preciznost	0.795	0.453	0.673	0.783	0.417	0.621	0.756	0.331	0.634	0.715	0.249	0.568
Odziv	0.050	0.050	0.051	0.051	0.050	0.050	0.050	0.049	0.498	0.051	0.049	0.050
Prag	0.776	0.524	0.576	0.760	0.496	0.560	0.732	0.392	0.452	0.696	0.296	0.360
Preciznost	0.611	0.367	0.443	0.590	0.341	0.430	0.501	0.238	0.332	0.397	0.168	0.228
Odziv	0.100	0.105	0.103	0.100	0.101	0.100	0.101	0.109	0.102	0.101	0.102	0.103
Prag	0.560	0.488	0.508	0.564	0.456	0.488	0.432	0.356	0.380	0.356	0.264	0.288
Preciznost	0.426	0.325	0.363	0.424	0.298	0.338	0.301	0.209	0.241	0.237	0.137	0.171
Odziv	0.197	0.197	0.203	0.205	0.206	0.206	0.206	0.208	0.200	0.195	0.203	0.196
Prag	0.504	0.456	0.468	0.492	0.416	0.436	0.380	0.328	0.344	0.308	0.236	0.256
Preciznost	0.350	0.296	0.320	0.336	0.268	0.290	0.239	0.191	0.207	0.173	0.121	0.140
Odziv	0.410	0.406	0.396	0.397	0.401	0.394	0.413	0.411	0.405	0.405	0.394	0.391
Prag	0.452	0.412	0.424	0.432	0.372	0.388	0.336	0.292	0.304	0.260	0.204	0.220
Preciznost	0.322	0.280	0.295	0.303	0.249	0.264	0.216	0.180	0.192	0.154	0.112	0.125
Odziv	0.600	0.609	0.609	0.603	0.593	0.615	0.608	0.596	0.610	0.601	0.608	0.608
Prag	0.416	0.376	0.384	0.388	0.336	0.344	0.304	0.264	0.272	0.228	0.176	0.188

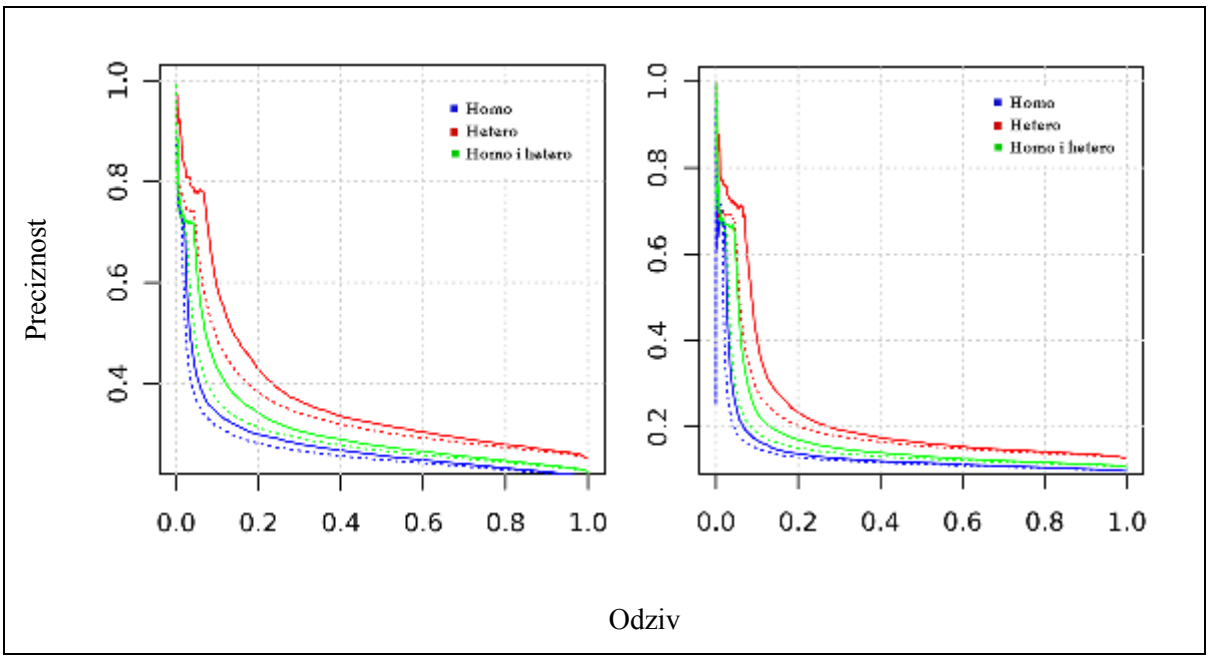
Tablica 5.16 Primjeri vrijednosti preciznosti i odziva izraženih u postotcima za klasifikaciju temeljem sekvence i profila slijeda

	Središnji ostatak u prozoru je u interakciji						Središnji ostatak u prozoru je u interakciji i još barem 4 na udaljenosti najviše 3 mjesta od njega					
	Maksimalna udaljenost			PIADA			Maksimalna udaljenost			PIADA		
	Hetero	Homo	Homo i hetero	Hetero	Homo	Homo i hetero	Hetero	Homo	Homo i hetero	Hetero	Homo	Homo i hetero
Preciznost	0.815	0.611	0.691	0.804	0.569	0.663	0.803	0.567	0.667	0.781	0.548	0.674
Odziv	0.050	0.051	0.050	0.051	0.052	0.050	0.051	0.050	0.050	0.050	0.050	0.050
Prag	0.596	0.452	0.496	0.600	0.416	0.476	0.516	0.348	0.404	0.496	0.276	0.348
Preciznost	0.601	0.478	0.527	0.549	0.442	0.512	0.526	0.378	0.448	0.498	0.293	0.399
Odziv	0.104	0.107	0.105	0.103	0.103	0.101	0.104	0.105	0.101	0.101	0.106	0.098
Prag	0.516	0.412	0.440	0.488	0.376	0.416	0.420	0.304	0.340	0.360	0.224	0.268
Preciznost	0.496	0.412	0.442	0.480	0.371	0.407	0.386	0.299	0.331	0.328	0.205	0.249
Odziv	0.203	0.202	0.202	0.198	0.206	0.207	0.204	0.198	0.199	0.198	0.206	0.206
Prag	0.464	0.380	0.400	0.444	0.340	0.368	0.368	0.272	0.296	0.304	0.192	0.220
Preciznost	0.421	0.352	0.376	0.404	0.320	0.338	0.317	0.240	0.267	0.254	0.160	0.184
Odziv	0.409	0.418	0.403	0.406	3.98	0.414	0.408	0.410	0.390	0.407	0.390	0.414
Prag	0.392	0.336	0.352	0.372	0.300	0.316	0.300	0.232	0.252	0.240	0.160	0.176
Preciznost	0.378	0.322	0.339	0.355	0.283	0.300	0.277	0.215	0.233	0.210	0.135	0.155
Odziv	0.610	0.608	0.596	0.601	0.610	0.618	0.597	0.603	0.578	0.599	0.603	0.615
Prag	0.336	0.304	0.312	0.316	0.264	0.276	0.244	0.204	0.216	0.188	0.132	0.144

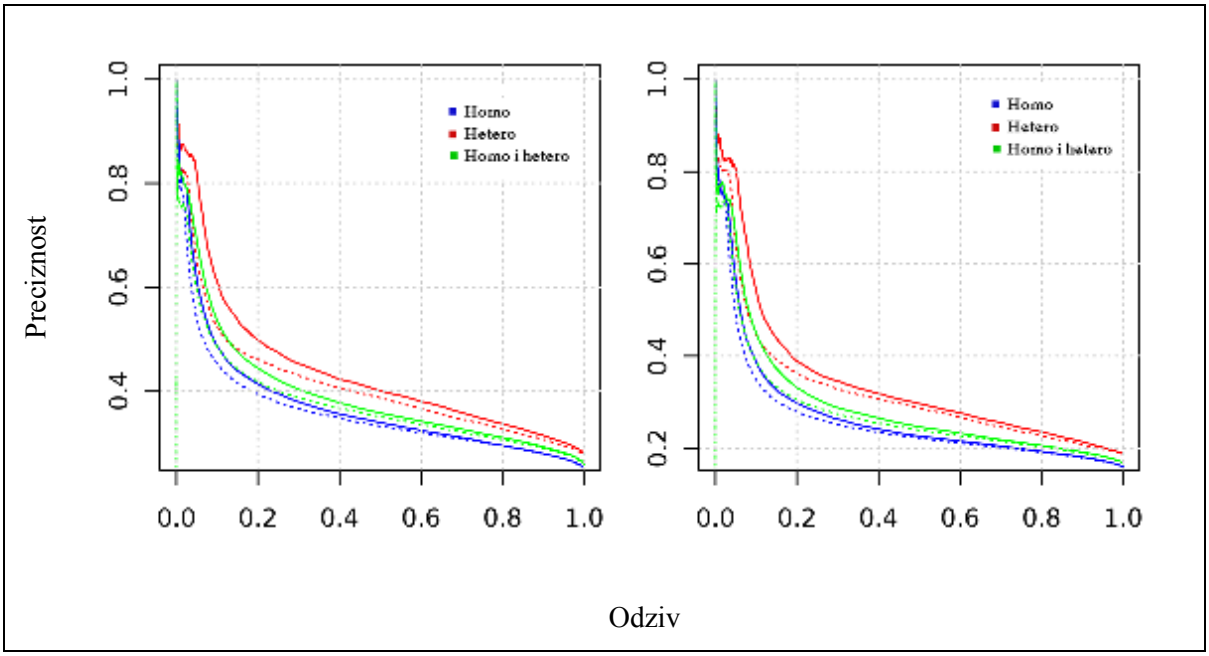
Rezultati krosvalidacije prikazani su u nastavku. Ovi rezultati predstavljaju kontrolu provedenog postupka i sugeriraju da je navedene informacije ključne u raspoznavanju moguće dobiti iz bilo kojih slučajno odabranih 70% uzoraka iz skupa, a ne baš onih koje je sadržavao testni skup kod svakog od prethodnih klasifikatora. Na slikama 5.16, 5.17, 5.18 i 5.19 punom je crtom označena karakteristika klasifikatora u slučaju testiranja na 30% skupa, a crtkanom je crtom označena karakteristika dobivena krosvalidacijom i to na prve dvije slike samo za klasifikaciju sekvencom amionkiselinskih ostataka, a na druga dva na temelju informacija o sekvenci i odgovarajućih profila.



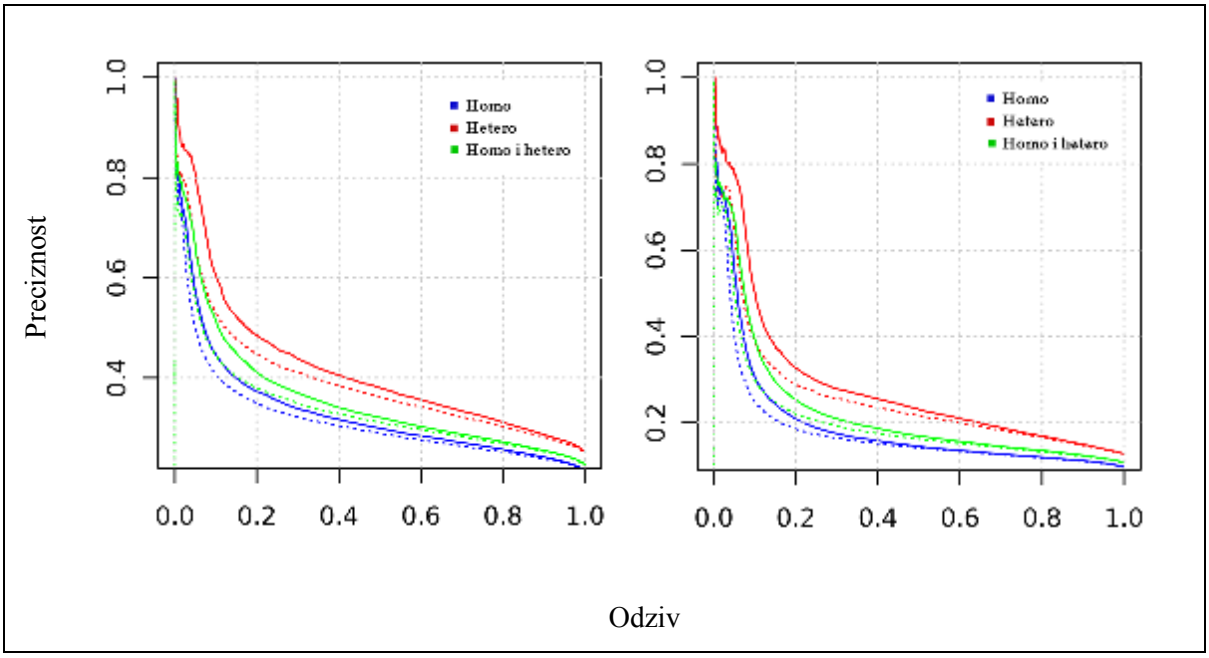
Slika 5.16 Graf *preciznost-odziv* za klasifikaciju temeljem sekvence aminokiselinskih ostataka za maksimalnu dozvoljenu udaljenost teških atoma, ako je kriterij klasifikacije postojanje interakcije središnjeg aminokiselinskog ostatka u prozoru (lijevo) ili uz to još barem 4 interakcije na udaljenosti ne više od 3 od središnjeg aminokiselinskog ostatka (desno)



Slika 5.17 Graf *preciznost-odziv* za klasifikaciju temeljem sekvence aminokiselinskih ostataka za PIADA algoritam, ako je kriterij klasifikacije postojanje interakcije središnjeg aminokiselinskog ostatka u prozoru (lijevo) ili uz to još barem 4 interakcije na udaljenosti ne više od 3 od središnjeg aminokiselinskog ostatka (desno)



Slika 5.18 Graf *preciznost-odziv* za klasifikaciju temeljem sekvence aminokiselinskih ostataka i odgovarajućih profila za maksimalnu dozvoljenu udaljenost teških atoma, ako je kriterij klasifikacije postojanje interakcije središnjeg aminokiselinskog ostatka u prozoru (lijevo) ili uz to još barem 4 interakcije na udaljenosti ne više od 3 od središnjeg aminokiselinskog ostatka (desno)



Slika 5.19 Graf *preciznost-odziv* za klasifikaciju temeljem sekvence aminokiselinskih ostataka i odgovarajućih profila za PIADA algoritam, ako je kriterij klasifikacije postojanje interakcije središnjeg aminokiselinskog ostatka u prozoru (lijevo) ili uz to još barem 4 interakcije na udaljenosti ne više od 3 od središnjeg aminokiselinskog ostatka (desno)

U cjelini, rezultati klasifikacije neosporivo ukazuju na to da sekvenca u velikoj i profili u nešto manjoj mjeri omogućavaju uspješnu predikciju mjesta proteinskih interakcija. Uz različite uvjete i metode definiranja proteinskih interakcija pojedini rezultati su malo bolji ili lošiji ali generalno slični. Rezultati provedene krosvalidacije to su i potvrdili: krivulja dobivena krosvalidacijom uvijek je pratila oblik originalne krivulje klasifikatora, ali sa blago slabijim rezultatima.

Optimalni rezultati klasifikacije postignuti su, dakle, za mjesta interakcije određena na temelju maksimalne udaljenosti teških atoma i promatranja kontakata samo središnjeg aminokiselinskog ostatka u pomičnom prozoru. Postignute vrijednosti preciznosti i odziva iznose 81.5% preciznosti za 5% odziva, 60.1% preciznosti za 10.4% odziva, 49.6% preciznosti za 20.3% odziva ili 37.8% za 61% odziva za hetero skup. Za homo skup rezultati su nešto slabiji i iznose 61.1% preciznosti za 5.1% odziva, 47.8% preciznosti za 10.7% odziva, 41.2% preciznosti za 20.2% odziva, 35.2% preciznosti za 41.8% odziva ili 32.2% preciznosti za 60.8% odziva. Konačno, za skup od homo i hetero lanaca rezultati su nešto bolji od homo skupa, ali lošiji od hetero skupa i iznose 69.1% preciznosti za 5%

odziva, 52.7% preciznosti za 10.5% odziva, 42.2% preciznosti za 20.2% odziva, 37.6% preciznosti za 40.3% odziva ili 33.9% preciznosti za 59.6% odziva.

6. Zaključak

Cilj ovog rada bio je proanalizirati statistička svojstva proteinskih interakcija koje ostvaruju proteinski lanci iz tri vrlo neredundantna skupa, kao i pokazati može li se ista mjesta uspješno predvidjeti na osnovu podataka o susjednim aminokiselinskim ostatcima oko promatranog i evolucijskih profila tog slijeda.

Budući da sam pojam proteinske interakcije nije sam po sebi jednoznačno definiran, bilo ga je potrebno pobliže odrediti, a i statistike interakcija i predikcije provjeriti stoga za više njihovih načina definiranja. U radu su stoga korištena dva načina određivanja mjesta proteinskih interakcija (maksimalna udaljenost dvaju teških atoma i PIADA algoritam), a za mjesta interakcije određena tim metodama primijenjen je kriterij da kako bi uzorak za predikciju bio svrstan u klasu 1 nužno je da mu je središnji aminokiselinski ostatak mjesto interakcije, odnosno da su to i još 4 aminokiselinska ostatka na udaljenosti ne većoj od 3 od njega u drugom slučaju.

Rezultati statističke analize ukazali su na sličnost u trendovima među interakcijama unutar sva tri skupa i u slučaju kada su mjesta interakcije određena PIADA algoritmom i kada su mjesta interakcije određena preko maksimalne dopuštene udaljenosti teških atoma lanaca u interakciji. Sličnost je prisutna glede svih ispitanih statistika: glede postotaka aminokiselinskih ostataka pojedine vrste koji sudjeluju u interakcijama općenito, glede postotaka aminokiselinskih ostataka pojedine vrste koji sudjeluju u interakcijama s svakim pojedinim aminokiselinskim ostatkom i glede parova aminokiselinskih ostataka koji su najčešće u interakciji.

Postupak predikcije pokazao je kako je moguće da klasifikator u određenoj mjeri nauči prepoznavati potencijalna mjesta interakcija. Rezultati predikcije, ipak, relativno su loši na što sigurno utječe vrlo malen postotak redundancije unutar korištenih skupova. S druge strane, oni potvrđuju da se i uz nisku redundanciju bitni zaključci za klasifikaciju mogu izvući iz susjednih 8 aminokiselinskih ostataka oko središnjeg ostatka u pomičnom prozoru, a također u primjetno manjoj mjeri iz evolucijskih profila za promatrani slijed aminokiselinskih ostataka u pomičnom prozoru. Najbolji ostvareni rezultati klasifikacije ostvareni su za mjesta interakcije određena temeljem maksimalne udaljenosti teških atoma i iznose primjerice 60.1% preciznosti i 10.5% odziva ili 49.6% preciznosti i 20.3% odziva za hetero skup, 47.8% preciznosti i 10.7% odziva ili 41.2% preciznosti i 20.2% odziva za

homo skup, te 52.7% preciznosti i 10.5% odziva ili 44.2% preciznosti i 20.2% odziva za kombinirani hetero i homo skup.

Ostvareni rezultati ostavljaju mnoštvo prostora za poboljšanja korištenjem nekih drugih svojstava proteinske strukture, ali i sugeriraju koban utjecaj uklanjanja redundancije na uspješnost klasifikacije. Rezultati statističke analize otvaraju nadu da bi uz još neke preinake ili uvjete skupovi homo i hetero lanaca zaista pokazali u potpunosti ista statistička svojstva.

7. Popis literature

[1] Šikić, M., Računalna metoda za predviđanje mjesta proteinskih interakcija, doktorska disertacija, Fakultet elektrotehnike i računarstva, Zagreb, 2008.

[2] Bioinformatics, http://www.ncbi.nlm.nih.gov/About/primer/bioinformatics.html, urešeno 29. ožujka 2004., pristupljeno 22. svibnja 2010.

[3] Arroyo, E., Measuring Proteins, the Building Blocks of our Bodies, http://www.articlesbase.com/health-articles/measuring-proteins-the-building-blocks-of-our-bodies-162383.html, uređeno 10. svibnja 2007. , pristupljeno 17. svibnja 2010.

[4] Shen, J., Zhang, J. , Luo, X., Zhu, W., Yu, K., Chen, K., Li, Y., Predicting protein–protein interactions based only on sequences information, 2006.

[5] Petrović, J., Automatizacija postupka izrade neredundantnog skupa proteina i izrada Web sjedišta za preuzimanje rezultata, završni rad, 2008.

[6] Dragosavljević, V., Predviđanje mjesta proteinskih interakcija iz profila slijeda i aminokiselinskih ostataka, diplomski rad, Fakultet elektrotehnike i računarstva, Zagreb, 2009.

[7] Ofran, Y., Rost, B., Predicted protein-protein interaction sites from local sequence information, FEBS *Lett*, vol. 544, pp. 236-19, 5. svibnja 2003.

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

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

[10] Breiman, L., Cutler, A., Random Forest, http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm#intro, pristupljeno 17.4.2010.

[11] Drummond, C., Holte, R. C., Costcurves: An improved method for visualizing classifier performance, MachLearn 65:95–130, SpringerScience+BusinessMedia, LLC 2006.

[12] Weng, G. C., Poon, J., A New Evaluation Measure for Imbalanced Datasets, School of Information Technologies, J12, UniversityofSydney, Sydney, NSW, Australia 2006.

[13] Davis, J., Goadrich, M., The relationship between Precision Recall and ROC curves," ACM International Conference Proceeding Series; vol 148, pp 233 1240, 2006.

Sažetak

Naslov: Predviđanje proteinskih interakcija koristeći algoritam slučajnih šuma

Predviđanje proteinskih interakcija aktualna je tema iz područja bioinformatike čiji je cilj predvidjeti potencijalne interakcije proteinskih lanaca te ih primijeniti u kreiranju lijekova, tumačenju složenih metaboličkih reakcija ili nekom drugom području. Zadatak ovog diplomskog rada bio je odrediti mjesta interakcija za postojeće neredundantne skupove proteina i pokušati ih predvidjeti klasifikatorom. Mjesta proteinskih interakcija određena su na temelju dvije definicije: PIADA algoritmom i preko maksimalne dozvoljene udaljenost od 6 Å između teških atoma s lanaca proteina u interakciji. Na temelju određenih mjesta interakcija provedena je detaljan statistička analiza i kreiran je binarni klasifikator za predviđanje interakcija. Klasifikacija je provođena temeljem slijeda aminokiselinskih ostataka koji tvore lanac proteina i odgovarajućih profila slijeda. Rezultati ostvareni u radu sugeriraju da iako i sekvenca u većoj i profili u manjoj mjeri sadrže informaciju koja je korisna za klasifikaciju, oni sami nisu dovoljni za visoke rezultate klasifikacije mjerljive preko preciznosti i odziva. Rezultati statističke analize ukazuju na to da su korišteni skupovi po sastavu i svojstvima vrlo slični, iako bi bilo pogrešno tvrditi da si potpuno odgovaraju po svojstvima.

Summary

Title: Prediction of protein-protein interactions using Random Forests

Prediction of protein-protein interactions is a topic related to the area of bioinformatics, that aims to predict protein-protein interactions in order to apply them in drug design, explanation of complex metabolic processes and elsewhere. The objective for this master thesis was to determine the protein-protein interaction sites in the current non-redundant protein data sets using two different definitions of interaction sites and to try to predict them using a binary classifier based on the random forests algorithm. The protein interaction sites were determined using the PIADA algorithm and the maximal allowed distance of 6 Å between two heavy atoms of the interacting chains. Random Forests classification was based on the protein sequence and evolution profiles information. Also, a statistical analysis of the contents and the interaction properties of the used data sets has been performed. The results of this work have showed that both sequence and evolution profiles contain very useful information for classifying, but not in a sufficient amount to obtain very good results regarding precision and recall. The results of the statistical analysis have shown many similarities within the data sets , but not enough to consider them equal.

Ključne riječi

- Predviđanje proteinskih interakcija
- Neredundantni skup
- Algoritam slučajnih šuma
- Sekvenca
- Evolucijski profili
- Redundancija
- Matrica pogreške

Keywords

- Prediction of protein-protein interactions
- Non-redundant dana set
- Random Forests algorithm
- Sequence
- Evolution profiles
- Redundancy
- Error matrix