SVEUČILIŠTE U ZAGREBU

FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

ZAVRŠNI RAD br. 233

Predviđanje mjesta proteinskih interakcija iz sekvence aminokiselina

Dragana Čolić

Zagreb, lipanj 2008.

Zadatak

Na osnovu podataka iz zadanog skupa podataka potrebno je koristeći metode slučajnih šuma u prvom koraku i Bayesovu mrežu u drugom koraku napraviti predviđanje mjesta proteinskih interakcija, te napraviti analizu dobivenih podataka.

Skup podataka koji se koristi je neredundantni skup proteina. Kao ulazne podatke u prvi klasifikator (metoda slučajnih šuma) koristiti prozor od devet aminokiselina u nizu koji predstavljaju devet atributa. Oznaka klase definira se tako da se pozitivna vrijednost (1) pridijeli u slučaju kada je središnja aminokiselina u prozoru mjesto kontakta. Mjestom kontakta između dva proteina se definira ona aminokiselina čija je udaljenost prema najbližoj aminokiselini susjednog proteina manja od šest angstrema (6 Å).

U drugom koraku Bayesovu se mrežu uči identificirati mjesto interakcije zavisno o oznaci klase (1 za mjesto interakcije, 0 inače) svojih susjeda. Ulaz za Bayesovu mrežu oznake su klasa osam aminokiselina koje okružuju ciljanu (po četiri sa svake strane).

Pri radu koristiti postojeće alate za strojno učenje kao što je, npr. WEKA.

Posebno hvala Mili Šikiću na toleranciji, strpljenju i spremnosti na pomoć koju je iskazivao kroz cijelo vrijeme suradnje.

SADRŽAJ

1 Uvod

Proteini su velike organske molekule čija je uloga u živom organizmu od iznimne važnosti. Oni omogućuju izgradnju i razvoj stanica te potpomažu odvijanje kemijskih reakcija. Građeni su od lanaca aminokiselina koji se u tekućem mediju savijaju u prostorne strukture. Proteini svoje funkcije obavljaju kroz interakciju s drugim molekulama (npr. vezanje hemoglobina za molekulu kisika), a način na koji će protein stupati u interakcije uvelike je određen njegovom strukturom. Zato je poznavanje i predviđanje struktura proteina jedno od najvažnijih područja istraživanja u bioinformatici.

Ovaj rad bavi se prepoznavanjem područja interakcije u lancu aminokiselina, odnosno prepoznavanjem pojedinačnih aminokiselina koje se vežu za drugi proteinski lanac. Uočavanje veze između specifičnosti pojedinih interakcija i aminokiselina koje u njima sudjeluju omogućuje bolje razumijevanje bioloških procesa i umjetnu sintezu proteina s ciljanim svojstvima u farmaceutskoj industriji. Budući da broj poznatih proteinskih lanaca raste mnogo brže nego što eksperimentalne metode analize mogu pratiti, javlja se potreba za pronalaženjem računalnih metoda koje će ubrzati proces prepoznavanja proteinskih kompleksa i interakcija.

Metode razmatrane u ovom radu temelje se na eksperimentalno utvrđenoj činjenici da se područja interakcije mogu svrstati u skupine sličnih svojstava. Riječ je o sličnosti aminokiselinskih sljedova u područjima koja tvore kemijske veze. Također utvrđeno je da se u bliskoj okolini aminokiseline koja stupa u interakciju vrlo vjerojatno nalazi još aminokiselina koje čine to isto. Na temelju ovih svojstava gradi se složeni sustav klasifikatora koji mjesta interakcije prepoznaje u dva koraka.

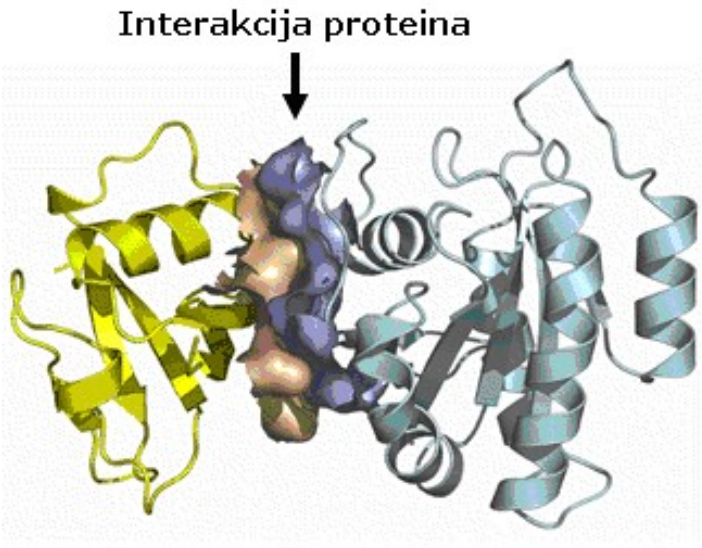
2 Analiza proteinskih kompleksa

Proteinski kompleksi prostorne su strukture koje se sastoje od jednog ili više lanaca aminokiselina. Dijelovi lanaca koji tvore međusobne veze i proteinski komplekse nazivaju se mjesta interakcije. Analizu proteinskih kompleksa možemo podijeliti na dva područja. Jedno je područje određivanja srodnosti lanaca na temelju sličnosti njihovih aminokiselinskih sljedova. Na temelju sličnosti proteinskih lanaca pretpostavlja se njihovo evolucijsko srodstvo i predviđaju se interakcije.

Npr. pretpostavimo:

- Ukoliko postoje dva lanca čija je sličnost veća od 90% smatraju se jednakima.
- Pretpostavimo postojanje dva lanca A i B čija je sličnost veća od 90%.
- lanac A je zaseban protein, a lanac B se javlja unutar kompleksa BC
- BC se sastoji od dva lanca B i C
- Postoji lanac D čija je mjera sličnosti lancu C veća od 90%
- Lanac D zaseban je protein.

Na temelju ovih pretpostavki može se zaključiti da će proteini A i D međusobno stupati u veze i tvoriti komplekse slične kompleksu BC.



Slika 1. Interakcija lanaca A i B proteina 1LFD

Drugi je cilj analize proteinskih kompleksa odrediti mjesta interakcije. Dijelovi lanaca koji mogu stupati u interakcije površinski su dijelovi strukture, a aminokiselinski ostatci (engl. *residue*) koje ih tvore nazivaju se površinski ostatci. Ti ostatci su moguća mjesta interakcije, a definiraju se takvima ukoliko je njihova udaljenost od između bilo koja dva teška atoma ostataka susjednog lanca manja od 6 Å (angstrem = 10^{-10} m) [3].

Isto tako utvrđeno je kako je sličnosti u sljedovima ostataka među kojima se nalazi jedno ili više mjesta interakcije najbolje tražiti na odsječcima proteinskih lanaca dugim 9 aminokiselina [2].

Već je spomenuto kako broj poznatih proteinskih lanaca raste mnogo brže nego što ih eksperimentalne metode mogu analizirati i utvrditi njihovo ponašanje (prostorne strukture koje poprimaju i interakcije u koje stupaju). Zato se javlja potreba za pronalaženjem novih, računalnih metoda kojima će se na temelju aminokiselinskih sljedova u proteinskim lancima moći predvidjeti njihovu trodimenzionalnu strukturu i interakcije u koje stupaju.

2.1 Traženje mjesta interakcije

Računalne metode analize proteinskih lanaca implementiraju matematički pristup obradi bioloških podataka. Pritom se ti podaci transformiraju u oblik pogodan za matematičku analizu. To se radi na temelju fizikalnih i kemijskih svojstava aminokiselina i lanaca kao što su hidrofobnost, polarnost, površina dostupna otapalu, elektrostatski potencijal i dr. svojstva. Sljedovi aminokiselina pretvaraju se u sljedove brojeva koji predstavljaju mjere promatranih svojstava. Takvi se sljedovi mogu promatrati kao diskretni signali te ih je moguće analizirati metodama obradbe informacija, npr. Fourierovom analizom.

Drugi, čest pristup u predviđanju mjesta interakcija oslanja se na metode strojnog učenja. Na temelju skupa podataka poznatih svojstava izgrađuje se klasifikatorski sustav. Jednom naučen klasifikatorski sustav u stanju je klasificirati novu, prvi put viđenu jedinku. Sustav je to bolji što više novih uzoraka točno klasificira.

U okviru ovog rada koriste se dvije različite metode klasifikacije uzoraka, metoda slučajnih šuma (eng. Random Forest - RF) i Bayesova metoda, koje se ukratko predstavljaju u idućem poglavlju.

2.2 Klasifikacija metodama raspoznavanja uzoraka

Raspoznavanje uzoraka područje je istraživanja umjetne inteligencije koje se bavi izgradnjom računalnih strojeva koji su u stanju odrediti neko svojstvo ulaznog objekta. U našem slučaju mogući objekt je aminokiselina, a traženo svojstvo bivanje mjestom interakcije. No, na temelju jedne značajke, u ovom slučaju imena aminokiseline teško je predvidjeti je li ona mjesto interakcije, ili nije. Iz tog razloga klasifikatori odluku donose na temelju niza značajki koje sadržavaju informacije vezane uz traženo svojstvo. Takve značajke nazivaju se diskriminatorne značajke, a njihov je odgovarajući izbor važan korak u izgradnji klasifikatorskog sustava. U našem slučaju diskriminatorne značajke predstavljaju aminokiseline koje okružuju ciljanu.

3 Algoritam slučajnih šuma

Algoritam slučajnih šuma (engl. *Random Forests*) je algoritam strojnog učenja koji se temelji na algoritmu stablo odlučivanja (eng. *Decision Tree*) [7]. Gradi se veći broj stabala odlučivanja, a odluka o pripadnosti razredu donosi se na temelju odluke većine stabala.

3.1 Stablo odlučivanja

Postupak izgradnje stabla odlučivanja rekurzivan je postupak, a stablo se od početnog čvora grana po različitim značajkama i njihovim vrijednostima. Grananje se zaustavlja onda kada se određeni skup (čvorovi u nizu) vrijednosti značajki može povezati s pripadnošću nekom razredu.

Ulazni skup je vektor je od N značajki, a izlaz je razred M kojem taj skup pripada. Prilikom izgradnje stabla koristi se skup od n uzoraka za učenje čiji je razred poznat. Koraci izgradnje stabla odluke su sljedeći:

1) U korijenu stabla čvor je koji sadrži sve uzorke iz skupa za učenje.

2) Ako svi uzorci iz skup promatranog čvora pripadaju istom razredu, vrati taj razred i ne razvijaj djecu.
3) Inače ako su sve ulazne vrijednosti iste vrati razred kojeg ima najviše i ne razvijaj djecu.
4) Inače skup uzoraka u promatranom čvoru dijeli se na podskupove određene vrijednostima značajke $\mathbf{N_i}$. $\mathbf{N_i}$ je pritom značajka koja nosi najveću količinu informacije.
5) Razvija se **k** novih čvorova iz promatranog čvora, **k** je broj različitih vrijednosti značajke $\mathbf{N_i}$ koje se javljaju u čvoru roditelju. Svaki čvor dijete poprima jednu od **k** vrijednosti i nasljeđuje one uzorke iz roditeljskog skupa koji imaju tu vrijednost značajke $\mathbf{N_i}$.
6) Korake 2) - 5) rekurzivno ponavljaj za svaki novi čvor.

Točnost klasifikatora ispituje se novim skupom podataka koji se spuštaju niz stablo. Ulazni skup uzoraka koji sadrži podatke pogrešne ili nebitne za odluku, može dovesti do izgradnje klasifikatora čija točnost neće biti najbolja moguća. Kako bi se izbjeglo uzimanje u obzir informacija koje predstavljaju šum, provodi se postupak podrezivanja stabla.

Podrezivanje stabla uklanja unutrašnje čvorove stabla i njegove pripadne nasljednike te ispituje točnost klasifikatora bez odrezanog dijela. Konačan oblik stabla onaj je koji daje najbolju točnost. Na ovaj način izbačene su sve nepotrebne odluke koje narušavaju točnost klasifikatora.

3.2 Algoritam slučajnih šuma

Algoritam slučajnih šuma gradi mnogo stabala odlučivanja..Skupovi za učenje svakog stabla su različiti, a dobivaju se tzv. „bootstrapping“ postupkom. Riječ je o statističkoj metodi koja se koristi za procjenu pogreške modela, a ideja je provesti učenje i ispitivanje u više koraka. Skup za učenje u svakom koraku dobiva se izborom podskupa uzoraka, pri čemu su dozvoljena ponavljanja elemenata. Preostali podaci i svakom koraku koriste se za ispitivanje modela. Svako stablo jedan je model, a konačna odluka donosi se većinom glasova, npr. ako su u šumi od

pet stabala, tri stabla odredila pripadnost uzorka razredu ***m1***, a dva pripadnost razredu ***m2***, konačna odluka šume je razred ***m1***.

Svako zasebno stablo u algoritmu slučajnih šuma gradi se sljedećim postupkom:

1. Skup uzoraka za učenje svakog stabla izabire se iz skupa svih uzoraka s mogućnošću ponavljanja nekog uzorka u više različitih skupova za učenje.
2. Odabire se broj **m**, znatno manji od ukupnog broja značajki **M**. Pri grananju svakog čvora uzima se u obzir **m** nasumičnih značajki među kojima se donosi odluka o značajki koja nosi najviše informacije i na temelju koje će se razviti novi čvorovi. Broj **m** jednak je za sve čvorove svih stabala u šumi.
3. Ne provodi se postupak podrezivanja stabla.

Skup uzoraka koji se izostavlja prilikom izgradnje svakog pojedinačnog stabla iznosi otprilike jednu trećinu. Ti se uzorci koriste kao ispitni skup. Prosječna vrijednost pogreške klasifikatora izračunata ovakvim ispitivanjem naziva se „*out of bag error*“ (OOB).

Pogreška klasifikatora izgrađenog algoritmom slučajnih šuma ovisi o:

- Korelaciji između svaka dva stabla - što je korelacija veća, veća je i pogreška klasifikatora.
- Snazi pojedinačnog stabla – što je pojedinačno stablo jače, veće točnosti to je i klasifikator bolji.

Izbor većeg broja značajki koje će se uzimati u obzir pri odluci o grananju čvora (m) povećava korelaciju između stabala, ali i snagu pojedinačnog stabla. Uravnoteživanje iznosa broja m doprinosi izgradnji optimalnog sustava za raspoznavanje uzoraka.

3.3 Svojstva algoritma slučajnih šuma

Metoda slučajnih šuma jedan je od najboljih poznatih algoritama za izgradnju klasifikatora, a posebno je pogodan za velike skupove podataka. To su dva osnovna

razloga zbog kojih je upravo ovaj algoritam korišten u prvom koraku raspoznavanja mjesta interakcije. Osim toga osobitosti algoritma slučajnih šuma su sljedeće:

- Daje procjenu o važnosti pojedinih značajki za klasifikaciju.
- Daje unutrašnju nepristranu procjenu o pogrešci klasifikatora.
- Čuva točnost i ukoliko je skup podataka nepotpun i neuravnotežen.
- Čuva izgrađene šume i računa prototipove koji se mogu koristiti u uspostavljanju odnosa između značajki i klasifikacije.
- Računa udaljenosti između svaka dva uzorka koje se mogu koristiti u nenadgledanom učenju (eng. *clustering*) i omogućuje eksperimentalno utvrđivanje interakcija pojedinih značajki.

4 Bayesov klasifikator

Bayesov klasifikator jednostavna je metoda strojnog učenja koja se temelji na statističkim podacima i koristi kod tzv. „učenja s učiteljem“ (eng. *supervised learning*). Riječ je o postupku izgradnje klasifikatora na temelju skupa za učenje koji sadrži uzorke čije su pripadnosti razredu poznate. Optimalna Bayesova metoda koja uzima u obzir sve uvjetne zavisnosti značajki predstavlja najbolju poznatu metodu raspoznavanja uzoraka . Njezina primjena je ograničena zbog računske zahtjevnosti algoritma, ali se uvijek može se koristiti kao standard za ocjenu uspješnosti drugih metoda. Pojednostavljene inačice ovog algoritma su Naivna Bayesova metoda koja pretpostavlja uvjetne nezavisnosti značajki i Bayesova mreža koja uzima u obzir samo neke zavisnosti značajki. Nakon Bayesovog teorema, dan je teoretski uvod u sve tri metode.

4.1 Bayesov teorem

Bayesov teorem važan je rezultat teorije vjerojatnosti koji povezuje uvjetne i potpune vjerojatnosti događaja, a u računarskoj se znanosti često koristi za izračunavanje aposteriornih vjerojatnosti događaja.

$P(A|B)$ vjerojatnost je događaja B ako je poznato da se ostvario događaj A i naziva se uvjetna vjerojatnost. Vjerojatnost umnoška zavisnih događaja A i B računa se formulom

$$P(AB) = P(A)P(B|A) \text{ i } (4.1)$$

$$P(AB) = P(B)P(A|B). \quad (4.2)$$

U slučaju nezavisnosti događaja A i B vrijedi

$$P(AB) = P(A)P(B) \quad (4.3)$$

Iz čega slijedi kako je za nezavisne događaje uvjetna vjerojatnost jednaka potpunoj vjerojatnosti

$$P(A) = P(A|B). \quad (4.4)$$

Iz relacija 2. i 3. izvodi se uvjetna vjerojatnost zavisnih događaja

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} \quad (4.5)$$

Potpuna vjerojatnost zavisnih događaja suma je svih zavisnih slučajeva. Za potpuni sustav događaja $\{B_1,...,B_n\}$ potpuna vjerojatnost događaja A iz skupa svih događaja Ω definira se kao

$$P(A) = \prod_{i=1}^{n} P(B_i)P(A|B_i) \quad (4.6)$$

Pretpostavimo postojanje skupa hipoteza $H = \{H_1,...,H_n\}$. Skup H podskup je skupa svih događaja Ω. Prije početka pokusa poznate su apriorne vjerojatnosti ostvarenja svake od hipoteza iz skupa H. Vjerojatnost događaja H_i označava se kao $P(H_i)$. Nakon pokusa poznato je koje su se hipoteze ostvarila, a koje nisu. Ako je ishod pokusa nepoznat, ali je poznato da se ostvario događaj A iz skupa svih događaja Ω mijenjaju se vjerojatnosti ostvarenja hipoteza iz skupa H. Na temelju formula 5. i 6. izvodi se Bayesova formula koja računa uvjetne vjerojatnosti svake od hipoteza H_i s obzirom na poznato ostvarenje događaja A

$$P(H_i \mid A) = \frac{P(H_i)P(A \mid H_i)}{\prod_{j=1}^{n} P(H_j)P(A \mid H_j)} \quad (4.7)$$

4.2 Naivni Bayesov klasifikator

Uzorak je predstavljen vektorom značajki $\vec{x} = (x_1,...,x_m)$, a svi razredi predstavljaju skup c. Događaji su poznate vrijednosti značajki, a hipoteze su pripadnosti razredima. Bayesov teorem kaže da se vjerojatnost pripadnosti uzorka razredu C, ako su poznate vrijednosti značajki uzorka $\vec{x}$, računa pomoću formule:

$$P(C \mid x_1 = X_1,...x_m = X_m) = \frac{P(C)P(x_1 = X_1,...x_m = X_m \mid C)}{P(x_1 = X_1,...x_m = X_m)} \quad (4.8)$$

Klasifikator se odlučuje za onu hipotezu čija je uvjetna vjerojatnost, s obzirom na poznate vrijednosti značajki, najveća. Riječ je o MAP (eng. *Maximum A Posteriori*) pravilu. Budući da je nazivnik neovisan o razredu C, izostavlja se u postupku odlučivanja. Kako bi smanjili složenost izračuna izraza $P(x_1 = X_1,...x_m = X_m \mid C)$, uvodimo naivnu pretpostavku o neovisnosti značajki

$$d(\vec{x}) = \max_{C \in c} \left(P(C) \prod_{i=1}^{m} P(x_i = X_i \mid C) \right). \quad (4.9)$$

Formula 8. sada postaje

$$P(C \mid x_1 = X_1, \ldots x_m = X_m) = P(C) \prod_{i=1}^{m} P(x_i = X_i \mid C) \quad (4.10)$$

A konačna funkcija odlučivanja sljedećeg je oblika

$$d(\vec{x}) = \max_{C \in c} \left(P(C) \prod_{i=1}^{m} P(x_i = X_i \mid C) \right). \quad (4.11)$$

Naivni Bayesov klasifikator po točnosti je približno jednak metodi Stabla odlučivanja, a omogućuje jednostavan izračun i radi s malo primjera za učenje. Budući da ne izbacuje hipoteze, već samo mijenja vjerojatnosti otporan je na šum.

Potrebna znanja za odlučivanje Bayesovom metodom su:

- Vjerojatnosti pojavljivanja razreda
- Vjerojatnosti pojavljivanja određenih vrijednosti značajki uz poznatu pripadnost razredu.

4.3 Optimalan Bayesov klasifikator

Naivni Bayesov klasifikator donosi odluku o najvjerojatnijoj hipotezi i toj hipotezu dodjeljuje vrijednost odluke. Optimalan Bayesov klasifikator uvodi novi skup svih razreda $V = \{v_1, \ldots, v_m\}$. Vjerojatnost ispravne klasifikacije uzroka $\vec{X}$ u razred v_i tada iznosi

$$P(v_i \mid \vec{X}) = \int_{C_j \in c} P(v_i \mid C_j) P(C_j \mid \vec{X}). \quad (4.12)$$

Izlaz klasifikatora je $\max_{v_i \in V} P(v_i \mid \vec{X})$.

Skup hipotezi u ovom slučaju ne mora odgovarati skupu mogućih klasifikacija. Optimalan Bayesov klasifikator prosječno je nenadmašiva metoda raspoznavanja uzoraka, ali

- Računski je složen i zahtjeva veliku količinu računalnih resursa.
- Potreban je velik ulazni skup podataka za proračun svih uvjetnih vjerojatnosti koje se koriste.

Iz navedenih razloga češće se koristi pojednostavljena inačica Bayesovog klasifikatora, opisana u prethodnom odjeljku, koja usprkos naivnim pretpostavkama polučuje izvrsne rezultate na stvarnim primjerima.

4.4 Bayesova mreža

Bayesove mreža inačica je Bayesovog klasifikatora koja uravnotežuje jednostavnost izračuna i točnost između naivnog i optimalnog Bayesovog klasifikatora. Prednost i „naivnost“ najjednostavnije inačice Bayesovog klasifikatora proizlazi iz pretpostavke o nezavisnosti diskriminatornih značajki svakog uzorka. Optimalan Bayesov klasifikator uzima u obzir sve međusobne zavisnosti značajki uzorka što znatno povećava složenost izračuna. Ukoliko je pretpostavka o nezavisnosti značajki točna, naivni i optimalni Bayes daju jednake rezultate, no ta pretpostavka u praksi najčešće nije točna.

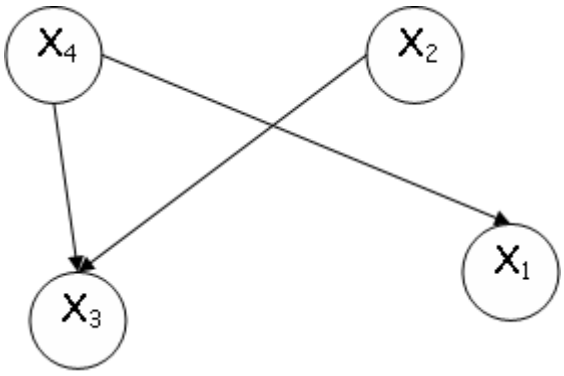
Bayesova mreža omogućuje pretpostavljanje uvjetnih zavisnosti i uvjetnih nezavisnosti među skupovima značajki. Pretpostavimo uvjetnu nezavisnost značajke x_1 i x_2 sa uvjetom x_4 uzorka $\vec{x} = (x_{1,\ldots}x_n)$. Ona se zapisuje na sljedeći način

$$\forall\, x_1 x_2 x_4 (P(x_1 \mid x_2, x_4) = P(x_1 \mid x_4). \quad (4.13)$$

Proširenjem na skupove značajki $X_1 = (x_i, x_j), X_2 = (x_k, x_l), X_4 = (x_m, x_n)$ dobiva se sljedeći izraz

$$\forall X_1 X_2 X_4 (P(X_1 \mid X_2, X_4) = P(X_1 \mid X_4) \quad (4.14)$$

U Bayesovoj mreži značajke se predstavljaju čvorovima. Sustavom putanja među čvorovima pretpostavljaju se uvjetne nezavisnosti, odnosno zavisnosti. Svaki čvor uvjetno je nezavisan od svih čvorova koji nisu njegovi roditelji uz uvjet roditelja. Uz svaki čvor veže se i razdioba vjerojatnosti pripadne značajke, uz uvjet poznatih vrijednosti roditeljskih značajki.



Slika 2. Primjer Bayesove mreže

Vjerojatnost pojave uzorka $\vec{X} = (x_1 = X_1, ..., x_n = X_n)$ u Bayesovoj mreži izračunava se pomoću sljedeće formule $P(\vec{X}) = \prod_{i=1}^{m} P(x_i = X_i \mid Roditelji(x_i))$

$$P(\vec{X}) = \prod_{i=1}^{m} P(x_i = X_i \mid Roditelji(x_i)) \quad (4.15)$$

U okviru ovog seminarskog rada korištena je upravo metoda Bayesove mreže za klasifikaciju uzoraka.

5 Klasifikator u dva koraka

Klasifikator izgrađen u okviru ovog rada raspoznaje uzorke u dva koraka, odnosno sastoji se od dva klasifikatorska podsustava.

Ulazni uzorak predstavljen je vektorom od devet značajki, a izlaz je razred uzorka. Radi se zapravo o prozoru od devet uzastopnih aminokiselina u lancu od kojih je svaka aminokiselina predstavljena svojom oznakom. Budući da postoji dvadeset različitih aminokiselina, svaka značajka može poprimiti jednu od dvadeset različitih imenovanih vrijednosti. Imena pridružena vrijednostima slova su koja predstavljaju skraćeni naziv aminokiseline.

Tablica 1. Puni i skraćeni nazivi aminokiselina

Aminokiselina	Troslovna skraćenica	Aminokiselina	Troslovna skraćenica
Alanin	**ALA**	Valin	**VAL**
Isoleucin	**ILE**	Leucin	**LEU**
Metionin	**MET**	Fenilalanin	**PHE**
Tirozin	**TYR**	Triptofan	**TRP**
Lizin	**LYS**	Arginin	**ARG**
Aspartat	**ASP**	Glutamat	**GLU**
Histidin	**HIS**	Glutamin	**GLN**
Aspargin	**ASN**	Serin	**SER**
Treonin	**THR**	Cistein	**CYS**
Glicin	**GLY**	Prolin	**PRO**

Izlaz klasifikatora je razred središnje, pete aminokiseline. Rečeno je već kako razred zapravo označuje je li aminokiselina mjesto interakcije, ili ne. Stoga postoje dva različita razreda:

- Aminokiselina **jest** mjesto interakcije – označava se s oznakom **1**.
- Aminokiselina **nije** mjesto interakcije – označava se s oznakom **0**.

Slika 3. Blok shema klasifikatorskog sustava

5.1 Klasifikatorski podsustavi

U prvom koraku vektor od devet aminokiselina prolazi kroz klasifikator, a izlaz iz klasifikatora je predviđanje razreda središnje aminokiseline.

ALA,ILE,ASP,TYR,LYS,VAL,SER,ILE,PHE -> 0

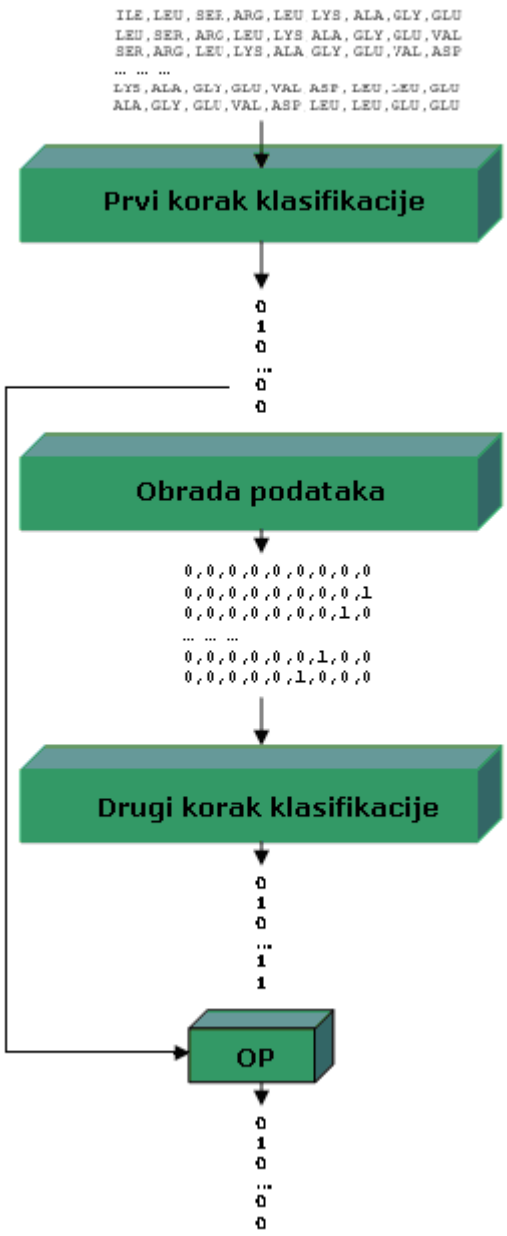
Ulaz u drugi klasifikator vektor je od osam značajki x[8] koje predstavljaju razrede okolnih aminokiselina, a izlaz je predviđanje razreda središnje aminokiseline.

0,1,1,1,0,0,0,1 -> 0

Izlaz prvog klasifikatora potrebno je preslikati u uzorke pogodne za ulaz u drugi klasifikator. To se čini tako da se svaki prozor u nizu jednog cjelovitog lanca zamijeni klasom središnje aminokiseline (izlaz prvog klasifikatora). Time je dobiven lanac klasa. Zatim se prozor od devet aminokiselina pomiče duž tog lanca i u svakom se koraku kao uzorak uzima osam aminokiselina koje okružuju središnju aminokiselinu u prozoru. Važno je primijetiti da prva i zadnja četiri prozora u svakom odsječku neće biti moguće popuniti jer prvi klasifikator nije obavio predviđanje za prve i zadnje četiri aminokiseline u lancu. Njima se pretpostavlja vrijednost 0 jer je veća apriori vjerojatnost pojave razreda 0.

Također očito je da se klasifikacija u dva koraka može obavljati samo nad nizom uzoraka koji predstavljaju jedan lanac, a ne po jedan uzorak. To i odgovara svrsi jer se analiziraju proteinski lanci, a ne zasebne aminokiseline.

Ako se takav lanac sastoji od *n* aminokiselina, tada je ulazni skup klasifikatora *n* prozora od devet aminokiselina.



Slika 4. Sustav dva klasifikatora

5.2 Metode klasifikacije

Ispitivano je nekoliko metoda klasifikacije koje su ostvarene u operacijom OP na slici Slika 4. Jedna metoda je izravno propuštanje izlaza drugog klasifikatora. Ona je ostvarena u dvije inačice:

1. Kada prvi klasifikator koristi metodu slučajnih šuma, a drugi Bayesove mreže – **RF-BY**
2. Kada i prvi i drugi klasifikator koristi metodu slučajnih šuma – **RF-RF**.

Druga je **OR** metoda koja izvodi „ILI“ operaciju nad izlazima prvog i drugog klasifikatora. Opet su ispitivane dvije inačice:

3. Kada prvi klasifikator koristi metodu slučajnih šuma, a drugi Bayesove mreže – **RForBY**
4. Kada i prvi i drugi klasifikator koristi metodu slučajnih šuma – **RForRF**.

Zadnja je **AND** metoda koja izvodi „I“ operaciju nad izlazima prvog i drugog klasifikatora. Dvije ispitivane inačice su:

5. Kada prvi klasifikator koristi metodu slučajnih šuma, a drugi Bayesove mreže – **RFandBY**
6. Kada i prvi i drugi klasifikator koristi metodu slučajnih šuma – **RFandRF**.

5.3 Zašto klasifikacija u dva koraka

Pretpostavka je da će klasifikator u dva koraka davati bolje rezultate nego samo jedan metoda predviđanja zato što se na ovaj način uzimaju u obzir dva različita svojstva mjesta interakcije:

- Postoje sličnosti u sastavu aminokiselina koje stupaju u slične interakcije – to je povezano s kemijsko - fizikalnim svojstvima tih aminokiselina koje se u sličnom okružju iskazuju na sličan način.
- Aminokiseline tvore nakupine u lancu – mjesta interakcije u proteinskom često se nalaze u grupama što se može objasniti pozicijama koje su izloženije i time je vjerojatnije da će stupiti u kemijsku vezu s drugim lancem.

Prvi korak analize lanaca uzima u obzir sličnost aminokiselinskog sastava u okolini mjesta interakcije, a drugi uzima u obzir raspored pozicija mjesta interakcije. Ispitivanja Yana, Dobbsa i Honavara [1] pokazala su da kod 97% mjesta interakcije okolina koju čine po četiri aminokiseline sa svake strane sadrži još barem jedno mjesto interakcije. Kod 70% mjesta interakcije ta okolina sadrži još najmanje četiri mjesta interakcije.

Drugi korak klasifikacije ne može se izvoditi samostalno, niti može primiti izravno podatke jer kod analize sljedova poznat je samo njihov aminokiselinski sastav, ali ne i raspored mjesta interakcije – on se zapravo traži. Uz pomoć prvog klasifikatora pretpostave se moguća mjesta interakcije na temelju sljedova aminokiselina, a

pomoću drugog klasifikatora ta se predviđanja dodatno poprave koristeći činjenicu da mjesta interakcije čine nakupine u lancu. Taj je podatak ugrađen u klasifikator prilikom njegovog učenja.

6 Metode

6.1 Priprema podataka

Za predviđanje mjesta proteinskih interakcija korišten je skup neredundantnih podataka koji se sastoji od 1500 lanaca 333 različita proteinska kompleksa. Podaci se temelje na zapisima proteinskih struktura PDB (eng. Protein Data Bank) bazi, a izrađeni su prethodno [2] i zapisani su u ARFF formatu. Priprema skupa podataka na kojima je obavljena analiza uključivala je izlučivanje bitnih značajki. Konačni oblik ARFF zapisa definiran je značajkama:

```
@ATTRIBUTE PDBId  STRING
@ATTRIBUTE lanac   STRING
@ATTRIBUTE residue1
{ALA,CYS,ASP,GLU,PHE,GLY,HIS,ILE,LYS,LEU,MET,ASN,PRO,GLN,ARG,SER,THR,VAL,TRP,
TYR}
…
@ATTRIBUTE residue9
{ALA,CYS,ASP,GLU,PHE,GLY,HIS,ILE,LYS,LEU,MET,ASN,PRO,GLN,ARG,SER,THR,VAL,TRP,
TYR}
@ATTRIBUTE pozicija NUMERIC
@ATTRIBUTE class {0,1}
```

Primjer jednog zapisa:

```
1A0O,A,ALA,ASP,LYS,GLU,LEU,LYS,PHE,LEU,VAL,6,0
```

Riječ je o prozoru devet aminokiselina iz lanca A proteina 1A0O. Središnja aminokiselina LEU šesta je aminokiselina u lancu i za nju je poznato da nije mjesto interakcije (0).

Prilikom pripreme podataka za učenje i testiranje klasifikatorskog sustava prvo su određeni svi neprekinuti odsječci lanaca. To su oni prozori koji su dobiveni pomicanjem duž jednog te istog lanca za po jednu aminokiselinu.

Takvi su odsječci korišteni kao nedjeljive jedinice prilikom podjele skupa podataka na podskupove za učenje i ispitivanje. Podskupovi su izabrani tako da budu podjednake veličine i da učestalost mjesta interakcije ostane približno jednaka ukupnom iznosu. Konačno su dobivena tri podskupa:

Tablica 2. Podjela podataka na podskupove

	Ukupno aminokiselina	Udio mjesta interakcije
Cijeli skup podataka	154680	0,260
Podskup1	51921	0,263
Podskup2	49533	0,263
Podskup3	53226	0,254

Udio mjesta interakcije pokazuje da je skup podataka neuravnotežen, odnosno da učestalost pojave razreda nije podjednaka, već je razred „mjesto interakcije“ - 1 otprilike četiri puta rjeđi od razreda „nije mjesto interakcije“ – 0.

Svaki je podskup stvoren u dvije inačice: jedna za učenje i testiranje RF klasifikatora, a druga za učenje i testiranje Bayesovog klasifikatora. Značajke prve inačice podataka slične su već navedenima, osim što su izuzeti atributi *PDBId*, *lanac* i *pozicija*, a umetnute su oznake početka i kraja pojedinog odsječka. Značajke druge inačice nešto su drukčije:

```
@ATTRIBUTE class1 {0,1}
@ATTRIBUTE class2 {0,1}
@ATTRIBUTE class3 {0,1}
@ATTRIBUTE class4 {0,1}
@ATTRIBUTE class  {0,1}
@ATTRIBUTE class6 {0,1}
@ATTRIBUTE class7 {0,1}
@ATTRIBUTE class8 {0,1}
@ATTRIBUTE class9 {0,1}
```

Razred zapisa je u ovom slučaju atribut *class*.

Na temelju ova tri skupa izgrađeni su skupovi za učenje Bayesovog i RF klasifikatora te skupovi za testiranje klasifikatora prema „leave-out“ principu. To znači da su u jednom koraku analize kao skup za učenje korištena dva podskupa, a skup za testiranje činio je izostavljeni, treći podskup. Skupovi za učenje Bayesovog i RF klasifikatora slični su po sastavu. Razlika je u tome što su kod učenja Bayesovog klasifikatora izostavljeni prozori koji sadrže aminokiseline nepoznate klase. To su prva četiri i zadnja četiri prozora u svakom punom odsječku.

6.2 Ispitivanje metoda

Prilikom ispitivanja metoda korištena su tri alata WEKA, R i PARF. R i PARF korišteni su za izvedbu algoritma slučajnih šuma jer podržavaju izgradnju šuma sa preko 100 stabala, dok WEKA ima prevelike memorijske zahtjeve te je iskoristiva za izgradnju do desetak stabala. WEKA je korištena za izgradnju i testiranje Bayesovog klasifikatora.

Obrada podataka koji se dobiveni kao izlaz iz prvog klasifikatora izvedena je u jeziku C#, a uključuje formiranje lanaca predviđenih klasa za svaki odsječak, izdvajanje prozora i pretpostavljanje vrijednosti nepoznatih klasa prva i zadnja četiri prozora svakog odsječka. Te su vrijednosti pretpostavljene na 0 jer je veća apriorna vjerojatnost pojave klase 0.

Nakon što je obavljen i drugi korak klasifikacije testnog skupa primjenjuje se *AND* i *OR* operacija na rezultate prvog i drugog klasifikatora. Te su metode također izvedene u programskom jeziku C#.

Korištena je BayesNet metoda u WEKA alatu i građene su slučajne šume od 100 stabala uz pomoću PARF programa.

Procjene uspješnosti obavljene su za više metoda:

- Za klasifikator slučajnih šuma
- Za Bayesov klasifikator
- Za kombinaciju klasifikatora u dvije faze
    - Klasifikator slučajnih šuma u oba koraka
    - Klasifikator slučajnih šuma u prvom i Bayesov klasifikator u drugom koraku

- AND operacija nad izlazima prvog i drugog koraka nad klasifikatorima u dva koraka
    - Klasifikator slučajnih šuma u oba koraka
    - Klasifikator slučajnih šuma u prvom i Bayesov klasifikator u drugom koraku
- OR operacija nad izlazima prvog i drugog koraka nad klasifikatorima u dva koraka
    - Klasifikator slučajnih šuma u oba koraka
    - Klasifikator slučajnih šuma u prvom i Bayesov klasifikator u drugom koraku

6.3 Mjere uspješnosti

Prilikom analize rezultata korištena je matrica greške (eng. *confusion matrix*) koja se sastoji od broja točno i netočno predviđenih uzoraka za svaki razred.

Tablica 3. Confusion matrica

TP (eng. true positives) predviđena mjesta interakcije koja to zaista jesu	**FP (eng. false positives)** predviđena mjesta interakcije koja to zapravo nisu
FN (eng. false negatives) predviđena mjesta ne-interakcije koja su zapravo mjesta interakcije	**TN (eng. true negatives)** predviđena mjesta ne-interakcije koja stvarno nisu mjesta interakcije

Mjere korištene za ocjenu rezultata su:

$$točnost = \frac{TP + TN}{TP + TN + FP + FN} \quad (6.1)$$

$$preciznost = \frac{TP}{TP + FP} \quad (6.2)$$

$$odziv = \frac{TP}{TP + FN} \quad (6.3)$$

$$F - mjera = \frac{2 \cdot (preciznost \cdot odziv)}{preciznost + odziv} \quad (6.4)$$

Točnost odražava koliko je ukupno aminokiselina točno klasificirano. Preciznost pokazuje koliko je aminokiselina klasificiranih kao mjesta interakcije dobro klasificirano, a odziv koliko je mjesta interakcije od ukupnog broja prepoznato. Preciznost i odziv mogu se računati za oba razreda (0 i 1), a ovdje smo ga računali samo za razred 1.

F-mjera je težinska harmonijska srednja vrijednost preciznosti i odziva. Uvodi se zato što kod neuravnoteženih skupova, kao što je naš (Tablica 2), točnost nije dobra mjera uspješnosti jer teži prema uspješnosti prepoznavanja većinske klase. Preciznost, odziv i F-mjera bolji su pokazatelji ukupne uspješnosti jer su osjetljivi na uspješnosti prepoznavanja oba razreda.

6.4 Grafovi

Osim numeričkih rezultata, prikazani su i *preciznost - odziv* grafovi. Crtani u programskom alatu R na temelju vjerojatnosti pojave klase 1 za svaki prozor. Za metode koje su rezultat davale izravno iz prvog ili drugog klasifikatora (RF, RF-BY, RF-RF) gotove su vjerojatnosti preuzete iz PARF [6] i WEKA [8] alata. Za AND i OR metode bilo je potrebno naknadno izračunati konačne vjerojatnosti što je učinjeno prema sljedećim algoritmima:

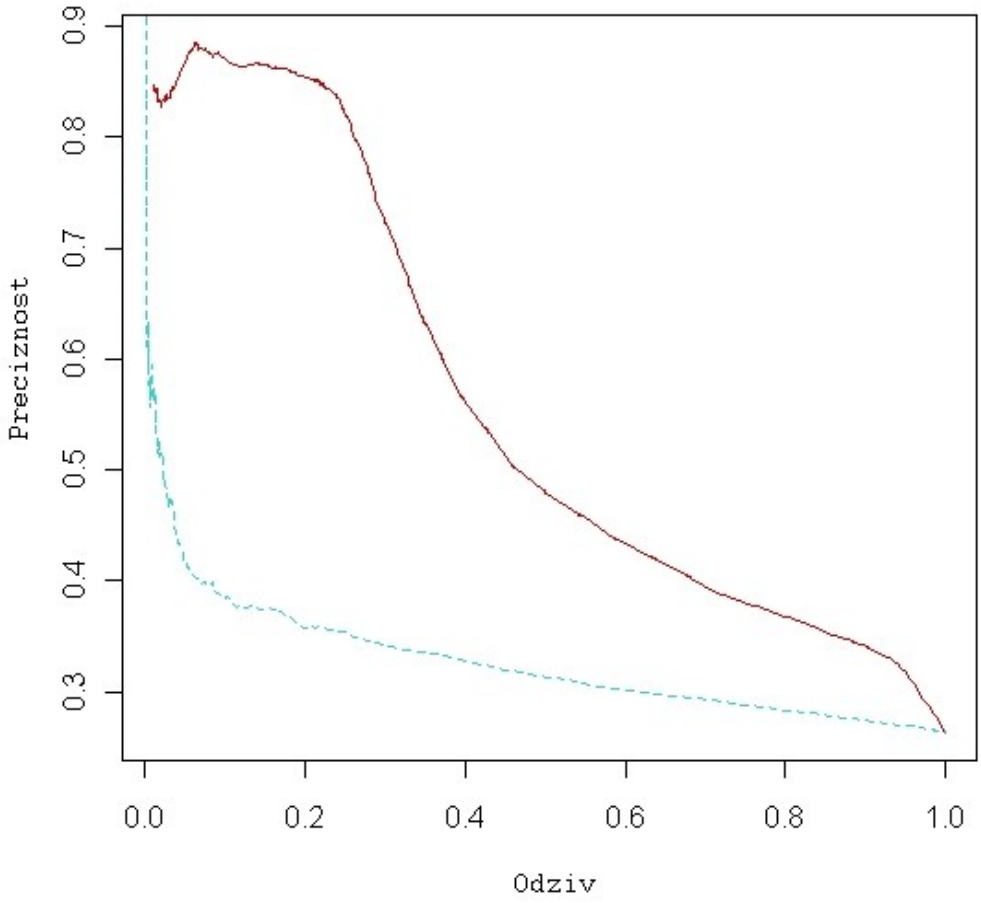
OR METODA

```
Klasa1 = izlaz_prvog_klasifikatora
Klasa2 = izlaz_drugog_klasifikatora
Izlazna_klasa

Za svaku Klasa1 klasu:
        ako je (Klasa2 > 0.5)
                Izlazna_klasa  = Klasa1 + Klasa2 - Klasa1*Klasa2
      inače
              Izlazna_klasa  = Klasa1
```

AND METODA

```
Klasa1 = izlaz_prvog_klasifikatora
Klasa2 = izlaz_drugog_klasifikatora
Izlazna_klasa

Za svaku Klasa1 klasu:
        ako je (Klasa2 < 0.5 && Klasa1 < 0.5)
                Izlazna_klasa  = Klasa1*Klasa2
      inače ako je ((Klasa1 > 0.5 && Klasa2 < 0.5) || (Klasa1 < 0.5 && Klasa2 >
0.5))
              Izlazna_klasa  = Klasa1 * Klasa2 + 0.25
      inače ako je (Klasa2 > 0 && Klasa1 > 0)
                Izlazna_klasa  = Klasa1*Klasa2 + 0.5
```

7 Rezultati

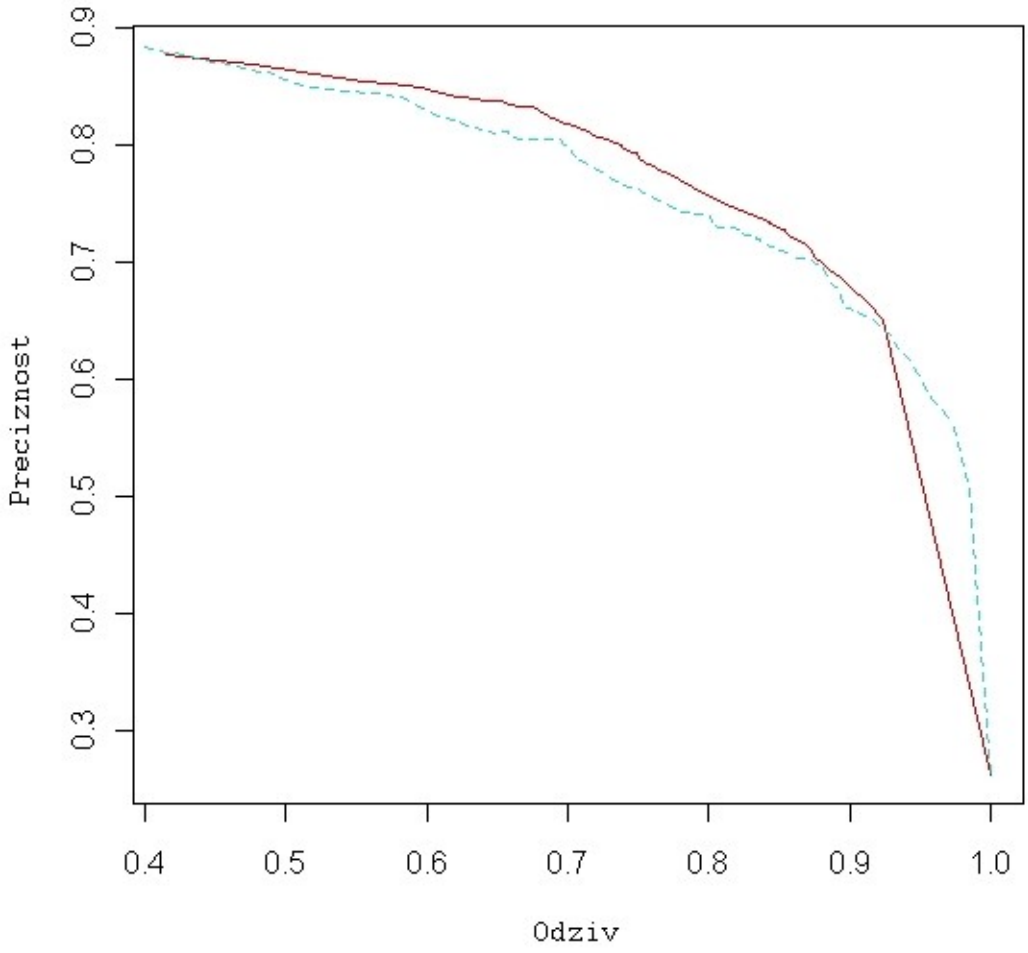


Slika 5. Preciznost-odziv graf za Bayes(plavo) i RF(crveno) klasifikaciju nad sekvencama aminokiselina

Tablica 4. Rezultati testiranja nad nizovima aminokiselina

	RF	**BY**
Točnost	0.780	0.737
Preciznost	0.667	0.591
Odziv	0.331	0.008
F-mjera	0.442	0.015

Slika 6. prikazuje Preciznost-odziv graf za Bayesov i RF klasifikator u jednom koraku. Ulazni podaci prozori su aminokiselina. Očito je da RF metoda postiže bolje rezultate od Bayes metode. Ukupna točnost je za oko 5% bolja od Bayesa, ali F-mjera je bitno bolja, tj. Bayes rezultate dobre postiže na temelju klasifikacije većinske klase, a manjinska klasa je pritom zanemarena. RF postiže bolju preciznost kod klasifikacije manjinske klase i bitno bolji odziv(broj prepoznatih mjesta interakcije u odnosu na njihov ukupan broj). Ovaj rezultat pokazuje kako je smisleno koristiti samo RF klasifikator u prvom koraku klasifikacije jer ga Bayes ne može nadmašiti.

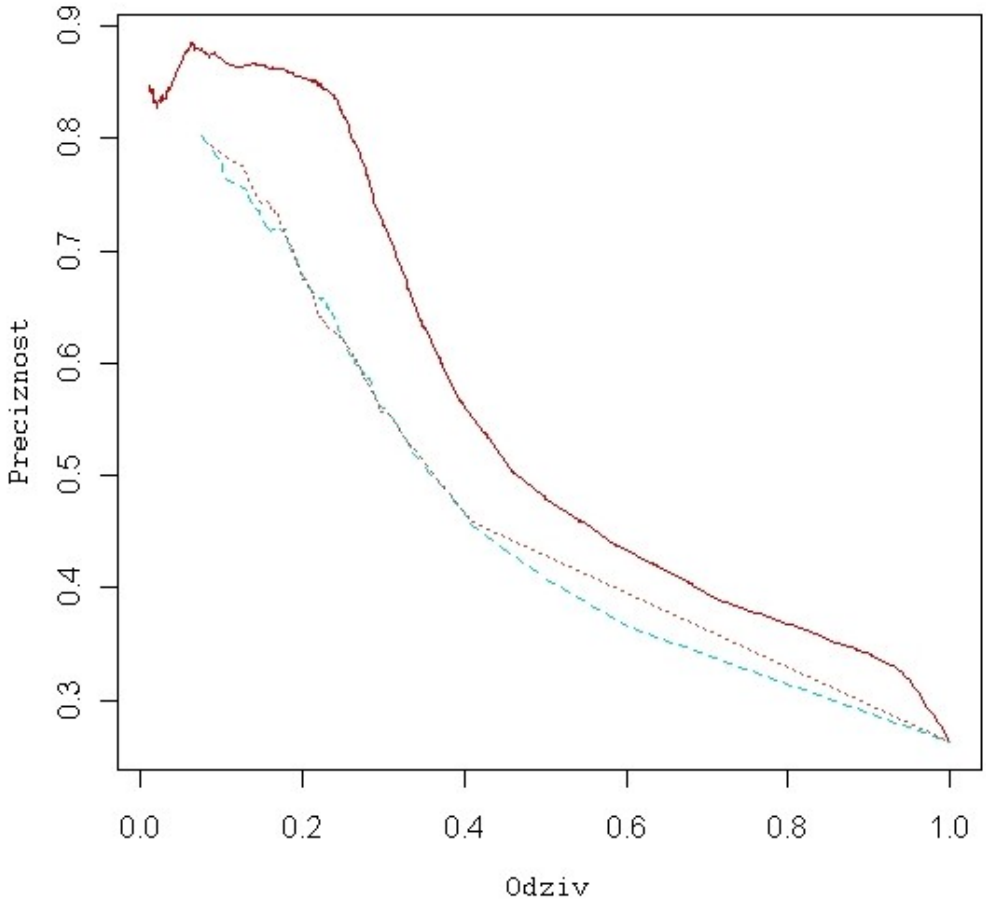


Slika 6. Usporedba Bayes (plavo) i RF (crveno) klasifikacije nad nizovima klasa

Tablica 5. Rezultati testiranja nad nizovima klasa

	RF	BY
Točnost	0,793	0.872
Preciznost	0,793	0.728
Odziv	0,747	0.821
F-mjera	0.770	0.772

Slika 7. prikazuje rezultate Bayes i RF klasifikacije za klasifikaciju u jednom koraku kada su ulazni podaci prozori klasa, odnosno kada na temelju poznatih razreda okoline (0,1) predviđamo razred središnje aminokiseline. Bayesove mreže prepoznaju uzorke dosta točnije od RF klasifikatora, ali taj je rezultat, kao i u prethodnom ispitivanju, vezan uz prepoznavanje većinske klase. Iz F-mjere vidljivo je da su rezultati za manjinsku klasu bliski za oba klasifikatora, s tim da je Bayes nešto malo bolji. Na temelju ovih rezultata dolazimo do zaključka da kod ispitivanja u dva koraka, u drugom koraku ima smisla koristiti i je dan i drugi klasifikator.

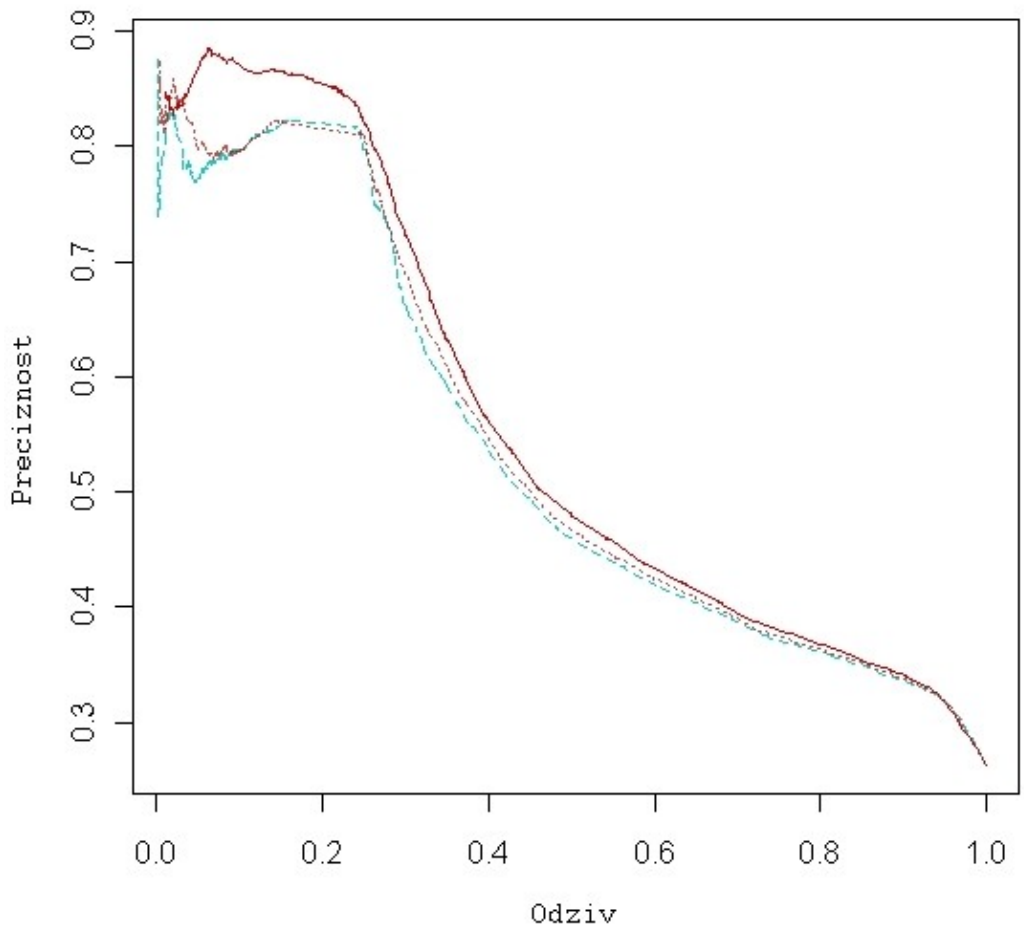


Slika 7. Usporedba Bayes (plavo) i RF (ružičasto crtkano) klasifikacije u drugom koraku s RF klasifikatorom

Tablica 6. Rezultati testiranja modela u dva koraka

	RF	RF-RF	RF-BY
Točnost	0.780	0,764	0,764
Preciznost	0.667	0,669	0.637
Odziv	0.331	0,206	0.241
F-mjera	0.442	0,315	0.350

Slika 8. Prikazuje usporedbu rezultata za RF metodu u jednom koraku i klasifikatora u dva koraka. Plava crtkana linija klasifikator je u dva koraka s Bayes metodom u drugom koraku, a ružičasta crtkana, klasifikator u dva koraka s RF metodom u drugom koraku. U oba slučaja u prvom koraku koristi se RF metoda. Očito je RF klasifikator neusporedivo bolji od obje metode u dva koraka i što se ukupne točnosti tiče i što se F-mjere (preciznost/odziv manjinske klase). U usporedbi RF-RF i RF-BY metoda ova druga je nešto bolja.

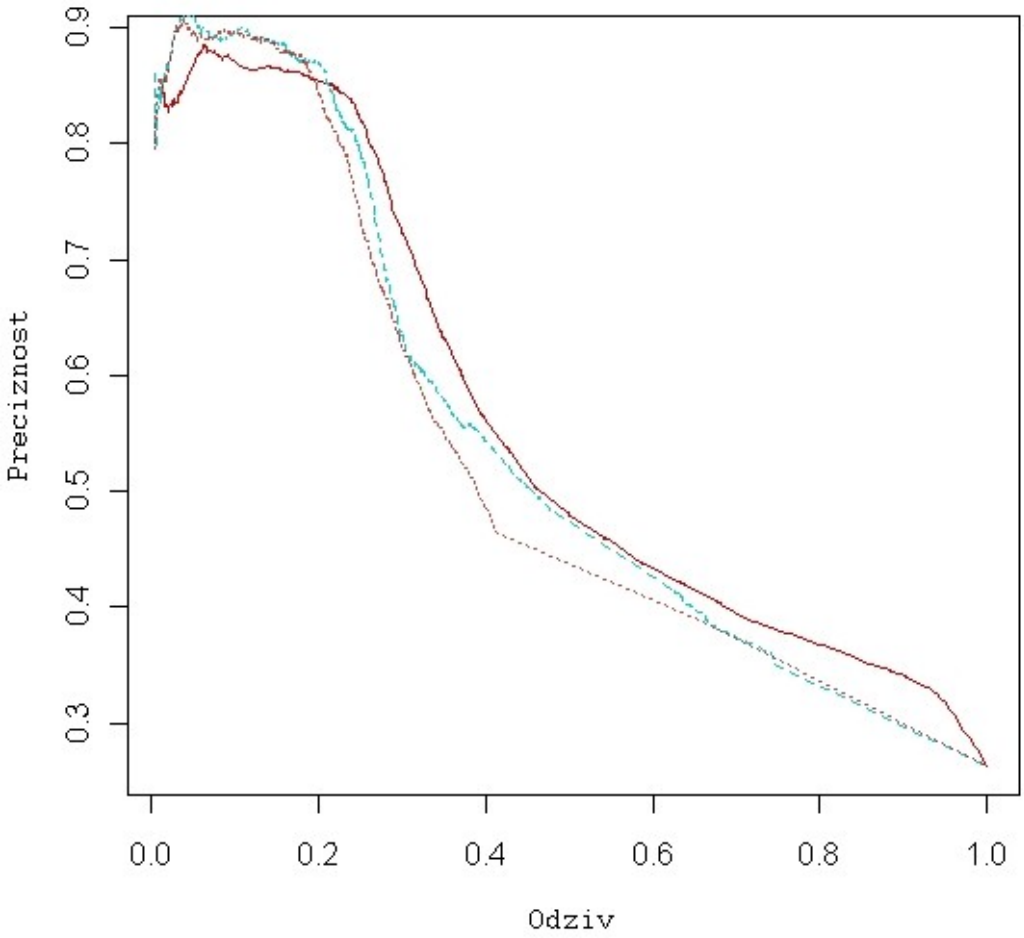


Slika 8. Usporedba Bayes (plavo) i RF (ružičasto crtkano) klasifikacije u OR metodi s RF klasifikatorom

Tablica 7. Rezultati testiranja OR metoda

	RF	**RForRF**	**RForBY**
Točnost	0.780	0,767	0,762
Preciznost	0.667	0,598	0.576
Odziv	0.331	0,354	0.363
F-mjera	0.442	0,445	0,445

Na slici Slika 9. uspoređene su RForRF i RForBY metode s RF klasifikatorom. Rezultati su bolji od RF-RF i RF-BY klasifikatora, no još uvijek nisu nadmašili RF klasifikator. Za visoke odzive preciznost se obaju OR metoda približava preciznosti RF klasifikatora, no za niske odzive preciznost značajno opada.



Slika 9. Usporedba Bayes (plavo) i RF (ružičasto crtkano) klasifikacije u AND metodi s RF klasifikatorom

Tablica 8. Rezultati testiranja AND metoda

	RF	**RFandRF**	**RFandBY**
Točnost	0.780	0,777	0.782
Preciznost	0.667	0,861	0.855
Odziv	0.331	0,183	0.209
F-mjera	0.442	0,302	0.336

Na slici Slika 10. vidljivi su rezultati AND metoda u odnosu na RF klasifikator. Ukupna točnost RFandBY metode malo je bolja od RF klasifikatora, ali F-mjera pokazuje da je to posljedica prepoznavanja većinske klase. Prepoznavanje manjinske klase i dalje pokazuje kako je RF klasifikator bolji. Ipak, poboljšanje preciznosti postoji za vrijednosti odziva između 0.02 do 0.22. AND metode pokazuju lošije rezultate na područjima visokog odziva.

8 Diskusija

Rezultati su pokazali kako je RF metoda u jednom koraku i dalje najbolja računalna metoda prepoznavanja mjesta interakcije. Iako Bayes daje vrlo visoko prepoznavanje razreda (točnost = 0.872, F-mjera = 0.772) na temelju informacija o razredima okoline taj doprinos nije uspio popraviti rezultate. Neuspjeh je vjerojatno posljedica toga što RF nije ponudio dovoljno dobre rezultate na temelju kojih je Bayes mogao predvidjeti klasu.

Ideja klasifikacije u dva koraka preuzeta je iz rada Yana, i ostalih [1] u kojem su korišteni SVM (eng. Support Vector Machines) klasifikator i Bayesove mreže. Točnost SVM klasifikatora od 0.66 i F-mjera od 0.44 poboljšane su na iznose 0.72 i na 0.47. Pritom nisu korišteni samo lanci aminokiselina, već i podaci iz strukture (eng. Accessible Surface Area - ASA) što je omogućilo dobivanje boljih rezultata..

Uspješnosti Gallet metode [5] koja se temelji na analizi hidrofobnosti aminokiselina iznose 0.51 (točnost) i 0.36 (F-mjera). Orfan i Rost [4] su kombinacijom informacija iz strukture (ASA, sekundarna struktura i evolucijski profili) i sekvenci pomoću neuronskih mreža postigli preciznosti između 0.6 i 0.7 za odzive iznad 0.1.

Očito su, ovdje ispitane, RF i metode predviđanja u dva koraka bolje od ostalih metoda, no uspjeh klasifikatora u dva koraka izravna je posljedica uspješnosti RF metode. Ona je zasad najbolja računalna metoda predviđanja mjesta interakcije. Ukoliko se podigne uspješnost RF klasifikatora u prvom koraku, moguće je da će primjena Bayesa u drugom koraku dovesti do još boljih rezultata.

9 Zaključak

Predviđanje mjesta interakcije na temelju nizova aminokiselina važno je područje bioinformatike jer doprinosi bržoj analizi struktura i funkcija proteinskih kompleksa. Poznavanje funkcija i struktura proteina važno je za razumijevanje bioloških procesa u živom organizmu, a iskoristivo je i u industriji lijekova.

Broj poznatih sekvenci već je daleko veći i raste brže nego što eksperimentalne metode analize mogu pratiti. Zato se javlja snažna potreba za razvojem računalnih metoda. U analizi računalnih metoda, uspješnost predviđanja nije dobra mjera jer se radi o nebalansiranim skupovima, odnosno mjesta interakcije čine otprilike četvrtinu aminokiselina u proteinu. Bolja mjera je F-mjera koja predstavlja usrednjenu vrijednost odziva i preciznosti manjinske klase. Najbolji rezultati u okviru ovog rada dobiveni su klasifikacijom algoritmom slučajnih šuma i iznose oko 0.66.

Kao mogućnost poboljšanja nameće se klasifikacija Bayesovim klasifikatorom u drugom koraku. Bayes naime daje F-mjeru u iznosu od 0.772 kada se primjeni na nizove klasa. Nažalost poboljšanje nije dobiveno nad ovim RF klasifikatorom, što je vjerojatno posljedica toga što RF bez obzira što je trenutno najbolji klasifikator za predviđanje mjesta interackijce ne daje dovoljno točne izlazne klase na temelju kojih bi Bayes mogao predvidjeti konačan izlaz.

10 Literatura

[1] Yan C., Dobbs D. , Honavar V. , A two-stage classifier for identification of protein-protein interface residues, Bioinformatics, Vol. 20, 2004., i371-i378.

[2] Šikić M., Računalna metoda za predviđanje mjesta proteinskih interakcija, doktorska disertacija, 2008, (rad u procesu ocjenjivanja)

[3] Y. Ofran and B. Rost, "Predicted protein-protein interaction sites from local sequence information," *FEBS Lett,* vol. 544, pp. 236-9, 5.6.2003.

[4] Y. Ofran and B. Rost, "ISIS: interaction sites identified from sequence," *Bioinformatics,* vol. 23, pp. e13-6, 15.1.2007.

[5] X. Gallet, B. Charloteaux, A. Thomas, and R. Brasseur, "A fast method to predict protein interaction sites from sequences," *J Mol Biol,* vol. 302, pp. 917-26, 29. 9. 2000.

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

[7] L. Breiman and A. Cutler, "Random Forests.", vol.2007.
http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm

[8] Weka 3: Data Mining Software in Java, datum pristupa 1.5.2007.
http://www.cs.waikato.ac.nz/~ml/weka/index.html

Predviđanje mjesta proteinskih interakcija iz sekvence aminokiselina

Predicting protein-protein interface residues based on amino acid sequences

Sažetak

U okviru ovog rada ispitane su računalne metode predviđanja mjesta interakcija u proteinskim lancima koje se temelje na klasifikacijskim algoritmima.

Proteini su biološke molekule koje imaju važnu ulogu u odvijanju staničnih funkcija. Njihova je uloga određena njihovom strukturom. Zbog velikog broja poznatih proteinskih lanaca čije strukture i funkcije nisu poznate potrebno je razviti metode koje će ih moći predvidjeti na temelju poznatih informacija – aminokiselinskog sastava. Eksperimentalne metode analize ne mogu pratiti brzi rast baze poznatih lanaca pa se javlja potreba za razvojem bržih, računalnih metoda.

Uz metode koje strukture i mjesta interakcije proteinskih lanaca predviđaju na temelju fizikalnih i kemijskih svojstava aminokiselina postoje i metode raspoznavanja uzoraka koje predviđanje obavljaju na temelju samih aminokiselinskih sljedova.

Dva algoritma raspoznavanja uzoraka ispitana u okviru ovog rada su Bayesove mreže i algoritam slučajnih šuma. Ispitane su pojedinačne uspješnosti algoritama nad

- uzorcima koji su predstavljeni prozorima od devet aminokiselina u nizu.
- uzorcima predstavljenim prozorima koji sadrže poznate klase osam aminokiselina koje okružuju ciljanu.

Pokazalo se da algoritam slučajnih šuma u prvom slučaju značajno nadmašuje Bayesov algoritam s F-mjerom od 0.66 dok su u drugom koraku uspješnosti podjednake, s malom prednošću Bayesa čija F-mjera iznosi 0.77.

Također ispitane su metode koje koriste dva algoritma raspoznavanja uzoraka uzastopce. Na izlazna predviđanja jednog algoritma primjenjuje se drugi algoritam. Nad takvim podacima izvođene su operacije *AND* i *OR*, a ispitana su i izravna

izlazna predviđanja drugog klasifikatora. Pokazalo se da niti jedna od ovih metoda ne nadmašuje sam algoritam slučajnih šuma. Algoritmi čiji je konačan izlaz jednak izlazu drugog klasifikatora bitno su lošiji od algoritma slučajnih šuma, dok mu se *AND* i *OR* algoritmi približavaju. *AND* algoritam za niske vrijednosti odziva manjinskog razreda (prepoznavanje mjesta interakcije) daje čak i višu preciznost od algoritma slučajnih šuma.

Pretpostavka je da je poboljšanje korištenjem Bayesovog algoritma u drugom koraku moguće, ali da predviđanja algoritma slučajnih šuma nisu dovoljno točna te da ruše uspješnost Bayesa u drugom koraku.

Summary

Protein–protein interactions play a central role in life processes. Structural information on these interactions is important for understanding their mechanisms and implications. The number of experimentally determined, primarily by X-ray crystallography, structures of protein–protein complexes is steadily growing. However, the crystal structures of protein–protein complexes are generally more difficult to obtain than those of the individual complexes. Thus, computational approaches are important as a source of protein-protein complexes structures and as a means to understand the principles of protein association.

We proposed and tested several methods for predicting protein-protein interaction residues. Methods are based on Random Forest (RF) and Bayesian Network classification algorithms. Tests showed that Bayesian Network classifier recognizes correctly around 72% of interface residues and the value of its F-measure is around 0.78. That is when patterns are represented as sequences of 8 known classes of residues that are surrounding the target residue.

When patterns are slide windows of 9 amino acids in a sequence, RF performs the best. Its F – measure value is 0.44. Other tested classifiers combine these two methods, with RF in the first step and Bayesian Network or RF in the second step.

It showed that neither output of the second classifier, nor the OR/AND combinations of outputs of both classifiers outperform one-step RF. However, AND combination of RF and Bayesian classifier (RFandBY) did achieve higher precisions than the RF classifier for low values of recall. The presumption is that RF still does not produce as correct and precise output as needed by Bayesian classifier to raise the final

results. Thus, if predictions in the first step could be improved, two stage classifying methods might show even more improvement in the second step.

Ključne riječi

- proteinske interakcije
- mjesto proteinske interakcije
- aminokiselinski sljedovi
- klasifikacija
- metoda slučajnih šuma
- Bayesov klasifikator
- neuravnotežen skup podataka
- F-mjera
- preciznost – odziv

Keywords

- protein – protein interactions
- interface residue
- amino acid sequences
- classification
- Radnom Forest
- Bayes classifier
- imbalanced data
- F-measure
- precision - recall