UNIVERZITET U BEOGRADU
BIOLOŠKI FAKULTET

Neven S. Šumonja

BIOINFORMATIČKI MODELI ZA AUTOMATSKO MAPIRANJE INTERAKCIJA IZMEĐU PROTEINA KOD ČOVEKA

doktorska disertacija

Beograd, 2019

UNIVERSITY OF BELGRADE
FACULTY OF BIOLOGY

Neven S. Šumonja

BIOINFORMATICS MODELS FOR AUTOMATIC PREDICTION OF HUMAN PROTEIN-PROTEIN INTERACTION

Doctoral Dissertation

Belgrade, 2019

Mentori:

dr Vladimir Perović
naučni saradnik Instituta za nuklearne nauke Vinča Univerziteta u Beogradu

dr Dušanka Savić Pavićević
redovni profesor Biološkog fakulteta Univerziteta u Beogradu

Komisija:

dr Dušanka Savić Pavićević
redovni profesor Biološkog fakulteta Univerziteta u Beogradu

dr Vladimir Perović
naučni saradnik Instituta za nuklearne nauke Vinča Univerziteta u Beogradu

dr Nevena Veljković
naučni savetnik Instituta za nuklearne nauke Vinča Univerziteta u Beogradu

Datum odbrane:

Zahvalnica

Ovaj rad je urađen u Centru za multidisciplinarna istraživanja i inženjerstvo Instituta za nuklearne nauke Vinča Univerziteta u Beogradu u okviru projekta Ministarstva prosvete, nauke i tehnološkog razvoja ON173001: „Primena EIIP/ISM bioinformatičke platforme u otkrivanju novih terapeutskih targeta i potencijalnih terapeutskih molekula“.

Najveću zahvalnost zelim da izrazim mojim mentorima dr Vladimiru Peroviću i dr Neveni Veljković za punu pomoć i podršku u svim aspektima izrade ove teze, od teorijskih postavki eksperimenata, izvođenja istih, razumevanje rezultata i pisanju teze. Završetak ove studije je omogućilo njihovo ogromno lično, strpljivo i posvećeno zalaganje i briga.

Dr Vladimiru Peroviću pripada moja ogromna zahvalnost u svim aspektima mog celokupnog rada u sklopu i van okvira ove teze, a posebno u tehničkom izvođenju eksperimenata i dizajnu softvera. Hvala na prijateljskoj i bratskoj moralnoj podršci, praktičnoj pomoći i savetima, zalaganju, strpljenju i poverenju. Hvala mnogo na svemu.

Posebno želim da zahvalim dr Neveni Veljković na preciznim i pravovremenim komentarima i usmeravanjima u pravcu planiranja i primene istraživanja provedenih u okviru ove teze, kao i svim drugim aspektma rada na ovoj disertaciji. Hvala i na svoj pomoći i zalaganju u svim aspektima mog rada u laboratoriji. Hvala mnogo.

Hvala i mojoj mentorki dr Dušanki Savić Pavićević profesorki Biološkog fakulteta Univerziteta u Beogradu za komentare i usmeravanje oko pisanja disertacije.

Hvala svim mojim kolegama iz laboratorije, posebno dr Branislavi Gemović, dr Rajku Davidoviću i dr Milanu Senćanskom i ostalima na strpljenju, pomoći oko izrade teze, novim idejama, savetima i konstruktivnim razgovorima koji su pomogli sprovođenje istraživanja u okviru ove teze.

Posebno hvala dr Veljku Veljkoviću za mogućnost da sarađujem sa laboratorijom kao i za korisne savete.

Najveća hvala mojoj porodici u prvom redu mojim roditeljima Slobodanu i Jovanci. Posebno hvala mom ocu Slobodanu na bezgraničnoj podršci u svakom smislu. Bez Vas roditelji, ne bih mogao ni zamisliti da završim ovu tezu. Posebno hvala i stricu Branku i dedi Dušanu kao i ostalim članovima moje familije.

Hvala dr Katarini Bačević na velikoj pomoći u pisanju teze kao i celokupnoj podršci.

Hvala mom prijatelju Olegu Lesmajsteru, nesuđenom biologu, na pomoći oko pisanja teze i na velikoj prijateljskoj podršci.

BIOINFORMATIČKI MODELI ZA AUTOMATSKO MAPIRANJE INTERAKCIJA IZMEĐU PROTEINA KOD ČOVEKA

Rezime

Interakcije između bioloških makromolekula imaju ključnu ulogu u osnovnim procesima u živim organizmima, posreduju u metaboličkim putevima, putevima prenosa signala, transkripciji, translaciji i drugim ćelijskim i sistemskim procesima. Veliki broj oboljenja uzrokovan je mutacijama proteina u regionima odgovornim za interakciju sa drugim proteinima koje mogu dovesti do ometanja interakcije protein-DNK, promene u obrascima savijanja proteina, novih nepoželjnih interakcija i omogućavanje interakcije protein-patogen. Mapiranje interaktoma, odnosno kompletne mape interakcija protein-protein (IPP) unutar organizma, je od suštinske važnosti za razumevanje kompleksnih molekularnih odnosa unutar živih sistema, kao i za rasvetljavanje raznih patoloških stanja ljudskog organizma. Bioinformatičke metode za automatsko predviđanje IPP, kao suplementi eksperimentalnim metodama za analizu IPP, omogućavaju bolje razumevanje bioloških procesa i funkcija, lakše otkrivanje potencijalnih meta za ciljanu terapiju i smanjenja vremena i troškova razvoja novih terapeutika. U ovoj studiji razvijeni su modeli i metode za automatsko predviđanje IPP bazirane na mašinskom učenju i proteinskoj sekvenci, koja predstavlja univerzalnu, visoko kvalitetnu i eksperimentalno potvrđenu informaciju o proteinu. Generisani su modeli za predviđanje IPP za specijalne slučajeve: (i) između transkripcionih regulatora, odnosno proteina koji učestvuju u kompleksnom procesu transkripcione regulacije koji kontroliše ekspresiju gena i značajan je za normalnu fiziologiju ćelije, i (ii) proteina sa neuređenom tercijarnom strukturom, koji su kao takvi uključeni u ključne biološke procese interakcijom sa višestrukim partnerima, imaju fleksibilnu strukturu, višestruke funkcije, centralnu ulogu u regulaciji signalnih puteva, procesu prepoznavanja i vezivanja za male molekule, i čine većinu proteina povezanih sa neprenosivim bolestima. Pored toga, kreirane su tri nove vrste atributa za predstavljanje proteina: (i) atributi zasnovani na primarnoj strukturi proteina, (ii) evolutivni atributi i (iii) mrežni atributi, kao i metode bazirane na genetskom algoritmu za (i) automatsko generisanje i selekciju atributa i (ii) za automatsko formiranje

i optimizaciju ansambla modela zasnovanim na mašinskom učenju, u svrhu proširenja prostora atributa i povećanja efikasnosti predviđanja IPP. Kao glavni rezultat studije, na osnovu toga, razvijen je opšti model HP-GAS za automatsko mapiranje IPP na nivou proteoma čoveka. Opsežna evaluacija i poređenje sa *state-of-the-art* metodama pokazali su da HP-GAS predstavlja trenutno najefikasniji metod za mapiranje IPP na nivou proteoma čoveka, sa efikasnošću predviđanja AUC= 0.93 i 0.85 tačnosti.

Ključne reči: interakcije protein-protein, proteom čoveka, proteinske sekvence, mašinsko učenje
Naučna oblast: Molekularna biologija
Uža naučna oblast: Bioinformatika
UDK broj: [57.087+57.088.6]:577.112(043.3)

BIOINFORMATICS MODELS FOR AUTOMATIC PREDICTION OF HUMAN PROTEIN-PROTEIN INTERACTION

Abstract

Interactions between biological macromolecules have a critical role in essential processes in living organisms, mediate the metabolic pathways, signaling pathways, transcription, translation and other cellular processes and systems. A large number of diseases caused by mutations in the regions of the protein responsible for the interactions with other proteins which can lead to interference of protein-DNA interaction, changes in the patterns of protein folding, new undesirable interaction and facilitate interaction of the protein-pathogen. Interactome mapping, i.e. mapping of complete network of protein-protein interactions (PPI) within the organism, is essential to an understanding of complex molecular relationships within the living system, as well as to elucidate the various human pathological conditions. Bioinformatics methods for automated PPI prediction, as addition to experimental methods for the analysis of PPI, allow a better understanding of biological processes and functions, easier detection of potential therapeutic targets, and reduce the time and cost of drug development. In this study, models have been developed and methods for the automated prediction of PPI based on machine learning and the protein sequence, which is a universal, high-quality and experimentally confirmed information on the protein. Models for the PPI prediction were generated for special cases: (i) between human transcriptional regulators, i.e. the proteins involved in the complex process of transcriptional regulation that controls the gene expression and they are important for normal cell physiology, and (ii) intrinsically disorder proteins, characterized by the lack of a fixed tertiary structure, which are as such involved in the regulation of key biological processes via binding to multiple protein partners, are malleable adapting to structurally different partners, have multiple functions, play a central roles in the regulation of signaling pathways, the process of molecular recognition and binding of small molecule, and are the prevailing protein class associated with noncommunicable diseases. In addition, three novel types of features for the representation of the proteins were created: (i) the features based on the protein sequence, (ii) the evolutionary features, and (iii) the graph features, as well as methods based on the

genetic algorithm for (i) automatic feature-engineering process and (ii) automatic ensembling of different machine learning algorithms, in order to expand the feature space and to improve the PPI prediction performance. Based on that, as a main result of the study, the general model HP-GAS for automatic mapping of PPIs at the human proteome level was created. Extensive evaluation and comparison with the state-of-the-art methods, show that the HP-GAS represents currently the most efficient method for proteome-wide forecasting of protein interactions, with prediction efficacy of AUC = 0.93 and 0.85 accuracy. HP-GAS method and PPI prediction models for special cases were implemented as free, time-efficient and easy-to-use software tools.

Key words: protein–protein interactions, human proteome, protein sequence, machine learning
Scientific field: Molecular biology
Scientific discipline: Bioinformatics
UDC number: [57.087+57.088.6]:577.112(043.3)

Spisak skraćenica

OTF - opšti transkripcioni faktori
LCR - kontrolni regioni lokusa (engl. *locus control region*)
TF - transkripcioni faktori
IPP - interakcija protein-protein
PNTS - proteini neuređene tercijarne strukture
RNTS - regioni neuređene tercijarne strukture
PDB - engl. *Protein Data Bank*
PUTS - proteini uređene tercijarne strukture
GO - genske ontologije (engl. *Gene Ontology*)
SSM - metode male skale (eng. *Small Scale Methods*)
LSM - metode velike skale (eng. *Large Scale Methods*)
IP - imunoprecipitacija
co-IP - engl. *complex-immunoprecipitation*
WB - engl. *Western Blot*
MS - Masena spektroskopija
SDS-PAGE - engl. *sodium dodecyl sulfate-polyacrylamide gel electrophoresis*
PLA - metoda ligacije usled blizine (engl. *Proximity Ligation Assay*)
FRET - Prenos fluorescentne rezonantne energije (engl. *Fluorescence Resonance Energy Transfer*)
BiFC - Bimolekularna fluorescentna komplementacija (engl. *Bimolecular fluorescence complementation*)
2D PAGE - 2D-poliakrilamidna gel elektroforeza
IEF - izolelektrično fokusiranje
AFM - engl. *Atomic Force Microscopy*
NMR - nuklearna magnetna rezonanca
TAP - engl. *Tandem Affinity Purification*
Y2H - kvaščev dvohibridni sistem (engl. *Yeast Two-Hybrid)*
BD - engl. *DNK binding domain*
AD - engl. *activating domain*
IGP - imidazol-glicerofosfat
X-gal - bromohloroindoksil-galaktozid
TEV - engl. *Tobacco Etch Virus*
CBR - engl. *Calmodulin-Binding Protein*
IMEx - engl. *International Molecular Exchange*
HIPPIE - engl. *Human Integrated Protein-Protein Interaction rEference*
HPRD - engl. *Human Protein reference Database*
MIPS - engl. *Munich information centre for protein sequences*
MPIDB - engl. *Microbial Protein Interaction database*
MINT - engl. *Molecular Interaction*
BIND - engl. Biomolecular Interaction Network Database
BLAST - engl. *Basic Local Alignment Search Tool*
AAC - minokiselinska kompozicija (engl. *Amino Acid Composition)*
DC - dipeptidna kompozicija (engl. *Dipeptide Composition*)

TC - tripeptidna kompozicija (engl. *Tripeptide composition*)
PSSM - engl. *Position Specific Scoring Matrix*
PAAC - engl. *Pseudo-Amino Acid Composition*
ML - mašinsko učenje (engl. *Machine learning*)
CV - unakrsna validacija - (engl. *cross-validation*)
RF - Nasumične šume (engl. *Random Forest*)
NIPP - skup parova proteina koji ne interaguju
SVM - Metod potpornih vektora (engl. *Support Vector Machine*)
kNB - K-najbliži sused (engl. *k-Nearest Neighbors*)
ACC - Auto kros-kovarijansa (eng. *auto cross-covariance function*)
PSI-BLAST - engl. *Position-Specific Iterative Basic Local Alignment Search Tool*
SP - pozitivni IPP
SN - stvarno negativni IPP
LN - lažno negativnom IPP
LP - lažno pozitivnom IPP
SSN - specifičnost (stopa stvarno negativnih)
SLP - stopa lažno pozitivnih
SLN - stopa lažno negativnih
MCC - Matejev koeficijent korelacije (engl. *Matthews correlation coefficient*)
ROC - engl. *Receiver Operator Curve*
AUROC – Površina ispod "receiver operator curve" (engl. *Area Under Receiver Operator Curve)*
PRC - engl. *Precision recall curve*
AUPRC – Površina ispod "precision recall curve" (engl. *Area Under Precision recall curve)*
PCA - Analiza glavnih komponenti (engl. *Principal component analysis*)
GA - Genetski algoritam
NB - Naivni Bejesov klasifikator (engl. *Naive Bayes*)
DNN – Duboke neuronske mreže (engl. *Deep neural network*)
DL - Duboko učenje (engl. *Deep learning*)
GPU - engl. *graphics processing unit*
GLM - engl. *Generalized linear model*
GBM - Metode gradijentnog pojačavanja (engl. *Gradient boosting machines*)
XGBoost - engl. *Extreme Gradient boosting*

Sadržaj

1 Uvod

1.1 **Interakcije protein-protein**

1.1.1 Proteom čoveka

Osnovna paradigma molekularne biologije je da je protein proizvod ekspresije informacije koja se nasleđuje u formi nukleinskih kiselina. Proteini čine preko 50% suve mase ćelija i ima ih više od bilo kojih makromolekula u ćeliji (Milo, 2013). Dok se procenjeni broj ljudskih protein-kodirajućih gena kreće od oko 20.000 do 22.000 (Salzberg, 2018), pretpostavlja se da je broj različitih proteina bar pet puta veći (Ponomarenko *et al.*, 2016). Rezultati detekcije proteina u ćeliji biohemijskim metodama ukazuju da postoji 150.000 različitih proteina koje ljudska ćelija može eksprimirati (Kessel and Ben-Tal, 2011). Proteini se nalaze u pozadini svih procesa u biološkom sistemu. Među značajnim ulogama proteina se ubrajaju: kataliza hemijskih reakcija (enzimske reakcije), energetski transfer, genska ekspresija, transport kroz ćelijske membrane, ćelijska komunikacija i molekulsko prepoznavanje. Nadalje, proteini vrše ključne funkcije u imunskom sistemu i ćelijskoj mreži signalnih puteva, formiranju intraćelijskih i ekstraćelijskih struktura, te sintezi novih molekula i njihovoj degradaciji, itd.

Strukturno, proteini su kompleksni makromolekuli sa visoko uređenom organizacijom.

Primarna struktura proteina je prvi nivo organizacije proteina i predstavlja sekvencu aminokiselina polipeptidnog lanca. Linearna sekvenca aminokiselina se formira kovalentnim vezivanjem aminokiselina preko peptidnih veza u linearni polipeptidni lanac. Biofizičke karakteristike aminokiselina koje ulaze u sastav proteina zavise od karakteristika bočnog lanca aminokiselina. Sekvenca proteina se sastoji od kombinacije 20 različitih aminokiselina, a svaki protein je definisan pomoću unikatne sekvence aminokiselinskih ostataka. Svi ostali nivoi organizacije proteina se oslanjanju na njegovu primarnu strukturu, a samim tim i funkcija proteina.

Sekundarna struktura proteina se formira u zavisnosti od lokalnih konformacija polipeptidnog lanca i prostornih odnosa između aminokiselinskih ostataka. Sekundarnu strukuru proteina čine alfa zavojnice i beta ploče.

Tercijarna struktura uključuje savijeni (engl. folding) polipetidni lanac, a definisana je kao prostorno (trodimenzionalno) uređenje aminokiselinskih ostataka.

Mnogi proteini sadrže više od jednog polipeptidnog lanca, a interakcija između ovih lanaca je osnova kvarterne strukture proteina (Whitford, 2005; Kessel and Ben-Tal, 2011).

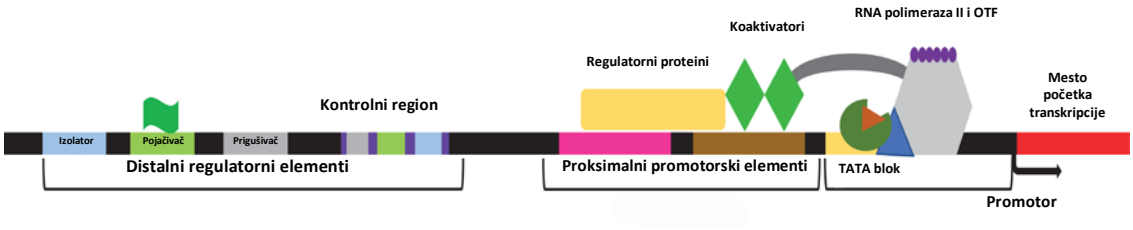
1.1.2 Transkripcioni faktori

Izvršavanje svih bioloških procesa u ćeliji poput rasta, proliferacije, apoptoze, starenja i diferencijacije zahteva skup precizno orkestriranih koraka koji prostorno i vremenski zavise od ekspresije gena (Maston, Evans and Green, 2006). Ekspresija eukariotskih, protein-kodirajućih gena može biti regulisana na dva međusobno povezana nivoa: (i) nivo transkripcione mašinerije i transkripcionih faktora i (ii) nivo hromatina i njegovih regulatora (Lee and Young, 2013). Veruje se da se najveći stepen regulacije odvija na prvom nivou i to posebno kod inicijacije transkripcije (Maston, Evans and Green, 2006). Transkripcija kod eukariota se izvršava posredstvom RNK polimeraze II, potpomognuta sa nekoliko opštih transkripcionih faktora (OTF), koji zajedno formiraju preinicijacioni kompleks. OTF uključuju TFIIA, TFIIB, TFIID, TFIIE, TFIIF i TFIIH proteine. Ključni uslov inicijacije transkripcije jeste formiranje preinicijacijonog kompleksa na promotoru, nekodirajućoj sekvenci DNK, sa kojom elementi preinicijacijonog kompleksa interaguju. Promotor-preinicijacijski kompleks je dovoljan za bazalni nivo transkripcije (Thomas and Chiang, 2008). Pored promotora drugi nekodirajući elementi DNK koji imaju važnu ulogu u regulaciji ekspresije gena nazivaju se *cis* regulatornim elementima. Regulatorni elementi u blizini gena, uzvodno od osnovnog promotora se označavaju kao proksimalni, dok distalni regulatorni elementi uključuju pojačivače (engl. enchancers), prigušivače (engl. silencers), izolatore (engl. insulators) i kontrolne regione lokusa (engl. locus control region, LCR) (Maston, Evans and Green, 2006; Coulon *et al.*, 2013; Hawkins, Al-attar and Storey, 2018) (Slika 1).

Ključni proteini transkripcione regulacije ekspresije gena se označavaju kao transkripcioni faktori (TF). U grupu transkripcionih faktora spadaju OTF i ostali proteini sposobni da se vežu za DNK regulatorne elemente i da deluju uslovno u svojstvu aktivatora ili represora (Lelli, Slattery and Mann, 2012). Pored TF ključnu ulogu u transkripcionoj regulaciji imaju koregulatori (Spiegelman and Heinrich, 2004). Koregulatori nemaju sposobnost direktnog fizičkog vezivanja za DNK, već svoju regulatornu funkciju ostvaruju interakcijama sa drugim regulatornim proteinima modifikujući njihovu aktivnost, i na taj način pozitivno ili negativno regulišu transkripciju (Lemon and Tjian, 2000; Maston, Evans and Green, 2006). Tokom regulacije transkripcije ostvaruje se sinergistički efekat više transkripcionih faktora preko međusobnog kooperativnog vezivanja, što predstavlja direktnu vezu između regulacije transkripcije i mnogobrojnih međusobnih interakcija protein-protein (IPP) (Lelli, Slattery and Mann, 2012).

Fino podešavanje kompleksnog sistema transkripcije u najvećoj meri zavisi od fizičkih interakcija između regulatornih proteina (Lee and Young, 2013). Međusobne interakcije transkripcionih regulatora dovode do promene u specifičnosti vezivanja za DNK, kao i do promena u afinitetu vezivanja za DNK oba člana kompleksa, što je pokazano na primeru interakcije između HOX i EXD proteina (Chan *et al.*, 1996), te ELK1/SRF kompleksa (Mo *et al.*, 2001). Ovakvi efekti navedenih IPP su posledica reorganizacije motiva za vezivanje za DNK i supresije autoinhibicije. Osim latentne specifičnosti vezivanja, gde proteini poput HOX proteina menjanju specifičnost vezivanja za DNK, formirajući IPP sa manjim brojem transkripcionih faktora, određeni transkripcioni faktori pokazuju znatno veću promiskutetnost u vezivanju koregulatora (Lelli, Slattery and Mann, 2012). Neki članovi SOX familije proteina interaguju sa više od 40 koregulatora u procesu finog naštimavanja specifičnosti vezivanja za DNK (Bernard and Harley, 2010). Ovakvo podešavanje sposobnosti vezivanja za DNK u zavisnosti od IPP omogućava preciznu tkivno-specifičnu kontrolu ekspresije u zavisnosti od dostupnosti koregulatora (Lelli, Slattery and Mann, 2012). Detaljno analiziran primer kooperativnog vezivanja regulatornih proteina je ekspresija ß-interferona, u kojoj se pet proteina vezuje za 55 baznih parova dugačak segment DNK, formirajući pojačivački kompleks (engl. enchanceosome) (Panne, 2008).

Slika 1. Regulatorni elemenata DNK kod eukariota (Hawkins, Al-attar and Storey, 2018).



Usled izrazite važnosti za regulaciju transkripcije, mutacije transkripcionih faktora se smatraju jednim od ključnih uzročnika brojnih oboljenja (Ulasov, Rosenkranz and Sobolev, 2018). Procenjuje se da je trećina poremećaja u razvoju kod čoveka uslovljena mutacijama transkripcionih faktora (Lee and Young, 2013). Vaquerizas i ostali (Vaquerizas *et al.*, 2009) su identifikovali 164 transkripciona faktora koji su povezani sa 277 monogenskih bolesti. Kao okosnica velikog broja signalnih puteva, promene u funkcionalnosti transkripcionih faktora mogu biti uzrok kompleksnih oboljenja poput: kancera, kardiovaskularnih bolesti, neuroloških, autoimunih i inflamatornih poremećaja, dijabetesa, neplodnosti i gojaznosti (Ulasov, Rosenkranz and Sobolev, 2018). Vremenski i prostorno koordinisane aktivnosti transkripcionih faktora, uslovljene su mrežom međusobnih interakcija. Mnogi transkripcioni faktori deluju kao ključni čvorovi mreže (engl. hubs), formirajući tkivno i vremenski specifične obrasce genske ekspresije regrutacijom velikog broja regulatornih proteina preko IPP (Francois, Donovan and Fontaine, 2018).

Pored značajnog uticaja na sve ćelijske procese, farmakološka modulacija transkripcionih faktora je veliki izazov (Arndt, 2006; Li *et al.*, 2014; Fontaine *et al.*, 2017). Izazovi upotrebe transkripcionih faktora kao meta za lekove su povezani sa opservacijama da modulacija pojedinih transkripcionih faktora u ćeliji može da dovede do značajnih promena na nivou ekspresije stotina ciljnih gena (Fontaine, Overman and François, 2015). Pored ovoga, transkripcioni faktori su funkcionalno redundantni. Na modelu kvasca (*Saccharomyces cerevisiae*) je pokazano da poremećaji funkcionalnosti transkripcionih faktora utiče samo na 3% gena koje regulišu (Wu and Lai, 2015).

Postoje indicije o visokom stepenu robusnosti transkripcione regulacije kod sisara. Ona se ogleda u nekoliko nivoa mehanizama zaštite koji omogućavaju funkcionalnost sistema usled remećenja ključnih elemenata transkripcione regulacije (Fontaine, Overman and

François, 2015). Pored toga, konvencionalni terapeutski pristup koji se oslanja na model „ključa i brave“ nije primenjiv na transkripcione faktore (Lazo and Sharlow, 2016). Pored farmakomodulacije pojedinačnih transkripcionih faktora, jedan od predloženih pravaca otkrivanja i razvoja novih terapeutika zahteva identifikaciju i karakterizaciju kompletne mreže IPP transkripcionih regulatora na nivou proteinskog interaktoma. To bi omogućilo ciljanu modulaciju većih delova mreže IPP transkripcionih regulatora, utičući tako na veći broj faktora transkripcione regulacije ciljnih gena u različitim ćelijskim stanjima (Francois, Donovan and Fontaine, 2018). Jedna od teškoća kod primene ovakvog pristupa jeste neophodnost značajnih materijalnih i vremenskih resursa za detekciju celokupne mreže IPP transkripcionih regulatora. Sa druge strane, potvrda direktne fizičke interakcije nekog transkripcionog faktora, koja bi poslužila za ciljanu terapiju, često zahteva dva ili više eksperimentalna pristupa (Francois, Donovan and Fontaine, 2018).

1.1.3 Humani proteini neuređene tercijarne strukture

Širok spektar funkcija koje transkripcioni regulatori vrše u ćeliji i visok stepen promiskuiteta u formiranju IPP, ukazivao je na potrebu za revizijom dogme da specifična funkcija proteina zahteva jedinstvenu trodimenzionalnu strukturu (Tsafou *et al.*, 2018; Uversky, 2018) Identifikacija proteina neuređene tercijarne strukture (PNTS) i proteina koji poseduju regione neuređene tercijarne strukture (RNTS) ukazala je na mogućnost stvaranja veze između biološke aktivnosti ovakvih proteina i stanja neuređenosti njihove trodimenizonalne strukture (Dunker and Uversky, 2010). PNTS proteini su definisani kao grupa proteina različitih strukturnih konformacija, koji su u stanju konstantne strukturne promene (Dunker *et al.*, 2013). Sa druge strane, mnogi proteini poseduju mozaičnu organizaciju uređenih i RNTS regiona. Ovakva hibridna kompozicija se smatra ključnim za multifunkcionalnost ovih proteina (Dunker *et al.*, 2013).

Kod PNTS i proteina sa RNTS uočena je izuzetna strukturna kompleksnost koja se ogleda u različitim nivoima i stepenu neuređenosti, na nivou manjih, većih regiona proteina ili na nivou čitave sekvence (Uversky, 2016, 2017). PNTS nemaju sposobnost samostalnog savijanja u stabilnu 3D strukturu. Veći stepen uređenosti dostižu samo kroz fizičku interakciju sa specifičnim proteinskim partnerom. Heterogena struktura PNTS i proteina sa RNTS dozvoljava postojanje delova proteina u različitom stepenu savijanja;

od spontano savijenih regiona, nesavijenih regiona, delimično savijenih regiona koji su konstantno u tom stanju, do delimično savijenih samo usled interakcija sa partnerom, kao i onih koji su povremeno nesavijeni da bi se izvršila određena proteinska funkcija. Zatim, odnos, distribucija i stepen neuređenosti različitih proteina sa RNTS se konstantno menjaju u vremenu i prostoru, kao i u zavisnosti od funkcionalnih potreba proteina (Uversky, 2018).

Transkripcioni faktori su među prvim proteinima kod kojih su aktivnosti u fiziološkim uslovima objašnjene specifičnostima u strukturi (Wright and Dyson, 2015). Odsustvo opisanih 3D struktura transkripcionih faktora u proteinskoj bazi podataka (engl. Protein Data Bank, PDB) ukazivalo je na činjenicu da veliki broj transkripcionih faktora zapravo pripada PNTS (Minezaki *et al.*, 2006). Kompjuterske analize pokazale su da od 82.63% do 94.13% eukariotskih transkripcionih faktora poseduje duže proteinske sekvence sa RNTS. Kod ljudi je identifikovan 401 transkripcioni faktor, od čega proteini sa RTNS zauzimaju 49% ukupne sekvence (Dunker and Uversky, 2010). Specifičnost strukture transkripcionih faktora ogleda se u učestaloj uređenosti regiona za vezivanje DNK i velikom broju RNTS van tih regiona (Sammak and Zinzalla, 2015). Sa druge strane, neuređenost velikih delova 3D strukture transkripcionih faktora omogućava izloženost relativno velike površine proteina, koje postaju lako dostupne za IPP i posttranslacione modifikacije. Kao rezultat povećane izloženosti, transkripcioni faktori imaju sposobnost visoko-specifičnog vezivanja za odgovarajuće regulatorne regione DNK, dok istovremeno ostvaruju preciznu interakciju sa velikim brojem proteinskih partnera, ali sa srednjim afinitetom vezivanja. Ovo dovodi do spontane disocijacije i brže terminacije interakcije (Rogers *et al.*, 2014). Na osnovu nekoliko kompjuterskih studija došlo se do zaključka da je neuređenost u strukturi proteina česta u prirodi. Sa kompleksnošću organizama povećava se i stepen neuređenosti proteina unutar proteoma (Xue *et al.*, 2010, 2014; Xue, Dunker and Uversky, 2012). Procenjuje se da polovina eukariotskih proteina poseduje duge RNTS (Peng *et al.*, 2014), a samo do sada su kod najmanje 6600 humanih proteina identifikovani RNTS (Afanasyeva *et al.*, 2018). Razlike između proteina uređene tercijarne strukture (PUTS) i PNTS, ogledaju se ne samo na nivou trodimenzionalne strukture, već i na nivou sekvence proteina uključujuću kompoziciju sekvence, fleksibilnost i kompleksnost polipeptidnog lanca, kao i hidrofobnost delova sekvence. U sekvenci PNTS i proteina sa RNTS proporcionalno se

nalazi više hidrofilnih u odnosu hidrofobne aminokiselinske ostatke (Williams *et al.*, 2001; Radivojac *et al.*, 2007). Razvijen je veliki broj metoda za predviđanje neuređenosti u tercijarnoj strukturi proteina na osnovu sekvence aminokiselina (He *et al.*, 2009; Deng, Eickholt and Cheng, 2012; Meng, Uversky and Kurgan, 2017). Poređenjem sekvenci PNTS proteina sa veštačkim sekvencama iste aminokiseline kompozicije utvrđeno je da su RNTS bioloških sekvenci znatno duže. Ovakve analize ukazale su na evolutivnu očuvanost PNTS sa dužim sekvencama. Pretpostavlja se da je to uticalo na povećanja funkcionalnog diverziteta eukariotskih proteina, te znatno kompleksniji odnos između sekvence, 3D strukture i funkcije proteina (Yu *et al.*, 2016; Uversky, 2018). PNTS i proteini sa RNTS obavljaju biološke funkcije koje su komplementarne PUTS proteinima. Specifične uloge PNTS i proteina sa RNTS ogledaju se u kontroli i regulaciji različitih signalnih puteva, pri čemu su istovremeno i sami precizno regulisani (Uversky and Dunker, 2010). Funcionalne prednosti PNTS i proteina sa RNTS povezane su sa sposobnošću formiranja velikog broja visoko specifičnih interakcija niskog afiniteta vezivanja (Oldfield *et al.*, 2005). Specifično savijanje proteina uslovljeno je okruženjem, funkcijom ili različitim interakcijama sa proteinima, membranama, nukleinskim kiselinama ili malim proteinima. Kao rezultat, opisana konformaciona plastičnost proteina omogućava brzu promenu funkcionalnih uloga. Veliki broj potencijalnih interakcija koje PNTS formiraju, čini ove proteine ključnim čvorovima u sklopu mreža IPP. PNTS proteini tako imaju ključnu ulogu u kontroli i održavanju vremenske i prostorne organizacije IPP mreža (Haynes *et al.*, 2006; Uversky, 2017, 2018).

1.1.4 Interakcije protein-protein kod čoveka

Od završetka projekta sekvenciranja ljudskog genoma (engl. Human Genome Project) (Craig Venter *et al.*, 2001; Lander *et al.*, 2001; Collins *et al.*, 2004) učinjen je značajan napredak u identifikaciji gena odgovornih za veliki broj različitih fenotipova, uključujući većinu naslednih monogenskih bolesti (engl. Mendelian diseases) (Hamosh *et al.*, 2005), različitih kompleksnih osobina (Hindorff *et al.*, 2009) i nekoliko hiljada vrsta tumora (Chin *et al.*, 2011). Identifikacija sekvenci gena ne odgovaraja na pitanje kojim mehanizmima promene u sekvenci DNK se menjaju funkcije gena i genskih produkata u kontekstu njihovih kompleksnih interakcija (Vidal, Cusick and Barabási, 2011). S

obzirom da je biološka funkcija proteina direktno vezana za njihove interakcije sa drugim proteinima, detekcija tih interakcija ima važnu ulogu u identifikaciji njihovih funkcija u ćeliji (Keskin *et al.*, 2008).

Interakcije između proteina su esencijalne za sve ćelijske procese (Titeca *et al.*, 2018). Potpuno razumevanje odnosa genotipa i fenotipa zahteva identifikaciju i opisivanje IPP (Rolland *et al.*, 2014). Međusobnim interakcijama velikog broja proteina se formiraju vremenski i prostorno zavisna mreža interakcija na nivou ćelije. Visoko kvalitetne i ekstenzivne mape mreža ljudskih IPP neophodne su za razumevanje promena unutar strukture mreža (Kuzmanov and Emili, 2013). Promene u strukturi mreža IPP su karakteristične za nasledna oboljenja (Rolland *et al.*, 2014). Kompletna mreža IPP deo je veće mreže fizičkih interakcija između makromolekula u ćeliji – interaktoma (Sanchez *et al.*, 1999; Vidal, Cusick and Barabási, 2011). Interaktom se najčešće odnosi na mrežu interakcija među proteinima ili između proteina i DNK. Interaktom može da predstavlja i interakcije unutar i između drugih grupa makromolekula (Yan *et al.*, 2018).

Procenjuje se da proteinski interaktom čoveka sadrži oko 650.000 IPP, mada ove procene ne uzimaju u obzir sve tranzijentne i specifične interakcije među proteinima (Stumpf *et al.*, 2008). Koristeći samo jednu varijantu alternativnog splajsovanja od procenjenih 20.000 protein-kodirajućih gena (Kim *et al.*, 2014) dobijamo ~200 miliona potencijalnih humanih IPP. Ukoliko se ima u vidu da prema proceni 92-94% ljudskih protein-kodirajućih gena daje više alternativnih transkripata (Wang *et al.*, 2008), broj potencijalnih interagujućih parova proteina višestruko raste.

1.1.5 Tipovi interakcija protein-protein

Biološki sistemi se mogu posmatrati kao kompleksne mreže različitih tipova odnosa među makromolekulama, među kojima su najvažniji interakcije među proteinima (Merico, Gfeller and Bader, 2009). Kao okosnica ćelijske kompleksnosti i dinamike, interakcije među proteinima se mogu podeliti u pet grupa (De Las Rivas and Fontanillo, 2012):

1. Fizičke interakcije
2. Regulatorne asocijacije

3. Genetičke interakcije
4. Funkcionalne asocijacije.

Fizičke interakcije protein-protein uključuju direktan ili indirektan fizički kontakt proteina, i čine osnovu za druge tipove asocijacija. Regulatorne asocijacije obuhvataju odnose aktivacije i inhibicije genske ekspresije, pri čemu transkripcioni regulatori imaju ulogu medijatora. Genetičke interakcije su prisutne ukoliko se analizom efekata kombinacije pojedinačnih gena utvrdi da postoje razlike u ispoljenom fenotipu u odnosu na očekivani fenotip. Funkcionalna asocijacija često je rezultat učestvovanja različitih proteina u istim signalnim ili metaboličkim putevima, kolokalizaciji unutar istog kompartmenta, organele ili ćelijske makrostrukture. Među grupom proteina istovremeno mogu da postoje različiti tipovi asocijacija sa različitim nivoima značaja za specifičan biološki proces (De Las Rivas and de Luis, 2004).

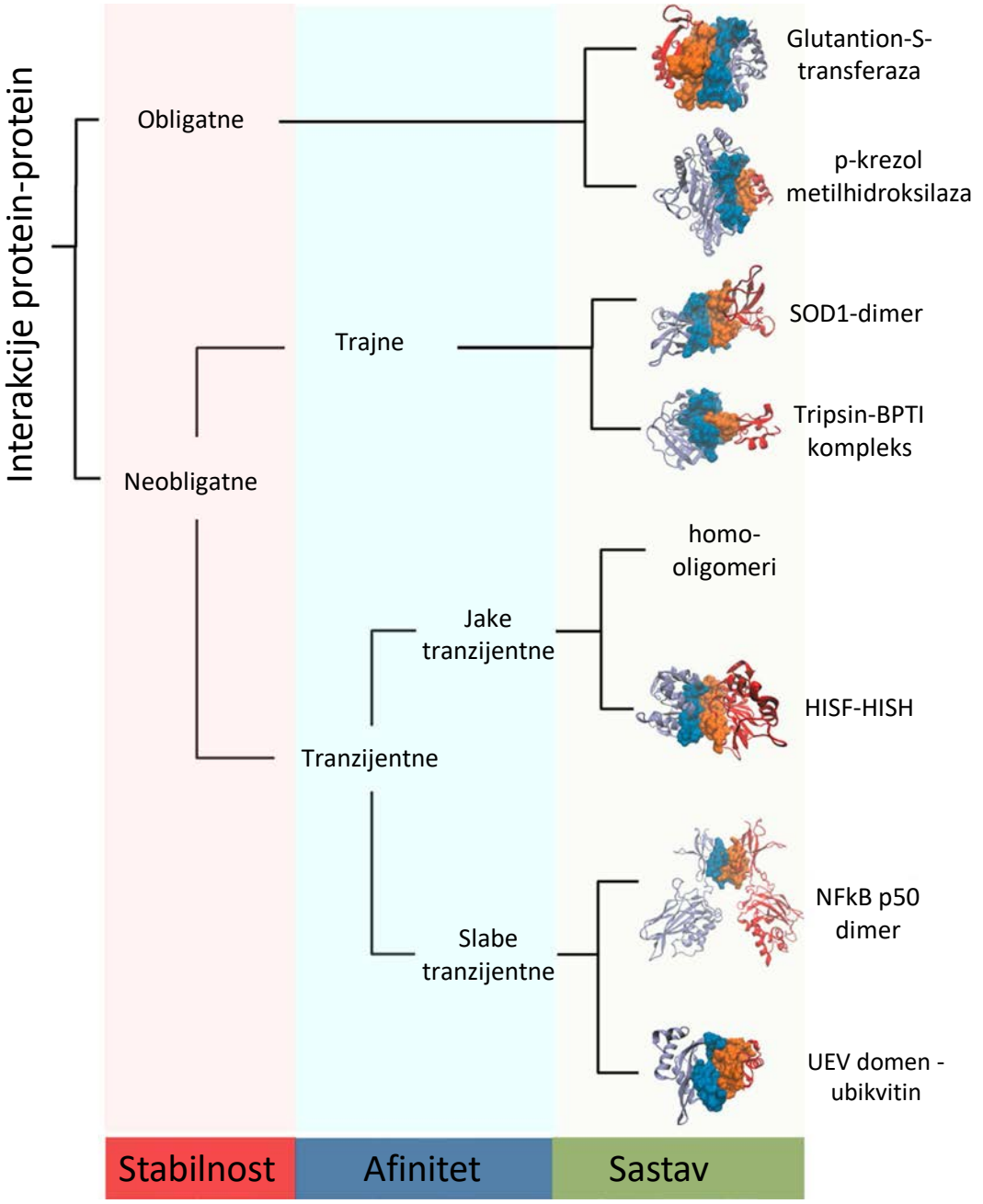
IPP se definišu kao fizički kontakt između proteina, koji uključuje molekularno ukotvljavanje i koji se ostvaruje *in vivo*. Fizički kontakt među proteinima može biti direktan ili indirektan. Direktan kontakt se ostvaruje ukoliko se proteini povezuju preko svojih molekulskih interfejsa, dok je indirektan ako se fizički kontakt ostvaruje preko jednog ili više proteina, koji vrše funkciju medijatora u sklopu većeg proteinskog kompleksa (De Las Rivas and Fontanillo, 2012). Proteini koji deluju u sklopu većeg proteinskog kompleksa, poput ribozoma ili osnovnog transkripcionog aparata, nalaze se u funkcionalnom međusobnom kontaktu (De Las Rivas and Fontanillo, 2010). Potvrda direktne fizičke interakcije između dva proteina unutar kompleksa često zahteva dokaze iz više eksperimentalnih metoda (Mackay *et al.*, 2007; Chatr-aryamontri *et al.*, 2008). IPP podrazumevaju specifične i neslučajne interakcije, čime se isključuju interakcije tokom procesa sinteze, savijanja ili degradacije proteina (Peng *et al.*, 2017). Većina proteina su visoko specifični u izboru interagujućeg partnera, iako se kod nekih pojavljuje kompeticija više interaktora za dostupna vezujuća mjesta (Nooren and Thornton, 2003). Specifičnost fizičke interakcije ostvaruje se delovanjem veoma preciznih biomolekularnih sila sa namerom formiranja posebnog interfejsa interakcije (De Las Rivas and Fontanillo, 2010).

IPP se mogu kategorizovati prema sastavu, efektu na proteinsku stabilnost i vremenskom trajanju interakcije (Keskin, Tuncbag and Gursoy, 2008).

Prema sastavu, proteinski kompleksi bilo da su binarni ili formirani interakcijom većeg broja proteina, mogu biti homo-oligomeri ili hetero-oligomeri. Homo-oligomeri se formiraju interakcijom istih proteinskih lanaca i čine osnovu stabilnih makromolekula. Hetero-oligomeri predstavljaju komplekse različitih proteinskih lanaca varijabilne stabilnosti (Keskin, Tuncbag and Gursoy, 2016).

Prema stabilnosti, proteinski kompleksi mogu biti obligatni ili neobligatni. Interakcija je obligatna ukoliko su monomeri proteinskog kompleksa nestabilni *in vivo*. Sa druge strane, komponente neobligatnih interakcija mogu nezavisno egzistirati u ćeliji (Levy and Teichmann, 2013). Veliki deo do sada opisanih interakcija su neobligatne i uključuju interakcije učesnika u signalinim i metaboličkim putevima, interakcijama antitelo-antigen, receptor-ligand, enzim-inhibitor i mnoge druge. Proteini koji su komponente neobligatnih kompleksa često pre interakcije nisu kolokalizovani i stabilne su strukture nezavisno od kompleksa (Nooren and Thornton, 2003). Proces formiranja obligatnih kompleksa uključuje savijanje i istovremeno vezivanje monomera, dok se kod neobligatnih kompleksa proteini prvo pakuju, pa tek onda ostvaruju fizičku interakciju (Keskin, Tuncbag and Gursoy, 2016).

Prema vremenskom trajanju interakcija IPP se dele na trajne i tranzijentne. Obligatni kompleksi po svojoj prirodi podrazumevaju stabilnu permanentnu interakciju među monomerima, pa ova klasifikacija ima smisla samo u kontekstu neobigatnih interakcija (Nooren and Thornton, 2003). Tranzijentne interakcije uključuju *in vivo* asocijaciju i disocijaciju homomera u veoma kratkom vremenskom roku, najčešće između 0.1 i 1 sekunde, dok trajne interakcije mogu da traju minutama ili satima (Rudolph, 2007). Tranzijetne interakcije se nalaze u osnovi brojnih važnih ćelijskiih procesa, posebno onih koji uključuju brz ćelijski odgovor na ekstraćelijski signal. Ovakav tip interakcije obuhvataju interakcije receptor-hormon, komplekse formirane u sklopu signalnih i metaboličkih puteva, proteinskog transporta, itd. (Keskin, Tuncbag and Gursoy, 2016). Trajne interakcije su neophodne za formiranje kompleksa koji učestvuju u procesima DNK replikacije i transkripcije, formiranje kompleksa nukleusnih pora, ribozoma i splajsozoma (Phizicky and Fields, 1995). Interakcije tipa antitelo-antigen su primer neobligatnih, ali permanentnih kompleksa (Keskin, Tuncbag and Gursoy, 2016).



Slika 2. Klasifikacija interakcija protein-protein prema stabilnost, afinitetu i kompoziciji. Prema stabilnosti, interakcije mogu biti obligatne ili neobligatne. Prema vremenu trajanja, interakcije mogu biti trajne ili tranzijentne. Prema afinitetu interakcije se klasifikuju na slabe i jake. Prema sastavu interakcija, proteini mogu formirati homodimere- (dimeri na gornjem panelu) ili heterodimere (dimeri na donjem panelu) (Keskin, Tuncbag and Gursoy, 2016).

1.1.6 Značaj mreža interakcija protein-protein

Mapiranje i karakterizacija mreže IPP na nivou proteoma omogućava predviđanje funkcije proteina, rasvetljavanje signalnih puteva, određenih proteinskih kompleksa, kao i proteina i gena ključnih za oboljenja čoveka (Peng *et al.*, 2017). Opisivanje mapa IPP se može iskoristiti u svrhu karakterizacije funkcije proteina koji je jedan od najvažnijih problema molekularne biologije (Peng *et al.*, 2017).

IPP u funkciji proteina

Matematička struktura IPP mreža formiranih na osnovu postojećih podataka o IPP može biti upotrebljena u svrhu predviđanja proteinskih funkcija (Nabieva *et al.*, 2005). Pri tome se analiziraju neposredni interaktori ciljnog proteina, njihove funkcionalne karakteristike i sličnosti (Chua, Sung and Wong, 2006). Analiza biohemijskih i fizičko-hemijskih karakteristika ciljnih proteina, zajedno sa analizom lokalnih i globalnih karakteristika njihovih IPP mreža, može voditi efikasnijem rasvetljavanju funkcija proteina u sklopu biološkog sistema (Hu *et al.*, 2011; Trivodaliev, Bogojeska and Kocarev, 2014).

IPP i signalni putevi

Ključna karakteristika višećelijskih organizama jeste kompleksna mreža signalnih i metaboličkih intercelularnih puteva. Ta mreža predstavlja strukturu na osnovu koje se formira mehanizam odgovora na promene i događaje unutar i izvan ćelije, kako u normalnim tako i u patološkim stanjima (Thompson and Giancotti, 2018). Objavljeni su različiti pristupi za identifikaciju članova signalnih puteva koristeći mreže IPP kao osnovu istraživanja. Upotrebom statističkih metoda moguće je na osnovu postojećih linearnih puteva u sklopu IPP mreže pretpostaviti nove članove homolognog signalnog puta (Shlomi *et al.*, 2006). Jedan od pristupa identifikuje elemente signalnog puta između dva ciljna proteina, korišćenjem dodatnih podataka o proteinima koji formiraju IPP mape, poput karakteristika sekvence, ćelijske lokalizacije i obrazaca ekspresije (Liu, Wong and Chua, 2009). Razvijeni su bioinformatički algoritmi koji osim mape IPP koriste i podatke o njihovoj pouzdanosti IPP, kao i druge izvore informacija kako bi pored identifikacije

ključnih članova signalnih puteva definisao smer rasprostiranja signala (Gitter *et al.*, 2011; Nguyen *et al.*, 2015).

IPP i identifikacija kompleksa i modula unutar mreže IPP

Identifikacija većih proteinskih kompleksa i funkcionalnih modula IPP mreža je još jedno od polja primene informacija o IPP i njihovoj pouzdanosti. Identifikacija funkcionalno povezanih modula mreže može biti od velikog značaja kod potpunog rasvetljavanja različitih ćelijskih mehanizama (Hakes *et al.*, 2008). Neki od metoda detekcije blisko povezanih grupa proteina u sklopu mreže koriste informaciju o semantičkoj sličnosti proteina koji formiraju IPP mrežu, zasnovane na genskim ontologijama (engl. Gene Ontology, GO), kako bi detektovali funkcionalno povezane članove lokalne grupe poroteina (Asthana *et al.*, 2004; Lubovac *et al.*, 2007). Jedna od tehnika detekcije lokalnih grupa uključuje otežavanje IPP mreža dodeljivanjem slojeva informacija na IPP mrežu na osnovu: analize sličnosti različitih IPP mreža (Sharan *et al.*, 2005), analize podataka o ekspresiji proteina (Chou and Cai, 2006) i statističke obrade podataka dobijenih masenom spektrometrijom (MS) (Krogan *et al.*, 2006). Kombinujući različite vrste informacija, interakcije među proteinima se kvantifikuju čime se mreža IPP otežava. Na ovakvim mrežama se onda primenjuju algoritmi pretrage lokalnih grupa proteina i primenom statističkih analiza izvodi zaključak o pripadnosti proteina definisanoj grupi (Liu, Wong and Chua, 2009; Zaki, Efimov and Berengueres, 2013).

IPP i obolenja

Jedna od ključnih primena analize mreža IPP je u produbljivanju biološkog uvida u mehanizme oboljenja. Kombinovana analiza informacija o obrascima ekspresije gena i mreže IPP kvasca (*Saccharomyces cerevisiae*), omogućila je prioritizaciju gena asociranih sa određenim genotipovima za detaljnije analize (Ma *et al.*, 2007). Pomoću algoritma za analizu mreža IPP otkriveni su novi geni, blisko povezani unutar mreže IPP sa već poznatim genima i njihovim produktima, koji igraju važnu ulogu u razvoju mozga (Chen *et al.*, 2015). Za identifikaciju novih proteina povezanih sa Alchajmerovom bolešću bilo je neophodno formiranje specifične mreže interakcija proteina koji su već

opisani kao faktori razvoja i progresije bolesti. Bioinformatičkom analizom ove specifične mreže IPP, zajedno sa informacijama o kvalitetu poznatih IPP, omogućena je prioritizacija ciljnih proteina i njihovih IPP koji su povezani sa Alchajmerovom bolesti (Malhotra *et al.*, 2015). Analize mapa IPP mogu doprineti boljem razumevanju kardiovaskularnih oboljenja. Na osnovu mrežne analize specifičnih IPP mreža i analize funkcionalne sličnosti proteina, razvijen je statistički metod prioritizacije daljnjeg testiranja proteina povezanih sa kardiomiopatijom (Li *et al.*, 2013).

Detekcija esencijalnih proteina

Esencijalni proteini su protein ključni za biološki sistem i njihovo odsustvo ili nepravilno funkcionisanje dovodi do smrti ili nemogućnosti reprodukcije organizma (Peng *et al.*, 2012). Identifikacija esencijalnih proteina u sklopu IPP mreža je ključna za proučavanje patologija, sintetičke biologije i u razvoju lekova i terapeutika. Razvijen je veći broj računarskih metoda koje koriste informacije sadržane u strukturi IPP mreža zajedno sa različitim tehnikama otežavanja istih u svrhu detekcije ključnih proteina (Peng *et al.*, 2017). Kao osnova za statističko modelovanje i predviđanje esencijalnih proteina mogu se koristiti topološke karakteristike otežanih IPP mreža. Ove karakteristike mogu biti dobijene analizom centralnosti IPP mreža (Li *et al.*, 2010), integracijom podataka o genskoj ekspresiji (Li *et al.*, 2012; Tang *et al.*, 2014), analizom homologih proteina i stepena grupisanja unutar mreže (Peng *et al.*, 2012) i ćelijske lokalizacije samih IPP (Peng *et al.*, 2015).

1.2 Eksperimentalne metode za detekciju IPP

Da bi se razumeli mehanizmi interakcija među proteinima na molekularnom nivou i mapirale IPP na globalnom nivou, došlo je do razvoja velikog broja eksperimentalnih tehnika za detekciju i karakterizaciju IPP. Proteinske interakcije mogu biti identifikovane korišćenjem različitih biohemijskih, genetičkih i fizičkih metoda. Najopštija klasifikacija metoda je zasnovana na obimu podataka koji se dobijaju – metode male skale (eng. Small Scale Methods, SSM) i metode velike skale (eng. Large Scale Methods, LSM). U zavisnosti od mesta gde se proteinske interakcije mere/detektuju, metode se dele na *in*

vitro i *in vivo*. *In vitro* analize se obavljaju izvan živog sistema i kao takve su pod strogo kontrolisanim uslovima, dok se *in vivo* eksperimenti izvode u živim sistemima, najčešće ćelijama (Phizicky and Fields, 1995; Browne *et al.*, 2010a). Takođe, metode mogu služiti isključivo za identifikaciju ili za vizualizaciju, odnosno, karakterizaciju interakcija.

1.2.1 Metode male skale

Metode male skale obuhvataju nekoliko metoda koje se koriste za (i) detekciju/identifikaciju proteinskih inerakcija, (ii) intracelularnu vizuelizaciju i (iii) verifikaciju (Brun *et al.*, 2003; Syafrizayanti *et al.*, 2014). Njihov princip je zasnovan na biohemijskim i biofizičkim osobinama proteinskih kompleksa (von Mering *et al.*, 2002a). Kako se u najvećem broju slučajeva metode koriste u kombinacijama, smatra se da su veoma pouzdane (von Mering *et al.*, 2002b; Stumpf *et al.*, 2008; Cusick *et al.*, 2009).

1.2.1.1 Imunoprecipitacija

Imunoprecipitacija (IP) je metoda kojom se izoluju specifični proteini od interesa/antigeni iz liziranih ćelija koristeći interakciju antitelo-antigen. Ćelijski lizati se inkubiraju sa antitelima koja su specifična za određene proteine od interesa. Uzorak se dalje inkubira sa magnetnim ili agaroznim kuglicama koje su najčešće obložene Proteinom A ili Proteinom G za koje se antitela vezuju. Zatim se kompleksi željenog proteina i antitela eluiraju sa kuglica odgovarajućim puferom. Proteini vezani za antitelo se odvajaju od ostatka uzorka i analiziraju. Ko-imunoprecipitacija (engl. complex-immunoprecipitation, co-IP) je varijanta imununoprecipitacije i koristi se za identifikaciju protein-protein interakcija - između željenog proteina i nepoznatog proteina koji je vezan za njega. Svi proteini se u daljem procesu analiziraju koristeći samo SDS-PAGE (engl. sodium dodecyl sulfate-polyacrylamide gel electrophoresis), ili u kombinaciji sa Western Blot (WB) metodom ili MS. Korišćenje eukariotskih ćelijskih lizata dodatno omogućava analizu post-translacionih promena, koje nisu prisutne u prokariotskim ćelijama (Rao *et al.*, 2014). Nedostaci ove metode su nemogućnost vizuelizacije svih subćelijskih komponenti, jer se ne izvodi *in vivo*, niti su membrane

očuvane. Takođe, neophodno je obeležavanje proteina ili korišćenje visoko specifičnih antitela, u suprotnom je pozadinski šum veliki (Persani, Calebiro and Bonomi, 2007).

1.2.1.2 Afinitetno prečišćavanje

Afinitetno prečišćavanje je tip afinitetne hromatografije koja se zasniva na interakciji proteina (liganada) koji su imobilisani na matriksu ili kolumni, tj. u čvrstoj fazi, sa proteinima iz ekstrakta u tečnoj fazi. Posle izlaganja ekstrakta ostaju vezani samo proteini ("plen") koji interaguju sa "mamcima". Na kraju se vezani proteini odvajaju od mamaca procesom elucije i analiziraju. Afinitetna hromatografija takođe može biti povezana sa SDS-PAGE i MS kako bi se generisali podaci velike skale (Rao et al). Najveći nedostatak ove metode je mnoštvo nespecifičnih interakcija između proteina, kao i činjenica da neke interakcije nisu prisutne *in vivo* (Gavin, Maeda and Kühner, 2011; Hubner and Mann, 2011; Rao *et al.*, 2014).

1.2.2 Metode za intracelularnu vizuelizaciju/lokalizaciju

1.2.2.1 Metoda ligacije usled blizine

Metoda ligacije usled blizine (engl. Proximity Ligation Assay, PLA) je inovativna metoda za preciznu detekciju intracelularnih proteina i proteinskih interakcija. Razlikuju se direktan i indirektan način detekcije. Za direktno prepoznavanje dva proteina koja interaguju koriste se dva različita antitela obavezno proizvedena u različitim vrstama životinja, za koja su konjugurane kratke DNK sekvence. Zatim se dodaju jos dve DNK sekvence, koje se nazivaju konektori. Sekvence koje su već prisutne na antitelima se ligiraju sa konektorima, što dovodi do formiranja kružne jednolančane DNK. Jedna od DNK sekvenci konjugovane za antitelo služi kao prajmer za amplifikaciju DNK. Dodavanjem DNK polimeraze, formira se DNK koja ostaje povezana sa antitelima. Kontinuirana sinteza DNK vodi ka nastanku dugačkog lanca DNK koji se sastoji od istih repetitivnih sekvenci (konkatamer). Ova sekvenca se može detektovati hibridizacijom sa različitim oligonukleotidima, koji mogu nositi fluorescentne probe i time se vizuelizovati pod mikroskopom (Fredriksson *et al.*, 2002; Söderberg *et al.*, 2006). Indirektna PLA

metoda uključuje sekundarna antitela koja imaju konjugirane sekvence i vezuju se za primarna antitela (Jiang *et al.*, no date). Najveći nedostatak metode je direktna zavisnost rezultata od kvaliteta antitela, kao i činjenica da se prikazuje samo kolokalizacija proteina, ali ne obavezno i interakcija (Titeca *et al.*, 2018)

1.2.2.2 Prenos fluorescentne rezonantne energije

Prenos fluorescentne rezonantne energije (engl. Fluorescence Resonance Energy Transfer, FRET) se koristi sa ciljem određivanja bliskosti molekula u opsegu od nekoliko nanometara, što može dovesti i do njihove interakcije. Zasniva se na transferu energije između dva fluorofora, ekscitiranog donora i odgovarajućeg akceptora. Transfer energije se dešava kada se ekscitaciona energija donora preklapa sa apsorpcionom energijom akceptora (Lakowicz, 2006; Padilla-Parra and Tramier, 2012). Pošto transfer energije zavisi od blizine molekula, signal koji se dobija vrlo precizno pokazuje njihovu udaljenost, pritom ne utičući na same molekule (Miyawaki, 2011). Prednosti ove metode su visoka specifičnost i preciznost, kao posledica analize u samo jednoj ćeliji (eng. single cell analysis), kao i analiza IPP *in vivo* u realnom vremenu. Iako je FRET odličan za prikazivanje proteina na veoma maloj razdaljini i čak direktnih interakcija, ipak ima najnižu senzitivnost od svih sličnih metoda; potrebna je ekstremna bliskost proteina da bi došlo do detekcije, a sam signal se ne amplifikuje. Osim toga, često dolazi do pozadinske fluorescencije, kao i foto-izbeljivanja (Syafrizayanti *et al.*, 2014; Titeca *et al.*, 2018)

1.2.2.3 Bimolekularna fluorescentna komplementacija

Bimolekularna fluorescentna komplementacija (engl. Bimolecular fluorescence complementation, BiFC) je praktična metoda za brzu *in vivo* vizuelizaciju IPP-a u različitim ćelijskim sistemima. Zasniva se na komplementaciji dva fragmenta fluorescentnog molekula, koja sama nisu fluorescentna, a čijim spajanjem dolazi do aktiviranja fluorescencije. Fragmenti su vezani za željene proteine, tako da interakcija ili bliskost dva proteina omogućavaju spajanje fragmenta i aktiviranje signala (Kerppola, 2008; Kodama and Hu, 2012). Glavni nedostatak ove metode je visok stepen lažno

pozitivnih IPP, kao i nestabilnost formiranog fluoresecentnog kompleksa (Xing *et al.*, 2016a).

1.2.2.4 Fluorescentna korelaciona spektroskopija

Fluorescentna korelaciona spektroskopija (engl. Fluorescence Correlation Spectroscopy, FCS) je visoko-specifična metoda za analizu IPP u uzorcima veoma malih zapremina. Metoda se bavi analizom mikroskopske fluktuacije fluorescencije obeleženog proteina. Za detekciju se koristi konfokalni mikroskop. Laserski zrak se usmerava ka ćeliji što dovodi do difuzije ili druge hemijske reakcije. Kako ove pojave menjaju fluktuacije intenziteta apsorbovane ili emitovane svetlosti u fluorescentno-obeleženom proteinu, detektuju se kretanja ili interakcije sa drugim proteinima. Ukoliko je obeleženi protein vezan za drugi protein (molekul), to utiče na fluorescenciju (Elson and Magde, 1974; Elson, 2011). Međutim, teško je detektovati potencijalne slabe interakcije ovom metodom, kao i interakcije između partnera koji imaju različite koncentracije. Osim toga, nije moguće razlikovati neposrednu interakciju od kolokalizacije (Macháň and Wohland, 2014).

1.2.3 Eksperimentalne metode za validaciju IPP

2D-poliakrilamidna gel elektroforeza

2D-poliakrilamidna gel elektroforeza (2D PAGE) je metoda za analiziranje ciljanih proteina i njihove koncentracije. Uzorak se prvo denaturiše koristeći specifične deterdžente, zatim se razdvaja na posebnom gelu u zavisnosti od svoje izoelektrične tačke gel za (izolelektrično fokusiranje, IEF) gel. Naime, IEF gel je poliakrilamidni gel koji ima pH gradijent, tako da proteini migriraju sve dok ne dostignu svoju izoelektirčnu tačku (pH na kome proteini nemaju neto naelektrisanje). Dalje se proteini razdvajaju po svojom molekularnoj masi na SDS-PAGE-u (prvo se tretiraju sa SDS, koji negativno naelektriše proteine, a zatim se razdvajaju pod električnim poljem isključivo prema masi). Na taj način se dobijaju proteini u 2D matriksu, razdvojeni prema izoelektričnoj tački i masi. Na kraju se vrši analiza uzoraka koristeći ili bojenje gela, Western blotting ili se uzorci

isecaju iz gela i šalju na analizu MS (Bjellqvist *et al.*, 1982; Büyükköroğlu *et al.*, 2018). Ova metoda je pristupačna i omogućava formiranje jasne slike. S druge strane, ima nisku reproducibilnost, nije dovoljno efikasna i vremenski je zahtevna (Syafrizayanti *et al.*, 2014).

Metoda mikroskopije atomskih sila

Metoda mikroskopije atomskih sila (engl. Atomic Force Microscopy, AFM) se koristi za merenje međumolekulskih sila između dva proteina. Dok je jedan protein imobilisan, drugi protein se nalazi na mobilnom nosaču čiji vrh skenira i meri jačinu interakcija (Kao *et al.*, 2012). Iako precizna, metoda zahteva skupu opremu i prečišćene, najčešće rekombinantne proteine.

Metode za strukturalnu vizuelizaciju proteinskih kompleksa

Najadekvatniji način analize proteinskih kompleksa je kroz vizuelizaciju njihove strukture koristeći metode kristalografije sa X-zracima, krio-elektromikroskopije i nuklearne magnetne rezonance.
Za **kristalografiju uz korišćenje X-zraka** potrebno je da je protein prečišćen i da ima visoku koncentraciju kako bi se kristalizovao. X-zraci se usmeravaju na kristal, i svetlost se difraktuje stvarajući sliku. Slika se analizira i na osnovu nje određuje struktura molekula (Smyth and Martin, 2000).

Krio-elektron mikroskopija je novija metoda čija je prednost u odnosu na kristalografiju sa X-zracima što ne zahteva 3D kristale, već duboko zaleđen uzorak, koji ostaje očuvan u nativnom stanju. Takav uzorak se izlaže elektronskom laseru i dobija se prvobitno 2D, a zatim i 3D slika uzorka, pomeranjem analizirane ravni. Metoda je trenutno u ekspanziji zbog mogućnosti analize pojedinačnih partikula (Frank, 2002; Murata and Wolf, 2018).

Nuklearna magnetna rezonanca (NMR) se koristi za otkrivanje slabih proteinskih interakcija. Zasniva se na nuklearnom Overhauser efektu, tj. prelasku polarizacije

nukleusovog spina sa jednog nukleusa na drugi, ili sa jedne populacije na drugu (Vinogradova and Qin, 2011). NMR se upotrebljava za analizu proteina manjih od 50 kD (Bonvin, Boelens and Kaptein, 2005).

1.2.4 Metode velike skale

Metode velike skale omogućavaju analizu velikog broja IPP i samih proteina u kratkom vremenskom roku. Najviše korišćene metode za analiziranje obuhvataju kvaščev dvohibridni test (*Saccharomyces cerevisiae*), tandemsko afinitetno prečišćavanje (engl. Tandem Affinity Purification, TAP) u kombinaciji sa MS, sintetičku letalnost i proteinski ereji (engl. protein arrays).

1.2.4.1 Kvaščev dvohibridni sistem

Kvaščev dvohibridni sistem (engl. Yeast Two-Hybrid, Y2H) je jednostavna i relativno pouzdana metoda, koja je zasnovana na rekonstituciji transkripcionih faktora kada dva proteina interaguju. Transkricioni faktori se najčešće vezuju uzvodno od gena koji regulišu, a sastoje se od domena koji se vezuje za DNK i domena koji aktivira transkripciju. Ukoliko domeni ne interaguju fizički, transkripcija se prekida. Ova metoda se zasniva na zasebnom kloniranju dva proteina; jedan koji se vezuje za DNK (vezujući domen – engl. DNK binding domain , BD) i drugi koji aktivira transkripciju (aktivirajući domen – engl. activating domain , AD). BD se označava kao ”mamac”, dok je AD “plen” (Uetz *et al.*, 2000; Ito *et al.*, 2001; Rual *et al.*, 2005; Stelzl *et al.*, 2005).

Protein od interesa se fuzioniše sa DNK-vezujućim segmentom tako sto se DNK konstrukt ove himere klonira u jednom od plazmida koji se koriste za ekspresiju. Postupak je identičan i za drugi protein, fuzionisan sa aktivirajućim segmentom. Zatim se plazmidi ubacuju u ćelije kvasca (*Saccharomyces cerevisiae*). Ukoliko proteini interaguju, “mamac” i “plen” sklopiće funkcionalni transkripcioni faktor, koji će aktivirati transkripciju ciljnog gena. Plazmidi imaju i selekcioni marker, koji je neka od esencijalnih aminokiselina. Ćelije kvasca se gaje u medijumu bez spomenutih aminokiselina, tako da samo one ćelije koje eksprimiraju neophodne aminokiseline mogu da prežive. Kao transkripcioni faktor obično se koristi Gal4, koji reguliše ekspresiju *His3*, *Ade2*, kao i

LacZ gena. *His3* kodira imidazol-glicerofosfat (IGP), enzim za biosintezu aminokiseline histidina, dok je *Ade2* zadužen za sintezu adenina. U ovom slučaju, ćelije kvasca se zaseju na agar bez histidina/adenina, što dozvoljava preživljavanje samo onih ćelija kod kojih dolazi do interakcije, odnosno, ćelija u kojima se vrši sinteza histidina. U drugom slučaju, aktivacija *LacZ* gena dovodi do sinteze beta-galaktozidaze, enzima koji aktivan hidrolizuje jedinjenje X-gal (bromohloroindoksil-galaktozid), pri čemu se otpušta jedinjenje koje boji kolonije u plavo. Plave kolonije su jasan znak da je gen aktivan (Fields and Song, 1989; Fashena, Serebriiskii and Golemis, 2000; Auerbach *et al.*, 2002). U prvom slučaju dva gena se koriste da odrede različite tipove interakcija: Gal4/*His3* – slabe interakcije, Gal4/*Ade2* – jake interakcije, dok Gal4/*LacZ* služi za kvantifikaciju interakcija merenjem LacZ aktivnosti enzimatskim esejima (Xing *et al.*, 2016b).

Kvaščev dvohibridni sistem je unapređen vremenom, tako da se danas koristi za skrining čitavog proteoma. Dva glavna pristupa su zasnovana na matrici i biblioteci.

Princip matrice se sastoji od dve matrice - matrice sa raspoređenim sojevima kvasaca koji eksprimiraju različite proteine "plena" i matrice sa sojem kvasca koji eksprimira jedan protein "mamac". Matrice se spoje i ćelije kvasca se reprodukuju. Na osnovu selekcionog markera se biraju diploidne ćelije, a ukoliko je došlo do ekspresije ciljnog gena, interakcija proteina je vidljiva na matrici (Uetz *et al.*, 2000).

U nasumičnom skriningu biblioteke koristi se jedna kvaščeva kultura koja eksprimira "mamac" i biblioteka ćelija sa različitim "plenom". Diploidi i pozitivne proteinske interakcije se selektuju na osnovu markera, ali se moraju proveriti sekvenciranjem kako bi se precizno identifikovao "plen" protein (Uetz *et al.*, 2000; Arndt and Vorberg, 2012). Glavni nedostatak ovog sistema je činjenica da su interakcije otkrivene ovim sistemom veštačke prirode i moraju biti potvrđene u *in vivo* uslovima nekog drugog model sistema ili nekim dodatnim metodama. Zatim, ukoliko je protein od interesa označen (tagovan) na N-kraju, može doći do gubitka funkcije proteina. Takođe, identifikuju se samo proteini koji su eksprimirani u nukleusu kvasca, što nije uvek optimalno i ne dozvoljava analizu transmembranskih proteina ili proteina u ostalim subcelularnim delovima (Uetz *et al.*, 2000; von Mering *et al.*, 2002b).

1.2.4.2 Pročišćavanje tandemskog afiniteta i masena spektrometrija

Tandemsko afinitetno prečišćavanje (engl. Tandem Affinity Purification, TAP) je metoda zasnovana na imunoprecipitaciji i prečišćavanju nepoznatih proteinskih kompleksa iz relativno malog uzorka (Rigaut *et al.*, 1999). Željeni protein se eksprimira kao fuzioni protein, tj. obeležen tagom (TAP tag) koji se sastoji od Proteina A (iz bakterije *Staphylococcus aureus*) i kalmodulin-vezujućeg proteina (engl. Calmodulin-Binding Protein, CBP) koji su međusobno povezani TEV insertom (engl. Tobacco Etch Virus, TEV) koji prepoznaje TEV proteaza. Obeležen protein se inkubira sa ćelijskim lizatom, iz koga se određeni protein veže za željeni protein i formira kompleks. Uzorak se propušta kroz kolonu sa kuglicama obloženim IgG antitelima koja se vezuju za Protein A. Ostatak nespecifično vezanih proteina se ispira, a vezani proteini tretiraju TEV proteazom koja će preseći most između Proteina A i CBP i ukloniti Protein A. U sledećem koraku uzorak se propušta kroz kolonu sa kuglicama obloženim kalmodulinom, tako da se vezuju svi kompleksi sa CBP. Nakon ispiranja i uklanjanja vezanih kuglica, proteinski kompleksi se razdvajaju na SDS-poliakrilamidnom gelu i dalje analiziraju MS (Rigaut *et al.*, 1999; Puig *et al.*, 2001; Abu-Farha, Elisma and Figeys, 2008).

Ova metoda je jednostavna i omogućava dobijanje velike količine informacija o proteinskim kompleksima za relativno kratko vreme. S druge strane, postoji mogućnost da tag utiče na formu ili ekspresiju željenog proteina, kao i gubitak potencijalnih kandidata usled čestih ispiranja kolone. Jedno od rešenja je zamena kompleksnog TAP taga jednostavnijim FLAG tagom (Abu-Farha, Elisma and Figeys, 2008).

1.2.4.3 Sintetička letalnost

Sintetička letalnost se zasniva na principu po kome delecija, inaktivacija ili mutacija jednog od dva izabrana neesencijalna gena ne utiče na vijabilnost, ali nedostatak oba funkcionalna gena je letalan. Ova pojava sugeriše da geni međusobno interaguju i može indirektno ukazati na interakciju između dva proteina kodirana analiziranim genima (Bender and Pringle, 1991; Rutherford, 2000; Ooi, Shoemaker and Boeke, 2003; Fraser *et al.*, 2004). Metoda je danas automatizovana, mutacije se sintetički dizajniraju, a koristi se za *in vivo* analize čitavih genoma (von Mering *et al.*, 2002; Shoemaker and Panchenko,

2007). Iako vrlo korisna, ova metoda ne daje konačnu potvrdu o IPP, već samo nagoveštaj, tako da mora biti praćena dodatnim ispitivanjima.

1.2.4.4 Proteinski čipovi

Proteinski čipovi su najbrža, najlakša i najekonomičnija metoda za analizu proteinskih interakcija (Mitchell, 2002). Iako postoji mnogo varijacija, osnovni princip je zasnovan na činjenici da su željeni proteini imobilisani na čvrstoj podlozi (staklo, membrana) u određenom redosledu i izloženi kontaktu sa obeleženim proteinima. Ukoliko dođe do interakcije između vezanih i slobodnih proteina, mesto na čipu postaje jasno vidljivo, jer se uglavnom koriste fluorescentni tagovi (Schweitzer, Predki and Snyder, 2003; Kaushansky *et al.*, 2010). Kako proteini mogu biti ili prečišćeni ili sintetisani koristeći cDNK biblioteke, moguće je pravljenje čipova sa hiljadama proteina (Angenendt *et al.*, 2006). Cilj korišćenja i unapređenja čipova bio je da se formira efikasan i senzitivan sistem koji bi davao velike količine podataka sa malom količinom inicijalnog uzorka (Rao *et al.*, 2014). Ograničenja ove metode se ogledaju u tome što ne reflektuje *in vivo* uslove, zatim postoji mogućnost gubitka funkcije proteina, kao i manjak post-translacionih modifikacija ukoliko se koriste sintetisani i rekombinantni proteini (Syafrizayanti *et al.*, 2014).

1.2.5 Ograničenja eksperimentalih metoda za detekciju IPP

Razvoj i primena eksperimentalnih tehnika za detekciju IPP na velikoj skali, omogućila je generisanje velike količine podataka o IPP. Sa druge strane, analize pouzdanosti ovako dobijenih podataka otkrile su veliki broj nekompletnih i kontradiktornih informacija (von Mering *et al.*, 2002). Eksperimentalne metode, iako efikasne, imaju svoja ograničenja. Bilo da se radi o visokoj ceni i utrošenom vremenu, kao u slučaju metoda male skale, ili činjenici da daju mnogo lažno pozitivnih rezultata, kao u slučaju metoda velike skale, sve ukazuje na potrebu za dopunskim *in silico* analizama (Rao *et al.*, 2014). Eksperimentalne metode mogu da otkriju samo jedan deo IPP koji se odigravaju u organizmu, tako da proteinski interaktom ostaje nepotpun. Takođe, postoje znatne poteškoće sa reproducibilnošću rezultata metoda velike skale (Jansen *et al.*, 2002; von Mering *et al.*,

2002). Metode male skale su obično pouzdanije od metoda velike skale, ali su zato skuplje i zahtevaju dosta vremena (Phizicky and Fields, 1995). Takođe, u mnogim slučajevima teško je razlikovati kolokalizaciju proteina od njihove međusobne interakcije (Macháň and Wohland, 2014).

Postoji potreba za napretkom tehnologija za detekciju IPP, kako objedinjavanjem postojećih, tako i razvijanjem novih metoda. Neophodno je razviti metode velike skale koje bi bile efikasne, štedeti vreme i novac, a istovremeno biti u mogućnosti da analiziraju dinamiku sistema, lokalizaciju i vremenske okvire IPP na nivou ćelije ili čitavog organizma. Za ostvarenje takvog cilja, jedan od potencijalnih pravaca razvoja je veći stepen automatizacije i robotizacije procesa datekcije IPP (Titeca *et al.*, 2018). Iako je skorašnji napredak u proteomici impresivan, biće potrebna decenija ili duže da bi se dostigao željeni nivo analiza IPP (Syafrizayanti *et al.*, 2014). U tom kontekstu, prisutna je potreba za razvojem računarskih metoda za predviđanje IPP, kako bi se proces mapiranja celokupnog proteinskog interaktoma ubrzao (Jansen and Gerstein, 2004; Browne *et al.*, 2010b; Titeca *et al.*, 2018).

1.3 Baze podataka IPP

Velika količina podataka o IPP dobijena eksperimentalnim metodama detekcije IPP je pohranjena u odgovarajuće baze podataka. Prema izvoru i vrsti podataka koje sadrže baze podataka se dele na:

1. Primarne baze podataka koje sadrže podatke o eksperimentalno potvrđenim IPP
2. Sekundarne baze (meta-baze) podataka koje sadrže podatke integrisane iz primarnih baza podataka.
3. Baze podataka specijalizovane (u osnovi sekundarne) u odnosu na određeni fokus ili biološku vrstu.
4. Baze podataka koje pored eksperimentalno potvrđenih IPP sadrže i IPP predviđene računarskim metodama.

Iako se primarne baze podataka oslanjaju na eksperimentalne podatke objavljene u naučnoj literaturi, različite baze podataka ne prijavljuju iste IPP iz istih studija. Prema

analizi 14,899 publikacija koje su korišćene kao izvor podataka u najmanje dve baze podataka IPP, u 39% slučajeva u različitim bazama podataka je prijavljen različiti broj IPP iz istih studija (Lehne and Schlitt, 2009). Analiza većeg broja primarnih baza podataka utvrdila je iznenađujuće mali presek IPP koje ove baze sadrže (Cusick *et al.*, 2009; Turinsky *et al.*, 2010). Jedan od mogućih uzroka ovih nepodudaranja je korišćenje različitih identifikatora proteina od strane različitih baza podataka. Dodatno, prijavljene su razlike u načinu računanja i dodeljivanja vrednosti mere pouzdanosti IPP kod različitih baza podataka (Lehne and Schlitt, 2009). U cilju standardizacije podataka o IPP u jedinstven format, nekoliko vodećih institucija je ujedinilo napore u formiranje IMEx konzorcijuma (engl. International Molecular Exchange, IMEx). Važan element pored standardizacije formata jeste redovno usaglašavanje i ažuriranje baza podataka. Prema Gemovic i sar. (Gemovic *et al.*, 2018), problem je što neke baze podataka poput HPRD (engl. Human Protein reference Database, HPRD) (Keshava Prasad *et al.*, 2009), baze podataka koju potpisuje MIPS (engl. Munich information centre for protein sequences, MIPS), Mpact (Güldener *et al.*, 2006) i MPPI (Pagel *et al.*, 2005) ne ažuriraju redovno ili su u potpunosti prestale da dodaju nove informacije o IPP. Neke prethodno održavane baze podataka poput, MPIDB (engl. Microbial Protein Interaction database, MPIDB) (Goll *et al.*, 2008) i MINT (engl. Molecular Interaction, MINT) su integrisane u IntAct (Hermjakob *et al.*, 2004) bazu podataka. Konačno, postoje slučajevi potpunog uklanjanja baze podataka kao u slučaju BIND (engl. Biomolecular Interaction Network Database, BIND) (Gilbert, 2005). Podaci o IPP interakcijama u prethodno pomenutim slučajevima čuvaju se u sklopu sekundarnih baza podataka.

Tabela 1. Primeri primarnih, sekundarnih i meta-baza IPP podataka (Gemovic *et al.*, 2018).

	Baza podatka	Broja IPP	Broj IPP čoveka	Broj organizama	Poslednja verzija	Veb link
Primarne Baze podataka	**DIP**	81.731	9.078	834	Feb 2017	http://dip.doe-mbi.ucla.edu/dip/Main.cgi
	IntAct	454.760	237.385	>850	Jan 2017 (mesečno ažuriranje)	http://www.ebi.ac.uk/intact/
	BioGRID	479.503	278.645	61	Feb 2017 (mesečno ažuriranje)	https://thebiogrid.org/
Sekundarne baze podataka	**iRefIndex**	797.994	472.494	>1400	Apr 2015	http://irefindex.org/wiki/index.php?title=iRefIndex
	iRefWeb	542.927	222.098	>1400	Jun 2014	http://wodaklab.org/iRefWeb/
	mentha	643.329	275.002	8+	Mar 2017 (sedmično ažuriranje)	http://mentha.uniroma2.it/
	APID	678.441	349.144	>400	Jun 2016	http://cicblade.dep.usal.es:8080/APID/init.action
Meta-baze podataka	**HIPPIE**	287.357	287.357	1	Jun 2016	http://cbdm-1.zdv.uni-mainz.de/~mschaefer/hippie/index.php
	IID	1.741.568	911.446	6	Mar 2016	http://iid.ophid.utoronto.ca/SearchPPIs/protein/
	STRING	~933 mil		2031	Apr 2016	http://string-db.org/
	GeneMANIA	~538 mil		9	Oct 2014	http://genemania.org/
	ConsensusPathDB	840.792	534.634	3	Jan 2017	http://cpdb.molgen.mpg.de/

1.4 Računarske metode za predviđanje IPP

Mnoge eksperimentalne metode za predviđanje IPP koriste i računarske analize u određenoj meri. Metode sintetičke letalnosti i koekspresije gena su samo neki od primera (Shoemaker and Panchenko, 2007). Računarske analize u sklopu ovih metoda koriste se u funkciji statističke obrade i predviđanja funkcionalnih interakcija među potencijalnim IPP. Eksperimentalne metode opisuju manje od 25% pretpostavljene veličine ljudskog proteinskog interaktoma (Stelzl and Wanker, 2006). U svrhu dopune skupih i vremenski zahtevnih eksperimentalnih metoda, došlo je do razvoja brojnih računarskih metoda za predviđanje IPP (Pitre *et al.*, 2008). Računarske metode mogu pomoći u skraćivanju

vremena za kompletiranje proteinskog interaktoma, rangirajući listu potencijalnih kandidata IPP (Schwartz *et al.*, 2008). Računarske metode za predviđanje IPP koriste statističke, mrežne ili algoritme zasnovane na mašinskom učenju. Ove metode se, osim prema računarskom pristupu, razlikuju i prema tipu informacija koje koriste za predviđanje IPP, uključujući i metode koje u sklopu svog pristupa integrišu razne tipove informacija. Ulazne informacije na osnovu kojih se vrši predviđanje IPP mogu se svrstati u tri šira konteksta: strukturni, genomski i biološki kontekst (Skrabanek *et al.*, 2008). Svaka podela računarskih metoda može se uzeti isključivo uslovno, usled tendencije kombinovanja različitih tipova informacija sa različitim tipovima računarskih analiza, u cilju preciznije predikcije IPP.

1.4.1 Metode za predviđanje IPP zasnovane ne genomskom kontekstu

1.4.1.1 Metode zasnovane na kolokalizaciji gena

Među prvim metodama za predviđanje IPP razvijeni su pristupi zasnovani na zapažanju da susedni geni kodiraju interagujuće proteine. Hromozomska blizina gena može da ukazuje na fizičku interakciju njihovih produkata ili na funkcionalnu asocijaciju (Dandekar *et al.*, 1998; Overbeek *et al.*, 1999). Na osnovu ovog fenomena prvo su razvijene metode predikcije bakterijskih operona, kao i operona arhea (lat. Archea) (Ermolaeva, White and Salzberg, 2001). Analiza tri genoma bakterija i arhea pokazala je da 63-75% ko-regulisanih gena fizički interaguje (Huynen *et al.*, 2000). Ovi rezultati ukazuju na značajan stepen konzerviranosti ko-regulisanih gena u sklopu operona u sistematski udaljenim grupama organizama usprkos efektima prirodne evolucije (Shoemaker and Panchenko, 2007). Iako su operoni retki u eukariotima (Blumenthal, 1998), pokazano je da se geni uključeni u iste biološke procese često nalaze u genomu u relativnoj blizini i da su zajedno ko-regulisani. (Dandekar *et al.*, 1998). Ovaj pristup uključuje identifikaciju i analizu proteina kodiranih susednim genima u referentnim genomima i uspostavljanje praga intergenske distance ciljnih gena. Na osnovu analize strukture grupa gena u referentnim genomima, pretpostavlja se fizička interakcija gena u genomu koji se ispituje (Muley and Ranjan, 2012). Metoda, u osnovi jednostavna, je neosetljiva na fizičke interakcije funkcionalno povezanih, ali udaljenih gena. Iz tog

razloga ova metoda zahteva pažljiv izbor i iscrpnu analizu većeg broja referentnih genoma. Kolokalizacija gena se kao pristup za predviđanje IPP najčešće koristi za manje proteome.

1.4.1.2 Metode zasnovane na analizi filogenetskih profila

Iza grupe metoda zasnovanih na analizi filogenetskih profila stoji hipoteza da se informacija o fizičkoj interakciji proteina može predvideti na osnovu analize homologije među genima. Homologni geni koji imaju analogne funkcije u različitim organizmima i imaju zajedničkog genskog pretka među evolutivno starijim vrstama se nazivaju ortolozi. Analiza prisutnosti gena u različitim vrstama može ukazivati na njihovu funkcionalnu povezanost ili fizičku interakciju (Skrabanek *et al.*, 2008). Za razliku od metoda zasnovanih na kolokalizaciji gena, prisustvo gena u ispitivanom genomu ne mora da uključuje hromozomska blizinu. Proučavanjem zajedničkog pojavljivanja ciljnih gena u većem broju genoma formiraju se filogenetski profili. Profili su predstavljeni u formi matrica kod kojih se prisutnost ili odsutnost ciljnih gena u ispitivanim genomima zapisuje u binarnom sistemu (Pellegrini *et al.*, 1999). Slični filogenetski profili mogu ukazivati na fizičku ili funkcionalnu asocijaciju između gena. Iako metode zasnovane na filogenetskim profilima mogu detektovati određene interakcije koje metode zasnovane na kolokalizaciji gena ne mogu, ovaj pristup značajno zavisi od broja i distribucije referentnih genoma, stoga nije primenljiv na esencijalne proteine, i ne može se koristiti na nekompletnim genomima (Panchenko, 2008). Specifičnosti eukariotskih genoma ograničavaju efikasnost ovog pristupa na IPP eukariota (Galperin and Koonin, 2000).

1.4.1.3 Fuzija gena (Rosetta Stone)

Komplementarne metodama zasnovanim na kolokalizaciji gena i filogentskim profilima, metode zasnovane na fuziji gena prevode asocijaciju među genima u fizičku asocijaciju njihovih produkata. Fuzija gena je fenomen povezivanja srodnih gena u jedan funkcionalnih gen, čiji se produkt označava kao „Rosetta Stone“ protein. Jedan od primera produkata fuzije gena je topoizomeraza II kod *Escherichia coli* (Wang, 1985). Fuzija gena se može posmatrati kao specijalan slučaj kolokalizacije gena malog početnog

hromozomskog rastojanja koji su naknadno spojeni u jedan gen, što se pripisuje evolutivnoj sili smanjenja nivoa regulacije pojedinačnih gena u odnosu na fuziju (Enright *et al.*, 1999). Korišćenjem „Rosetta stone“ metoda pronađeno je 6809 potencijalno interagujućih parova kod *Escherichia coli*. Pristup uključuje pronalaženje parova nehomolognih proteina koji u referentnim genomima imaju visok stepen homologije sa pojedinačnim proteinom. Dodatno analiza ovakvih parova ukazala je da su u više od 50% pronađenih slučajeva nehomologni proteini funkcionalno asocirani (Marcotte *et al.*, 1999). Na smanjenu pouzdanost ovog pristupa kod predviđanja IPP eukariota utiče pojavljivanje velikog broja *promiskuitetnih* domena (Enright, Van Dongen and Ouzounis, 2002) i relativno retki slučajevi fuzije proteina (Panchenko, A.; Przytycka, 2008). Zajednička karakteristika svih metoda za predviđanje IPP zasnovanih na geneomskom kontekstu jeste visok stepen računarske zahtevnosti.

1.4.2 Metode za predviđanje IPP zasnovane na biološkom kontekstu

1.4.2.1 Metode zasnovane na ekspresiji gena , kolokalizaciji i funkcionalnoj asocijaciji

Analizom obrazaca ekspresije velikog broja gena bakteriofaga T7 i kvasca (*Saccharomyces cerevisae*) utvrđena je statistički značajna korelacija između zajedničke ekspresije gena i fizičke interakcije proteina koje kodiraju (Grigoriev, 2001). Za ostvarenje fizičke interakcije među proteinima neophodno je da interagujući proteini budu simultano prisutni u istom delu ćelije, što je najčešće i vidljivo u obrascima ekspresije gena koji ih kodiraju. Vodeći se ovom opservacijom, razvijene su metode koji analiziraju podatke dobijene analizom DNK mikroereja u cilju predikcije novih, i evaluacije pouzdanosti postojećih IPP. U prvoj fazi, ove metode računaju koeficijent korelacije koekspresije gena korišćenjem informacija iz DNK mikroereja. Ukoliko ovaj koeficijent prelazi unapred definisani prag, geni se označavaju kao koekspresovani. Primenom statističkih tehnika grupisanja, geni koji se zajednički eksprimiraju, zajedno se grupišu i analiziraju se njihovi funkcionalni odnosi. Na bazi ovog pristupa razvijeni su metode za evaluaciju postojećih skupova IPP (Deng, Sun and Chen, 2002), kao i statistički modeli za predviđanje novih IPP (Rhodes *et al.*, 2005). Neka od ograničenja

ovog pristupa se odnose na tip IPP koji se pomoću njega mogu pretpostaviti. Stepen korelacije između ekspresije gene i IPP je znatno veći u slučaju permanentnih IPP nego u slučajevima tranzijentnih IPP (Jansen, Greenbaum and Gerstein, 2002). Dodatno, nivo proteina u ćeliji ne pokazuje savršenu korelaciju sa nivoom ekspresije gena (Keskin, Tuncbag and Gursoy, 2016). Nadalje, interakcija među ciljnim proteinima se pretpostavlja indirektno, pa je ovaj pristup adekvatan kao dopuna postojećim računarskim ili eksperimentalnim metodama za pretpostavljanje IPP.

Studija poređenja IPP čoveka, kvasca (*Saccharomyces cerevisiae*), mušice (*Drosophila melanogaster*) i crva (*Caenorhabditis elegans)* pokazala je da proteini lokalizovani u istom delu ćelije imaju veću verovatnoću formiranja interakcije (Gandhi *et al.*, 2006). Zajednička lokalizacija (kolokalizacija) proteina može da ukazuje na njihovo potencijalno fizičko povezivanje, ali korelacija između ova dva biološka fenomena nije potvrđena u određenim delovima ćelije, kao i za određene proteine (Gandhi *et al.*, 2006). Iako se može koristiti kao indicija postojeće IPP, zajednička lokalizacija proteina u određenom ćelijskom delu nije dovoljno pouzdana kao samostalan prediktor IPP. Sa druge stran, IPP mogu biti korišćeni kao osnova za predviđanje subćelijske lokalizacije (Shin *et al.*, 2009).

Interagujući proteini često, pored zajedničke lokalizacije, učestvuju u istom biološkom procesu pa je za očekivati da imaju zajedničke funkcionalne anotacije (Voter, Manthei and Keck, 2016). Nekoliko studija je sličnost GO anotacije okarakterisalo kao jedan od najpouzdanijih prediktora interakcije među proteinima. Čitav niz metoda je razvijen na principu korišćenja GO sličnosti među proteinima, kao samostalne ili jedne od izvora informacija, u predikciji IPP, metodama zasnovanim na analizama GO mreža ili korišćenjem mašinskog učenja (Maetschke *et al.*, 2012; Zhang and Tang, 2016). Peng i sar. (Peng *et al.*, 2017) navode niz metoda koje analiziraju sličnost GO termina proteina čije interakcije su prijavljene u postojećim skupovima IPP u svrhu evaluacije pouzdanosti njihovih IPP.

1.4.3 Metode za predviđanje IPP zasnovane na strukturnom kontekstu

Računarsko predviđanje IPP zasnovana na strukturnom kontekstu koristi informacije raznih nivoa proteinske strukture: primarne strukture, tercijarne strukture i proteinskih

domena. Korišćenjem strukturnog konteksta, računarske metode za predviđanje IPP imaju dva cilja: (i) predviđanje da li, i sa kakvim stepenom pouzdanosti, dva proteina fizički interaguju i (ii) identifikacija delova proteina i aminokiselinskih ostataka, kojima se ostvaruje povezivanje (Skrabanek *et al.*, 2008).

1.4.3.1 Metode koje koriste informacije o proteinskim domenima

Proteinski domen je stabilna, fundamentalna jedinica tercijarne strukture proteina koja može da ostvaruje svoju funkciju, evoluira i pakuje se nezavisno od ostatka proteina. Proteini svoje funkcije ostvaruju preko IPP, u čijoj osnovi je često interakcija među specifičnim proteinskim domenima (Deng *et al.*, 2002; Khor, 2014). Razvijen je veći broj metoda za predviđanje IPP, koje se oslanjaju na detektovanje i kvantifikaciju interakcija među proteinskim domenima (Sprinzak and Margalit, 2001; Wan, Park and Suh, 2002; Ng, Zhang and Tan, 2003). Međudomenske interakcije se kvantifikuju merenjem snage interakcije među proteinskim domenima na osnovu 3D kristalne strukture proteina (Singhal and Resat, 2007). Ovakve, eksperimentalno potvrđene, međudomenske interakcije se smatraju visoko kvalitetnim skupom podataka. Pored eksplicitne potvrde, interakcije među proteinskim domenima se mogu detektovati implicitno. U nizu studija o eksperimentalno potvrđenim IPP analiziraju se frekvencije pojavljivanja parova proteinskih domena u parovima proteina koji interaguju, pri čemu se analizira interakcija domena d1 proteina X sa domenom d2 proteina Y. Korelacija među frekvencijom pojavljivanja interagujućih domena kod parova interagujućih proteina, kao i verovatnoća njihove interakcije se koristi kao osnov za statističku analizu sa ciljem predviđanja potencijalnih IPP (Singhal and Resat, 2007; Khor, 2014). Izazovi za korišćenje interakcija među proteinskim domenima kao osnove za pretpostavljanje IPP se ogledaju u opservaciji da se veliki broj ne formira preko interakcija domenima (Schelhorn, Lengauer and Albrecht, 2008), a kod multidomenskih proteina, gde domeni jesu medijatori IPP, ne učestvuju svi domeni u formiranju interakcije (Khor, 2014). Pored toga, formiranje IPP može da zahteva učešće nekoliko domena IPP jednog proteina sa jednim ili više domena drugog proteina, mada većina studija pretpostavlja 1-na-1 interakciju među domenima.

1.4.3.2 Metode koje koriste tercijarnu strukturu proteina

Kod grupe metoda koje koristi tercijarnu strukturu proteina kao osnov za predviđanje IPP, interakcija dva ciljna proteina se predviđa na osnovu 3D strukture eksperimentalno okarakterisanih proteinskih kompleksa. Jedna od korišćenih strategija je analiza sličnosti tercijarnih struktura potencijalnih interaktora sa 3D strukturama proteina poznatog interagujućeg kompleksa (Hue *et al.*, 2010). Neke metode koriste sličnost proteinske sekvence kao osnov za pretpostavljanje sličnosti strukture. Homologija sekvenci ispitivanog proteinskog para sa proteinskim parom opisanog kompleksa interakcije se uzima kao baza za statističko modeliranje (Mukherjee and Zhang, 2011). Jedan od metoda koji koristi takav pristup je InterPreTS (Aloy and Russell, 2003), koji koristi program BLAST (engl. Basic Local Alignment Search Tool) za pronalaženje homolognih sekvenci među interagujućim parovima poznate 3D strukture. Slično, HOMCOS omogućava (Fukuhara and Kawabata, 2008) modeliranje potencijalne IPP na osnovu homologije sekvenci sa proteinima opisanog 3D kompleksa, uz dodatak analize potencijalne energije među kontaktnim površinama proteina. Modeliranje zasnovano na homologiji je primenjivo jedino u slučaju visokog stepena sličnosti sekvenci između ciljnih proteina i proteina sa poznatom tercijarnom strukturom. U slučaju 30-40% sličnosti sekvence, parovi proteina mogu posedovati potpuno različitu strukturnu topologiju interaktivne površine (Gemovic *et al.*, 2018). Da bi se omogućilo modeliranje interakcija proteinskih parova u slučajevima niskog stepena homologije sa poznatim strukturama, uvedena je tehnika tredovanja (engl. threding). U posebnu grupu metoda koje koriste 3D strukturu za modeliranje strukture interagujućeg kompleksa su doking (engl. docking) metode (Smith and Sternberg, 2002).

Doking metode se oslanjaju na fizičko-hemijske karakteristike i 3D strukturu interaktora u traženju optimalnog kompleksa (Vreven *et al.*, 2014). Osnovni cilj aplikacije doking metoda jeste pronalaženje optimalne interagujuće strukture između dva proteina. Znatno ređe se upotrebljavaju za predviđanje IPP (Aloy and Russell, 2006). Razvijene su metode koje pomoću statističke analize doking kompleksa potencijalnih interaktora pokušavaju diskriminisati prave interaktore od manje verovatnih (Wass *et al.*, 2011). Metode zasnovane na informacijama o proteinskoj strukturi omogućavaju detaljniju analizu proteinskih interakcija u odnosu na metode zasnovane na genomskom

kontekstu. Pored predikcije IPP, strukturno zasnovani pristupi omogućavaju analizu fizičkih karakteristika interakcije i preciznu detekciju aminokiselinskih ostataka proteinskog interfejsa, koje ostvaruju fizički kontakt (Skrabanek *et al.*, 2008).

Strukturno zasnovani metodi za predviđanje IPP, pokazuju ograničenu efikasnost u modeliranju interakcija među nestruktuiranim proteinima (Aloy and Russell, 2006). Nadalje, metodi zasnovani na strukturi se za svoje analize oslanjanju na poznate 3D strukture proteina pohranjene u dostupnim bazama podataka. Zbog visoke zahtevnosti resursa potrebnih da bi se eksperimentalnim metodama rešila 3D struktura proteina, broj poznatih 3D struktura je relativno malen u odnosu na broj sekvenciranih proteina. Dodatno, visoka računarska zahtevnost metoda koje koriste 3D strukturu proteina, onemogućava njihovu efikasnu upotrebu za predviđanje IPP na velikoj skali (Zahiri, Bozorgmehr and Masoudi-Nejad, 2013).

1.4.4 Metode za predviđanje IPP zasnovane na sekvenci proteina

Prednost korišćenja primarne strukture proteina za predviđanje IPP jeste u njenoj dostupnosti i univerzalnosti, te je stoga jedan od najčešće korišćenih osnova pristupa za predviđanje IPP (Shen *et al.*, 2007; Valente *et al.*, 2013). Sekvenca proteina sadrži dovoljno informativnosti za efikasnu predviđanje IPP bez uključivanja informacija o biološkom i genetskom kontekstu, kao i bez drugih strukturnih podataka (Csermely *et al.*, 2013). Metode za predviđanje IPP sastoje se najčešće iz dva elementa: matematičke reprezentacije sekvence proteina i metoda statističkog učenja.

1.4.4.1 Modeliranje primarne strukture proteina

Korišćenje metoda mašinskog učenja zahteva predstavljanje podataka koji se modeliraju u formi skupa instanci, koje su predstavljene numeričkim nizovima istih dužina. Kod metoda mašinskog učenja proteinske sekvence su predstavljene numeričkom sekvencom (vektorom) čija dužina predstavlja broj dimenzija, a vrednosti pojedinačnog elementa vektora predstavlja poziciju tog vektora u višedimenzionom prostoru. Koordinate vektora se još označavaju kao atributi (engl. features). Svaka instanca skupa za učenje je opisana vektorom vrednosti definisanih atributa. Najvažniji faktor u formiranju prediktivnog

modela visoke diskriminativne moći je izbor atributa kojim se primeri opisuju (Guyon, 2003; Domingos, 2012).

Za formiranje efikasnog metoda za predviđanje IPP zasnovanog na sekvenci proteina krucijalno je efikasno modeliranje proteinskih sekvenci varijabilnih dužina, predstavljenih alfabetskim nizom kombinacije 20 aminokiselina, u numeričke vektore fiksne dužine. Najveći broj metoda za predviđanje IPP koristi prosto algebarsko spajanje pojedinačnih vektora interagujućih proteina u cilju formiranja numeričke reprezentacije interagujućeg para (Park, 2009). Metode mašinskog učenja zasnovane na kernelima (engl. kernel methods) omogućavaju upotrebu kompleksnijih pristupa za analizu relacija između numeričkih vektora interagujućih proteina.

Pristupe numeričkoj reprezentaciji proteinskih sekvenci možemo podeliti u 4 grupe:

a. Pristupi zasnovani na k-mer reprezentaciji
b. Pristupi zasnovani na odnosima fizičko-hemijskih karakteristika aminokiselina
c. Pristupi zasnovani na evolutivnim profilima
d. Kombinovani pristupi (kombinacija više tipova atributa)

1.4.4.1.1 Pristupi zasnovani na k-mer reprezentaciji

K-mer reprezentacija podrazumeva prevođenje proteinske sekvence u vektor normalizovanih frekvencija subsekvenci (Ben-Hur and Noble, 2005). Najčešće se koriste frekvencije pojedinačnih aminokiselina (1-mer) , duada (2-mer) i trijada (3-mer) iako se pojavljuju reprezentacije sa $k \geq 4$.

U 1-mer modelu proteinske sekvence, označen kao aminokiselinska kompozicija - AAC (engl. Amino Acid Composition, AAC), računa se broj pojavljivanja svake od 20 aminokiselina u sklopu proteinske sekvence i predstavlja u formi relativnih frekvencija, odnosno brojeva pojavljivanja aminokiselina podeljenih sa ukupnim brojem aminokiselinskih ostataka. Na ovaj način proteinska sekvenca se predstavlja vektorom vrednosti 20 atributa (Chou, 2001).

Slično tome, dipeptidna kompozicija (engl. Dipeptide Composition, DC) predstavlja frekvenciju pojavljivanja svih kombinacija parova aminokiselina unutar sekvence

proteina. Na ovaj način proteinska sekvenca je predstavljena vektorom od 400 numeričkih vrednosti.

Analiza interagujućih domena proteina koji interaguju otkrila je specifičan profil pojavljivanja pojedinačnih AAC monomera i dimera. Analiza statističke značajnosti frekvencija pojavljivanja određenih AAC monomera i dimera u interagujućim domenima ukazala je na potrebu korišćenja čitavog spektra frekvencija AAC proteinske sekvence za optimalne rezultate u problemu predikcije IPP (Roy *et al.*, 2009). Model AAC je korišćen kao osnova za formiranje univerzalnog prediktora IPP, formiranog koristeći IPP iz više vrsta eukariota (Valente *et al.*, 2013).

Usled značajnog povećanja broja atributa, a samim tim i dimenzionalnosti skupa za učenje, K-mer reprezentacije veće od k-3 reprezentacije, TC (engl. tripeptide composition) u nemodifikovanoj formi su retke. Martin i sar. (Martin, Roe and Faulon, 2005), u svom pristupu koriste frekvencije svih k-3 tipova subsekvenci u predstavljanju pojedinačnih proteina. Proteinski par se reprezentuje u formi skalarnog umnožka proteina uključenih u ciljnu interakciju, kako bi se obezbedila računarska efikasnost. Nekih autori (Shen *et al.*, 2007; Pan, Zhang and Shen, 2010), 20 aminokiselina grupišu u 7 klasa prema karakteristikama bočnih lanaca. Aminokiseline koje pripadaju istoj klasi se posmatraju identično. Na ovaj način se efektivno posmatraju trijade klasa grupisanih aminokiselina umjesto trijada pojedinačnih aminokisleina gde svaku od tri pozicije može zauzeti bilo koja od 20 potencijalnih aminokiselina. Sekvenca proteina se tako predstavlja vektorom dužine 343 elementa (7x7x7) u odnosu na predstavljanje koje bi uključivalo 8000 atributa (20x20x20) (Xiao, Xu and Cao, 2014). Da bi se formirao vektor koji predstavlja interagujući par, autori su koristili prosto spajanje dva vektora atributa interagujućih proteina (Shen *et al.*, 2007). Komparativna analiza je pokazala da korišćenje trijada klasa u odnosu na trijade 20 mogućih aminokiselina ima značajan uticaj na smanjenje tačnosti modeliranja i predviđanja IPP (Park, 2009). U cilju poboljšanja pristupa trijada klasa aminokiselina korišćeni su različiti kriterijumi grupisanja 20 aminokiseline u klase (Shoyaib and Abdullah-Al-Wadud, 2010), kao i upotreba mehanizama za procenu značajnosti frekvencija trijada zasnovanog na verovatnoći (Yu, Chou and Chang, 2010). Jedan od daljnjih pokušaja smanjenja vremena modeliranja velikog broja IPP uz poboljšanje predikcione sposobnosti je opisali su Guarracino i sar. (Guarracino *et al.*,

2010), koji su vektor klasa trijada opisan prema Shen i sar. (Shen *et al.*, 2007) filtrirali metodama selekcije atributa.

Za optimizaciju veličine vektora frekvencija grupa aminokiselina koji predstavlja protein, uz zadržavanje informacije o interakciji udaljenih delova proteinske sekvence, koristi se segmentacioni pristup. Umesto računanja k-mer frekvencija na čitavoj sekvenci, uzimaju se nasumično izabrani segmenti sekvence (Ren *et al.*, 2011). Kontinuirani i diskontinuirani segmenti aminokiselina različite dužine u sklopu proteinske sekvence igraju važnu ulogu u karakterizaciji IPP (You *et al.*, 2014; You, Chan and Hu, 2015).

1.4.4.1.2 Pristupi zasnovani na odnosima fizičko-hemijskih karakteristika aminokiselina

Interakcije među proteinima se često ostvaruju fizičkim kontaktima među udaljenim delovima sekvence (Guo *et al.*, 2008). Sa druge strane, upotreba većih segmenata u k-mer predstavljanju proteina u svrhu predviđanja IPP je često računarski neizvodljivo.
Deskriptori zasnovani na korelaciji omogućavaju uključivanje informacije o relacijama između međusobno udaljenih aminokiselina u proteinskoj sekvenci. Korelacioni deskriptori predstavljaju proteinsku sekvencu u formi vektora korelacija između fizičko-hemijskih karakteristika aminokiselina u sklopu pomerajućeg prozora uzduž sekvence proteina (Xiao *et al.*, 2015). Na konačan broj ovako izdvojenih atributa utiču veličina prozora koji se posmatra i broj osobina aminokiseline čijim numeričkim vrednostima se aminokiseline predstavljaju. Guo i sar. (Guo *et al.*, 2008) računaju autokorelacije između 7 fizičko-hemijskih karakteristika aminokiseline u okviru pomerajućeg prozora širine 20 aminokiseline uzduž sekvence u svrhu vektorskog predstavljanja proteina. Xia i sar. posmatraju relacije između 6 fizičko-hemijskih osobina aminokiselina u pomerajućem segmentu dužine 30 aminokiseline opisujući tako proteine numeričkim nizom od 180 elemenata (Xia, Han and Huang, 2010). Shi i sar. (Shi *et al.*, 2010) posmatraju efekat relacija računanjem koeficijenta korelacija između 12 fizičko-hemijskih i topoloških karakteristika aminokiseline uzduž sekvence.

1.4.4.1.3 Pristupi zasnovani na evolutivnim profilima

Za analizu evolutivnih profila proteina, koriste se PSSM (engl. Position Specific Scoring Matrix) matrice. PSSM čuva informacije o verovatnoći supstitucije svake od 20 mogućih aminokiselina na specifičnoj poziciji duž sekvence (Altschul, 1997). Frekvencije supstitucija na specifičnoj poziciji unutar proteinske familije, čuvane u formi PSSM, su u formi pozitivnih i negativnih vrednosti. Negativne vrednosti ukazuju da je specifična aminokiselina ređe zamjenjena drugom aminokiselinom u sklopu proteinske familije (Dehzangi *et al.*, 2017). Svaka proteinska sekvenca se na ovaj način može predstaviti L x 20 PSSM matricom, gde L predstavlja dužinu proteinske sekvence. Proces formiranja PSSM matrica zahteva korišćenje poravnavanje (multiple alingments) proteinskih sekvenci čuvanih u bazama podataka proteina, detektujući tako informacije o homologiji proteina i pripadnosti proteinskim familijama (Waris *et al.*, 2016).

Budući da PSSM matrica zadržava dužinu sekvence proteina, efikasno korišćenje ovako čuvane evolutivne informacije zahteva neki od metoda ekstrakcije atributa iz PSSM matrice. Evolutivni profili formirani na osnovu informacije o sličnostima sekvence su se pokazali osnovom za visoko efikasne skupove atributa u raznim bioinformatičkim problemima (Rangwala and Karypis, 2005; Ye, Wang and Altschul, 2011; Hamp and Rost, 2015a).

1.4.4.1.4 Kombinovani pristupi (kombinacija više tipova osobina sekvenci)

Kombinovanje više skupova atributa u jedan ili više vektorskih reprezentacija je čest pristup u cilju poboljšavanja efikasnosti razdvajanja klasa modelom mašinskog učenja za predviđanje IPP. Jedan od često korišćenih kombinovanih deskriptora u predikciji IPP su deskriptori pod nazivom PAAC (engl. Pseudo-Amino Acid Composition) (Chou, 2001). Kombinovanjem aminokiselinske kompozicije i autokorelacije između fizičko-hemijskih karakteristika, važnih za interakciju među proteinima, povećava se informativnost predikcionog modela u odnosu na korišćenje pojedinačnih grupa atributa.

Dong i sar. (Dong, Zhou and Liu, 2010) upotrebljavaju kombinaciju k-mer reprezentacija, motiva i evolutivnih profila dobijenih poravnavanjima proteinskih sekvenci uz statističke tehnike smanjenja broja atributa u svrhu numeričkog predstavljanja proteinskih sekvenci.

Slično prethodnom pristupu, Zhao i sar. (Zhao, Ma and Yin, 2012) kombinuju nekoliko tipova deskriptora ekstrahovanih iz sekvenci uključujući korelacione, k-mer i PAAC deskriptore, gde su od ukupno 930 atributa metodama statističke selekcije odabrali 67 za formiranje finalnog vektora reprezentacije. U sklopu DXEC metode za predviđanje IPP, integrisano je šest grupa atributa zasnovanih na sekvenci, koristeći statističke metode za odabir najefikasnijih atributa (Du *et al.*, 2014). Slično, nekoliko grupa atributa zasnovanih na k-mer reprezentaciji i autokorelaciji čine osnovu DeepPPI metoda (Du *et al.*, 2017). Jedan od izazova u dizajniranju informativnog načina predstavljanja proteinskih sekvenci numeričkim vektorima, posebno izraženog kod kombinovanih atributa, jeste zadržati nizak stepen dimenzionalnosti reprezentacije.

1.4.5 Metode za predviđanje IPP zasnovane na mašinskom učenju

1.4.4.2 Osnove i pristupi mašinskog učenja

Generisanje velike količine eksperimentalnim podataka omogućilo je razvoj bioinformatičkih metoda zasnovanih na mašinskom učenju u cilju transformacije heterogenih podataka u relevantno biološko znanje o ćelijskim mehanizmima i procesima.

Mašinsko učenje (engl. Machine learning, ML) je oblast veštačke inteligencije koja se bavi razvojem računarskih metoda i algoritama sposobnih da poboljšavaju svoje performanse učeći na osnovu dostupnih informacija (Jordan and Mitchell, 2015). Protokoli mašinskog učenja omogućavaju automatsko učenje računarskih programa iz podataka i željenog izlaza, za razliku od tradicionalnog programiranja gde se kreira program koji nad željenim podacima formira ciljni izlaz (Domingos, 2012). Sposobnost automatskog učenja identifikacijom strukture podataka se upotrebljava u slučajevima kada je: (i) ekspertsko znanje nepotpuno ili netačno, (ii) količina podataka u toj meri velika da onemogućuje efikasno ručno modeliranje i (iii) kada generalna pravila nisu primenjiva na specifične slučajeve (Yip, Cheng and Gerstein, 2013). Ovakav pristup automatskom učenju omogućava da se naučeni programi (modeli) koriste u svrhu predikcije na podacima koji nisu korišćeni u procesu učenja. Automatizacija procesa kreiranja modela algoritmom mašinskog učenja minimizira mogućnost ljudske greške i

naklonjenost određenom rešenju. Na ovaj način algoritmi mašinskog učenja omogućavaju ponovljivo i automatsko kreiranje predikcionih modela zasnovanih na konkretnim podacima (Tarca *et al.*, 2007). Zahvaljujući tome, mašinsko učenje se primenjuje na velikom broju bioinformatičkih problema uključujući: analize signalnih puteva i mreža; analizu i pretprocesiranje podataka mikroereja; predikciju strukture funkcije i interakcija proteina, anotacije gena i proteina, detekciju i identifikaciju kodirajućih regiona, mesta vezivanja transkripcionih faktora i mesta vezivanja promotora; funkcionalne analize gena, predviđanje funkcije gena i događaja alternativnog splajsovanja, itd. (Larrañaga *et al.*, 2006).
U oblasti mašinskog učenja izdvajaju se dve paradigme pristupa učenju:

- Nadgledano učenje (engl. supervised learning) podrazumeva vrstu učenja gde se algoritmu za učenje pored podataka na kojima uči obezbeđuju i željeni ishodi.
- Nenadgledano učenje (engl. unsupervised learning) predstavlja pristup automatskom učenju gde se algoritmu učenja pružaju podaci ali bez izlaza, pri čemu algoritam pretražuje podatke u cilju pronalaženja sličnosti među primerima.

Kod nadgledanog učenja, algoritam uči na podacima koji se sastoje od kolekcije primera za učenje (instanci) opisanih skupom atributa i ciljnim ishodom. Svaka instanca iz skupa za učenje je opisana vrednostima atributa u vektorskoj formi. Ciljni ishod kod problema klasifikacije može biti pripadnost određenoj klasi, gde je broj klasa ceo broj veći od nule (Domingos, 2012). Ovakav problem nadgledanog učenja se označava klasifikacijom, a u oblasti bioinformatike najčešće se koristi dvoklasni klasifikacioni sistem, gde se algoritmom mašinskog učenja formira model za klasifikaciju (klasifikator) instanci u jednu od dve klase (Larrañaga *et al.*, 2006; Kelchtermans *et al.*, 2014). Cilj nadgledanog učenja jeste da, pronalaženjem zakonitosti (obrazaca) u podacima, algoritam mašinskog učenja generiše model sposoban da klasifikuje podatke koji nisu korišćeni u procesu učenja, samo na osnovu vrednosti njihovih atributa (Tarca *et al.*, 2007). Pored pripadnosti klasi, moguće je formirati modele ML za predviđanje ciljnog ishoda u formi kontinuiranih brojeva, gde se ovakav proces naziva regresija. Učenje algoritmima mašinskog učenja je vrsta induktivnog učenja gde je osnovni cilj generalizacija znanja naučenog na manjem skupu podataka na njegov nadskup (Shalev-Shwartz and Ben-David, 2014). Ovakva vrsta

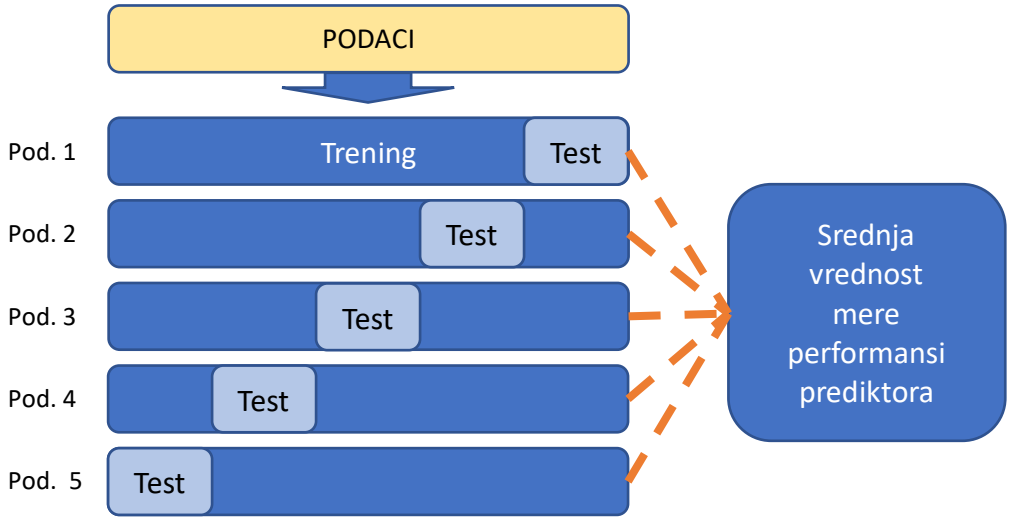
učenja se naziva modeliranje. Iako su razvijeni različiti algoritmi mašinskog učenja, od čega se mnogi primjenjuju u oblasti bioinformatike (Larrañaga *et al.*, 2006), svaki pristup mašinskog učenja uključuje tri elementa: reprezentaciju, evaluaciju i optimizaciju (Domingos, 2012). Za proces modeliranja neophodno je definisati ciljnu funkciju (engl. target function) koja u slučaju binarnog klasifikacionog problema predikcije IPP podrazumeva svrstavanje instanci IPP, definisanih vektorom vrednosti atributa, u jednu od dve klase: grupa interagujućih ili neinteragujućih proteina. Reprezentacija podrazumeva formalno predstavljanje klasifikatora u formi koju računarski hardver može da interpretira. Svaka reprezentacija klasifikatora ograničava potencijalni prostor hipoteza (engl. hypothesis space) na one hipoteze koje mogu biti predstavljene datim pristupom. Svaka, od hipoteza (modela) koje mogu biti naučene, predstavlja aproksimaciju ciljne funkcije (Domingos, 2012). U bioinformatici najčešće korišćenje reprezentacije su modeli potpornih vektora, neuronske mreže, stabla odlučivanja i aproksimacije zasnovane na Bajesovoj teoremi (Larrañaga *et al.*, 2006). Kako je svaki klasifikator iz prostora hipoteza samo aproksimacija ciljne funkcije, naučene hipoteze se razlikuju prema stepenu greške u odnosu na ciljnu funkciju. Prema tome zadatak učenja modela se sastoji iz minimazacije greške izračunate internom evaluacionom funkcijom (engl. scoring function) (Larrañaga *et al.*, 2006; Domingos, 2012). Kod klasifikacionog problema predikcije IPP, cilj je pronaći hipotezu sa najmanjom greškom kategorisanja instanci IPP iz skupa za učenje u odnosu na ciljnu funkciju. Proces traženja hipoteze sa minimalnom greškom pri aproksimaciji ciljne funkcije, u prostoru hipoteza, se označava optimizacionim korakom u sklopu učenja algoritmom ML (Domingos, 2012).

Iako je proces traženja optimalnog klasifikatora zavisan od podataka na kome se uči, konačan cilj pretrage je sposobnost generalizacije klasifikatorom na podacima koji nisu korišćeni u procesu učenja, jer u suprotnom bi dovoljno bilo memorisati podatke na kojima se uči (Domingos, 2012; Shalev-Shwartz and Ben-David, 2014; Chicco, 2017).

1.4.4.3 Procena modela mašinskog učenja

Da bi se nepristrasno evaluirala sposobnost klasifikatora predviđanja na novim podacima koristi se šema podele podataka za učenje na podatke za trening i podatke za testiranje (Efron, 1983). Za objektivnu procenu kvaliteta naučenog modela neophodno je

da test skup bude disjunktan u odnosu na trening skup (Bhaskar, Hoyle and Singh, 2006). Najčešće korišćena šema višestruke podele na trening i test skupove je unakrsna validacija. U procesu unakrsne validacije (Ng, 1997) (engl. cross-validation, CV) se skup za učenje podeli na n jednakih delova. U toku treniranja koristi se n-1 delova, dok se na preostalom podskupu podataka vrši provera kvaliteta modela. Proces se ponavlja dok svaki od n disjunktnih delova ukupnog skupa za učenje ne bude iskorišćen kao test skup (Slika 3). Broj n zavisi kako od broja tako i karakteristika ulaznih podataka (Larrañaga *et al.*, 2006; Chicco, 2017). Kod problema predikcije IPP najčešće vrednosti su n = 3, 5, 10, pri čemu se šeme onda odgovarajućih unakrsnih validacija označavaju sa CV-3, CV-5, CV-10 (Bhaskar, Hoyle and Singh, 2006).



Slika 3. Šema unakrsne validacije sa podelom originalnih podataka na pet skupova za treniranje i testiranje. Računa se srednja vrednost prediktivnih performansi modela formiranih u svakoj podeli (engl. fold) na odgovarajućem skupu za treniranje i testiranih na odgovarajućem skupu za testiranje. U svakom deljenju proces treniranja i testiranja se vrši na različitim delovima ulaznih podataka.

Iako šema podele na podskup instanci za treniranje i testiranje omogućava objektivnu proveru sposobnosti generalizacije modela, pronađena hipoteza ne mora biti najbolja aproksimacija ciljne funkcije. Veći broj algoritama mašinskog učenja sadrži dodatne parametre (meta-parametre) čije fino podešavanje može značajno uticati na prediktivne performanse konačnog modela (Olson *et al.*, 2017). Pored toga, ni jedan algoritam

mašinskog učenja ne omogućava pronalaženje optimalne hipoteze na svakom problemu, što je formalizovano u formi „no-free-lunch“ teoreme (Wolpert, 2002). Budući da bi korišćenje test skupa u fazi finog nameštanja parametara algoritma, u okviru faze učenja ili izbora algoritma, bila kontaminacija, neophodno je izdvojiti dodatni test skup za finalno testiranje izabranog modela (Chicco, 2017). Čest pristup je korišćenje unakrsne validacije na podskupu za trening u svrhu izbora algoritma mašinskog učenja ili podešavanje njegovih parametara (Domingos, 2012). Na ovaj način, kod problema mašinskog učenja greška na test skupu (varijansa, engl. variance) se pokušava aproksimirati na osnovu greške na skupu za trening (sistemsko odstupanje, engl. bias). Obučavanje efikasnog modela mašinskog učenja za efikasnu generalizaciju zavisi od balansa između varijanse i sistemskog odstupanja. Složeniji algoritmi smanjuju sistemsko odstupanje na uštrb povećanja varijanse i obrnuto (Domingos, 2012), pa proces selekcije modela podrazumeva prvo biranje jednostavnijih modela (Chicco, 2017). Formiranje kvalitetnog modela zavisi od izbora algoritma i podešavanja njegovih meta-parametara (Bhaskar, Hoyle and Singh, 2006; Shervashidze *et al.*, 2011; Chicco, 2017; Olson *et al.*, 2017).

1.4.4.4 Kreiranje atributa

Kreiranje i selekcija optimalnog skupa atributa je najvažniji korak kod formiranja efikasnog prediktivnog modela (Chicco, 2017). U bioinformatici je za efikasnu formalnu reprezentaciju algoritmom mašinskog učenja često neophodna selekcija relevantnih atributa. Praktična primena algoritama mašinskog učenja ukazuje da najveći uticaj na sposobnost generalizacije klasifikatora ima generisanje i selekcija atributa (Larrañaga *et al.*, 2006; Domingos, 2012; Chicco, 2017). Ručno generisanje atributa zahteva najveći deo procesa formiranja kvalitetnog predikcionog modela i u najvećoj meri zavisi od nivoa domenskog znanja i stepena ekspertize istraživača. Selekcija atributa je poseban korak kojim se smanjuje dimenzionalnost podataka, pronalaženjem optimalnog podskupa ulaznih atributa uz minimalno smanjenje ili moguće i povećanje tačnosti modela (Saeys, Inza and Larranaga, 2007).

1.4.4.5 Ansambliranje modela mašinskog učenja

Generisanje efikasnog modela mašinskog učenja za rešavanje problema klasifikacije predstavlja izazov. U idealnom slučaju, model sa maksimalnom sposobnošću generalizacije na novim podacima bi se formirao algoritmom mašinskog učenja uz minimalan utrošak vremena i računarskih resursa. Nijedan algoritam mašinskog učenja nije najbolje rešenje za svaki problem nadgledanog učenja (Wolpert, 2002; Smith-Miles, 2008). Statističke analize su pokazale da kombinovanje većeg broja podoptimalnih modela može omogućiti veći stepen tačnosti predviđanja nego bilo koji od datih modela pojedinačno (Rokach, 2010). Algoritmi mašinskog učenja koji omogućavaju automatsko spajanje većeg broja modela (ansambl algoritmi), se u poslednje vreme standardno upotrebljavaju u bioinformatici i drugim oblastima (Larrañaga *et al.*, 2006; Domingos, 2012).

1.4.4.6 Heterogene infomacije

Pored ekstrakcije više skupova atributa iz jednog tipa informacija u cilju predviđanja IPP, poput korišćenja primarne strukture proteina, često se koristi integrisani pristup formiranju atributa za predviđanje IPP. Budući da svaka vrsta informacija ima svoje nedostatke, čest pristup je integracija više tipova informacija pri generisanju grupa atributa (Jansen, 2003; Ben-Hur and Noble, 2005; Qi, Bar-Joseph and Klein-Seetharaman, 2006; Elefsinioti *et al.*, 2011). Ben Hur i Noble (Ben-Hur and Noble, 2005) koriste GO anotacije i karakteristike bioloških mreža kao izvor informacija za poboljšanje efikasnosti. Više vrsta atributa formiranih na osnovu sekvence uključuju k-mer, motive i proteinske domene detektovanih na osnovu analize sekvenci. Vektori atributa kvaščevih IPP se koriste za učenje SVM (engl. Support Vector Machine) modela, pomoću posebno dizajniranog *kernela* za parove proteina (engl. kernel).

U sklopu hPRINT pristupa (Elefsinioti *et al.*, 2011) autori integrišu heterogene biološke podatke u cilju predikcije IPP. Za potrebe učenja klasifikatora autori su izdvojili 18 grupa atributa koji uključuju mrežne atribute na osnovu potvrđenih i predviđenih interakcija STRING baze podataka, filogenetske profile, koekspresiju, fuziju gena itd. Atributi koji opisuju povezanost proteina unutar GO mreža, KEGG puteva i topološke karakteristike

bioloških mreža su korišćene za formiranje SVM modela u sklopu ppiPre metoda (Deng, Gao and Wang, 2013). Zubek i sar. (Zubek *et al.*, 2015) su razvili MLPPi pristup koji koristi heterogene informacije iz tercijarne i primarne strukture proteina posmatrajući relacije između svih segmenata sekvence interagujuća proteina. Model mašinskog učenja je formiran na IPP kvasca pomoć algoritma nasumičnih šuma (engl. Random Forest, RF) LocFuse (Zahiri *et al.*, 2014) koristi kombinacije 8 vrsta atributa uključujući post-translacione modifikacije, evolutivne profile i GO sličnost, formiranih na osnovu analiza iz raznih izvora, za predstavljanje proteina 582 dimenzionim vektorom.

1.4.4.7 Negativni primeri interakcija protein-protein

Kod najvećeg broja metoda za predviđanje IPP, zasnovanih na modeliranju IPP, metodama mašinskog učenja na nivou sekvenci, problem predviđanje IPP se posmatra kao problem binarne klasifikacije (Liu, 2009; Elefsinioti *et al.*, 2011; Hashemifar *et al.*, 2018). Učenje binarnog klasifikatora zahteva, pored pouzdanih informacija o eksperimentalno verifikovanim IPP koje se označavaju pozitivnim skupom IPP, skup parova proteina koji ne interaguju (NIPP). Budući da ne postoje eksperimentalne metoda za detekciju NIPP na nivou proteoma, uz izraženu tendenciju prijavljivanja samo pozitivnih rezultata eksperimenata za detekciju IPP u biomedicinskoj literaturi, broj prijavljenih eksperimentalno potvrđenih NIPP u odnosu na IPP je veoma mali (Smialowski *et al.*, 2010). Automatskim prikupljanjem podataka, zasnovanom na pristupu *data mining* i espertskom proverom u Negatome 2.0 bazi podataka, arhivirano je 6532 NIPP koje pored ljudskih uključuju i proteine nekoliko drugih bioloških vrsta (Blohm *et al.*, 2014). Jedan od parametara koji utiču na efikasnost formiranog modela za mašinsko učenje je količina podataka na kojima se uči (Domingos, 2012). Sa druge strane disbalans u broju pozitivnih i negativnih primera negativno utiče na sposobnost generalizacije prediktora IPP na novim podacima (Park and Marcotte, 2011). Usled toga, razvijeni su pristupi za formiranje skupova NIPP sa malom verovatnoćom interakcije. Među najčešće korišćenim strategijama izdvajaju se: (i) nasumično povezivanje proteina u NIPP za koje interakcija nije detektovana u bazama podataka IPP, (ii) korišćenje pretpostavke da se NIPP nalaze u različitim ćelijskim kompartmentima i (iii) formiranje

skupa NIPP od sintetički formiranih proteinskih sekvenci generisanih nasumičnim kombinovanjem aminokiselina (Liu, 2009).

1.4.4.8 Tipovi test parova IPP

Matematičko modeliranje binarnih IPP podrazumeva da se svaka proteinska interakcija AB sastoji od dve komponente, proteina A i proteina B, bez obzira na način spajanja originalnih vektora u konačni vektor koji predstavlja interakciju. U toku formiranja modela ML zasnovanog na podacima o IPP, informacija o komponentama (proteinima) je očuvana u sklopu predikcionog modela. U fazi testiranja efikasnosti formiranog modela u predviđanju novih interakcija, test skup može sadržati sledeće kombinacije (tip proteinskog para):

1. da su oba proteina (komponente) prisutna u skupu na kome se treniralo (C1 tip),
2. samo jedan protein iz binarne IPP je prisutan u trening skupu (C2 tip),
3. niti jedan od proteina koji formiraju IPP instancu test skupa se ne nalazi u bilo kojoj kombinaciji IPP u trening skupu (C3 tip).

Pokazalo se da sposobnost modela da tačno predviđa IPP značajno zavisi od stepena prisutnosti proteina iz trening skupa u test skupu. Najveća tačnost je primećena u slučaju C2, a najmanja u C3 (Park and Marcotte, 2012).

1.4.4.9 Metode mašinskog učenja za predviđanje IPP

Metod potpornih vektora (engl. Support Vector Machine, SVM) se pokazao kao jedan od najefikasnijh u formiranju modela za predviđanje IPP visokih performansi (Gomez, Noble and Rzhetsky, 2003; Ben-Hur and Noble, 2005; Martin, Roe and Faulon, 2005; Shen *et al.*, 2007; Guo *et al.*, 2008; Dong, Zhou and Liu, 2010; Shoyaib and Abdullah-Al-Wadud, 2010; Wang *et al.*, 2010; Shi *et al.*, 2010; Ren *et al.*, 2011; Zhao, Ma and Yin, 2012; Hamp and Rost, 2015b; Huang *et al.*, 2016). Sposobnost klasifikacije IPP SVM modela, najvećim delom je omogućen pažljivo dizajniranim *kernelima*. U osnovi funkcije preslikavanja, kerneli omogućuju inkorporaciju različitih tipova informacija izabranih na osnovu domenskog znanja o IPP.

Jednostavnost aplikacije, visoka prediktivna tačnost formiranih modela sa standardnim podešavanjem parametara algoritma, doveli su do raširene upotrebe RF algoritma u problemima predikcije IPP (Pan, Zhang and Shen, 2010; Xia, Han and Huang, 2010; Valente *et al.*, 2013; Du *et al.*, 2014; Zahiri *et al.*, 2014; You, Chan and Hu, 2015).
Potreba za poboljšanjem performansi, uz smanjenje vremena modeliranja i predviđanja na proteomskom nivou, zahtevala je upotrebu efikasnijih algoritama mašinskog učenje poput K-najbližeg suseda (engl. k-Nearest Neighbors, kNB) (Guarracino *et al.*, 2010).
Povećanje količine dostupnih podataka o IPP zajedno sa povećanjem računarskih potencijala omogućilo je širu upotrebu DL u predikciji IPP (Du *et al.*, 2017; Hashemifar *et al.*, 2018).

1.4.6 Metode za predviđanje IPP zasnovane na topologiji bioloških mreža

Za potpuno razumevanje ćelijskih procesa neophodno je analizirati ćeliju kao kompleksan biološki sistem (Albert, 2007). Kompleksne sisteme je moguće predstaviti kao mrežu odnosa između elemenata i analizirati standardnim metodama (Voit, 2013). Oslanjajući se na teoriju složenih sistema, sistemska biologija konceptualizuje ćeliju kao kompleksnu mrežu relacija među makromolekulima (Vidal, Cusick and Barabási, 2011). Ovakva reprezentacija interakcija unutar ćelije uobičajeno se formalizuje pomoću teorije grafova. Teorija grafova je deo diskretne matematike koja se bavi odnosom između apstraktnih objekata (Diestel, 2018). Mrežom se složeni sistem simplifikuje u formi grafa pri čemu su elementi mreže predstavljeni čvorovima a njihove međusobne interakcije granama (Pavlopoulos *et al.*, 2011).

Ovakvo predstavljanje omogućava analizu dinamike i strukture bioloških sistema u celini, za koje je utvrđeno da se strukturno značajno razlikuju od nasumično generisanih mreža (Vidal, Cusick and Barabási, 2011). Mrežno predstavljanje biološkog sistema je moguće na bilo kom nivou, od ćelije do biosfere (Kim et al., 2019), dok se ćelijski interaktom često analiza mrežama IPP, metaboličkim, mrežama genske regulacije i provođenja signala (Gosak *et al.*, 2018). Nezavisno koji deo ćelijskog interaktoma se reprezentuje, izdvajaju se tri načina formiranja mreža u odnosu na izvor informacija: (i) korišćenje informacija o potvrđenim interakcijama na osnovu literature, (ii) korišćenje računarskih metoda u svrhu predviđanja interakcija i (iii) korišćenje sistemskih

eksperimenata velike skale za utvrđivanje interakcija na nivou genoma ili proteoma (Vidal, Cusick and Barabási, 2011).

Grafovi predstavljaju nelinearne i negeometrijske strukture definisane teorijom grafova i najčešće grafički predstavljene u formi mreže čvorova povezanih granama. Graf G se definiše preko skupa čvorova V koji su povezani granama E, $G = (V,E)$. Matematičkim analizama obrazaca povezivanja čvorova moguće je utvrditi zakonitosti koje vladaju u globalnoj i lokalnoj strukturi mreže i na osnovu toga vršiti predviđanja promene konektivnosti mreže, uključujući i nove veze između čvorova (Huang, 2010). Jedna grana između čvorova predstavlja jednu vrstu informacije na osnovu koje se uspostavlja asocijacija između elemenata mreže. Mreže koje predstavljaju jednu vrstu IPP se predstavljaju jednostavnim grafom, dok je moguće predstaviti i različite vrste konekcija, poput koekspresije i evolutivne povezanosti, između susednih čvorova. Za ovakvo predstavljanje IPP u bazi STRING (Szklarczyk *et al.*, 2017) koristi se multigraf (Pavlopoulos *et al.*, 2011). Komutativna priroda fizičkih binarnih interakcija proteina se predstavlja neusmerenim granama između susednih čvorova.

Usled nekompletnosti humanog proteinskog interaktoma, grafovi formirani na osnovu postojećih podatka o interakcijama su podgrafovi kompletne mreže interakcija (Aittokallio and Schwikowski, 2006). Tumačenje karakteristika nadgrafa zavisi od načina uzorkovanja podgrafa, a interpretacija karakteristika kompletnog grafa na osnovu karakteristika podgrafa može biti pogrešna (Stumpf, Wiuf and May, 2005). Analiza do danas rasvetljenih bioloških mreža na različitim nivoima, od biohemijskih mreža do biosfere, ukazuje na zajedničke karakteristike mrežne organizacije celokupnog života (Kim *et al.*, 2019).

Karakteristike interne strukture i organizacije mreža se otkrivaju mrežnom analizom. Povećanjem broja sekvenciranih proteina i informacija o IPP povećava se ne samo broj čvorova u mreži IPP već i ukupan broj grana, koji se označava ukupnom konektivnošću grafa (Pavlopoulos *et al.*, 2011). Broj konekcija koje čvor poseduje sa susednim čvorovima se označava stepenom čvora (Diestel, 2018). Raspodela stepena čvorova je jedna od najvažnijih karakteristika pomoću koje se opisuju razlike između bioloških i nasumično formiranih mreža (Vidal, Cusick and Barabási, 2011). Različito od Erdos-Renyi modela nasumičnog grafa (Erdős and Rényi, 1959; Newman, 2002) čiji čvorovi imaju približno jednak stepen, mnoge stvarne mreže, od socijalnih do bioloških,

poseduju stepenu raspodelu (engl. power law) stepena čvorova (Barabási and Albert, 1999). Ovakva raspodela podrazumeva postojanje malog broja čvorova sa relativno velikim brojem konekcija i znatno većeg broja proteina sa relativno malim brojem interakcija. Stepena raspodela je pronađena u IPP i metaboličkim mrežama svih ispitivanih organizama, koje se usled toga označavaju i kao *mreže bez skale* (engl. scale free) (Barabási and Oltvai, 2004). Ovakvo svojstvo IPP mreža se objašnjava 'rich-get-richer' fenomenom (Barabási and Albert, 1999) koji podrazumeva veću verovatnoću formiranja novih interakcija kod proteina sa relativno velikim brojem interakcija u odnosu na prosečan stepen čvora u mreži. Smatra se da je ovaj mehanizam u osnovi formiranja proteinskih *habova*. Biološka osnova ovog mehanizma se verovatno ogleda u procesu duplikacije gena i hromozoma. U toku evolucije produkti dupliranih gena i njihovi potomci zadržavaju partnere za interakciju (Pastor-Satorras, Smith and Solé, 2003; Vázquez *et al.*, 2003). Sa druge strane, uloga prostorne organizacije unutar ćelije u ograničavanju IPP ukazuje na zaključak o visokom stepenu sličnosti IPP mreža sa geometrijskim mrežama, poput mreža elektordistribucije ili železničkog transporta (Pržulj, Corneil and Jurisica, 2004; McGillivray *et al.*, 2018). Raspodela stepeni usmerenih grana od transkripcionih faktora ka genima kod regulatorne mreže analizirane kod kvasca je približna stepenoj raspodeli, dok je raspodela stepeni usmerenih grana od gena ka transkripcionim faktorima, eksponencijalna (Deplancke *et al.*, 2006). Ovakve kakakteristika regulatornih mreža se objašnjava činjenicom da mali broj transkripcionih faktora reguliše veliki broj gena, dok su sa druge strane geni istovremeno retko regulisani velikim broje transkripcionih faktora (Vidal, Cusick and Barabási, 2011).

Razvijen je veliki broj metoda koje implementiraju različite pristupe za predviđanju nedostajućih ivica u sklopu ispitivane mreže. Neki od ovih algoritama su zasnovani na tehnikama lokalne ili globalne analize mrežnih karakteristika (Lichtenwalter, Lussier and Chawla, 2010; Ahmed, Elkorany and Bahgat, 2016; Yang and Zhang, 2016), rekonstrukcije mreža probabilističkim metodama (Guimerà and Sales-Pardo, 2009), korelacije između čvorova (Liao, Zeng and Zhang, 2015), korelacije između vremenski ili na drugi način povezanih mreža (Wu, Zhang and Wu, 2017). Učinjeni su različiti napori ka integraciji tehnika mašinskog učenja na mrežnim strukturama (Lee and Seung, 2001; Al Hasan *et al.*, 2006; Benchettara, Kanawati and Rouveirol, 2010; Ahmed, Elkorany and Bahgat, 2016). Efikasna primena mašinskog

učenja na mrežne strukture ostaje jedan od važnih izazova i pravaca razvoja (Latouche and Rossi, 2015).

Često se teorijske postavke grafa koriste u rešavanju problema predviđanje IPP. Jedan od pristupa uključuju formiranje IPP mreža oko proteina partnera u sklopu ispitivane interakcije. Ovakve submreže se formiraju na osnovu postojećih podataka o IPP koju ispitivani proteini formiraju. Pretraživanjem susedstva u sklopu formiranih submreža i njihovim poređenjem se formira skor na osnovu koga se vrši predviđanje interakcije (Li, Liu and Burge, 2012). Murakami i Mizuguchi (Murakami and Mizuguchi, 2014) mrežnim algoritmima pretražuju udaljenosti između homolognih proteina u sklopu IPP mreže. Homologija među parovima interagujućih proteina može dovesti do detekcije novih veza unutar IPP mreže. L3 pristup (Luck *et al.*, 2018) predviđanja nedostajućih ivica je zasnovan na teoretskoj osnovi da je potrebno posmatranje parova proteina i njihovog susedstva na većoj udaljenosti. L3 metoda posmatra i analizira puteve u sklopu mreže između ispitivanih proteina na udaljenosti 3, raspravljajući da ovakav pristup je više zasnovan na biološkoj realnosti koju IPP mreže treba da predstavljaju (Luck *et al.*, 2018). Jedan od najvećih izazova različitih mrežnih metoda predstavlja dostupnosti informacija. Nedostatak informacije o potvrđenim IPP ciljnog proteina značajno smanjuje verovatnoću i pouzdanost predviđanja drugih interakcija tog proteina (Luck *et al.*, 2018).

2 Ciljevi istraživanja:

1) Razvoj modela IPP koji učestvuju u procesu transkripcione regulacije i metoda za automatsko predviđanje ovih IPP. Sastavni deo ovog cilja je i povećanje dostupnosti novog pristupa za predviđanje adaptiranjem u veb alat.

2) Razvoj modela IPP čoveka koji obuhvata proteine sa neuređenom tercijarnom strukturom. U sklopu ovog cilja je implementacija razvijenog metoda u formi veb alata.

3) Razvoj modela IPP čoveka i bioinformatičkog metoda za predviđanje IPP. U okviru ovog cilja planirano je implementiranje razvijenog predikcionog modela i metode u formi samostalne aplikacije.

3 Materijali i metode

3.1 Podaci

3.1.1 Baze podataka proteina i proteinskih sekvenci

U svrhu ove studije, informacije o ljudskim proteinima i proteinskim sekvencama preuzete su iz UniProtKB/Swiss-Prot (Consortium, 2015) baze podataka. Za analizu i modeliranje ljudskih transkripcionih regulatora i njihovih IPP, korišćena je verzija 2015_09 UniProtKB/Swiss-Prot baze podataka. Sekvence proteina analizirane u svrhu formiranja predikcionog modela PNTS proteina su preuzete iz verzije 2017_11 UniProtKB/Swiss-Prot baze podataka. U svrhu generisanja opšteg prediktivnog modela ljudskih IPP korišćena je 2018_02 verzija UniProtKB/Swiss-Prot.

DisProt 7 baza podataka, verzija 0.5 iz 05.11.2017, sadrži 237 ljudskih PNTS i proteina sa RNTS (Piovesan *et al.*, 2017). Ova verzija DisProt baze je služila kao resurs za proteinske sekvence PNTS i proteine sa RNTS.

3.1.2 Baze podataka IPP

HIPPIE (engl. Human Integrated Protein-Protein Interaction rEference) baza podataka ljudskih IPP (Schaefer *et al.*, 2012) je korišćena kao osnovni izvor eksperimentalno potvrđenih i funkcionalno anotiranih IPP. HIPPIE baza podataka verzija 2.0 iz juna 2016. (Alanis-Lobato, Andrade-Navarro and Schaefer, 2017) sadrži 287357 eksperimentalno potvrđenih IPP. HIPPIE 2.0 je korišćena kao izvor IPP za formiranja predikcionog modela transkripcionih regulatora i modela za predviđanje PNTS IPP. Za formiranje opšteg predikcionog modela ljudskih IPP korišćena je HIPPIE verzija 2.1 koja sadrži 299247 IPP dok je za evaluaciju predikcionih performansi opšteg modela ljudskih IPP korišćena novija verzija baze podataka HIPPIE 2.2 sa opisanih 347139 eksperimentalno potvrđenih IPP.

MENTHA baza podataka (Calderone, Castagnoli and Cesareni, 2013) trenutno sadrži 741337 IPP 8 bioloških vrsta. U svrhu analize efikasnosti modela ljudskih IPP za

predviđanje IPP pacova (*Rattus norvegicus*) korišćena verzija MENTHA baze podataka od 04.07.2016. godine.

Negatome baza podataka je služila kao izvor eksperimentalno potvrđenih NIPP anotiranih iz literature (Blohm *et al.*, 2014). Negatome verzija 2.0 sadrži 6532 intraspecijskih i interspecijskih NIPP nekoliko bioloških vrsta.

IntAct baza podataka (Hermjakob *et al.*, 2004) je korišćena kako bi se pronašli sve IPP koje formiraju proteini koji učestvuju u NIPP prethodno pronađeni u Negatome 2.0. Kao deo IMEX conzorcijuma IntAct pruža iscrpan resurs za preuzimanje IPP za 20 bioloških vrsta.

3.1.3 Baza podataka fizičko-hemijskih karakteristika aminokiselina

AAindex je specijalizovana baza podataka fizičko-hemijskih i biohemijskih karakteristika aminokiselina (Kawashima, Ogata and Kanehisa, 1999). Modeliranje proteinskih sekvenci u vektorskoj formi izvršeno je korišćenjem 532 mere aminokiselina preuzete iz verzije 9.0 AAindeks baze podataka. U cilju numeričkog predstavljanja proteinskih sekvenci u sklopu procesu formiranja generalnog predikcionog modela ljudskih IPP korišćena je verzija 9.1 AAindex baze podataka.

3.1.4 Anotacijski resursi ontologija gena (GO i AMIGO)

Genska ontologija (GO, www.geneontology.org) je projekt razvoja preciziranih rečnika (ontologija) GO termina kako bi standardizovano opisale karakteristike gena i genskih produkata bez obzira na biološku vrstu. Projekat je realizirao Konzorcijum genskih ontologija (engl. Gene Ontology Consortium, GOC, 2000). U sklopu GO, GO termini su predstavljeni čvorovima povezanih u formu direktnog acikličnog grafa i svrstavaju se u jednu od tri klase ontologija: molekularne funkcije, biološke procese i ćelijske lokacije. GO imaju hijerarhijsku strukturu, gde u grafu termini preci obuhvataju šire značenje u odnosu na termine potomke, u sklopu date GO. GO terminima su pridružene GO anotacije u formi tvrdnji potkrepljenih dokazima koje opisuju odnose između GO termina i genskih produkata i reference koji opisuju dati odnos.

AmiGO je veb alat razvijen i održavan od strane GOC u cilju pregledanja, pretraživanja i vizuelizacije GO i GO anotacija. AmiGO omogućava ciljna pretraživanja gena i genskih produkata prema različitim identifikacionim sinonimima iz različitih baza podataka gena i genskih produkata. U toku naše analize korišćena je AmiGO2 verzija veb alata (Carbon *et al.*, 2009; Blake *et al.*, 2015). Za identifikaciju proteina povezanih sa GO terminima korišćeni su UniprotAC identifikatori.

Za vizuelizaciju mreže ili dela mreže GO termina korišćenih u ovom radu, korišćen je QuickGO alat (Binns *et al.*, 2009).

BINGO je alat razvijen u svrhu pronalaženja GO kategorija koje su karakteristične za skup gena ili podgraf biološke mreže, na osnovu statističkih analiza (Maere, Heymans and Kuiper, 2005).

REVIGO je veb servis koji korisniku omogućava različite načine vizuelizacije skupa GO termina u cilju olakšavanje interpretacije (Supek *et al.*, 2011).

3.2 Modeliranje proteina i IPP

3.2.1 Grupisanje proteina (CD-HIT alat)

CD-HIT je program za grupisanje proteina sa visokim računarskim performansama u vidu brzine i male potrošnje računarskih resursa. CD-HIT omogućava grupisanje proteina prema sličnosti njihovih sekvenci bez iscrpnog poređenja svake sa svakom sekvencom. U sklopu CD-HIT algoritma sekvence se prvo sortiraju prema dužini, tako da se najduže postavljaju kao predstavnici grupa. Sve ostale sekvence se porede sa predstavnikom postojeće grupe. Ukoliko stepen sličnosti sekvence sa predstavnikom grupe prelazi unapred definisani prag, protein se svrstava u datu grupu dok u suprotnom sam postaje predstavnik druge grupe (Fu *et al.*, 2012). U toku analize proteina prema sličnosti sekvence korišćene su standardne postavke CD-HIT programa. Program je korišćen u paralelnom modu na računaru sa Xeon procesorom (E5-2630 V3 2.4 GHz) i 32 GB RAM. Pomoćni alati za filtriranje IPP od proteina sa stepenom sličnosti sekvence većom od 40% su razvijeni u programskom jeziku AWK (Aho, Kernighan and Weinberger, 1979).

3.2.2 Pseudo kompozicija aminokiselina

Pseudo kompozicija aminokiselina (engl. pseudo amino acid composition, PAAC) je koncept modeliranja proteinske sekvence koji uključuje kompoziciju aminokiselina (engl. Amino acid composition, AAC) i informacije o redosledu i udaljenostima uzduž sekvence.

Kompozicija aminokiselina je forma numeričkog predstavljanja primarne strukture proteina (Chou, 2009). AAC predstavlja numerički vektor koji sadrži normalizovane frekvencije pojavljivanja 20 aminokiseline u formi diskretnih brojeva. AAC numerički vektor frekvencija se računa pomoći jednačine:

$$AAC_i = \frac{n_i}{L},\ i = 1..20 \qquad (1)$$

gdje je n_i broj pojavljivanja i-te aminokiseline u sekvenci dužine L.

Roy i sar. (Roy *et al.*, 2009) su pokazali da kompozicija aminokiseline u formi numeričkog vektora dužine 20 može biti upotrebljena za efikasnu reprezentaciju sekvenci proteina u svrhu predviđanja IPP metodama ML. Autori su ukazali na osnovne prednosti AAC u predikciji IPP: (i) jednostavno i veoma brzo računanje čak i u slučaju veoma dugačkih sekvenci proteina, (ii) nezavisnost od informacije o domenima budući da se AAC računa isključivo iz primarne strukture proteina bez potrebe za dodatnim informacijama i (iii) veoma dobre performanse u zadacima klasifikacije u čiji okvir ulazi predviđanje IPP. Nedostaci AAC se očituju u odsustvu informacije o redosledu pojedinačnih aminokiselina unutar proteinske sekvence. Druga komponenta PAAC modela je numerički vektor dobijen računanjem korelacije između tri fizičko-hemijske karakteristike aminokiselina: hidrofobnost, hidrofilnost i masa bočnih lanaca uzduž sekvence proteina. Dužina druge komponente PAAC, lambda (λ), zavisi od udaljenosti na kojoj se posmatra korelacija i broja numerički predstavljenih osobina aminokiseline između kojih se korelacija računa. Pored λ, značajan faktor za optimizaciju je i faktor otežanja ω za dodatno podešavanje druge komponente PAAC vektora (Chou, 2001). Za računanje PAAC vektora korišćena je implementacija u R paketu *protr* (Xiao, Xu and Cao, 2014). U toku razvijanja algoritma za formiranje i evaluaciju modela za predviđanje IPP transkripcionih regulatora, modifikovana PAAC reprezentacija je implementirana u

R statističkom jeziku. U svrhu formiranja web alata zasnovanog na razvijenom algoritmu za predviđanje IPP transkripcionih regulatora modifikovani PAAC je implementiran u JAVA programskom jeziku. U toku razvoja algoritma, evaluacije algoritma i izrade veb alata za predviđanje IPP PNTS proteina, modifikovani PAAC pristup je implementiran u JAVA programskom jeziku.

Tabela 2. Vrednost fizičko-hemijskih karakteristika u sklopu PAAC4 modela proteinskih sekvenci za 20 aminokiselina: hidrofobnost, hidrofilnost, masa bočnog lanca i potencijal elekto-jon interakcije. Ovaj model je korišćen pri modeliranju IPP transkripcionih regulatora.

AA	Hidrofobnost	Hidrofilnost	Masa bočnog lanca	EIIP
A	0.62	-0.5	15	0.0373
R	-2.53	3	101	0.0959
N	-0.78	0.2	58	0.0036
D	-0.9	3	59	0.1263
C	0.29	-1	47	0.0829
E	-0.74	3	73	0.0057
Q	-0.85	0.2	72	0.0761
G	0.48	0	1	0.0050
H	-0.4	-0.5	82	0.0242
I	1.38	-1.8	57	0.0000
L	1.06	-1.8	57	0.0000
K	-1.5	3	73	0.0371
M	0.64	-1.3	75	0.0823
F	1.19	-2.5	91	0.0946
P	0.12	0	42	0.0198
S	-0.18	0.3	31	0.0829
T	-0.05	-0.4	45	0.0941
W	0.81	-3.4	130	0.0548
Y	0.26	-2.3	107	0.0516
V	1.08	-1.5	43	0.0058

Tabela 3. Lestvica karakteristika aminokiselina korišćenih u modeliranju proteinskih sekvenci u formi PAAC5 modela za predviđanje IPP PNTS.

AA lestvica	Top-IDP	Net charge	B-vrednost	FoldUnfold	DisProt
W	-0.884	0	0.938	28.48	-0.465
F	-0.697	0	0.934	27.18	-0.381
Y	-0.51	0	0.981	25.93	-0.427
I	-0.486	0	0.977	25.71	-393
M	-0.397	0	0.963	24.82	0.197
L	-0.326	0	0.982	25.36	-0.26
V	-0.121	0	0.968	23.93	-0.302
N	0.007	0	1.022	18.49	-0.106
C	0.02	0	0.939	23.52	-0.546
T	0.059	0	0.998	19.81	-0.116
A	0.06	0	0.994	19.89	0.042
G	0.166	0	1.018	17.11	0.095
R	0.18	1	1.026	21.03	0.211
D	0.192	-1	1.022	17.41	0.127
H	0.303	0	0.967	21.72	-0.127
Q	0.318	0	1.041	19.23	0.381
K	0.586	1	1.029	18.19	0.37
S	0.341	0	1.025	17.67	0.201
E	0.736	-1	1.052	17.46	0.469
P	0.987	0	1.05	17.43	0.419

3.2.3 Auto kros-kovarijansa

Auto kros-kovarijansa (eng. auto cross-covariance function, ACC) (Sjöström, Rännar and Wieslander, 1995) je funkcija kojom se može predstaviti sličnosti između ciljnih osobina aminokiselina i definisana je jednačinom:

$$ACC_{j,k,l} = \frac{1}{L-l}\sum_{i=1}^{L-l} z_{j,i} z_{k,i+l} \quad j,k = 1..n,\ l = 1..m \qquad (2)$$

gde je m dužina pomerajućeg prozora, odnosno dužina podsekvenci, n je broj osobina koje se posmatraju, $Z_{i,j}$ j-ta vrednosti i-te komponente. Auto kros-kovarijansa omogućava

generisanja vektora jednake dužine kojim se mogu predstaviti sekvence proteina. U sklopu ove funkcije posmtraju se međusobne relacije između svih posmatranih osobina aminokiselina u sklopu pomerajućeg prozora sekvence proteina.

3.2.4 Filogenetski profili (PSI-BLAST alat)

PSI-BLAST (engl. Position-Specific Iterative Basic Local Alignment Search Tool) je alat za računanje PSSM matrica (engl. Position-scpecific scoring matrix) koristeći BLAST alat (Altschul, 1997). BLAST ja algoritam za poravnavanje i upoređivanje proteinskih ili nukleotidnih sekvenci u cilju detekcije sličnosti među sekvencama. Pojedinačne sekvence se porede sa sekvencama u ciljnoj bazi podataka formirajući lokalna poravnanja. Lokalna poravnanja sa skorom sličnosti većim od zadatog praga u odnosu na matricu vrednosti (BLOSUM62 matrica je najčešće korišćena) se prijavljuju. Za matrice vrednosti se koriste PAM i BLOSUM matrice pri čemu je standardno korišćena BLOSUM62 matrica. PSI-BLAST iterativno koristi BLAST algoritam za pronalaženje svih sekvenci sa skorom većim od postavljenog praga. Na osnovu višestrukog poravnanja homolognih sekvenci pronađenih prethodnim BLAST pretragama, izvodi se profil poravnanja ili PSSM. Profil poravnanja predstavljen PSSM matricom predstavlja matricu substitucije aminokiselina gde se visoke vrednosti dodeljuju aminokiselinama u visoko konzerviranim regionima, dok se pozicijama niskog stepena očuvanja pridodaju vrednosti blizu nuli. Novoformirana PSSM matrica se koristi kao matrica vrednosti u narednim pretragama, tako da je čitav proces iterativan i zaustavlja se ili kada se postigne konvergencija algoritama ili nema više novih sekvenci čija sličnost zadovoljava prethodno postavljene kriterijume.

Računanje PSSM matrica je izvršeno korišćenjem alata PSI_BLAST iz programskog paketa BLAST+ verzija 2.7.1 (Altschul, 1997; Altschul and Koonin, 1998) preuzetog sa veb sajta Američkog nacionalnog centra za biotehnološke informacije (engl. National Center for Biotechnology Information, NCBI) (Benson *et al.*, 1990), korišćenjem sledećih parametara: *num_iterations* = 3, *inclusion_ethresh* = 0.002, *evalue* = 10 i *matrix* = BLOSUM62, dok su ostali parametri podešeni kako je savetovano u studijama (Jones and Swindells, 2002; Bhagwat and Aravind, 2007)

3.2.5 Topološke karakteristike grafa

Za formiranje grafova IPP i mrežnu analizu korišćen je iGraph softverski paket implementiran u formi R paketa „igraph“ (Csardi and Nepusz, 2006).

Čvorovi neusmerenog neotežanog grafa IPP, koji predstavljaju proteine, su predstavljeni vrednostima 21 atributa čije su vrednosti dobijene računanjem topoloških karakteristika grafa. Korišćeni mrežni atributi se mogu svrstati u tri grupe: (i) lokalne, (ii) globalne mere konektivnosti čvorova i (iii) pripadnost mrežnim modulima ili motivima (Paladugu *et al.*, 2008). Lokalne karakteristike čvorova u okviru mreže su predstavljene stepenom čvora. Globalne mere uključuju mere centralnosti čvora (engl. centrality), bliskosti (engl. closeness) i meru relacije (engl. betweenes). Za računanje vrednosti globalnih mera čvora, odnosno Alfa centralnost (engl. Alpha centrality) (Bonacich and Lloyd, 2001), Klainbergova autoritativna centralnost (engl. Kleinberg's authority centrality) (Jon M. Kleinberg, 1999) i Sopstveni vektor čvora (engl. Eigenvector centrality) (Bonacich, 2002), u obzir se pored stepena tog čvora uzimaju i stepeni njime susednih čvorova. Veličina susedstva čvora je predstavljena vrednostima mere Broj suseda na udaljenosti n (engl. Ego size) i komplementarne mere Bartovo ograničenje (engl. Burt's constraint) (Burt, 2004). Koristeći drugačiji metodološki pristup, mera Statistika lokalnih suseda (engl. Local scan statistics) (Priebe, 2006) i mera razvijena na osnovu istoimenog algoritma za pretraživanje Interneta, PageRank centralnost (engl. PageRank centrality) (Brin and Page, 2012) predstavljaju statistiku povezanosti okruženja ciljnog čvora. Globalna pozicija čvora u mreži može biti opisana funkcijom broja najkraćih puteva (engl. shortest path) koji povezuju ciljni čvor sa svim ostalim čvorovima mreže ili prolaze kroz njega. Ovakav pristup je korišćen kod računanja vrednosti ekscentričnosti (engl. Eccentricity) čvora (Barnes and Harary, 1983), Relaciona centralnost (engl. Betweenness centrality) i Centralnost po bliskosti (engl. Closeness centrality) vrednosti (Freeman, 1978). Jedan od pristupa uključuje upotrebu algoritma K najbližih suseda u cilju predstavljanja globalne uloge čvora u širem okruženju (Barrat *et al.*, 2004).

Mere pripadnosti distinktnim modulima ili motivima se zasnovaju na različitim pristupima detekcije i karakterizacije zajednica grupisanih čvorova. Pripadnost specifičnog čvora definisanoj grupi se može posmatrati kao atribut. Vrednost Broja trouglova kojima čvor pripada (engl. Count triangles) je kvantitivna mera učestvovanja

čvora u formiranju maksimalno povezanih trijada čvorova. Lokalne zajednice gusto povezanih podgrafova se mogu detektovati algoritmom koji primenjuje metodu nasumične šetnje (engl. random walks) (Pons and Latapy, 2005) i omogućava računanje vrednosti mera Indeks pripadajuće zajednice čvora po slučajnom obilasku *(*engl. Cluster walktrap*)* i Indeks pripadajuće zajednice čvora po optimizaciji modularnosti (engl. Cluster edge betweenes) (Newman and Girvan, 2004). Na Tabeli 4. su prikazani nazivi mrežnih mera korišćenih za predstavljanje proteinskih sekvenci kako se navode u „igraph“ R paketu, zajedno sa punim nazivima

Tabela 4. Lista mrežnih atributa sa kratkim opisima dobijeni analizom mreža IPP.

Nazivi atributa alata	Puni nazivi i opisi atributa
Alpha_centrality	Alfa centralnost (engl. Alpha centrality)
Authority_score	Klainbergova autoritativna centralnost (engl. Kleinberg's authority centrality)
Betweenness	Relaciona centralnost (engl. Betweenness centrality)
Centr_clo	Centralnost po bliskosti (engl. Closeness centrality)
Closeness	Blizina čvorova
Cluster_fast_greedy	Indeks pripadajuće zajednice čvora po optimizaciji modularnosti
Cluster_walktrap	Indeks pripadajuće zajednice čvora po slučajnom obilasku
Components	Indeks pripadajuće komponente čvora
Constraint	Bartovo ograničenje (engl. Burt's constraint)
Coreness	*K-core* dekompozicija grafa
Count_triangles	Broj trouglova kojima čvor pripada
Degree	Stepen čvora (engl. Degree centrality)
Eccentricity	Ekscentričnost čvora
Ego_2	Broj suseda na udaljenosti reda 2
Ego_3	Broj suseda na udaljenosti reda 3
Eigen_centrality	Sopstveni vektor čvora (engl. Eigenvector centrality)
Knn	Prosečan stepen najbližih suseda
Local_scan	Statistika lokalnih suseda (engl. Local scan statistics)
Max_cardinality	Maksimalna kardinalnost (engl. Maximal cardinality)
Page_rank	PageRank centralnost (engl. PageRank centrality)
Strength	Otežani stepen čvora (engl. Weighted degree)

3.3 Mašinsko učenje

3.3.1 Statističke mere predikcionih performansi modela

Da bi se procenile i poredile sposobnosti predviđanja modela mašinskog učenja na novim podacima koriste se standardne mere za evaluaciju. Vrednovanje kvaliteta modela i izbor mere zavise od tipa evaluiranih modela: modeli za klasifikaciju ili modeli za regresiju. U toku ove studije bavimo se problemima klasifikacije i stoga su korišćene adekvatne mere evaluacije otkrivenog znanja. Cilj evaluacije modela za predviđanje IPP je izmeriti u kojoj meri se predviđanja modela razlikuju od stvarne klasifikacije IPP u pozitivne i negativne. Evaluacija se uvek vrši na skupu IPP koji nije učestvovao u izgradnji modela (test skup). Pri procesu evaluacije stvarna podela IPP na pozitivne i negativne je eliminisana iz testnog skupa. Stvarno stanje klasifikacije se koristi da bi se poredila predviđanja sugerisana modelom. Razlika u prognozi klase u odnosu na stvarnu klasu IPP se označava klasifikacionom greškom. Različite mere evaluacije ukupan broj grešaka na test skupu prezentuju u različitoj formi i međusobno su komplementarne. Problem prognoziranja IPP je problem binarne klasifikacije. Razlikuju se dve grupe: proteini koji interaguju i koji ne interaguju. Razlikujemo dve vrste grešaka koje se mogu prikazati u formi matrice grešaka kod klasifikacionog problema (Tabela 5). Potreba za primenom različitih mera evaluacije se javlja usljed različite osetljivosti na razne tipove grešaka.

Tabela 5. Matrica grešaka za klasifikacioni problem predviđanja IPP.

		Stvarna klasa	
		Pozitivni IPP	Negativni IPP
Predviđena klasa	Pozitivni IPP	Stvarno pozitivni IPP	Lažno pozitivni IPP
	Negativni IPP	Lažno negativni IPP	Stvarno negativni IPP

Kao ishod poređenja stvarnog stanja u evaluacionom skupu u odnosu na klasu predviđenu generisanim modelom možemo imati sledeće situacije. Ukoliko je model ispravno pretpostavio postojanje ili odsustvo interakcije, rezultat prognoze su stvarno pozitivni IPP (SP) i stvarno negativni IPP (SN), retrospektivno. U slučaju greške pri klasifikaciji,

ukoliko je postojeća, verifikovana interakcija označena negativnom, prognoza se označava lažno negativnom IPP (LN), a u obrnutom slučaju lažno pozitivnom IPP (LP).

Postoje dve grupe mera za evaluaciju performansi prediktora:

a) Mere koje se računaju na osnovu matrice grešaka
b) Mere nezavisne od matrice grešaka (Površina ispod krive)

3.3.1.1 Mere koje se računaju na osnovu matrice grešaka

Tačnost predstavlja udeo tačnih predviđanja u ukupnom broju predviđanja. Vrednost se kreće od 0 do 1, pri čemu savršen prediktor ima tačnost 1. Tačnost je data jednačinom:

$$AAC = \frac{SP + SN}{SP + SN + LP + LN} \tag{3}$$

Odziv (senzitivnost ili stopa stvarno pozitivnih - SSP), predstavlja udeo stvarno pozitivnih u ukupnom broju onih koji su originalno pozitivni, i definisana je jednačinom:

$$Odziv = \frac{SP}{SP + LN} = 1 - SLN \tag{4}$$

Odziv se može definisati i preko stope lažno negativnih (SLN). Vrednosti se kreću od 0 do 1.

Specifičnost (stopa stvarno negativnih - SSN) je definisana kao udeo slučajeva gde je negativna klasa ispravno predviđena, a data je formulom:

$$Specifičnost = \frac{SN}{SN + LP} = 1 - SLP \tag{5}$$

Specifičnost se može definisati preko stope lažno pozitivnih (SLP). Vrednosti se kreću od 0 do 1.

Stopa lažno negativnih (SLN) predstavlja udeo slučajeva gde je pripadnost negativnoj klasi pogrešno predviđena u odnosu na ukupan broj pozitivnih slučajeva. Vrednosti SLN se kreću od 0 do 1, i računa se formulom:

$$SLN = \frac{LN}{SP + LN} = 1 - SSP \tag{6}$$

Stopa lažno pozitivnih (SLP) se opisuje kao odnos negativnih primera koje su pogrešno označeni kao pozitivni u odnosu na ukupan broj negativnih. Vrednosti SLP se kreću od 0 do 1, a računa se formulom:

$$SLP = \frac{LP}{SN + LP} = 1 - SSN \tag{7}$$

Preciznost predstavlja udeo predviđenih pozitivnih primera koji su tačni u ukupnom broju pozitivno predviđenih primera. Vrednosti se kreću od 0 do 1. Veća vrednost označava precizniji model, sa maksimalnom vrednosti 1.

$$Preciznost = \frac{SP}{SP + LP} \tag{8}$$

F mera se predstavlja harmonijsku sredinu preciznosti i odziva. Vrednosti se kreću od 0 do 1. Klasifikator perfektnog odziva i preciznosti (vrednost 1) ima takođe maksimalnu vrednost F mere = 1.

$$F\ mera = \frac{2SP}{2SP + LP + LN} \tag{9}$$

Matejev koeficijent korelacije (engl. Matthews correlation coefficient - MCC), balansirana mera performansi binarnog klasifikatora, osetljiva je na odnose ukupnog

broja pozitivnih i negativnih primera u test skupu. MCC uzima u obzir vrednosti svih elemenata matrice grešaka, i u finalnom skoru MCC vrednosti se kreću od -1 do 1. Perfektan klasifikator ima vrednost MCC = 1.

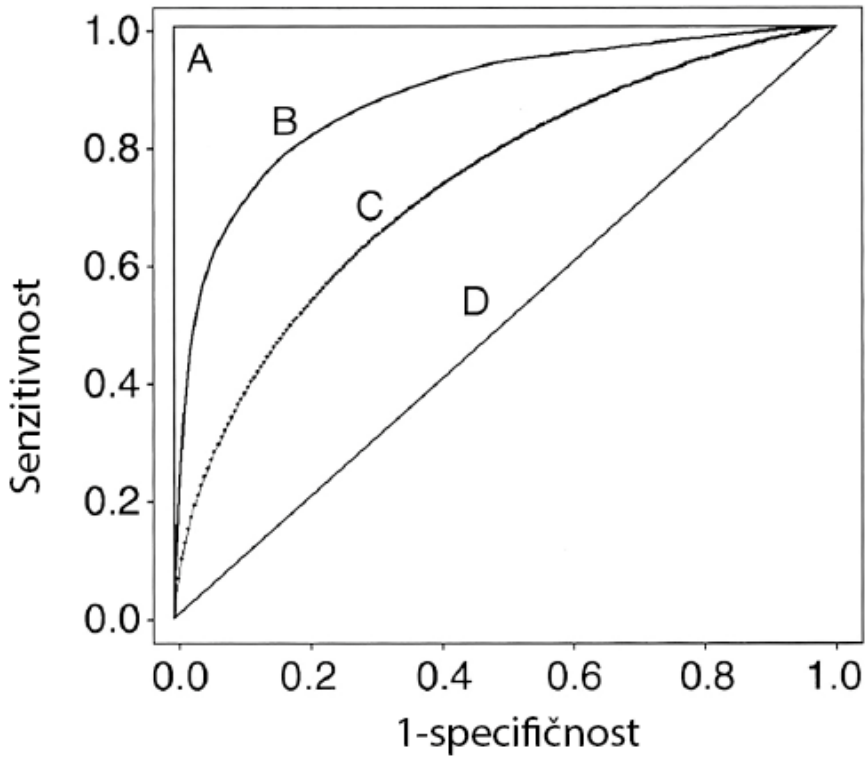
$$MCC = \frac{SP * SN - LP * LN}{\sqrt{(SP + LP)(SP + LN)(SN + LP)(SN + LN)}} \quad (10)$$

3.3.1.2 Mere koje se ne zasnivaju na matrici grešaka

Većina klasifikatora predikcije izveštava u vidu broja, odnosno unutrašnje mere verovatnoće ili poverenja. Ova "verovatnoća" se kreće od 0 od 1. Da bi određena predikcija bila kategorisana u jednu ili drugu klasu, standardno se uzima prag "verovatnoće" 0.5. Svaka predviđena IPP gde je sigurnost klasifikatora bila ≥ 0.5 se svrstava u klasu pozitivnih, a ≤ 0.5 u klasu negativnih interakcija.

Mere koje se zasnivaju na matrici grešaka ignorišu vrednosti ove verovatnoće, uzimajući u obzir samo da li je ta verovatnoća iznad ili ispod određenog praga. Analiza grešaka klasifikacionog algoritma i formiranje matrice grešaka se vrši nakon primene praga "verovatnoće" na dobijene predikcije. Menjanjem praga "verovatnoće" kojom se razdvajaju klase moguće je povećati vrednost jedne mere performansi prediktora umanjujući druge. Može doći do povećanja senzitivnosti modela ali će to smanjiti njegovu specifičnost. Mere performansi modela koje uzimaju u obzir sve elemente matrice greške, nezavisne su od praga "verovatnoće" i koriste vrednosti "verovatnoće" prediktora su mere površine ispod krive (engl. Area under curve).

Površina ispod "receiver operator curve" (engl. Receiver Operator Curve, ROC), odnosno AUROC vrednost (engl. Area Under Receiver Operator Curve) je najčešće korišćena mera performansi klasifikatora. ROC je dvodimenzionalna kriva prikazana u formi grafika gde su na X osi predstavljene SLP vrednosti (1-specifičnost) dok Y osa predstavlja senzitivnost. Tačke koje formiraju ROC krivu predstavljaju vrednosti SLP za odgovarajuću vrednosti senzitivnosti. Povezivanjem svih pragovima "verovatnoće" formira se ROC kriva (Fawcett, 2004; Park, Goo and Jo, 2004). Površina ispod ove krive se kreće od 0 do 1, gde 1 predstavlja savršen klasifikator (Slika 4 A).



Slika 4. Primer ROC krive u prikazivanju performansi prediktora. D. Dijagonalna kriva označava prediktor čija sposobnost predviđanja je jednaka nasumičnom nagađanju (AUROC = 0.5). A. AUROC = 1, prediktor perferktno razlikuje dve klase na test skupu. B i C leže između dva ekstrema pri čemu klasifikator sa krivom B bolje predviđa od klasifikatora sa krivom C.

Površina ispod krive preciznost/odziv (engl. Precision recall curve, PRC), odnosno AUPRC vrednost (engl. Area Under Precision recall curve, AUPRC), je mera performance prediktora, koja se predstavlja u formi grafa i sličnih je osobina kao AUROC. AUPRC u obzir uzima sve elemente matrice greške, nezavisna je od praga "verovatnoće" i koristi "verovatnoće" prediktora pri formiranju PRC. Za razliku od AUROC grafa, kod AUPRC grafa na X osi predstavljen je odziv, a na Y osi preciznost. Primenjuje se komplementarno AUROC meri, pogotovo u slučajevima nejednakog broja primera jedne klase u odnosu na drugu u test skupu. U slučajevima neizbalansiranosti klasa, pri čemu jedne klase ima daleko više nego druge, AUROC može dati preterano optimističan prikaz performansi prediktora jer je neosetljiv na ukupan broj pripadnika veće klase. Pri merenju performansi IPP klasifikatora čest slučaj je da negativnih IPP

primera ima više nego pozitivnih unutar test grupe. U takvim slučaju potrebno je imati realističan prikaz performansi prediktora, te se stoga AUROC i AUPRC koriste kao komplementarne mere (Davis and Goadrich, 2006; Park, 2009; Saito and Rehmsmeier, 2015).

3.3.2 Selekcija atributa: Analiza glavnih komponenti

Analiza glavnih komponenti (engl. Principal component analysis, PCA) je tehnika za redukciju broja atributa i ekstrakciju relevantnih informacija iz skupova podataka. Cilj PCA procedure je transformisati korelisane ulazne atribute u manji broj nekorelisanih sintetičkih varijabli koje se označavaju *glavnim komponentama.* Efikasnost redukcije broja atributa se ogleda u očuvanju velikog dela informativnosti velikog skupa atributa u novom sintetičkom skupu. Statistički ova informativnost podataka se definiše varijansom. Prva *glavna komponenta* sadrži najveći stepen informativnosti u podacima, dok svaka subsekventna komponenta sadrži deo preostale informativnosti. U PCA analizi, slično kao i kod ML algoritama, ulazni podaci se analiziraju u formi matrica, gde su primeri predstavljeni vektorima vrednosti atributa. Vektori primera se posmatraju u algebarskom prostoru čiji je broj dimenzija određen brojem atributa, a pozicija vektora njihovim vrednostima. Osnovni zadatak PCA metoda je, primenom matematičkih transformacija originalnih podataka, pronaći nove koordinatne ose (*glavne komponente)* takve da se sa manjim brojem osa opišu ulazni podaci i tako smanji broj atributa kojima se podaci opisuju. Budući da su koordinatne ose ortogonalne, nove sintetičke varijable koje ove ose predstavljaju su međusobno nekorelisane. Nove koordinatne ose predstavljaju linearnu kombinaciju originalnih atributa. Analizom otežanja (doprinosa) originalnih atributa definiše se njihov doprinos informativnosti koju opisuje određena *glavna komponenta* (Jolliffe, 2011; Hu and Tsay, 2014; Shlens, 2014). U toku analiza predstavljenih u ovom istraživanju korišćena je PCA metoda implementirana u R programskom jeziku.

3.4 Algoritmi mašinskog učenja

3.4.1 Nasumične šume

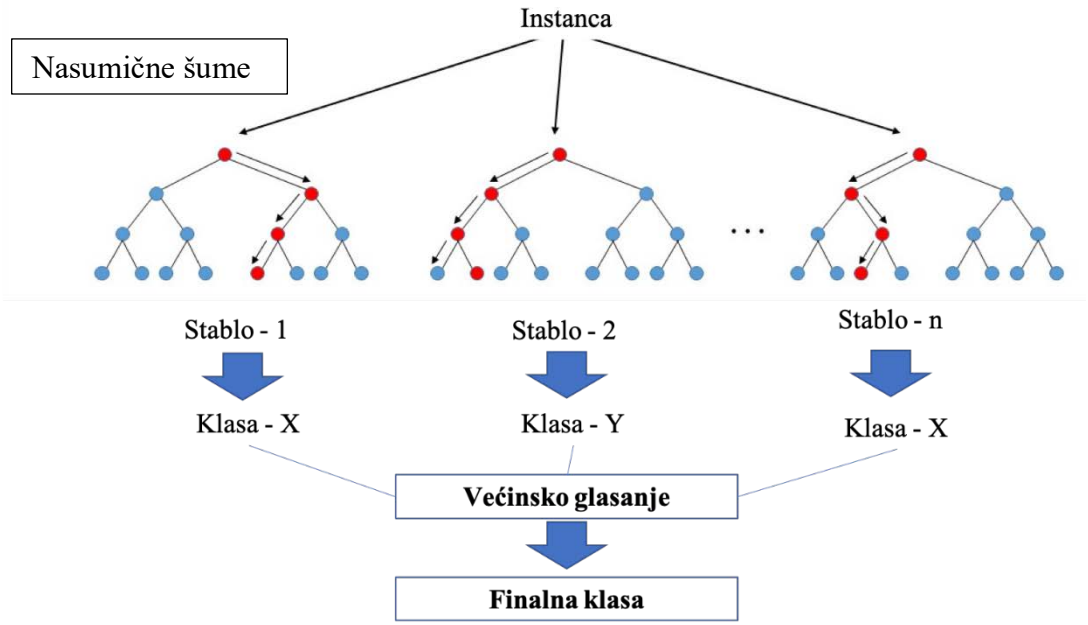
Nasumične šume (engl. Random Forest, RF) je ansambl algoritam mašinskog učenja za klasifikaciju i regresiju, koji je razvio Leo Breiman (Breiman, 2001). Nasumična šuma se formira od većeg broja nerazgranatih (engl. unpruned) stabala odlučivanja. Svako stablo odlučivanja u „šumi“ se trenira na slučajnom uzorku od ukupnog broja primera gde je dozvoljeno njihovo ponavljanje. Visok stepen generalizacije modela formiranih RF algoritmom je obezbeđen tehnikom nasumičnog biranja atributa. Ovim pristupom se svako drvo odlučivanja uči pomoću nasumično izabranog podskupa atributa. Proces konstrukcije stabala odlučivanja na podskupu atributa je kontrolisan kriterijumom za izbor atributa. Kod stabala odlučivanja za klasifikaciju i regresiju (engl. classification and regression tree, CART), u problemima klasifikacije, kao kriterijum za izbor najboljeg atributa se koristi smanjenje stepena pogrešnog klasifikovanih primera, takozvana *Gini impurity* mera. Odlučivanje, primenom naučenog znanja u formi RF modela na skupu novih podataka, se vrši “glasanjem” većine stabala odlučivanja, u slučaju klasifikacije, za jednu od dve klase (Slika 5).

Tokom razvoja i testiranja metoda za predviđanje IPP transkripcionih regulatora korišćena je implementacija RF u ‘randomForest’ paketu R (http://cran.r-project.org/) programskog jezika (Liaw and Wiener, 2002). Da bi se omogućilo korišćenje prednosti u brzini izvršenja algoritma koje nudi računarska paralelizacija, izvršena je paralelizacija originalnog koda iz ‘randomForest’ paketa. Detektovano je višestruko povećanje vremena izvršenja RF algoritma. Jedna od ključnih prednosti RF algoritma i razlog njegove široke primene u raznim oblastima bioinformatike, jesu visoke prediktivne performanse sa standardnim postavkama algoritma. Dva parametra RF algoritma sa najvećim uticajem na sposobnost generalizacije jesu broj nasumično izabranih atributa po stablu odlučivanja (m_{try}) i ukupan broj stabala u ansamblu (*ntrees*) (Breiman, 2001). Upotrebljena je standardna vrednost ntrees parametra (*ntrees* = 500 stabala) dok je optimalan mtry (m_{try} = 9) pronađen primjenom mrežne pretrage (engl. grid search) koristeći CV-10 .

Za razvoj alata za predviđanje IPP PNTS, korišćen je RF algoritam implementiran u JAVA programskom jeziku. Kontrola i podešavanje algoritma vršeno je u R programskom jeziku uz upotrebu RStudio razvojnog okruženja (RStudio Team, 2016). Ova paralelna implementacija RF algoritma je omogućila fino podešavanje dodatnih parametara algoritma u cilju što bolje generalizacije na test skupovima velikog disbalansa pozitivnih i negativnih IPP. Fino tjunirani parametri algoritma pored broja stabala u ansamblu i broja nasumično izabranih atributa po stablu odlučivanja su: stepen razgranatosti pojedinačnih stabala, minimalan broj primera u listovima, tip histograma, stepen diskretizacije podataka i veličina podskupa primera na kojima se uči. Pronalaženje optimalnog skupa meta-parametara za predviđanje PNTS IPP vršeno je pomoću nasumične pretrage (Bergstra and Bengio, 2012). Algoritam nasumične pretrage je implementiran u R programskom jeziku.

U toku razvoja pristupa za predviđanje IPP proteoma čoveka korišćen je distribuirani RF algoritam implementiran u JAVA programskom jeziku.

U sklopu ove studije, a u svrhu razvoja modela za predviđanje IPP čoveka, razvijen je automatski proces za generisanje i selekciju atributa čiji supervizovani deo čini genetski algoritam (GA). U sklopu GA koristi se algoritam ML. U svrhu minimizacije računarskog vremena i resursa potrebnih za izvršavanje GA, izabrana je C++ implementacija RF algoritama koja se poziva u formi R paketa *ranger*. Prema analizi Wright i Ziegler (Wright and Ziegler, 2017), *ranger* je najbrža i memorijski najefikasnija RF implementacija u odnosu na testirane implementacije.



Slika 5. Šematski prikaz osnova algoritma nasumičnih šuma.

3.4.2 Naivni Bajes

Naivni Bejesov klasifikator (engl. Naive Bayes, NB) je familija probabilističkih algoritama mašinskog učenja zasnovan na Bajesovoj teoremi (Hand and Yu, 2001). Osnovna pretpostavka za korišćenje NB algoritma je nezavisnost pojedinačnih atributa u skupu za treniranje. NB pronalazi optimalni klasifikator na skupu podataka za učenje, računajući frekvencije pojavljivanja vrednosti određenog atributa u ukupnom broju primera skupa za trening. Pretpostavljajući nezavisnost atributa, pronalazi se optimalna kombinacija verovatnoća pojedinačnih atributa za svaku zadatu klasu. Prednosti BN algoritma su visoka jednostavnost, visoka brzina treniranja, testiranja i visoka stepen neosetljivosti na irelevantne atribute u skupovima (Hand and Yu, 2001; Glickman and van Dyk, 2007).

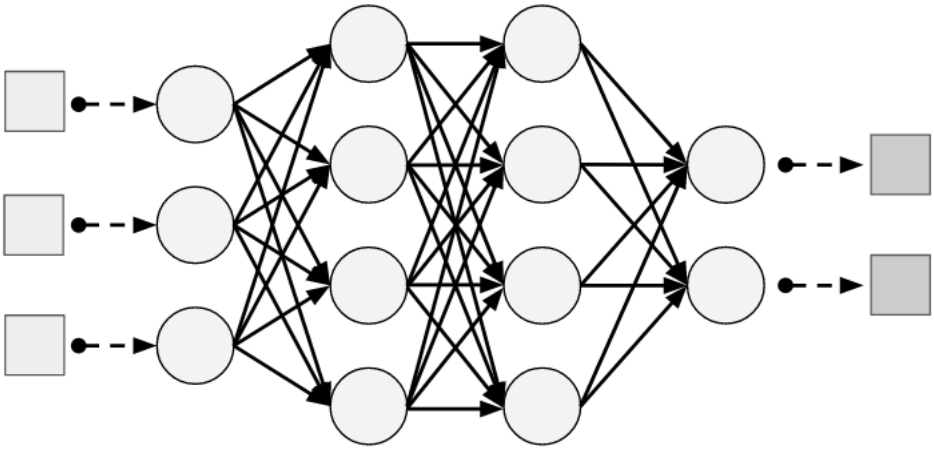
U sklopu ove studije korišćena je implementacija NB u JAVA programskom jeziku u sklopu „h2o“ paketa (Landry, 2018) R programskog jezika.

3.4.3 Neuronske mreže i Duboko učenje

U pokušaju simuliranja rada ljudskog mozga matematičkim strukturama, došlo je do razvoja neuronskih mreža. Neuronske mreže se ne mogu direktno svrstati u metode mašinskog učenja iako im je zajednička sposobnost generalizacije kao rezultata nadgledanog učenja. Pod određenim uslovima, određene neuronske mreže konačnog broja neurona se mogu koristiti za aproksimaciju funkcija (Hornik, 1991). U toku godina ubrzanog razvoja i širokih primena neuronskih mreža, predstavljeni su različiti tipovi neuronskih mreža. Ipak, osnovu strukture svake neuronske mreže čine međusobno povezani veštački neuroni. U osnovi, neuron je parametrizovana funkcija kojom se računa linearna kombinacija ulaznog signala u formi sume produkata komponenti signala sa otežanjima. Ulazni signal je predstavljen u vektorskoj formi. Nelinearna aktivaciona funkcija sumu otežanog ulaznog signala zajedno sa parametrom modela prevodi u vrednosti od 0 do 1. Ukoliko rezultat nelinearne transformacije aktivacionom funkcijom pređe unapred zadati prag, neuron se smatra aktiviranim i signal se prostire dalje. Prostiranje signala u neuronskoj mreži je hijerarhijsko i kreće se od neurona ulaznog sloja prema neuronima izlaznog sloja. Proces formiranja modela neuronske mreže podrazumeva pronalazak kombinacija otežanja ulaznih signala svakog neurona, koji se posmatraju kao parametri sistema. Za optimizaciju neuronske mreže se koriste gradijentni metodi. Za podešavanja parametara neuronske mreže, u svrhu smanjenja greške predikcije, upotrebljava se algoritam *propagacije unazad* (engl. backpropagation). Ukoliko neuronska mreža između ova dva sloja sadrži jedan ili više sakrivenih slojeva neurona označava se kao duboka neuronska mreža (engl. Deep neural network, DNN) (Slika 6). Formiranje modela dubokih neuronskih mreža uključuje učenje reprezentacije ulaznog signala u formi svojstvenih reprezentacija bez neophodnog domenskog znanja, što se označava kao Duboko učenje (engl. Deep learning, DL) (Lecun, Bengio and Hinton, 2015; Schmidhuber, 2015). Primena DL zahteva visok stepen neophodnog stručnog znanja za odabir najadekvatnije arhitekture mreže za ciljni problem i podešavanje velikog broj meta-parametara mreže neophodnih za efikasnu generalizaciju. Pored ovoga, za efikasan DL model potrebni su veliki broj primera za učenje, visoke računarske performanse hardvera i prevazilaženje niza problema u procesu kombinovanja raznih meta-parametara.

Za potrebe poređenja modela formiranih različitim ML metodama na skupovima proteoma čoveka korišćena je implementacija DL u JAVA programskom jeziku u sklopu *h2o* R paketa (Cook, 2016). Optimalna arhitektura mreže i meta-parametri sistema podešavani su nasumičnom pretragom (Bergstra and Bengio, 2012) i primenom heuristike.

U sklopu algoritma za automatski izbor ML algoritama i njihovo ansembliranje, razvijen u ovoj studiji, korišćena je *Deep Water* implementacija DL algoritma, u sklopu h2o softverskog paketa (Phan *et al.*, 2017). *Deep Water* implementacija omogućava korišćenje grafičkih procesora (engl. graphics processing unit**,** GPU) u cilju višestrukog ubrzanja procesa formiranja i optimizacije dubokih neuronskih mreža. Za osnovu arhitekture neuronske mreže korišćena je standardna MXnet postavka (T. Chen *et al.*, 2015).



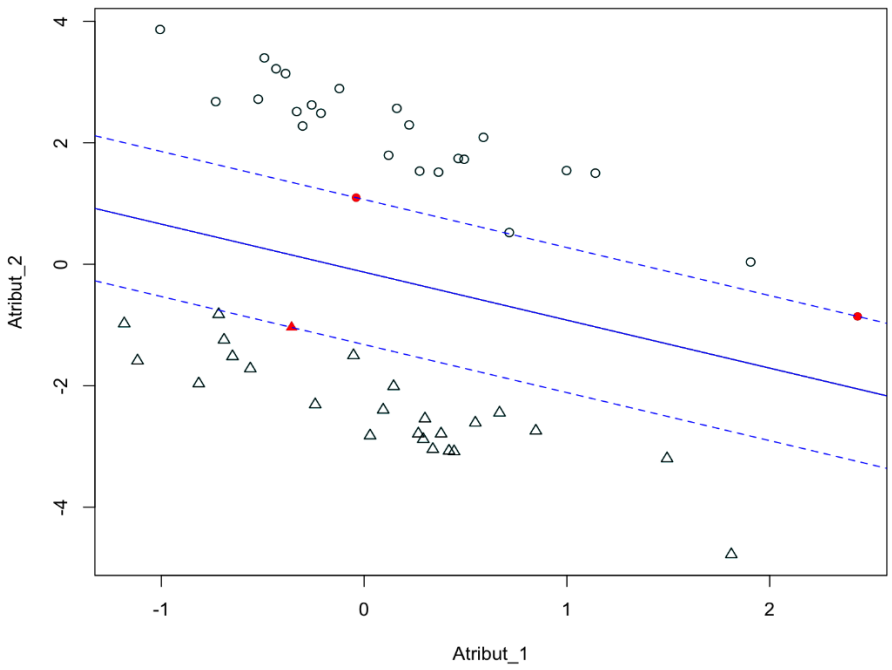
Slika 6. Šematski prikaz duboke neuronske mreže za binarnu klasifikaciju sa dva sakrivena sloja (Gibson and Patterson, 2017).

3.4.4 Metoda potpornih vektora

Metoda potpornih vektora (engl. Support Vector Machine, SVM) je u osnovi linearni algoritam mašinskog učenja razvijen 1999. godine od strane Vapnika (Vapnik, 1999). SVM algoritam ima za cilj da pronađe maksimalnu marginu razdvajanja između hiper-ravni i njoj najbližih tačaka. Svaka instanca trening skupa je predstavljena u

višedimenzionom prostoru vrednostima svojih atributa. SVM algoritam konstruiše hiper-ravan sa maksimalnom marginom razdvajanja između instanci koje pripadaju različitim klasama. Formiranje hiper-ravni je definisano instancama za učenje koji su joj najbliži, te se takve instance označavaju potporni vektori (Slika 7). U slučaju nemogućnosti linearnog razdvajanja tačkama predstavljenih primera, koristi se *kernel trik* (engl. Kernel trick). U proceduri kernel trika ulazni vektor vrednosti atributa kojima je svaka instanca predstavljena se preslikava u višedimenzioni prostor *kernel* funkcijom (engl. kernel). Za pronalaženje optimalne hiper-ravni (hipoteze) nije neophodno eksplicitno računanje lokacije instanci u preslikanom višedimenzionom prostoru, već samo njihova međusobna udaljenost, koja se predstavlja u vidu skalarnog proizvoda vektora instanci.

U toku rada na ovoj studiji korišćena je LibSVM implementacija u C++ programskom jeziku (Chang and Lin, 2013). U eksperimentima poređenja ML algoritama na istim skupovima korišćen je Gausova kernel funkcija (engl. Radial Basis Function, RBF). Meta-parametri SVM algoritma sa RBF kernelom C i γ, su optimizovani su na osnovu mrežne pretrage.



Slika 7. Primer razdvajanja linearno razdvojivih primera algoritmom potpornih vektora. SVM algoritam pronalazi optimalnu hiper-ravan (puna plava linija), sa maksimalnim marginama (isprekidane plave linije) razdvajanja koje se oslanjanju na potporne vektore (crvene tačke i trougao).

3.4.5 Uopšteni linearni modeli

Uopšteni linearni modeli (engl. Generalized linear model, GLM) (Nelder and Wedderburn, 1972), predstavljaju širu klasu algoritama od kojih su najčešće korišćeni linearna regresija i logistička regresija. Jedan od najjednostavnijih i najšire korišćenih linearnih modela je linearna regresija koja ima za cilj da predstavi linearnu zavisnost između zavisne i nezavisne varijable. Ukoliko modeliranje podrazumeva pronalaženje linearnog odnosa zavisne varijable i više nezavisnih varijabli (atributa) naziva se višestruka linearna regresija. U slučaju da modeliranje osim kontinuiranih atributa uključuje i kategoričke, i da raspodela aminokiselinskih ostataka sledi normalnu distribuciju, takvi modeli se nazivaju generalnim linearnim modelima (engl. General linear model). Svaki GLM model obuhvata tri komponente: raspodelu zavisne varijable, linearni prediktor (linearnu kombinaciju nezavisnih varijabli) i *funkciju veze* (engl. Link function). Za analize u ovom radu je korišćena logistička regresija u sklopu šire GLM familije algoritama. Osnovna razlika logističke regresije i raznih vrsta linearne regresije jeste u binarnoj prirodi zavisne varijable kod logističke regresije, što omogućuje binarnu klasifikaciju. Zavisna varijabla sledi binomnu raspodelu. Kod logističke regresije verovatnoća pripadnosti nekoj klasi je predstavljena kao linearna funkcija kombinacije atributa koji mogu biti kontinuirani ili kategorički, stoga se ona naziva mešoviti linearni prediktor. Kao *funkcija veze* između zavisne varijable i linearnog prediktora se koristi određena sigmoidna funkcija (Olsson, 2002; Agresti, 2003).
Za svrhu ove studije korišćena je implementacija GLM algoritma programskom jeziku JAVA u sklopu „h2o“ softverskog paketa (Cook, 2016).

3.4.6 Gradijentno pojačavanje

Metode gradijentnog pojačavanja (engl. Gradient boosting machines, GBM) (Friedman, 2001, 2002) predstavljaju klasu algoritama mašinskog učenja zasnovanih na iterativnom ansambliranju *slabih* (engl. weak) modela. Breiman je demonstrirao da kombinovanje više *slabih* modela može dovesti do poboljšanja predikcionih performansi (Breiman, 2001). Za razliku od RF, gde se za konačno predviđanje uzima prosek predikcija svih

baznih modela, metodi koji se oslanjanju na *boosting* tehniku, bazne modele dodaju u odnosu na stepen ukupne greške do tada formiranog anasambla. Osnove GBM metoda čine tri elementa: *funkcija greške* (engl. loss function), *slabi* prediktori i aditivni model. Zadatak aditivnog modela je sekvencijalno dodavanje baznih modela kako bi se minimizovala *funkcija greške.* U zavisnosti od tipa nadgledanog učenja (klasifikacija ili regresija) i zadatka konačnog modela definiše se *funkcija greške.* Optimizacija *funkcije greške* se postiže tehnikom gradijentnog spusta. Pošto su u toku analiza ove studije uključeni isključivo klasifikacioni zadaci, korišćena je binomna *funkcija greške.* Kao bazni modeli u implementacijama GBM metoda koje smo aplicirali, koriste se stabla odlučivanja. *Slabi* modeli se formiraju na primerima koji su pogrešno klasifikovani prethodnim modelima u ansamblu. Na ovaj način subsekventni modeli se „fokusiraju" na teško klasifikovane primere kako bi se otežavanjem predikcija svih modela u konačnom ansamblu minimizovala *funkcija greške.* Aditivni model predstavlja algoritam računanja *funkcije greške* i bazni modeli se generišu tako da se predikciona greška smanjuje ka minimumu *funkcije greške*. Algoritam se zaustavlja kada se dostigne definisani broj sekvencijalnih baznih modela ili u slučaju da dodavanje novih modela u konačni ansambl ne doprinosi smanjenju predikcione greške (Natekin and Knoll, 2013; Chen and Guestrin, 2016) Posljednji pristup je korišćen u procesu formiranja GBM modela u ovoj tezi. U cilju sprečavanja preteranog prilagođavanja modela podacima iz treninga i povećanja sposobnosti generalizacije GBM algoritmi sadrže veći broj meta-parametara za podešavanje.

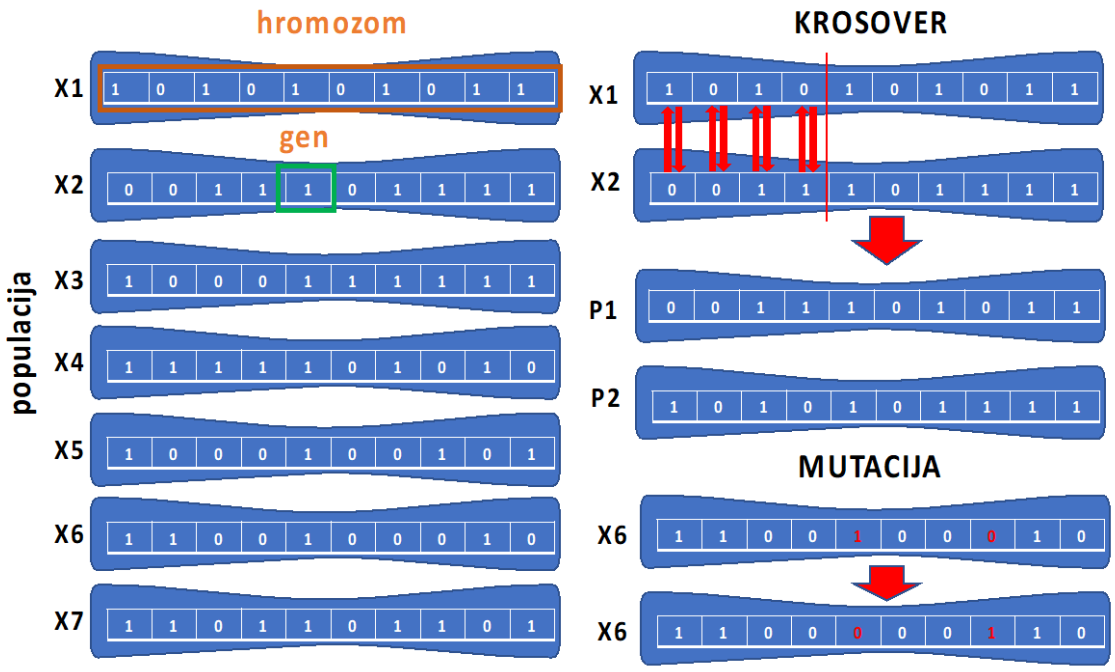
Dve implementacije GBM modela su korišćene u ovoj studiji: GBM metoda implementirana u JAVA programskom jeziku u sklopu u sklopu „h2o" softverskog paketa (Cook, 2016) i XGBoost (engl. Extreme Gradient boosting) implementiran u C++ programskom jeziku (Chen and Guestrin, 2016). XGBoost, koji je reimplementiran u JAVA programskom jeziku u sklopu „h2o" alata, apliciran je primenom CUDA tehnologiju za hiper-paralelizaciju.

Optimizacija meta-parametara oba algoritma je vršena tehnikom *nasumične pretrage* (Bergstra and Bengio, 2012).

3.5 Genetski algoritmi

Genetski Algoritam (GA) je stohastička metoda rešavanja optimizacionih problema, dizajnirana da oponaša biološku evoluciju (Mitchell, 1996). U analizama predstavljenim u ovoj tezi GA je primenjen za pronalaženje optimalnog podskupa atributa za predviđanje IPP metodama mašinskog učenja. Cilj je bio naći optimalnu kombinaciju atributa nekog skupa IPP za predviđanje novih interakcija među proteinima. GA simulira proces prirodne selekcije gde se iz populacije jedinki jedne generacije, izaberu najsposobniji organizmi za razmnožavanje, kako bi se formiralo potomstvo u sljedećoj generaciji. GA se izvodi na populaciji jedinki i bez obzira na implementaciju i optimizacioni problem izdvajamo 5 faza: formiranje početne populacije, određivanje stepena prilagođenosti jedinke (odnosno *fitnes* mere), selekcija, ukrštanje (krosover) i mutacija. Proces započinje formiranjem populacije jedinki. Svaka jedinka predstavlja jedinstven skup rešenja. U kontekstu selekcije atributa, jedinka je predstavljena hromozomom a pojedinačni atribut genom. Svaka jedinka poseduje kombinaciju od ukupnog skupa gena (atributa) gde su prisutni atributi označeni sa 1 a odsutni sa 0. Drugi element GA je funkcija fitnesa. Za analize u ovoj studiji AUROC je izabrana kao funkcija fitnesa koja se maksimizuje. Fitnes mera jedinke se računa unakrsnom validacijom modela mašinskog učenja formiranog na binarno predstavljenom podskupu atributa koji čine tu jedinku. Proces modeliranja i računanja AUROC vrednosti, zatim izbor individua sa najvećom vrednosti fitnesa predstavlja fazu selekcije. U toku ukrštanja između jedinki sa najvećom vrednosti fitnesa dolazi do razmjene genetskog materijala. Nasumičnim mešanjem binarnih vektora kojima je predstavljena prisutnost atributa jedinki formira se potomstvo. Pored nasleđenih promena, u toku faze mutacije potomstvo nove generacije se nasumično podvrgava nasumičnim mutacijama niske verovatnoće pojavljivanja u populaciji. Čitav proces od pet faza se ponavlja u toku zadatog broja generacija dok se ne pronađu jedinke dovoljno visoke vrednosti fitnesa (Tsai, Eberle and Chu, 2013; Oluleye and Armstrong, 2014) (Slika 8).

Korišćena je implementacija GA algoritma u R paketu „caret“ (Kuhn, 2008).



Slika 8. Šematski prikaz genetskog algoritma. Populaciju čine jedinke (X1-X7) predstavljene hromozomima. Hromozomi su binarni vektori prisustva atributa koji se optimizuju. U toku krosover faze nasumično izabrane jedinke (X1,X2) formiraju potomstvo (P1,P2). U toku krosovera nasumično se uspostavlja tačka do koje delovi hromozoma se mješaju kod potomaka. Faza mutacije uključuje nasumične promjene na nasumično izabranom članu populacije. Stopa mutacije se odlikuje malom verovatnoćom pojavljivanja u populaciji.

3.5.1 GAFT algoritam

GAFT algoritam, koji je razvijen u ovoj studiji u svrhu automatskog generisanja i selekcije atributa za treniranje modela za predviđanje IPP kod čoveka, implementiran je u R programskom jeziku. Za implementaciju korelacione analize korišćen je FSelector R paket (Romanski and Kotthoff, 2009). Za optimizaciju R koda i izvršavanje zahtevnih aritmetičkih operacija u C jeziku korišćeni su R paketi „plyr“ (Wickham, 2018) i „dplyr“ (Wickham *et al.*, 2016). Posebni moduli algoritma su implementirani „doMC“ (Weston, 2017) paket i druge funkcije za paralelizaciju baznog R paketa, kako bi se omogućilo izvršavanje GAFT algoritma.

3.5.2 GA-STACK algoritam

Algoritam GA-STACK, koji je razvijen i primenjen u sklopu razvoja metode za predviđanje IPP čoveka, kao i protokol *nasumične pretrage* svih algoritama ML korišćenih u sklopu GA-STACK algoritma su implementirane u R jeziku. Algoritmi mašinskog učenja čiji modeli formiraju GA-STACK su implementirani u JAVA programskom jeziku u sklopu „h2o“ softverskog paketa (Cook, 2016).

4 Rezultati

4.1 Predviđanje IPP transkripcionih regulatora

4.1.1 Skupovi podataka

Ljudski geni uključeni u transkripcionu regulaciju identifikovani su pretraživanjem GO baze podataka. Pretraživanje je izvršeno pomoću AMIGO web alata. Izdvojeni su geni asocirani sa terminom genske ontologije GO:0006355. Ovaj termin obuhvata gene i njihove produkte uključene u DNK-zavisnu ćelijsku transkripciju. U sklopu ovog termina, transkripciona regulacija se definiše kao bilo koji proces kod koga se modulira frekvencija, stopa ili stepen DNK-zavisne transkripcije u ćeliji.

Pronađeno je 5337 proteina asociranih sa terminom transkripcione regulacije. Sve proteinske sekvence su preuzete iz UniProtKB/Swiss-Prot baze podataka. Proteinske sekvence koje u nazivu sadrže riječi ‘putative’, ‘potential’ ili ‘uncharacterized’ se smatraju nepouzdanim, pa je preporučljivo da se izostave iz dalje analize (Patthy, 2016). Ukupno 135 ovakvih sekvenci je eliminisano iz skupa od 5337 proteina. Takođe, 177 proteina i proteinskih fragmenta kraćih od 50 aminokiselina su izuzeti iz konačne liste proteina. Dužina sekvence od ≤ 50 aminokiselina se najčešće smatra granicom koja razdvaja proteine od peptida (Moss, Smith and Tavernier, 2007). Metode za predviđanje IPP zasnovani na sekvenci proteina mogu demonstrirati preterano optimistične prediktivne performanse ukoliko skupovi za učenje sadrže homologne proteine (Park, 2009). Proteini sa sličnošću sekvence većom od 40% pronađeni su koristeći CD-HIT alat. Konačna lista sekvenci, nakon svih faza filtriranja, sadrži 1515 proteina, prosečne dužine od 640 aminokiselina.

U cilju pronalaženja binarnih fizičkih IPP u kojima učestvuje 1515 regulatora transkripcije, pretraživana je HIPPIE baza podataka. Od preuzetih 12224 IPP formiran je neredundantan pozitivni skup IPP. Pozitivni IPP skup nije uključivao interakcije između istih proteina (autointerakcije). Skup neinteragujućih parova proteina NIPP, odnosno negativnih IPP, je formiran koristeći tehniku nasumičnog negativnog uzorkovanja. Formirani NIPP skup je balansiran, tako da je broj pozitivnih jednak broju negativnih IPP. Konačan skup IPP sastavljen od negativnog i pozitivnog skupa IPP sa ukupno 24448

interakcija. Ovaj skup je poslužio za testiranje performansi metoda i formiranje konačnog modela predviđanja IPP transkripcionih regulatora.

4.1.2 PAAC4 atributi

Kao osnova za naš statistički model izabranih sekvenci transkripcionih regulatora poslužila je pseudo kompozicija aminokiselina, odnosno PAAC pristup (Chou, 2001). U osnovi PAAC modela su fizičko-hemijske karakteristike aminokiselina: hidrofobnosti, hidrofilnosti i mase bočnih lanaca koristi. Novi PAAC4 model za predstavljanje proteinskih sekvenci, pored prethodno pomenute tri fizičko-hemijske karakteristike aminokiselina, koristi i informaciju o dalekosežnim interakcijama između molekula u formi deskriptora, potencijal elektron-jon interakcija (engl. Electron-Ion Interaction Potential, EIIP) (Veljkovic, 1980) (Tabela 2). U sklopu PAAC4 modela posmatra se korelacija između svake fizičko-hemijske karakteristike aminokiselina na udaljenosti od 50 aminokiselina duž sekvence proteina. Maksimalna dužina λ faktora korelacije je određena minimalnom dužinom sekvence, koja za naš skup proteina iznosi 50 aminokiselina. Veličina λ faktora od 50 unutar PAAC4 modela omogućava posmatranje korelacije između fizičko-hemijskih karakteristika aminokiseline duž sekvence na maksimalnoj udaljenosti. Svaka proteinska sekvenca je transformisana PAAC4 pristupom u numerički vektor dužine 70. Prvih 20 elemenata vektora uključuje AAC vrednosti. Informacija o poretku aminokiseline niza proteina je sačuvana u formi numeričkog vektora dužine 50.

Interakcija dva proteina je predstavljena numeričkim vektorom formiranim spajanjem njihovih PAAC4 vektora. Operacija spajanja vektora || nije komutativna. Interagujući ili neinteragujući par proteina, A i B, je predstavljen je sa dva vektora dužine 140, PAAC4(A)||PAAC4(B) and PAAC4(B)||PAAC4(A).

4.1.3 PAAC4_RF model

Koristeći PAAC4 model, 24448 selektovanih IPP pretvoreno je u numeričke vektore dužine 140, pogodne za formiranje predikcionog modela. Koristeći CV-3 validaciju na izabranom skupu transkripcionih regulatora, testirane su performanse tri algoritma

mašinskog učenja: SVM, NB i RF. RF algoritam je u proseku formirao klasifikatore sa boljom pouzdanošću predikcije novih interakcija između transkripcionih regulatora posmatrajući sve mere performansi prediktora (Tabela 6). Finalni predikcioni model mašinskog učenja formiran je pomoću RF algoritma (PAAC4_RF). Optimizacija meta-parametara RF algoritma je vršena tehnikom mrežne pretrage koristeći CV-3 validaciju. Pronađeni optimalni parametri su: maksimalan broj stabala = 500, i *mtry* = 9.

Tabela 6. Poređenje predikcionih performansi algoritama mašinskog učenja: Naivni bajes, Metoda potpornih vektora, Nasumične šume koristeći 10-stepenu unakrsnu validaciju na skupu IPP transkripcionih regulatora. Predstavljene su srednje vrednosti za osam mera performansi klasifikatora.

	NB	**SVM**	**RF**
AUROC	0.642	0.782	0.878
AUPRC	0.626	0.77	0.879
Tačnost	0.597	0.714	0.798
F1	0.644	0.717	0.796
Preciznost	0.576	0.71	0.803
Specifičnost	0.464	0.703	0.807
Senzitivnost	0.729	0.725	0.789
MCC	0.2	0.428	0.596

4.1.4 Evaluacija PAAC4_RF modela

4.1.4.1 Efikasnost predikcije i poređenje sa standardnim metodama

Metod PAAC4_RF upoređen je sa standardnim, metodama za predviđanje IPP zasnovanim na sekvencama: Guo_M (Guo *et al.*, 2008), PIPE (Pitre *et al.*, 2006) i Shen_M (Shen *et al.*, 2007). PAAC4_RF pristup omogućava predviđanje parova transkripcionih regulatora koji interaguju međusobno sa većom pouzdanošću u odnosu na prethodno testirane metode upoređujući sve statističke mere (Tabela 7), osim u slučaju senzitivnosti kod PIPE metode, gde su autori povećali senzitivnost na uštrb specifičnosti (promena ne utiče na vrednosti AUROC i AUPRC).

Tabela 7. Poređenje klasifikacione sposobnosti PAAC4_RF pristupa u odnosu na standardne metode za predviđanje IPP zasnovane na sekvenci. Srednja vrednosti i standardna devijacija su prikazane za AUROC, AUPRC, tačnost, preciznost, senzitivnost, specifičnost, F1 i MCC.

	PAAC4_RF	**Guo_M**	**PIPE**	**Shen_M**
AUROC	0.878 ± 0.003	0.789 ± 0.005	0.762 ± 0.002	0.690 ± 0.012
AUPRC	0.879 ± 0.003	0.788 ± 0.007	0.762 ± 0.004	0.661 ± 0.017
Tačnost	0.800 ± 0.002	0.715 ± 0.004	0.679 ± 0.001	0.663 ± 0.009
Prec.	0.803 ± 0.000	0.713 ± 0.006	0.644 ± 0.002	0.687 ± 0.012
Senz.	0.791 ± 0.005	0.737 ± 0.004	0.802 ± 0.003	0.597 ± 0.008
Spec.	0.806 ± 0.001	0.694 ± 0.011	0.556 ± 0.004	0.728 ± 0.013
F1	0.796 ± 0.003	0.719 ± 0.001	0.714 ± 0.000	0.639 ± 0.000
MCC	0.597 ± 0.004	0.431 ± 0.009	0.370 ± 0.002	0.328 ± 0.018

4.1.4.2 Intraspecijska evaluacija - predviđanje transkripcionih faktora pacova pomoću PAAC4_RF pristupa

U cilju ispitivanja specifičnosti modela za predviđanje IPP transkripcionih regulatora, PAAC4_RF pristup smo koristili za predviđanje intraspecijskih i interspecijskih IPP TR pacova (*Rattus norvegicus*). Geni pacova asocirani sa terminom genske ontologije GO:006355 su pronađeni pretraživanjem GO baze podataka pomoću AMIGO web alata. Sekvence proteina pacova su preuzete iz UniProtKB/Swiss-Prot baze podataka. U MENTHA meta-bazi podataka pronađeno je 160 primera interakcija između proteina pacova i 107 interakcija gde je jedan od interagujućih partnera protein čoveka, a drugi protein pacova. Koristeći PAAC4_RF model, ustanovili smo da je efikasnost predikcije IPP pacov-pacov 50%, odnosno od 160 intraspecijskih IPP, detektovano je tačno 80. Predviđanje interspecijskih IPP PAAC4_RF modelom testirano je na 107 IPP pacov-čovek, i tačnost predviđanja je iznosila 0.68.

4.1.4.3 Brzina metoda

Pored predikcionih performansi, testirali smo i brzinu algoritma kojim se formira PAAC4_RF model uključujući i brzinu predikcije primera u test skupu (Tabela 8). Dobijene rezultate smo poredili sa brzinom predviđanja novih interakcija na istom test skupu standardnim metodama za predviđanje IPP: Guo_M, PIPE i Shen_M, što je relevantno za formiranje efikasnog veb alata (Tabela 9). Testiranja svih metoda su vršena na trening skupu od 24488 IPP koristeći CV-3 validaciju. Sve metode su implementirane i testirane na PC-u sa CPU 6 jezgara 3.3 GHz , 16 GB RAM i sa LINUX operativnim sistemom.

Tabela 8. Vreme (s) izvođenja različitih faza PAAC4_RF algoritma.

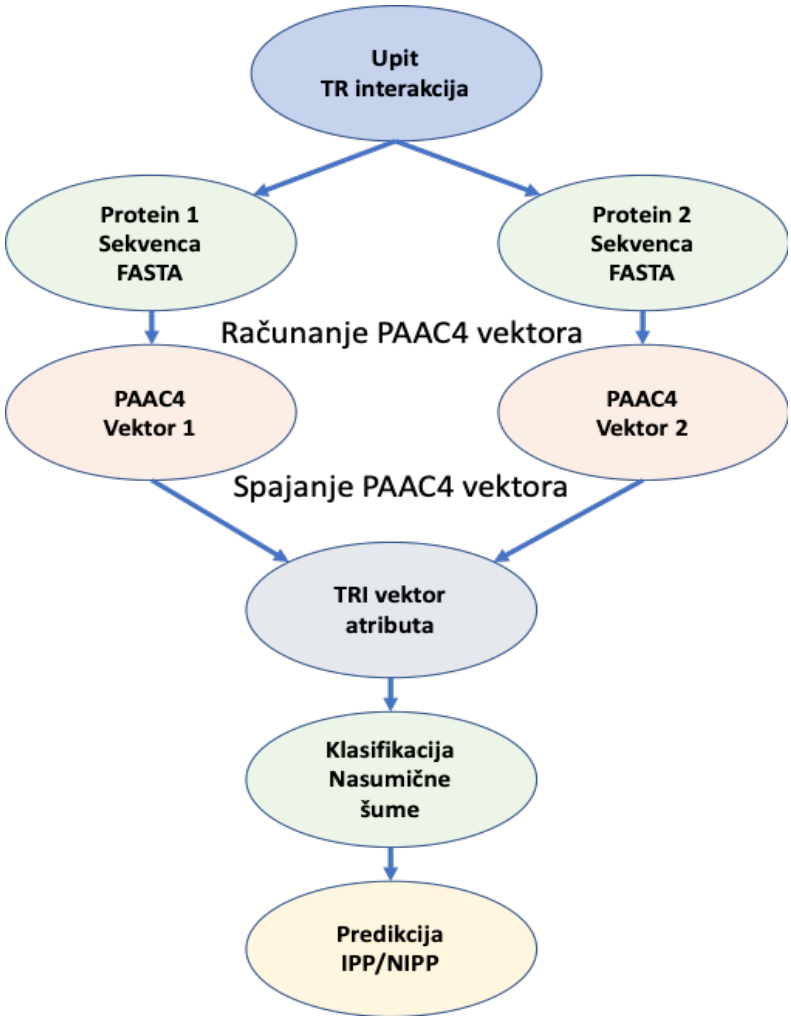
	Računanje PAAC4	**Konstrukcija modela**	**Predviđanje testnog skupa**
Vreme (s)	5s	45s	1s

Tabela 9. Poređenje vremena potrebnog da PAAC4_RF metod izvrši predviđanje novih IPP sa standardnim pristupima. Prikazano je vreme izvršenja u sekundama zajedno sa standardnom devijacijom koristeći CV-3 na 24488 IPP.

	PAAC4_RF	**Guo_M**	**Shen_M**	**PIPE**
Vreme (s)	1s ± 0	636s ± 6	1282s ± 49	47167s ± 377

4.1.5 TRI_tool veb alat

Testirani PAAC4_RF model i algoritam njegova primena za predviđanje IPP transkripcionih regulatora je implementiran u vidu TRI_tool veb alata. TRI_tool algoritam (Slika 9) je, zadržavajući performanse PAAC4_RF algoritma, omogućio dostupnost istog široj naučnoj javnosti. TRI_tool je implementiran u JAVA programskom jeziku. Poseban fokus bio ja na zadržavanju jednakih performansi predviđanja sa očuvanjem visoke računarske brzine PAAC4_RF pristupa. TRI_tool omogućava mapiranje proteinskih interakcija koristeći samo sekvencu željenog proteina u FASTA formatu. Veb alat je dostupan na adresi: *https://www.vin.bg.ac.rs/180/tools/tfpred.php*



Slika 9. Algoritam TRI_tool veb alata zasnovanog alata za predviđanje interakcija među transkripcionim regulatorima.

4.1.6 WT1 studija slučaja

Efikasnost TRI_tool alata, odnosno PAAC4_RF modela, je testirana u slučaju predviđanja potencijalnih interakcija transkripcionog faktora WT1 (engl. Wilm's tumor) proteina. Overekspresija WT1 proteina je detektovana u raznim ljudskim malignim oboljenjima (Huff, 2011). Kompleksna uloga WT1 je najvećim delom regulisana brojnim interakcijama sa drugim proteinima. Koristeći TRI_tool, najpre smo mapirali potencijalne interakcije WT1 sa skupom proteinskih kinaza koje učestvuju u transkripcionoj regulaciji (Tabela 10).

Tabela 10. Predikcije dobijene TRI_tool alatom: WT1 interakcije sa proteinskim kinazama koje učestvuju u regulaciji transkripcije.

Proteinske kinaze	**Predviđena interakcija (DA/NE)**	**Verovatnoća predikcije**
ABL1_HUMAN	YES	0.6733
M3K7_HUMAN	YES	0.6667
TIF1B_HUMAN	YES	0.6283
TGFR1_HUMAN	YES	0.6233
DYR1B_HUMAN	YES	0.6200
IKKA_HUMAN	YES	0.6150
CDK8_HUMAN	YES	0.6100
IRAK1_HUMAN	YES	0.5933
CDK1_HUMAN	YES	0.5883
CDK9_HUMAN	YES	0.5850
PRKDC_HUMAN	YES	0.5850
PLK1_HUMAN	YES	0.5833
RIPK1_HUMAN	YES	0.5750
KSYK_HUMAN	YES	0.5700
NLK_HUMAN	YES	0.5633
KPCZ_HUMAN	YES	0.5567
CCND1_HUMAN	YES	0.5417

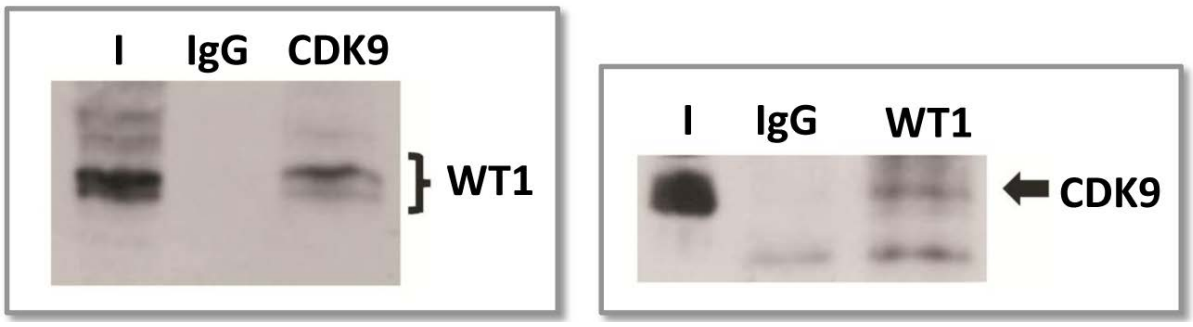
E2AK2_HUMAN	YES	0.5350
E2AK3_HUMAN	YES	0.5350
M3K2_HUMAN	YES	0.5200
CHK1_HUMAN	YES	0.5150
KS6A5_HUMAN	YES	0.5067
CD5R1_HUMAN	NO	0.4933
CD11A_HUMAN	NO	0.4917
TBK1_HUMAN	NO	0.4750
STK3_HUMAN	NO	0.4683
RN5A_HUMAN	NO	0.4583
RIPK3_HUMAN	NO	0.4517
MP2K5_HUMAN	NO	0.4183
PKN1_HUMAN	NO	0.3967
M3K10_HUMAN	NO	0.3883

Pretraživanjem literature identifikovali smo više od 40 interaktora za WT1. Među tim partnerima, interakcija između WT1 i Transkripcionog faktora II B (TFIIB), iako opisana u literaturi (McKay, Carpenter and Roberts, 1999), nije bila uključena u verziju HIPPIE baze podataka koja je korišćena za formiranje trening skupa. Uzimajući u obzir tu činjenicu, testirali smo efikasnost predviđanja WT1-TFIIB interakcije pomoću TRI_tool alatom (Tabela 11).

Tabela 11. Predviđanje interakcije WT1 sa transkripcionim regulatorom TFIIB dobijena TRI_tool alatom.

	Intrakcija	**verovatnoća(p)**
WT1-TFIIB	YES	0.52

Na osnovu predikcije TRI_tool alatom eksperimentalno je ispitivana interakcija između WT1 i ciklin-zavisne kinaze, CDK9. Interakcija WT1 i CDK9 je potvrđena metodom koimunoprecipitacije koja je izvedena u saradnji sa Profesorom Robertsom sa Univerziteta u Bristolu, Velika Britanija (Perovic *et al.*, 2017) (Slika 10).

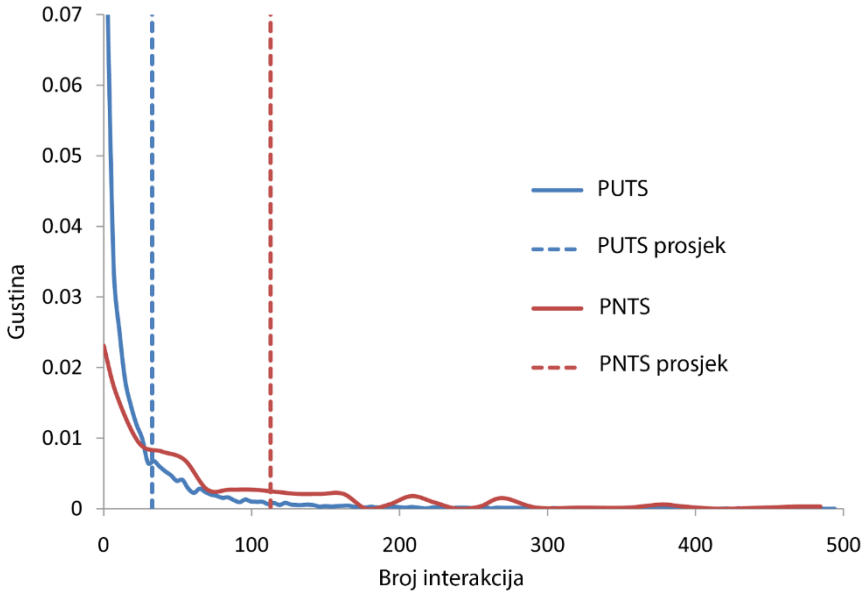


Slika 10. Koimunoprecipitacija WT1 i CDK9 proteina. Levi panel: Ekstrakt nukleusa K562 je precipitiran sa anti-CDK9 antitelima ili kontrolnim anti-IgG antitelima. Precipitati su podvrgnuti SDS-PAGE i izvršen je imunoblot sa anti-WT1 antitelima. Izoforme WT1 su obeležene vitičastom zagradom. Desni panel: Ekstrakt nukleusa K562 je precipitiran sa anti-WT1 antitelima ili kontrolnim anti-IgG antitelima . Precipitati su podvrgnuti SDS-PAGE i izvršen je imunoblot sa anti-CDK9 antitelima. CDK9 je označen strelicom (Perovic *et al.*, 2017).

4.2 Predviđanje IPP PNTS proteina

4.2.1 Skup podataka

Pretraživanjem DisProt baze podataka pronađeno je 237 ljudskih PNTS. U ovu grupu spadaju proteini čija je trodimenzionalna struktura u potpunosti neuređena i proteini koji sadrže regione sa neuređenom strukturom. Lista proteina sa kojima PNTS ostvaruju interakcije formirana je pretraživanjem HIPPIE baze podataka IPP čoveka. U toku ove faze pronađeno je 24994 IPP, u kojima je bar jedan od interaktora PNTS. Prosečan broj interakcija PNTS je 112,77, dok je prosečan broj interakcija PUTS 31,58. U proseku, PNTS se u 3.5 puta većoj meri povezuju sa drugim proteinima u odnosu na PUTS (Slika 11).

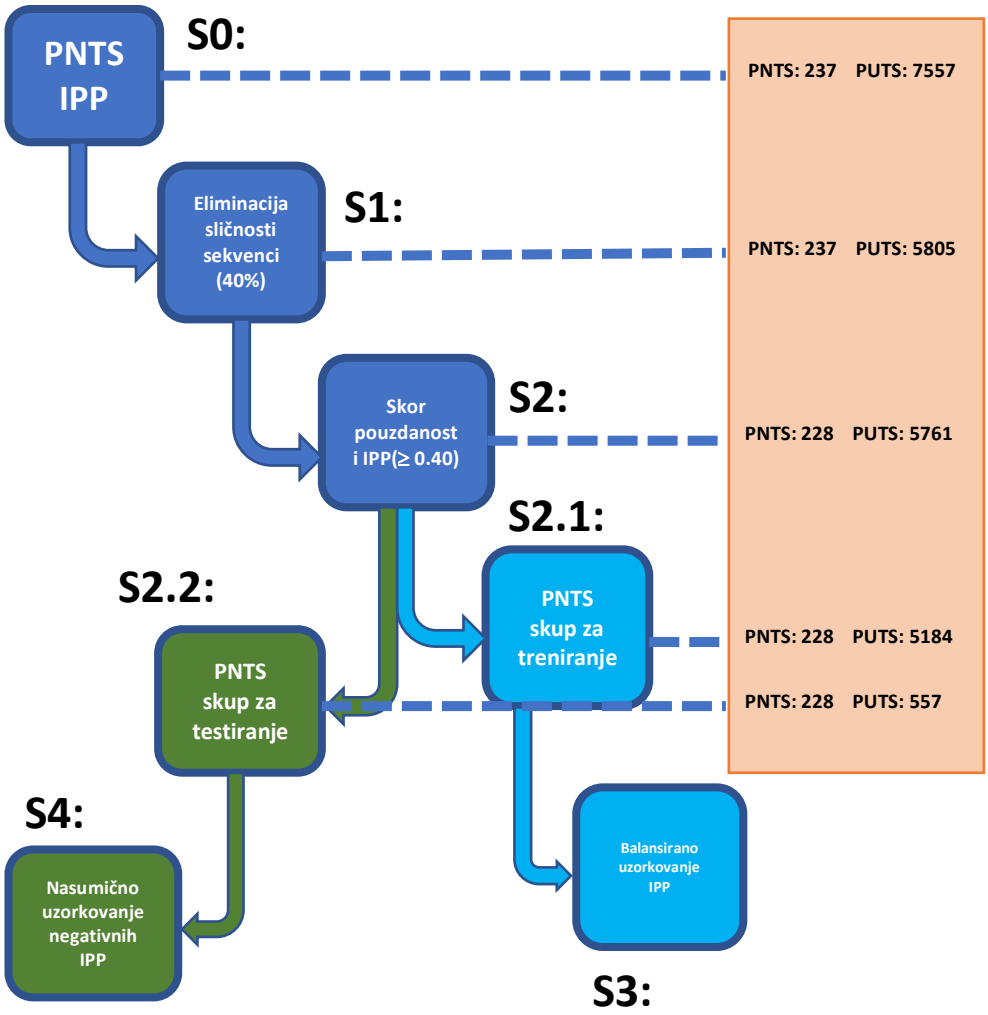


Slika 11. Krive gustine broja interaktora pronađenih u HIPPIE bazi podataka za PNTS i PUTS. PNTS su predstavljeni crvenom, a PUTS plavom krivom zajedno sa prosekom za svaku grupu. Prosečan broj interakcija za PNTS je 112,77 dok je za PUTS 31,58.

Sekvence proteina čiji naziv je sadržavao 'putative', 'potential' ili 'uncharacterized' su odstranjene (Patthy, 2016). U cilju izbegavanja preterano optimističnog predviđanja potencijalno interagujućih proteina, redundantne sekvence (Park, 2009) sa stepenom sličnosti većim od 40% su odstranjene iz skupa veličine 24994 IPP (Slika 12. skup S0). Suvišnim su se smatrale sekvence sa stepenom sličnosti većim od 40%. Sličnosti sekvence među proteinima u ciljnom skupu analizirane su koristeći CD-HIT alat za grupisanje proteinskih sekvenci prema sličnosti. Unutar inicijalnog skupa od 24994 IPP pronađeno je 7557 proteina sa redundantnim sekvencama. Eliminacijom ovih proteina i njihovih interakcija, formiran je skup od 20216 IPP (Slika 12. skup S1). Ovaj skup je pored 237 PNTS sadržavao i 5805 jedinstvenih PUTS. U sledećoj fazi eliminisane su proteinske interakcije sa manjim stepenom pouzdanosti u eksperimentalnoj anotaciji. Kao osnovu za sortiranje IPP prema stepenu pouzdanosti, upotrebljen je HIPPIE sistem evaluacije pouzdanosti IPP koji omogućava merenje nivoa pouzdanosti IPP na osnovu rangiranja raznih eksperimentalnih metoda za detekciju IPP, broja prijavljenih eksperimentalnih verifikacija po pojedinačnoj IPP i broj studija u kojima je interakcija potvrđena. IPP sa HIPPIE stepenom pouzdanosti ≥ 0.40 su posmatrane kao interakcije sa visokim stepenom sigurnosti. Finalni skup je sadržao 19835 visoko pouzdanih i

neredundantnih IPP. U ovom skupu 228 PNTS interaguju međusobno i sa 5761 PUTS (Slika 12. skup S2). Za evaluaciju razvijenog metoda za predviđanje interakcija PNTS, pet puta su nasumičnim uzorkovanjem iz skupa od 19837 potvrđenih IPP, formirani skupovi za treniranje (Slika 12. S2.1) i testiranje (Slika 12. S2.2) budućeg modela.

Uobičajen pristup formiranju skupa negativnih IPP je da se nasumično uzorkuje željeni broj interakcija iz skupa maksimalnog potencijalnog skupa IPP, iz koga su eliminisane poznate interakcije (Ben-Hur and Noble, 2006; Park and Marcotte, 2012). Maksimalni potencijalni skup sadrži svaku moguću kombinaciju proteina čije interakcije se nalaze u grupi potvrđenih IPP. Usled visokog afiniteta PUTS proteina u formiranju interakcija, neophodno je balansirano uzorkovanje negativnih IPP za formiranje skupa za učenje (Slika 12. skup S3). Skup negativnih IPP formiran je na način da se frekvencija interakcija koje su prijavljene za protein u sklopu IPP liste zadržava u skupu negativnih IPP. Na ovaj način se izbegava učenje modela neželjenim karakteristikama skupa za treniranje. Cilj je formirati model koji minimizuje predviđanje IPP samo na osnovu velike zastupljenosti proteina koji formiraju datu interakciju u skupu za treniranje (Ben-Hur and Noble, 2006; Park and Marcotte, 2011). Za formiranje skupa za testiranje, uzorkovanje NIPP je izvršeno nasumično u cilju testiranja pristupa za predviđanje IPP na nivou populacije (Slika 12. S4). Za svaki od test skupova, formirani su NIPP tehnikom nasumičnog uzorkovanja, dok su kod skupova za treniranje izabrani tehnikom balansiranog uzorkovanja. U svakom od test skupova samo jedan od interaktora, odnosno PNTS, se nalazi u trening skupu.



Slika 12. Proces generisanja skupova IPP. Pravougaonicima su predstavljene faze koje su vodile do formiranja konačnih skupova za učenje i skupova za testiranje algoritama za klasifikaciju. Levi blok (belo) prikazuje i veličinu skupa (S) IPP u sklopu faze formiranja skupa predstavljenih pravougaonicima. Sa desne strane, unutar narandžastog bloka, prikazani su brojevi PNTS i PUTS koje dati skup (S) sadrži.

4.2.2 DP_PAAC5 atributi

Upoređivanjem primarne i tercijarne strukture PNTS i PUTS, otkrivene su specifičnosti u kompoziciji aminokiselina sekvence PNTS proteina (Uversky and Dunker, 2010). Specifičnosti PNTS su posebno izražene u načinu implementacije hemijskih i fizičkih principa interakcije proteina. Posebnost karakteristika vezivanja PNTS proteina u tranzijentne komplekse ogleda se u sastavu aminokiselinskih ostataka, veličini i

strukturnim karakteristikama interagujućeg interfejsa, kao i strukturnim karakteristikama partnera za interakciju (Mészáros *et al.*, 2007). Jedan od ciljeva analize je bio, pronalaženje adekvatnog matematičkog modela sekvenci PNTS, koji bi omogućio efikasniju reprezentaciju specifičnosti PNTS u formiranju IPP i tačnijeg predviđanja novih IPP. Strukturne karakteristike sekvence predstavljene su sa dva matematička modela.

Prvo, sekvenca PUTS je predstavljena normalizovanom 2-mer reprezentacijom, pri čemu je meren broj pojavljivanja svakog dipeptida, gde je dipeptid podsekvenca proteina sastavljena od dve uzastopne aminokiseline. Svaka proteinska sekvenca je na ovaj način tako predstavljena numeričkom vektorom sa 400 elemenata (DP), što je ukupan broj svih mogućih dipeptida.

Drugo, PAAC je modifikovana da pored AAC očuva strukturnu informaciju sadržanu u redosledu aminokiseline unutar proteinske sekvence, specifičnu za PNTS. Za karakterisanje pseudo-aminokiselinske kompozicije korišćeno je pet lestvica aminokiselina (Tabela 3).

1. Poredak aminokiselina prema njihovoj sklonosti ka uređenosti ili neuređenosti tercijarne strukture proteina optimizovan je u formi TOP-IDP lestvice. Aminokiseline su poređane prema vrednostima TOP-IDP skale i podeljeni prema uticaju na uređenost proteina na aminokiseline koje promovišu uređenost i aminokiseline koje promovišu neuređenost. Rangiranje je formirano na osnovu analize 517 objavljenih atributa iz AAindex baze podataka (Campen *et al.*, 2008).
2. Lestvica DisProt je zasnovana na statističkoj razlici u kompoziciji aminokiselina kod PNTS i PUTS.
3. Analizom strukturne fleksibilnosti proteina formirana je lestvica aminokiselina, B -vrednost. Za svaku aminokiselinu analizirana je fleksibilnost aminokiselinskih ostataka u odnosu na dva nefleksibilna suseda. Aminokiselinski ostaci su karakterizovane i poređane u skladu sa prosečnom fleksibilnošću za čitav protein (Vihinen, Torkkila and Riikonen, 1994).
4. Lestvica FoldUnfold je nastala kao rezultat merenja srednje gustine pakovanja (engl. Mean packing density). Kriterijum kvantifikacije je broj parnih kontakata koji aminokiselinske ostatke ostvaruju sa susedima. FoldUnfold lestvica sadrži

aminokiseline rangiranih prema srednjoj gustini pakovanja (Galzitskaya, Garbuzynskiy and Lobanov, 2006).

5. Neto naelektrisanje (engl. Net charge) (Klein, Kanehisa and DeLisi, 1984)

Na osnovu PAAC, razvijen je PAAC5 model koji, pored AAC, sadrži strukturnu informaciju sekvence PNTS. Svaka od pet lestvica aminokiselina PNTS, važnih za njihovo neuređeno stanje, posmatrana je unutar klizećeg prozora dužine 50 aminokiselina. Unutar ovog prozora, vektor atributa za svaku od pet karakteristika je formiran primenom funkcija korelacije, u skladu sa standardnim PAAC modelom. Veličina prozora je određena minimalnom dužinom sekvence u sklopu skupa sekvenci od 50 AA.

Na ovaj način, svaka proteinska sekvenca je modelirana sa dva posebna vektora: DP veličine 400 i PAAC5 veličine 70, koji spojeni predstavljaju DP_AAC5 atribute proteina. Interakcija dva proteina je predstavljena spajanjem vektora pojedinačnih proteina, tako da je sklopu ove analize svaka IPP je karakterisana 940-dimenzionim vektorom.

4.2.3 DP_PAAC5_RF model

U cilju određivanja najadekvatnijeg algoritma ML za problem predviđanja PNTS IPP, poredili smo pet algoritama ML. Algoritmi su poređeni koristeći CV-10 validaciju na skupu od 35830 IPP koristeći DP_PAAC5 model sekvenci. Efikasnost razdvajanja IPP od NIPP PUTS testirana je modelima formiranim sledećim algoritmima: RF, SVM, GLM i GBM (Tabela 12). Za svaki algoritam odgovarajući skup meta-parametara podešen je pomoću nasumične pretrage. Algoritam koji je dao najveći procenat tačnih predviđanja na osnovu DP_PAAC5 modela sekvenci, bio je zasnovan na algoritmu RF. Novi predikcioni model formiran na osnovu DP_PAAC5 i algoritama RF na našem PNTS skupu označen je kao DP_PAAC5_RF.

Tabela 12. Poređenje RF, GLM, GBM, SVM algoritama na S3 skupu PNTS koristeći CV-10 kao šemu validacije. IPP su kodirane DP_PAAC5 modelom. Srednja vrednosti i standardna devijacija su prikazane za AUROC, AUPRC, tačnost, preciznost, senzitivnost, specifičnost, F1 i MCC.

	RF	**GLM**	**GBM**	**SVM**
AUROC	0.827 ± 0.004	0.697 ± 0.005	0.811 ± 0.004	0.728 ± 0.010
AUPR	0.841 ± 0.003	0.697 ± 0.007	0.825 ± 0.003	0.741 ± 0.009
Tačnost	0.758 ± 0.003	0.644 ± 0.003	0.742 ± 0.003	0.673 ± 0.007
F1	0.747 ± 0.003	0.645 ± 0.003	0.733 ± 0.003	0.658 ± 0.007
Preciznost	0.783 ± 0.002	0.644 ± 0.004	0.758 ± 0.003	0.690 ± 0.007
Specifičnost	0.803 ± 0.002	0.642 ± 0.006	0.774 ± 0.004	0.718 ± 0.006
Senzitivnost	0.713 ± 0.004	0.647 ± 0.003	0.709 ± 0.004	0.628 ± 0.008
Stopa lazno poz.	0.197 ± 0.002	0.358 ± 0.006	0.226 ± 0.004	0.282 ± 0.006
MCC	0.518 ± 0.005	0.289 ± 0.006	0.484 ± 0.006	0.348 ± 0.013

4.2.4 Evaluacija DP_PAAC5_RF modela

4.2.4.1 Efikasnost DP_PAAC5_RF modela u predikciji novih IPP i poređenje sa standardnim metodama za predviđanje IPP

Da bi se testiralo koliko efikasno DP_PAAC5_RF model predviđa izvesnost interakcije između PNTS proteina i novog proteina koji nije prisutan u skupu za treniranje, DP_PAAC5_RF model je evaluiran na pet test skupova C2 tipa. Koristeći iste skupove za treniranje kao i iste C2 skupovima za testiranje, DP_PAAC5_RF je upoređen sa standardno upotrebljivanim metodama za predviđanje IPP zasnovanim na sekvenci proteina: Martin_M (Martin, Roe and Faulon, 2005) , Guo_M (Guo *et al.*, 2008), Shen_M (Shen *et al.*, 2007) (Tabela 13).

DP_PAAC5_RF modeli su demonstrirali visok stepen pouzdanosti u predviđanju novih IPP PNTS sa drugim proteinima, i u poređenju sa testiranim, standardno korišćenim metodama za predviđanje IPP, pokazali veću efikasnost predviđanja.

Tabela 13. Poređenje DP_PAAC5_RF modela u odnosu na standardne metode za predviđanje IPP zasnovane na sekvenci u poređenju novih interakcija između PNTS i proteina nepoznatog trening skupu. Srednja vrednosti i standardna devijacija su prikazane za AUROC, AUPRC, F1 i MCC.

Metod	**AUROC**	**AUPRC**	**Tačnost**	**F1**	**MCC**
DP_PAAC5_RF	0.746 ± 0.017	0.734 ± 0.020	0.670 ± 0.015	0.633 ± 0.021	0.348 ± 0.028
Martin_M	0.688 ± 0.017	0.697 ± 0.018	0.638 ± 0.013	0.590 ± 0.022	0.285 ± 0.025
Guo_M	0.637 ± 0.014	0.613 ± 0.012	0.593 ± 0.010	0.553 ± 0.019	0.190 ± 0.021
Shen_M	0.627 ± 0.011	0.643 ± 0.014	0.599 ± 0.008	0.518 ± 0.013	0.211 ± 0.017

4.2.4.2 Efikasnost DP_PAAC5_RF modela u predikciji novih IPP u slučaju disbalansa NIPP u odnosu na IPP test skupova

U biološkom sistemu broj primera odsustva interakcije među proteinima daleko premašuje broj stvarnih interakcija, u odnosu na maksimalni broj interakcija. Iz tog razloga važno je ispitati predikcionu sposobnost metoda za predviđanje IPP u slučajevima disbalansa u broju NIPP, u odnosu na potvrđene IPP. DP_PAAC5_RF model je dodatno evaluiran na novim test skupovima gde je broj pozitivnih IPP 10 i 100 puta manji od broja NIPP (Tabela 14).

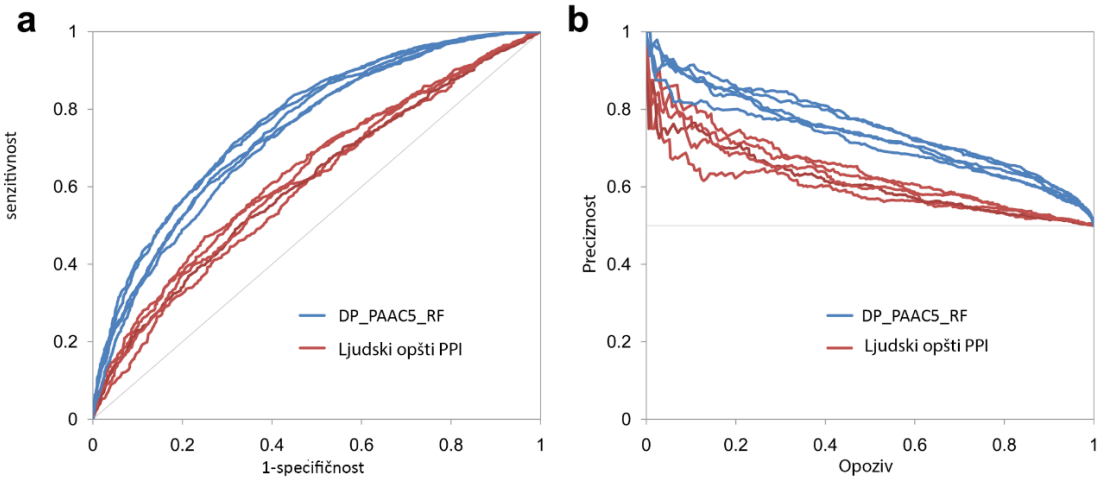
Tabela 14. Poređenje DP_PAAC5_RF modela u odnosu na standardne metode za predviđanje IPP zasnovane na sekvenci na test skupovima sa 10 puta (10N) i 100 puta (100N) većim brojem NIPP u odnosu na IPP. Korišćene su AUROC, AUPRC i Tačnost kao mere performansi klasifikatora.

Metod	**10N**			**100N**		
	AUROC	AUPRC	Tačnost	AUC	AUPRC	Tačnost
DP_PAAC5_RF	0.745	0.237	0.740	0.748	0.050	0.757
Martin_M	0.691	0.217	0.724	0.692	0.048	0.737
Guo_M	0.645	0.140	0.648	0.646	0.025	0.657
Shen_M	0.624	0.163	0.740	0.624	0.032	0.763

DP_PAAC5_RF model je demonstrirao bolje predikcione performanse u odnosu na standardno korišćene metode kod nebalansiranih test skupova (10N i 100N). Značajno povećanje broja negativnih IPP od 10 i 100 puta ne utiče značajno na promenu AUROC u odnosu na balansirani test skup. Značajno pogoršanje performansi je detektovano pomoću AUPRC mere koja je znatno osetljivija na ukupan broj lažno pozitivnih rezultata (Park and Marcotte, 2011).

4.2.4.3 Poređenje DP_PAAC5_RF i modela formiranih sa opštim ljudskim IPP

Interakcije koje PNTS proteini sadrže specifičan način primene opštih principa IPP (Mészáros *et al.*, 2007). Postavlja se pitanje da li je potrebno generisati modele fino naštimovane za razdvajanje pozitivnih od negativnih IPP, u kojima učestvuju PNTS, ili su dovoljni opšti modeli za predviđanje ljudskih IPP. Da bismo demonstrirali neophodnost da se specifičnosti na molekularnom nivou IPP u kojima PNTS učestvuju, prenesu na nivo statističkog modeliranja IPP, poredili smo efikasnost: (i) DP_PAAC5_RF modela formiranih na pet trening skupova IPP čoveka (Park and Marcotte, 2012) formiranih na osnovu potvrđenih interakcija iz PINA baze podataka i (ii) DP_PAAC5_RF modela formiranim na specifično pripremljenim primerima IPP sa PNTS za treniranje evaluiranim na odgovarajućim test skupovima (Slika 13).



Slika 13. Poređenje pet DP_PAAC5_RF modela formiranih na posebno pripremljenim PNTS skupovima za trening (plavo), sa modelima formiranim na skupovima za treniranje modela za predviđanje opštih ljudskih IPP prema AUROC (a) i AUPRC (b) merama efikasnosti klasifikatora.

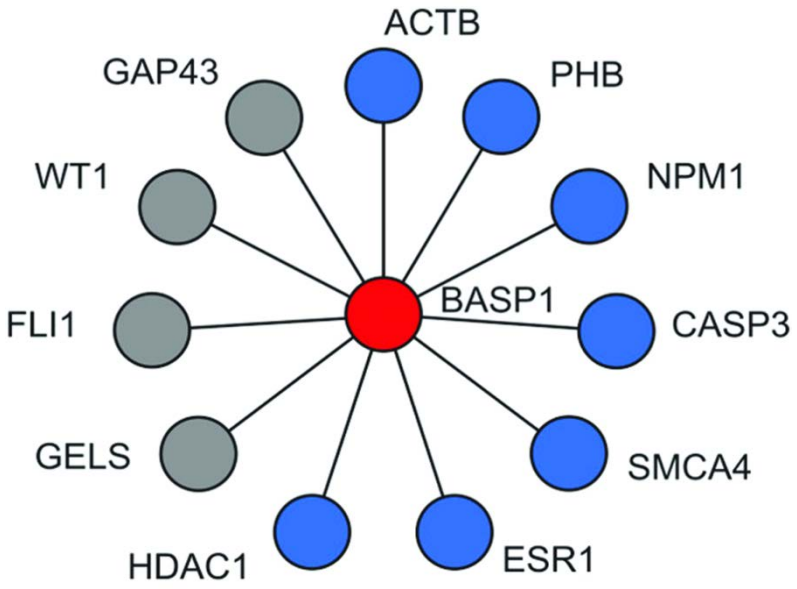
Poređenje ova dva pristupa pokazuje poboljšanje efikasnosti predviđanja specifičnog modela treniranog sa, i testiranom na, specifičnim NTPS interakcijama u odnosu na model zasnovanom na IPP iz čitavog proteoma čoveka.

4.2.5 IDPpi_tool veb alat

DP_PAAC5_RF algoritam učinjen je dostupnim u formi veb alata (IDPpi_tool), dostupnog široj naučnoj javnosti. Alat omogućava testiranje potencijalnih interakcija ljudskih PNTS iz DisProt baze podataka sa proteinima od interesa na osnovu njihove sekvence. Korisnik može izabrati željeni PNTS protein iz padajućeg menija na levoj strani korisničkog okruženja. Sa desne strane, korisnik ima mogućnost da unese do 100 potencijalnih interaktora u FASTA formatu. Izabrani PNTS protein je dinamički povezan sa njegovom referencom u UniProtKB/Swiss-Prot i DisProt bazi podataka. Na ovaj način, detaljna informacija o regionima, varijantama i vezujućim domenima izabranog PNTS proteina je lako dostupna korisniku. Veb alat je dostupan na adresi: http:// www.vin.bg.ac.rs/180/tools/dispred

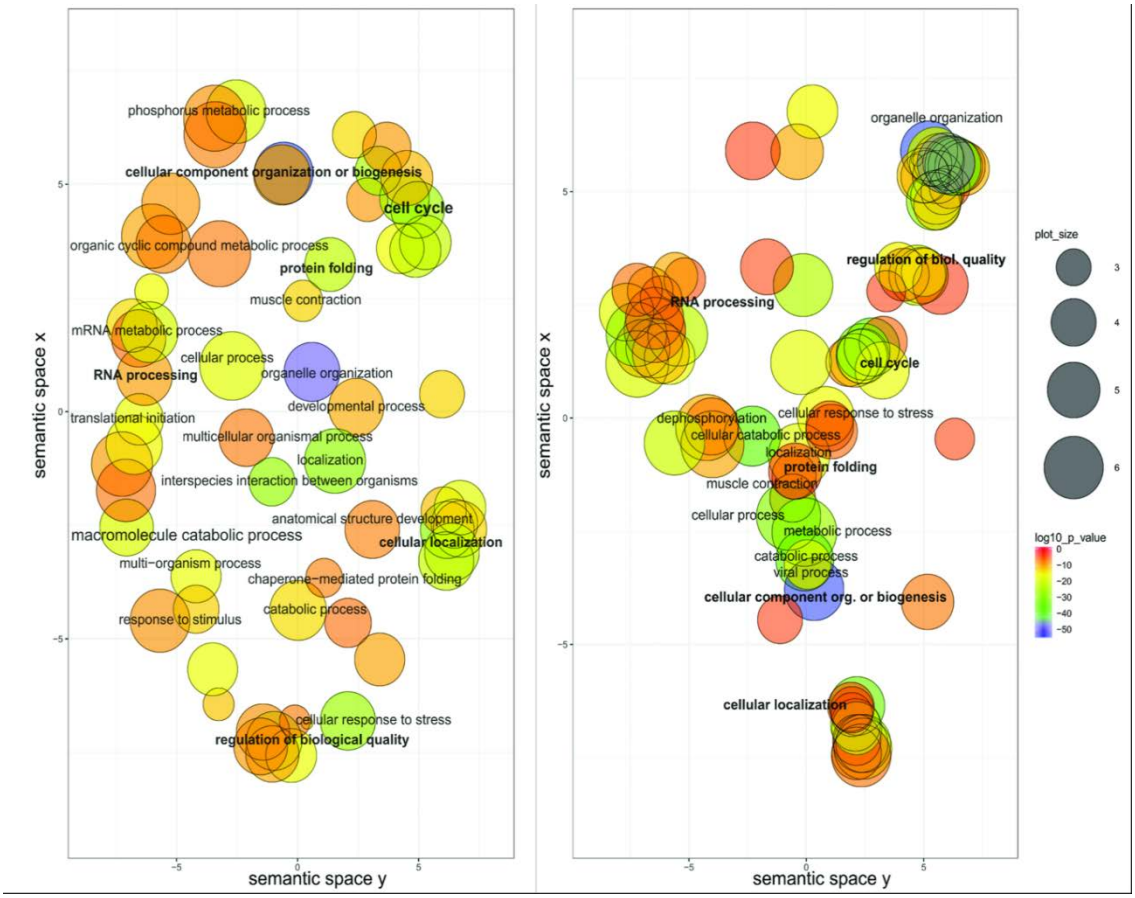
4.2.6 Studija slučaja BASP1 transkripcionog koregulatora

Protein BASP1 (engl. Brain Acid-Soluble Protein-1) je značajan transkripcioni koregulator (Forsova and Zakharov, 2016) specifičan po strukturnoj neuređenosti čitave sekvence (Toska and Roberts, 2014). Uloga BASP1 u ćelijskim procesima zdravih i obolelih osoba nije do kraja razjašnjena. Otkrivanje do sada nepoznatih interaktora BASP1 bi omogućilo dodatan uvid u njegovu složenu funkciju u ćeliji. U prvom koraku detektovano je 11 interaktora BASP1 prijavljenih u stručnoj literaturi, čije interakcije nisu korišćene za formiranje predikcionog modela u sklopu IDPpi_tool alata. IDPpi_tool je tačno predvideo 7 od 11 potvrđenih interakcija među kojima su: HDAC1 (engl. Histone Deacetylase 1), ACTB (engl. Actin Beta), NPM (engl. Nucleophosmin 1) i CASP3 (engl. Caspase 3) (Slika 14).



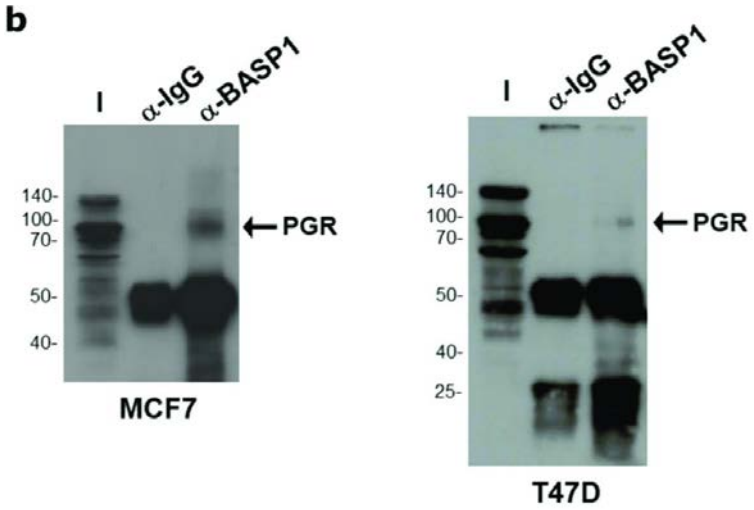
Slika 14. Predviđanje poznatih interaktora IDPpi_tool alatom u studiji slučaja BASP1 proteina. Potvrđene interakcije BASP1 koje su tačno identifikovane IDPpi_tool alatom su HDAC1, ACTB, PHB, CASP3, SMCA4, ESR1 i NPM1 (označene plavom bojom). Sa druge strane, interakcije BASP1 i WT1, GELS, FLI1 i GAP43 proteina predviđene su kao negativne.

U cilju funkcionalne analize BASP1 proteina korišćenjem informacija iz predviđenog proteinskog BASP1 interaktoma, izvršeno je obogaćivanje GO termina (engl. Enrichment Analysis) na osnovu BASP1 interaktora predviđenih IDPpi_tool alatom. Za proces obogaćivanja korišćen je BINGO alat, dok je vizuelna prezentacija obavljena REVIGO alatom. Lista obogaćenih GO termina BASP1 interaktora prethodno poznatih, poređena je sa listom obogaćenih GO termina na osnovu novih interaktora predviđenih IDPpi_tool alatom. Analizirani GO termini pripadaju pod-ontologiji 'biološki procesi' (BP). Vizuelni prikaz termina prikazan je na Slici 15. GO termini su prikazani u formi čvorova čija je međusobna udaljenost direktno povezana sa sličnosti termina u semantičkom prostoru. Stepen obogaćivanja termina je prikazan određenom bojom, čija je legenda prikazana na desnoj strani slike.



Slika 15. Analiza obogaćivanja GO termina koji pripadaju grupi 'biološki proces' u skupu prethodno utvrđenih BASP1 (levo) i predviđenih interaktora (desno). Značajno obogaćeni GO termini veće semantičke sličnosti su prikazani bliži jedni drugima u grafičkom prikazu. Kružni markeri su skalirani i obojeni prema log10 p-vrednosti značajnosti termina. Plavi krugovi su značajniji od crvenih.

U cilju daljnjeg testiranja prediktivne efikasnosti IDPpi_tool eksperimentalno je potvrđena , tehnikom koimunoprecipitacije, IDPpi_tool predviđena interakcija između BASP1 i PRGR (engl. Progesterone Receptor) u saradnji sa Profesorom Robertsom sa Univerziteta u Bristolu, Velika Britanija (Perovic *et al.*, 2018).



Slika 16. Koimunoprecipitacija BASP1 i PRGR proteina. Ekstrakt nukleusa MCF7 ćelija (levi panel) ili T47D ćelija kancera dojke (desni panel) je precipitiran sa anti-BASP1 antitelima ili kontrolnim anti-IgG antitelima. Precipitati su podvrgnuti SDS-PAGE i izvršen je imunoblot sa anti-PGR antitelima. Markeri molekularne težine (kDa) su prikazani skroz levo na oba panela (Perovic *et al*., 2018).

4.3 Predviđanje IPP na nivou proteoma kod čoveka

4.3.1 Skupovi podataka ljudskih IPP

Informacije o ljudskim IPP preuzete su iz HIPPIE baze podataka verzija 2.1 (07/2017). Pronađeno je 15804 ljudskih proteina koji formiraju 299247 IPP. Redundantne interakcije i interakcije između proteina čoveka i proteina drugih vrsta nisu uključene u ovaj skup. Proteinske sekvence su preuzete iz UniProtKB/Swiss-Prot baze podataka. Iz skupa su eliminisani proteini čije sekvence imaju manje od 50 aminokiselina. Zatim su proteini sa međusobnom sličnošću sekvence većom od 40 % detektovani pomoću CD-HIT alata. Redundantni proteini su eliminisani iz skupa zajedno sa IPP koje formiraju. Finalni skup je uključivao 196071 IPP. Ista procedura je primenjena na novu verziju HIPPIE baze podataka (v2.2), objavljenu nakon generisanja finalnih modela za predviđanje IPP. U sklopu HIPPIE v2.2 baze pronađeno je 47892 IPP koje nisu prisutne u HIPPIE v2.1, i na osnovu njih formiran je skup razlike dve verzije HIPPIE baze podataka, označen kao HIPPIE_new skup.

Da bi se izabrao skup za formiranje referentnog modela za predviđanje ljudskih interakcija na velikoj skali, ispitivan je uticaj pouzdanosti IPP, predstavljen u vidu mere

pouzdanosti IPP (engl. confidence score) iz HIPPIE baze podataka. Izdvojeno je pet skupova ljudskih IPP gde svaki skup sadrži IPP koje imaju jednaku ili veću meru pouzdanosti HIPPIE u odnosu na definisani prag.

1. C80 sadrži 14621 IPP sa merom pouzdanosti ≥ 0.80
2. C70 sadrži 56749 IPP sa merom pouzdanosti ≥ 0.70
3. C60 sadrži 161558 sa merom pouzdanosti ≥ 0.60
4. C50 sadrži 172962 IPP sa merom pouzdanosti ≥ 0.50
5. C00 sadrži 196071 IPP sa merom pouzdanosti ≥ 0.0

Skup C00 sadrži sve IPP bez obzira na vrednost mere pouzdanosti HIPPIE baze podataka. Gore opisani skupovi su korišćeni kao pozitivni deo skupova za formiranje i testiranje budućih modela ML za predviđanje humanih IPP. Skupovi NIPP formirani su nasumičnim uzorkovanjem proteinskih parova čije interakcije nisu potvrđene u celokupnoj HIPPIE bazi podataka (Park and Marcotte, 2011). Finalni skupovi su sadržali jednak broj pozitivnih i negativnih primera, odnosno proteinskih parova za koje postoji mala verovatnoća da interaguju. Koristeći isti pristup generisanja negativnih podskupova, formiran je HIPPIE_new' skup.

4.3.1.1 Nezavisnost performansi modela od različitih skupova NIPP

Sposobnost modela da efikasno razdvaja ciljne parove proteina prema prisustvu ili odsustvu interakcije, značajno zavisi od skupa na kome se uči (Domingos, 2012). Budući da se negativna polovina skupa za formiranje predikcionog modela formira na osnovu stohastičkog procesa, neophodno je ispitati uticaj načina generisanja NIPP na performanse prediktivnog modela. U tu svrhu korišćen je C80 skup IPP. Pet puta je ponovljeno različito nasumično uzorkovanje 14621 parova proteina čije interakcije su posmatrane kao NIPP. Spajanjem sa C80 skupom, dobijeno je pet C80` skupova koji su ukupno sadržavali 29242 IPP. Proteini su kodirani na bazi PCA_ACC modela, a za algoritam ML je izabran GBM. Šema validacije CV-10, je primenjena za podelu skupova na podskupove za testiranje i treniranje. Srednje vrednosti i standardne devijacija vrednosti sedam mera performansi klasifikatora predstavljene su na Tabeli 15.

Standardna devijacija AUROC mere performansi modela zasnovanih na istom uzorku negativnih je veća od standardne devijacije (0.002) između srednjih vrednosti AUROC mere između različitih uzoraka NIPP. To ukazuje potencijal replikativnosti formiranja modela za predviđanje IPP zasnovanih na nasumičnom uzorkovanju NIPP kakvo je provedeno u sklopu ove studije.

Tabela 15. Poređenje uticaja nasumičnog uzorkovanja negativnog dela skupa za učenje na performanse formiranog GBM modela mašinskog učenja. Pet C80' skupova IPP je formirano od C80 pozitivnog skupa i jednakog broja nasumično uzorkovanih NIPP. Uzorkovanje je ponovljeno za svaki od C80'. Kao šema evaluacije upotrebljena je CV-10 validacija. Srednja vrednost (sv) i standardna devijacija (sd) vrednosti sedam mera evaluacije performansi prediktora su prikazane.

	C80′$_1$		**C80′$_2$**		**C80′$_3$**		**C80′$_4$**		**C80′$_5$**	
	sv	sd	sv	sd	sv	sd	sv	sd	sv	sd
Tačnost	0.780	0.004	0.778	0.006	0.780	0.004	0.781	0.004	0.775	0.004
AUROC	0.856	0.003	0.856	0.005	0.858	0.004	0.861	0.003	0.856	0.003
F1	0.786	0.003	0.785	0.006	0.785	0.004	0.789	0.004	0.783	0.003
MCC	0.561	0.008	0.557	0.011	0.560	0.008	0.564	0.008	0.553	0.008
Preci.	0.766	0.010	0.760	0.010	0.766	0.009	0.761	0.011	0.757	0.010
Odziv	0.807	0.010	0.812	0.013	0.805	0.009	0.821	0.014	0.812	0.009
Specif.	0.753	0.016	0.744	0.017	0.754	0.013	0.741	0.019	0.739	0.016

4.3.1.2 Formiranje visokokvalitetnog skupa IPP

Performanse klasifikacionog modela mašinskog učenja značajno zavise od veličine skupa za treniranje i kvaliteta primera koji sačinjavaju trening skup (Figueroa *et al.*, 2012). Razvijeni su različiti pristupi za evaluaciju kvaliteta eksperimentalno potvrđenih IPP (Rao *et al.*, 2014). U sklopu HIPPIE baze podataka razvijen je sistem analize i kvantifikacije pouzdanosti detektovanih IPP interakcija. Kao rezultat analiza, svakoj IPP anotirane u HIPPIE bazi podataka se dodeljuje vrednost mere pouzdanosti (Alanis-Lobato, Andrade-Navarro and Schaefer, 2017). Kvalitet IPP predstavljen vrednostima mere pouzdanosti HIPPIE, utiče na sposobnost modela treniranog na ovakvim podacima

da detektuje nove IPP (Hamp and Rost, 2013). Za formiranje efikasnog finalnog modela potrebno je ispitati uticaj broja i pouzdanosti IPP koje sačinjavaju skup za treniranje. Odabir optimalnog skupa za treniranje izvršen je prema rezultatima iz Tabele 16. Pet skupova zasnovanih na vrednostima HIPPIE skora pouzdanosti IPP služili su kao osnov za evaluaciju. Za svaki od skupova IPP C00, C50, C60, C70, C80 formirani su odgovarajući NIPP, nasumičnim uzorkovanjem, i njihovim uparivanjem formirani su skupovi IPP C00', C50', C60', C70', C80'. Interakcije ovako formiranih skupova su predstavljene PCA_AAC modelom (poglavlje 4.3.2.1), dok je za algoritam mašinskog učenja izabran GBM, a CV-10 je korišćena kao šema validacije. Skup C50', sa najvećom vrednosti AUROC, je izabran za dalje analize.

Tabela 16. Rezultati CV-10 na pet IPP skupova iz HIPPIE baze podataka formiranih prema vrednostima mere pouzdanosti HIPPIE. Za predstavljanje IPP, korišćen je AAC_PCA model predstavljanja sekvenci. Za metod mašinskog učenja korišćen je GBM. Predstavljene su srednje vrednosti (sv) i standardna devijacija (sd) za vrednosti sedam mera performansi prediktora.

	C00'		**C50'**		**C60'**		**C70',**		**C80'**	
	sv	**sd**	**sv**	**sd**	**sv**	**sd**	**sv**	**sd**	**sv**	**sd**
Tačnost	0.8191	0.001	0.8189	0.001	0.8154	0.002	0.7992	0.002	0.7792	0.005
AUROC	0.9029	0.001	0.9033	0.001	0.9007	0.001	0.8813	0.002	0.8577	0.003
F1	0.8277	0.001	0.8271	0.001	0.8247	0.001	0.8078	0.002	0.7862	0.003
MCC	0.6415	0.002	0.6407	0.002	0.6346	0.003	0.6010	0.004	0.5598	0.008
Preciz.	0.7903	0.003	0.7910	0.003	0.7854	0.005	0.7746	0.006	0.7623	0.010
Odziv	0.8689	0.004	0.8667	0.003	0.8683	0.006	0.8442	0.008	0.8122	0.011
Spec.	0.7693	0.006	0.7710	0.005	0.7625	0.009	0.7542	0.010	0.7459	0.019

Testiranje na skupu IPP sa eksperimentalno potvrđenim NIPP

Diskriminacija IPP prema vrednosti mere pouzdanosti HIPPIE baze podataka utiče na veličinu selektovanih skupova. Selekcijom IPP sa većom vrednosti mere pouzdanosti smanjuje se broj izabranih IPP u skupu. Skupovi koji uključuju širi dijapazon vrednosti mere pouzdanosti su veći. Koristeći CV-10 šemu razdvajanja na trening i test skup, trening skup se uzima kao fiksni razlomak ukupnog skupa (9/10). Da bi se predikciona sposobnost modela pripisala preciznom izboru IPP trening skupa prema vrednosti mere

pouzdanosti, a ne veličini skupa, potrebno je testirati modele, inicijalno formirane na različitim skupovima za treniranje, na istom test skupu. U svrhu formiranje posebnog test skupa, iz Negatome 2.0 baze podataka izdvojeno je 5004 eksperimentalno potvrđenih NIPP. Verifikovane IPP (5965) preuzete su iz IntAct baze podataka. Budući da Negatome baza podataka sadrži relativno mali broj NIPP, IntAct baza je pretraživana prema proteinima koji formiraju skup od 5004 NIPP. Iz IntAct baze podataka su izdvojene samo interakcije čiji proteini formiraju interakcije u negativnom skupu. Kombinujući pozitivne primere iz IntAct baze podataka i negativne iz Negatome baze podataka formiran je IntNeg test skup. U odnosu na skupove IPP koji su korišćeni za treniranje modela, IntNeg predstavlja mešavinu C2 i C3 tipa IPP. Proteinske sekvence su predstavljene PCA_AAC reprezentacijom. Predikcioni modeli su formirani na C00',C50',C60', C70',C80' skupovima za učenje koristeći GBM algoritam ML i testirani na IntNeg test skupu (Tabela 17). Poređenje performansi modela formiranih na C00',C50',C60', C70',C80' skupovima za učenje i testiranih na IntNeg test skupu, pokazuje da, kao i u slučaju analize CV-10 validacijom (Tabela 16), skup C50' najoptimalniji za generisanje prediktivnog modela.

Tabela 17. Poređenje GBM modela, treniranih na C00',C50',C60', C70',C80' skupovima koristeći PCA_AAC reprezentaciju IPP, prema sposobnosti predviđanja eksperimentalno potvrđenog skupa IntNeg. Test set formiran kombinacijom 5965 potvrđenih IPP iz IntAct i 5004 potvrđenih negativnih IPP iz Negatome 2.0. Kao mere performansi su korišćeni AUROC, AUPRC i Tačnost.

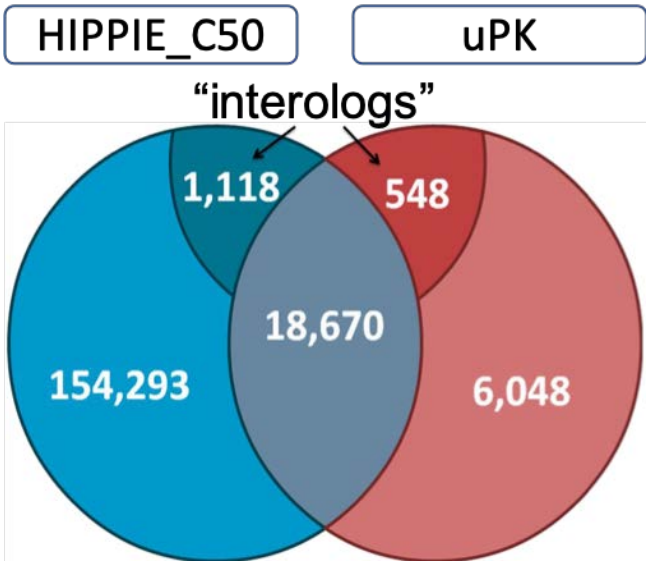
	AUROC	**AUPRC**	**TAČNOST**
C00'	0.775	0.816	0.684
C50'	0.788	0.829	0.695
C60'	0.765	0.814	0.690
C70'	0.750	0.799	0.689
C80'	0.735	0.788	0.681

Poređenje C50 skupa sa referentnim skupom IPP

Pored prethodno izabranih skupova zasnovanih na HIPPIE bazi podataka, za komparativnu analizu korišćeni su i skupovi IPP (Park and Marcotte, 2012) formirani na osnovu potvrđenih interakcija iz PINA baze podataka (Wu *et al.*, 2009) i nasumično generisanih negativnih primera. Prethodno izabrani pozitivni C50 je upoređen sa pozitivnim IPP iz unije *Park i Marcotte* skupova (uPK) radi utvrđivanje međusobnog nivoa sličnosti. Skup uPK je generisan spajanjem svih eksperimentalno potvrđenih IPP iz 40 skupova IPP za treniranje i testiranje modela mašinskog učenja, prethodno opisanih (Park and Marcotte, 2012). Pored poređenja ukupnog broja proteina i opisanih IPP (Tabela 18), utvrđen je broj zajedničkih IPP, interakcija i „interologa" između dva skupa (Slika 12). U ovom kontekstu „interolog" podrazumevamo IPP C50 skupa čiji interaktori su homolozi sa proteinima referentnog skupa koji takođe međusobno interaguju. Analiza pokazuje da je 11,44% proteina iz C50 skupa identično ili homologno sa proteinima iz uPK skupa, dok je 75,53% proteina iz uPK skupa identično ili homologno sa proteinima iz C50 skupa. Na osnovu rezultata analize za izgradnju finalnog visokokvalitetnog modela izabran je C50' skup usled većeg ukupnog broja IPP i proteina koje obuhvata, odnosno veće pokrivenosti, u odnosu na referentni uPK skup.

Tabela 18. Poređenje C50' i uPK skupova prema ukupnom broju proteina i IPP.

	uPK	C50
IPP	24718	172963
Proteini	7033	10984



Slika 17. Sličnosti i presek između C50' i uPK skupa.

4.3.2 Atributi

4.3.2.1 PCA_AAC atributi zasnovani na primarnoj strukturi proteina

Za efikasno predstavljanje sekvence proteina u formi pogodnoj za statističko modeliranje, poželjno je apstrahovati informacije o sastavu i linearnoj strukturi sekvenci (Chou, 2001, 2009). Fizičko-hemijske i druge karakteristike aminokiselina čuvaju informaciju o potencijalu za interakciju među proteinima (Guo *et al.*, 2008).

Novi matematički model za numeričku reprezentaciju proteinskih sekvenci, PCA_AAC, se sastoji iz dve komponente: kompozicije aminokiselina (AAC, jednačina 1) i vektora auto kros-kovarijanse glavnih komponenti duž proteinske sekvence.

Analiza glavnih komponenti primenjena je na 532 karakteristika aminokiselina opisanih u AAindex bazi podataka. Na ovaj način formirani su novi sintetički atributi (Tabela 19), koji su linearne kombinacije izvornih 532 karakteristika iz AAindex baze podataka. Ukupan broj ulaznih karakteristika je kompresovan u manji broj novih varijabli, koje nisu međusobno linearno povezane i redundantne, ali sadrže visok stepen informativnosti početnog skupa karakteristika aminokiselina. Ovako dobijene sintetičke varijable sadrže esenciju varijabilnosti ulaznog skupa i njihove međusobne relacije se

posmatraju dužinom čitave proteinske sekvence preko jednačine auto kros-kovarijanse (ACC, jednačina 2), gde je n broj prvih PCA komponenti, a parametar m dužina prozora. Spajanjem vrednosti atributa AAC i vrednosti dobijenih računanjem ACC funkcije na novim varijablama dobijenim PCA, formira se numerički vektor dužine $40 + 2mn^2$, koji reprezentuje aminokiselinsku sekvencu proteina. Proteinski par koji interaguje predstavljen je spajanjem numeričkih nizova dva proteina iz para.

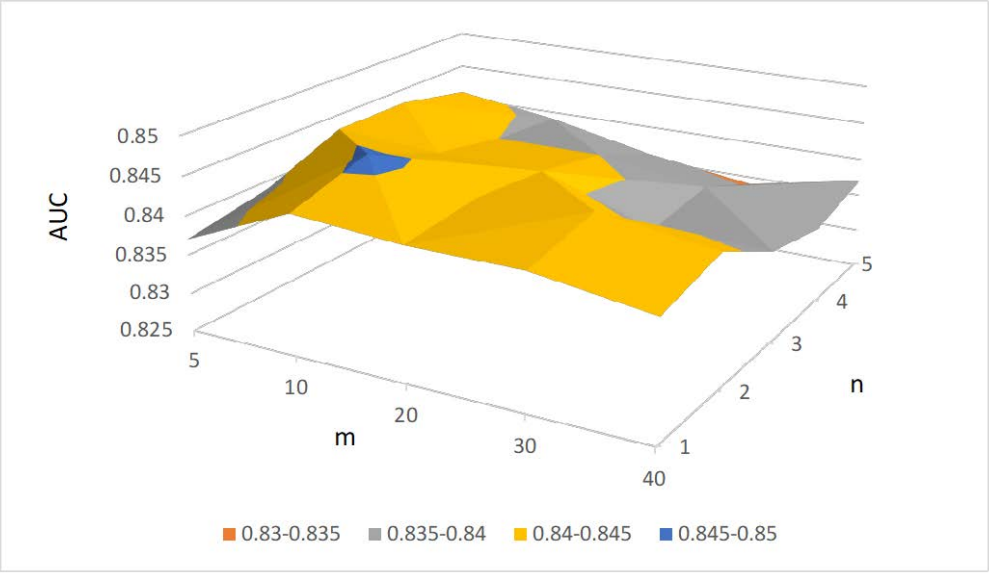
Tabela 19. Sintetičke karakteristike aminokiselina formirane PCA tehnikom koristeći 532 karakteristika aminokiselina opisanih u AAindex bazi podataka. Prikazane su prve dve sintetičke varijable.

	PC1	**PC2**
A	-0.97	-0.32
R	8.54	-13.79
N	14.87	1.57
D	18.13	-2.15
C	-8.37	8.30
E	12.03	-13.30
Q	7.92	-8.67
G	14.83	19.24
H	0.69	-6.18
I	-20.34	4.14
L	-17.64	-0.35
K	11.70	-13.64
M	-15.60	-5.75
F	-18.60	0.92
P	16.22	15.09
S	11.85	6.88
T	4.53	5.13
W	-16.30	-3.90
Y	-7.50	1.31
V	-16.01	5.49

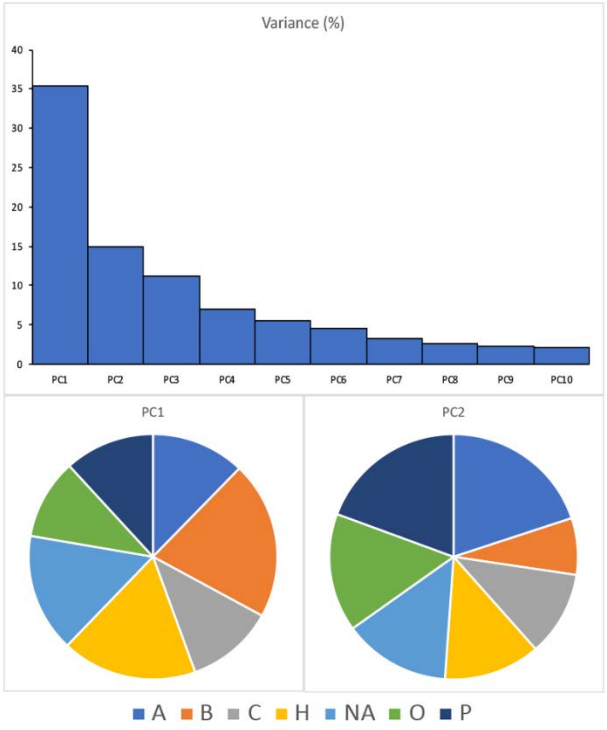
Broj sintetičkih varijabli dobijenih PCA (varijabla n) i veličina subsekvence na kojoj se računa ACC (varijabla m) određuju konačni broj atributa kojima se protein predstavlja. Da bi se pronašle optimalne vrednosti varijabli m i n u kontekstu predviđanja novih proteinskih interakcija, izvršena je analiza iscrpne pretrage. Na *Park i Marcotte* skupovima, zasnovanim na PINA bazi podataka IPP, testirana je efikasnost različitih kombinacija n i m parametara u predstavljanju IPP. Za generalizaciju IPP odabran je RF

algoritmom mašinskog učenja. Usled varijabilnog broja atributa kojim se IPP predstavljaju, variranjem m i n parametra, RF algoritam je izabran zbog svoje robusnosti i kompetativnih performansi korišćenjem podrazumevanih vrednosti njegovih meta-parametara. Za svaku od pet vrednosti n= (1,2,3,4,5) i m = (5, 10, 20, 30, 40) srednje vrednosti AUROC mere performansi modela testiranih na pet različitih skupova, prikazane su na Slici 18.

Na osnovu testiranja utvrđeno je da optimalan rezultat kod predikcije IPP ima numerički vektor proteina formiran računanjem ACC funkcije na dve komponente PCA (n = 2) (Tabela 19) u veličini posmatranog prozora od 10 aminokiselina (m = 10). Izabrani sintetički atributi i doprinos pojedinih grupa fizičko-hemijskih i drugih karakteristika iz AAindex baze podatka u formiranju ovih atributa prikazani su na Slici 19. Na ovaj način svaka IPP je predstavljena numeričkim vektorom PCA_AAC dužine 120.



Slika 18. Optimizacija parametara PCA_AAC modela proteinskih sekvenci. Parametar m predstavlja veličinu prozora u kome se posmatraju relacije između odabranih broja PCA komponenti (n parameter). Za svaku od pojedinačnih m,n kombinacija testirano je pet modela mašinskog učenja. Srednje vrednosti AUC mere performansi pet modela su prikazane.



Slika 19. Gornji panel: prve dve PCA komponente odabrane optimizacijom, omogućuju reprezentaciju 50% varijabilnosti 532 karakteristike aminokiselina AAindex baze podataka. Donji panel: na formiranje prve dve PCA komponente najveći uticaj su imale hidrofobnost i sklonost ka formiranju beta ploča (engl. beta propensity*). Boje predstavljaju grupe karakteristika aminokiselina klasifikovanih prema (Tomii and Kanehisa, 1996). Plava: sklonost ka formiranju alfa heliksa i zavonica (engl. alpha and turn propensities) (A), narandžasta: sklonost ka formiranju beta ploča (B), siva: kompozicija (C), žuta: hidrofobnost (H), tamno plava: fizičko-hemijske karakteristike (P), zelena: ostale karakteristike (O). Svetlo plavom su predstavljene karakteristike uključene u AAindex bazu podataka nakon verzije 3.0 (Tomii and Kanehisa, 1996).*

4.3.2.2 PSSM_AAC evolutivni atributi

Evolutivna sličnost proteinskih sekvenci je opisana u formi atributa evolutivnih profila. Za formiranje evolutivnih atributa korišćen je sledeći algoritam:

1. Za svaku sekvencu izračunata je PSSM matrica pomoću PSI-BLAST alata, sa podešenim parametrima: *num_iterations* = 3, *inclusion_ethresh* = 0.002, *evalue* = 10 i matrix = BLOSUM62. Ostali parametri su u podrazumevanim postavkama.
2. Na osnovu PSSM matrice, 20 atributa kompozicije aminokiselina je generisano za svaku proteinsku sekvencu pomoću Jednačine 11:

$$PSSM^{AAC}{}_{l} = \frac{1}{L}\sum_{j=1}^{L} p_{j,i} \quad i = 1..20 \tag{11}$$

gde L predstavlja dužinu proteinske sekvence, a $p_{j,i}$ je verovatnoća pojavljivanja i-te aminokiseline na j-toj poziciji unutar *L x 20* dimenzione PSSM matrice. Za svaku proteinsku sekvencu na ovaj način je formirano 20 PSSM_AAC atributa.

3. Spajanjem PSSM_AAC atributa pojedinačnih proteina para formiran je numerički vektor dužine 40 za predstavljanje IPP.

Kod testiranja performansi modela formiranih pomoću evolutivnih atributa, PSSM matica je računata samo na osnovu upoređivanja sa sekvencama proteina koji formiraju interakcije iz skupa za trening. Ovom procedurom se sprečava da podaci na kojima se testira model budu uključeni u postupak kojim se generiše klasifikator. Na ovaj način se omogućava objektivna evaluacija performansi prediktora zasnovanog na PSSM_AAC atributima.

4.3.2.3 Mrežni atributi

Interakcije među proteinima mogu biti predstavljene u formi grafa. Topološke karakteristike grafa nose informaciju o sposobnostima i preferencijama vezivanja između čvorova mreže (Eisenberg and Levanon, 2003). Topološke i strukturne karakteristike grafa proteinskih interakcija iskorišćene su za numeričko predstavljanje proteinskih sekvenci. Potvrđene IPP su predstavljene u formi neusmerenog i neotežanog grafa. Proteini su predstavljeni čvorovima, a poznate interakcije među njima granama. Analizirajući topološku strukturu formirane IPP mreže iskorišćene su 21 karakteristike čvora koji opisuju njegovu ulogu u strukturi mreže (Tabela 4). Svaka od karakteristika je predstavljena u numeričkoj formi, i predstavlja rezultat različitog pristupa analizi grafa. Vrednosti ovih mrežnih atributa su računate za svaki pojedinačni čvor (protein) u grafu IPP. Imena izračunatih karakteristika i opisi su predstavljeni u Tabeli 4. Svaki protein je opisan numeričkim nizom dužine 21, dok je IPP predstavljen spajanjem vektora interagujućih proteina (Graph_21). Da bi se izbegla kontaminacija informacijama i

netačno preterano optimistične performanse prediktora pri merenju efikasnosti pristupa, IPP mreža na osnovu koje su vrednosti mrežnih atributa računate, je formirana isključivo na osnovu IPP koji učestvuju u formiranju skupa za treniranje.

4.3.2.4 GAFT algoritam za automatsko generisanje i selekciju atributa

Skup atributa kreiran na osnovu domenskog znanja ne garantuje maksimalan predikcioni rezultat modela mašinskog učenja. Efikasna apstrakcija podataka opisanih skupom atributa zahteva prethodno filtriranje atributa. U svrhu inženjeringa atributa, razvijen je GAFT algoritam za automatsko formiranje i selekciju atributa u predikciji IPP. U osnovi GAFT algoritma se nalazi genetski algoritam. GAFT algoritam uključuje faze transformacije i selekcije inicijalnih skupova atributa u nove atribute većeg predikcionog potencijala. U toku procesa selekcije cilj je da se skup originalnih atributa filtrira i zameni novim, predikciono efikasnijim atributima, sa minimalnim povećanjem ukupnog skupa atributa. GAFT algoritam uključuje sljedeće faze:

1. Unarno generisanje.
   Faza generisanja novih atributa transformacijom ulaznog skupa atributa unarnim matematičkim operatorima: sin, exp, inv, log, sqr i sqrt
2. Unarna selekcija.
   Selekcija formiranih atributa zasnovana na korelaciji sa klasom.
3. Binarno generisanje.
   Formiranje atributa binarnim aritmetičkim operacijama (+, *, /) na skupu originalnih i izabranih unarnih atributa iz faze 2.
4. Binarna selekcija.
   Selekcija formiranih binarnih atributa na osnovu stepena međusobne korelacije.
5. Filtriranje predikciono najefikasnijih atributa za predviđanje IPP genetskim algoritmom za selekciju atributa.

Tabela 20. Brojevi atributa u toku različitih faza izvršenja GAFT algoritma na tri grupe ulaznih atributa: PCA_AAC, PSSM_AAC, Graph_21

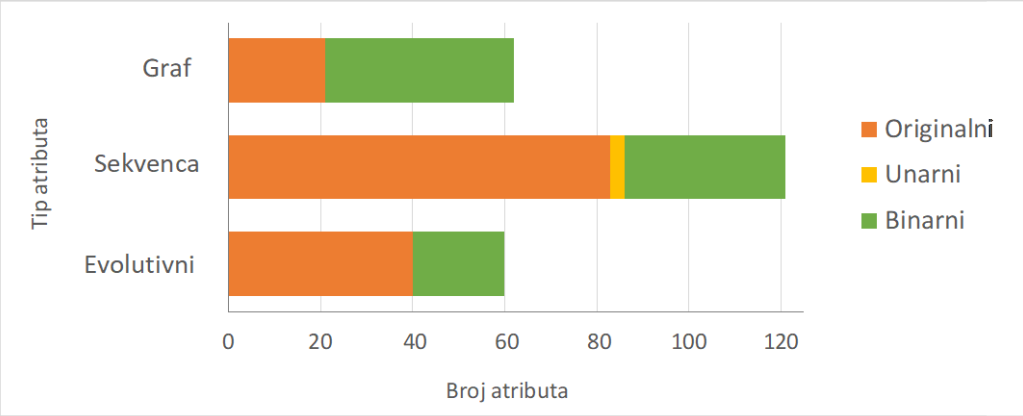
Atributi	Sekvenca	Evolutivni	Graf
Ulazni	120	40	42
Unarni	720	240	252
Izabrani unarni	6	54	17
Binarni	23,625	13,113	5,133
Izabrani binarni	67	81	110
Ulaz u GA algoritam	193	175	169
Izabrani GA algoritmom	121	60	62

Kao ulazni skup atributa za GAFT algoritam korišćene su posebno svaka od 3 grupe atributa: PCA_AAC, PSSM_AAC i Graph_21. Kao rezultat generacije i selekcije, originalni skup svake grupe atributa je izmenjen i proširen novim supervizovano selektovanim atributima.

Broj ulaznih atributa za predviđanje IPP i atributa u toku raznih faza izvođenja GAFT algoritma prikazani su u Tabeli 20.

U toku prve faze proverava se prisustvo nedostajućih vrednosti u originalnom skupu atributa. (Kuhn and Johnson, 2013). U cilju lakšeg i bržeg pronalaženja optimalnog skupa atributa u prostoru svih mogućih skupova atributa, originalni skup ulaznih atributa za predviđanje IPP je transformisan korišćenjem sedam unarnih matematičkih operatora. Selekcija novo dobijenih atributa je izvršena koristeći implicitno unarno filtriranje. Za svaki atribut izmeren je stepen korelacije sa klasom korišćenjem Spirmanovog koeficijenta korelacije (Conover, 1982). Ukoliko stepen korelacije ne prelazi unapred definisani prag, novokreirani atribut se eliminiše. Da bi se ispitale interakcije između različitih atributa, u sljedećoj fazi svi originalni i filtrirani unarni atributi su kombinovani binarnim aritmetičkim operacijama sabiranja, množenja i deljenja. Između svakog novoformiranog para atributa izračunat je Spirmanov koeficijent korelacije. Za svaku kombinaciju novogenerisanih atributa popunjena je matrica korelacije. Atributi čiji koeficijent ukupne korelacije prelazi unapred definisani prag se eliminišu. U poslednjoj

fazi originalni, filtrirani unarni i filtrirani binarni atributi su podvrgnuti nadgledanoj selekciji atributa genetskim algoritmom. Selekcija genetskim algoritmom pronalazi nadskup originalnih atributa većeg potencijala za generalizaciju modelom mašinskog učenja. Udeo originalnih atributa u novim skupovima atributa za tri grupe atributa za predviđanje IPP prikazan je na Slici 20.



Slika 20. Finalni skupovi atributa. Prikazan je udeo orginalnih i atributa kreiranih GAFT algoritmom (unarni i binarni) u konačnim skupovima atributa prema kategorijama.

4.3.3 HP-GAS metod i generisani modeli mašinskog učenja IPP proteoma čoveka

4.3.3.1 Izbor algoritma mašinskog učenja za odabir skupova za treniranje

U svrhu odabira skupova za formiranje finalnog modela izvršen je odabir metoda mašinskog učenja. Odabir algoritma mašinskog učenja kao i poređenje sa standardnim metodama za predviđanje IPP vršen je na standardnim trening i test skupovima (Park and Marcotte, 2012) zasnovanim na PINA (Wu *et al.*, 2009) bazi podataka IPP interakcija. Za odabir najadekvatnijeg algoritma mašinskog učenja, korišćeno je pet standardnih skupova. Test skupovi na kojima je testirana efikasnost modela imaju oba proteina iz interagujućih i neinteragujućih parova prisutna i u trening skupu (C1 tip). Za predstavljanje IPP korišćena je PCA_AAC reprezentacija. Da bi pronašli algoritam mašinskog učenja kojim bi smo formirali najefikasnije modele za predviđanje IPP,

testirali smo šest algoritama: GBM, RF, GLM, NB, DL, SVM. Pomoću svakog algoritma, pet modela je formirano na pet trening PINA standardnim skupova ljudskih IPP i testirano na odgovarajućim test skupovima. Srednje vrednosti i standardne devijacije sedam mera performansi klasifikatora na šest metoda mašinskog učenja prikazane su u Tabeli 21.

Tabela 21. Poređenje GBM, RF, GLM, NB, GLM, DL i SVM algoritama mašinskog učenja na pet parova (trening i test) skupova IPP PINA baze podataka formiranih od strane Park i Marcotte. IPP su reprezentovane PCA_AAC atributima. Srednja vrednost (SV) i standardna devijacija (SD) sedam mera performansi prediktora (AUROC, AUPRC, tačnost, senzitivnost, specifičnost, MCC i F1) su prikazane.

		AUROC	**AUPRC**	**Tačn.**	**Senz.**	**Spec.**	**MCC**	**F1**
GBM	SV	0.863	0.870	0.791	0.761	0.821	0.583	0.785
	SD	0.009	0.008	0.008	0.009	0.009	0.016	0.008
RF	SV	0.840	0.850	0.771	0.735	0.807	0.543	0.762
	SD	0.008	0.007	0.007	0.009	0.010	0.015	0.008
NB	SV	0.585	0.574	0.560	0.652	0.468	0.122	0.597
	SD	0.009	0.015	0.006	0.015	0.012	0.013	0.008
GLM	SV	0.579	0.560	0.559	0.561	0.557	0.118	0.560
	SD	0.005	0.008	0.007	0.014	0.008	0.015	0.010
DL	SV	0.624	0.619	0.514	0.509	0.490	-0.537	0.369
	SD	0.011	0.011	0.007	0.102	0.093	0.089	0.072
SVM	SV	0.806	0.808	0.741	0.722	0.722	0.487	0.738
	SD	0.007	0.007	0.005	0.005	0.005	0.007	0.004

Na osnovu rezultata u Tabeli 21, GBM algoritam mašinskog učenja je izabran za testiranje kvaliteta skupova IPP za formiranje finalnog modela, kao najefikasniji algoritam mašinskog učenja.

4.3.3.2 GA-STACK algoritam za automatsku formiranje i optimizaciju ansambla modela

Efikasna generalizacija podataka na kojima se uči zavisi i od odabira optimalnog algoritma mašinskog učenja za specifičan problem i podešavanja njegovih unutrašnjih parametara (Domingos, 2012). Spajanje različitih modela u ansambl može da obezbedi veći stepen predikcione efikasnosti nego bilo koji pojedinačni model (Rokach, 2010). U svrhu optimizacije spajanja različitih modela u nadgledanom maniru, razvijen je GA-STACK algoritam. GA-STACK je algoritam za automatsko spajanje i selekciju heterogeno formiranih modela mašinskog učenja, zasnovan na genetskom algoritmu. Različiti modeli se generišu dvostruko: (i) modifikacijom parametara pojedinačnog algoritma mašinskog učenja i (ii) primenom različitih algoritama mašinskog učenja za problem predikcije IPP.
Algoritmi mašinskog učenja koji formiraju modele u sklopu GA-STACK algoritma su: GLM, NB, DL, RF, GBM i XGBoost. Za svaki algoritam se na početku generiše se skup kombinacija meta-parametara iz koga se nasumično bira finalni skup kombinacija za korak selekcije genetskim algoritmom.

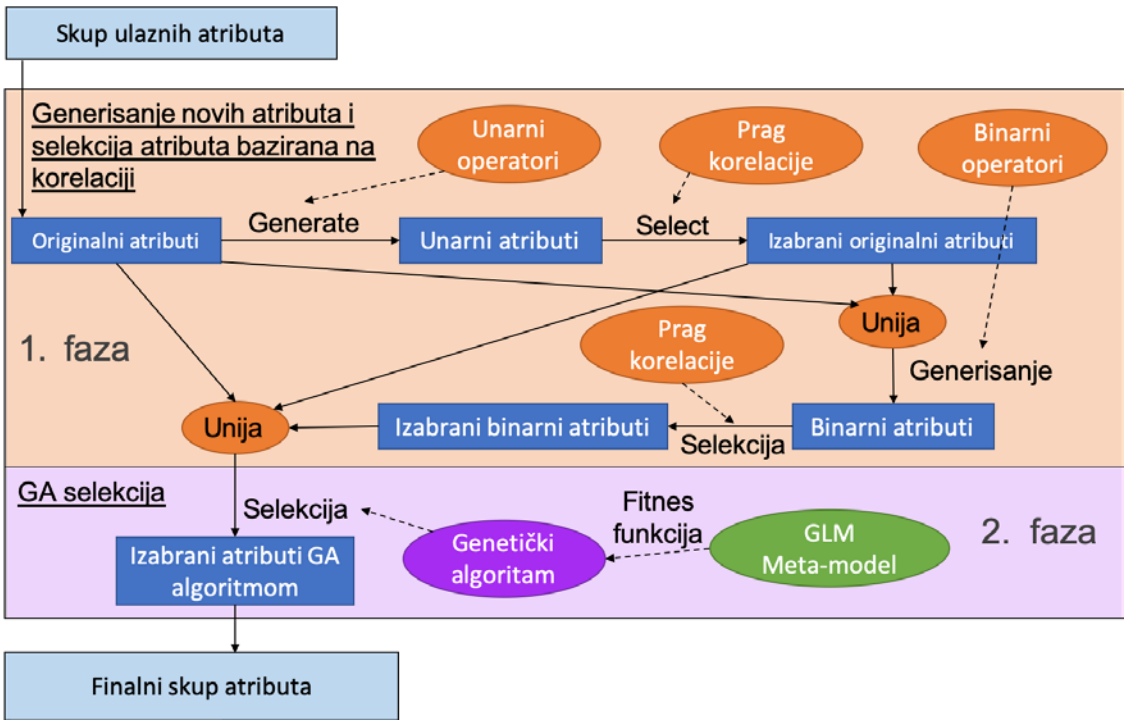
Protokol GA-STACK uključuje sledeće korake:

1. Generišu se skupovi kombinacija meta-parametara za svaki algoritam mašinskog učenja da bi se izgradili bazni modeli.
2. Skup za treniranje se deli korišćenjem CV-5 validacije.
3. Koristeći CV-5 šemu prikupe se predikcije na čitavom skupu za treniranje.
4. Eliminišu se kombinacije algoritama i njihovih meta-parametara čije predikcije na trening skupu imaju stepen Pirsonove korelacije veći od 0.95.
5. Posmatrajući nekorelisane predikcije različitih modela na trening skupu kao sintetičke atribute za meta-algoritam mašinskog učenja, primenjuje se genetski algoritam za selekciju atributa, odnosno modela. Cilj ove faze je pronaći optimalnu kombinaciju baznih prediktora koja će maksimizovati prediktivne performanse meta-algoritma mašinskog učenja.
6. Generalizacija mata algoritmom GLM na izabranim članovima ansambla.

7. Klasifikacija primera iz test skupa primenom GLM algoritma. Bazne predikcije na test skupu se formiraju primenom izabranih modela GA algoritmom na ulazni skup za testiranje.

4.3.3.3 HP-GAS protokol

GA-STACK algoritam je primenjen posebno na svaki od tri skupa IPP podataka opisanih atributima generisanim GAFT algoritmom na odgovarajućem ulaznom skupu atributa: PCA_AAC, PSSM_AAC i GRAPH_21. Za svaki skup atributa GA-STACK je generisao jedan model. Ovako dobijena finalna 3 modela su iskorišćena za formiranje konačnog ansambla zasnovanog na GLM meta-algoritmu. Novi protokol koji za predviđanje IPP koji uključuje formiranje 3 skupa atributa, GAFT algoritam i GA-STACK algoritama nazvan je HP-GAS. Šema HP-GAS protokola je prikazana na Slici 21.



Slika 21. Šema HP-GAS protokola. HP-GAS protokol uključuje GAFT algoritam i GA-STACK algoritam, koji su primenjeni na tri ulazna skupa atributa za predstavljanje IPP.

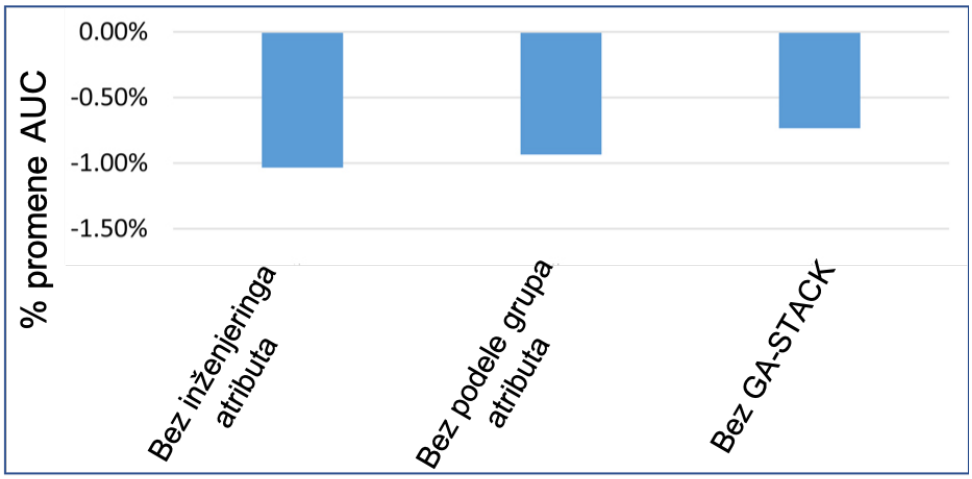
4.3.4 Evaluacija HP-GAS modela

4.3.4.1 Evaluacija efikasnosti strategija korišćenih u formiranja HP-GAS algoritma

HP-GAS metod je zasnovan na tri strategije:

1. Modeliranje proteinskih sekvenci pomoću tri pristupa (sekvenca, evolutivni profili i graf)
2. Automatsko generisanje i selekcija atributa (GAFT)
3. Automatsko formiranje ansambla modela (GA-STACK)

Kako bi se analizirao pojedinačni doprinos svake od ovih strategija, upotrebljen je sledeći eksperiment. Na prethodno odabranom C50' skupu generisani su i testirani predikcioni modeli svaki put eliminišući jedan postupak zasnovan na datoj strategiji. Osim odsustva jedne od tri pomenute strategije ostale procedure HP-GAS algoritma su ostale nepromenjene. Rezultati tri eksperimenta pokazuju da sve tri strategije imaju pozitivan uticaj na povećanje prediktivnih performansi finalnog modela, dok automatsko generisanje i selekcija atributa ima najznačajniji uticaj (Slika 24).



Slika 22. Promene performansi pri izostavljanju jedne od tri procedure HP-GAS pristupa. Intenzitet promene performansi nepotpunog HP-GAS pristupa je izražen u % promene u vrednosti AUROC u odnosu na potpun HP-GAS pristup. Procedure u ovim eksperimentima su: automatsko generisanje i selekcija atributa, modeliranje proteinskih sekvenci pomoću tri pristupa i automatsko formiranje ansambla modela. Negativna promena u % AUROC prikazuje smanjenje AUROC vrednost nakon izostavljanja jedne od tri procedure.

4.3.4.2 *Evaluacija efikasnosti strategije razdvajanja atributa prema distinktnim grupama*

U sklopu HP-GAS pristupa, automatsko generisanje i selekcija atributa kao i generisanje modela od kojih će se formirati konačni ansambl, vrši se odvojeno na tri grupe atributa. Rezultati poređenja ovakvog pristupa u odnosu na spajanje svih atributa u sklopu jednog skupa primera za treniranje modela predstavljeni su u Tabeli 22. Pokazuje se da je formiranje modela na odvojenim grupama atributa koji su prethodno podvrgnuti GAFT proceduri i njihovo spajanje GA-STACK protokolom, superiorniji pristup.

Tabela 22. Poređenje HP-GAS pristupa sa odvojenim i spojenim (HP-GAS-spojeni) atributima iz tri distinktne grupe. Vrednosti 7 mera performansi prediktora su predstavljene.

	HP-GAS	HP-GAS-spojeni
AUROC	0.928	0.919
AUPRC	0.927	0.918
Tačnost	0.853	0.844
F vrednost	0.853	0.840
Preciznost	0.858	0.857
Specifičnost	0.859	0.864
Senzitivnost	0.848	0.824
MCC	0.707	0.687

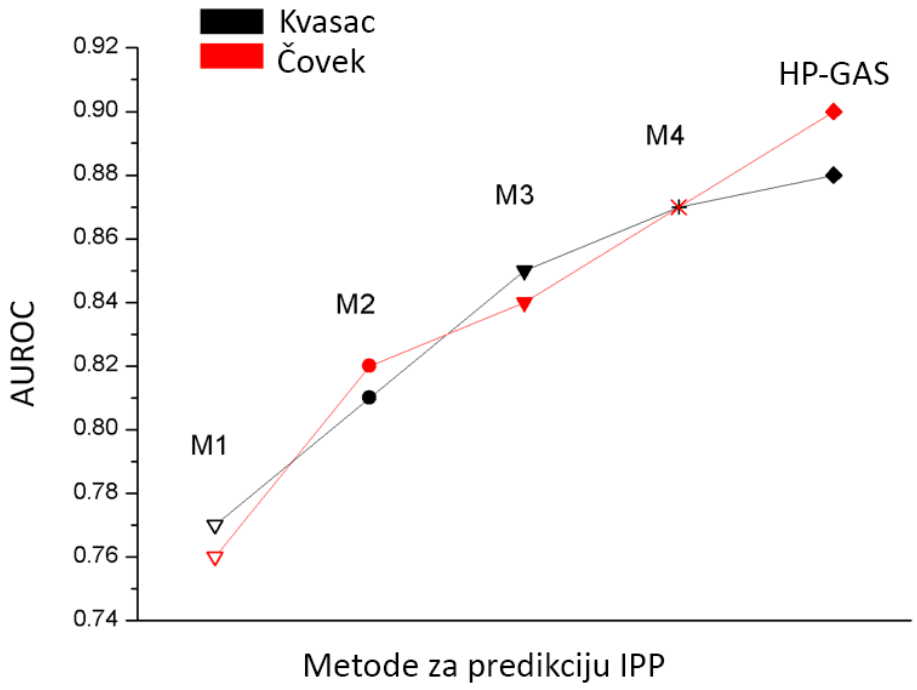
4.3.4.3 *Evaluacija performansi HP-GAS pristupa i poređenje sa standardnim metodama*

Da bi se ispitala tačnost predikcije ljudskih IPP, HP-GAS pristupom, na prethodnom odabranom C50` skupu, CV-10 validacija je primenjena. Srednje vrednosti pet mera performansi klasifikatora generisanih od različitih algoritama mašinskog učenja kao i prediktivna sposobnost finalnog HP-GAS modela su prikazane u Tabeli 23.

Tabela 23. Poređenje različitih klasifikacionih algoritama na C50` skupu koristeći CV-10 validaciju. Srednje vrednosti i standardne devijacije pet mera performansi prediktora su prikazane za pet algoritama mašinskog učenja i HP-GAS pristup.

	AUROC	**AUPRC**	**Tačnost**	**F1**	**MCC**
GLM	0.680 ± 0.010	0.668 ± 0.013	0.633 ± 0.007	0.643 ± 0.004	0.267 ± 0.015
XGB	0.863 ± 0.077	0.863 ± 0.077	0.788 ± 0.069	0.784 ± 0.074	0.576 ± 0.137
DL	0.697 ± 0.093	0.688 ± 0.095	0.648 ± 0.074	0.659 ± 0.059	0.297 ± 0.144
RF	0.697 ± 0.095	0.696 ± 0.107	0.649 ± 0.082	0.643 ± 0.068	0.302 ± 0.169
GBM	0.888 ± 0.015	0.889 ± 0.014	0.807 ± 0.016	0.806 ± 0.017	0.615 ± 0.032
GA-STACK	**0.928 ± 0.001**	**0.927 ± 0.002**	**0.853 ± 0.002**	**0.853 ± 0.001**	**0.707 ± 0.004**

Da bi se poredile performanse predikcije IPP HP-GAS pristupa sa standardnim metodama za predviđanje IPP, koristili smo standardne skupove za evaluaciju formirane od strane autora *Park i Marcotte* (Park and Marcotte, 2012) zasnovanim na PINA (Wu *et al.*, 2009) bazi podataka IPP interakcija (Hamp and Rost, 2015b). Poređenje je izvršeno sa SigProd metodom (M1), metodom sa najboljim predikcionim performansama testiranim i implementiranim od strane autora *Park i Marcotte* (Park and Marcotte, 2012)(M3) i *state-of-the-art* metodom, PPI-PK (M3) (Hamp and Rost, 2015b). HP-GAS modeli su generisani i testirani na 40 parova skupova za trening i testiranje formiranih od ljudskih IPP i 40 uparenih test i trening skupova formiranih od IPP kvasca (*Saccharomyces cerevisiae*). Autori Park i Marcotte su prijavili rezultate evaluacije M1 i M2 metoda, a Hump i Rost M3 metode na istim skupovima. Srednje vrednosti AUROC mere HP-GAS, M1, M2 i M3 metoda na ljudskim i skupovima IPP kvasca (*Saccharomyces cerevisiae*) prikazane na Slici 23.



Slika 23. Poređenja HP-GAS pristupa sa M1, M2 i M3 standardnim metodama za detekciju IPP prema srednjim vrednostima AUROC mere performansi prediktora na 40 ljudskih i 40 skupova IPP kvasca.

Na osnovu Slika 23, vidljiv je veći stepen prediktivne efikasnosti HP-GAS metoda u odnosu na state-of-the-art i standardne metode za predviđanje IPP, testirano na referentim skupovima IPP čoveka i kvasca.

Pored poređenja na *Park i Marcotte* skupovima, poređen je HP-GAS pristup sa PPI-PK (M3) metodom na C50' skupu koristeći prethodno formirane skupove za treniranje i testiranje (CV-10 validacija). Rezultati pokazuju da je HP-GAS superioran u odnosu na PPI-PK (M3) u vrednostima svih mera performansi predviđanja novih IPP na nivou proteoma čoveka (Tabela 24).

Tabela 24. Rezultati poređenja HP-GAS pristupa i state-of-the-art PPI-PK(M3) metode na C50' skupu koristeći CV-10 validaciju.

	HP-GAS	**PPI-PK(M3)**
AUROC	0.928 ± 0.001	0.906 ± 0.002
AUPRC	0.927 ± 0.002	0.905 ± 0.003
Tačnost	0.853 ± 0.002	0.834 ± 0.004
F vrednost	0.853 ± 0.001	0.836 ± 0.002
Preciznost	0.857 ± 0.006	0.834 ± 0.009
Specifičnost	0.859 ± 0.007	0.829 ± 0.011
Senzitivnost	0.848 ± 0.004	0.839 ± 0.006
MCC	0.707 ± 0.004	0.668 ± 0.007

Dodatno poređenje tačnosti HP-GAS i state-of-the-art pristupa PPI-PK (M3) izvršeno je na HIPPIE_new skupu IPP, koji sadrži eksperimentalno potvrđene IPP ažurirane u noviju verziju HIPPIE baze podataka od one korišćene za formiranje C50' skupa. HP-GAS pristup je IPP iz HIPPIE_new skupa predviđao tačnije i sa većim ukupnim brojem detektovanih IPP (Tabela 25). Pored ovoga HP-GAS metod je predvideo veći broj ekskluzivno pozitivnih IPP. Ekskluzivne pozitivne interakcija predstavljaju one IPP koje su predviđene ekskluzivno samo jednom od metoda, HP-GAS ili PPI-PK (M3).

Tabela 25. Poređenje prediktivnih performansi HP-GAS i PPI-PK (M3) na HIPPIE_new skupu. Ekskluzivne pozitivne interakcija predstavljaju one IPP koje su predviđene ekskluzivno samo jednom od metoda, HP-GAS ili PPI-PK (M3).

	HP-GAS	**PPI-PK(M3)**
Tačnost	0.702	0.663
F mera	0.825	0.797
Ukupan broj pozitivnih IPP	33,618	31,738
Broj ekskluzivno pozitivnih IPP	5,326	3,446

Dodatno su testirane performanse HP-GAS pristupa za predviđanje IPP u slučajevima kada broj negativnih primera nadmašuje broj pozitivnih. Za ovakav scenario formirana je serija skupova sastavljena od pozitivnih primera test skupova C50` i negativnih u razmeri: 1:10 (C50'$_{10}$-test), 1:100 (C50'$_{100}$-test) i 1:400 (C50'$_{400}$-test). U slučaju sva tri test skupa AUROC je zadržao jednaku vrednost zbog svoje neosetljivosti na relativan broj negativnih u test skupu (Park, 2009; Park and Marcotte, 2011). Sa druge strane, vidljivo je značajno smanjenje vrednosti AUPRC (Tabela 26).

Tabela 26. Testiranje HP-GAS pristupa na test skupovima sa nebalansiranim odnosom pozitivnih i negativnih primera. Vrednosti tri mere performansi prediktora su predstavljene.

	C50'$_{10}$-test	**C50'$_{100}$-test**	**C50'$_{400}$-test**
AUROC	0.925	0.925	0.925
AUPRC	0.644	0.258	0.115
Tačnost	0.848	0.849	0.849

4.3.5 Studija slučaja EGFR proteina

Detekcija proteinskih interakcija koje čine osnovu signalne kaskade u procesu maligne transformacije ostaje jedan od najvećih izazova molekularne biologije. Jedan od ključnih proteina prenosa signala, EGFR (engl. Epidermal growth factor receptor) receptor tirozinska kinaza, igra važnu ulogu u ćelijskom razviću, diferencijaciji i apoptozi. Poremećaji EGFR posrednog prenosa signala se povezuju sa razvojem niza tumora. Stoga, identifikacija EGFR interaktora predstavlja važan korak u razvoju novih terapija (Seshacharyulu *et al.*, 2012; Sigismund, Avanzato and Lanzetti, 2018).

HP-GAS je upotrebljen za pretragu potencijalnih EGFR interaktora EGFR u ljudskom proteomu. Među naviše rangiranim potencijalnim kandidatima na osnovu HP-GAS predikcije se ističu SUMO (engl. Small Ubiquitin-like Modifier) 1, 2 , 3 i 4 regulatorni faktori (Tabela 27) koji igraju važnu ulogu u rastu i diferencijaciji ćelija kao i ćelijskom odgovoru na stres (Wilkinson and Henley, 2010; Eifler and Vertegaal, 2015). Posttranslaciona modifikacija koja se izvodi posredstvom SUMO proteina, sumoilacija je značajan instrument u ćelijskom odgovoru na stres, te abnormalne promene ovog mehanizma su prisutne u mnogim kancerima (Seeler and Dejean, 2017). Pretragom literature pronađena je eksperimentalna potvrda EGFR interakcije sa SUMO-1, koja prethodno nije bila deo HP-GAS modela (Packham *et al.*, 2015). Nadalje, u sklopu studije endogenih posttranslacionih modifikacija EGFR je pokazana njegova interakcija sa SUMO 2/3 (Horita *et al.*, 2017).

Tabela 27. Predviđanja HP-GAS pristupom interakcije između EGFR i SUMO 1,2,3 i 4 proteina sa relativnim verovatnoćama.

Protein A	**Protein B**	**Verovatnoća**
EGFR_HUMAN	SUMO1_HUMAN	0.98988779
EGFR_HUMAN	SUMO2_HUMAN	0.98981107
EGFR_HUMAN	SUMO3_HUMAN	0.98748174
EGFR_HUMAN	SUMO4_HUMAN	0.97683854

4.3.6 Implementacija u formi alata

HP-GAS pristup je implementiran u formi jednostavnog za korišćenje i široko dostupnog, samostalnog alata. Softverski alat HP-GAS je formiran na celom C50` skupu IPP. Alat i uputstvo za korišćenje su dostupni na veb stranici: https://www.vinca.rs/180/tools/HP-GAS.php

5 Diskusija

Najvažniji zadaci kod razvoja algoritma za predviđanje IPP su: (i) izbor i obrada adekvatnih podataka kao osnova za formiranje prediktivnog modela i njegovo testiranje, (ii) pronalaženje opšteg pristupa za predstavljanje proteinskih sekvenci, (iii) pronalaženje efikasnog algoritma za učenje modela ML i (iv) razvoj alata zasnovanog na prethodno generisanom algoritmu.

Priprema ulaznih podataka je prvi i jedan od najvažnijih koraka od kojeg zavisi efikasnost konačnog modela za predikciju IPP (Domingos, 2012). Priprema podataka na kojima se uči budući model zasniva se na pažljivom izboru izvora podataka i procedura za filtriranje tih podataka. Ovo se u prvom redu odnosi na baze podataka proteinskih sekvenci. Kod modeliranja IPP transkripcionih regulatora u ovoj studiji, kao osnova za sekvence koristila se UniProtKB/Swiss-Prot baza podataka. Filtriranje proteinskih sekvenci je dovelo do višestrukog smanjenje broja proteina čije interakcije se mogu koristiti za formiranje modela za predviđanje. Iako se na ovaj način smanjuje broj dostupnih informacija za učenje modela, on omogućuje da se model uči samo na pouzdanim informacijama. Sličan kompromis bio je neophodan i u sklopu procedure izbora visoko pouzdanih proteinskih sekvenci za predviđanje IPP PNTS. Povećanja broja dostupnih sekvenci proteina u bazama podataka, ne znači automatski i povećanje broja informacija neophodnih za formiranje prediktivnih modela.

Izbor adekvatnih baza podataka eksperimentalno potvrđenih IPP i različite procedure filtriranja, nalaze se u osnovi najvećeg broja do sada razvijenih pristupa za predviđanje IPP (Martin, Roe and Faulon, 2005; Shen *et al.*, 2007; Guo *et al.*, 2008; Hamp and Rost, 2015a). Usled relativno malog preseka različitih baza podataka koje sadrže eksperimentalno potvrđene IPP (Cusick *et al.*, 2009b; Turinsky *et al.*, 2010) i različite brzine ažuriranja pojedinih baza podataka, izbor adekvatne baze podataka IPP može predstavljati izazov (Gemovic *et al.*, 2018). U tu svrhu smo kao osnovu za izbor kvalitetnih IPP u sklopu ove studije izabrali HIPPIE meta-bazu podataka, koja integriše deset drugih baza podataka IPP i redovno se ažurira. HIPPIE integriše informaciju o pouzdanosti informacije o svakoj pojedinačnoj IPP. Analogno gore spomenutom slučaju sekvenci proteina, neophodno je osigurati neredundantnost informacije sadržane u formiranim skupovima IPP pripremljenim za treniranje klasifikatora. Jedan od

potencijalnih faktora koji može pružiti preterano optimistične performance prediktora su različite forme redundantnosti informacija poput inverznih IPP ili visokog stepena homologije proteina u skupovima za treniranje i testiranje. Baze podataka IPP često sadrže IPP iz različitih izvora, koje se razlikuju samo prema redosledu proteina u sklopu IPP (AB i BA). Zatim, homologija proteina koji formiraju različite IPP (protein A homologan sa proteinom C, protein B sa proteinom D, kod IPP AB i CD) je poznat uzrok prijavljivanja preterano optimističnih performansi prediktora (Park, 2009). U sklopu procedura za formiranje modela za predikciju IPP transkripcionih regulatora, PNTS i opštih IPP čoveka, posebna pažnja posvećena je uklanjanju različitih vrsta redundantnosti u podacima. Filtriranje podataka o IPP neminovno dovodi do smanjenja broja IPP dostupnih za formiranje modela za predviđanje, zarad poboljšanje prediktivnih performansi modela. Jedan od budućih pravaca studija koje se bave generisanjem modela za predikciju IPP, može biti razvoj alata i procedura koje bi dovele do povećanja iskoristivosti već dostupnih informacija o proteinima i njihovim IPP.

Pored filtriranja podataka preuzetih iz baza podataka proteina i IPP, izbor i dizajn skupova za učenje može predstavljati značajan faktor uticaja na efikasnost formiranog modela kod specifičnog problema. Poređenje modela za predikciju IPP PNTS formiranih identičnim algoritmom, ali na različitim skupovima podataka ukazuje na neophodnost specijalizacije podataka na kojima se model trenira.

Pored odabira podataka za učenje klasifikatora, jedan od najvažnijih faktora u formiranju efikasnog prediktivnog modela je izbor atributa (Domingos, 2012). Metode za predviđanje IPP zasnovane na sekvenci, imaju prednosti poput dostupnosti podataka za treniranje i testiranje. U tom slučaju jedina informacija neophodna za formiranje modela i predviđanje IPP je sekvenca proteina (Shen *et al.*, 2007). Studije su pokazale da je proteinska sekvenca dovoljno informativnosti za uspešno modeliranje IPP (Park and Marcotte, 2011). Iako metode koje svoje modele za predviđanje IPP zasnivaju samo na sekvenci proteina imaju univerzalnu primenu, jedan od najvećih problema ovih metoda je razvoj uspešnog predstavljanja proteinskih sekvenci varijabilne dužine u formi adekvatnoj za treniranje modela ML. Predstavljanje sekvenci proteina različitih dužina u formi vektora istih dužina, uz minimalni gubitak informacije očuvane u sekvenci proteina, predstavlja ključan korak u procesu razvoja efikasnog klasifikatora IPP. Pored

ovoga, model kojim se proteinske sekvence predstavljaju numerički mora da obuhvati generalne karakteristike skupa sekvenci koji se na ovaj način modeliraju.

Transkripciona regulacija je precizno vremenski i prostorno regulisana velikim brojem IPP transkripcionih regulatora (Lee and Young, 2013). Robusnost elemenata mreže IPP koji imaju ulogu u transkripciji, ukazuje na visok stepen evolutivne zaštite čitavog sistema transkripcije, uspostavljanjem velikog broja međusobnih IPP ključnih učesnika regulacije (Francois, Donovan and Fontaine, 2018). Čuov PAAC model pored histograma aminokiselinskog sastava sekvence proteina, uključuje i informaciju o strukturi i redosledu aminokiselina (Chou, 2009). Kako bi se iskoristila potencijalna informacija o dalekosežnim interakcijama, pored karakteristika hidrofilnosti i hidrofobnosti mase bočnog lanca inicijalno obuhvaćenih PAAC modelom, uključili smo i EIIP karakteristiku (Veljkovic, 1980). Modifikovani PAAC4 model omogućava očuvanje informacije o redosledu aminokiselina proteinske sekvence na maksimalnoj udaljenosti, što može biti posebno značajno za proteine visokog promiskuiteta u formiranju IPP, poput transkripcionih regulatora.

Slično tome, visok stepen specifičnosti PNTS proteina u broju interaktora i načinu vezivanja, kao i strukturi sekvence (Uversky, 2013), zahteva odgovarajući matematički model sekvenci kojim bi se predstavile ove specifičnosti u primeni principa prepoznavanja interagujućih partnera i formiranja IPP (Mészáros *et al.*, 2007). Kod velikog broja transkripcionih faktora čoveka pronađeni su RNTS regioni (Dunker and Uversky, 2010). Poređenje broja IPP ciljne grupe PNTS i PUTS na osnovu informacija iz HIPPIE baze podataka pokazalo je da PNTS proteini formiraju u proseku 3.5 više interakcija u odnosu na PUTS. U tu svrhu korišćen je sličan model matematičkog predstavljanja kao u slučaju transkripcionih regulatora. Ipak, visok stepen specifičnosti neuređenosti struktura PNTS proteina zahtevao je korišćenje posebno izabranih karakteristika aminokiselina, u sklopu već predstavljenog PAAC5 modela. Na osnovu analize dostupnih podataka o karakteristikama sekvence PNTS koje su povezane sa neuređenošću tercijarne strukture, izdvojeno je pet karakteristika aminokiselina povezanih sa neuređenošću u tercijarnoj strukturi. Slično kao u slučaju PAAC4 modela, kod PAAC5 modela relacija između ovih karakteristika posmatrana je na maksimalnoj udaljenosti, kako bi se obuhvatili što veći regioni sekvence koji mogu biti odgovorni za interakciju sa ciljnim partnerom. Nestabilnost 3D strukture PNTS proteina nosi određene

specifičnosti u sekvenci ovih proteina, poput specifične kompozicije aminokiselina (Radivojac *et al.*, 2007). Iako PAAC5 model u svom sklopu sadrži informaciju o 1-mer kompoziciji aminokiselina, u sklopu ove studije uključili smo i informaciju o 2-mer kompoziciji. Ovako formirani DP_PAAC5 model sekvenci PNTS sposoban je da očuva znatno više informativnosti o specifičnostima kompozicije ovih proteina.

Korišćenje heterogene informacije zasnovane na sekvenci proteina, u svrhu formiranja PAAC4 i složenijeg DP_PAAC5 modela, zahteva izbor i fino naštimavanje algoritma ML koji može uspešno da vrši generalizaciju na ovim podacima. Poređenje različitih algoritama mašinskog učenja, potvrdilo je superiornost ansambl algoritma RF u formiranju efikasnog modela za predviđanje IPP. Na sličan zaključak navodi i rezultat poređenja šest različitih algoritama mašinskog učenja formiranjem modela za predviđanje različitih fizičkih, funkcionalnih i regulatornih interakcija među proteinima (Qi, Bar-Joseph and Klein-Seetharaman, 2006). Znatno veća studija, koja je analizirala performanse 14 familija klasifikacionih algoritama mašinskog učenja na 115 skupova diskutuje da će primena RF, SVM, i GBM algoritama najverovatnije dovesti do najveće predikcione tačnosti kod različitih klasifikacionih problema (Wainer, 2016).

Poređenje sa standardnim metodama za predviđanje opštih IPP čoveka, ukazalo je na prednosti finog podešavanja matematičkog modela sekvenci biološkoj realnosti. Standardne metode za predviđanje IPP čoveka pokazale su znatno manju prilagođenost specifičnosti IPP transkripcionih regulatora od PAAC4_RF modela. Pored boljih prediktivnih performansi PAAC4_RF modela, algoritam formiranja PAAC4 i PAAC4_RF modela demonstrirao je značajno brže vreme izvođenja procedura, u odnosu na standardne metode za predviđanje IPP. Ova računarska efikasnost je omogućila kreiranje, efikasnog TRI_tool veb servisa. Na ovaj način olakšan je pristup našem alatu široj naučnoj zajednici. Osim brzine predviđanja potencijalnih IPP, radno okruženje TRI_tool servisa dizajnirano je kako bi minimiziralo neophodno vreme za efikasno korišćenje alata i tumačenje rezultata. Značajan stepen poboljšanja prediktivnih performansi u odnosu na standardne metode je postignut i u slučaju DP_PAAC5_RF prediktivnog modela PNTS IPP. Potencijalni broj NIPP daleko prevazilazi potencijalni broj IPP u sklopu biološkog sistema (Park, 2009). U tom kontekstu, postoji mogućnost da će se u realnom test skupu ispitivanih IPP naći veći broj NIPP u odnosu na IPP. Poređenje performansi DP_PAAC5_RF u odnosu na standardne metode, ukazalo je na

razliku u performansama u korist našeg pristupa, zadržane i u slučajevima nebalansiranih test skupova (10 i 100 puta veći broj NIPP u odnosu na IPP), u odnosu na testirane standardne metode. Algoritam za formiranje DP_PAAC5_RF modela je učinjen dostupnim u formi IDPi_tool veb servisa. Slično kao kod TRI_tool, posebna pažnja je bila usmerena kako bi se osim dostupnosti, alat učinio što više prilagođen korisniku.

TRI_tool je dodatno testiran kroz ispitivanje slučaja Vilmsovog tumor proteina, WT1. Transkripcioni faktor WT1 povezan je sa čitavim nizom malignih oboljenja kod čoveka (Huff, 2011). WT1 svoje funkcije ostvaruje kroz interakcije sa nizom drugih proteina, od kojih je p53 tumor supresor jedan od njegovih najpoznatijih interaktora (Maheswaran *et al.*, 1993). Praktična primena TRI_tool alata, odnosno PAAC4_RF modela, pokazana je kroz eksperimentalno potvrđivanje interakcije koja nije bila uključena u treniranje modela, ali je poznata iz literature. TRI_tool je uspešno predvideo takvu interakciju, WT1-TFIIB.

Proteinske kinaze igraju važne uloge u različitim ćelijskim procesima. Pored ove poznata je uloga proteinskih kinaza u mnogim patološkim procesima, pre svega malignim oboljenjima kod čoveka (Shchemelinin, Sefc and Necas, 2006). Svoje funkcije u svim ovim procesima ostvaruju kroz mrežu IPP, pa se često ističu kao primarni ciljevi potencijalnih terapeutika. TRI_tool je korišćen za predviđanje potencijalnih interakcija WT1 i skupa od 31 proteinske kinaze kod čoveka. Oslanjajući se na predviđanje uz korišćenje našeg alata, interakcija između WT1 i ciklin-zavisne kinaze CDK9 je i eksperimentalno potvrđena (Perovic *et al.*, 2017). Ova dva primera su poslužila za demonstraciju praktične aplikacije našeg alata.

BASP1 protein je karakterističan zbog RNTS duž čitave sekvence (Forsova and Zakharov, 2016) i stoga adekvatan primer za praktično testiranje IDPi_tool alata, odnosno DP_PAAC5_RF modela. BASP1 je transkripcioni koregulator koji je eksprimiran u različitim stadijumima razvića, a bitno je istaći da je BASP1 gen utišan u nekoliko tipova tumora (Yeoh *et al.*, 2002; Moribe *et al.*, 2008; Guo *et al.*, 2016). Budući da uloga BASP1 u različitim oboljenjima nije u potpunosti razjašnjena, pokušali smo da unapredimo uvid u BASP1 funkcije proširujući listu njegovih potencijalnih interaktora. Prema literaturi, postoji 11 poznatih interaktora BASP1, a IDPi_tool je uspeo da tačno predvidi 7 od 11 interaktora. Koristeći IDPi_tool predviđen je PRGR (engl. Progesterone Receptor) progesteronski receptor, kao jedan od novih interaktora BASP1. Ova interakcija je

naknadno eksperimentalno potvrđena (Perovic *et al.*, 2018). Prethodno je eksperimentalno pokazana interakcija između BASP1 (Carpenter *et al.*, 2004) i WT1 , v-myc (mielocitomatozni viralni onkogenski homolog) (Han *et al.*, 2013) i ESR1 estrogenskog receptora (Marsh *et al.*, 2017). Interakcija BASP1 i PRGR kao i predviđene interakcije sa nekoliko drugih transkripcionih faktora ukazuju na potencijalnu ulogu BASP1 kao važnog transkripcionog koregulatora. Prethodne studije su opisale da BASP1 interaguje sa proteinima HDAC1 i SMCA4, koji imaju ulogu u remodelovanju hromatina (Toska *et al.*, 2014). Slično tome, nalazi ove studije vezani za predviđene interaktore BASP1, takođe ukazuju na njegovu vezu sa ovom grupom proteina. Rezultati prethodnih analiza i rezultati predstavljeni u sklopu ove studije, ukazuju na potencijalni značaj BASP1 kao regulatora transkripcije. Analizirali smo razlike skupa GO termina predstavljenih već poznatim interaktorima BASP1, u odnosu na skup GO termina predviđenih našim alatom, u domenu GO ontologije bioloških procesa (Slika 15). Posmatrajući obe grupe GO termina, ističu se funkcije u organizaciji ćelijskih komponenti, procesima RNK metabolizma i uloge u ćelijskom ciklusu. Analizirajući predviđene BASP1 interakcije i njihov GO semantički prostor, ističu se dve još uvek nedovoljno istražene funkcije BASP1: savijanje proteina i ćelijski procesi vezani za viruse. Poznato je da su obe funkcije usko povezane sa proteinima nestabilne tercijarne strukture (Wright and Dyson, 2015).

Razvoj pristupa za predviđanje IPP na nivou proteoma uključuje dodatne izazove u odnosu na razvoj specijalizovanih pristupa, poput algoritama za predviđanje IPP transkripcionih regulatora i PNTS proteina. Najvažnije izazove pri ispunjenju ovih zadataka predstavljaju: (i) razvoj modela superiornih predikcionih performansi, (ii) održavanje zadovoljavajućeg nivoa računarske efikasnosti i (iii) mogućnost da se čitav sistem pretvoriti u efikasan i široko dostupan alat. Ispunjavanje postavljenih zadataka uz prevazilaženje prethodno pomenutih izazova, često je nelinearan proces.

Prevazilaženje ovih problema i na manjim zadacima, kao u slučaju razvoja metoda za predviđanje IPP transkripcionih regulatora i PNTS proteina, može predstavljati značajan izazov. Često prevazilaženje određenog problema podrazumeva kompromis u odnosu na drugi. Mada su često za superiorne prediktivne performanse neophodni kompleksniji modeli ML, kompleksniji modeli ML podrazumevaju veću potrošnju računarskih i materijalnih resursa u svakoj fazi razvoja takvog algoritma (Domingos,

2012). Potrošnja resursa dramatično raste sa povećanjem količine podataka koja se modelira, kao sto je slučaj sa razvojem modela za predviđanje IPP na nivou proteoma kod čoveka. Dok je u slučaju modeliranja transkripcionih regulatora konačni skup za modeliranje sadržao 24448 IPP, a kod modeliranja interakcija PNTS proteina već uključivao skoro duplo više IPP, konačni skup za formiranje opšteg modela za predviđanje IPP čoveka sadržao je oko četrnaest puta više IPP (345924). Za formiranje efikasnog opšteg modela za predviđanje IPP ljudskih proteina na nivou proteoma, neophodan je veliki broj primera za učenje. Ipak, konstantno povećanje broja IPP ne garantuje simultano povećanje efikasnosti prediktivnog modela formiranog na tim podacima. Izbor kvalitetnih podataka među eksperimentalno potvrđenim IPP igra važnu ulogu. Više od dvadeset hiljada novih IPP C00 skupa, u odnosu na C50 skup IPP, nižeg stepena pouzdanosti prema normama HIPPIE baze podataka, dovode do smanjenja tačnosti predviđanja novih IPP. To ukazuje na neophodnost pažljivog konstruisanja skupa za učenje koji sadrži dovoljno raznolikosti među proteinima i njihovim IPP, potvrđenim iz više izvora i eksperimentalnih metoda. Na sličan zaključak navode rezultati studije dizajniranja M3 (PPI-PK) metoda za predikciju IPP (Hamp and Rost, 2015a). Pažljivo pripremljen skup HIPPIE_C50, u sklopu naše studije, pokazao se superiornim u kontekstu broja proteina koje obuhvata i njihovih IPP u odnosu na referentni uPK skup IPP (Park and Marcotte, 2012).

Pored neophodnog i značajnog povećanja broja IPP primera potrebnih za treniranje efikasnog modela, bilo je nophodno pronaći skup atributa koji najbolje opisuje IPP u kontekstu njihove raznolikosti u ljudskom interaktomu. Nekoliko prethodno objavljenih studija ukazalo je da efikasnost prediktivnog modela IPP može biti povećana integracijom različitih tipova informacija (Jansen, 2003; Ben-Hur and Noble, 2005; Qi, Bar-Joseph and Klein-Seetharaman, 2006; Elefsinioti *et al.*, 2011). Informacije kao što su GO anotacije proteina i opisane 3D strukture, često nisu dostupne za novo-sekvencirane proteine ili proteine koji nisu dovoljno istraženi u literaturi. Imajući u vidu prednosti razvoja metoda koji se oslanjaju samo na sekvencu proteina i njihove interakcije (Shen *et al.*, 2007), u sklopu ove studije je razvijen pristup koji različite tipove informacija bez uključivanje dodatnih izvora informacija osim proteinske sekvence.

PAAC4 i DP_PAAC5 reprezentacije proteinskih sekvenci su uključivali različite grupe atributa zasnovanih na sekvenci proteina. U sklopu ovih modela su sadržane informacije o sastavu proteinske sekvence, kao i o međusobnim odnosima između udaljenih delova sekvence. Za predstavljanje znatno većeg broja proteina na nivou ljudskog proteoma upotrebljena je PCA_AAC reprezentacija. U odnosu na PAAC4 i PAAC5 koji su specifično modifikovani za predstavljanje dve grupe proteina, PCA_AAC reprezentacija obuhvata širi spektar različitih proteina ljudskog proteoma. PCA_AAC poput PAAC4 i DP_PAAC5 reprezentacija sadrži AAC predstavljanje proteinskih sekvenci za koje je pokazan visok potencijal kod formiranja prediktivnih modela IPP (Roy *et al.*, 2009). Umesto četiri ili pet osobina aminokiselina iz AAindex baze podataka, koje retrospektivno sadrže PAAC4 i PAAC5 modeli, PCA_AAC model proteinskih sekvenci obuhvata čitavu AAindex bazu podataka. Da bi se izbegla zamka prevelike veličine vektora kojim se IPP predstavljaju, upotrebljena je PCA tehnika za ekstrakciju varijabilnosti AAindeks baze podataka i njeno pretvaranje u nekoliko sintetičkih varijabli. PCA ima široku primenu u bioinformatici najčešće u svrhu redukcije broja attributa korišćenih za formiranje prediktivnih modela (You *et al.*, 2013; Mahmoudian, 2015; Perez-Riverol *et al.*, 2017). U cilju očuvanja informacija o redosledu aminokiselina proteina varijabilne dužine i predstavljanja u formi vektora jednakih dužina, kod PCA_AAC je upotrebljena auto kros-korelaciona funkcija. Broj sintetičkih varijabli i maksimalna dužina sekvence u čijem okviru bi se posmatrali odnosi između karakteristika aminokiselina su prilagođeni što informaciono bogatijem predstavljanju proteoma čoveka sa jedne strane, i potrebi za minimizacijom računarske zahtevnosti sa druge strane. Optimizacija rešenja predstavljenih u PCA_AAC modelu omogućila je da se očuva informativnost 532 osobine AAindex baze podataka i informacija o poretku aminokiselina proteina u vektoru dužine 60. Analiza prve dve PCA komponente kao nove sintetičke karakteristike aminokiselina upotrebljene u PCA_AAC modelu, ukazuje da: (i) prve dve komponente PCA predstavljaju 50% varijabilnosti 532 posmatrane osobine aminokiselina iz AAindeks baze podataka i (ii) od sedam grupa osobina aminokiselina (Tomii and Kanehisa, 1996), kod prve i najvažnije komponente, se ističu hidrofobnost i sklonost pojedinih aminokiselina ka beta pločama (engl. *beta propensity),* dok kod formiranja druge komponente aminokiselinske karakteristike svih sedam grupa imaju podjednak uticaj. Dok je značaj hidrofobnosti za formiranje IPP dobro poznat u literaturi,

značajan uticaj sklonosti beta pločama u formiranju prve sintetičke varijable može ukazivati na visok stepen važnosti ove osobine u predviđanju IPP interakcija (Watkins and Arora, 2014).

Potencijal evolutivnih atributa za visoko efikasnu reprezentaciju proteina i njihovih IPP je prethodno potvrđen u nekoliko istraživanja (Hamp and Rost, 2015a; Li *et al.*, 2017; Wang *et al.*, 2017). Matrice substitucije aminokiselina u sekvenci proteina (PSSM) čuvaju informaciju evolutivne udaljenosti proteina sa ostalim proteinima ispitivane baze podataka. Mada PSSM matrice čine osnovu za formiranje evolutivnih atributa, proces ekstrakcije vektora jednakih dužina kojima se predstavljaju proteini na osnovu PSSM matrica, često predstavlja izazov. Hamp i Rost (Hamp and Rost, 2015a) formulišu kernel funkciju koja omogućava implicitno predstavljanje IPP bez direktnog računanja 16000 atributa. Računanje kernel funkcije je jedno od računarski najzahtevnijih procedura pri formiranju svakog SVM metoda. Osim toga ovakav pristup reprezentacije IPP može biti korišćen skoro isključivo sa SVM algoritmom ML. Modeliranje ovako velikog broja atributa u svrhu predviđanja IPP na nivou proteoma bi zahtevalo znatne računarske resurse. U sklopu ovog istraživanja smo stoga obezbedili korišćenje evolutivne informacije za različite algoritme ML, minimizacijom broja atributa na samo 40, za predstavljanje IPP (PSSM_AAC model). Jedna od važnih prednosti PSSM_AAC i PCA_AAC modela proteinskih sekvenci je upotreba samo informacije sadržane u sekvencama proteina kao osnov za formiranje klasifikatora IPP.

Analiza bioloških mreža je omogućila uvid u ponašanje različitih bioloških procesa na sistemskom nivou. Sa druge strane, upotreba strukture mreža u predviđanju nedostajućih veza je posebno razvijena u socijalnim naukama, ali i drugim oblastima (Nowell *et al.*, 2003; Huang, 2010; Yang and Zhang, 2016; Ma, Bao and Zhang, 2017; Yin *et al.*, 2017). Analogno socijalnim mrežama, interaktom može biti predstavljen mrežom u kojoj su proteini predstavljeni čvorovima, a njihove međusobne interakcije granama. Nepostojeće veze unutar mreže mogu da predstavljaju potencijalne, a do sada nepoznate IPP. Zabeleženi su pokušaji pretpostavki nedostajućih veza (interakcija) među proteinima u sklopu IPP mreže, korišćenjem različitih pristupa na osnovu lokalne i globalne strukture IPP mreža (Hu *et al.*, 2011; Li, Liu and Burge, 2012; Han *et al.*, 2016; Luck *et al.*, 2018). Često pristupi koji upotrebljavaju mrežne algoritme za predviđanje nedostajućih veza, nisu kompatibilni sa algoritmima mašinskog učenja, te se moraju

koristiti odvojeno. U cilju prevazilaženja ovog problema u okviru ovog rada integrisali smo karakteristike mreže IPP na takav način da mogu biti korišćeni pri formiranju klasifikatora IPP zajedno sa drugim atributima. Da bi ostvarili taj cilj, bilo je neophodno ispuniti dva uslova: (i) naći efikasan način reprezentacije pojedinačnih proteina (čvorova mreže), koji čuvaju informacije o mrežnoj topologiji i (ii) koristiti samo informacije neophodne za formiranje prediktivnog modela IPP zasnovanog na sekvenci. Sekvence proteina i informacije o poznatim IPP čine osnovu formiranja skupa za treniranje. U tu svrhu, pri formiranju mreže IPP korišćene su samo poznate interakcije iz skupa za treniranje i samo one osobine mreže koje se mogu računati za pojedinačne čvorove (proteine). Formirani Graph_21 atributi su takođe i štedljivi u kontekstu računarskih resursa; IPP je predstavljena sa 42 mrežna atributa.

Kombinujući tri modela različitih tipova informacija, uključujući informacije o strukturi i redosledu aminokiselina, evolutivnim odnosima i strukturi mreža IPP, omogućeno je predstavljanje IPP sa vektorom dužine 202 atributa.

Dodavanje nove informacije u formi novih atributa ne garantuje poboljšanja prediktivnih sposobnosti klasifikatora, već može biti i kontraproduktivno usled povećanja veličine vektora i potencijalnih interakcija među atributima (Domingos, 2012). U svrhu pronalaženja optimalne kombinacije atributa, koriste se razne tehnike selekcije atributa (Saeys, Inza and Larranaga, 2007). Selekcija atributa može rezultovati značajnim povećanjem računarskih performansi, uz male gubitke u prediktivnim performansama klasifikatora, a u nekim slučajevima može dovesti i do njihovog povećanja. Pored upotrebe domenskog znanja i heuristike u konstrukciji atributa, kao i primeni tehnika selekcije atributa, jedan od značajnih postupaka generisanja efikasnog klasifikatora je i primena tehnika matematičke transformacije atributa. Ovaj postupak ima za cilj generisanje novih atributa koji pružaju veći stepen informativnosti algoritmu ML. Pored informativnosti, za visoke prediktivne performanse potrebno je da skup atributa sadrži atribute visoko korelisane sa klasom, a niskog stepena međusobne korelacije (Michalak and Kwasnicka, 2006).

Matematičke transformacije atributa se mogu aplicirati na pojedinačne atribute: originalne, formirane na osnovu domenskog znanja, ili prethodno generisane pojedinim transformacijama originalnih. Ovakve novogenerisane atribute označavamo unarnim atributima. Dor i Reich (Dor and Reich, 2012) su istakli prednosti aplikacije unarnih

transformacija u kreiranju novih unarnih atributa za proces formiranja modela ML. Pored transformacija pojedinačnih ulaznih atributa, transformacije mogu uključivati formiranje novih atributa matematičkim kombinovanjem dva ili više atributa. Ovi novoformirani atributi mogu da nose novu informaciju interakcije među dva (binarni atributi) ili više atributa. Izbor originalnih atributa i matematičkih transformacija koje bi se na njima primenile, često zavise od nivoa stručnosti istraživača i njegovih pretpostavki. Sa druge strane, računanje svake moguće kombinacije matematičkih transformacija između različitih atributa je računarski teško izvodljivo (binarni atributi). Čak i mali broj transformacija samo jednog ulaznog atributa može da dovede do generisanja ogromnog broja novih atributa (eksplozija atributa) (Khurana, Samulowitz and Turaga, 2017). Pored toga, kako bi se osiguralo da novi atributi ispunjavaju zahteve informativnosti, pored međusobne niske korelisanosti i visoke diskriminativne moći, neophodna je primena metoda selekcije atributa. Ove metode mogu biti eksplicitne (nadgledane), što uključuje modeliranje svake moguće kombinacije atributa uz ekstremnu računarsku zahtevnost, ili implicitne (nenadgledane), koje uz manji stepen računarske zahtevnosti pretpostavljaju efikasnost modela formiranog na datom podskupu atributa, a bez direktnog formiranja takvog modela.

Pokušaj rešavanja problema generisanja novih atributa, uz minimizaciju računarskih resursa, kao i uz računarski efikasnu selekciju atributa bez oslanjanja na heuristiku ili stručnost istraživača, dovelo je do razvoja GAFT algoritma za automatsko generisanje i selekciju (konstrukciju, engl. engineering) atributa. Upotrebom tehnika optimizacije i paralelizacije, GAFT algoritam dozvoljava povećanje broja ulaznih atributa, bez povećanja računarskih resursa. Dodatno, kombinovanjem nenadgledane i nadgledane selekcije GA algoritmom, omogućena je selekcija optimalnog podskupa atributa bez značajnog povećanja neophodnog vremena za testiranje i računarskih resursa. Upotrebom GAFT algoritma procesuirali smo 43083 originalnih i automatski generisanih atributa od kojih je konačno izabran podskup atributa sadržao svega 2.3‰ tog broja. Proces selekcije atributa omogućio je zamenu nekih od originalnih atributa novoformiranim atributima, koji su u kombinaciji sa ostalima korisniji u kreiranju efikasnijeg modela za predviđanje IPP. Finalni model za predviđanje IPP čoveka je tako sadržao oko 40% novoformiranih atributa GAFT algoritmom. Fino podešavanje GAFT algoritma omogućava rešavanje optimizacionih problema pronalaženja maksimalno

efikasnog podskupa atributa u sklopu određenog vremenskog ograničenja ili ograničenja određenih računarskih resursa. Ovakav fleksibilan pristup omogućava potencijalnu integraciju novih tipova atributa, koji bi mogli dovesti do poboljšanja prediktivne sposobnosti modela IPP u budućnosti. Strukturni atributi za sada imaju ograničenu upotrebljivost u formiranju modela za predviđanje IPP na nivou proteoma. Samo 6500 ljudskih proteina ima rešenu 3D strukturu (Gaudet *et al.*, 2017). Ipak, povećanjem broja proteina opisane 3D strukture, u budućnosti je moguće iskoristiti demonstrirani potencijal koje atributi zasnovani na 3D strukturi imaju u predviđanju IPP (Wass *et al.*, 2011; Planas-Iglesias *et al.*, 2013).

Pored izbora i obrade adekvatnih podataka, kao osnova za formiranje prediktivnog modela i pronalaženje opšteg pristupa za predstavljanje proteinskih sekvenci, jedan od važnih faktora u generisanju efikasnog modela za predviđanje IPP je pronalaženje adekvatnog algoritma za učenje modela ML. U svrhu izbora najadekvatnijeg broja i kvaliteta IPP na osnovu kojih će se generisati konačni model IPP korišćen je GBM algoritam mašinskog učenja. GBM se pokazao superiornim u odnosu na performansama najbliži algoritam RF pri modeliranju IPP čoveka standardnim skupovima (Park and Marcotte, 2012). U odnosu na znatno robusniji RF na različitim tipovima podataka, GBM zahteva fino naštimavanje znatno većeg broja meta-parametara u odnosu na RF algoritam. Ansambl algoritmi poput GBM i RF su pokazali visok stepen efikasnosti kod modeliranja IPP transkripcionih faktora, PNTS kao i kod formiranja opšteg modela IPP čoveka u odnosu na ostale ispitivane algoritme u sklopu ove studije. Visok stepen efikasnosti ansambl algoritama mašinskog učenja potvrđen je i u rešavanju drugih klasifikacionih problema u bioinformatici (Yang *et al.*, 2010).

Formiranje ansambla raznih algoritama uključujući i one koji su u osnovi sami zasnovani na ansambliranju, može dovesti do superiornih performansi u odnosu na pojedinačni algoritam ansambla (Divina *et al.*, 2018). Izbor algoritama koji će se ansamblirati, tehnika ansambliranja kao i podešavanje meta-parametara pojedinačnih algoritama ansambla je zahtevan zadatak, a ishod često zavisi od vremena uloženog u testiranje i stručnosti istraživača. Formiranje svih mogućih kombinacija meta-parametara, čak i nekoliko algoritama sa ciljem optimizacije konačnog ansambla, računarski je teško izvodljiva. U cilju automatizacije čitavog procesa razvili smo GA-STACK algoritam za automatsko ansambliranje i optimizaciju modela mašinskog učenja. GA-STACK se

zasnova na nasumičnom formiranju velikog broja klasifikatora, nasumičnim uzorkovanjem meta-parametara šest algoritama mašinskog učenja i supervizovanom selekcijom GA algoritmom superiorne kombinacije modela koja se koristi kao ulaz za generisanje meta-modela. Testiranje GA-STACK algoritma pokazalo je da je ovakav način ansambliranja heterogenih modela superiorniji od bilo kog pojedinačnog modela korišćenog za formiranje ansambla. Promovisanjem nekorelisanosti među formiranim modelima omogućava se da svaki pojedinačni model u sklopu ansambla doprinese meta-modelu različitom informacijom. GA-STACK omogućuje potpunu automatizaciju čitavog procesa, a obezbeđuje i odvajanje efikasnosti meta-modela od ekspertize istraživača. Algoritam je podešen da obezbeđuje generisanje optimalnog modela u definisanom vremenskom okviru ili sa ograničenim resursima.

Kombinovanjem GAFT i GA-STACK algoritama u jedinstveno okruženje, HP-GAS metod, omogućili smo odvajanje procesa konstrukcije atributa od formiranja konačnog modela. State-of-the-art algoritmi konstrukcije atributa, poput ExploreKit (Katz, Shin and Song, 2017; Kaul, Maheshwary and Pudi, 2017), ne obuhvataju proces modeliranja. ProFET algoritam (Ofer and Linial, 2015) kombinuje ova dva pristupa na način gde je sam proces konstrukcije atributa pod uticajem ručne optimizacije meta-parametara algoritama koji se koristi za nadgledanu selekciju. Odvajanje ova dva procesa u sklopu HP-GAS algoritma omogućava potpunu automatizaciju i kontrolu oba procesa, uz minimizaciju utroška računarskih resursa. Korišćenje GA-STACK algoritma nam je omogućilo odvojeno formiranje modela na pojedinačnim grupama atributa: evolutivnim, mrežnim i atributima zasnovanim na sekvenci. U pogledu efikasnosti različitih elementa HP-GAS protokola, GAFT algoritm ima najznačajniji relativni doprinos. Ovim se dodatno potvrđuje u literaturi opisani značaj postupaka generisanja i selekcije atributa u procesu formiranja efikasnog prediktora (John, Kohavi and Pfleger, 1994).

Jedan od ključnih elemenata u sklopu razvoja metoda za predviđanje IPP zasnovanog na ML je adekvatna evaluacija i poređenje sa dostupnim metodama. Jedan od pristupa zahteva formiranje skupova za treniranje i testiranje i implementaciju drugih metoda na ovim skupovima. Drugi način poređenja uključuje evaluaciju razvijenog metoda na prethodno konstruisanim standardnim skupovima, koji su se koristili za evaluaciju metoda za predikciju IPP. Ipak, izbor skupova za realističnu procenu prediktivnih performansi metoda baziranog na mašinskom učenju ostaje jedan od važnih

izazova. U ovome, tri elementa su primarna: broj instanci unutar skupova za testiranje i treniranja, ukupan broj skupova za evaluaciju i njihove karakteristike. Neki algoritmi mašinskog učenja zahtevaju veći broj instanci u sklopu skupa za treniranje u odnosu na druge. Za uspešnu generalizaciju algoritmi bazirani ne neuronskim mrežama uglavnom zahtevaju veći broj instanci u skupu za trening, u praksi često 100 hiljada ili više (Gibson and Patterson, 2017). Sa druge strane, algoritmi poput SVM omogućavaju formiranje efikasnih modela i sa nekoliko desetina primera u skupu za trening (Burges, 1998; Scholkopf and Smola, 2001). Iako u principu veći broj uzorka instanci u skupovima za treniranje i testiranje omogućava formiranje boljih prediktivnih modela i pouzdanije evaluacije, usled boljeg predstavljanja čitave populacije, optimalan broj instanci je različit za različite metode mašinskog učenja (Figueroa *et al.*, 2012). Opsežna testiranja algoritama mašinskog učenja na velikom broju skupova su pokazala da čak i algoritmi koji u proseku omogućuje najbolje prediktivne performanse, na određenim skupovima demonstriraju skoro optimalnu tačnost (Wainer, 2016). Ovo se tumači boljom generalizacijom karakteristika specifičnih skupova na kojima se trenira (Domingos, 2012).

U svrhu što objektivnijeg poređenja iskoristili smo oba pristupa. U prvom redu poređenja na 40 standardnih skupova IPP čoveka (uPK skupova) (Park and Marcotte, 2012), demonstrirana je dominantnost HP-GAS pristupa u odnosu na standardne metode za predviđanje IPP. Sve metode sa kojima smo poredili naš pristup su zasnovane na SVM algoritmu. Metode M1 i M2 upotrebljavaju specifičnu k-mer reprezentaciju proteinskih sekvenci sa razlikom u kernel funkciji. State-of-the-art metoda M3 koristi posebno razvijenu kernel funkciju kako bi omogućila efikasniju reprezentaciju IPP evolutivnim atributima proteina (Hamp and Rost, 2013). Ovo ukazuje na neophodnost razvoja i finog podešavanja svakog pojedinog elementa u procesu generisanja efikasnog modela za predviđanje IPP. Pored poređenja na IPP čoveka, u cilju ispitivanja konzistentnosti rezultata na drugom proteomu, HP-GAS pristup smo poredili sa M1, M2 i M3 metodama na IPP kvasca (*Saccharomyces cerevisiae*). Demonstracija poboljšanih prediktivnih sposobnosti našeg pristupa i u ovom slučaju ukazuje na potencijalnu sposobnost HP-GAS algoritma da uspešno generalizuje modele i na drugim organizmima. Kako je naš finalni model formiran na izabranom skupu HIPPIE baze podataka, drugi pristup uporednom testiranju uključivao je poređenje performansi HP-GAS algoritma i state-of-the-art M3

pristupa na C50' skupu IPP. I u ovom slučaju pokazana je veća prediktivna tačnost HP-GAS metoda. Dodatno poređenje dva pristupa je izvršeno na HIPPE_new skupu formiranom na osnovu razlike između dve verzije HIPPE baze podataka. HP-GAS model se pokazao sposobnim da tačnije predviđa eksperimentalno potvrđene IPP integrisane u noviju verziju HIPPIE baze podataka. Pored ovoga HP-GAS pristupom je predviđen veći ukupan broj IPP kao i veći broj IPP koje M3 pristup nije predvideo.

HP-GAS pristup primenili smo u cilju pretrage potencijalnih interaktora EGFR proteina. EGFR je transmembranski receptor tirozinska kinaza, koji se povezuje sa nizom ljudskih oboljenja, u prvom redu karcinomom pluća (Paez *et al.*, 2004; Maemondo *et al.*, 2010). Među visoko rangiranim interaktorima EGFR pronašli smo SUMO-1, pri čemu njihova interakcija mada poznata u literaturi, nije bila uključena u skup za treniranje finalnog modela HP-GAS pristupa (Packham *et al.*, 2015). Pored ovoga, pretpostavljene su interakcije sa SUMO 2, 3 i 4 regulatornim faktorima. Ove pretpostavljene interakcije mogu činiti osnovu detaljnijeg ispitivanje konekcija između ovih proteina, što može biti relevantno za onkogenu aktivnost EGFR proteina.

Jedan od važnih izazova je formirati takav metod za predviđanje koji se može predstaviti u formi efikasnog alata prilagođenog korisniku, a istovremeno zadržati visok stepen prediktivnih performansi i brzine predviđanja. State-of-the-art metoda PPI-PK (M3) (Hamp and Rost, 2015a) kao i metode M1 (Martin, Roe and Faulon, 2005) i M2 (Park and Marcotte, 2012) su efikasne metode za predviđanje IPP. Za praktičnu primenu, M3 i M1 metode zahtevaju prevođenje originalnog i dostupnog koda koji u slučaju M1 zahteva pronalaženje i instalaciju dodatnog softvera. U slučaju M2 dostupne su samo informacije o tipu algoritma ML i tipa reprezentacije proteinskih sekvenci. Dodatno, upotreba bioinformatičkog softvera je često ograničena ne samo dostupnošću i softverskim ograničenjima, već i nedostatkom adekvatne dokumentacije neophodne za instalaciju i korišćenje (Seemann, 2013; Karimzadeh and Hoffman, 2017). Instalacija i korišćenje bioinformatičkog softvera često zahteva tehničke i veštine programiranja koje mogu umanjiti upotrebnu vrednost predstavljenog alata istraživačima. HP-GAS algoritam je dizajniran kako bi se mogao implementirati u vidu HP-GAS korisnički-orijentisanog alata, omogućavajući širok stepen dostupnosti i jednostavnosti korišćenja. U cilju brze i jednostavne upotrebe, HP-GAS alat za prioritizaciju IPP kandidata za eksperimentalno istraživanje IPP, sadrži sledeće karakteristike: (i) prekompajliran alat koji ne zahteva

dodatnu instalaciju, (ii) korisnički-orijentisano okruženje sa primerima upotrebe na različitim operativnim sistemima i (iii) jasna i iscrpna dokumentacija sa uputstvom za korišćenje, objašnjenjima uzroka mogućih grešaka pri korišćenju, primerima ulaznih podataka i uputstva za tumačenje izlaznih rezultata.

6 Zaključak

U ovoj disertaciji razvijeni su i prikazani sledeći pristupi i bioinformatički modeli za automatsko predviđanje interakcija između proteina kod čoveka:

1) Automatsku pristup za predviđanje IPP transkripcionih regulatora koji je: (i) baziran na sekvenci proteina, (ii) razvijen korišćenjem tehnike mašinskog učenja, (iii) prediktivno i računarski efikasan, (iv) u poređenje sa standardnim modelima za predikciju IPP baziranim na sekvenci proteina, pokazao prednosti u vidu bolje tačnosti predviđanja novih interakcija između transkripcionih regulatora uz manji utrošak računarskih resursa i (v) adaptiran u formu efikasnog veb alata.

2) Predikcioni metod za predviđanje interakcija proteina sa neuređenom tercijarnom strukturom koji: (i) uključuje različite informacije o strukturi sekvence i fizičko-hemijskim karakteristikama aminokiselina, (ii) kreiran izborom optimalnog algoritma mašinskog učenja, (iii) demonstrirao veći stepen tačnosti u predikciji interakacija proteina neuređene tercijarne strukture u odnosu na standardne metode za predikciju IPP i (iv) implementiran u formi široko dostupnog veb alata za brzo i efikasno predviđanje novih interaktora proteina sa neuređenom tercijarnom strukturom.

3) Metod za predviđanje IPP na nivou proteoma, uz koji su razvijeni: (i) pouzdan i izbalansiran skup IPP veličine neophodne za treniranje efikasnog modela, (ii) tri nove grupe atributa, atributi bazirani na sekvenci proteina, evolutivnim profilima i topološkim karakteristikama mreže IPP, sa dovoljnom informacijom za njihovo formiranje sadržanom u sekvenci proteina i identitetu njihovih interaktora, (iii) računarsko i prediktivno efikasni algoritmi za automatsko generisanje i selekciju atributa, kao i za supervizovano spajanje više modela mašinskog učenja, (iv) integrisani opšti model HP-GAS za automatsko mapiranje IPP na nivou proteoma čoveka koji je prikazao veći stepen tačnosti predikcije u odnosu na state-of-the-art metode za predikciju IPP, sa efikasnošću predviđanja AUC= 0.93 i 0.85 tačnosti i (v) implemetacija u formi samostalnog softverskog alata koji se izdvaja većim stepenom prilagođenosti korisniku i jednostavnošću instalacije i upotrebe.

Razvijeni bioinformatički metodi predstavljeni u ovoj disertaciji, kao suplementi eksperimentalnim metodama za analizu IPP, omogućavaju: (i) bolje razumevanje

kompleksnih molekularnih odnosa unutar živih sistema, nove uvide u molekularne mehanizme i razumevanje bioloških procesa i funkcija, (ii) rasvetljavanje raznih patoloških stanja ljudskog organizma, (iii) lakše otkrivanje potencijalnih meta za ciljanu terapiju u procesu nalaženja lekova, (iv) nova saznanja o signalnim putevima i otkrivanju mogućih terapeutskih meta, predviđanjem novih interaktora transkripcionih regulatora i (v) smanjenje vremena i troškova razvoja novih terapeutika.

7 Literatura

Abu-Farha, M., Elisma, F. and Figeys, D. (2008) ‘Identification of Protein–Protein Interactions by Mass Spectrometry Coupled Techniques’, in *Advances in biochemical engineering/biotechnology*, pp. 67–80. doi: 10.1007/10_2007_091.

Afanasyeva, A. *et al.* (2018) ‘Human long intrinsically disordered protein regions are frequent targets of positive selection’, *Genome Research*, 28(7), pp. 975–982. doi: 10.1101/gr.232645.117.

Agresti, A. (2003) *Categorical data analysis*. John Wiley & Sons.

Ahmed, C., Elkorany, A. and Bahgat, R. (2016) ‘A supervised learning approach to link prediction in Twitter’, *Social Network Analysis and Mining*. Springer Vienna, 6(1), pp. 1–11. doi: 10.1007/s13278-016-0333-1.

Aho, A. V., Kernighan, B. W. and Weinberger, P. J. (1979) ‘Awk — a pattern scanning and processing language’, *Software: Practice and Experience*, 9(4), pp. 267–279. doi: 10.1002/spe.4380090403.

Aittokallio, T. and Schwikowski, B. (2006) ‘Graph-based methods for analysing networks in cell biology.’, *Briefings in bioinformatics*, 7(3), pp. 243–55. doi: 10.1093/bib/bbl022.

Alanis-Lobato, G., Andrade-Navarro, M. A. and Schaefer, M. H. (2017) ‘HIPPIE v2.0: Enhancing meaningfulness and reliability of protein-protein interaction networks’, *Nucleic Acids Research*, 45(D1), pp. D408–D414. doi: 10.1093/nar/gkw985.

Albert, R. (2007) ‘Network inference, analysis, and modeling in systems biology.’, *The Plant cell*, 19(11), pp. 3327–3338. doi: 10.1105/tpc.107.054700.

Aloy, P. and Russell, R. B. (2003) ‘InterPreTS : protein Interaction Prediction through Tertiary Structure’, 19(1), pp. 161–162.

Aloy, P. and Russell, R. B. (2006) ‘Structural systems biology: modelling protein interactions’, *Nature Reviews Molecular Cell Biology*. Nature Publishing Group, 7(3), p. 188.

Altschul, S. F. (1997) ‘Gapped BLAST and PSI-BLAST: a new generation of protein database search programs’, *Nucleic Acids Research*. Oxford University Press, 25(17), pp. 3389–3402. doi: 10.1093/nar/25.17.3389.

Altschul, S. F. and Koonin, E. V (1998) ‘Iterated profile searches with PSI-BLAST—a tool for discovery in protein databases’, *Trends in Biochemical Sciences*. Elsevier, 23(11), pp. 444–447. doi: 10.1016/S0968-0004(98)01298-5.

Angenendt, P. *et al.* (2006) ‘Generation of High Density Protein Microarrays by Cell-free *in Situ* Expression of Unpurified PCR Products’, *Molecular & Cellular Proteomics*, 5(9), pp. 1658–1666. doi: 10.1074/mcp.T600024-MCP200.

Arndt, H. D. (2006) ‘Small molecule modulators of transcription’, *Angewandte Chemie - International Edition*. doi: 10.1002/anie.200600285.

Arndt, V. and Vorberg, I. (2012) ‘Defining the Cellular Interactome of Disease-Linked Proteins in Neurodegeneration’, in *Protein Interactions*. InTech. doi: 10.5772/37751.

Asthana, S. *et al.* (2004) ‘Predicting protein complex membership using probabilistic network reliability’, *Genome Research*. doi: 10.1101/gr.2203804.

Auerbach, D. *et al.* (2002) ‘The post-genomic era of interactive proteomics: Facts and perspectives’, *PROTEOMICS*, 2(6), pp. 611–623. doi: 10.1002/1615-9861(200206)2:6<611::AID-PROT611>3.0.CO;2-Y.

Barabási, A.-L. and Oltvai, Z. N. (2004) 'Network biology: understanding the cell's functional organization.', *Nature reviews genetics*. Nature Publishing Group, 5(2), pp. 101–13. doi: 10.1038/nrg1272.

Barabási, A. L. and Albert, R. (1999) 'Emergence of scaling in random networks', *Science*. doi: 10.1126/science.286.5439.509.

Barnes, J. A. and Harary, F. (1983) 'Graph theory in network analysis', *Social Networks*. doi: 10.1016/0378-8733(83)90026-6.

Barrat, A. *et al.* (2004) 'The architecture of complex weighted networks', *Proceedings of the national academy of sciences*. National Acad Sciences, 101(11), pp. 3747–3752.

Ben-Hur, A. and Noble, W. S. (2005) 'Kernel methods for predicting protein-protein interactions.', *Bioinformatics (Oxford, England)*, 21 Suppl 1(SUPPL. 1), pp. i38-46. doi: 10.1093/bioinformatics/bti1016.

Ben-Hur, A. and Noble, W. S. (2006) 'Choosing negative examples for the prediction of protein-protein interactions.', *BMC bioinformatics*, 7 Suppl 1, p. S2. doi: 10.1186/1471-2105-7-S1-S2.

Benchettara, N., Kanawati, R. and Rouveirol, C. (2010) 'Supervised machine learning applied to link prediction in bipartite social networks', *Proceedings - 2010 International Conference on Advances in Social Network Analysis and Mining, ASONAM 2010*, pp. 326–330. doi: 10.1109/ASONAM.2010.87.

Bender, A. and Pringle, J. R. (1991) 'Use of a screen for synthetic lethal and multicopy suppressee mutants to identify two new genes involved in morphogenesis in Saccharomyces cerevisiae.', *Molecular and cellular biology*, 11(3), pp. 1295–305.

Benson, D. *et al.* (1990) 'The National Center for Biotechnology Information', *Genomics*. doi: 10.1016/0888-7543(90)90583-G.

Bergstra, J. and Bengio, Y. (2012) 'Random Search for Hyper-Parameter Optimization', *Journal of Machine Learning Research*, 1(1), pp. 281–305. doi: 10.1162/153244303322533223.

Bernard, P. and Harley, V. R. (2010) 'Acquisition of SOX transcription factor specificity through protein-protein interaction, modulation of Wnt signalling and post-translational modification', *International Journal of Biochemistry and Cell Biology*. doi: 10.1016/j.biocel.2009.10.017.

Bhagwat, M. and Aravind, L. (2007) 'PSI-BLAST tutorial', *Methods in Molecular Biology*. doi: 10.1385/1-59745-514-8:177.

Bhaskar, H., Hoyle, D. C. and Singh, S. (2006) 'Machine learning in bioinformatics: A brief survey and recommendations for practitioners', *Computers in Biology and Medicine*, 36(10), pp. 1104–1125. doi: 10.1016/j.compbiomed.2005.09.002.

Binns, D. *et al.* (2009) 'QuickGO: a web-based tool for Gene Ontology searching', *Bioinformatics*, 25(22), pp. 3045–3046. doi: 10.1093/bioinformatics/btp536.

Bjellqvist, B. *et al.* (1982) 'Isoelectric focusing in immobilized pH gradients: principle, methodology and some applications.', *Journal of biochemical and biophysical methods*, 6(4), pp. 317–39.

Blake, J. A. *et al.* (2015) 'Gene ontology consortium: Going forward', *Nucleic Acids Research*, 43(D1), pp. D1049–D1056. doi: 10.1093/nar/gku1179.

Blohm, P. *et al.* (2014) 'Negatome 2.0: a database of non-interacting proteins derived by literature mining, manual annotation and protein structure analysis.', *Nucleic acids research*, 42(Database issue), pp. D396-400. doi: 10.1093/nar/gkt1079.

Blumenthal, T. (1998) 'Gene clusters and polycistronic transcription in eukaryotes',

Bioessays. Wiley Online Library, 20(6), pp. 480–487.

Bonacich, P. (2002) 'Power and Centrality: A Family of Measures', *American Journal of Sociology*. doi: 10.1086/228631.

Bonacich, P. and Lloyd, P. (2001) 'Eigenvector-like measures of centrality for asymmetric relations', *Social Networks*, 23(3), pp. 191–201. doi: 10.1016/S0378-8733(01)00038-7.

Bonvin, A. M., Boelens, R. and Kaptein, R. (2005) 'NMR analysis of protein interactions', *Current Opinion in Chemical Biology*, 9(5), pp. 501–508. doi: 10.1016/j.cbpa.2005.08.011.

Breiman, L. (2001) 'Random forests', *Machine learning*, pp. 5–32. doi: 10.1023/A:1010933404324.

Brin, S. and Page, L. (2012) 'Reprint of: The anatomy of a large-scale hypertextual web search engine', *Computer Networks*. doi: 10.1016/j.comnet.2012.10.007.

Browne, F. *et al.* (2010a) 'From Experimental Approaches to Computational Techniques: A Review on the Prediction of Protein-Protein Interactions', *Advances in Artificial Intelligence*. Hindawi, 2010, pp. 1–15. doi: 10.1155/2010/924529.

Browne, F. *et al.* (2010b) 'From Experimental Approaches to Computational Techniques: A Review on the Prediction of Protein-Protein Interactions', *Advances in Artificial Intelligence*, 2010, pp. 1–15. doi: 10.1155/2010/924529.

Brun, C. *et al.* (2003) 'Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network.', *Genome Biology*, 5(1), p. R6. doi: 10.1186/gb-2003-5-1-r6.

Burges, C. J. C. (1998) 'A tutorial on support vector machines for pattern recognition',

Data mining and knowledge discovery. Springer, 2(2), pp. 121–167.

Burt, R. S. (2004) 'Structural Holes and Good Ideas', *American Journal of Sociology*. doi: 10.1086/421787.

Büyükköroğlu, G. *et al.* (2018) 'Techniques for Protein Analysis', *Omics Technologies and Bio-Engineering*. Academic Press, pp. 317–351. doi: 10.1016/B978-0-12-804659-3.00015-4.

Calderone, A., Castagnoli, L. and Cesareni, G. (2013) 'mentha: a resource for browsing integrated protein-interaction networks', *Nature Methods*. doi: 10.1038/nmeth.2561.

Campen, A. *et al.* (2008) 'TOP-IDP-Scale: A New Amino Acid Scale Measuring Propensity for Intrinsic Disorder', *Protein & Peptide Letters*, 15(9), pp. 956–963. doi: 10.2174/092986608785849164.

Carbon, S. *et al.* (2009) 'AmiGO: Online access to ontology and annotation data', *Bioinformatics*, 25(2), pp. 288–289. doi: 10.1093/bioinformatics/btn615.

Carpenter, B. *et al.* (2004) 'BASP1 is a transcriptional cosuppressor for the Wilms' tumor suppressor protein WT1.', *Molecular and Cellular Biology*, 24(2), pp. 537–49. doi: 10.1128/MCB.24.2.537–549.2004.

Chan, S.-K. *et al.* (1996) 'An extradenticle-induced conformational change in a HOX protein overcomes an inhibitory function of the conserved hexapeptide motif', *The EMBO Journal*.

Chang, C. and Lin, C. (2013) 'LIBSVM : A Library for Support Vector Machines', *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2, pp. 1–39. doi: 10.1145/1961189.1961199.

Chatr-aryamontri, A. *et al.* (2008) 'Protein interactions: integration leads to belief',

Trends in Biochemical Sciences, 33(6), pp. 241–242. doi: 10.1016/j.tibs.2008.04.002.

Chen, L. *et al.* (2015) 'Discovery of new candidate genes related to brain development using protein interaction information', *PLoS ONE*. doi: 10.1371/journal.pone.0118003.

Chen, T. *et al.* (2015) 'MXNet: A Flexible and Efficient Machine Learning Library for Heterogeneous Distributed Systems', pp. 1–6. doi: 10.1145/2532637.

Chen, T. and Guestrin, C. (2016) 'XGBoost: A Scalable Tree Boosting System', *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*. New York, New York, USA: ACM Press, pp. 785–794. doi: 10.1145/2939672.2939785.

Chicco, D. (2017) 'Ten quick tips for machine learning in computational biology', *BioData Mining*. BioData Mining, 10(1), pp. 1–17. doi: 10.1186/s13040-017-0155-3.

Chin, L. *et al.* (2011) 'Making sense of cancer genomic data', *Genes and Development*. doi: 10.1101/gad.2017311.

Chou, K.-C. (2009) 'Pseudo Amino Acid Composition and its Applications in Bioinformatics, Proteomics and System Biology', *Current Proteomics*, 6(November), pp. 262–274. doi: 10.2174/157016409789973707.

Chou, K. C. (2001) 'Prediction of protein cellular attributes using pseudo-amino acid composition.', *Proteins*, 43(3), pp. 246–55. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11288174.

Chou, K. C. and Cai, Y. D. (2006) 'Predicting protein-protein interactions from sequences in a hybridization space', *Journal of Proteome Research*, 5(2), pp. 316–322. doi: 10.1021/pr050331g.

Chua, H. N., Sung, W. K. and Wong, L. (2006) 'Exploiting indirect neighbours and

topological weight to predict protein function from protein-protein interactions', *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 3916 LNBI(13), p. 1. doi: 10.1007/11691730_1.

Collins, F. S. *et al.* (2004) 'Finishing the euchromatic sequence of the human genome', *Nature*. doi: 10.1038/nature03001.

Conover, W. J. (1982) 'A Distribution-Free Approach to Inducing Rank Correlation Among Input Variables', *Communications in Statistics - Simulation and Computation*. doi: 10.1080/03610918208812265.

Consortium, T. U. (2015) 'UniProt: a hub for protein information', *Nucleic Acids Research*, 43(D1), pp. D204–D212. doi: 10.1093/nar/gku989.

Cook, D. (2016) *Practical Machine Learning with H2O: Powerful, Scalable Techniques for Deep Learning and AI*. Available at: https://books.google.com/books?hl=en&lr=&id=nJWmDQAAQBAJ&oi=fnd&pg=PP1&dq=Practical+Machine+Learning+with+H2O&ots=9v8d5pRyMu&sig=k7zFUm_uxz3BJHyhc1c7_2UJGDU (Accessed: 23 November 2018).

Coulon, A. *et al.* (2013) 'Eukaryotic transcriptional dynamics: From single molecules to cell populations', *Nature Reviews Genetics*. Nature Publishing Group, 14(8), pp. 572–584. doi: 10.1038/nrg3484.

Craig Venter, J. *et al.* (2001) 'The sequence of the human genome', *Science*. doi: 10.1126/science.1058040.

Csardi, G. and Nepusz, T. (2006) 'The igraph software package for complex network research', *InterJournal, Complex Systems*, 1695(5), pp. 1–9. doi: 10.4236/jsea.2012.54028.

Csermely, P. *et al.* (2013) ‘Structure and dynamics of molecular networks: A novel paradigm of drug discovery’, *Pharmacology & Therapeutics*, 138(3), pp. 333–408. doi: 10.1016/j.pharmthera.2013.01.016.

Cusick, M. E. *et al.* (2009) ‘Literature-curated protein interaction datasets’, *Nature Methods*, 6(1), pp. 39–46. doi: 10.1038/nmeth.1284.

Dandekar, T. *et al.* (1998) ‘Conservation of gene order: a fingerprint of proteins that physically interact’, *Trends in biochemical sciences*. Elsevier, 23(9), pp. 324–328.

Davis, J. and Goadrich, M. (2006) ‘The Relationship Between Precision-Recall and ROC Curves’, *Proceedings of the 23rd International Conference on Machine learning - - ICML’06*, pp. 233–240. doi: 10.1145/1143844.1143874.

Dehzangi, A. *et al.* (2017) ‘PSSM-Suc: Accurately predicting succinylation using position specific scoring matrix into bigram for feature extraction’, *Journal of Theoretical Biology*, 425, pp. 97–102. doi: 10.1016/j.jtbi.2017.05.005.

Deng, M. *et al.* (2002) ‘Inferring domain-domain interactions from protein-protein interactions’, in *Proceedings of the sixth annual international conference on Computational biology*. ACM, pp. 117–126.

Deng, M., Sun, F. and Chen, T. (2002) ‘Assessment of the reliability of protein-protein interactions and protein function prediction’, in *Biocomputing 2003*. World Scientific, pp. 140–151.

Deng, X., Eickholt, J. and Cheng, J. (2012) ‘A comprehensive overview of computational protein disorder prediction methods’, *Molecular BioSystems*. doi: 10.1039/c1mb05207a.

Deng, Y., Gao, L. and Wang, B. (2013) ‘ppiPre: predicting protein-protein interactions by combining heterogeneous features.’, *BMC systems biology*, 7 Suppl 2(Suppl 2), p.

S8. doi: 10.1186/1752-0509-7-S2-S8.

Deplancke, B. *et al.* (2006) ‘A Gene-Centered C. elegans Protein-DNA Interaction Network’, *Cell*. doi: 10.1016/j.cell.2006.04.038.

Diestel, R. (2018) *Graph theory*. Springer Publishing Company, Incorporated.

Divina, F. *et al.* (2018) ‘Stacking Ensemble Learning for Short-Term Electricity Consumption Forecasting’, *Energies*, 11(4), p. 949. doi: 10.3390/en11040949.

Domingos, P. (2012) ‘A few useful things to know about machine learning’, *Communications of the ACM*, 55(10), p. 78. doi: 10.1145/2347736.2347755.

Dong, Q., Zhou, S. and Liu, X. (2010) ‘Prediction of protein protein interactions from primary sequences’, *International Journal of Data Mining and Bioinformatics*, 4(2), p. 211. doi: 10.1504/IJDMB.2010.032151.

Dor, O. and Reich, Y. (2012) ‘Strengthening learning algorithms by feature discovery’, *Information Sciences*. Elsevier Inc., 189, pp. 176–190. doi: 10.1016/j.ins.2011.11.039.

Du, X. *et al.* (2014) ‘A Novel Feature Extraction Scheme with Ensemble Coding for Protein–Protein Interaction Prediction’, *International journal of molecular sciences*, 15(7), pp. 12731–12749. doi: 10.3390/ijms150712731.

Du, X. *et al.* (2017) ‘DeepPPI: Boosting Prediction of Protein-Protein Interactions with Deep Neural Networks’, *Journal of Chemical Information and Modeling*, 57(6), pp. 1499–1510. doi: 10.1021/acs.jcim.7b00028.

Dunker, A. K. *et al.* (2013) ‘What’s in a name? Why these proteins are intrinsically disordered’, *Intrinsically Disordered Proteins*. doi: 10.4161/idp.24157.

Dunker, A. K. and Uversky, V. N. (2010) ‘Drugs for “protein clouds”: Targeting

intrinsically disordered transcription factors', *Current Opinion in Pharmacology*. Elsevier Ltd, 10(6), pp. 782–788. doi: 10.1016/j.coph.2010.09.005.

Efron, B. (1983) 'Estimating the error rate of a prediction rule: improvement on cross-validation', *Journal of the American statistical association*. Taylor & Francis, 78(382), pp. 316–331.

Eifler, K. and Vertegaal, A. C. O. (2015) 'SUMOylation-mediated regulation of cell cycle progression and cancer', *Trends in biochemical sciences*. Elsevier, 40(12), pp. 779–793.

Eisenberg, E. and Levanon, E. Y. (2003) 'Preferential attachment in the protein network evolution.', *Physical review letters*, 91(13), p. 138701. doi: 10.1103/PhysRevLett.91.138701.

Elefsinioti, A. *et al.* (2011) 'Large-scale De Novo Prediction of Physical Protein-Protein Association', *Molecular & Cellular Proteomics*, 10(11), p. M111.010629. doi: 10.1074/mcp.M111.010629.

Elson, E. L. (2011) 'Fluorescence correlation spectroscopy: past, present, future.', *Biophysical journal*. The Biophysical Society, 101(12), pp. 2855–70. doi: 10.1016/j.bpj.2011.11.012.

Elson, E. L. and Magde, D. (1974) 'Fluorescence correlation spectroscopy. I. Conceptual basis and theory', *Biopolymers*. John Wiley & Sons, Ltd, 13(1), pp. 1–27. doi: 10.1002/bip.1974.360130102.

Enright, A. J. *et al.* (1999) 'Protein interaction maps for complete genomes based on gene fusion events', *Nature*. Nature Publishing Group, 402(6757), p. 86.

Enright, A. J., Van Dongen, S. and Ouzounis, C. A. (2002) 'An efficient algorithm for large-scale detection of protein families', *Nucleic acids research*. Oxford University

Press, 30(7), pp. 1575–1584.

Erdős, P. and Rényi, A. (1959) ‘On random graphs 1.’, *Publ. Math. Debrecen*. doi: 10.2307/1999405.

Ermolaeva, M. D., White, O. and Salzberg, S. L. (2001) ‘Prediction of operons in microbial genomes’, *Nucleic acids research*. Oxford University Press, 29(5), pp. 1216–1221.

Fashena, S. J., Serebriiskii, I. and Golemis, E. A. (2000) ‘The continued evolution of two-hybrid screening approaches in yeast: how to outwit different preys with different baits.’, *Gene*, 250(1–2), pp. 1–14.

Fawcett, T. (2004) ‘ROC Graphs : Notes and Practical Considerations for Researchers’, *ReCALL*, 31(HPL-2003-4), pp. 1–38. doi: 10.1.1.10.9777.

Fields, S. and Song, O. (1989) ‘A novel genetic system to detect protein–protein interactions’, *Nature*. Nature Publishing Group, 340(6230), pp. 245–246. doi: 10.1038/340245a0.

Figueroa, R. L. *et al.* (2012) ‘Predicting sample size required for classification performance’, *BMC Medical Informatics and Decision Making*. BioMed Central Ltd, 12(1), p. 8. doi: 10.1186/1472-6947-12-8.

Fontaine, F. *et al.* (2017) ‘Small-Molecule Inhibitors of the SOX18 Transcription Factor’, *Cell Chemical Biology*. doi: 10.1016/j.chembiol.2017.01.003.

Fontaine, F., Overman, J. and François, M. (2015) ‘Pharmacological manipulation of transcription factor protein-protein interactions: opportunities and obstacles’, *Cell Regeneration*, 4(1), p. 2. doi: 10.1186/s13619-015-0015-x.

Forsova, O. S. and Zakharov, V. V. (2016) ‘High-order oligomers of intrinsically

disordered brain proteins BASP1 and GAP-43 preserve the structural disorder', *FEBS Journal*, 283(8), pp. 1550–1569. doi: 10.1111/febs.13692.

Francois, M., Donovan, P. and Fontaine, F. (2018) 'Modulating transcription factor activity: Interfering with protein-protein interaction networks', *Seminars in Cell and Developmental Biology*. Elsevier Ltd. doi: 10.1016/j.semcdb.2018.07.019.

Frank, J. (2002) 'Single-Particle Imaging of Macromolecules by Cryo-Electron Microscopy', *Annual Review of Biophysics and Biomolecular Structure*. Annual Reviews 4139 El Camino Way, P.O. Box 10139, Palo Alto, CA 94303-0139, USA , 31(1), pp. 303–319. doi: 10.1146/annurev.biophys.31.082901.134202.

Fraser, H. B. *et al.* (2004) 'Coevolution of gene expression among interacting proteins', *Proceedings of the National Academy of Sciences*, 101(24), pp. 9033–9038. doi: 10.1073/pnas.0402591101.

Fredriksson, S. *et al.* (2002) 'Protein detection using proximity-dependent DNA ligation assays', *Nature Biotechnology*, 20(5), pp. 473–477. doi: 10.1038/nbt0502-473.

Freeman, L. C. (1978) 'Centrality in social networks conceptual clarification', *Social Networks*. North-Holland, 1(3), pp. 215–239. doi: 10.1016/0378-8733(78)90021-7.

Friedman, J. H. (2001) 'Greedy function approximation: A gradient boosting machine.', *The Annals of Statistics*. JSTOR, 29(5), pp. 1189–1232. doi: 10.1214/aos/1013203451.

Friedman, J. H. (2002) 'Stochastic gradient boosting', *Computational Statistics & Data Analysis*. Elsevier, 38(4), pp. 367–378. doi: 10.1016/S0167-9473(01)00065-2.

Fu, L. *et al.* (2012) 'CD-HIT: accelerated for clustering the next-generation sequencing data.', *Bioinformatics (Oxford, England)*, 28(23), pp. 3150–3152. doi: 10.1093/bioinformatics/bts565.

Fukuhara, N. and Kawabata, T. (2008) ‘HOMCOS: a server to predict interacting protein pairs and interacting sites by homology modeling of complex structures’, *Nucleic acids research*. Oxford University Press, 36(suppl_2), pp. W185–W189.

Galperin, M. Y. and Koonin, E. V (2000) ‘Who’s your neighbor? New computational approaches for functional genomics’, *Nature biotechnology*. Nature Publishing Group, 18(6), p. 609.

Galzitskaya, O. V., Garbuzynskiy, S. O. and Lobanov, M. Y. (2006) ‘FoldUnfold: Web server for the prediction of disordered regions in protein chain’, *Bioinformatics*, 22(23), pp. 2948–2949. doi: 10.1093/bioinformatics/btl504.

Gandhi, T. K. B. *et al.* (2006) ‘Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets’, *Nature genetics*. Nature Publishing Group, 38(3), p. 285.

Gaudet, P. *et al.* (2017) ‘The neXtProt knowledgebase on human proteins: 2017 update’, *Nucleic Acids Research*, 45(D1), pp. D177–D182. doi: 10.1093/nar/gkw1062.

Gavin, A.-C., Maeda, K. and Kühner, S. (2011) ‘Recent advances in charting protein–protein interaction: mass spectrometry-based approaches’, *Current Opinion in Biotechnology*, 22(1), pp. 42–49. doi: 10.1016/j.copbio.2010.09.007.

Gemovic, B. *et al.* (2018) ‘Mapping of Protein-Protein Interactions: Web-Based Resources for Revealing Interactomes’, *Current Medicinal Chemistry*, 25, pp. 1–19. doi: 10.2174/0929867325666180214113704.

Gibson, A. and Patterson, J. (2017) *Deep Learning A Practitioner’s Approach, O’Reiley Media*. doi: 10.1038/nature14539.

Gilbert, D. (2005) ‘Biomolecular interaction network database’, *Briefings in Bioinformatics*. doi: 10.1093/bib/6.2.194.

Gitter, A. *et al.* (2011) 'Discovering pathways by orienting edges in protein interaction networks', *Nucleic Acids Research*. doi: 10.1093/nar/gkq1207.

Glickman, M. E. and van Dyk, D. a (2007) 'Basic Bayesian methods.', *Methods in molecular biology (Clifton, N.J.)*, 404, pp. 319–38. doi: 10.1007/978-1-59745-530-5_16.

Goll, J. *et al.* (2008) 'MPIDB: The microbial protein interaction database', *Bioinformatics*. doi: 10.1093/bioinformatics/btn285.

Gomez, S. M., Noble, W. S. and Rzhetsky, a. (2003) 'Learning to predict protein-protein interactions from protein sequences', *Bioinformatics*, 19(15), pp. 1875–1881. doi: 10.1093/bioinformatics/btg352.

Gosak, M. *et al.* (2018) 'Network science of biological systems at different scales: A review', *Physics of Life Reviews*, 24, pp. 118–135. doi: 10.1016/j.plrev.2017.11.003.

Grigoriev, A. (2001) 'A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage T7 and the yeast Saccharomyces cerevisiae', *Nucleic acids research*. Oxford University Press, 29(17), pp. 3513–3519.

Guarracino, M. R. *et al.* (2010) 'Efficient Prediction of Protein-Protein Interactions Using Sequence Information', *Complex, Intelligent and Software Intensive Systems (CISIS), 2010 International Conference on*. doi: 10.1109/CISIS.2010.161.

Guimerà, R. and Sales-Pardo, M. (2009) 'Missing and spurious interactions and the reconstruction of complex networks', *Proceedings of the National Academy of Sciences*, 106(52), pp. 22073–22078. doi: 10.1073/pnas.0908366106.

Güldener, U. *et al.* (2006) 'MPact: the MIPS protein interaction resource on yeast',

Nucleic acids research. Oxford University Press, 34(suppl_1), pp. D436–D441. doi: 10.1093/nar/gkj003.

Guo, R.-S. *et al.* (2016) 'Restoration of Brain Acid Soluble Protein 1 Inhibits Proliferation and Migration of Thyroid Cancer Cells.', *Chinese medical journal*, 129(12), pp. 1439–46. doi: 10.4103/0366-6999.183434.

Guo, Y. *et al.* (2008) 'Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences.', *Nucleic acids research*, 36(9), pp. 3025–30. doi: 10.1093/nar/gkn159.

Guyon, I. (2003) 'An Introduction to Variable and Feature Selection.pdf', 3, pp. 1157–1182. doi: 10.1023/A:1012487302797.

Hakes, L. *et al.* (2008) 'Protein-protein interaction networks and biology--what's the connection?', *Nature biotechnology*, 26(1), pp. 69–72. doi: 10.1038/nbt0108-69.

Hamosh, A. *et al.* (2005) 'Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders', *Nucleic Acids Research*. doi: 10.1093/nar/gki033.

Hamp, T. and Rost, B. (2013) 'Improved protein-protein interaction prediction from sequence', 2, pp. 1–4.

Hamp, T. and Rost, B. (2015a) 'Evolutionary profiles improve protein–protein interaction prediction from sequence', *Bioinformatics*, 31(12), pp. 1945–1950. doi: 10.1093/bioinformatics/btv077.

Hamp, T. and Rost, B. (2015b) 'More challenges for machine-learning protein interactions', *Bioinformatics*, 31(10), pp. 1521–1525. doi: 10.1093/bioinformatics/btu857.

Han, M.-H. *et al.* (2013) 'The Novel Caspase-3 Substrate Gap43 is Involved in AMPA Receptor Endocytosis and Long-Term Depression', *Molecular & Cellular Proteomics*, 12(12), pp. 3719–3731. doi: 10.1074/mcp.M113.030676.

Han, Y. C. *et al.* (2016) 'Prediction and characterization of protein-protein interaction network in Bacillus licheniformis WX-02', *Scientific Reports*. Nature Publishing Group, 6(September 2015), pp. 1–11. doi: 10.1038/srep19486.

Hand, D. J. and Yu, K. (2001) 'Idiot's Bayes---Not So Stupid After All?', *International Statistical Review*, 69(3), pp. 385–398. doi: 10.1111/j.1751-5823.2001.tb00465.x.

Al Hasan, M. *et al.* (2006) 'Link prediction using supervised learning', in *SDM06: workshop on link analysis, counter-terrorism and security*, pp. 556–562. doi: 10.1109/IJCNN.2008.4634046.

Hashemifar, S. *et al.* (2018) 'Predicting protein-protein interactions through sequence-based deep learning', *Bioinformatics*, 34(17), pp. i802–i810. doi: 10.1093/bioinformatics/bty573.

Hawkins, L. J., Al-attar, R. and Storey, K. B. (2018) 'Transcriptional regulation of metabolism in disease: From transcription factors to epigenetics', *PeerJ*, 6(3), p. e5062. doi: 10.7717/peerj.5062.

Haynes, C. *et al.* (2006) 'Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes', *PLoS Computational Biology*, 2(8), pp. 0890–0901. doi: 10.1371/journal.pcbi.0020100.

He, B. *et al.* (2009) 'Predicting intrinsic disorder in proteins: An overview', *Cell Research*. doi: 10.1038/cr.2009.87.

Hermjakob, H. *et al.* (2004) 'IntAct: an open source molecular interaction database.', *Nucleic acids research*, 32(Database issue), pp. D452–D455. doi: 10.1093/nar/gkh052.

Hindorff, L. A. *et al.* (2009) 'Potential etiologic and functional implications of genome-wide association loci for human diseases and traits', *Proceedings of the National Academy of Sciences*. doi: 10.1073/pnas.0903103106.

Horita, H. *et al.* (2017) 'A simple toolset to identify endogenous post-translational modifications for a target protein: a snapshot of the EGFR signaling pathway', *Bioscience Reports*. doi: 10.1042/bsr20170919.

Hornik, K. (1991) 'Approximation capabilities of multilayer feedforward networks', *Neural Networks*. doi: 10.1016/0893-6080(91)90009-T.

Hu, L. *et al.* (2011) 'Predicting functions of proteins in mouse based on weighted protein-protein interaction network and protein hybrid properties', *PLoS ONE*, 6(1). doi: 10.1371/journal.pone.0014556.

Hu, Y. P. and Tsay, R. S. (2014) 'Principal Volatility Component Analysis', *Journal of Business and Economic Statistics*, 32(2), pp. 153–164. doi: 10.1080/07350015.2013.818006.

Huang, Y.-A. *et al.* (2016) 'Sequence-based prediction of protein-protein interactions using weighted sparse representation model combined with global encoding', *BMC Bioinformatics*. BMC Bioinformatics, 17(1), p. 184. doi: 10.1186/s12859-016-1035-4.

Huang, Z. (2010) 'Link Prediction Based on Graph Topology: The Predictive Value of Generalized Clustering Coefficient', *SSRN Electronic Journal*, pp. 1–31. doi: 10.2139/ssrn.1634014.

Hubner, N. C. and Mann, M. (2011) 'Extracting gene function from protein–protein interactions using Quantitative BAC InteraCtomics (QUBIC)', *Methods*, 53(4), pp. 453–459. doi: 10.1016/j.ymeth.2010.12.016.

Hue, M. *et al.* (2010) 'Large-scale prediction of protein-protein interactions from structures.', *BMC bioinformatics*, 11, p. 144. doi: 10.1186/1471-2105-11-144.

Huff, V. (2011) 'Wilms' tumours: about tumour suppressor genes, an oncogene and a chameleon gene', *Nature Reviews Cancer*, 11(2), pp. 111–121. doi: 10.1038/nrc3002.

Huynen, M. *et al.* (2000) 'Predicting protein function by genomic context: quantitative evaluation and qualitative inferences', *Genome research*. Cold Spring Harbor Lab, 10(8), pp. 1204–1210.

Ito, T. *et al.* (2001) 'A comprehensive two-hybrid analysis to explore the yeast protein interactome', *Proceedings of the National Academy of Sciences*, 98(8), pp. 4569–4574. doi: 10.1073/pnas.061034498.

Jansen, R. *et al.* (2002) 'Integration of genomic datasets to predict protein complexes in yeast', in *Journal of Structural and Functional Genomics*. doi: 10.1023/A:1020495201615.

Jansen, R. (2003) 'A Bayesian Networks Approach for Predicting Protein-Protein Interactions from Genomic Data', *Science*. American Association for the Advancement of Science, 302(5644), pp. 449–453. doi: 10.1126/science.1087361.

Jansen, R. and Gerstein, M. (2004) 'Analyzing protein function on a genomic scale: the importance of gold-standard positives and negatives for network prediction.', *Current opinion in microbiology*, 7(5), pp. 535–45. doi: 10.1016/j.mib.2004.08.012.

Jansen, R., Greenbaum, D. and Gerstein, M. (2002) 'Relating whole-genome expression data with protein-protein interactions', *Genome Research*. doi: 10.1101/gr.205602.

Jiang, X. *et al.* (no date) 'The development of an indirect competitive immunomagnetic-proximity ligation assay for small-molecule detection The development of an indirect competitive immunomagnetic-proximity ligation assay for small-molecule detection †'.

doi: 10.1039/c2an36447f.

John, G. H., Kohavi, R. and Pfleger, K. (1994) 'Irrelevant features and the subset selection problem', in *Machine Learning Proceedings 1994*. Elsevier, pp. 121–129. Available at: http://machine-learning.martinsewell.com/feature-selection/JohnKohaviPfleger1994.pdf.

Jolliffe, I. (2011) 'Principal component analysis', in *International encyclopedia of statistical science*. Springer, pp. 1094–1096.

Jon M. Kleinberg (1999) 'Authoritative Sources in a Hyperlinked Environment', *Journal of the ACMthe ACM*, 46(5), pp. 604–632. doi: 10.1.1.120.3875.

Jones, D. T. and Swindells, M. B. (2002) 'Getting the most from PSI-BLAST', *Trends in Biochemical Sciences*, 27(3), pp. 161–164. doi: 10.1016/S0968-0004(01)02039-4.

Jordan, M. I. and Mitchell, T. M. (2015) 'Machine learning: Trends, perspectives, and prospects', *Science*. doi: 10.1126/science.aaa8415.

Kao, F.-S. *et al.* (2012) 'Chip-based protein-protein interaction studied by atomic force microscopy', *Biotechnology and Bioengineering*, 109(10), pp. 2460–2467. doi: 10.1002/bit.24521.

Karimzadeh, M. and Hoffman, M. M. (2017) 'Top considerations for creating bioinformatics software documentation', *Briefings in Bioinformatics*, 19(November 2016), p. bbw134. doi: 10.1093/bib/bbw134.

Katz, G., Shin, E. C. R. and Song, D. (2017) 'ExploreKit: Automatic feature generation and selection', *Proceedings - IEEE International Conference on Data Mining, ICDM*, pp. 979–984. doi: 10.1109/ICDM.2016.176.

Kaul, A., Maheshwary, S. and Pudi, V. (2017) 'AutoLearn — Automated Feature

Generation and Selection', *2017 IEEE International Conference on Data Mining (ICDM)*, pp. 217–226. doi: 10.1109/ICDM.2017.31.

Kaushansky, A. *et al.* (2010) 'Quantifying protein–protein interactions in high throughput using protein domain microarrays', *Nature Protocols*, 5(4), pp. 773–790. doi: 10.1038/nprot.2010.36.

Kawashima, S., Ogata, H. and Kanehisa, M. (1999) 'AAindex: Amino acid index database', *Nucleic Acids Research*, 27(1), pp. 368–369. doi: 10.1093/nar/27.1.368.

Kelchtermans, P. *et al.* (2014) 'Machine learning applications in proteomics research: How the past can boost the future', *Proteomics*, 14(4–5), pp. 353–366. doi: 10.1002/pmic.201300289.

Kerppola, T. K. (2008) 'Bimolecular Fluorescence Complementation (BiFC) Analysis as a Probe of Protein Interactions in Living Cells', *Annual Review of Biophysics*, 37(1), pp. 465–487. doi: 10.1146/annurev.biophys.37.032807.125842.

Keshava Prasad, T. S. *et al.* (2009) 'Human Protein Reference Database--2009 update', *Nucleic Acids Research*. Oxford University Press, 37(Database), pp. D767–D772. doi: 10.1093/nar/gkn892.

Keskin, O. *et al.* (2008) 'Principles of Protein−Protein Interactions: What are the Preferred Ways For Proteins To Interact?', *Chemical Reviews*, 108(4), pp. 1225–1244. doi: 10.1021/cr040409x.

Keskin, O., Tuncbag, N. and Gursoy, A. (2008) 'Characterization and prediction of protein interfaces to infer protein-protein interaction networks.', *Current pharmaceutical biotechnology*, 9(2), pp. 67–76. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18393863.

Keskin, O., Tuncbag, N. and Gursoy, A. (2016) 'Predicting Protein-Protein Interactions

from the Molecular to the Proteome Level', *Chemical Reviews*, 116(8), pp. 4884–4909. doi: 10.1021/acs.chemrev.5b00683.

Kessel, A. and Ben-Tal, N. (2011) *Introduction to Proteins: Structure, Function, and Motion*. CRC Press.

Khor, S. (2014) 'Inferring domain-domain interactions from protein-protein interactions with formal concept analysis', *PLoS ONE*, 9(2). doi: 10.1371/journal.pone.0088943.

Khurana, U., Samulowitz, H. and Turaga, D. (2017) 'Feature Engineering for Predictive Modeling using Reinforcement Learning'. Available at: http://arxiv.org/abs/1709.07150.

Kim, H. *et al.* (2019) 'Universal scaling across biochemical networks on Earth', *Science Advances*, 5(1), p. eaau0149. doi: 10.1126/sciadv.aau0149.

Kim, M. S. *et al.* (2014) 'A draft map of the human proteome', *Nature*. doi: 10.1038/nature13302.

Klein, P., Kanehisa, M. and DeLisi, C. (1984) 'Prediction of protein function from sequence properties. Discriminant analysis of a data base.', *Biochimica et biophysica acta*, 787(3), pp. 221–6. Available at: http://www.ncbi.nlm.nih.gov/pubmed/6547351.

Kodama, Y. and Hu, C.-D. (2012) 'Bimolecular fluorescence complementation (BiFC): A 5-year update and future perspectives', *BioTechniques*, 53(5), pp. 285–98. doi: 10.2144/000113943.

Krogan, N. J. *et al.* (2006) 'Global landscape of protein complexes in the yeast Saccharomyces cerevisiae', *Nature*. doi: 10.1038/nature04670.

Kuhn, M. (2008) 'Caret package', *Journal of statistical software*, 28(5), pp. 1–26. Available at: http://www.jstatsoft.org/v28/i05/.

Kuhn, M. and Johnson, K. (2013) *Applied Predictive Modeling*, *Springer*. doi: 10.1007/978-1-4614-6849-3.

Kuzmanov, U. and Emili, A. (2013) 'Protein-protein interaction networks: Probing disease mechanisms using model systems', *Genome Medicine*, 5(4), pp. 1–12. doi: 10.1186/gm441.

Lakowicz, J. R. (2006) 'Introduction to Fluorescence', in *Principles of Fluorescence Spectroscopy*. Boston, MA: Springer US, pp. 1–26. doi: 10.1007/978-0-387-46312-4_1.

Lander, E. S. *et al.* (2001) 'Initial sequencing and analysis of the human genome', *Nature*. doi: 10.1038/35057062.

Landry, M. (2018) 'Machine learning with R and H2O.', *Mountain View, CA*. H2O.ai. Available at: https://www.h2o.ai.

Larrañaga, P. *et al.* (2006) 'Machine learning in bioinformatics', *Briefings in Bioinformatics*. Elsevier B.V., 7(1), pp. 86–112. doi: 10.1093/bib/bbk007.

De Las Rivas, J. and Fontanillo, C. (2010) 'Protein–Protein Interactions Essentials: Key Concepts to Building and Analyzing Interactome Networks', *PLoS Computational Biology*. Edited by F. Lewitter, 6(6), p. e1000807. doi: 10.1371/journal.pcbi.1000807.

De Las Rivas, J. and Fontanillo, C. (2012) 'Protein-protein interaction networks: unraveling the wiring of molecular machines within the cell', *Briefings in Functional Genomics*, 11(6), pp. 489–496. doi: 10.1093/bfgp/els036.

De Las Rivas, J. and de Luis, A. (2004) 'Interactome data and databases: Different types of protein interaction', *Comparative and Functional Genomics*, 5(2), pp. 173–178. doi: 10.1002/cfg.377.

Latouche, P. and Rossi, F. (2015) ‘Graphs in machine learning: an introduction’. Available at: http://arxiv.org/abs/1506.06962.

Lazo, J. S. and Sharlow, E. R. (2016) ‘Drugging Undruggable Molecular Cancer Targets’, *Annual Review of Pharmacology and Toxicology*. doi: 10.1146/annurev-pharmtox-010715-103440.

Lecun, Y., Bengio, Y. and Hinton, G. (2015) ‘Deep learning’, *Nature*, 521(7553), pp. 436–444. doi: 10.1038/nature14539.

Lee, D. D. and Seung, H. S. (2001) ‘Link Prediction using Supervised Learning’, *Advances in Neural Information Processing Systems 13 (NIPS)*, pp. 556–562. doi: 10.1109/IJCNN.2008.4634046.

Lee, T. I. and Young, R. A. (2013) ‘Transcriptional regulation and its misregulation in disease’, *Cell*. Elsevier Inc., 152(6), pp. 1237–1251. doi: 10.1016/j.cell.2013.02.014.

Lehne, B. and Schlitt, T. (2009) ‘Protein-protein interaction databases: keeping up with growing interactomes.’, *Human genomics*, 3(3), pp. 291–297. doi: E54180217L822J23 [pii].

Lelli, K. M., Slattery, M. and Mann, R. S. (2012) ‘Disentangling the Many Layers of Eukaryotic Transcriptional Regulation’, *Annual Review of Genetics*, 46(1), pp. 43–68. doi: 10.1146/annurev-genet-110711-155437.

Lemon, B. and Tjian, R. (2000) ‘Orchestrated response: A symphony of transcription factors for gene control’, *Genes and Development*. doi: 10.1101/gad.831000.

Levy, E. D. and Teichmann, S. A. (2013) ‘Structural, evolutionary, and assembly principles of protein oligomerization’, in *Progress in molecular biology and translational science*. Elsevier, pp. 25–51.

Li, H. *et al.* (2014) 'Discovery of small-molecule inhibitors selectively targeting the DNA-binding domain of the human androgen receptor', *Journal of Medicinal Chemistry*. doi: 10.1021/jm500802j.

Li, H., Liu, C. and Burge, L. (2012) 'Predicting Protein-Protein Interactions Based on PPI Networks', 2(6), pp. 794–797.

Li, J.-Q. Q. *et al.* (2017) 'PSPEL: In Silico Prediction of Self-Interacting Proteins from Amino Acids Sequences Using Ensemble Learning', *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 14(5), pp. 1165–1172. doi: 10.1109/TCBB.2017.2649529.

Li, M. *et al.* (2010) 'Essential proteins discovery from weighted protein interaction networks', in *International Symposium on Bioinformatics Research and Applications*. Springer, pp. 89–100.

Li, M. *et al.* (2012) 'A new essential protein discovery method based on the integration of protein-protein interaction and gene expression data', *BMC systems biology*. BioMed Central, 6(1), p. 15.

Li, Wan *et al.* (2013) 'Prioritizing Disease Candidate Proteins in Cardiomyopathy-Specific Protein-Protein Interaction Networks Based on "Guilt by Association" Analysis', *PLoS ONE*. doi: 10.1371/journal.pone.0071191.

Liao, H., Zeng, A. and Zhang, Y. C. (2015) 'Predicting missing links via correlation between nodes', *Physica A: Statistical Mechanics and its Applications*. Elsevier B.V., 436, pp. 216–223. doi: 10.1016/j.physa.2015.05.009.

Liaw, A. and Wiener, M. (2002) 'Classifcation and Regression by randomForest', *R news*, 2(December), pp. 18–22. doi: 10.1177/154405910408300516.

Lichtenwalter, R. N., Lussier, J. T. and Chawla, N. V. (2010) 'New perspectives and methods in link prediction', *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '10*, p. 243. doi: 10.1145/1835804.1835837.

Liu, G., Wong, L. and Chua, H. N. (2009) 'Complex discovery from weighted PPI networks', *Bioinformatics*. doi: 10.1093/bioinformatics/btp311.

Liu, H. (2009) 'Protein-Protein Interaction Detection By SVM From Sequence Information', *The Third International Symposium on Optimization and Systems Biology*, pp. 198–206.

Lubovac, Z. *et al.* (2007) 'Weighted cohesiveness for identification of functional modules and their interconnectivity', in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*.

Luck, K. *et al.* (2018) 'Network-based prediction of protein interactions', *bioRxiv*. doi: 10.1101/275529.

Ma, C., Bao, Z. K. and Zhang, H. F. (2017) 'Improving link prediction in complex networks by adaptively exploiting multiple structural features of networks', *Physics Letters, Section A: General, Atomic and Solid State Physics*. Elsevier B.V., 381(39), pp. 3369–3376. doi: 10.1016/j.physleta.2017.08.047.

Ma, X. *et al.* (2007) 'CGI: A new approach for prioritizing genes by combining gene expression and protein-protein interaction data', *Bioinformatics*. doi: 10.1093/bioinformatics/btl569.

Macháň, R. and Wohland, T. (2014) 'Recent applications of fluorescence correlation spectroscopy in live systems.', *FEBS letters*, 588(19), pp. 3571–84. doi: 10.1016/j.febslet.2014.03.056.

Mackay, J. *et al.* (2007) 'Protein interactions: is seeing believing?', *Trends in Biochemical Sciences*, 32(12), pp. 530–531. doi: 10.1016/j.tibs.2007.09.006.

Maemondo, M. *et al.* (2010) 'Gefitinib or Chemotherapy for Non–Small-Cell Lung Cancer with Mutated EGFR', *New England Journal of Medicine*. doi: 10.1056/NEJMoa0909530.

Maere, S., Heymans, K. and Kuiper, M. (2005) 'BiNGO: A Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks', *Bioinformatics*, 21(16), pp. 3448–3449. doi: 10.1093/bioinformatics/bti551.

Maetschke, S. R. *et al.* (2012) 'Gene Ontology-driven inference of protein-protein interactions using inducers', *Bioinformatics*, 28(1), pp. 69–75. doi: 10.1093/bioinformatics/btr610.

Maheswaran, S. *et al.* (1993) 'Physical and functional interaction between WT1 and p53 proteins', *Proceedings of the National Academy of Sciences*. National Acad Sciences, 90(11), pp. 5100–5104.

Mahmoudian, S. (2015) 'Protein-Protein Interaction Prediction using PCA and SVR-PHCS', *Open Bioinformatics* …, pp. 1–12. Available at: http://benthamopen.com/contents/pdf/TOBIOIJ/TOBIOIJ-9-1.pdf.

Malhotra, A. *et al.* (2015) 'Exploring novel mechanistic insights in Alzheimer's disease by assessing reliability of protein interactions', *Scientific Reports*. doi: 10.1038/srep13634.

Marcotte, E. M. *et al.* (1999) 'Detecting protein function and protein-protein interactions from genome sequences', *Science*. American Association for the Advancement of Science, 285(5428), pp. 751–753.

Marsh, L. A. *et al.* (2017) 'BASP1 interacts with oestrogen receptor α and modifies the

tamoxifen response', *Cell Death and Disease*. Nature Publishing Group, 8(5), pp. e2771-10. doi: 10.1038/cddis.2017.179.

Martin, S., Roe, D. and Faulon, J. L. (2005) 'Predicting protein-protein interactions using signature products', *Bioinformatics*, 21(2), pp. 218–226. doi: 10.1093/bioinformatics/bth483.

Maston, G. A., Evans, S. K. and Green, M. R. (2006) 'Transcriptional regulatory elements in the human genome.', *Annual review of genomics and human genetics*, 7, pp. 29–59. doi: 10.1146/annurev.genom.7.080505.115623.

McGillivray, P. *et al.* (2018) 'Network Analysis as a Grand Unifier in Biomedical Data Science', *Annual Review of Biomedical Data Science*, 1(1), pp. 153–180. doi: 10.1146/annurev-biodatasci-080917-013444.

McKay, L. M., Carpenter, B. and Roberts, S. G. E. (1999) 'Regulation of the Wilms' tumour suppressor protein transcriptional activation domain', *Oncogene*. Nature Publishing Group, 18(47), p. 6546.

Meng, F., Uversky, V. and Kurgan, L. (2017) 'Computational Prediction of Intrinsic Disorder in Proteins', in *Current Protocols in Protein Science*. doi: 10.1002/cpps.28.

Merico, D., Gfeller, D. and Bader, G. D. (2009) 'How to visually interpret biological data using networks', *Nature Biotechnology*, 27(10), pp. 921–924. doi: 10.1038/nbt.1567.

von Mering, C. *et al.* (2002) 'Comparative assessment of large-scale data sets of protein–protein interactions', *Nature*, 417(6887), pp. 399–403. doi: 10.1038/nature750.

Mészáros, B. *et al.* (2007) 'Molecular Principles of the Interactions of Disordered Proteins', *Journal of Molecular Biology*, 372(2), pp. 549–561. doi: 10.1016/j.jmb.2007.07.004.

Michalak, K. and Kwasnicka, H. (2006) ‘Correlation – Based Feature Selection Strategy’, *Int. J. Appl. Math. Comput. Sci.*, 16(4), pp. 503–511.

Milo, R. (2013) ‘What is the total number of protein molecules per cell volume? A call to rethink some published values’, *BioEssays*. doi: 10.1002/bies.201300066.

Minezaki, Y. *et al.* (2006) ‘Human Transcription Factors Contain a High Fraction of Intrinsically Disordered Regions Essential for Transcriptional Regulation’, *Journal of Molecular Biology*. doi: 10.1016/j.jmb.2006.04.016.

Mitchell, M. (1996) ‘An introduction to genetic algorithms’, *Computers & Mathematics with Applications*, 32(6), p. 133. doi: 10.1016/S0898-1221(96)90227-8.

Mitchell, P. (2002) ‘A perspective on protein microarrays’, *Nature Biotechnology*, 20(3), pp. 225–229. doi: 10.1038/nbt0302-225.

Miyawaki, A. (2011) ‘Development of Probes for Cellular Functions Using Fluorescent Proteins and Fluorescence Resonance Energy Transfer’, *Annual Review of Biochemistry*, 80(1), pp. 357–373. doi: 10.1146/annurev-biochem-072909-094736.

Mo, Y. *et al.* (2001) ‘Crystal structure of a ternary SAP-1/SRF/c-fos SRE DNA complex’, *Journal of Molecular Biology*. doi: 10.1006/jmbi.2001.5138.

Moribe, T. *et al.* (2008) ‘Identification of novel aberrant methylation of BASP1 and SRD5A2 for early diagnosis of hepatocellular carcinoma by genome-wide search.’, *International journal of oncology*, 33(5), pp. 949–58. doi: 10.3892/ijo_00000082.

Moss, G. P., Smith, P. A. S. and Tavernier, D. (2007) ‘Glossary of class names of organic compounds and reactivity intermediates based on structure (IUPAC Recommendations 1995)’, *Pure and Applied Chemistry*. doi: 10.1351/pac199567081307.

Mukherjee, S. and Zhang, Y. (2011) ‘Protein-protein complex structure predictions by multimeric threading and template recombination’, *Structure*. Elsevier Ltd, 19(7), pp. 955–966. doi: 10.1016/j.str.2011.04.006.

Muley, V. Y. and Ranjan, A. (2012) ‘Effect of reference genome selection on the performance of computational methods for genome-wide protein-protein interaction prediction’, *PLoS ONE*, 7(7). doi: 10.1371/journal.pone.0042057.

Murakami, Y. and Mizuguchi, K. (2014) ‘Homology-based prediction of interactions between proteins using Averaged One-Dependence Estimators.’, *BMC bioinformatics*, 15(1), p. 213. doi: 10.1186/1471-2105-15-213.

Murata, K. and Wolf, M. (2018) ‘Cryo-electron microscopy for structural analysis of dynamic biological macromolecules’, *Biochimica et Biophysica Acta (BBA) - General Subjects*. Elsevier, 1862(2), pp. 324–334. doi: 10.1016/J.BBAGEN.2017.07.020.

Nabieva, E. *et al.* (2005) ‘Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps’, *Bioinformatics*, 21(SUPPL. 1), pp. 302–310. doi: 10.1093/bioinformatics/bti1054.

Natekin, A. and Knoll, A. (2013) ‘Gradient boosting machines, a tutorial’, *Frontiers in Neurorobotics*, 7(DEC). doi: 10.3389/fnbot.2013.00021.

Nelder, J. A. and Wedderburn, R. W. M. (1972) ‘Generalized Linear Models’, *Journal of the Royal Statistical Society. Series A (General)*. Hoboken, NJ, USA: John Wiley & Sons, Inc., 135(3), p. 370. doi: 10.2307/2344614.

Newman, M. E. J. (2002) ‘Random graphs as models of networks’, p. 35. doi: 10.1002/3527602755.ch2.

Newman, M. E. J. and Girvan, M. (2004) ‘Finding and evaluating community structure

in networks', *Physical review E*. APS, 69(2), p. 26113.

Ng, A. Y. (1997) 'Preventing" overfitting" of cross-validation data', in *ICML*, pp. 245–253.

Ng, S.-K., Zhang, Z. and Tan, S.-H. (2003) 'Integrative approach for computationally inferring protein domain interactions', in *Proceedings of the 2003 ACM symposium on Applied computing*. ACM, pp. 115–121.

Nguyen, H. A. *et al.* (2015) 'Discovery of pathways in protein-protein interaction networks using a genetic algorithm', in *Data and Knowledge Engineering*. doi: 10.1016/j.datak.2015.04.002.

Nooren, I. M. A. and Thornton, J. M. (2003) 'Diversity of protein-protein interactions', *EMBO Journal*, 22(14), pp. 3486–3492. doi: 10.1093/emboj/cdg359.

Nowell, D. L. *et al.* (2003) 'The link prediction problem for social networks', *Proceedings*, (November 2003), pp. 556–559. doi: http://doi.acm.org/10.1145/956863.956972.

Ofer, D. and Linial, M. (2015) 'ProFET: Feature engineering captures high-level protein functions', *Bioinformatics*, 31(21), pp. 3429–3436. doi: 10.1093/bioinformatics/btv345.

Oldfield, C. J. *et al.* (2005) 'Comparing and combining predictors of mostly disordered proteins', *Biochemistry*. doi: 10.1021/bi047993o.

Olson, R. S. *et al.* (2017) 'Data-driven Advice for Applying Machine Learning to Bioinformatics Problems', pp. 192–203. doi: 10.1142/9789813235533_0018.

Olsson, U. (2002) 'Generalized linear models', *An applied approach. Studentlitteratur, Lund*, 18.

Oluleye, B. and Armstrong, L. (2014) ‘A genetic Algorithm-Based feature selection’, *British Journal …*, (July). Available at: http://researchrepository.murdoch.edu.au/23706/.

Ooi, S. L., Shoemaker, D. D. and Boeke, J. D. (2003) ‘DNA helicase gene interaction network defined using synthetic lethality analyzed by microarray’, *Nature Genetics*. Nature Publishing Group, 35(3), pp. 277–286. doi: 10.1038/ng1258.

Overbeek, R. *et al.* (1999) ‘The use of gene clusters to infer functional coupling’, *Proceedings of the National Academy of Sciences*. National Acad Sciences, 96(6), pp. 2896–2901.

Packham, S. *et al.* (2015) ‘The Nucleus-Localized Epidermal Growth Factor Receptor is SUMOylated’, *Biochemistry*. doi: 10.1021/acs.biochem.5b00640.

Padilla-Parra, S. and Tramier, M. (2012) ‘FRET microscopy in the living cell: Different approaches, strengths and weaknesses’, *BioEssays*, 34(5), pp. 369–376. doi: 10.1002/bies.201100086.

Paez, J. G. *et al.* (2004) ‘EGFR mutations in lung, cancer: Correlation with clinical response to gefitinib therapy’, *Science*. doi: 10.1126/science.1099314.

Pagel, P. *et al.* (2005) ‘The MIPS mammalian protein-protein interaction database’, *Bioinformatics*. doi: 10.1093/bioinformatics/bti115.

Paladugu, S. R. *et al.* (2008) ‘Mining protein networks for synthetic genetic interactions’, *BMC Bioinformatics*, 9, pp. 1–14. doi: 10.1186/1471-2105-9-426.

Pan, X. Y., Zhang, Y. N. and Shen, H. Bin (2010) ‘Large-scale prediction of human protein-protein interactions from amino acid sequence based on latent topic features’, *Journal of Proteome Research*, 9(10), pp. 4992–5001. doi: 10.1021/pr100618t.

Panchenko, A.; Przytycka, T. (2008) *Protein-protein Interactions and Networks - Identification, Computer Analysis, and Prediction*, *Springer*. Springer. Available at: http://www.springer.com/computer/bioinformatics/book/978-1-84800-124-4 (Accessed: 27 August 2018).

Panne, D. (2008) 'The enhanceosome', *Current Opinion in Structural Biology*. doi: 10.1016/j.sbi.2007.12.002.

Park, S. H., Goo, J. M. and Jo, C.-H. (2004) 'Receiver operating characteristic (ROC) curve: practical review for radiologists.', *Korean Journal of Radiology*, 5(March), pp. 11–8. doi: 10.3348/kjr.2004.5.1.11.

Park, Y. (2009) 'Critical assessment of sequence-based protein-protein interaction prediction methods that do not require homologous protein sequences', *BMC Bioinformatics*, 10(1), p. 419. doi: 10.1186/1471-2105-10-419.

Park, Y. and Marcotte, E. M. (2011) 'Revisiting the negative example sampling problem for predicting protein-protein interactions', *Bioinformatics*, 27(21), pp. 3024–3028. doi: 10.1093/bioinformatics/btr514.

Park, Y. and Marcotte, E. M. (2012) 'Flaws in evaluation schemes for pair-input computational predictions', *Nature Methods*, 9(12), pp. 1134–1136. doi: 10.1038/nmeth.2259.

Pastor-Satorras, R., Smith, E. and Solé, R. V. (2003) 'Evolving protein interaction networks through gene duplication', *Journal of Theoretical Biology*. doi: 10.1016/S0022-5193(03)00028-6.

Patthy, L. (2016) 'Identification and Correction of Erroneous Protein Sequences in Public Databases', in *Methods in Molecular Biology*, pp. 179–192. doi: 10.1007/978-1-4939-3572-7_9.

Pavlopoulos, G. a *et al.* (2011) ‘Using graph theory to analyze biological networks.’, *BioData mining*. BioMed Central Ltd, 4(1), p. 10. doi: 10.1186/1756-0381-4-10.

Pellegrini, M. *et al.* (1999) ‘Assigning protein functions by comparative genome analysis: protein phylogenetic profiles’, *Proceedings of the National Academy of Sciences*. National Acad Sciences, 96(8), pp. 4285–4288.

Peng, W. *et al.* (2012) ‘Iteration method for predicting essential proteins based on orthology and protein-protein interaction networks’, *BMC Systems Biology*. doi: 10.1186/1752-0509-6-87.

Peng, X. *et al.* (2015) ‘An efficient method to identify essential proteins for different species by integrating protein subcellular localization information’, in *Proceedings - 2015 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2015*. doi: 10.1109/BIBM.2015.7359693.

Peng, X. *et al.* (2017) ‘Protein-protein interactions: detection, reliability assessment and applications’, *Briefings in bioinformatics*, 18(5), pp. 798–819. doi: 10.1093/bib/bbw066.

Peng, Z. *et al.* (2014) ‘Exceptionally abundant exceptions: Comprehensive characterization of intrinsic disorder in all domains of life’, *Cellular and Molecular Life Sciences*. doi: 10.1007/s00018-014-1661-9.

Perez-Riverol, Y. *et al.* (2017) ‘Accurate and fast feature selection workflow for high-dimensional omics data’, *PLoS ONE*, 12(12), pp. 1–14. doi: 10.1371/journal.pone.0189875.

Perovic, V. *et al.* (2017) ‘TRI_tool: a web-tool for prediction of protein–protein interactions in human transcriptional regulation’, *Bioinformatics*, 33(2), pp. 289–291. doi: 10.1093/bioinformatics/btw590.

Perovic, V. *et al.* (2018) 'IDPpi: Protein-Protein Interaction Analyses of Human Intrinsically Disordered Proteins', *Scientific Reports*. Springer US, 8(1), p. 10563. doi: 10.1038/s41598-018-28815-x.

Persani, L., Calebiro, D. and Bonomi, M. (2007) 'Technology Insight: modern methods to monitor protein–protein interactions reveal functional TSH receptor oligomerization', *Nature Clinical Practice Endocrinology & Metabolism*. Nature Publishing Group, 3(2), pp. 180–190. doi: 10.1038/ncpendmet0401.

Phan, W. *et al.* (2017) 'Deep Learning with Deep Water: First Edition Deep Learning with Deep Water'. Available at: http://h2o-release.s3.amazonaws.com/h2o/rel-wright/1/docs-website/h2o-docs/booklets/DeepWaterBooklet.pdf (Accessed: 23 November 2018).

Phizicky, E M and Fields, S. (1995) 'Protein-protein interactions: methods for detection and analysis.', *Microbiological reviews*. Am Soc Microbiol, 59(1), pp. 94–123. Available at: http://mmbr.asm.org/content/59/1/94.short (Accessed: 14 August 2013).

Phizicky, EM M and Fields, S. (1995) 'Protein-protein interactions: methods for detection and analysis.', *Microbiological reviews*, 59(1), pp. 94–123.

Piovesan, D. *et al.* (2017) 'DisProt 7.0: A major update of the database of disordered proteins', *Nucleic Acids Research*, 45(D1), pp. D219–D227. doi: 10.1093/nar/gkw1056.

Pitre, S. *et al.* (2006) 'PIPE: a protein-protein interaction prediction engine based on the re-occurring short polypeptide sequences between known interacting protein pairs.', *BMC bioinformatics*, 7, p. 365. doi: 10.1186/1471-2105-7-365.

Pitre, S. *et al.* (2008) 'Global investigation of protein-protein interactions in yeast Saccharomyces cerevisiae using re-occurring short polypeptide sequences', *Nucleic Acids Research*, 36(13), pp. 4286–4294. doi: 10.1093/nar/gkn390.

Planas-Iglesias, J. *et al.* (2013) 'Understanding Protein–Protein Interactions Using Local Structural Features', *Journal of Molecular Biology*. Elsevier Ltd, 425(7), pp. 1210–1224. doi: 10.1016/j.jmb.2013.01.014.

Ponomarenko, E. A. *et al.* (2016) 'The Size of the Human Proteome: The Width and Depth', *International Journal of Analytical Chemistry*. doi: 10.1155/2016/7436849.

Pons, P. and Latapy, M. (2005) 'Computing communities in large networks using random walks', in *International symposium on computer and information sciences*. Springer, pp. 284–293.

Priebe, C. E. (2006) 'Scan Statistics on Enron Graphs', *Computational & Mathematical Organization Theory*.

Pržulj, N., Corneil, D. G. and Jurisica, I. (2004) 'Modeling interactome: Scale-free or geometric?', *Bioinformatics*. doi: 10.1093/bioinformatics/bth436.

Puig, O. *et al.* (2001) 'The Tandem Affinity Purification (TAP) Method: A General Procedure of Protein Complex Purification', *Methods*, 24(3), pp. 218–229. doi: 10.1006/meth.2001.1183.

Qi, Y., Bar-Joseph, Z. and Klein-Seetharaman, J. (2006) 'Evaluation of different biological data and computational classification methods for use in protein interaction prediction', *Proteins: Structure, …*. Wiley Online Library, 63(3), pp. 490–500. doi: 10.1002/prot.20865.

Radivojac, P. *et al.* (2007) 'Intrinsic disorder and functional proteomics', *Biophysical Journal*. doi: 10.1529/biophysj.106.094045.

Rangwala, H. and Karypis, G. (2005) 'Profile-based direct kernels for remote homology detection and fold recognition', *Bioinformatics*, 21(23), pp. 4239–4247. doi: 10.1093/bioinformatics/bti687.

Rao, V. S. *et al.* (2014) 'Protein-Protein Interaction Detection: Methods and Analysis', *International Journal of Proteomics*, 2014(ii), pp. 1–12. doi: 10.1155/2014/147648.

Ren, X. *et al.* (2011) 'Improving accuracy of protein-protein interaction prediction by considering the converse problem for sequence representation.', *BMC bioinformatics*. BioMed Central Ltd, 12(1), p. 409. doi: 10.1186/1471-2105-12-409.

Rhodes, D. R. *et al.* (2005) 'Probabilistic model of the human protein-protein interaction network', *Nature Biotechnology*. Nature Publishing Group, 23(8), pp. 951–959. doi: 10.1038/nbt1103.

Rigaut, G. *et al.* (1999) 'A generic protein purification method for protein complex characterization and proteome exploration', *Nature Biotechnology*, 17(10), pp. 1030–1032. doi: 10.1038/13732.

Rogers, J. M. *et al.* (2014) 'Interplay between partner and ligand facilitates the folding and binding of an intrinsically disordered protein', *Proceedings of the National Academy of Sciences*. doi: 10.1073/pnas.1409122111.

Rokach, L. (2010) 'Ensemble-based classifiers', *Artificial Intelligence Review*, 33(1–2), pp. 1–39. doi: 10.1007/s10462-009-9124-7.

Rolland, T. *et al.* (2014) 'A proteome-scale map of the human interactome network', *Cell*, 159(5), pp. 1212–1226. doi: 10.1016/j.cell.2014.10.050.

Romanski, P. and Kotthoff, L. (2009) 'Fselector: selecting attributes', *Vienna: R Foundation for Statistical Computing*.

Roy, S. *et al.* (2009) 'Exploiting amino acid composition for predicting protein-protein interactions', *PLoS ONE*, 4(11). doi: 10.1371/journal.pone.0007813.

RStudio Team, - (2016) ‘RStudio: Integrated Development for R’, *[Online] RStudio, Inc., Boston, MA URL http://www. rstudio. com*, p. RStudio, Inc., Boston, MA. doi: 10.1007/978-81-322-2340-5.

Rual, J.-F. *et al.* (2005) ‘Towards a proteome-scale map of the human protein–protein interaction network’, *Nature*, 437(7062), pp. 1173–1178. doi: 10.1038/nature04209.

Rudolph, J. (2007) ‘Inhibiting transient protein–protein interactions: lessons from the Cdc25 protein tyrosine phosphatases’, *Nature Reviews Cancer*. Nature Publishing Group, 7(3), p. 202.

Rutherford, S. L. (2000) ‘From genotype to phenotype: buffering mechanisms and the storage of genetic information’, *BioEssays*, 22(12), pp. 1095–1105. doi: 10.1002/1521-1878(200012)22:12<1095::AID-BIES7>3.0.CO;2-A.

Saeys, Y., Inza, I. and Larranaga, P. (2007) ‘A review of feature selection techniques in bioinformatics’, *Bioinformatics*, 23(19), pp. 2507–2517. doi: 10.1093/bioinformatics/btm344.

Saito, T. and Rehmsmeier, M. (2015) ‘The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets’, *Plos One*, 10(3), p. e0118432. doi: 10.1371/journal.pone.0118432.

Salzberg, S. L. (2018) ‘Open questions: How many genes do we have?’, *BMC Biology*. BMC Biology, 16(1), pp. 10–12. doi: 10.1186/s12915-018-0564-x.

Sammak, S. and Zinzalla, G. (2015) ‘Targeting protein-protein interactions (PPIs) of transcription factors: Challenges of intrinsically disordered proteins (IDPs) and regions (IDRs)’, *Progress in Biophysics and Molecular Biology*. Elsevier Ltd, 119(1), pp. 41–46. doi: 10.1016/j.pbiomolbio.2015.06.004.

Sanchez, C. *et al.* (1999) ‘Grasping at molecular interactions and genetic networks in

Drosophila melanogaster using FlyNets, an internet database', *Nucleic Acids Research*. doi: 10.1093/nar/27.1.89.

Schaefer, M. H. *et al.* (2012) 'Hippie: Integrating protein interaction networks with experiment based quality scores', *PLoS ONE*, 7(2), pp. 1–8. doi: 10.1371/journal.pone.0031826.

Schelhorn, S.-E., Lengauer, T. and Albrecht, M. (2008) 'An integrative approach for predicting interactions of protein regions', *Bioinformatics*. Oxford University Press, 24(16), pp. i35–i41.

Schmidhuber, J. (2015) 'Deep Learning in neural networks: An overview', *Neural Networks*. Elsevier Ltd, 61, pp. 85–117. doi: 10.1016/j.neunet.2014.09.003.

Scholkopf, B. and Smola, A. J. (2001) *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press.

Schwartz, A. S. *et al.* (2008) 'Cost-effective strategies for completing the interactome', *Nature methods*. Nature Publishing Group, 6(1), p. 55.

Schweitzer, B., Predki, P. and Snyder, M. (2003) 'Microarrays to characterize protein interactions on a whole-proteome scale', *PROTEOMICS*, 3(11), pp. 2190–2199. doi: 10.1002/pmic.200300610.

Seeler, J.-S. and Dejean, A. (2017) 'SUMO and the robustness of cancer', *Nature Reviews Cancer*. Nature Publishing Group, 17(3), p. 184.

Seemann, T. (2013) 'Ten recommendations for creating usable bioinformatics command line software', *GigaScience*, 2(1), pp. 2–4. doi: 10.1186/2047-217X-2-15.

Seshacharyulu, P. *et al.* (2012) 'Targeting the EGFR signaling pathway in cancer therapy', *Expert opinion on therapeutic targets*. Taylor & Francis, 16(1), pp. 15–31.

Shalev-Shwartz, S. and Ben-David, S. (2014) *Understanding Machine Learning: From Theory to Algorithms*. doi: 10.1017/CBO9781107298019.

Sharan, R. *et al.* (2005) 'Conserved patterns of protein interaction in multiple species', *Proceedings of the National Academy of Sciences*. doi: 10.1073/pnas.0409522102.

Shchemelinin, I., Sefc, L. and Necas, E. (2006) 'Protein kinases, their function and implication in cancer and other diseases.', *Folia biologica*.

Shen, J. *et al.* (2007) 'Predicting protein-protein interactions based only on sequences information', *Proceedings of the National Academy of Sciences*, 104(11), pp. 4337–4341. doi: 10.1073/pnas.0607879104.

Shervashidze, N. *et al.* (2011) 'Weisfeiler-Lehman Graph Kernels', *Journal of Machine Learning Research*, 12, pp. 2539–2561. doi: 10.1.1.232.1510.

Shi, M.-G. G. *et al.* (2010) 'Predicting protein-protein interactions from sequence using correlation coefficient and high-quality interaction dataset.', *Amino acids*, 38(3), pp. 891–9. doi: 10.1007/s00726-009-0295-y.

Shin, C. J. *et al.* (2009) 'Protein-protein interaction as a predictor of subcellular location', *BMC systems biology*. BioMed Central, 3(1), p. 28.

Shlens, J. (2014) 'A Tutorial on Principal Component Analysis', *Online Note httpwww snl salk edushlenspubnotespca pdf*, 2, pp. 1–16. doi: 10.1.1.115.3503.

Shlomi, T. *et al.* (2006) 'QPath: a method for querying pathways in a protein-protein interaction network', *BMC bioinformatics*. BioMed Central, 7(1), p. 199.

Shoemaker, B. a. and Panchenko, A. R. (2007) 'Deciphering protein-protein interactions. Part II. Computational methods to predict protein and domain interaction

partners.', *PLoS computational biology*, 3(4), p. e43. doi: 10.1371/journal.pcbi.0030043.

Shoemaker, B. A. and Panchenko, A. R. (2007b) 'Deciphering Protein–Protein Interactions. Part I. Experimental Techniques and Databases', *PLoS Computational Biology*. Public Library of Science, 3(3), p. e42. doi: 10.1371/journal.pcbi.0030042.

Shoyaib, M. and Abdullah-Al-Wadud, M. (2010) 'Predicting Protein-protein Interaction Using Amino Acid Sequence Information: A Computational Approach', *Plant Tissue Culture* …, 20(1), pp. 37–45. Available at: http://www.banglajol.info/index.php/PTCB/article/viewArticle/5963 (Accessed: 14 August 2013).

Sigismund, S., Avanzato, D. and Lanzetti, L. (2018) 'Emerging functions of the EGFR in cancer', *Molecular oncology*. Wiley Online Library, 12(1), pp. 3–20.

Singhal, M. and Resat, H. (2007) 'A domain-based approach to predict protein-protein interactions', *BMC Bioinformatics*, 8(1), p. 199. doi: 10.1186/1471-2105-8-199.

Sjöström, M., Rännar, S. and Wieslander, Å. (1995) 'Polypeptide sequence property relationships in Escherichia coli based on auto cross covariances', *Chemometrics and Intelligent Laboratory Systems*, 29(2), pp. 295–305. doi: 10.1016/0169-7439(95)80104-H.

Skrabanek, L. *et al.* (2008) 'Computational Prediction of Protein–Protein Interactions', *Molecular Biotechnology*, 38(1), pp. 1–17. doi: 10.1007/s12033-007-0069-2.

Smialowski, P. *et al.* (2010) 'The Negatome database: a reference set of non-interacting protein pairs.', *Nucleic acids research*, 38(Database issue), pp. D540-4. doi: 10.1093/nar/gkp1026.

Smith-Miles, K. A. (2008) 'Cross-disciplinary perspectives on meta-learning for

algorithm selection', *ACM Computing Surveys*, 41(1), pp. 1–25. doi: 10.1145/1456650.1456656.

Smith, G. R. and Sternberg, M. J. E. (2002) 'Prediction of protein-protein interactions by docking methods', *Current Opinion in Structural Biology*, 12(1), pp. 28–35. doi: 10.1016/S0959-440X(02)00285-3.

Smyth, M. S. and Martin, J. H. (2000) 'x ray crystallography.', *Molecular pathology : MP*, 53(1), pp. 8–14.

Söderberg, O. *et al.* (2006) 'Direct observation of individual endogenous protein complexes in situ by proximity ligation', *Nature Methods*, 3(12), pp. 995–1000. doi: 10.1038/nmeth947.

Spiegelman, B. M. and Heinrich, R. (2004) 'Biological control through regulated transcriptional coactivators', *Cell*. doi: 10.1016/j.cell.2004.09.037.

Sprinzak, E. and Margalit, H. (2001) 'Correlated sequence-signatures as markers of protein-protein interaction', *Journal of molecular biology*. Elsevier, 311(4), pp. 681–692. doi: 10.1006/jmbi.2001.4920.

Stelzl, U. *et al.* (2005) 'A Human Protein-Protein Interaction Network: A Resource for Annotating the Proteome', *Cell*, 122(6), pp. 957–968. doi: 10.1016/j.cell.2005.08.029.

Stelzl, U. and Wanker, E. E. (2006) 'The value of high quality protein–protein interaction networks for systems biology', *Current opinion in chemical biology*. Elsevier, 10(6), pp. 551–558.

Stumpf, M. M. P. H. *et al.* (2008) 'Estimating the size of the human interactome', *Proceedings of the …*, 105(19), pp. 6959–6964. doi: 10.1073/pnas.0708078105.

Stumpf, M. P. H., Wiuf, C. and May, R. M. (2005) 'Subnets of scale-free networks are

not scale-free: Sampling properties of networks', *Proceedings of the National Academy of Sciences*, 102(12), pp. 4221–4224. doi: 10.1073/pnas.0501179102.

Supek, F. *et al.* (2011) 'Revigo summarizes and visualizes long lists of gene ontology terms', *PLoS ONE*, 6(7). doi: 10.1371/journal.pone.0021800.

Syafrizayanti *et al.* (2014) 'Methods for analyzing and quantifying protein–protein interaction', *Expert Review of Proteomics*, 11(1), pp. 107–120. doi: 10.1586/14789450.2014.875857.

Szklarczyk, D. *et al.* (2017) 'The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible', *Nucleic Acids Research*. Oxford University Press, 45(D1), pp. D362–D368. doi: 10.1093/nar/gkw937.

Tang, X. *et al.* (2014) 'Predicting essential proteins based on weighted degree centrality', *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*. IEEE Computer Society Press, 11(2), pp. 407–418.

Tarca, A. L. *et al.* (2007) 'Machine Learning and Its Applications to Biology', *PLoS Computational Biology*, 3(6), p. e116. doi: 10.1371/journal.pcbi.0030116.

Thomas, M. C. and Chiang, C.-M. (2008) 'The general transcription machinery and general cofactors.', *Critical reviews in biochemistry and molecular biology*, 41(3), pp. 105–178. doi: 10.1080/10409230600648736.

Thompson, B. J. and Giancotti, F. G. (2018) 'Editorial overview: Cell signalling: Signal transduction to the nucleus, cytoskeleton, and organelles', *Current Opinion in Cell Biology*. Elsevier Ltd, 51, pp. iv–vii. doi: 10.1016/j.ceb.2018.04.005.

Titeca, K. *et al.* (2018) 'Discovering cellular protein-protein interactions: Technological strategies and opportunities', *Mass Spectrometry Reviews*, (June 2017), pp. 79–111. doi: 10.1002/mas.21574.

Tomii, K. and Kanehisa, M. (1996) ‘Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins’, *Protein Engineering, Design and Selection*. Oxford University Press, 9(1), pp. 27–36.

Toska, E. *et al.* (2014) ‘Prohibitin is required for transcriptional repression by the WT1–BASP1 complex’, *Oncogene*. Nature Publishing Group, 33(43), pp. 5100–5108. doi: 10.1038/onc.2013.447.

Toska, E. and Roberts, S. G. E. (2014) ‘Mechanisms of transcriptional regulation by WT1 (Wilms’ tumour 1)’, *Biochemical Journal*, 461(1), pp. 15–32. doi: 10.1042/BJ20131587.

Trivodaliev, K., Bogojeska, A. and Kocarev, L. (2014) ‘Exploring function prediction in protein interaction networks via clustering methods’, *PLoS ONE*, 9(6). doi: 10.1371/journal.pone.0099755.

Tsafou, K. *et al.* (2018) ‘Targeting Intrinsically Disordered Transcription Factors: Changing the Paradigm’, *Journal of Molecular Biology*. Elsevier Ltd, 430(16), pp. 2321–2341. doi: 10.1016/j.jmb.2018.04.008.

Tsai, C. F., Eberle, W. and Chu, C. Y. (2013) ‘Genetic algorithms in feature and instance selection’, *Knowledge-Based Systems*. Elsevier B.V., 39, pp. 240–247. doi: 10.1016/j.knosys.2012.11.005.

Turinsky, A. L. *et al.* (2010) ‘Literature curation of protein interactions: measuring agreement across major public databases’, *Database*. Oxford University Press, 2010.

Uetz, P. *et al.* (2000) ‘A comprehensive analysis of protein–protein interactions inSaccharomyces cerevisiae’, *Nature*, 403(6770), pp. 623–627. doi: 10.1038/35001009.

Ulasov, A. V., Rosenkranz, A. A. and Sobolev, A. S. (2018) ‘Transcription factors:

Time to deliver', *Journal of Controlled Release*, 269(September 2017), pp. 24–35. doi: 10.1016/j.jconrel.2017.11.004.

Uversky, V. N. (2013) 'Intrinsic Disorder-based Protein Interactions and their Modulators', *Current Pharmaceutical Design*, 19(23), pp. 4191–4213. doi: 10.2174/1381612811319230005.

Uversky, V. N. (2016) 'Paradoxes and wonders of intrinsic disorder: Complexity of simplicity', *Intrinsically Disordered Proteins*. doi: 10.1080/21690707.2015.1135015.

Uversky, V. N. (2017) 'Intrinsically disordered proteins in overcrowded milieu: Membrane-less organelles, phase separation, and intrinsic disorder', *Current Opinion in Structural Biology*. doi: 10.1016/j.sbi.2016.10.015.

Uversky, V. N. (2018) 'Intrinsic Disorder, Protein–Protein Interactions, and Disease', in *Advances in Protein Chemistry and Structural Biology*. 1st edn. Elsevier Inc., pp. 85–121. doi: 10.1016/bs.apcsb.2017.06.005.

Uversky, V. N. and Dunker, A. K. (2010) 'Understanding protein non-folding', *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, 1804(6), pp. 1231–1264. doi: 10.1016/j.bbapap.2010.01.017.

Valente, G. T. *et al.* (2013) 'The development of a universal in silico predictor of protein-protein interactions.', *PloS one*, 8(5), p. e65587. doi: 10.1371/journal.pone.0065587.

Vapnik, V. N. (1999) 'An overview of statistical learning theory', *IEEE Transactions on Neural Networks*, 10(5), pp. 988–999. doi: 10.1109/72.788640.

Vaquerizas, J. M. *et al.* (2009) 'A census of human transcription factors: Function, expression and evolution', *Nature Reviews Genetics*, 10(4), pp. 252–263. doi: 10.1038/nrg2538.

Vázquez, A. *et al.* (2003) 'Modeling of protein interaction networks', *Complexus*. Karger Publishers, 1(1), pp. 38–44. doi: 10.1159/000067642.

Veljkovic, V. (1980) *A theoretical approach to the preselection of carcinogens and chemical carcinogenesis*. Gordon & Breach Publishing Group.

Vidal, M., Cusick, M. E. and Barabási, A. L. (2011) 'Interactome networks and human disease', *Cell*, 144(6), pp. 986–998. doi: 10.1016/j.cell.2011.02.016.

Vihinen, M., Torkkila, E. and Riikonen, P. (1994) 'Accuracy of protein flexibility predictions', *Proteins: Structure, Function, and Bioinformatics*, 19(2), pp. 141–149. doi: 10.1002/prot.340190207.

Vinogradova, O. and Qin, J. (2011) 'NMR as a Unique Tool in Assessment and Complex Determination of Weak Protein–Protein Interactions', in *Topics in current chemistry*, pp. 35–45. doi: 10.1007/128_2011_216.

Voit, E. O. (2013) 'Biochemical Systems Theory: A Review', *ISRN Biomathematics*. doi: 10.1155/2013/897658.

Voter, A. F., Manthei, K. A. and Keck, J. L. (2016) 'A high-throughput screening strategy to identify protein-protein interaction inhibitors that block the Fanconi Anemia DNA repair pathway', *Journal of biomolecular screening*. SAGE Publications Sage CA: Los Angeles, CA, 21(6), pp. 626–633.

Vreven, T. *et al.* (2014) 'Evaluating template-based and template-free protein-protein complex structure prediction', *Briefings in Bioinformatics*, 15(2), pp. 169–176. doi: 10.1093/bib/bbt047.

Wainer, J. (2016) 'Comparison of 14 different families of classification algorithms on 115 binary datasets', (2014). Available at: http://arxiv.org/abs/1606.00930.

Wan, K. K., Park, J. and Suh, J. K. (2002) 'Large scale statistical prediction of protein-protein interaction by potentially interacting domain (PID) pair', *Genome Informatics*. Japanese Society for Bioinformatics, 13, pp. 42–50.

Wang, E. T. *et al.* (2008) 'Alternative isoform regulation in human tissue transcriptomes', *Nature*. doi: 10.1038/nature07509.

Wang, J. C. (1985) 'DNA topoisomerases', *Annual review of biochemistry*. Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA, 54(1), pp. 665–697.

Wang, L. *et al.* (2017) 'An ensemble approach for large-scale identification of protein-protein interactions using the alignments of multiple sequences', *Oncotarget*, 8(3), pp. 5149–5159. doi: 10.18632/oncotarget.14103.

Wang, Y. *et al.* (2010) 'Sequence-based protein-protein interaction prediction via support vector machine', *Journal of Systems Science and Complexity*, 23(5), pp. 1012–1023. doi: 10.1007/s11424-010-0214-z.

Waris, M. *et al.* (2016) 'Identification of DNA binding proteins using evolutionary profiles position specific scoring matrix', *Neurocomputing*, 199, pp. 154–162. doi: 10.1016/j.neucom.2016.03.025.

Wass, M. N. *et al.* (2011) 'Towards the prediction of protein interaction partners using physical docking', *Molecular Systems Biology*. Nature Publishing Group, 7(469), pp. 1–8. doi: 10.1038/msb.2011.3.

Watkins, A. M. and Arora, P. S. (2014) 'Anatomy of β-strands at protein-protein interfaces', *ACS Chemical Biology*. doi: 10.1021/cb500241y.

Weston, S. (2017) 'Getting Started with doMC and foreach'.

Whitford, D. (2005) *Proteins : structure and function*. J. Wiley & Sons. Available at: https://www.wiley.com/en-us/Proteins%3A+Structure+and+Function-p-9780471498940 (Accessed: 27 August 2018).

Wickham, H. *et al.* (2016) 'dplyr: A Grammar of Data Manipulation. R package version 0.5. 0'. R Core Development Team Vienna.

Wickham, H. (2018) 'Package "plyr"'.

Wilkinson, K. A. and Henley, J. M. (2010) 'Mechanisms, regulation and consequences of protein SUMOylation', *Biochemical Journal*. Portland Press Limited, 428(2), pp. 133–145.

Williams, R. M. *et al.* (2001) 'The protein non-folding problem: amino acid determinants of intrinsic order and disorder.', *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 100, pp. 89–100. doi: 10.1.1.32.6125.

Wolpert, D. H. (2002) 'The Supervised Learning No-Free-Lunch Theorems', in Roy R., Köppen M., Ovaska S., Furuhashi T., H. F. (ed.) *Soft Computing and Industry*. London: Springer London, pp. 25–42. doi: 10.1007/978-1-4471-0123-9_3.

Wright, M. N. and Ziegler, A. (2017) 'ranger : A Fast Implementation of Random Forests for High Dimensional Data in C++ and R', *Journal of Statistical Software*, 77(1). doi: 10.18637/jss.v077.i01.

Wright, P. E. and Dyson, H. J. (2015) 'Intrinsically disordered proteins in cellular signalling and regulation', *Nature Reviews Molecular Cell Biology*. Nature Publishing Group, 16(1), pp. 18–29. doi: 10.1038/nrm3920.

Wu, J. *et al.* (2009) 'Integrated network analysis platform for protein-protein interactions.', *Nature methods*, 6(1), pp. 75–7. doi: 10.1038/nmeth.1282.

Wu, S. yao, Zhang, Q. and Wu, M. (2017) ‘Cold-start link prediction in multi-relational networks’, *Physics Letters, Section A: General, Atomic and Solid State Physics*. Elsevier B.V., 381(39), pp. 3405–3408. doi: 10.1016/j.physleta.2017.08.046.

Wu, W. S. and Lai, F. J. (2015) ‘Functional redundancy of transcription factors explains why most binding targets of a transcription factor are not affected when the transcription factor is knocked out’, *BMC Systems Biology*. doi: 10.1186/1752-0509-9-S6-S2.

Xia, J.-F., Han, K. and Huang, D.-S. (2010) ‘Sequence-based prediction of protein-protein interactions by means of rotation forest and autocorrelation descriptor.’, *Protein and peptide letters*, 17(1), pp. 137–145. doi: 10.2174/092986610789909403.

Xiao, N. *et al.* (2015) ‘protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences’, *Bioinformatics*, 31, pp. 1857–1859. doi: 10.1093/bioinformatics/btv042.

Xiao, N., Xu, Q. and Cao, D. (2014) ‘protr : R package for generating various numerical representation schemes of protein sequence’.

Xing, S. *et al.* (2016a) ‘Techniques for the Analysis of Protein-Protein Interactions in Vivo.’, *Plant physiology*. American Society of Plant Biologists, 171(2), pp. 727–58. doi: 10.1104/pp.16.00470.

Xing, S. *et al.* (2016b) ‘Topical Review on Protein-Protein Interaction Techniques Techniques for the Analysis of Protein-Protein Interactions in Vivo 1[OPEN] Critical discussion of technological limitations and advantages of the most prominent in vivo protein-protein interaction ’, *Plant Physiology Ò*, 171, pp. 727–758. doi: 10.1104/pp.16.00470.

Xue, B. *et al.* (2010) ‘Archaic chaos: Intrinsically disordered proteins in Archaea’, *BMC Systems Biology*. doi: 10.1186/1752-0509-4-S1-S1.

Xue, B. *et al.* (2014) ‘Structural disorder in viral proteins’, *Chemical Reviews*. doi: 10.1021/cr4005692.

Xue, B., Dunker, A. K. and Uversky, V. N. (2012) ‘Orderly order in protein intrinsic disorder distribution: Disorder in 3500 proteomes from viruses and the three domains of life’, *Journal of Biomolecular Structure and Dynamics*. doi: 10.1080/07391102.2012.675145.

Yan, S. *et al.* (2018) *Application of Systems Biology in the Research of TCM Formulae*, *Systems Biology and Its Application in TCM Formulas Research*. Elsevier Inc. doi: 10.1016/B978-0-12-812744-5.00003-5.

Yang, J. and Zhang, X. D. (2016) ‘Predicting missing links in complex networks based on common neighbors and distance’, *Scientific Reports*. Nature Publishing Group, 6(November), pp. 1–10. doi: 10.1038/srep38208.

Yang, P. *et al.* (2010) ‘A Review of Ensemble Methods in Bioinformatics’, *Current Bioinformatics*, 5(4), pp. 296–308. doi: 10.2174/157489310794072508.

Ye, X., Wang, G. and Altschul, S. F. (2011) ‘An assessment of substitution scores for protein profile–profile comparison’, *Bioinformatics*, 27(24), pp. 3356–3363. doi: 10.1093/bioinformatics/btr565.

Yeoh, E.-J. *et al.* (2002) ‘Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling’, *Cancer Cell*, 1(2), pp. 133–143. doi: 10.1016/S1535-6108(02)00032-6.

Yin, L. *et al.* (2017) ‘An evidential link prediction method and link predictability based on Shannon entropy’, *Physica A: Statistical Mechanics and its Applications*. Elsevier B.V., 482, pp. 699–712. doi: 10.1016/j.physa.2017.04.106.

Yip, K. Y., Cheng, C. and Gerstein, M. (2013) ‘Machine learning and genome annotation: a match meant to be?’, *Genome biology*. BioMed Central, 14(5), p. 205.

You, Z.-H. *et al.* (2013) ‘Prediction of protein-protein interactions from amino acid sequences with ensemble extreme learning machines and principal component analysis.’, *BMC bioinformatics*. BioMed Central Ltd, 14 Suppl 8(Suppl 8), p. S10. doi: 10.1186/1471-2105-14-S8-S10.

You, Z.-H. *et al.* (2014) ‘Prediction of protein-protein interactions from amino acid sequences using a novel multi-scale continuous and discontinuous feature set’, *BMC Bioinformatics*. BioMed Central Ltd, 15(Suppl 15), p. S9. doi: 10.1186/1471-2105-15-S15-S9.

You, Z.-H., Chan, K. C. C. and Hu, P. (2015) ‘Predicting Protein-Protein Interactions from Primary Protein Sequences Using a Novel Multi-Scale Local Feature Representation Scheme and the Random Forest’, *PLOS ONE*. Edited by F. Fraternali, 10(5), p. e0125811. doi: 10.1371/journal.pone.0125811.

Yu, C.-Y. Y., Chou, L.-C. C. and Chang, D. T.-H. (2010) ‘Predicting protein-protein interactions in unbalanced data using the primary structure of proteins.’, *BMC bioinformatics*, 11, p. 167. doi: 10.1186/1471-2105-11-167.

Yu, J. F. *et al.* (2016) ‘Natural protein sequences are more intrinsically disordered than random sequences’, *Cellular and Molecular Life Sciences*. doi: 10.1007/s00018-016-2138-9.

Zahiri, J. *et al.* (2014) ‘LocFuse: Human protein–protein interaction prediction via classifier fusion using protein localization information’, *Genomics*. Elsevier B.V., 104(6), pp. 496–503. doi: 10.1016/j.ygeno.2014.10.006.

Zahiri, J., Bozorgmehr, J. H. and Masoudi-Nejad, A. (2013) ‘Computational Prediction of Protein-Protein Interaction Networks: Algo-rithms and Resources.’, *Current*

genomics, 14(6), pp. 397–414. doi: 10.2174/1389202911314060004.

Zaki, N., Efimov, D. and Berengueres, J. (2013) 'Protein complex detection using interaction reliability assessment and weighted clustering coefficient', *BMC Bioinformatics*. doi: 10.1186/1471-2105-14-163.

Zhang, S. B. and Tang, Q. R. (2016) 'Protein-protein interaction inference based on semantic similarity of Gene Ontology terms', *Journal of Theoretical Biology*. Elsevier, 401, pp. 30–37. doi: 10.1016/j.jtbi.2016.04.020.

Zhao, X.-W., Ma, Z.-Q. and Yin, M.-H. (2012) 'Predicting Protein-Protein Interactions by Combing Various Sequence- Derived Features into the General Form of Chou's Pseudo Amino Acid Composition', *Protein and Peptide Letters*, 19(5), pp. 492–500. doi: 10.2174/092986612800191080.

Zubek, J. *et al.* (2015) 'Multi-level machine learning prediction of protein–protein interactions in *Saccharomyces cerevisiae*', *PeerJ*, 3, p. e1041. doi: 10.7717/peerj.1041. ectrometry Coupled Techniques', in *Advances in biochemical engineering/biotechnology*, pp. 67–80. doi: 10.1007/10_2007_091.

8 Objavljeni radovi

U toku rada i na osnovu rezultata postignutih u ovoj disertaciji objavljeni su sledeći radovi u vrhunskim i istaknutim međunarodnim časopisima:

1. Neven Sumonja, Branislava Gemovic, Nevena Veljkovic, Vladimir Perovic, „Automated feature engineering improves prediction of protein–protein interactions“, Amino acids, 2019 Jul 5:1-4. https://doi.org/10.1007/s00726-019-02756-9

2. Vladimir Perovic, Neven Sumonja, Lindsey A. Marsh, Sandro Radovanovic, Milan Vukicevic, Stefan G. Roberts, Nevena Veljkovic, „IDPpi: Protein-protein interaction analyses of human intrinsically disordered proteins“, Scientific reports, 2018 Jul 12;8(1):10563. https://doi.org/10.1038/s41598-018-28815-x

3. Branislava Gemovic, Neven Sumonja, Radoslav Davidovic, Vladimir Perovic, Nevena Veljkovic, “Mapping of Protein-Protein Interactions: Web-Based Resources for Revealing Interactomes”, Current Medicinal Chemistry (2018) 25: 1. https://doi.org/10.2174/0929867325666180214113704

4. Vladimir Perovic, Neven Sumonja, Branislava Gemovic, Eneda Toska, Stefan G Roberts, Nevena Veljkovic, „TRI_tool: a web-tool for prediction of protein–protein interactions in human transcriptional regulation“, Bioinformatics, Volume 33, Issue 2, 15 January 2017, Pages 289–291, https://doi.org/10.1093/bioinformatics/btw590

Biografija autora

Neven Šumonja je rođen 26.05.1985. godine u Zagrebu (Republika Hrvatska). Osnovnu školu i srednju školu Gimnaziju završio je u Banjoj Luci (Republika Srpska, BiH).

Školske 2004/2005. godine upisao je studije na Prirodno-matematičkom fakultetu, Univerziteta u Banjoj Luci, Odsek za biologiju, opšti smjer. Diplomirao 26.09.2011. godine na Prirodno-matematičkom fakultetu, Odsek za biologiju sa ocenom 10 na diplomskom ispitu, odbranivši diplomski rad pod nazivom „Genski markeri na X hromozomu u humanoj identifikaciji", pod rukovodstvom mentora Prof. dr Stojka Vidovića, i prosečnom ocenom 8,70.

Školske 2011/2012. godine upisao je prvu godinu doktorskih akademskih studija na Biološkom fakultetu Univerziteta u Beogradu, na studijskom programu Molekularna biologija, modul Molekularna biologija eukariota.

Od marta 2015. do marta 2017. volontirao je u Centru za multidisciplinarna istraživanja i inženjering, u Institutu za nuklearne nauke "Vinča" na projektu OI173001, rukovodilac projekta dr Veljko Veljković, naučni savetnik.

Od marta 2017. godine, zaposlen je u Laboratoriji za bioinformatiku i računarsku hemiju, u Institutu za nuklearne nauke "Vinča", gde je trenutno angažovan na projektu „Primena EIIP/ISM bioinformatičke platforme u otkrivanju novih terapeutskih targeta i potencijalnih terapeutskih molekula", broj 173001, Ministarstva prosvete, nauke i tehnološkog razvoja Republike Srbije. Iste godine je izabran u zvanje istraživač saradnik.

Od 2016. godine učestvuje u COST akciji BM1405 "Non-globular proteins - from sequence to structure, function and application in molecular physiopathology (NGP-NET)". U sklopu ove akcije ostvaruje kratku posetu CRBM institutu, Montpellier, Francuska pod rukovodstvom dr Andrey Kajave.

Do sada je objavio 6 naučnih radova, od toga su 3 u vrhunskim međunarodnim časopisima, jedan u istaknutom međunarodnom časopisu, jedan u međunarodnom časopisu i jedan u domaćem časopisu. Pored toga objavio je i 13 saopštenja sa međunarodnih konferencija.

Spisak slika i tabela

Slike

Tabele

Prilog 1

Изјава о ауторству

Име и презиме аутора Невен Шумоња

Број индекса Б 3039, 2011

Изјављујем

да је докторска дисертација под насловом

Биоинформатички модели за аутоматско мапирање интеракција између протеина код човека

- резултат сопственог истраживачког рада;
- да дисертација у целини ни у деловима није била предложена за стицање друге дипломе према студијским програмима других високошколских установа;
- да су резултати коректно наведени и
- да нисам кршио/ла ауторска права и користио/ла интелектуалну својину других лица.

Потпис аутора

У Београду, 21.06.2019.

Невен Шумоња

Prilog 2

Изјава о истоветности штампане и електронске верзије докторског рада

Име и презиме аутора Невен Шумоња

Број индекса Б 3039, 2011

Студијски програм Молекуларна биологија

Наслов рада Биоинформатички модели за аутоматско мапирање интеракција између протеина човека

Ментор др Душанка Савић Павићевић, редовни професор Биолошког факултета Универзитета у Београду

др Владимир Перовић, научни сарадник Института за нуклеарне науке Винча Универзитета у Београду

Изјављујем да је штампана верзија мог докторског рада истоветна електронској верзији коју сам предао/ла ради похрањена у **Дигиталном репозиторијуму Универзитета у Београду.**

Дозвољавам да се објаве моји лични подаци везани за добијање академског назива доктора наука, као што су име и презиме, година и место рођења и датум одбране рада.

Ови лични подаци могу се објавити на мрежним страницама дигиталне библиотеке, у електронском каталогу и у публикацијама Универзитета у Београду.

У Београду, 21.06.2019.

Потпис аутора

Невен Шумоња

Prilog 3

Изјава о коришћењу

Овлашћујем Универзитетску библиотеку „Светозар Марковић“ да у Дигитални репозиторијум Универзитета у Београду унесе моју докторску дисертацију под насловом:

Биоинформатички модели за аутоматско мапирање интеракција између протеина код човека

која је моје ауторско дело.

Дисертацију са свим прилозима предао/ла сам у електронском формату погодном за трајно архивирање.

Моју докторску дисертацију похрањену у Дигиталном репозиторијуму Универзитета у Београду и доступну у отвореном приступу могу да користе сви који поштују одредбе садржане у одабраном типу лиценце Креативне заједнице (Creative Commons) за коју сам се одлучио/ла.

1. Ауторство (CC BY)

2. Ауторство – некомерцијално (CC BY-NC)

3. Ауторство – некомерцијално – без прерада (CC BY-NC-ND)

4. Ауторство – некомерцијално – делити под истим условима (CC BY-NC-SA)

5. Ауторство – без прерада (CC BY-ND)

6. Ауторство – делити под истим условима (CC BY-SA)

(Молимо да заокружите само једну од шест понуђених лиценци.
Кратак опис лиценци је саставни део ове изјаве).

Потпис аутора

У Београду, 21.06.2019.

1. **Ауторство**. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце, чак и у комерцијалне сврхе. Ово је најслободнија од свих лиценци.

2. **Ауторство – некомерцијално**. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца не дозвољава комерцијалну употребу дела.

3. **Ауторство – некомерцијално – без прерада**. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, без промена, преобликовања или употребе дела у свом делу, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца не дозвољава комерцијалну употребу дела. У односу на све остале лиценце, овом лиценцом се ограничава највећи обим права коришћења дела.

4. **Ауторство – некомерцијално – делити под истим условима**. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце и ако се прерада дистрибуира под истом или сличном лиценцом. Ова лиценца не дозвољава комерцијалну употребу дела и прерада.

5. **Ауторство – без прерада**. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, без промена, преобликовања или употребе дела у свом делу, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца дозвољава комерцијалну употребу дела.

6. **Ауторство – делити под истим условима**. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце и ако се прерада дистрибуира под истом или сличном лиценцом. Ова лиценца дозвољава комерцијалну употребу дела и прерада. Слична је софтверским лиценцама, односно лиценцама отвореног кода.