Nastavno-naučnom veću
Matematičkog fakulteta
Univerziteta u Beogradu

Odlukom Nastavno-naučnog veća Matematičkog fakulteta Univerziteta u Beogradu donetom na sednici održanoj 23.01.2015. imenovani smo u Komisiju za pregled i ocenu doktorske disertacije "Strukturna predikcija funkcije proteina i odnos funkcionalnih kategorija i neuređenosti" kandidata Jovane Kovačević, diplomiranog matematičara. Posle pregledanja podnetog rukopisa podnosimo sledeći

IZVEŠTAJ

• Biografski podaci

Jovana Kovačević je rođena 6.5.1983. godine u Beogradu. Završila je Petu beogradsku gimnaziju i diplomirala na Matematičkom fakultetu, smer Računarstvo i informatika. Nakon završenih osnovnih studija, 2007. godine upisala je doktorske studije na istom smeru. Položila je sve ispite predviđene planom i programom doktorskih studija.

Od oktobra 2007. godine angažovana je na Matematičkom fakultetu kao saradnik u nastavi, a od oktobra 2009. godine kao asistent. U svom dosadašnjem radu na Matematičkom fakultetu držala je vežbe iz niza predmeta na osnovnim i master studijama:

- Programiranje I i II
- Programske paradigme
- Arhitektura i operativni sistemi
- Primena računara u biologiji
- Funkcionalno programiranje
- Bioinformatika

Osnovna oblast interesovanja joj je Istraživanje podataka u bioinformatici. Od 2011. godine učesnik je naučno istraživačkog projekta "Automatsko rezonovanje i istraživanje podataka", Ministarstva prosvete, nauke i tehnološkog razvoj Republike Srbije, 174021. Od 2010.-2011. godine učestvovala je na projektu „Automatsko rezonovanje i istraživanje velikih količina podataka i teksta”,144030, Ministarstva prosvete, nauke i tehnološkog razvoja Republike Srbije.
Objavila je (ili pripremila za objavljivanje) veći broj naučnih radova i učestvovala na nekoliko međunarodnih i domaćih konferencija.

• Naučni radovi

- Gordana Pavlović-Lažetić, Nenad Mitić, Jovana Kovačević, Zoran Obradović, Saša Malkov, Miloš Beljanski, **Bioinformatics analysis of disordered proteins in prokaryotes**, *BMC Bioinformatics (izdavač: BioMed Central Ltd.,* ISSN 1471-2105, *IF 2.58), 12:66*, 2011.

- Jovana Kovačević, **Computational Analysis of Position-dependent Disorder Content in DisProt Database**, *Genomics, Proteomics, Bioinformatics (izdavač: Elsevier,* ISSN: 1672-0229*), volume 10, number 3, pages 158-165*, 2012.
- J. Graovac, J. Kovačević, G. Pavlović-Lažetić, **Language Independent n-Gram-Based Text Categorization with Weighting Factors: A Case Study,** *JIDM - Journal of Information and Data Management, Vol. 6, No. 1, Pages 4-17, 2015.*
- **Jovana Kovačević, Jelena Graovac, Application of a Structural Support Vector Machine method to N-gram based text classification in Serbian,** Journal for Digital Humanities Infotheca, ISSN: 1450-9687 (print edition) ISSN: 2217-9461 (online edition) (prihvaćen za štampu**)**
- Jelena Graovac, Jovana Kovačević, and Gordana Pavlović-Lažetič, **Hierarchical vs. flat n-gram-based text categorization: can we do better?** under review in Computer Science and Information Systems, ISSN: 1820-0214 (Print) 2406-1018 (Online)
- Nenad S. Mitić, Saša N. Malkov, Jovana J. Kovačević, Gordana M. Pavlović-Lažetić, Miloš V. Beljanski**, Intrinsically disordered proteins /protein regions: implications for genomic and environmental adaptation of Archaea and Bacteria,** under review in BMC Bioinformatics**,** ISSN: 1471-2105 (electronic version)

• Učešće na konferencijama

- Jovana Kovačević, Predrag Radivojac, Gordana Pavlović-Lažetić **On protein function prediction methods**, *Proceedings of the Theoretical Approaches to Bioinformation Systems, 17-22 September 2013, Belgrade, Institute of Physics, ISBN: 978-86-82441-37-3*
- Jovana Kovačević, **Bioinformatics study of protein disorder content with respect to its position and protein function,** *Proceedings of the Data mining in bioinformatics*, *Beograd, Srbija, 26-28.6.2012., ISBN: 978-86-7589-085-0, strane 29-33*
- Gordana Pavlović Lažetić, Vesna Pajić, Nenad Mitić, Jovana Kovačević, Miloš Beljanski **Mining correlations and associations for organism characteristics in prokaryotes – an integrative approach**, *Proceedings of 2nd International Work-Conference on Bioinformatics and Biomedical Engineering, April 7-9 2014, Granada, Spain*
- J.Kovacevic and G.Pavlovic-Lazetic. **Predictive models based on support vector machines for structured outputs**. In Mathematical Data Science-Book of Abstracts, Belgrade, Serbia, Mathematical Institute SASA, 22. 6. 2015., 2015.

• Studijski boravci i letnje škole

- Studijski boravak na Univerzitetu u Indijani, Fakultet za informatiku i računarstvo, mart-april 2013. rad u okviru bioinformatičke istraživačke grupe prof. Predraga Radivojca
- Letnja škola „Bioinformatics and structural biology of intrinsically disordered proteins“, Budimpešta, Mađarska, 10-15.10.2011.
- Letnja škola „Computational methods in molecular biology“, ICGEB, Trst, Italija, 20-26.6.2010.

• Predmet disertacije

Disertacija "Strukturna predikcija funkcije proteina i odnos funkcionalnih kategorija i neuređenosti" pripada interdisciplinarnoj oblasti računarstva poznatoj kao istraživanje podataka (engl. Data mining) koja se nalazi u preseku veštačke inteligencije, mašinskog učenja, statistike i sistema baza podataka. Domen na koji se primenjuje otkrivanje znanja koje predstavlja suštinu istraživanja podataka jeste funkcija proteina kao jedna od centralnih tema nauka o životu (engl. life sciences); njihov razvoj u savremenom svetu ne može se zamisliti bez primene računarskih tehnologija, a problemi koje otvaraju povratno utiču na metodološki razvoj računarskih nauka. Predmet ove disertacije je upravo jedna takva sprega.

Proteini predstavljaju najvažniju grupu biomolekula u živom svetu. Poznavanje i razumevanje njihovih funkcija je esencijalno u istraživanju bilo kog biološkog procesa, sa posebnim naglaskom na oboljenja ljudi, s obzirom da se mnoga od njih mogu pojaviti zbog funkcionalnih mutacija.

Aktuelne eksperimentalne metode za funkcionalnu anotaciju proteina su suviše vremenski i materijalno zahtevne za veliki priliv novootkrivenih proteinskih sekvenci čiji broj raste sa svakim sekvencionisanim genomom. Zbog toga je poslednjih godina intenziviran razvoj softverskih alata za automatsku predikciju funkcije proteina, koji mogu predstavljati prvi korak u usmeravanju skupih laboratorijskih resursa.

Klasičan pristup problemu predikcije funkcije podrazumeva takozvano prenošenje funkcionalne anotacije sa sličnih proteina, za koje su funkcije eksperimentalno utvrđene, na dati protein, pri čemu se sličnost može utvrđivati globalnim, lokalnim i višestrukim poravnanjem, na osnovu zajedničkih šablona u sekvenci proteina, na osnovu evolutivne povezanosti, na osnovu slične sekundarne strukture itd. Napredniji pristup se sastoji u primeni različitih algoritama mašinskog učenja koji se treniraju na skupovima već funkcionalno anotiranih proteina, gde na osnovu odabranih karakteristika proteina pokušavamo da zaključimo koje funkcije ima dati protein. U zavisnosti od toga da li koristimo nadgledane ili nenadgledane metode mašinskog učenja, problemu predviđanja funkcije možemo pristupiti kao problemu klasifikacije ili klasterovanja. U okviru prvog pristupa korišćeni su sledeći algoritmi: metod podržavajućih vektora (*A.Sokolov et al*, . *Journal of Bioinformatics and Computational Biology 2010.*), neuralne mreže (*W.Clark i P.Radovijac, Proteins, Structure, Function, Bioinformatics, 2011.*), višeslojni perceptroni (*Mateos et al, Genome Research, 2010.*), Markovljeva slučajna polja (*Deng et al, Journal of Computational Biology, 2003.*), bajesovske mreže (*Chen i Xu, Nucleic Acid Research, 2004.*), kernel logistička regresija (*Lee et*

al, Journal of Integrative Biology, 2006.) i druge. U okviru drugog pristupa korišćeni su sledeći algoritmi: Markovljevo klasterovanje (*Pereira-Leal et al, Proteins, 2004*), hijerarhijsko klasterovanje (*Rives i Galitski,Proceedings of the National Academy of Sciences USA, 2003*), verovatnosni grafički modeli (*Segal et al, Bioinformatics, 2003*), model skrivenog modularnog slučajnog polja (*Shiga et al, Bioinformatics, 2007*) i drugi. Naučna zajednica koja se bavi predviđanjem funkcije proteina korišćenjem tehnika mašinskog učenja svake dve godine održava takmičenje najnovijih prediktora funkcije (*CAFA – Critical Assessment of Function Annotation experiment*).

Uobičajeni način predstavljanja funkcije proteina definisan je kroz *Gene Ontology (GO)* projekat. GO razdvaja sve moguće funkcije proteina na tri različita usmerena aciklička grafa: ontologija molekulskih funkcija, ontologija bioloških procesa i ontologija ćelijskih komponenti. Svaki čvor u ontologiji predstavlja jednu funkciju, a svaka grana predstavlja vezu između čvorova – funkcija koje povezuje. Svaki čvor definiše specifičniju funkciju nego njegov predak. Time se određivanje funkcije proteina svodi na nalaženje najpogodnijeg podgrafa određene ontologije na osnovu kompatibilnosti sa proteinskom sekvencom koja predstavlja ulazni podatak. Svaka ontologija ima nekoliko hiljada čvorova i u okviru svake anotirano je po nekoliko desetina hiljada proteina.

Predmet disertacije u ovom aspektu jeste usavršavanje postupka predviđanja funkcije proteina primenom metode stukturalnih podržavajućih vektora koja predstavlja jedno proširenje metode podržavajućih vektora za strukturalni izlaz. Razvijen je model i programski sistem za automatsko predviđanje funkcije proteina na osnovu njegove sekvence korišćenjem ove metode i izvršena je pozitivna evaluacija metode i poređenje sa rezultatima prediktora koji su učestvovali na prestižnoim CAFA takmičenju.

Drugi način dodeljivanja funkcije proteinu je njegovo klasifikovanje u neku od funkcionalnih kategorija proteina koje predstavljaju rezultat klasterovanja proteina iz kompletnih genoma različitih organizama na osnovu njihove evolutivne povezanosti. Na ovaj način je nastala javno dostupna baza proteina pod nazivom COG (*Cluster of Orthologous Groups*). Proteini su tako podeljeni na 25 klastera, u svakom klasteru su proteini koji obavljaju istu funkciju, a oni su dalje grupisani u četiri funkcionalne kategorije: *Cellular processes and signaling*, *Information storage and processing*, *Metabolism* i *Poorly characterized*.

Funkcija proteina je, prema paradigmi struktura-funkcija, vezana za strukturu koju protein zauzima u prostoru - najčešće uređenu prostornu strukturu u obliku spirale ili ravni. Međutim, za neke proteine eksperimentalno je pokazano da su neuređeni, što znači da nemaju fiksiranu 3D strukturu, u pojedinim regionima proteina ili kompletno. Pored eksperimentalnih metoda, postoji veliki broj alata za automatsko određivanje neuređenosti proteinske sekvence (*Oates et al, Nucleic Acid Research, 2013*). Poznato je da postoji veza neuređenosti proteina sa njihovom funkcijom (ulogom u organizmu) kao i sa različitim genomičkim, metaboličkim i ekološkim karakteristikama organizma kom protein pripada, kao i veza sa lokacijom pojavljivanja neuređenosti u proteinu. Istraživanja nad neuređenim proteinima su od velikog značaja jer je utvrđeno da su oni povezani sa nekim od najtežih savremenih bolesti.

Predmet disertacije u ovom aspektu jeste utvrđivanje odnosa neuređenosti proteina i funkcionalnih kategorija kojima pripadaju. Izvršena je opsežna analiza velikog skupa prokariotskih i eukariotskih proteina u odnosu na njihove funkcionalne kategorije i u odnosu na sastav neuređenih regiona sa rezultatima koji se mogu primeniti kako na razvoj sofisticiranih prediktora funkcije tako i prediktora neuređenosti.

- **Prikaz disertacije**

Doktorska disertacija "Strukturna predikcija funkcije proteina i odnos funkcionalnih kategorija i neuređenosti" sastoji se iz pet poglavlja osnovnog teksta i zaključka. Iscrpan spisak literature iz oblasti, korišćene u ovoj disertaciji, sastoji se od 109 bibliografskih jedinica.
U Uvodu se opisuje značaj teme disertacije, prikazuju se postojeće metode za određivanje funkcije proteina i način rešavanja ovog problema u ovoj disertaciji, kao i ciljevi i organizacija teksta disertacije. Preostali tekst disertacije može da se grupiše u dve celine.

Prvu celinu čine poglavlja 2-4. U Glavi 2 uvodi se strukturna klasifikacija kao pristup klasifikaciji u kome je rezultat klasifikacije struktuirani objekat kao što je, na primer, niz, niska, stablo ili graf. Uvodi se pojam margine i maksimizacije margine u binarnoj i strukturnoj klasifikaciji na primeru metode strukturnih podržavajućih vektora (engl. Structured Support Vector Machine, SSVM). Formulišu se optimizacioni problemi za SSVM, i prikazuje algoritam odsecajućih ravni za treniranje SSVM sa strukturnim izlazom koji se u prilagođenom obliku koristi u ovoj disertaciji za rešavanje problema određivanja funkcije proteina.

Glava 3 prikazuje problem predviđanja funkcije proteina, postojeće načine za njegovo rešavanje i rešavanje odgovarajućeg optmizacionog problema primenom SSVM metode. Prikazana je najznačajnija baza proteinskih funkcija Gene Ontology (GO) koja se sastoji od tri ontologije struktuirane u obliku usmerenog acikličkog grafa gde se u čvorovima nalaze funkcije a grane koje ih povezuju definišu relaciju "is-a". Predstavljen je način rešavanja problema predviđanja funkcije proteina metodom SSVM kroz razvoj zajedničke reprezentacije ulaznih podataka – proteinskih sekvenci – i izlaznih rezultata – podgrafova GO koji opisuju funkcije proteina, funkcije sličnosti odnosno različitosti grafova funkcija (predviđenih i stvarnih) sa kojima su vršeni elsperimenti kao i algoritam rešavanja odgovarajućeg optimizacionog problema. Opisani su izvedeni eksperimenti nad skupom funkcionalno anotiranih proteina koji je preuzet iz javno dostupne baze podataka Swiss-Prot.

Glava 4 opisuje rezultate izvršenih eksperimenata. Trenirano je i testirano 20 klasifikacionih modela - za proteine 5 organizama i za 4 mere različitosti. Modeli sa najvećim skorom F mere primenjeni su na CAFA proteine i upoređeni sa drugim prediktorima i utvrđen je rang za svaki od testiranih skupova proteina. Zaključeno je da su rezultati predloženih prediktivnih modela uporedivi sa aktuelnim rezultatima prikazanim na poslednjem CAFA takmičenju, iako su trenirani na znatno manjim skupovima podataka znatno manjeg skupa organizama.

Glava 5 predstavlja drugu celinu ove disertacije i odnosi se na analizu neuređenosti proteina u odnosu na njihove funkcionalne kategorije kao i u odnosu na poziciju neuređenih regiona u proteinu. Analize su sprovedene na dva različita skupa proteina: na velikom skupu prokariotskih proteina gde je neuređenost određena predikcijom i na malom skupu proteina iz raznih organizama gde je neuređenost određena eksperimentalno. Dobijeni rezultati imaju veliki praktični značaj jer mogu doprineti unapređenju metoda za automatsko predviđanje funkcije proteina, analizi interakcija neuređenih i uređenih regiona proteina i mogućoj zameni neuređenih regiona malim molekulima leka.

Glavni naučni doprinosi u priloženom radu su:

- Defiisanje nove metode za automatsko predviđanje funkcije proteina na osnovu njegove primarne sekvence korišćenjem metoda strukturne klasifiacije, konkretno metode strukturalnih podržavajućih vektora (eng. Structured Support Vector Machines, SSVM)
- Prilagođavanje metode strukturalnih podržavajućih vektora za konkretan problem
    - definisanjem strukturalnog modela problema, tj. definisanjem funkcije koja predstavlja zajednički zapis ulaza i izlaza,
    - definisanjem mere sličnosti između dva grafa kao funkcije gubitka (tzv. loss funkcije) kao i
    - definisanjem, konstrukcijom i implementacijom algoritma za određivanje maksimuma funkcije cilja po svim mogućim izlazima, što s obzirom da je broj mogućih grafova ogroman predstavlja poseban izazov.
- Evaluacija rezultata koji se postižu primenom ove metode kao i njihovo poređenje sa aktuelnim metodama predviđanja funkcije proteina,
- Utvrđivanje uticaja porekla proteina nad kojima je treniran klasifiacioni model na njegove performanse.
-

- **Zaključak**

U rukopisu "Strukturna predikcija funkcije proteina i odnos funkcionalnih kategorija i neuređenosti" kandidat Jovana Kovačević je pokazala sistematično poznavanje oblasti istraživanja podataka i mašinskog učenja, prvenstveno klasifikacije podataka uopšte i posebno metoda strukturne klasifikacije. Takođe, ovladala je značajnim bioinformatičkim domenom primene ovih metoda – domenom predikcije funkcije proteina. Izvršila je prilagođavanje metode podržavajućih vektora sa strukturnim izlazom rešavanju postavljenog problema. Implementiran je programski sistem kojim se realizuje konstruisani algoritam optimizacije i realuzuju rešenja uoprediva sa rezultatima prediktora najuticajnijeg CAFA takmičenja u ovoj oblasti. Razvijena metoda ima široke mogućnosti primene i u drugim

oblastima, na primer, klasifikaciji tekstova na prirodnom jeziku. Gde kandidat takođe već ima značajne rezultate.

Kandidat je kroz ovaj rad dao metodološki i praktični doprinos rešavanju problema predikcije funkcije proteina, kao i odnosa funkcije proteina i aspekata njegove strukture što otvara put primenama u oblasti farmakologije i medicinske prakse. Predlažemo stoga Nastavno-naučnom veću Matematičkog fakulteta da rukopis "Strukturna predikcija funkcije proteina i odnos funkcionalnih kategorija i neuređenosti" kandidata Jovane Kovačević, prihvati kao doktorsku disertaciju i odredi komisiju za javnu odbranu.

U Beogradu, 6. novembra 2015.

dr Gordana Pavlović-Lažetić
redovni profesor Matematičkog fakulteta

dr Miloš Beljanski, naučni savetnik
Institut za opštu i fizičku hemiju, Beograd

dr Nenad Mitić, vanredni profesor
Univerzitet u Beogradu, Matematički fakultet

dr Predrag Radivojac, redovni profesor
Fakultet za informatiku i računarstvo, Univerzitet Indijana,
Blumington, Indijana, SAD

dr Mladen Nikolić, docent
Univerzitet u Beogradu, Matematički fakultet