

НАСТАВНО-НАУЧНОМ ВЕЋУ УНИВЕРЗИТЕТА У БЕОГРАДУ - БИОЛОШКОГ ФАКУЛТЕТА

На X редовној седници Наставно-научног већа Универзитета у Београду Биолошког факултета, одржаној 06.09.2019. године, на основу молбе ментора, др Владимира Перовића и др Душанке Савић-Павићевић, одређена је Комисија за преглед и оцену докторске дисертације Невена С. Шумоње, истраживача сарадника Института за нуклеарне науке ВИНЧА, под насловом: „Биоинформатички модели за аутоматско мапирање интеракција између протеина код човека” у саставу: др Владимир Перовић, научни сарадник, Институт за нуклеарне науке ВИНЧА, Универзитет у Београду, др Душанка Савић-Павићевић, редовни професор, Биолошки факултет, Универзитет у Београду и др Невена Вељковић, научни саветник, Институт за нуклеарне науке ВИНЧА, Универзитет у Београду.

Комисија је прегледала урађену докторску дисертацију кандидата и Наставно-научном већу Универзитета у Београду - Биолошког факултета подноси следећи

ИЗВЕШТАЈ

Општи подаци о докторској дисертацији

Докторска дисертација Невена С. Шумоње под насловом: „Биоинформатички модели за аутоматско мапирање интеракција између протеина код човека“ написана је на 194 стране и садржи 27 табела, 23 слике и 449 библиографске јединице. Докторску дисертацију чине: Насловна страна на српском и енглеском језику, Подаци о менторима и члановима комисије, Захвалница, Резиме на српском и енглеском језику са кључним речима, Списак скраћеница, Садржај, Текст по поглављима и Прилози (Биографија аутора, Списак слика и табела, Изјава о ауторству, Изјава о истоветности штампане и електронске верзије докторске дисертације и Изјава о коришћењу). Дисертација је подељена на 9 поглавља: Увод (1-49. стране), Циљеви истраживања (50. страна), Материјали и методе (51-77. стране), Резултати (78-123. стране), Дискусија (124-139. стране), Закључак (140-141. страна), Литература (142-193. стране) и Објављени радови (194. страна).

Анализа докторске дисертације

Кандидат је у поглављу **УВОД** кроз четири одељка приказао систематично податке из литературе који представљају и описују досадашња сазнања релевантна за предмет истраживања своје докторске дисертације. У одељку **Интеракције протеин-протеин** описане су структуре протеина и протеом човека, затим транскрипциони фактори, као кључни протеини транскрипционе регулације и експресије гена, хумани протеини неуређене терцијарне структуре (ПНТС), који имају вишеструке функције, вишеструке партнере у интеракцији и који су укључени у кључне биолошке процесе и повезани са непреносивим болестима. Дат је приказ интеракција протеин-протеин (ИПП) код човека и истакнут је значај мапирања мреже ИПП. У наредном одељку, **Експерименталне методе за детекцију ИПП**, описане су експерименталне методе мале и велике скале за детекцију, методе за интрацелуларну локализацију и валидацију ИПП. На крају одељка објашњена су ограничења експерименталних метода ИПП која ће у наредном вишегодишњем периоду условљавати потребу за развојем рачунарских метода за предикцију ИПП. У трећем одељку **Базе података ИПП** су наведени извори информација и базе података о експериментално потврђеним протеинским интеракцијама, док су у последњем одељку под насловом **Рачунарске методе за предвиђање ИПП** описане постојеће рачунарске методе за предвиђање ИПП засноване на различитим врстама информација, генском, биолошком и структурном контексту, и посебно секвенци протеина која је истакнута као високо квалитетна, универзална, доступна, и експериментално потврђена информација, као поуздани извор биолошке информације за развој метода за анализу и предвиђање ИПП. Затим су презентоване методе засноване на машинском учењу за предикцију ИПП, поред којих су објашњене основе

машинског учења, приступ процени ефикасности модела, креирање атрибута и специфичне карактеристике приступа машинског учења на проблему предвиђања ИПП. На крају овог поглавља представљене су и методе засноване на топологији биолошких мрежа.

У поглављу **ЦИЉЕВИ ИСТРАЖИВАЊА** концизно и јасно су представљени циљеви докторске дисертације. За општи циљ ове тезе је постављен развој биоинформатичког модела за аутоматско предвиђање ИПП код човека, на глобалном нивоу протеома као и за специфичне случајеве. У сврху реализације општег циља задата су посебно три циља: 1) развој метода за аутоматско предвиђање ИПП који учествују у процесу транскрипционе регулације, уз адаптирање новог приступа у широко доступан веб алат; 2) развој модела за предикцију ИПП код човека који обухватају протеине са неуређеном терцијарном структуром, уз имплементацију модела у форми веб алата и 3) развој модела за предвиђање ИПП на нивоу протеома човека, заједно са имплементацијом развијеног модела у форми самосталне апликације.

У поглављу **МАТЕРИЈАЛИ И МЕТОДЕ** детаљно су описани подаци, алати, методе и алгоритми који су коришћени у раду. У оквиру ове тезе коришћени су подаци из јавно доступних база података протеина, протеинских секвенци, протеинских интеракција, физичко-хемијских карактеристика аминокиселина и анотацијских ресурса онтологија гена, чији су извори дати и описане њихове верзије, коришћене током рада на овој дисертацији, у одељку **Подаци**. Софтверски алати и математичке методе за представљање протеина и интеракција на основу секвенце протеина, физичко-хемијских карактеристика аминокиселина, филогенетских профила и тополошких особина графа детаљно су представљени у одељку **Моделирање протеина и ИПП**. Формуле и описи статистичких мера за евалуацију предикционих перформанси модела, односно бинарних класификатора, заснованих на машинском учењу дати су у одељку **Машинско учење**, док су сви класификациони алгоритми машинског учења у процесу генерисања модела за предикцију интеракција коришћени у овом раду, а засновани на стаблима одлучивањима, пробабилистичком приступу, неуронским мрежама, линеарним моделима, потпорним векторима и градијентном појачавању, наведени и детаљно описани у одељку **Алгоритми машинског учења**. На крају поглавља, у одељку **Генетски алгоритми**, описани су генетски алгоритми за оптимизацију који су примењени у овој тези за развој метода за аутоматско креирање атрибута и ансамбла модела у финалном моделу за предвиђање ИПП на нивоу протеома човека.

У поглављу **РЕЗУЛТАТИ** детаљно, прецизно, јасно и прегледно су наведени добијени резултати. Резултати ове тезе, у складу са постављеним циљевима, приказани у оквиру три одељка. У првом одељку **Предвиђање ИПП транскрипционих регулатора** приказани су новоразвијени РААС4 атрибути за представљање транскрипционих регулатора и нови модел РААС4_RF за предвиђање ИПП на специјалном случају транскрипционих регулатора. Приказано је и упоређивање модела РААС4_RF који је надмашио по ефикасности предвиђања стандардне методе за предикцију ИПП засноване на секвенци. Поред представљене имплементације РААС4_RF у форми веб алата, приказана је и WT1 студија случаја где је предвиђена интеракција WT1-CDK9 алатом РААС4_RF и експериментално потврђена. У другом одељку под насловом **Предвиђање ИПП ПНТС** презентовани су нови DP_PAAC5 атрибути за представљање протеина неуређене терцијарне структуре, као и новогенерисани DP_PAAC5_RF модел за предвиђање интеракција у којима учествују протеини неуређене терцијарне структуре. Затим је приказана евалуација DP_PAAC5_RF модела у студијама поређења са стандардним методама за предвиђање ИПП на скуповима балансираних и дисбалансираних негативних интеракција, где је DP_PAAC5_RF модел показао бољу ефикасност. Поређење DP_PAAC5_RF са моделом формираним на основу општих људских ИПП потврдио је неопходност креирања посебног модела за специјални случај. На крају одељка је представљена имплементација у форми веб алата, као и студија случаја неуређеног протеина BASP1 транскрипционог корегулатора, у којој је интеракција са прогестеронским рецептором BASP1-PRGR предвиђена DP_PAAC5_RF алатом и експериментално потврђена. У последњем одељку **Предвиђање ИПП на нивоу протеома код човека** приказане су три нове врсте атрибута за представљање протеина и интеракција, засноване на: 1) примарној структури протеина, у оквиру којих су развијене две нове синтетичке карактеристике аминокиселина, 2) еволутивним профилима и 3) мрежама; као

и нови GAFT алгоритам за аутоматско креирање и селекцију атрибута. Затим, приказан је GA-STACK алгоритам за аутоматско формирање и оптимизацију ансамбла модела машинског учења, на основу кога, заједно са три новокреиране врсте атрибута, је представљен HP-GAS метод за предикцију ИПП на нивоу целог протеома човека. Поред тога, приказана је евалуација HP-GAS метода на 1) случају анализе ефикасности стратегија у формирању HP-GAS, где је инжењеринг атрибута показао највећи допринос, 2) раздвајања атрибута према три групе атрибута које је побољшало финални модел и 3) поређење са стандардним методама за предикцију ИПП, где је HP-GAS надмашио остале и приказао високу ефикасност предвиђања $AUC=0,93$ и тачност $0,85$. На крају овог одељка је, поред имплементација HP-GAS методе у форми широко доступног и једноставног за коришћење самосталног алата, приказана и студија случаја EGFR протеина, у којој су интеракције, које нису биле у скупу за тренирање, EGFR са SUMO 1, 2, 3 и 4, регулаторним факторима, предвиђене са високом вероватноћом, које су претрагом литературе и експериментално потврђене.

У поглављу **ДИСКУСИЈА** поступно су тумачени резултати ове докторске дисертације у контексту постојећих литературних података. У оквиру овог поглавља кандидат детаљно, аналитички и критички дискутује и тумачи добијене резултате и доступне литературне податке у светлу релевантних истраживања о методама за предвиђање интеракција протеин-протеин. Детаљно су протумачени, продискутовани и упоређени резултати посебно за сваки од три развијена метода, за специјалне случајеве транскрипционих регулатора и протеина неуређене терцијарне структуре, као и на општем нивоу протеома код човека. Истакнута је значајност развоја специјалних модела поред општег. Поступно су анализирани и продискутовани изазови у сваком од корака у процесу развоју метода за предвиђање ИПП: прикупљање поузданих података, креирање квалитетних скупова за тренирање, формирање атрибута за представљање протеина на основу секвенце као поуздане информације, аутоматска селекција и комбиновање атрибута, проблем величина скупова и комплексности алгоритама, избор алгоритама машинског учења њихова оптимизација и генерализација као и аутоматска оптимизована комбинација у ансамбл, поуздана евалуација модела, поређење са state-of-the-art методама, и имплементација метода у ефикасне, брзе, једноставне за коришћење, документоване и широко доступне софтверске алате. Дискутована је детаљно и примена развијених метода на примерима Вилмсовог тумор протеина WT1, транскрипционог корегулатора BASP1 и EGFR трансмембранског рецептора тирозинске киназе, повезаног са низом људских обољења, где су предвиђене WT1-TFIIВ, WT1-CDK9, BASP1-PRGR, EGFR-SUMO1,2,3,4 интеракције, а затим и литературно или експериментално потврђене.

У поглављу **ЗАКЉУЧАК** јасно су наведени закључци изведени на основу добијених и дискутованих резултата а према претходно постављеним конкретним циљевима. Истакнута су три главна закључка: 1) развијен је биоинформатички модел за аутоматско предвиђање ИПП транскрипционих регулатора, базиран на секвенци протеина и машинском учењу, који се показао предиктивно и рачунарски ефикаснији у односу на стандардне моделе за предикцију ИПП, и адаптиран је у форми веб алата; 2) развијен метод за предвиђање интеракција протеина са неуређеном терцијарном структуром, заснован на секвенци протеина, физичко-хемијским карактеристикама аминокиселина и машинском учењу, показао је већу тачност у односу на стандардне методе, и имплементиран је у форми ефикасног и широко доступног веб алата; 3) развијен је метод за предвиђање ИПП на нивоу протеома човека, уз три нове врсте атрибута и алгоритме за аутоматско креирање атрибута и надгледано спајање модела машинског учења, који је приказао већи степен тачности предвиђања ИПП у односу на стандардне методе, и имплементиран је у форми самосталног софтверског алата. Поред ова три, наведен је и општи закључак да биоинформатичке методе развијене у овој тези имају велики значај и самостално и као суплементи експерименталним методама за анализу ИПП, јер омогућавају боље разумевање комплексних молекуларних механизма биолошких процеса и функција. Поред тога, развијени модели омогућавају расветљавање патолошких стања организма, лакше откривање мета за циљану терапију и доносе нова сазнања о сигналним путевима, терапеутским метама, транскрипционим регулаторима уз смањење времена и трошкова развоја нових терапеутика.

На крају ове докторске дисертације налазе се поглавља **ЛИТЕРАТУРА** у коме је дата опсежна листа библиографских јединица коришћених у овој дисертацији које су адекватно и на одговарајућим местима цитирани и поглавље **ОБЈАВЉЕНИ РАДОВИ** где су наведени радови који су објављени у врхунским и истакнутим међународним часописима на основу резултата из ове дисертације.

Радови и конгресна саопштења из докторске дисертације

Б1. Радови у часописима међународног значаја

1. **Neven Sumonja**, Branislava Gemovic, Nevena Veljkovic, Vladimir Perovic, „Automated feature engineering improves prediction of protein–protein interactions“, *Amino acids*, 2019;1-4.
<https://doi.org/10.1007/s00726-019-02756-9> (M22)
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;8(1):10563.
<https://doi.org/10.1038/s41598-018-28815-x> (M21)
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> (M21)
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*, 2017; 33 (2):289–291
<https://doi.org/10.1093/bioinformatics/btw590> (M21a)

Б2. Конгресна саопштења на скуповима међународног значаја

1. **Neven Sumonja**, Vladimir Perovic, Nevena Veljkovic. Uncovering human transcriptional regulation through the sequence information. Biocuration 2016, The Ninth International Biocuration Conference, Geneva, Switzerland, April 10-14, 2016. (M34)
2. **Neven Sumonja**, Nevena Veljkovic, Sanja Glisic, Vladimir Perovic. Protein-protein interaction prediction method based on principle component analysis of amino acid physicochemical properties. Belgrade Bioinformatics Conference, BELBI 2016, Belgrade, Serbia, June 20th-24th. (M34)
3. Gemović B, Davidović R, **Šumonja N**, Veljković N, Perović V. 2017. Contribution of features based on sequence, predicted PPIs and GO similarities to the prediction of gene-HPO associations. ISMB/ECCB 2017, Prague, Czech Republic, July 21-25, 2017. (M34)
4. Vladimir Perovic, **Neven Sumonja**, Sandro Radovanovic, Branislava Gemovic, Milan Vukicevic, Nevena Veljkovic. Intrinsically disordered proteins protein-protein interaction prediction method based on various sequence representations. 3rd NGP-Net Symposium on Non-Globular Proteins, Kosice, Slovakia. August 28 - September 1, 2017. (M34)
5. Gemović B, Davidović R, Stanković K, **Šumonja N**, Perović V, Veljković N. 2018. Function Annotation Algorithm Based on Sequence Spectral Features: Application on Human Transcription Factors. The 11th International Biocuration Conference (IBC), Shanghai, China, April 8-11, 2018. (M34)
6. **Šumonja N**, Veljković N., Perović V. 2018. Exploring the usefulness of graph properties in protein protein interaction predictions. Belgrade Bioinformatics Conference 2018 (BelBi 2018), Belgrade, Serbia, June 18-22 2018. In: *Biologia Serbica*, 2018, 40(1): 129. (M34)
7. Perovic V, Davidovic R, Gemovic B, **Sumonja N**, Veljkovic N. 2018. Prediction of Human Phenotype Ontology Terms for Intrinsically Disordered Proteins. Belgrade Bioinformatics Conference 2018 (BelBi 2018), Belgrade, Serbia, June 18-22 2018. In: *Biologia Serbica*, 2018, 40(1): 90. (M34)
8. Gemović B, Davidović R, **Šumonja N**, Perović V, Veljković N. 2018. Function Annotation Algorithm Based on Sequence Spectral Features: Evaluation on Human Transcription Factor.

Провера оригиналности докторске дисертације

На основу Правилника о поступку провере оригиналности докторских дисертација које се бране на Универзитету у Београду и налаза у извештају из програма iThenticate којим је извршена провера оригиналности докторске дисертације „Биоинформатички модели за аутоматско мапирање интеракција између протеина код човека“, аутора Невена С. Шумоње, утврђено је да дисертација садржи 55184 речи од којих је 105 подударана на основу 72 извора и да индекс подударана текста износи 3%. Овај степен подударности последица је општих места и појмова, претходно публикованих резултата истраживања докторанда, који су проистекли из његове дисертације, личних имена и назива институција. Све поменуто је у складу са чланом 9. Правилника и сматра се да докторска дисертација кандидата Невена С. Шумоње представља оригинално научно дело.

Мишљење и предлог Комисије

Докторска дисертација под насловом „Биоинформатички модели за аутоматско мапирање интеракција између протеина код човека“ представља оригиналан научни рад кандидата Невена С. Шумоње и написана је систематично и јасно, са добро дефинисаним циљевима. Резултати приказани у оквиру ове дисертације представљају значајан допринос у циљу разумевања и расветљавања интерактома код човека, развојем ефикасних аутоматских метода за предвиђања нових интеракција и интерактора. Мапирање интерактома је од суштинске важности за разумевање комплексних молекуларних односа унутар живих система и расветљавање разних патолошких стања људског организма, а рачунарски методи за предвиђање интеракција имају примену и у биомедицини. Кандидат је показао добро познавање литературе и проблематике релевантне научне области као и висок степен истраживачке зрелости. Остварени резултати објављени су у оквиру четири оригинална научна рада, што потврђује актуелност и значајност добијених резултата.

Имајући у виду квалитет докторске дисертације Невена С. Шумоње, под насловом „Биоинформатички модели за аутоматско мапирање интеракција између протеина код човека“, Комисија, са задовољством, предлаже Наставно-научном већу Биолошког факултета Универзитета у Београду да прихвати позитивну оцену Комисије и одобри јавну усмену одбрану докторске дисертације кандидата Невена С. Шумоње.

КОМИСИЈА:

У Београду, 10.09.2019. године

др Владимир Перовић, научни сарадник
Институт за нуклеарне науке ВИНЧА, Универзитет у Београду

др Душанка Савић-Павићевић, редовни професор
Биолошки факултет, Универзитет у Београду

др Невена Вељковић, научни саветник
Институт за нуклеарне науке ВИНЧА, Универзитет у Београду