

Наставно-научном већу
Математичког факултета
Универзитета у Београду

Одлуком Наставно-научног већа Математичког факултета Универзитета у Београду донетом на 368. седници одржаној 24.01.2020. године именовани смо за чланове комисије за преглед и оцену докторске дисертације „Рачунарске методе партиционисања и груписања у биолошким мрежама“ кандидата Милане Грбић, мастер математичара. После прегледања поднетог рукописа подносимо следећи

ИЗВЕШТАЈ

БИОГРАФИЈА КАНДИДАТА

Милана (Милан) Грбић је рођена у Бањој Луци 15. августа 1989. године, где је завршила основну школу и Гимназију. Основне студије је завршила на Природно-математичком факултету Универзитета у Бањој Луци 2012. године са просечном оценом 9.24, стекавши звање Дипломирани математичар и информатичар - 240 ЕСПБ. Мастер студије студијског програма Рачунарство и информатика Математичког факултета Универзитета у Београду завршила је 2016. године са просечном оценом 10.00, одбравивши мастер рад под називом „Груписање организама помоћу различитих метода класификације у зависности од генотипских и фенотипских карактеристика“ под руководством проф. др Ненада Митића. Докторске студије студијског програма Информатика на Математичком факултету је уписала 2016. године. Положила је све испите предвиђене планом и програмом докторских студија са просечном оценом 10.00.

Милана Грбић је од 2013. године запослена на Природно-математичком факултету Универзитета у Бањој Луци, прво на позицији асистента, а од 2017. године на позицији вишег асистента. До сада је изводила вежбе из више информатичких предмета: Увод у програмирање, Основе рачунарских система 1, Основе рачунарских система 2, Основе рачунарских система 3, Основе рачунарских система 4, Оперативни системи, Биоинформатика и др.

Учесник је више домаћих истраживачких пројеката и једног међународног истраживачког пројеката из области биоинформатике. У току студирања на докторским студијама боравила је на више летњих и зимских школа, као што су: летња школа ICGEB Course "Bioinformatics: Computer Methods in Molecular and Systems Biology", Трст, Италија, студијски боравак у Лабараторији за биоинформатику и рачунарску хемију, Институт за нуклеарне науке Винча, зимска школа „2nd NGP-net Winter School on computational methods for non-globular proteins“, Никозија, Кипар и зимска школа „NGP-net Winter School on experimental methods for protein disorder and aggregation“, Марсеј, Француска. Учествовала је на више међународних конференција из области биоинформатике.

ОБЈАВЉЕНИ НАУЧНИ РАДОВИ И САОПШТЕЊА СА НАУЧНИХ СКУПОВА

- [1] Grbić, M., Kartelj, A., Janković, S., Matić, D., Filipović, V. **Variable neighborhood search for partitioning sparse biological networks into the maximum edge-weighted k -plexes**. IEEE/ACM Transactions on Computational Biology and Bioinformatics, DOI 10.1109/TCBB.2019.2898189, 2019. (M21, IF2018 = 2.896)
- [2] Grbić, M. **Conditional Random Fields-based Approach to Classification: Application to Life Sciences**, IPSI BgD Transactions on Advanced Research (TAR), Special issue - "ICT Research at the University of Belgrade and at its Foreign Guests", Vol. 15, No. 1, pp. 1-9, 2019, ISSN 1820 - 4511. (M53)
- [3] Grbić, M. **A three-phase mapreduce-based algorithm for searching biomedical document databases**, "International Journal of Electrical Engineering and Computing", Vol. 3, No. 1, pp. 1-8, 2019, e-ISSN: 2566-3682. (M53)
- [4] Grbić, M., Kartelj, A., Matić, D., Filipović, V. **Improving 1NN strategy for classification of some prokaryotic organisms**, Belgrade Bioinformatics Conference (BelBI) 2016, Proceedings Belgrade Bioinformatics Conference 2016, ISBN: 978-86-7589-124-6, pp. 43-54. (M33)
- [5] Grbić, M., Kartelj, A., Matić, D., Janković, S., Filipović, V. **A heuristic approach for clustering metabolic networks into highly connected components**, 4th NGP-NET symposium, Druskininkai, Litvanija, septembar 2018. (M34)
- [6] Grbić, M., Janković, S., Matić, D., Pavlović-Lažetić, G. **Conditional Random Fields based approach for classification of the reactants in some metabolic reactions**, Belgrade Bioinformatics Conference 2018, Beograd, Srbija, jun 2018. (M34)
- [7] Grbić, M., Kartelj, A., Matić, D., Filipović, V. **A local search based heuristic for clustering large biological networks into highly connected components**, First Congress of Molecular Biologists of Serbia, Beograd, Srbija, septembar 2017. (M34)
- [8] Grbić, M., Kartelj, A., Matić, D., Filipović, V. **Partitioning biological networks in k -plex subnetworks with maximum edge weights**, 3rd NGP-NET symposium, Košice, Slovačka, septembar 2017. (M34)

ПРЕДМЕТ ДИСЕРТАЦИЈЕ

У рукопису „Рачунарске методе партиционисања и груписања у биолошким мрежама“ разматрано је неколико проблема дефинисаних над биолошким мрежама, који су решавани оптимизационим методама и методама машинског учења. У рукопису су решавани следећи проблеми: партиционисање ретких биолошких мрежа на k -plex подмреже (енгл. Maximum Edge-weight k -plex Partitioning Problem - Max-EkPP), предвиђање улоге метаболита у метаболичким реакцијама, партиционисање биолошких мрежа на високо повезане компоненте (енгл. Highly Connected Deletion problem – HCD) и идентификација значајних група протеина додавањем нових грана у

тежинску мрежу протеинских интеракција (енгл. protein-protein interaction – PPI). Рачунарски најзахтевнија фаза последњег разматраног проблема укључује проблем одређивања минималног броја додатних PPI, које треба укључити у PPI мрежу, да би познати протеински комплекси, посматрани као подграфови у тој мрежи, били повезани.

Познато је да кохезивни биолошки елементи (као што су гени, протеини, метаболити и сл.) у мрежама формирају густе подструктуре (подмреже). Због тога, партиционисање (кластеровање) биолошке мреже у циљу препознавања важних, међусобно повезаних структура може бити користан приступ за добијање нових информација о биолошким елементима и њиховим међусобним односима. С друге стране, бројне класе биолошких мрежа садрже ретке мреже, чија подела на густе подмреже (на пример клике) може бити врло рестриктивна, а многе потенцијално важне информације о интерференцији биолошких објеката се могу изгубити. Дакле, приступ код кога је услов да се мреже партиционису на клике релаксиран, би могао водити ка биолошки смисленијим партицијама. Релаксирањем клика до подграфова који нису потпуно повезани, биолошки објекти се на бољи и биолошки оправданији начин повезују у семантичке или функционалне логичке групе, које могу бити од користи за даља истраживања.

Први проучавани проблем, Max-EkPP, има значајну биолошку интерпретацију. Подскуп од n чворова мреже се назива k -plex ако је степен сваког чвора у подмрежи индукованој тим подскупом најмање $n-k$. Max-EkPP подразумева партиционисање улазне мреже у k -plex подмреже тако да је сума тежина свих грана у добијеним подмрежама максимална. Проблем партиционисања ретких биолошких мрежа на k -plex подмреже је NP тежак проблем и у циљу решавања овог проблема у рукопису је развијена метода променљивих околина. Метода је даље тестирана на реалним биолошким и синтетичким инстанцама из литературе. Анализа перформанси методе и поређење са другим методама из литературе указују на висок квалитет добијених резултата. k -plex подмреже, које су добијене партиционисањем метаболичких мрежа, су даље анализиране са биолошког аспекта, чиме је додатно демонстрирана употребна вредност развијене методе и добијених резултата.

Други проблем који се истражује у овој дисертацији је проблем предвиђања улоге метаболита у метаболичким реакцијама. Идентификацијом улоге метаболита у реакцији може се доћи до нових сазнања о вези између метаболита, на основу њиховог учешћа у заједничкој реакцији. У рукопису је за решавање овог проблема примењена предиктивна метода машинског учења заснована на методи статистичког моделирања условних случајних поља. За примену ове методе формирана су три различита скупа карактеристичних функција које укључују информације о елементима у најближем окружењу и информације о њиховим ознакама. Добијени резултати се могу користити за детаљнију анализу метаболизма појединачних организама.

У рукопису је проучаван и проблем брисања грана ради обезбеђивања високе повезаности. Граф са n чворова се сматра високо повезаним (енгл. highly connected) ако је степен сваког чвора већи од $n/2$. Решавање HCD проблема се своди на брисање што је могуће мање грана из графа, тако да добијене компоненте буду високо повезане. Наведени проблем има неколико примена у рачунарској биологији, нпр. за проналажење група протеина са сличном GO анотацијом или проналажење групе гена са сличним профилем експресије. У рукопису је за решавање овог NP тешког проблема развијена метода променљивих околина. Предложена метода је примењена на PPI мреже и метаболичке мреже, а за неке од добијених високо повезаних компоненти

(подграфова) дата је одговарајућа биолошка интерпретација. Поред тога, приказана је и метода предвиђања GO термина који се додељују протеинима, а која се заснива на високо повезаним компонентама добијеним партиционисањем PPI мреже. Високо повезане компоненте представљају полазне групе протеина, које се даље проширују на основу суседства (интеракција) у полазној PPI мрежи. Након тога, протеинима који се налазе у истој групи се додељују GO термини које имају остали протеини из те групе.

Последњи проблем који је анализиран и решаван у рукопису је проблем идентификације значајних група протеина додавањем нових грана у тежинску PPI мрежу. Предложена метода се састоји од три фазе. У првој фази се решава NP тежак проблем додавања минималног броја интеракција у PPI мрежу, тако да протеински комплекси постану повезани. Протеински комплекс чини неколико протеина (бар два), који су физички међусобно повезани. Предложено решење се заснива на методи променљивих околина и тестирано је над PPI мрежама. Друга фаза методе идентификације значајних група протеина подразумева издвајање подмрежа које су индуковане новододатим гранама. Протеини који су крајеви новододатих грана чине базне групе протеина, које се, у трећој фази предложене методе, проширују протеинима који су индиректни суседи протеина из базне групе и имају просечну вредност генске ко-експресије изнад унапред задатог прага. У рукопису се посматрани проблем разматра и над тежинским и нетежинским мрежама. Тако идентификоване групе протеина су даље предмет анализе обогаћивања информацијама (енгл. enrichment analysis), којом је потврђена семантичка значајност добијених протеинских група.

ПРИКАЗ ДИСЕРТАЦИЈЕ

Рукопис има 137 страна и обухвата 6 поглавља основног текста, списак коришћене литературе од 169 библиографских јединица, 30 табела и 63 слике. Структура рукописа је следећа.

У уводном поглављу дат је преглед биолошких мрежа и проблема дефинисаних над биолошким мрежама. Поред тога, дат је опис коришћених метода машинског учења и оптимизационих метода. У последњој секцији уводног поглавља описани су биоинформатички алати и ресурси, који су коришћени у изради ове дисертације.

Друго поглавље је посвећено анализи и решавању Max-EkP проблема. Дат је детаљан опис проблема, преглед ранијих резултата и сличних проблема. Затим је детаљно описана развијена метода променљивих околина. У наставку је дат приказ резултата над метаболичким мрежама и синтетичким мрежама из литературе. У посебном делу овог поглавља дата је биолошка интерпретација и визуелни приказ добијених *k-plex*-а. На крају су дата завршна разматрања у решавању овог проблема.

У трећем поглављу се анализира и решава проблем предвиђања улоге метаболита у метаболичким реакцијама. Дат је преглед резултата над сродним проблемима, детаљно је описана метода условних случајних поља и њена примена на овај проблем. На крају овог поглавља приказани су резултати тестирања и дата су завршна разматрања овог проблема.

Четврто поглавље садржи приказ проблема партиционисања биолошких мрежа на високо повезане компоненте, односно детаљан опис проблема и приказ методе променљивих околина за његово решавање. У посебном делу овог поглавља приказани су резултати тестирања. Део овог поглавља је посвећен биолошкој евалуацији високо повезаних компоненти, са нагласком на примену предложене методе за предвиђање GO термина за протеине са неуређеном структуром.

У петом поглављу је приказана метода идентификације значајних група протеина додавањем нових грана у тежинску PPI мрежу. Дат је приказ ранијих резултата, а затим и описана метода у три фазе. Прва фаза подразумева подржавање комплекса додавањем PPI у постојећу тежинску PPI мрежу метахеуристичким методом променљивих околина. У другој фази се врши спајање одређених делова различитих комплекса комбинацијом постојећих PPI и нових PPI додатих у првој фази. На крају се у трећој фази, на основу индиректних интеракција, додају нови протеини у протеинске групе које су формиране у другој фази. Приказани су резултати тестирања и дат је приказ случаја употребе дате методе на конкретној мрежи протеинских интеракција.

У последњем, шестом поглављу, дат је закључак, наведени су научни доприноси дисертације и правци даљих истраживања.

НАУЧНИ ДОПРИНОС ДИСЕРТАЦИЈЕ

- Решавање NP тешког проблема партиционисања графа у k -plex-е структуре, који је примењен на партиционисање биолошких мрежа. Развијени алгоритам је заснован на методи променљивих околина. Предложена метода је примењена на метаболичке биолошке мреже и представља нови приступ у откривању нових информација у биолошким структурама.
- Решавање NP тешког проблема брисања грана уз очување високе повезаности. Предложена метода је примењена на протеинске и метаболичке мреже и представља нови приступ у анализи односа у биолошким структурама.
- Развој методе за предвиђање улоге метаболита у метаболичким реакцијама. Метода је заснована на условним случајним пољима и представља нови приступ за решавање проблема предвиђања у области биолошких наука.
- Развој новог приступа за идентификовање значајних група протеина у PPI мрежи. Проучавани проблем је еквивалентан математичком NP тешком проблему реконструкције мреже. Приступ укључује додавање нових протеин-протеин интеракција у постојећу мрежу и груписање протеина комбиновањем биолошких информација из различитих извора (PPI мреже, генска ко-експресија и алати за обогаћивање информацијама). Идентификовање група протеина на описани начин представља нови приступ у разумевању унутрашњих структура и функција биолошких података.

ЗАКЉУЧАК

У рукопису „Рачунарске методе партиционисања и груписања у биолошким мрежама“ кандидат Милана Грбић је показала широко и систематично знање из области рачунарске биологије, биоинформатике и рачунарства. Посматрани проблеми који су предмет дисертације су веома важни за изучавање биолошких структура са теоријског и практичног аспекта. За разматране проблеме су анализирани и развијани одговарајући математички модели, који су даље решавани различитим рачунарским методама. Предложене методе партиционисања и груписања у биолошким мрежама представљају нови приступ у решавању разматраних проблема, којима се на основу полазних података идентификују значајне биолошке структуре, корисне за даља биолошка истраживања. Кроз темељно тестирање, које је вршено на реалним биолошким и синтетичким тест примерима, показано је да предложене методе машинског учења и оптимизације успешно решавају анализирани проблеми.

Кандидат Милана Грбић је кроз истраживачки рад приказан у овој дисертацији дала значајан допринос у решавању неколико рачунарски сложених проблема који имају велику примену у биоинформатици и рачунарској биологији. Стога предлажемо Наставно-научном већу Математичког факултета Универзитета у Београду, да рукопис „Рачунарске методе партиционисања и груписања у биолошким мрежама“, кандидата Милане Грбић, прихвати као докторску дисертацију и одреди комисију за јавну одбрану.

У Београду, 17.02.2020. године

Чланови комисије:

др Владимир Филиповић, редовни професор,
Математички факултет, Универзитет у Београду

др Александар Картељ, доцент,
Математички факултет, Универзитет у Београду

др Драган Матић, ванредни професор,
Природно-математички факултет, Универзитет у Бањој Луци

др Бранислава Гемовић, научни сарадник,
Институт за нуклеарне науке Винча, Универзитет у Београду