Наставно-научном већу
Математичког факултета
Универзитета у Београду

Одлуком Наставно-научног већа Математичког факултета у Београду донетом на 405. седници одржаној 23.6.2023. године именовани смо за чланове комисије за оцену докторске дисертације „**Семантичко обједињавање и претраживање биоинформатичких база података коришћењем метода истраживања података**" (енг. „***Semantic Unification And Searching Of Bioinformatics Databases Using Data Mining Methods***") кандидата **Александра Вељковића**, Мастер информатичара. После прегледа поднетог рукописа подносимо следећи

Извештај

Биографски подаци

Кандидат Александар Вељковић је рођен 23. семтембра 1992. године у Пожаревцу. Основну школу је завршио у Голупцу, као носилац Вукове дипломе, а гимназију у Великом Градишту, као ученик генерације и носилац Вукове дипломе. Школске 2011/2012. године уписао је основне академске студије на Математичком факултету у Београду (ст. програм Информатика). Дипломирао је у року, школске 2013/2014. године са просеком 9,77.

Школске 2014/2015. године уписао је мастер академске студије на Математичком факултету, ст. програм Информатика. Мастер академске студије завршио је 2016. године одбраном мастер рада под насловом „Нова метода за асемблирање генома на основу PFG електрофорезе" под руководством проф. др Ненада Митића. Просечна оцена на мастер студијама је била 9,69.

Школске 2016/2017. уписао је докторске академске студије на Математичком факултету, ст. програм Информатика. Положио је све испите предвиђене планом и програмом докторских студија са просечном оценом 10,00.

Од 2015. године је био запослен на Математичком факултету Универзитета у Београду као сарадник у настави, а од 2017. године до октобра 2023. године као асистент у настави на Катедри за рачунарство и информатику. Учествовао је у

извођењу вежби из више предмета на основним и мастер студијама (Програмирање 1, Програмирање 2, Алгоритми и структуре података, Веб програмирање, Вештачка интелигенција, Истраживање података 2, Биоинформатика (Математички факултет, Машински факултет), Истраживање података у биоинформатици, Криптографија, Управљање пројектима у индустрији и науци и Основи програмирања за биохемичаре (Хемијски факултет)).

Од 2023. ради на позицији сениор истраживача у области криптографије у оквиру истраживачког одељења фирме *MVP Workshop*.

Учествовао је у развоју више техничких решења у области децентрализованих система, криптографије и управљања и контроле лета беспилотних летелица. Основне области интересовања су му криптографија, истраживање података и биоинформатика. Објављени научни радови и саопштења са научних скупова

Александар Вељковић има седам радова у часописима са SCI листе (M21, M22, M21a, M23, M23, M23, M23), један самосталан рад у домаћем часопису, десет излагања на међународним конференцијама (седам штампаних у изводу) и једно излагање на домаћој конференцији.

Радови у међународним часописима са SCI листе (пет):

1. Veljković, A. et al.: *BioGraph: Data Model for Linking and Querying Diverse Biological Metadata*, International Journal of Molecular Sciences, (2023) 24(8):6954. https://doi.org/10.3390/ijms24086954 (M21, IF=5.6)

2. Milutinović, B. et al.: *VLSI for SuperComputing: Creativity in R+D from applications and algorithms to masks and chips*, Advances in Computers, (2022). https://doi.org/10.1016/bs.adcom.2022.01.001 (M22, IF=3.067)

3. Čokić, V.P. et al.: *A comprehensive mutation study in wide deep-rooted R1b Serbian pedigree: mutation rates and male relative differentiation capacity of 36 Y-STR markers,* Forensic Science International: Genetics, Volume 41 (2019), 137-144 https://doi.org/10.1016/j.fsigen.2019.04.007 (M21a, IF=4.884)

4. Alempijević, T.M. et al.: *Change in the incidence and anatomic distribution of colorectal adenoma and cancer over a period of 20 years – A single center experience,* Vojnosanitetski pregled, Volume 75, Issue 3 (2018), 260-266 https://doi.org/10.2298/VSP160409207A (M23, IF=0.418)

5. Alempijevic T. et al.: *Doppler ultrasonography combined with transient elastography improves the non-invasive assessment of fibrosis in patients with*

chronic liver diseases, Medical Ultrasonography, Volume 19 (2017), 7-15 http://dx.doi.org/10.11152/mu-921 (M23, IF=1.651)

6. Arsenijević V. e al.: *Erythropoietin in the Evaluation of Treatment Outcomes in Patients with Polytrauma,* Acta clinica Croatica, Vol. 56. No. 4 (2017), 581-587 https://doi.org/10.20471/acc.2017.56.04.01 (M23, IF=0.452)

7. Alempijevic T. et al.: *Erythropoietin in predicting prognosis in patients with acute-on-chronic liver failure,* Journal of Gastrointestinal and Liver Diseases, Voluma 25, Number 4, (2016), 473-479 https://doi.org/10.15403/jgld.2014.1121.254.jev (M23, IF=4.884)

Радови у домаћим часописима:

8. Veljković A., *Algorithm for Document Authorship Identification and Plagiarism Evaluation Based on Generalized Suffix Tree*, Pregled NCD 37 (2020), 46–51 http://www.ncd.matf.bg.ac.rs/issues/37/5.pdf

Саопштења на међународним конференцијама штампана у целини или изводу:

9. Veljković A, Stojanović B, Malkov S, Beljanski M, Pavlović-Lažetić G, Mitić M. *Codon Usage-based SARS-CoV-2 Protein Classification.* Book of Abstracts Belgrade BioInformatics Conference 2021. Biologia serbica. 2021;43(1) ISSN: 2334-6590, 21-25. Jun, 2021, Belgrade. http://belbi.bg.ac.rs/wp-content/uploads/2021/06/Book_of_Abstracts_2021-1.pdf

10. Stojanović B, Veljković A, Malkov A, Beljanski M, Pavlović-Lažetić G, Mitić N. *Codon Usage Polymorphism in SARS-CoV-2 Protein Coding Sequences.* Book of Abstracts Belgrade BioInformatics Conference 2021. Biologia serbica. 2021;43(1) ISSN: 2334-6590, 21-25. Jun, 2021, Belgrade. http://belbi.bg.ac.rs/wp-content/uploads/2021/06/Book_of_Abstracts_2021-1.pdf

11. Ćirić N, Veljković A. *Identification of Differentially Expressed Genes in SARS-Cov-2 Infected Cells Using Bayesian Network Models.* Book of Abstracts Belgrade BioInformatics Conference 2021. Biologia serbica. 2021;43(1) ISSN: 2334-6590, 21-25. Jun, 2021, Belgrade. http://belbi.bg.ac.rs/wp-content/uploads/2021/06/Book_of_Abstracts_2021-1.pdf

12. Veljković A, et al. *Classification of Single Cell Types using Small Sets of Expressed Genes: Comparative Analysis of Supervised Machine Learning Methods.* 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), DOI: 10.1109/bibm52615.2021.9669844

13. Veljković A, Mitić N, *Semantic unification and search of bioinformatics databases.* Book of Abstracts Belgrade BioInformatics Conference 2023. (2023) ISBN: 978-86-82679-14-1, 19-23. Jun, 2023, Belgrade. https://belbi.bg.ac.rs/wp-content/uploads/2023/07/BelBi2023-Book-of-Abstracts.pdf

14. Malkov, S. et al. *Clustering and classification of SARS-CoV-2 isolates using RSCU.* Book of Abstracts Belgrade BioInformatics Conference 2023. (2023) ISBN: 978-86-82679-14-1, 19-23. Jun, 2023, Belgrade. https://belbi.bg.ac.rs/wp-content/uploads/2023/07/BelBi2023-Book-of-Abstracts.pdf

15. Veljković, A, et al. *Analysis of nucleotide sequence repeats in coronaviruses.* Book of Abstracts Belgrade BioInformatics Conference 2023. (2023) ISBN: 978-86-82679-14-1, 19-23. Jun, 2023, Belgrade. https://belbi.bg.ac.rs/wp-content/uploads/2023/07/BelBi2023-Book-of-Abstracts.pdf

Саопштења на међународним конференцијама:

16. Veljković A. *Zero Knowledge Machine Learning.* Data Science Conference DSC Europe 24, 2023.

17. Veljković A, et al. *BioGraph: Data Model for Linking and Querying Diverse Biological Metadata.* World congress "Systems theory, algebraic biology, artificial intelligence: mathematical foundations and applications", 2023.

18. Veljković A, et al. *BioGraph: Data Framework for Linking and Querying Diverse Biological Metadata*. VII Congress of Russian Biophysicists, 2023.

Саопштења на домаћим конференцијама штампана у изводу:

19. Veljković A. *Digitization of church records for creating knowledge database of personal historical records.* NCD XIX Conference 2022.

Предмет дисертације

Предмет дисертације је развој метода за препознавање и унификацију семантичких веза између метаподатака из различитих биолошких база података и примена метода истраживања података за омогућавање и остваривање претраге и анализе семантичких веза.

Биолошки подаци су представљени у великом броју отворених и затворених колекција података, које користе велики број различитих типова и формата записивања података. Сваки биолошки податак обухвата како саме податке тако и скуп метаподатака који описују својства конкретног биолошког појма или објекта. При томе различите базе података садрже различите подскупове метаподатака који се односе на исте биолошке појмове. Начин приступа овим подацима и њихово претраживање специфични су за сваку од база података у којима се налазе, што додатно отежава њихово повезивање и анализу.

Идентификатори истих појмова се често разликују између база података те је семантичка претрага у неким случајевима не само одговарајуће него и једино решење за претраживање појмова. Проналажењем и унификацијом семантичких веза између (мета)података који долазе из различитих извора са потенцијално различитом структуром и применом метода истраживања података на тако установљене везе, би могла да се обезбеди њихова агрегација у унифковану структуру метаподатака која подржава семантичку претрагу.

Кандидат Александар Вељковић у свом раду предлаже нов модел повезивања биоинформатичких база података, који обухвата и предложен аутоматизовани протокол за откривање нових семантичких односа између објеката. Предлаже се и нов аутоматизовани протокол за откривање нових односа семантичке сличности између објеката различитих база података, који је заснован на методама истраживања података који су добијени из дефинисаних модела. Као потврду исправности предложених модела и протокола, кандидат је на основу тог модела имплементирао систем *BioGraph* и повезао га са пет биоинформатичких база података уз коришћење графовске базе података. Поред тога што представља вид потврде валидности модела, ова имплементација има и конкретну биоинформатичку примену.

Приказ дисертације

Рукопис има 129 (113 + XVI) страна, писан је на енглеском језику и има следећу структуру:

1. *Introduction* (срп. Увод)
2. *Cluster analysis and association rules mining* (срп. Анализа кластера и истраживање података о придруживању)

3. *Searching bioinformatics databases* (срп. Претраживање биоинформатичких база података)
4. *New Data Joining Model Proposal* (срп. Предлог новог модела повезивања података)
5. *Model implementation and validation* (срп. Имплементација и валидација модела)
6. *Results and discussion* (срп. Резултати и дискусија)
7. *Conclusion* (срп. Закључак)

Насловна страна, резиме и подаци о ментору и члановима комисије су наведени на енглеском и српском језику. Рад обухвата и *Contents* (срп. Садржај), *List of Figures* (срп. Списак слика), *List of Tables* (срп. Списак табела), *Bibliography* (срп. Списак литературе) од 99 библиографских јединица, додатак са изворним кодом модула за увоз података и Биографију кандидата.

У уводном поглављу представљен је контекст проблема и његов значај. Истакнута је потреба за аутоматизацијом повезивања садржаја различитих биоинформатичких база података. Њихов велики број и разноликост формата и садржаја чине тај проблем и тежим и значајнијим. Уведен је концепт мерења удаљености објеката, односно њихове сличности, најпре да појединачним једноставним атрибутима, а затим и на нивоу сложенијих објеката. Представљен је концепт семантичке сличности објеката, на основу упоређивања њихових метаподатака. Изложени су и познати резултати у области, као и повезане теме.

У другом поглављу се уводе алгоритми кластеровања и истраживања правила придруживања, који могу да се употребљавају за препознавање одређених образаца у подацима и који могу да буду од помоћи при извођењу и описивању односа семантичке сличности између података.

Поглавље 3 обухвата преглед неких од најрепрезентативнијих биоинформатичких база података, међу којима су *UniProt*, *NCBI Gene*, *EMBL* и *Enssembl*. Посебно се бави проблемом претраживања биоинформатичких података и њихових хетерогених схема података и различитих начина представљања и моделирања одређених ентитета. Истичу се тешкоће при покушајима њиховог повезивања, као и различити формати података и протокола који се користе за представљање и описивање објеката у овим базама података, попут *CSV*, *JSON*, *RDF*, *TSV*, *PDB*, *FASTA*, *FASTQ* и други. Представљене су различите врсте објеката који се појављују у тим базама података, као и различити начини идентификовања појединачних објеката.

Поглавље 4 представља предлог новог модела за повезивање биоинформатичких база података. Модел је пројектован тако да може да послужи као основа за обједињавање биоинформатичких података из више извора и за развој система за

извођење нових односа семантичке сличности између објеката различитих база података.

Основни циљ новог модела повезивања је да се омогући семантичко претраживање на основу података из различитих база података, при чему се за њихово повезивање користе првенствено расположиви метаподаци. Мотиви за то су да се за проналажење веза међу објектима не користи прикупљање свих података на једном месту, већ прављење релативно мањих мрежа за индексирање, али и богатство различитих информација које се могу пронаћи у метаподацима. Додатни мотив представља њихова величина, зато што биомедицински подаци могу да буду веома велики, па иако за неке мање податке обим метаподатака може да буде упоредив, па чак и већи, у већини случајева су метаподаци много мањи. Нови модел је назван *BioGraph*. Модел *BioGraph* је обликован тако да за његову имплементацију могу да се користе СУБП који почивају на различитим моделима (пре свега, важно је да може да се користи релациони модел), али је у даљем раду (у поглављу 5) употребљавана графовска база података.

Основу модела чине три типа објеката (ентитети, идентификатори и подаци) и односи који повезују те објекте. Ентитети могу да имају различите типове, који зависе од домена конкретних база података, на пример, могу да буду протеини, гени, болести, антигени, епитопи, региони и друго. У идеалном случају се за ентитет, који се појављује у различитим базама података, прави један објекат (са потенцијално више идентификатора), али често то није могуће, па се прави више различитих објеката, међу којима се касније успостављају одговарајући односи (међу којима и однос „*IS_EQUAL_TO*“). Објекти идентификатора се повезују са ентитетима путем односа („*HAS_ID*“). Додатни метаподаци о ентитетима се чувају у објектима-подацима, са којима се повезују односом „*HAS_DATA*“. Један ентитет може да се повеже са више података, као што и један податак може да се повеже са више ентитета. Подаци могу да представљају различите врсте информација о ентитетима, од њиховог назива, преко специфичних структурних или општих карактеристика па до описа порекла и техника за издвајање ентитета. И међу ентитетима могу да постоје разчичите врсте односа, чијим се установљавањем значајно унапређују могућности претраживања. У оквиру описа модела је у грубим цртама представљено како би модел могао да се имплементира графовском или релационом базом података.

Саставни део модела чини метод за извођење семантичких односа међу објектима. Односи између објеката се успостављају на основу анализе постојећих односа тих објеката са другим објектима и на основу сличности и разлика међу њиховим метаподацима. Уопштену основу поступка чини примена алгоритама истраживања података на подацима о постојећим односима (матрице изабраних врста односа и повезаних ентитета) и одабирање, прављење и додавање нових

односа у колекцију (граф знања) на основу резултата примене тих алгоритама. Избор конкретних скупова и алгоритама зависи од циља. У опису модела је представљено како могу да се користе кластеровање, граф најближих суседа, правила придруживања и латентна семантичка анализа.

У оквиру модела је представљен и основни скуп врста односа (*IS_EQUAL*, *IS_INSTANCE*, *IS_VARIANT*, *FROM*, *CONTAINS*, *HAS_ROLE*, *RELATED_WITH*, *SIMILAR_TO*), али тај скуп може да се проширује према потреби и према доменима конкретних база података које се претражују.

Поглавље 5 описује појединости имплементације предложеног модела и система за извођење нових односа семантичке сличности на основу расположивих података. Имплементација је повезена са пет база података (*DisProt*, *DisGeNET*, *Tantigen*, *IEDB* и *HGNC*). Имплементација језгра система пружа *REST-API* за претраживање. Упити се описују на интерном специфичном и за ове потребе развијеном упитном језику. Кориснички интерфејс је имплементиран у облику веб-интерфејса према језгру система. Све је развијено на језику *JavaScript* и уз употребу окружења (сервера) *NodeJS*.

Архитектуру система чине основни сервиси, увозници података, индексери, адаптери за базе података и *REST* интерфејс:

- Основни сервиси представљају централну компоненту система *BioGraph*. Они омогућавају другим компонентама да додају податке, повезују их или их претражују. Користе трансакционе механизме за одржавање података, чиме се одржава интегритет система.

- Увозници података представљају улазну тачку система. За сваку базу података која се повезује имплементира се по бар један увозник података. Увозник преузима податке из базе података (преузимајући припремљене пакете или обилазећи веб сајт базе или на неки други начин), трансформише их у одговарајући облик и попуњава њима базу *BioGraph* користећи основне сервисе.

- Индексери су сервиси за чување и индексирање свих података у систему. Посебно се индексирају подаци о увозу података, ентитети, идентификатори и описи (метаподаци).

- Адаптери за базе података омогућавају да се систем имплементира помоћу различитих СУБП. Они представљају основу преносивости и флексибилности система. У конкретном случају имплементација користи графовску базу *Neo4J* и одговарајући адаптер. Ради експеримената и провере исправности концепата развијен је и адаптер за *MySQL*.

- Апликативни интерфејс система је развијен применом *REST-a*.

За потребе имплементације је обликован и интерни упитни језик, по узору на *Cypher*. Упити се записују у формату *JSON*.

У 6. поглављу се разматрају карактеристике предложеног модела и конкретне имплементације. Модел се пореди са другим сличним постојећим моделима и истичу се разлике и сличности међу њима. У односу на већину значајних критеријума предложен модел има значајних предности, што се посебно види када се разматрају флексибилност и проширивост модела. Иако је модел у основи независан од употребљеног модела података (релациони или неки други), имплементација се за неки други модел може значајно разликовати.

Научни допринос дисертације

У својој дисертацији кандидат Александар Вељковић представља предлог потпуно новог модела за повезивање различитих биоинформатичких база података, као и опис имплементације тог модела на конкретном примеру.

Кључни научни доприноси дисертације су (1) предложен нов модел повезивања биоинформатичких база података, који обухвата и предложен аутоматизовани протокол за откривање нових семантичких односа између објеката, (2) предложен аутоматизовани протокол за откривање нових односа семантичке сличности између објеката различитих база података, а заснованих на методама истраживања података који су добијени из дефинисаних модела и (3) јавно расположива имплементација предложеног модела (https://github.com/aleksandar-veljkovic/biograph) на примеру повезивања пет биоинформатичких база података уз коришћење графовске базе података, која осим практичне применљивости представља и потврду употребљивости модела и његове добре заснованости.

Закључак

Резултати до којих је Александар Вељковић дошао током истраживања у области семантичког обједињавања биоинформатичких база података, и који су представљени у рукопису "**Семантичко обједињавање и претраживање биоинформатичких база података коришћењем метода истраживања података**" (енг. „***Semantic Unification And Searching Of Bioinformatics Databases Using Data Mining Methods***") представљају вредан научни допринос у областима биоинформатике, база података и примене истраживања података у биоинформатици.

Тема и предмет истраживања су из уже научне области Биоинформатика, а методи истраживања су из уже научне области Истраживање података. Неки од

остварених резултата могу да имају и ширу примену у другим врстама база података. Кандидат Александар Вељковић је показао одлично познавање ових области и оспособљеност за обављање самосталног научног рада. Добијени резултати имају значајан потенцијал за даљу примену и наставак истраживања.

Имајући у виду све претходно наведено предлажемо Наставно-научном већу Математичког факултета да рукопис "**Семантичко обједињавање и претраживање биоинформатичких база података коришћењем метода истраживања података**" (енг. „***Semantic Unification And Searching Of Bioinformatics Databases Using Data Mining Methods***") кандидата Александра Вељковића прихвати као докторску дисертацију и одреди комисију за њену одбрану.

У Београду, 25.12.2023.

Чланови комисије за оцену

_____

(др Саша Малков, ванредни професор)

_____

(др Јована Ковачевић, доцент)

_____

(др Александар Картељ, ванредни професор)

_____

(проф. др Наталија Половић, редовни професор
Универзитет у Београду – Хемијски факултет)

_____

(проф. др Јурij Орлов,
професор Руске академије наука,
Институт за дигитално здравље, Први московски државни медицински
универзитет И.М.Сеченов, Москва, Русија)