Наставно-научном већу
Математичког факултета
Универзитета у Београду

Одлуком Наставно-научног већа Математичког факултета донетом на 372 седници одржаној 03.07.2020. године, именовани смо у комисију за преглед и оцену докторске дисертације под називом **Развој метода за анализу сличности биолошких секвенци на основу карактеристика поновака** који је предат као докторска дисертација кандидата, мастер математичара, Јасмине Јовановић. Након прегледа рукописа подносимо Наставно-научном већу следећи

Извештај

1 Биографија кандидата

Јасмина Јовановић рођена је 19. октобра 1987. године у Пожаревцу. Основну школу "Иво Лола Рибар" и Гимназију у Великом Градишту завршила је као носилац дипломе Вук Караџић и ђак генерације. Математички факултет у Београду, смер Рачунарство и информатика, уписала је 2006. године. Дипломирала је у јунском испитном року 2010. године са просечном оценом 9.82. Мастер студије, на студијском програму Математика, модул Рачунарство и информатика, завршила је 2011. године са просечном оценом 10.00. Мастер рад под називом "Дизајн и имплементација апликације за мобилна плаћања рачуна путем SMS и USSD сервиса" одбранила је под менторством проф. Владимира Филиповића. Добитник је стипендије "Доситеја", фонда за младе таленте Републике Србије за студенте завршних година студија 2009. године, као и 2010. године као студент мастер студија.

Докторске студије на студијском програму Информатика уписала је 2011. године. Све испите предвиђене планом студија положила је са просечном оценом 9.83. Њена основна област интересовања је истраживање података у биоинформатици. Током докторских студија била је на стручном усавршавању у Лондону на Унивезитету "Imperial College London", где је добила сертификат за "Студије асоцијације целокупног генома".

2 Списак научних радова

Јасмина Јовановић је до сада била аутор четири рада у међународним часописима на СЦИ листи, од којих један самосталан, и имала четири саопштења на научним скуповима.

Радови у међународним часописима са СЦИ листе

1. Jovanović J. New Method for Sequence Similarity Analysis Based on the Position and Frequency of Statistically Significant Repeats. Current Bioinformatics. 2021; 16(10):1299- 1310, doi: 10.2174/1574893616999210805165628, (M21; IF2020= 3.543).

2. Zeljic K, Jovanovic I, Jovanovic J, Magic Z, Stankovic A, Supic G. MicroRNA meta-signature of oral cancer: evidence from a meta-analysis. Ups J Med Sci. 2018; 123(1): 43- 49, doi: 10.1080/03009734.2018.1439551, (M21; IF2018= 2.747).

3. Jovanović I, Zivković M, Jovanović J, Djurić T, Stanković A. The Co-Inertia approach in identification of specific microRNA in early and advanced atherosclerosis plaque. Med Hypotheses. 2014;83(1):11-5, doi: 10.1016/j.mehy.2014.04.019, (M23; IF2014= 1.074).

4. Stojkovic G; Jovanovic I; Dimitrijevic M; Jovanovic J; Tomanovic N; Stankovic A; Arsovic N; Boricic I; Zeljic K. The meta-signature guided investigation of miRNA candidates as potential biomarkers of oral cancer. Oral Diseases, 2022, 1– 15, doi: 10.1111/odi.14185, (M21; IF2014= 3.511).

Саопштења на научним скуповима

1. Jovanović I, Zivković M, Jovanović J, Djurić T, Stanković A. Could integrative bioinformatic approach predict the circulating miRs that have significant role in pancreatic tissue in type 2 diabetes?. Belgrade Bioinformatic Conference BelBi 2016 Proceedings, 2017; 82-87.

2. Jovanović I, Zivković M, Jovanović J, Djurić T, Stanković A. The improvement of microRNA activity prediction: The integration of Co Expression Meta Analysis of microRNA Targets into Co-Inertia analysis, NGS and non-coding RNA data analysis COST workshop, 15-16 May 2014 Plovdiv, Bulgaria.

3. Jovanović I, Zivković M, Jovanović J, Djurić T, Stanković A. The prediction of potentially new microRNA biomarkers for advanced atherosclerosis using Co-inertia analysis, V Congress of the Serbian Genetic Society, September 28th – October 2nd 2014, Kladovo, Serbia.

4. Zeljić K, Jovanović I, Jovanović J, Magić Z, Stanković A, Šupić G. Identification of miRNA meta-signature for discrimination between oral cancer and normal tissue: meta-analysis approach. The Third Congress of the Serbian Association for Cancer Research SDIR-3, Belgrade, Serbia, 6-7 October 6th – 7th, 2017.

3 Предмет и садржај дисертације

Предмет докторске дисертације је анализа биолошких (нуклеотидних и аминокиселинских) секвенци и њихових статистички значајних поновака различитих типова и дужина у

циљу развоја нових модела за одређивање сличности секвенци на основу идентификованих поновака. Понављајуће (репетитивне) секвенце (ниске) представљају делове секвенци (подсеквенце) који се јављају два или више пута. У зависности од тога да ли је подсеквенца копије идентична оригиналној, или заједно са оригиналном чини палиндром, поновци могу бити директни или обрнути. Такође поновци могу бити подељени у комплементарне и некомплементарне поновке у зависности од тога да ли испуњавају функцију пресликавања свих карактера у њихове комплементарне карактере. Статистички значајни поновци представљају подскуп поновака секвенце за које није очекивано да ће се појавити у насумичној секвенци исте дужине.

Биолошки макромолекули полимерне природе (ДНК, РНК, протеини) се могу посматрати као ниске карактера. ДНК се може представити као ниска карактера над азбуком A = A, C, G, T, док се аминокиселинска ниска може посматрати као ниска карактера над 20-ословном азбуком сачињеном од ознака амино киселина. Дужина нуклеотидних секвенци варира од неколико нуклеотида до више стотина милиона нуклеотида, док дужина аминокиселинксих секвенци варира од неколико десетина до више хиљада аминокиселина. Анализа сличности нуклеотидних и протеинских секвенци је важна у одређивању функционалних, структурних и еволуционих односа између различитих таксономских категорија и/или других карактеристика организама. Њена значајност се такође огледа у одређивању категорија новооткривених секвенци, поређењем са секвенцама које имају познате функције. Ове анализе су најчешће коришћене и примењиване у биоинформатици. Постоје различити алгоритми за поређење секвенци, који се могу поделити на алгоритме засноване на поравнању (глобална и локална поравнања) или без поравнања секвенци. Такође се могу класификовати на методе за поравнање парова (две секвенце) и за вишеструко поравнање (са више од две секвенце). Више алгоритама које користе технике динамичког програмирања (као што су Needleman-Wunsch и Smith-Waterman), као и хеуристички алгоритми (FASTA, BLAST и ClustalW) су успешно развијени у циљу решавања проблема поравнања секвенци и анализе сличности, који се могу, поред нуклеотидних и протеинских, применити и на друге секвенце (на пример у лингвистици за анализу природних језика).

До сада развијени и коришћени алгоритми нису засновани на различитим типовима статистички значајних поновака варијабилних дужина. Предмет ове дисертације је анализа нуклеотидних и протеинских секвенци и њихових поновака у циљу развоја нових модела за идентификовање сличности секвенци на основу изабраних поновака. Класификација (до сада некласификованих) улазних секвенци је вршена на два начина: формирањем вектора карактеристичних тачака секвенце и формирањем њихових профила на основу поновака и поређењем са базом секвенци (за први начин) односно профила секвенци (за други начин) које имају познате карактеристике. У формирању класификационог модела за идентификацију сличности секвенци коришћене су методе истраживања података (класификација, кластеровање) на основу којих је омогућена провера прецизности добијених резултата.

4 Приказ дисертације

Рукопис се састоји од 114 страница (vi+108) и има следећу структуру:

1. Увод
2. Основе и сродни приступи истраживања сличности биолошких секвенци
3. Нове методе за анализу сличности биолошких секвенци
4. Резултати и дискусија
5. Закључак и даљи рад

уз Резиме (на енглеском и српском језику), Садржај, Додатак, Списак литературе који се састоји од 86 библиографских јединица и Биографију кандидата.

У уводном поглављу је дат приказ основних појмова, описан предмет истраживања и циљ дисертације. Такође је приказана и организација тезе по поглављима.

Друго поглавље садржи кратак опис база биолошких података из којих је узет материјал обрађиван у дисертацији, детаљан опис поновака и њихових карактеристика, и дат приказ постојећих метода за одређивање сличности секвенци. Поглавље се завршава описом метода, техника, и навођењем метрика/мера истраживања података које су коришћене у анализи сличности секвенци.

У трећем поглављу које заједно са четвртим чини централни део рада су описане нове методе предложене за анализу сличности биолошких секвенци. Предложене су метода заснована на позицији и локалној учесталости поновака и методе засноване на потписима секвенци и профилима категорија.

У методи заснованој на позицији и локалној учесталости поновака (*R-P/F* метода (енгл. *Repeats-Position/Frequency method*) одређивање сличности биолошких секвенци се заснива на израчунавању вредности ентропије засноване на локалној учесталости улазне секвенце и поновка, узимајући у обзир број појављивања, позицију понављајуће секвенце, као и чињеницу да није очекивано истоветно појављивање поновака у случајно изабраним секвенцама исте дужине. Секвенце су представљене нумеричким векторима у вишедимензионом векторском простору чија димензија одговара броју различитих поновака у секвенци. Односи између секвенци су идентификовани користећи мере за одређивање растојања вектора у векторском простору. На основу добијених резултата формира се матрица сличности, која се даље користи у алгоритмима хијерархијског кластеровања, где се сличност секвенци очитава из дендограма који се формира при кластеровању.

Идеја коришћена у другој групи метода се заснива на коришћењу (мањег) скупа карактеристичних тачака секвенци, уместо читавих секвенци. У овој групи метода се поређење врши са секвенцама које припадају познатим таксономским категоријама. Методе засноване на потписима секвенци и профилима садрже:

1. Методу за одређивање сличних секвенци поређењем потписа секвенци. Ова метода омогућава одређивање сличности секвенци, која као резултат може да произведе сличност са једном или више секвенци које могу да припадају и различитим таксономским категоријама. На основу идентификованих парова поновака и њихових растојања се формира скуп потписа секвенци и профила категорија. Потпис једне секвенце се дефинише као скуп свих уређених парова (поновак r, растојање леве и десне компоненте d_r) који су садржани у датој секвенци. Потпис секвенце је вектор уређених парова, који може да садржи већи број уређених парова са идентичним вредностима за поновак r и растојање d_r. Овакав концепт одређивања потписа секвенце је независан од дужине саме секвенце. Сличност улазних секвенци се израчунава одређивањем сличности између потписа секвенци. Сврха овако конципиране методе је идентификовање сличних секвенци као и њихових односа. Такође, дата метода може да се користи у идентификовању таксономске категорије непознате секвенце (која може да буде потпуна или делимична) поређењем потписа те секвенце са базом потписа познатих секвенци.

2. Методу за класификацију секвенци засновану на профилима категорија. Ова метода омогућава добијање прецизнијих резултата, тачније таксономске категорије којој припада улазна секвенца. Профил категорије се формира на основу скупа потписа секвенци које припадају тој категорији. Скуп потписа секвенци се формира идентификовањем свих уређених парова (r, d_r) који су присутни у одређеном броју потписа секвенци које припадају категорији чији се профил одређује. Профил категорије се конструише узимајући у обзир све уређене парове (r, d_r) који се појављују у неком проценту посматраног броја секвенци те категорије, уз услов да се уређени парови који припадају профилу једне категорије не појављују у профилу друге категорије (ако дође до те ситуације, такви парови се искључују из оба профила). Категорија непознате секвенце се одређује поређењем потписа секвенце са профилима категорија у бази података. Сличност између потписа секвенце и профила категорије се израчунава користећи изабрану меру сличности.

У четвртом поглављу *Резултати и дискусија* су приказане примене предложених модела на различите скупове улазних података за различите мере сличности. Прва метода је тестирана на скупу митоходнријалних ДНК, секвенце РНК вируса еболе, маргбург вируса и бетакоронавируса, док је улазни тест материјал за другу групу мтода укључивао 4.231 секвенцу вируса преузетих из ICTV базе података (*International Committee on Taxonomy of Viruses (ICTV) database*. Добијени резултати су потврдили коректност предложених модела при чему је прецизност била упоредива са резултатима до сада широко коришћених метода (нпр. са BLAST методом).

У петом поглављу *Закључак и даљи рад* је дат сумарни приказ садржаја дисертације и наведени планови за даљу обраду који укључују проширење базе података која садржи скупове секвенци и њихове поновке, унапређење предложених метода у циљу побољшања брзине извршавања, могућност коришћења нових мера сличности, као и неколико других проширења која би повећала прецизност и расположивост базе.

У Додатку су приказане табеле и слике које садрже резултате за различите типове поновака који нису приказани у четвртом поглављу: резултати методе засноване на позицији и локалној учесталости поновака, као и табеле са резултатима методе засноване на потписима секвенци и профилима категорије.

5 Закључак

Рукопис **Развој метода за анализу сличности биолошких секвенци на основу карактеристика поновака** садржи вредан научни допринос у области примене истраживања података у биоинформатици.

У раду је разматран проблем одређивања сличности биолошких (нуклеотидних и аминокиселинских) секвенци. Предложене су нове методе засноване на идентификацији поновака у секвенцама. Употребом поновака, секвенце се представљају преко нумеричких вектора у вишедимензионом простору. Поређење се врши са скупом познатих секвенце чије се карактеристике чувају у бази, односно са карактеристима таксономских категорија скупова секвенци. Добијени резултати су упоредиви по прецизности са до сада постојећим методама, због чувања само мањег скупа карактеристика у односу на комплетне секвенце захтева мање простоара од до сада коришћених метода, а у зависности од величине скупа података чије су карактеристике претходно израчунате и смештене у базу, извршавање у појединим случајевима може да буде знатно брже.

Имајући у виду претходно наведено предлажемо Наставно-научном већу Математичког факултета да рукопис **Развој метода за анализу сличности биолошких секвенци на основу карактеристика поновака** кандидата Јасмине Јовановић прихвати као докторску дисертацију и одреди комисију за њену одбрану.

У Београду, 04.07.2022.

Чланови комисије за преглед и оцену

(проф. др Ненад Митић, редовни проф.)

(проф. др Гордана Павловић-Лажетић, редовни проф.)

(др Јована Ковачевић, доцент)

(др Зоран Огњановић, научни саветник)