Наставно-научном већу
Математичког факултета
Универзитета у Београду

Одлуком Наставно-научног већа Математичког факултета у Београду донетом на 383. седници одржаној 18.6.2021. године именовани смо за чланове комисије за оцену докторске дисертације „***Prediction of alphabets of local protein structures using data mining methods***" (срп. „**Предвиђање алфабета локалне структуре протеина применом метода истраживања података**") кандидаткиње **Мирјане Маљковић**, Мастер информатичара. После прегледа поднетог рукописа подносимо следећи

Извештај

Биографски подаци

Мирјана Маљковић је рођена 13. новембра 1986. године у Канбери, Аустралија. Основну школу и гимназију завршила је у Београду. Школске 2005/2006. године уписала је основне академске студије на Математичком факултету у Београду (смер Информатика), а школске 2008/2009. године уписала је мастер академске студије на Математичком факултету, смер Информатика. На истом факултету уписала је докторске академске студије, смер Информатика, у оквиру којих је положила све испите предвиђене планом и програмом докторских студија са просечном оценом 10,00.

Од 2009. године запослена је на Математичком факултету Универзитета у Београду као сарадник у настави на Катедри за рачунарство и информатику, а од 2011. године као асистент у настави на истој Катедри. До сада је учествовала у извођењу вежби из следећих предмета на основним и мастер студијама:

Релационе базе података, Програмирање база података, Развој софтвера, Истраживање података и Истраживање података 1.

Основне области интересовања су јој базе података, истраживање података и биоинформатика. Учествовала је као истраживач на пројекту „Развој нових информационо-комуникационих технологија, коришћењем напредних математичких метода, са применама у медицини, телекомуникацијама, енергетици, заштитити националне баштине и образовању" Министарства просвете и науке Републике Србије.

Мирјана Маљковић је била члан организационог одбора међународних конференција *Belgrade Bioinformatics Conference 2016* (BelBi2016) и *Belgrade Bioinformatics Conference 2018* (BelBi2018). Члан је техничког уредништва часописа "*Преглед Националног центра за дигитализацију*", који издаје Математички факултет Универзитета у Београду.

Објављени научни радови и саопштења са научних скупова

Мирјана Маљковић има једанаест радова и саопштења са научних скупова, од којих су два у часопису са SCI листе (M21, M22), а један од њих је самосталан. Мирјана је у седам радова први аутор.

Радови у међународним часописима са SCI листе (два, од тога један самосталан):

1. Maljković M. *DistAA: Database of amino acid distances in proteins and web application for statistical review of distances*. Computational Biology and Chemistry. 2019. doi:10.1016/j.compbiolchem.2019.107130 (M22, IF2019=1.850)
2. Akhila M, at al. *A structural entropy index to analyse local conformations in intrinsically disordered proteins. Journal of Structural Biology*. 2020. doi:10.1016/j.jsb.2020.107464 (M21, IF2019=3.071)

Радови у међународним часописима:

3. Akhila M, at al. *Data set of intrinsically disordered proteins analysed at a local protein conformation level*. Data in Brief. 2020. doi:10.1016/j.dib.2020.105383

Радови у домаћим часописима:

4. Maljković M, Stojanović B, Mijajlović Ž. *Digital Legacy Of Mihailo Petrović Alas*. Review of the National Center for Digitization. 2017;31:10 - 17. issn: 1820-0109
5. Maljković M, Stojanović B, Mijajlović Ž. *Digital Legacy Of Professor Slaviša Prešić*. Review of the National Center for Digitization. 2017;30:1-6. issn: 1820-0109
6. Maljković M, Malbašić D. *Digitalization of Serbian Heritage in Village Osredak in Bosnia and Hercegovina*. Review of the National Center for Digitization. 2016;28:32 - 42. issn: 1820-0109

Саопштења на међународним конференцијама штампана у целини или изводу:

7. Maljkovic M, Mitic N, De Brevern A. *Models for Prediction of Structural Alphabet Protein Blocks*. Book of Abstracts: Book of Abstracts Belgrade BioInformatics Conference 2021. 21-25 June 2021, Vinča, Serbia
8. Mitić N, Pavlović-Lažetić G, Malkov S, Beljanski M, Maljković M, Jelović A. *Accuracy of disorder predictors results –comparison on DisProt DB data*. Book of Abstracts: International Multiconference on "Bioinformatics of Genome Regulation and Structure / Systems Biology" – BGRS/SB-2020. Novosibirsk, 06. – 10 Jul, 2020
9. Maljković M, Mitić N, Beljanski M. *Analysis of Amino Acid Interactions Based on Geometric Distances*. Book of Abstracts Belgrade BioInformatics Conference 2018. Biologia Serbica. 2018;40(1):116 - 116. issn: 2334-6590, 18. - 22. Jun, 2018, Belgrade
10. Maljković M, Beljanski M, Mitić N. *Analysis of amino acid distances in protein chains*. 14th International Conference on Fundamental and Applied Aspects of Physical Chemistry - Physical Chemistry 2018. 2018;1:543 - 546. isbn: 978-86-82475-36-1, 24. - 28. Sep, 2018, Belgrade

Саопштења на домаћим конференцијама штампана у изводу:

11. Mitić N, Malkov S, Stojanović B, Maljković M. *Nacionalni centar za digitalizaciju - prošlost, sadašnjost i budućnost*. Book of Abstracts: The Seventeenth National Conference Digitization of Cultural Heritage, Old Records from the Natural and Social Sciences and Digital Humanities.17 Sep 2019, Belgrade

Предмет дисертације

Предмет докторске дисертације чине прављење модела за предвиђање прототипова структурног алфабета за задату аминокиселинску секвенцу коришћењем метода истраживања података, анализа корисности развоја новог структурног алфабета, као и прављење модела за предвиђање прототипова новог структурног алфабета применом метода истраживања података. Предмет докторске дисертације припада научној области Рачунарство и информатика и ужим научним областима Биоинформатика и Истраживање података.

Структура протеина се описује на три нивоа: (1) примарном структуром која се назива и аминокиселинска секвенца, (2) секундарном структуром и (3) тродимензионалном структуром (3D). Пошто је експериментално одређивање тродимензионалне структуре протеина скупо и временски захтевно, постојање програма који на основу аминокиселинске секвенце предвиђају особине 3D

структуре је изузетно значајно за истраживања заснована на структури протеина, као што је функција протеина или развој лекова. 3D структура главног ланца протеина (енг. *backbone*) може да се приближно опише и коришћењем прототипова локалне структуре протеина. Прототипови се користе за апроксимацију локалних савијања која се јављају у структури познатих протеина, а одређују се на основу издвојених фрагмената узастопних аминокиселина у полипептидним ланцима чија је 3D структура позната. Скуп дефинисаних прототипова локалне структуре чини библиотеку локалних структура протеина, која се још назива и структурни алфабет (СА) (енг. *structural alphabet*).

Један од најпознатијих структурних алфабета, чији је оригинални рад цитиран више од 350 пута, се назива Протеински блокови (ПБ). Структурни алфабет Протеински блокови се састоји од 16 прототипова који су издвојени применом алгоритама кластеровања на фрагментима од 5 узастопних аминокиселина из протеина са познатом структуром.

У дисертацији су предложени модели за предвиђање прототипова структурног алфабета за задату аминокиселинску секвенцу засновани на структурном алфабету Протеински блокови и различитим приступима истраживања података према формату података и алгоритмима за класификацију (дрвета одлучивања и неуронске мреже) који могу да повећају прецизност у односу на постојеће предикторе који се користе у те сврхе. Поред аминокиселинске секвенце, улаз у моделе за предвиђање прототипова Протеинских блокова је обогаћен додатним информацијама о протеинском ланцу које се могу одредити или предвидети на основу аминокиселинске секвенце (предвиђене структурне особине протеина, позиције директних и инверзних поновака у аминокиселинској секвенци и резултати осам предиктора за одређивање позиција могућих неуређених региона у протеину). У раду је такође анализирана корисност развоја новог структурног алфабета применом алгоритма за кластеровање *TwoStep* и модела за предвиђање прототипова добијених структурних алфабета у циљу коришћења новог структурног алфабета као основе за развој нових предиктора локалне структуре протеина.

Приказ дисертације

Рукопис има 165 (149 + XVI) страна, писан је на енглеском језику и има следећу структуру:

1. *Introduction* (срп. Увод)
2. *Data mining methods* (срп. Методе истраживања података)
3. *Structural alphabets* (срп. Структурни алфабети)
4. *New Protein Blocks predictors* (срп. Нови предиктори за Протеинске блокове)

5. *Development of new structural alphabets* (срп. Развој нових структурних алфабета)
6. *Conclusion* (срп. Закључак)

уз *Abstract* (срп. Резиме) на енглеском и српском језику, *Contents* (срп. Садржај), *List of Figures* (срп. Списак слика), *List of Tables* (срп. Списак табела), *Appendix* (срп. Додатке), *Bibliography* (срп. Списак литературе) од 88 библиографских јединица и Биографију кандидата.

У уводном поглављу описан је шири контекст проблема и његов значај. Приказана је и организација тезе по наредним поглављима.

Друго поглавље садржи приказ метода истраживања података и формата података који су коришћени у дисертацији за развој предложених модела за предвиђање прототипова стуктурних алфабета и њихово одређивање. Описана су два формата за припрему података у оквиру дисертације: формат клизајућег прозора фиксне дужине (енг. fixed-length sliding window) и секвенцијални формат. Дат је приказ шест алгоритама дрвета одлучивања (*C5.0*, *CART*, *XGBoost Tree*, *CHAID*, *SPRINT*, *Random Forests*) и два алгоритма вештачких неуронских мрежа (*Multilayer Perceptron*, *LSTM-Bidirectional recurrent neural networks*), као и два алгоритма за кластеровање: Кохоненове самоорганизујуће мапе (енг. *Kohonen Self-Organizing Feature Map*) и *TwoStep*. Алгоритам Кохоненове самоорганизујуће мапе је коришћен за развој постојећег структурног алфабета Протеински блокови, а *TwoStep* је коришћен у анализи развоја нових структурних алфабета у оквиру дисертације.

Треће поглавље садржи опис структуре протеина, мотив за дефинисање структурних алфабета и опис неколико структурних алфабета. Структурни алфабети представљају један начин описа 3D структуре главног ланца протеина. Развијени су са циљем да се превазиђе проблем непостојања стриктних правила за одређивање секундарне структуре протеина, што доводи до постојања већег броја приступа за одређивање секундарне структуре чији се резултати разликују. Структурни алфабет представља библиотеку прототипова локалне структуре протеина који се одређују на основу издвојених фрагмената узастопних аминокиселина у полипептидним ланцима чија је 3D структура позната. Структурни алфабет је дефинисан као скуп од N прототипова дужине l аминокиселина. У овом поглављу је дат преглед осам структурних алфабета који се разликују по својствима коришћеним за опис главног ланца протеина ($C\alpha$ координате, $C\alpha$ растојања, торзиони углови), методама коришћеним за дефинисање прототипова (алгоритми кластеровања K-средина, хијерархијско кластеровање, Кохоненове мапе, вештачке неуронске мреже или метод за кластеровање заснован на доменском знању и развијен за потребе истраживања), броју аминокиселина у фрагментима и броју прототипова. Сваком прототипу

структурног алфабета се додељује посебна ознака, те се применом структурних алфабета главни ланац протеина представља као низ ознака прототипова структурних алфабета. Детаљније је описан најпознатији структурни алфабет Протеински блокови који је коришћен као основа за нове моделе, описане у оквиру дисертације, за предвиђање прототипова структурног алфабета. Структурни алфабет Протеински блокови се састоји од 16 прототипова који су означени словима од *a* до *p* и одређени на основу фрагмената од 5 узастопних аминокиселина применом Кохоненове самоорганизујуће мапе. Дат је преглед неколико постојећих алата за предвиђање прототипова структурног алфабета Протеински блокови на основу аминокиселинске секвенце који су развијени применом различитих приступа. Алати се могу поделити у три групе:

(1) алати засновани на Бајесовом приступу који поред аминокиселинске секвенце користе и дефинисане фамилије секвенци,

(2) алат *PBk-PRED* који користи приступ заснован на доменском знању и базу података са фрагментима дужине 5 из протеина са експериментално одређеном 3D структуром и придруженим прототиповима Протенских блокова, и

(3) алати засновани на алгоритмима машинског учења и информацијама из хомологих протеина које се добијају на основу аминокиселинске секвенце применом алата *PSI-BLAST*.

Према објављеним подацима у научним радовима, најбољу прецизност међу алатима за предвиђање прототипова структурног алфабета Протеински блокови постиже алат *SVMprat* и она износи 68,9% (новије процене показују да је ова вредност око 55%).

У четвртом поглављу *New Protein Blocks predictors* (срп. Нови предиктори за Протеинске блокове), које представља централни део рада, детаљно су описани новоразвијени модели за предвиђање прототипова структурног алфабета Протеински блокови на основу аминокиселинске секвенце. Поред аминокиселинске секвенце, као улаз у моделе за предвиђање Протеинских блокова, користе се и информације о протеинском ланцу које се могу одредити или предвидети на основу аминкиселинске секвенце. Метод *Spider3* је коришћен за предвиђање неких структурних особина протеина (торзионих углова, секундарне структуре и приступачне површине аминокиселина); програм *StatRepeat* за проналажење позиција директних и инверзних поновака (понављајућих ниски) у аминокиселинској секвенци и осам предиктора за одређивање позиција могућих неуређених региона у протеину. Коришћена су два формата за припрему података за прављење модела: формат клизајућег прозора фиксне дужине и секвенцијални формат. У тези је предложено 11 модела за предвиђање прототипова Протеинских блокова; 10 модела за податке у формату

клизајућег прозора фиксне дужине и један модел за податке у секвенцијалном формату. За изградњу модела заснованих на скупу података у формату клизајућег прозора фиксне дужине коришћени су алгоритми дрвета одлучивања (*C5.0*, *CART*, *XGBoost Tree*, *CHAID*, *SPRINT*, *Random Forests*) и алгоритам неуронских мрежа *Multilayer Perceptron*, а за изградњу модела заснованог на подацима у секвенцијалном формату алгоритам *LSTM-Bidirectional recurrent neural networks*. Параметри модела су оптимизовани применом унакрсне провере. Прецизност сваког оптималног модела је проверена на скупу који није учествовао у изградњи модела. Детаљније су анализирани резултати за 4 најбоља модела према прецизности. Најбољи добијени модели према прецизности имају прецизност у опсегу од 74% до 80%, а најбољи модел према прецизности је изграђен применом алгоритма *C5.0*. Извршено је поређење нових модела, анализа значајности коришћења позиција неуређених региона у протеину и поновака у аминокиселинској секвенци као део улаза, као и анализа перформанси најбољег модела. Уочено је да се боља прецизност за неке прототипове постиже применом модела који нису најбољи према глобалној прецизности. Најбољи добијени модел према прецизности је упоређен са публикованим резултатима постојећих алата за предвиђање прототипова Протеинских блокова. Уочено је да нови модел има бољи одзив за све прототипове осим за прототип *g*.

У петом поглављу је описана анализа корисности развоја нових структурних алфабета применом алгоритма за кластеровање *TwoStep*. Описано је више структурних алфабета који су развијени за потребе анализе. Нови структурни алфабети се разликују према броју прототипова које садрже (од 10 до 100) и дужини фрагмената у броју аминокиселина (од 4 до 10) који су коришћени за проналажење прототипова. Торзиони углови су коришћени за опис фрагмената. За сваки издвојен структурни алфабет направљен је модел за предвиђање прототипова применом вештачких неуронских мрежа. Добијени структурни алфабети су поређени према способности да апроксимирају структуру протеина и прецизности направљеног модела за предвиђање прототипова на основу аминокиселинске секвенце. Закључак анализе добијених структурних алфабета је да би прецизни модели за предвиђање структурног алфабета са 16 или више прототипова дужине 4 био користан за предвиђање угла ψ. Модели за предвиђање прототипова нових структурних алфабета имају прецизност између 49% и 77%. Структурни алфабет развијен у оквиру истраживања који има 16 прототипова дужине 5 аминокиселина (CA 5_16) је упоређен са структурним алфабетом Протеински блокови. Уочено је да за 6 прототипова у CA 5_16 постоји по један веома сличан прототип у Протеинским блоковима са *rmsda* до 6,8°, а за 10 прототипова у CA 5_16 по један прототип у Протеинским блоковима са *rmsda* до 16,8°. *rmsda* (*root mean square deviation on angular values*) је мера различитости заснована на угловима. Такође, дистрибуција прототипова оба структурна алфабета је слична према класификацији прототипова заснованој на секундарним

структурама са којима су прототипови повезани. Сличност СА 5_16 и Протеинских блокова показује да се примењени поступак за прављење структурних алфабета може користити као основа за развој нових структурних алфабета са прототиповима погодним за представљање локалне структуре протеина.

У шестом поглављу је дат сумарни приказ садржаја дисертације и планова за будући рад.

У одељку *Appendix* (срп. Додатак) су у графичком облику приказани резултати кластеровања применом алгоритма *TwoStep* који су коришћени за издвајање прототипова структурних алфабета у петом поглављу и подаци о прототиповима издвојених структурних алфабета. Дат је приказ (а) величине кластера према броју придружених фрагмената за свако кластеровање на основу кога је издвојен структурни алфабет, (б) углова главног ланца за прототипове добијених структурних алфабета, и (ц) заступљености сваког од стања секундарне структуре по позицијама аминокиселина у фрагментима покривеним прототиповима нових структурних алфабета са 20 прототипова.

Научни допринос дисертације

У раду је разматран проблем предвиђања прототипова алфабета локалне структуре протеина. У току рада су дефинисани нови модели за предвиђање прототипова најзначајнијег структурног алфабета Протеински блокови. Дефинисани модели су засновани на алгоритмима класификације. Прецизност добијених модела је већа од прецизности постојећих модела за решавање овог проблема. Најбољи модел према прецизности, који је развијен применом алгоритма *C5.0*, има презнизност 80% (најбољи познати постојећи модел има прецизност 68,9%).

У раду је приказана и анализа корисности развоја нових структурних алфабета применом алгоритма за кластеровање TwoStep над фрагментима описаним помоћу углова главног ланца протеина, као и модела за предвиђање њихових прототипова. Закључак је да би нови структурни алфабет са 16 или више прототипова дужине 4 био користан као основа за прављење модела за предвиђање угла ψ главног ланца.

Закључак

Резултати до којих је Мирјана Маљковић дошла током истраживања и који су представљени у рукопису "***Prediction of alphabets of local protein structures using data mining methods***" (срп. „Предвиђање алфабета локалне структуре протеина применом метода истраживања података") представљају вредан научни допринос у областима биоинформатике 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. Кандидаткиња Мирјана Маљковић је показала одлично познавање обе 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 „***Prediction of alphabets of local protein structures using data mining methods***" (срп. „*Предвиђање алфабета локалне структуре протеина применом метода истраживања података*") кандидаткиње Мирјане Маљковић прихвати као докторску дисертацију и одреди комисију за њену одбрану.

У Београду, 28.7.2021.

Чланови комисије за оцену

(др Саша Малков, ванредни професор)

(др Јована Ковачевић, доцент)

(др Alexandre G. de Brevern, виши научни сарадник)