УНИВЕРЗИТЕТ У БЕОГРАДУ
МАТЕМАТИЧКИ ФАКУЛТЕТ
КАТЕДРА ЗА РАЧУНАРСТВО И ИНФОРМАТИКУ

УНИВЕРЗИТЕТ У БЕОГРАДУ
МАТЕМАТИЧКИ ФАКУЛТЕТ
Бр. 1038/1
16.12. 20 22. год.
Београд, Студентски трг 16
тел. 20 27 801, ФАКС: 26 30 151

Предмет: Пријава теме за израду докторске дисертације

Молим да ми се одобри израда докторске дисертације под насловом „**Примена метода истраживања података на употребу кодона ради класификације протеинских секвенци и одређивања међусобне сличности вируса**". Тему предлажем у договору са ментором проф. др Ненадом Митићем.

У Београду, 09.12.2022. године

Подносилац пријаве:

_____
Биљана Стојановић
број индекса 2005/2017

Комисија за оцену научне заснованости теме докторске дисертације

_____
др Саша Малков, ванредни професор Математичког факултета Универзитета у Београду

_____
др Мирјана Маљковић, доцент Математичког факултета Универзитета у Београду

_____
др Марија Гавровић-Јанкуловић, редовни професор Хемијског факултета Универзитета у Београду

У Београду, 09.12.2022. године

Наставно-научном већу Математичког факултета Универзитета у Београду

Пријава теме за израду докторске дисертације кандидата
Биљане Стојановић

"Примена метода истраживања података на употребу кодона ради класификације протеинских секвенци и одређивања међусобне сличности вируса "

Образложење теме

1. Научна област из које је тема

Предложена тема припада научној области Рачунарство и информатика, а ужој научној области Биоинформатика.

2. Предмет и садржај докторске дисертације

Протеини или беланчевине су главни извршни биолошки макромолекули. Они се састоје од једне или више ниски (секвенци) аминокиселина. Редослед аминокиселина у протеинским нискама одређује структуру и функцију датог протеина. Редослед аминокиселина је пак, одређен редоследом три узастопна нуклеотида који чине триплет, генетски код или кодон. Кодони чине гене који кодирају протеине.

Четири врсте нуклеотида из ДНК молекула могу да дају 64 могућа триплета – кодона, који кодирају 20 аминокиселина у молекулу протеина. Од тога, 61 кодон кодира аминокиселине, а три кодона су интерпункција, односно означавају крај низа аминокиселина. Пошто 61 кодон кодира 20 аминокиселина, многе аминокиселине су кодиране са више кодона (2-6), који се називају синонимним или истозначним кодонима. Из тог разлога се каже да је генетски код редундантан, тј. дегенерисан. Установљено је да функција која врши пресликавање триплета у аминокиселине није идентична за све организме. Често се појављује корелација аминокиселина са одређеним кодонима унутар гена који кодирају протеине, као и унутар скупа таквих гена у организмима, или врстама. Овај феномен се означава као употреба кодона (енг. *Codon Usage*).

Уставновљено је да склоност ка употреби одређених кодона зависи од многих фактора, као што су:

- заступљеност, односно проценат *GC* нуклеотида у ДНК (ако организми садрже већи проценат *GC* или *AT* нуклеотида, неки кодони за дату аминокиселину могу бити заступљенији од других);
- ниво генске експресије (постоји врло јака корелација између нивоа генске експресије и склоности као одређеним кодонима; у јако експримиране гене убрајају се сви они гени чији су продукти критични за брзу производњу велике количине исправно структуираних протеина; постоји селекцијски притисак да се користе управо такви "оптимални кодони");
- хоризонтални пренос гена (тј. размена генетичког материјала између врста), и
- физиолошке и еко карактеристике

Одређивање склоности појединачних (група) организама (или њихових вируса) као и функционалних карактеристика протеина унутар организма према употреби кодона представља проблем који још увек није решен. Развијене су бројне биоинформатичке методе (мере) за анализу склоности према одређеним кодонима. Неке од мера су: релативна употреба синонимних кодона (енг. *Relative Synonymous Codon Usage, RSCU, Sharp and Li 1987*), адаптациони индекс кодона (енг. *Codon Adaptation Index, CAI, Sharp and Li 1987*), ефективни број кодона (енг. *Effective Number of Codons,* N_C*, Wright 1990*), фреквенција оптималних кодона (енг. *Frequency of Optimal Codons,* F_{OP}*, Ikemura 1981*), индекс пристрасности кодона (енг. *Codon Bias Index,CBI, Bennetzen and Hall 1982)* и друге.

У биоинформатичким истраживањима један од највећих проблема представља количина података која је везана за предмет истраживања. Количина произведених и сакупљених података онемогућава или знатно отежава примену класичних метода за анализу података и издвајање информација и знања које је скривено у тим подацима. Нека од таквих истраживања која су предмет предложене теме су и утврђивање утицаја промене коришћења кодона и мутација на функционалне карактеристике протеина. За добијање поузданих резултата неопходна је примена метода истраживања података - математички заснованих рачунарских метода којима се одређују модели који су по карактеристикама најближи карактеристикама посматраних података, при чему може да се оцени квалитет добијених резултата.

Методе истраживања података примењиве на одређивање склоности употребе кодона и њихову примену на карактеризацију протеина, откривање потенцијалних мутација (замене или уметања/брисања кодона) и испитивање сличности карактеристика вируса обухватају одређивање правила придруживања, класификацију, кластеровање, истраживање образаца

кодона, итд. Претходно наведене мере могу да се користе као једна од карактеристика при одређивању параметара модела.

3. Научни циљ докторске дисертације

Основни циљ истраживања је утврђивање утицаја коришћења кодона на класификацију протеинских (аминокиселинских) ниски, или делова ниски, на модел-систему појединачних вируса и/или група вируса, као и склоност вируса или њихових компоненти (гена) ка одређеним кодонима. Вируси су инфективни агенси који се репликују само унутар живих ћелија сва три суперкраљевства (Архее, Бактерије, Еукарие). Пошто не поседују све критеријуме којима се одликује живот, вируси се дефинишу као репликатори, или (квази)/организми, на граници живог-неживог. У раду ће бити анализирано више фамилија вируса, при чему ће посебна пажња бити поклоњена вирусима из фамилије коронавируса. Биће анализиране постојеће методе за карактеризацију употребе кодона.

Досадашњи резултати су добијени анализом употребе кодона код појединачних вируса или скупова вируса исте фамилије. При томе, досадашња истраживања нису обухватала испитивање употребе кодона у зависности од припадности протеинима који имају исту или сличну функцију код сродних фамилија организама. Такође, у раду ће бити испитивано постојање мутација у сличним протеинима вируса и утицај тих мутација на евентуалну промену карактеристика протеина или организма.

Очекује се да добијени рачунарски модели допринесу бољем разумевању карактеристика вируса, предвиђању сродних региона у вирусима, као и бољем резумевању утицаја мутација на нуклеотидном нивоу на промену карактеристика самог вируса.

4. Хипотезе од којих се полази

Хипотезе од којих се полази у овом раду су:

- Употреба кодона је специфична за појединачни протеин у оквиру вируса. Такође, употреба кодона за исте протеине/протеине са сличном функцијом у различитим или истим фамилијама вируса може да се разликује у зависности од (под)врсте вируса. Ова различитост може да се односи и на регионе протеина, односно на његове подсеквенце.
- На основу употребе кодона може да се формира карактеристични профил секвенце који може да се представи вектором вредности сваког кодона. Могуће је формирати више различитих вектора, у зависности од мере која се користи за изражавање употребе кодона.
- Добијени профили (вектори вредности) могу да се искористе као материјал за прављење класификационих модела помоћу којих би се одредила припадност (а самим тим и оквирна функционалност) улазне (непознате) ниске.
- Кластер анализом материјала може да се одреди хоризонтална и вертикална блискост појединих вируса или њихових делова.

- У случају мутација, промењени генетски материјал задржава употребу кодона из оригиналног вируса.

5. Научни методи који ће бити коришћени у раду

Основне научне методе које ће бити коришћене у истраживању засноване су на моделима и алгоритмима у области истраживања података.

За одабране скупове вируса из различитих фамилија, користећи различите мере, биће одређени вектори добијених вредности за употребу кодона код ниски које кодирају протеине. Добијени вектори биће коришћени као материјал за формирање модела применом метода истраживања података (класификације, кластеровања и правила придруживања). Модели ће бити засновани на одређивању правилности појављивања континуалних или дископтинуалних ниски кодона у зависности од карактеристика протеина/подсеквенци које се обрађују.

Развој модела ће бити заснован на употреби кодона у скупу од преко 750.000 вируса, различитих фамилија.

Квалитет добијених резултата биће приказан у зависности од примењених математичких модела и мера који се користе у дефинисаним моделима. На основу резултата поређења ће бити процењени доприноси новодефинисаног приступа.

6. Закључак и очекивани резултати

Област која је предмет ове докторске дисертације је релевантна и актуелна и њоме се тренутно бави велики број истраживача како у академском окружењу, тако и у великим истраживачким лабораторијама. Као последица пандемије, анализа карактеристика коронавируса је у последње две године област на којој је радио велики број истраживача у различитим областима – медицини, биологији, биоинформатици, и фармацији. Предложено истраживање је засновано на новом и до сада ретко примењиваном приступу који комбинује употребу кодона и методе истраживања података.

У претходном периоду, кандидат је самостално, или у сарадњи са ментором и другим колегама учествовао у истраживањима која су посвећена одређивању повезаности употребе кодона и различитих протеинских структура. Очекује се да ће даља обрада предложене теме довести до значајних резултата који су потпуно оригинални и представљају новину у овој области.

7. Списак литературе која ће се користити

Tan PN, Steinbach M, Karpatne A, Kumar V.: *Introduction to Data Mining*. 2nd ed. Pearson Education, 2019.

Aggarwal C. (eds.): *Data Classification Algorithms and Applications*. Chapman and Hall/CRC, 2014.

Chen JY, Lonardi S (eds.): *Biological Data Mining*. Chapman and Hall/CRC, 2017.

Abu-Jamous B, Fa R, Nandi AK: *Integrative Cluster Analysis in Bioinformatics*. John Wiley & Sons, Ltd, 2015.

Giovanni S, Elder JF: *Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions*. Morgan & Claypool, 2010.

Xu R, Wunsch DC: *Clustering*. IEEE Press, 2009.

Aggarwal C.: *Machine Learning for Text*. Springer International Publishing, 2018.

NCBI вирусне секвенце, https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/

Dmitrij Frishman and Manja Marz, (Eds.), *Virus Bioinformatics*, Chapman and Hall/CRC, 2021, ISBN: 978-1-003-09767-9

Ikemura, T., (1981). *Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli system*. Journal of Molecular Biology 151: 389-409.

Jane Flint, Vincent R. Racaniello, Glenn F. Rall, Theodora Hatziioannou, Anna Marie Skalka *Principles of Virology*, ASM Press, 2020, ISBN: 978-1-683-67358-3

Malik YS, Ansari MI, Kattoor JJ, Kaushik R, Sircar S, Subbaiyan A, Tiwari R, Dhama K, Ghosh S, Tomar S, Zhang KYJ. *Evolutionary and codon usage preference insights into spike glycoprotein of SARS-CoV-2*. Briefings in Bioinformatics. 2021;22(2):1006–1022. doi: 10.1093/bib/bbaa383.

Sharp PM, Li W-H (1987) *The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications*. Nucleic Acids Res 15(3):1281–1295

Sun X., Yang Q., Xia X. *An improved implementation of effective number of codons (nc)*. Mol Biol Evol. 2013;30(1):191-6. doi: 10.1093/molbev/mss201.

Bennetzen, J. L., and B. D. Hall, (1982). *Codon selection in yeast*. Journal of Biological Chemistry 257: 3026-3031.

Wright F. *The "effective number of codons" used in a gene*. Gene. 1990;87:23–29. doi: 10.1016/0378-1119(90)90491-9.

Подаци о кандидату

Биографски подаци

Биљана Стојановић рођена је 6. септембра 1982. године у Лесковцу. 2001. године била је матурант Гимназије у Лесковцу на природно-математичком смеру као носилац Вукове дипломе. Школске 2001/2002. године уписала је основне академске студије на смеру Рачунарство и информатика студијског програма Математика на Математичком факултету Универзитета у Београду и дипломирала октобра 2007. године са просечном оценом 9,61.

Докторске академске студије студијског програма Информатика на Математичком факултету Универзитета у Београду уписала је први пут школске 2009/2010. године. Након краће паузе у студирању, поново је уписала докторске студије 2017. године, Положила је све предмете предвиђене програмом студија са просечном оценом 10,00.

Од октобра 2007. године била је запослена на Математичком факултету Универзитета у Београду у звању сарадника у настави за ужу научну област Рачунарство и информатика. Године 2009. изабрана је у звање асистента за исту научну област и радила је у том звању до 2020. године. Држала је наставу из следећих предмета: Програмирање 1, Програмирање 2, Увод у организацију рачунара, Објектно-оријентисано програмирање, Увод у организацију и архитектуру рачунара 1, Трансакциона обрада, Дистрибуиране и објектне базе података.

Од 2020. године запослена је на Математичком институту Српске академије наука и уметности (МИСАНУ) у звању истраживача приправника.

У претходном периоду учествовала је као истраживач на научним пројектима Министарства просвете, науке и технолошког развоја Републике Србије: Математички модели и методе оптимизације са применама, бр.144007 (2009-2010) и на потпројекту пројекта бр.ИИИ-044006 {Примене ИКТ у дигитализацији и научно-историјско-архивском третману културног наслеђа} (2012-2019). Коаутор је 9 техничких решења везаних за развој и унапређење техничких решења Дигиталних легата (http://legati.matf.bg.ac.rs/) и Виртуелне библиотеке Математичког факултета Универзитета у Београду (http://elibrary.matf.bg.ac.rs/). Члан је техничког уредништва националног часописа „Преглед националног центра за дигитализацију". Била је члан организационих одбора више међународних и националних научних скупова.

Област научног интересовања је истраживање података са применама у биоинформатици.

Библиографија

Радови у националним часописима

1. Maljković M., Stojanović B., Mijajlović Ž.: *Digital Legacy of Mihailo Petrović Alas.* Review of the National Center for Digitization, Faculty of Mathematics, University of Belgrade, 31, pp. 10 – 17, 2017.

2. Maljković M., Stojanović B., Mijajlović Ž.: *Digital Legacy of professor Slaviša Prešić.* Review of the National Center for Digitization, Faculty of Mathematics, University of Belgrade, 30, pp. 1 – 6, 2017.

3. Stojanović B.: *Digitizing the Archive of Student Services of the Faculty of Mathematics*. Review of the National Center for Digitization, Faculty of Mathematics, University of Belgrade, 29, pp. 33 – 41, 2016.

Саопштења са међународних скупова

1. Mitić N., Pavlović-Lažetić G., Beljanski M., Malkov S., Maljković M., Stojanović B., Veljković A., Kapunac S.: *Large scale clustering in structural and evolutionary analysis of SARS-CoV-2 proteins*. Bioinformatics of Genome Regulation and Structure/Systems Biology (BGRS/SB-2022), The 13th International Multiconference, 04–08 July, 2022, Novosibirsk, Russia, pp. 888-889, DOI 10.18699/SBB-2021-000, 2022.
2. Stojanović B., Veljković A., Malkov S., Beljanski M., Pavlović-Lažetić G., Mitić N.: *Codon usage polymorphism in SARS-2 CoV protein coding sequences*. Biologia Serbica, Book of Abstracts, Belgrade BioInformatics Conference 2021, Department of Biology and Ecology, Faculty of Sciences, University of Novi Sad, Serbia, 43(1), pp. 97 – 97, 2021.
3. Veljković A., Stojanović B., Malkov S., Beljanski M., Pavlović-Lažetić G., Mitić N.: *Codon Usage-based SARS-CoV-2 protein classification*. Biologia Serbica, Book of Abstracts, Belgrade BioInformatics Conference 2021, Department of Biology and Ecology, Faculty of Sciences, University of Novi Sad, Serbia, 43(1), pp. 92 – 92, 2021.

Саопштења са скупова националног значаја

1. Mitić N., Malkov S., Stojanović B., Maljković M.: *Nacionalni centar za digitalizaciju - prošlost, sadašnjost i budućnost*. Book of Abstracts: XVII National Conference Digitization of Cultural Heritage, Old Records from the Natural and Social Sciences and Digital Humanities, Belgrade, Serbia, pp. 6 – 6, 2019.
2. Mitić N., Malkov S., Stojanović B., Maljković M.: *Digitalizacija sabranih dela Nedeljka Parezanovića*. Book of Abstracts: XVI National Conference Digitization of Cultural Heritage, Old Records from the Natural and Social Sciences and Digital Humanities, Belgrade, Serbia, pp. 5 – 5, 2018.
3. Maljković M., Stojanović B., Mijajlović, Ž.: *Digitalni Legat Mihaila Petrovića Alasa*. Book of Abstracts: XV National Conference Digitization of Cultural Heritage, Old Records from the Natural and Social Sciences and Digital Humanities, Belgrade, Serbia, pp. 21 – 21, 2017.
4. Stojanović B., Mijajlović Ž., Maljković M.: *Digitalni Legat profesora Slaviše Prešića*. Book of Abstracts: XIV National Conference, Digitization of Cultural Heritage, Old Records from the Natural and Social Sciences and Digital Humanities, Belgrade, Serbia, pp. 5 – 5, 2016.
5. Mitić N., Malkov S., Vukša N., Stojanović B.: *Digitalizacija arhive studentske službe Matematičkog fakulteta u Beogradu*, XIII Conference Digitization of Cultural Heritage and Digital Humanities, Belgrade, Serbia, pp. 7 – 8, 2015.

Научно-наставном већу Математичког факултета Универзитета у Београду

И З Ј А В А

Овим путем изјављујем под материјалном и кривичном одговорношћу да тему докторске дисертације „*Примена метода истраживања података на употребу кодона ради класификације протеинских секвенци и одређивања међусобне сличности вируса* ” нисам пријавила ни на једном другом факултету, као ни на било ком другом универзитету у Републици Србији и иностранству.

У Београду, 09.12.2022. године

Биљана Стојановић

Образац 2.

ПРИЈАВА

ТЕМЕ ДОКТОРСКЕ ДИСЕРТАЦИЈЕ

Име (име родитеља) и презиме Биљана (Тихомир) Стојановић

Студијски програм Информатика

Школска година уписа на студијски програм 2017

Број индекса 2005/2017

Претходно образовање кандидата (основне и мастер студије): Дипломирани математичар, смер Рачунарство и информатика, Математички факултет, Универзитет у Београду

Радни наслов теме докторске дисертације Примена метода истраживања података на употребу кодона ради класификације протеинских секвенци и одређивања међусобне сличности вируса

Научне области које обухвата тема докторске дисертације Рачунарство и информатика, ужа научна област Биоинформатика

Контакти (телефон, мобилни телефон, e-mail): телефон: 0693523500, email: bstojanovic@turing.mi.sanu.ac.rs

Прилози:

- Образложење теме (научна област из које је тема, предмет научног истраживања, основне хипотезе, циљ истраживања и очекиване резултате, методе истраживања и списак стручне литературе која ће се користити)
- Биографија кандидата
- Библиографија кандидата
- Изјава да предложену тему кандидат није пријављивао на другој високошколској установи у земљи или иностранству
- Мишљење одговарајућих етичких комитета о етичким аспектима истраживања, уколико је предвиђено посебним прописима.

Подносилац пријаве

Биљана Стојановић