Наставно-научном већу
Математичког факултета
Универзитета у Београду

На 312. седници Наставно-научног већа Математичког факултета Универзитета у Београду одржаној 14.02.2014. године, одређени смо за чланове комисије за преглед и оцену рукописа **Истраживање података на протеинским нискама: н-грамска анализа уређених и неуређених региона протеина** који је предат као докторска дисертација кандидата Самира Алсхафах. Након прегледа рукописа подносимо Наставно-научном већу следећи

Извештај

1 Биографија кандидата

Самира Алсхафах је рођена 29. децембра 1978. године у Завии, Либија. Основну и средњу школу завршила је у Харши (Либија), а основне студије из области рачунарства на Факултету техничких наука, на оделењу за електротехнику, на Универзитету у Завии (Либија). Мастер студије из области рачунарства је завршила 2007. године на Либијској академији за последипломске студије Триполију (Либија).

Била је запослена од 2002. до 2004. године као наставник из области рачунарства у средњој школи у Харши (Либија), а од 2004. године је запослена на Факултету техничких наука, на оделењу за електротехнику, на Универзитету у Завии, прво као предавач, а после завршетка мастер студија као наставник. Држала је наставу из Јава програмирања на Институту за високо образовање у области рачунарских наука у Енјели (Либија). Област интерсовања су јој препознавање руком писаног текста на Арапском језику, Биоинформатика, и Истраживање података. Има два објављена рада (један самосталан и један коауторски на СЦИ листи). Учествовала је на две међународне конференције.

2 Предмет и садржај дисертације

Предмет докторске дисертације припада области примене истраживања података у биоинформатици. Истраживање података је једна од области рачунарства која се најбрже развијала у последње две деценије. Методе истраживања података укључују велики број различитих (најчешће математички заснованих) алгоритама помоћу којих се врши провера података и одређује модел који је по карактеристикама најближи карактеристикама података који се посматрају, при чему може да се оцени квалитет добијених резултата. Истраживање података обухвата широк скуп метода које укључују класификацију, кластеровање, одређивање правила придруживања, истраживање образаца, итд. Развој рачунарства омогућио је и рађање и развој нове дисциплине - биоинформатике. Рачунарска обрада података омогућила је велике продоре у биоинформатици са применом у различитим областима: медицини, фармакологији, пољопривреди, итд.

Тема рада повезује истраживање података и биоинформатику и везана је за конструкцију модела за карактеризацију неуређених и уређених региона протеина. Познато је да неуређени региони протеина имају велику улогу различитим (често везаним за болести) процесима у ћелији. Такође је познато да садржај неуређености протеина има утицаја на његову функцију. Сам процес одређивања да ли у протеину постоје неуређени региони је скуп и временски захтеван, и до сада је само за око 800 протеина експериментално утврђена позиција неуређених региона. Због тога је јако значајан развој нових рачунарских модела и метода који скраћују време обраде и повећавају прецизност одређивања позиције неуређених региона у протеинима. У дисертацији је представљена нова метода помоћу које могу да се приближно одреде позиције неуређених региона у протеинима. Метода је заснована на одређивању карактеристичних н-грама за сваки тип региона - н-грама који се јављају искључиво или највећим делом у регионима одређеног типа. Дефинисана метода не претендује да одреди тачне границе региона у протеину, већ омогућује њихово локализовање и знатно скраћење времена потребног за друге врсте анализа.

3 Кратак приказ дисертације и оригиналних доприноса

Рукопис се састоји од 118 страница (VI+112) и има следећу структуру:

1. Увод
2. Методе за одређивање карактеристичних ниски у регионима протеина
3. Материјал
4. Резултати
5. Закључак

уз Резиме (на енглеском и српском језику), Садржај, Додатак, Списак литературе и Биографију кандидата.

У уводном поглављу је дат приказ основних појмова и проблема који је обрађен у дисертацији.

Друго поглавље садржи приказ метода за одређивање карактеристичних ниски у регионима протеина: н-грамска анализа, анализа помоћу понављајућих ниски, молских фракција, фракцијских разлика, и z-вредности. У овом поглављу су описане и технике истраживања података које су коришћене у дисертацији: класификација и откривање правила придруживања. Такође, дат је и приказ програма за предвиђање неуређених региона и укратко наведене њихове карактеристике. На крају овог поглавља дефинисан је модел који ће бити коришћен за одређивање карактеристичних ниски за уређене/неуређене регионе протеина.

У трећем поглављу је дат преглед скупа података који је коришћен као улазни материјал. Истраживање је вршено на аминокиселинским и нуклеотидним секвенцама преко 190000 протеина који припадају скупу од 4076 вируса. Као материјал за проверу модела за аминокиселинске секвенце узет садржај ДисПрот базе података експериментално утврђених неуређених региона протеина. У овом поглављу је дат и приказ одређивања граничних вредности за појаву н-грама и понављајућих секвенци и њихов утицај на резултате.

У четвртом поглављу *Резултати* које представља централни део рада, су приказани и дискутовани резултати примене дефинисаних модела. Показано је да је улазни скуп података репрезентативан, тј. да без обзира што материјал садржи само вирусне протеине, његов садржај не одступа у великом проценту од садржаја скупа података из ДисПрот базе (који је врло хетероген по саставу) и која представља јединu узорак експериментално потврђених неуређених региона. Изложени су резултати добијени применом сваког појединачног метода, а затим и резултати добијени узимањем пресека тих резултујућих скупова.

Н-грами који се налазе у пресеку скупова добијених помоћу модела заснованих на молским фракцијама и фракцијским разликама имају минималну прецизност карактеризације од 85%, при чему прецизност расте са повећањем дужине н-грама. Слични проценти се јављају и за резултате других метода уз уочену правилност да краћи н-грам (дужине до 5) боље карактеришу уређене, а дужи боље неуређене регионе. Ако се посматра свака метода појединачно, н-грами добијени методом заснованом на правилима придруживања имају највећу прецизност. Прецизност карактеризације (упоређивањем са резултатима истог поступка спроведеног на ДисПрот бази) се повећава до 95% комбиновањем резултата добијених појединачним методама. При одређивању карактеристичних н-грама нађени су и карактеристични обрасци који се јављају у њима. Пронађено је да уређене регионе карактеришу обрасци у понављајућим секвенцама амино-киселина VV, FF, WW, YY, LLL, CC и II, хоморипити и тандем рипити амино-киселина које преферирају уређене регионе. Слично, неуређене регионе карактеришу палиндроми, хоморипити и тандем рипити амино-киселина које преферирају неуређене регионе, као и комбинације хоморипита и амино-киселина које преферирају неуређене регионе. У поглављу су наведени неки од н-грама и образаца, док се детаљнији списак откривених н-грама налази у табелама у Додатку.

Идентичан поступак је спроведен и за нуклеотидне секвенце протеина из материјала, при чему резултати нису могли да буду верификовани на подацима из ДиспПрот базе јер у њој не постоје подаци за нуклеотидне секвенце.

У петом поглављу *Закључак* је дат сумарни приказ садржаја дисертације и наведени су н-грами и обрасци који имају највећу поузданост у карактеризацији уређених/неуређених региона протеина.

У Додатку су приказани табела амино киселина, списак метода за предвиђање неуређених региона протеина, дистрибуција протеина који су коришћени у раду по фамилијама и класама вируса, и 22 табеле са списковима карактеристичних н-грама за регионе протеина. Н-грами у табелама су класификовани по коришћеним методама, типу региона који карактеришу, дужини, и поузадности.

Списак литературе се састоји од 33 библиографске јединице.

4 Радови

Резултате повезане са темом изложеним у овом рукопису кандидат је публиковала у два рада (један на СЦИ листи и један самосталан).

1. Samira Almokhtar Alshafah (2013). *Bioinformatics analysis of Archaeal viruses' genomes*, Journal of Advanced Computer Science and Technology Research, Vol.3 No.4, December 2013, 173-182

2. A. Jelović, N. Mitić, S. Eshafah, M. Beljanski: Finding statistically significant repeats in

nucleic acids and proteins, Journal of Computational Biology, DOI: 10.1089/cmb.2017.0046, (SCI - IF 1.032)

5 Закључак

Рукопис **Истраживање података на протеинским нискама: н-грамска анализа уређених и неуређених региона протеина** садржи вредан научни допринос у области истраживања података и његове примене у биоинформатици. У раду је разматран проблем карактеризације уређених и неуређених региона протеина помоћу аминокиселинских и нуклеотидних н-грамских ниски. У току рада су дефинисани нови модели од којих су најзначајнији модел заснован на комбинацији фракцијских разлика и z-вредности, модел заснован на правилима придруживања, као и комбинација ова два модела. Резултати теста модела на експериментално утврђеним неуређеним/уређеним регионима протеина показали су врло висок степен прецизности дефинисаног модела, који расте са повећањем дужине н-грама. Као резултат рада наведене су н-грамске ниске и обрасци који карактеришу уређене и неуређене регионе протеина, као и прелазе између њих. У раду је дефинисан и класификациони модел за карактеризацију региона, који даје нешто слабије резултате.

Имајући у виду претходно наведено предлажемо Наставно-научном већу Математичког факултета да рукопис **Истраживање података на протеинским нискама: н-грамска анализа уређених и неуређених региона протеина** кандидата Самире Алсхафах прихвати као докторску дисертацију и одреди комисију за њену одбрану.

У Београду, 28.05.2018.

Чланови комисије за преглед и оцену

(проф. др Ненад Митић, ванр. проф.)

(проф. др Саша Малков, ванр. проф.)

(др Милош Бељански, научни саветник)