Наставно-научном већу
Математичког факултета
Универзитета у Београду

На 310. седници Наставно-научног већа Математичког факултета Универзитета у Београду одржаној 14.10.2013. године, одређени смо за чланове комисије за преглед и оцену рукописа **Истраживање образаца у одређивању карактеристика протеина** који је предат као докторска дисертација кандидата Улфете Маровац. Након прегледа рукописа подносимо Наставно-научном већу следећи

Извештај

1 Биографија кандидата

Улфета Маровац рођена је 1980. године у Чачку. Основну школу и гимназију завршила је у Новом Пазару. Школске 1998/1999. године уписала је студије на Математичком факултету у Београду (смер Рачунарство и информатика). Дипломирала је школске 2003/2004. године са просечном оценом 9,24. Школске 2007/2008. уписала је докторске студије на Математичком факултету, смер Рачунарство и информатика.

Од септембара 2004. до септембра 2007. године радила је као наставник математике у Гимназији у Новом Пазару. Од септембра 2007. године запослена је на Државном универзитету у Новом Пазару као сарадник у настави. Године 2009. изабрана је у звање асистента. До сада је држала вежбе из следећих предмета: Основи информатике, Програмирање I, Програмирање II, Објектно програмирање, и Информатика. У претходном периоду је учествовала на пројектима технолошког развоја ТР-13012 и ТР-11002. Тренутно је истраживач на пројекту Нове информационе технологије за аналитичко одлучивање базиране на организацији експеримента и њихова примена у биолошким, економским, и социолошким системима, бр.ИИИ-44007, који финансира Министарство просвете, науке и технолошког развоја Републике Србије.

2 Предмет и садржај дисертације

Предмет докторске дисертације припада области истраживања података. 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, методе које раде са подржавајућим векторима, итд.

Тема рада повезује истраживање података и биоинформатику и везана је за конструкцију модела за одређивање карактеристика протеина помоћу методе истраживања образаца. Експериментално одређивање карактеристика протеина је сувише скупо, дуготрајно и

често немогуће у реалном времену због великог броја протеина чије се карактеристике одређују. Због тога је јако значајан развој нових рачунарских модела који скраћују време обраде и повећавају прецизност одређивања карактеристика. Конструисани модел одређује ЦОГ карактеристике протеина које омогућују њихову класификацију у групе ортологих протеина према њиховим функцијама. Познавање карактеристика протеина је важно за боље разумевање организације и функције биолошких система са крајњом применом у фармацији и медицини.

3 Кратак приказ дисертације и оригиналних доприноса

Рукопис се састоји од 107 страница (XII+95) и има следећу структуру:

1. Увод
2. Модели и методе за одређивање карактеристика протеина
3. Модел за одређивање карактеристичних n-грама за ЦОГ-ове протеина
4. Тестирање и примена модела
5. Закључак

уз Резиме (на српском и енглеском језику), Садржај, Додатке (додатак А, Б, В, Г, Д), Списак литературе и Биографију кандидата.

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

Друго поглавље садржи приказ постојећих метода за одређивање ЦОГ категорије протеина, као и приказ метода истраживања података које су коришћене у дисертацији.

У трећем поглављу, које заједно са четвртим поглављем представља централни део рада, дефинисан је модел за одређивање карактеристичних n-грама за појединачне ЦОГ категорије. У првом кораку у конструкцији модела улазни (протеински) n-грами се трансформишу у ОАК (основне амино киселинске) ниске. ОАК ниска n-грама садржи уређени низ слова која су елементи скупа слова садржаних у полазном n-граму. С обзиром да се у ОАК ниски свако слово из скупа јавља тачно једном, без обзира на број његовог појављивања у оригиналном n-граму, трансформација представља специфичан облик димензионе редукције улазних података. Пресликавањем се скуп различитих преклапајућих аминокиселинских n-грама који се се јављају у протеину замењује скупом одговарајућих ОАК ниски чиме се значајно смањује димензија вектора који описује састав n-грама у протеину. Добијене ОАК ниске се сматрају карактеристичним за појединачну ЦОГ категорију уколико су њихова подршка, поверење и бројач асполутне подршке већи од унапред дефинисаних вредности. Секвенцијални обрасци који карактеришу само једну појединачну ЦОГ категорију се називају дескриптори. Проблем одређивања ниски које карактеришу појединачну ЦОГ категорију протеина своди се на проблем одређивања скупа дескриптора те категорије. У том циљу је над добијеним скупом ОАК ниски дефинисана Булова алгебра и показано је да се одређивање ниске која представља секвенцијални образац карактеристичан за одређену ЦОГ категорију своди на решавање генерализованог система Булових једначина. У раду је описан и начин решавања одговарајућег система једначина. На крају овог поглавља су дефинисани алгоритми за издвајање скупа секвенцијалних образаца придружених одговарајућој ЦОГ категорији протеина, изградњу модела за класификацију протеина по функционалним категоријама ЦОГ-ова, и реализацију предложеног модела у облику програма за предвиђање ЦОГ категорије протеина. До сада познате методе класификације протеина по функционалним категоријама су вршиле поређење сваког новог протеина (коме треба одредити функцију)

са скупом свих протеина који су већ класификовани према функцијама ради одређивања групе која садржи протеине који су најсличнији протеину који се класификује. Предност нове методе у односу на претходне је што не врши секвенца-секвенца поређење већ се у протеину траже обрасци (n-грами) који су карактеристични за одговарајућу ЦОГ категорију, чиме се добија на уштеди меморијског простора и времена обраде протеина. Уз то, с обзиром на математичку основу која се налази у позадини одређивања дескриптор ниски гарантује се тачност добијених резултата у зависности од скупа улазних података над којима се конструише модел.

У четвртом поглављу *Тестирање и примена модела* приказани су резултати провере предложеног модела као и оцена квалитета добијеног модела за различите скупове материјала. За проверу је коришћен предиктор развијен на основу алгоритма описаног у претходном поглављу. Приказане су и дискутоване карактеристике модела у зависности од улазних података и утицај различитих параметара (на пример, дужине n-грама) на квалитет издвојених секвенцијалних образаца. На крају поглавља је дат резултат предвиђања ЦОГ категорије за групу протеина за коју није успешно извршена класификација постојећим методама.

У петом поглављу *Закључак* је дат сумарни приказ садржаја дисертације и описани могући правци у даљем раду.

У додатку А су наведене карактеристике скупа организама који су коришћени у тренинг и тест фази при имплементацији предложеног модела. Додатак Б садржи приказ карактеристика класификационог модела прављеног за појединачне класе бактерија, док Додатак В садржи приказ карактеристика класификационог модела прављеног групно за комплетан улазни материјал. У додатку Г су приказани резултати предиктора примењеног на скуп до сада некласификованих протеина, и у додатку Д су наведене скраћенице коришћене у раду.

Списак литературе се састоји од 59 библиографских јединица.

4 Радови

Резултате приказане у овом рукопису кандидат је публиковала у четири рада (радови су објављени или прихваћени за објављивање) од којих су два самостална и два коауторска. Сва четири рада су публикована у часописима на SCI листи:

1. Banković Dragić, Marovac Ulfeta: *System of two Boolean inequations*, Journal of Multiple Valued Logic and Soft Computing 24:5-6 (2015), pp. 521-528

2. Marovac Ulfeta: *System of k Boolean inequations*, Journal of Multiple Valued Logic and Soft Computing, volume 25(5), D401-MVLSC, 2015, прихваћен за објављивање (http://www.oldcitypublishing.com/journals/mvlsc-home/mvlsc-forthcoming-issuesaccepted-papers/)

3. Marovac Ulfeta, Mitić Nenad: *N-gram analysis of COG categorized protein sequences*, MATCH: Communications in Mathematical and in Computer Chemistry, прихваћен за објављивање

4. Marovac Ulfeta: *Disjunction of Boolean equations*, Publications de'l Institut Mathematique, Beograd, прихваћен за објављивање

5 Закључак

Рукопис **Истраживање образаца у одређивању карактеристика протеина** садржи вредан научни допринос у области истраживања података и његове примене у биоинформатици. У раду је разматран проблем класификације протеина према припадајућој ЦОГ категорији помоћу методе истраживања n-грамски образаца. У току рада је добијено више нових резултата који су укључени у нови класификациони модел. Резултати теста модела су показали да његове карактеристике у појединим случајевима надмашују моделе који се тренутно користе за решавање овог проблема.

Имајући у виду претходно наведено предлажемо Наставно-научном већу Математичког факултета да рукопис **Истраживање образаца у одређивању карактеристика протеина** кандидата Улфете Маровац прихвати као докторску дисертацију и одреди комисију за њену одбрану.

У Београду, 01.06.2015.

Чланови комисије за преглед и оцену

_____

(проф. др Ненад Митић, ванр. проф.)

_____

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

_____

(др Мирјана Павловић, виши научни сарадник)