Р Е Ц Е Н З И Ј А

НА ДОКТОРСКАТА ДИСЕРТАЦИЈА СО НАСЛОВ „АЛГОРИТМИ ЗА ПОРАМНУВАЊЕ И ПРЕБАРУВАЊЕ НА ДНК СЕКВЕНЦИ“ ИЗРАБОТЕНА ОД М-Р ДОНЕ СТОЈАНОВ, ФАКУЛТЕТ ЗА ИНФОРМАТИКА, УНИВЕРЗИТЕТ „ГОЦЕ ДЕЛЧЕВ“ ВО ШТИП

Со Одлука број 0206-1006/4 од 30 септември 2015 година, донесена од Наставно-научниот совет на докторски студии на Кампус 2 (биотехнички, техничко-технолошки и природно-математички науки) при Универзитет „Гоце Делчев” во Штип, формирана е Комисија за оценка и одбрана на докторската дисертација со наслов „Алгоритми за порамнување и пребарување на ДНК секвенци“ пријавена и изработена од кандидатот м-р Доне Стојанов, асистент на Факултет за информатика во Штип, во состав:

– вон. проф. д-р Александра Милева – претседател;
– вон. проф. д-р Сашо Коцески, член;
– доц. д-р Зоран Утковски, член;
– проф. д-р Ана Мадевска-Богданова, член, екстерен ментор и
– проф. д-р Цвета Мартиновска Банде, член, ментор.

Комисијата во наведениот состав, по прегледувањето на докторската дисертацијата, го поднесува следниов

И З В Е Ш Т А Ј

Докторската дисертација со наслов **„Алгоритми за порамнување и пребарување на ДНК секвенци“** од м-р Доне Стојанов е напишана на 152 страници, со вкупно 69 слики/графикони, 16 табели, 100 цитирани трудови и ги содржи следниве поглавја: *Поими од биоинформатиката*, *Истражувања во областа*, *Опис на предложените алгоритми*, *Софтверска библиотека*, *Експериментални резултати*, *Дискусија*, *Заклучок*, *Додаток – листа на алгоритми и Референци*.

Во првото поглавје со наслов *Поими од биоинформатиката* кандидатот дава краток осврт на основните концепти од молекуларна биологија кои се поврзани со процесите што се предмет на докторската дисертација. Се разгледува структурата на ДНК која содржи информација за развој и функционирање на живите суштества, улогите на различни видови РНК (рибозомска, транспортна и гласник РНК), функцијата на структурните гени како протеин кодирачки фрагменти, кодоните како триплети од нуклеотиди кои кодираат аминокиселини и протеините како линеарни вериги од аминокиселини. Во истото поглавје кандидатот го истакнува и значењето на порамнувањето на секвенци и пребарувањето на генетска база на податоци како основни пристапи за анализа на генетски податоци.

Поглавјето *Истражувања во областа* дава преглед и детален опис на најзначајните алгоритми за порамнување и пребарување на ДНК секвенци. Во почетокот на поглавјето е направена класификација на алгоритмите за порамнување на ДНК секвенци и тоа, според бројот на секвенци кои се порамнуваат, опсегот на порамнување, применетиот математички приод и оптималноста на решението. Кандидатот во оваа поглавје детално ги опишува особеностите на алгоритмите базирани на динамичко програмирање, како: Needleman-Wunsch, Sellers, Wagner-Fischer, Waterman-Eggert, Mayer-Miller, Ficket, Ukkonen и др. Овие алгоритми имаат нелинеарна временска комплексност, пропорционална на производот од должините на ДНК секвенците кои се порамнуваат, така што методот на динамичко програмирање не е применлив за порамнување на долги секвенци како хромозоми и целосни геноми или за пребарување врз база на секвенци. За таа цел се применуваат хевристички алгоритми кои не гарантираат наоѓање на оптимално решение како: FASTP, FASTA и BLAST. Во оваа група на алгоритми спаѓаат и Pattern Hunter, BLAT, FLASH, YASS алгоритмите за порамнување базирани на стебло на суфикси MUMmer и AVID, GLASS и LALIGN. Опишани се и алгоритамот за пребарување базиран на споредба на

ДНК фрагменти (DALIGN), алгоритамот за суперпарово порамнување кој е временски линеарен алгоритам за парово порамнување на хомологни секвенци, пристапи базирани на компресија на индекси и др. На крај од поглавјето се опишани современите хеш базирани пристапи за индексирање и пребарување на ДНК база на податоци: методот SSAHA и методот на Renker и Shyu.

Во третото поглавје *Опис на предложените алгоритми* кандидатот ги образложува концептите на предложените алгоритми за празнинско и беспразнинско парово порамнување на ДНК секвенци и предложениот алгоритам за временски и мемориски ефикасно индексирање и пребарување на ДНК база на податоци по ДНК прашалник. Во оваа поглавје покрај тезите за пресметковна подобреност на предложените решенија во однос на постоечките алгоритми, кандидатот ги разгледува и тезите за предностите на предложените алгоритми од биолошки аспект, како на пример:

- *поточна детекција на непрекинати мутации од вид додавања/бришења на триплети во фрагменти на гени кои се во релација со заболувања кои се последица на прекумерно повторување на специфичен кодон во однос на алгоритмот на Smith и Waterman;*
- *можност за детекција на блиска и далечна хомологија, независно од изборот на метриката на порамнување;*
- *зголемена точност на пребарување на индексираната ДНК база на податоци во однос на SSAHA и алгоритмот на Reneker и Shyu.*

Тезите изнесени во описот на алгоритмите се елаборирани со примери, проследени со математички издржана анализа.

Поглавјето *Софтверска библиотека* претставува имплементација на предложените алгоритми во хибридна околина C#/C++. Апликацијата има можност за преземање, филтрирање и прикажување на ДНК секвенци. ДНК секвенците со кои се извршуваат програмите се преземаат од Европската нуклеотидна архива, во FASTA податочен облик и локално се зачувуваат во текстуални датотеки. При примена на алгоритмот за брзо порамнување корисникот ја избира метриката на порамнување, внесувајќи награда за совпаднатите нуклеотиди и казна за различните нуклеотиди. Покрај структурата на оптималното порамнување се печати и процентот на сличност на секвенците, временската комплексност изразена во број на споредби и резервираната меморија во тек на извршување на алгоритмот.

При примена на алгоритмот за празнинско порамнување апликацијата прво го утврдува множеството на конзитентни совпаѓања со споредба на зборови со еднаква должина се до првото несовпаѓање. На тој начин се минимизира бројот на извршени базни споредби. Се прикажуваат структурите на порамнувањето со примена на моделот за додавање празнини по совпаѓање и со примена на моделот за додавање празнини меѓу несовпаѓачки парови со најмала фреквенција на појавување.

Во програмата за индексирање корисникот ја задава должината на индексирање k, ги внесува идентификаторите на ДНК секвенците и структурата на ДНК прашалникот. За индексирање постои можност да се избере формулата на Reneker и Shyu за единечно базно пресликување и предложената формула за брзо пресликување на преклопувачки зборови од ДНК секвенца. Предмет на споредбена анализа се и времињата на индексирање и пребарување кај SSAHA и алгоритмот на Renker и Shyu и предложениот алгоритам, како и бројот на пронајдени совпаѓања на ДНК прашалникот во индексираната ДНК база на податоци со примена на предложениот алгоритам и алгоритмот на Reneker и Shyu.

Во петтото поглавје *Експериментални резултати* се дадени резултатите од тестирање на мемориската и временската комплексност на имплементираните алгоритми во однос на сродни алгоритми. Тестовите се извршени со секвенци преземени од Европската нуклеотидна архива.

Во поглавјето *Дискусија* е дадена табела која ги сумира карактеристиките на предложените алгоритми и нивните предности во однос на сродни програми.

На крајот од докторската дисертација е даден заклучок во кој концизно се претставени карактеристиките на предложените алгоритми, како и подобрувањата во однос на сродни

алгоритми. Како прилог на докторската дисертација се дадени псевдокодовите на сите новопредложени алгоритми.

Главниот *научен придонес* на докторската дисертација се состои во три нови приоди за порамнување и пребарување на ДНК секвенци презентирани преку три алгоритми.

Предложениот алгоритам за брза идентификација на оптимална беспразнинска хомологија меѓу пар на ДНК секвенци работи на принцип на поместување на пократка секвенца долж подолга секвенца, така што во првата фаза од извршувањето решението се бара во рамки на преклопувачки прозорци со должина еднаква на должината на пократката секвенца, по што во зависност од вредноста на локалниот максимум за резултат на порамнување се ограничуваат левите и десни поместувања на пократката секвенца долж подолгата со кои се образуваат преклопувачки прозорци со должина помала од должината на пократката секвенца. Воведен е пристап за мемориска репрезентација на секое совпаѓање со податочна тројка која ги бележи почетните положби на совпаѓање во рамки на секвенците и должината на совпаѓање. Предложениот алгоритам не е само временски поефикасен од алгоритмот на Smith и Waterman, туку е и мемориски поефикасен во споредба со просторно линеарните алгоритми за парово порамнување на ДНК секвенци како: Hirschberg, Myers и Miller, Huang et al. и др.

Во дисертацијата се презентира нов пристап за порамнување на ДНК секвенци со додавање на празнини, кој е побрз од стандардните пристапи базирани на динамичко програмирање. Алгоритмот се базира на рекурзивна идентификација на ДНК фрагменти на совпаѓање, кои се порамнуваат со додавање на празнини, во зависност од нивната положба во ДНК секвенците. Тезите за подобрувањата со предложениот алгоритам се потврдени со експериментални резултати добиени со примена на софтверска библиотека изработена од кандидатот.

Алгоритмот за празнинско порамнување е тестиран на парови албумин протеин-кодирачки секвенци од Европската нуклеотидна архива. Се споредува бројот на базни совпаѓања меѓу предложениот алгоритам и алгоритмот на Smith и Waterman и просечниот број на совпаѓања при порамнувања со примена на хевристичките алгоритми.

Предложениот алгоритам генерира порамнувања со поголем број на совпаѓања во однос на алгоритмот на Smith и Waterman или хевристичките алгоритми MUMmer, Avid, Glass, Lalign. Совпаѓањата кои ги отфрла алгоритмот на Smith и Waterman би го намалиле резултатот на порамнување, додека совпаѓањата кои ги отфрлаат евристичките алгоритми се пократки совпаѓања на поголема оддалеченост од регионите на ДНК со најголема густина на совпаѓање. Предложениот алгоритам не отфрла кое било совпаѓање со што овозможува откривање како на блиска, така и на далечна хомологија на ДНК секвенци.

Пристапот базиран на додавање на празнини помеѓу неусогласени базни парови со минимална честота на појавување е клучната особина која го прави предложениот алгоритам применлив за поточна детекција на мутации во гени кои содржат непрекинати тринуклеотидни повторувања (Глутамин, кај фрагменти од хомо сапиенс хантингтин ген) во однос на алгоритмот на Smith и Waterman.

Основна идеја кај алгоритамот за индексирање и пребарување на ДНК база на податоци е користење на сортиран речник како индексна структура наместо хеш табела. Со употреба на сортиран речник се овозможува идентификација на сите совпаѓања на ДНК прашалник во база на ДНК секвенци без да се обработат сите записи, на што во принцип се должи намалувањето на времето на пребарување во споредба со алгоритмот на Renker и Shyu. Со примена на пристапот за додавање само за зборови кои постојат во базата се намалува мемориската зафатнина на индексираната структура во споредба со алгоритмот SSAHA.

Главното подобрување на предложениот алгоритам од биолошки аспект е можноста за детекција на сите совпаѓања, без оглед на нивната почетна положба во секвенците од базата на податоци.

Со цел да се зголеми апликативниот придонес на докторската дисертација, кандидатот ги има имплементирано сите предложени алгоритми во хибридна C#/C++ околина и истите се достапни за истражувачите од оваа област.

Докторската дисертација е напишана на начин кој е разбирлив и сите делови се конзистентни во рамките на целината. Референтната литература е соодветна и правилно цитирана, а материјата е изнесена јасно, читливо и со факти.

Резултатите од истражувањата на м-р Доне Стојанов се публикувани во следниве научни списанија со фактор на влијание:

- ***Stojanov Done***, Sašo Koceski, Aleksandra Mileva, Nataša Koceska, and Cveta Martinovska Bande. "*Towards computational improvement of DNA database indexing and short DNA query searching*". Biotechnology & Biotechnological Equipment. ISSN: 1310-2818 (Print), 1314-3530 (онлајн). Волумен: 28, број 5, 2014 год, страници: 958-967. (**Impact Factor=0,3**);
- ***Stojanov Done***, Saso Koceski, and Aleksandra Mileva. "FLAG: Fast Local Alignment Generating Methodology" Romanian Biotechnological Letters. ISSN: 1224-5984. Волумен: 18, број 1, 2013 год, страници: 7881-7888. (**Impact Factor=0,4**);

што претставува дополнителна потврда за валидноста и издржаноста на тезите презентирани во докторската дисертација.

ЗАКЛУЧОК И ПРЕДЛОГ

Комисијата за оценка и одбрана на докторската дисертација **„Алгоритми за порамнување и пребарување на ДНК секвенци",** пријавена и поднесена од кандидатот м-р Доне Стојанов, асистент на Факултетот за информатика, го разгледа поднесениот текст и донесе заклучок дека истиот претставува ***оригинален, самостоен и издржан*** научен труд, во кој се сумирани резултатите од долгогодишната научноистражувачка работа на кандидатот м-р Доне Стојанов, која резултираше со публикација на два научни труда со научни списанија со фактор на влијание: (*Biotechnology & Biotechnological Equipment*, **i.f=0,3**) и (*Romanian Biotechnological Letters*, **i.f=0,4**) и два труда од областа на истражување кои се објавени во меѓународните списанија: *Advanced studies in biology* и *Annals of West university of Timisoara, ser. Biology*.

Врз основа на тоа, Комисијата има чест да **му предложи на Наставно-научниот совет на докторски студии на Кампус 2 да ја прифати позитивната рецензија на докторската дисертација** со наслов „**Алгоритми за порамнување и пребарување на ДНК секвенци**", изработена од кандидатот м-р Доне Стојанов и да одобри јавна одбрана на истата.

КОМИСИЈА ЗА ОЦЕНКА И ОДБРАНА НА ДОКТОРСКА ДИСЕРТАЦИЈА

Вон. проф. д-р Александра Милева - претседател, с.р.
Вон. проф. д-р Сашо Коцески - член, с.р.
Доц. д-р Зоран Утковски - член, с.р.
Проф. д-р Ана Мадевска-Богданова - член, екстерен ментор, с.р.
Проф. д-р Цвета Мартиновска-Банде - член, ментор, с.р.