Рецензия

**на дисертационния труд за придобиване на
образователна и научна степен „доктор”
по направление 4.6 „Информатика и компютърни науки”
научна специалност 01.01.12 „Информатика“
на Валерия Николаева Симеонова
на тема “Методи на soft-computing в изчислителната биология:
Асемблиране на данни от геномно секвениране“**

Рецензент: доц. Стефка Фиданова

Утвърдена съм със заповед No РД38-59/05.02.2014 на Ректора на СУ „Св. Климент Охридски“ проф. дин. Иван Илчев за член на Научно жури във връзка с процедура за придобиване на образователна и научна степен „доктор“ по специалност 01.01.12 „Информатика“, професионално направление 4.6 „Информатика и компютърни науки“ от Валерия Николаева Симеонова с дисертация на тема „Методи на soft-computing в изчислителната биология: Асемблиране на данни от геномно секвениране“, научни ръководители , проф д-р.Красен Стефанов, доц. д-р Димитър Василев.

Като член за Научното жури съм получила:

1. Дисертация за придобиване на образователна и научна степен „доктор“ по направление 4.6 „Информатика и компютърни науки“;
2. Автореферат;
3. Професионална биография;
4. Копия от публикациите на автора свързани с дисертацията.

При оценяването на дисертационния труд се взема под внимание изискванията на Закона за развитие на академичния състав в Република България (ЗРАСРБ), Правилника за неговото прилагане (ППЗ) и Правилника за условията и реда за придобиване на научни степени и за заемане на академични длъжности във ФМИ на СУ. Основните норми, които трябва да бъдат спазени са:

1. Съгласно чл. 6(3) от ЗРАСРБ „дисертационният труд трябва да съдържа научни или научно приложни резултати, които представляват оригинален принос в науката. Дисертационният труд трябва да показва, че кандидатът притежава задълбочени теоретични знания по съответната специалност и способности за самостоятелни научни изследвания“.
2. Според чл. 27(2) от ППЗ дисертационният труд трябва да се представи във вид и обем, съответстващи на специфичните изисквания на първичното звено. Дисертационният труд трябва да съдържа: заглавна страница; съдържание; увод; изложение; заключение – резюме на получените резултати; библиография.

Дисертацията се състои от Увод (глава първа), три глави, заключение (глава пета), библиография, списък на авторските публикации по дисертационния труд, списък на фигурите и списък на таблиците в дисертацията.

1. Актуалност на проблема и целесъобразност на поставените цели и задачи

Биоинформатика е област от науката, която съчетава в себе си нови технологии, като използва методи на софт-компютинга и статистиката, които предлагат различни възможности за оптимизации и решения. Това е област на изследвания, които генерират нови знания и нови теории. Паралелното секвениране на геномни данни решава задачата за асемблиране на геном, независимо дали за него съществува или не референтен такъв, който да се възприема за относително верен и точен.

Представеният за рецензиране дисертационен труд е насочен основно към методи за асемблиране на геномни данни. На страница 11 в Увода е формулирана целта на дисертационния труд:

Основната цел на настоящата работа е разработването на метод и алгоритъм за асемблиране на данни от паралелно секвениране, за които съществуват референтни секвенции.

За постигането на така поставените цели се решават следните по-специфични задачи:

- Дефиниране на задачата за анализ, с ясно определяне на ограниченията които се налагат от характера на данните
- Определяне на общата схема от проблеми, които следва да се решат, заедно с връзките и преходите между тях. Всеки проблем съставлява отделен алгоритъм, който може да се изпълни и самостоятелно върху подходящ тип данни:

2. Познаване състоянието на проблема от страна на дисертанта

Няма съмнение, че дисертантката е навлязла много добре в научната проблематика. Тематиката е сравнително нова. Списъкът на цитираните литературни източници е актуален – преобладават публикации от последните 10 години, но от друга страна в него присъстват и позовавания на по-стари източници. Общият брой цитирани източници е 212, като всички са на английски и са от чуждестранни автори.

Познаването на проблема от страна на дисертантката е много добре илюстрирано от глава 2 на дисертационния труд. Там е представен подробен литературен преглед на съществуващите методи и алгоритми за секвениране и асемблиране на геномни данни.

3. Методика на изследването

Методиката за провеждане на изследването, избрана от дисертантката, произтича от поставената цел и съответства на произтичащите от целта задачи. Тя е напълно адекватна на целта на работата. Авторът предлага метод за обработка на секвенираните данни, разглежда случаите на непълнота на информацията. Предложен е алгоритъм за асемблиране. Подходът е алтернатива на съществуващи приложения. Използван е нестандартен метод за графичен анализ на статистически данни. Постигната е оптимизация на функции за статистически анализ от средата R. Настоящата разработка позволява паралелизация на процесите за статистическо профилиране, както и цялостно асемблиране. Това дава възможност за намаляване на времето за изпълнение.

4. Характеристика и оценка на приносите на дисертационния труд

Дисертацията се състои от пет глави (включващи увод и заключение), списък на цитираните публикации, списък на авторските публикации по дисертацията и списък на научните форуми на които са докладвани резултатите включени към дисертацията. Приложени са таблици и цветни фигури илюстриращи разглежданите задачи и постигнатите резултати.

Глава първа представлява увод в предметната област на изследването. Там е дадена мотивация за избор на проблема и обекта на изследването. Направен е кратък преглед на използваните в дисертацията методи. Посочени са целите и задачите на дисертацията.

В Глава 2 е направен подробен преглед на съществуващите методи за секвениране и асемблиране на геномни данни. Както и методи за намаляване на грешките във входните данни.

Глава 3 разглежда различните формати данни от паралелно геномно секвениране. Там е предложен общ модел на алгоритъм за асемблиране. Даден е метод за откриване и филтриране на фоновия шум, както и метод за генериране на консенсусни секвенции. Разгледан е и случаят при припокриване на последователности. Използва се граф за представяне на задачата. Предложен е метод за оценка на пътя в графа, както и метод за запълване на интервалите в пътищата.

Глава 4 е посветена на приложението на разработените алгоритми и на валидацията им. Построява се мрежа от изкуствени реферативни „прочити – идентификатори“, която служи като гръбнак на графа, по който ще се извършват процедурите по съставянето на главния път и под-пътищата в графа. Алгоритъмът е реализиран в средата R.

Глава 5 представлява заключение. Там са дискутирани получените в дисертацията резултати и са дадени основните приноси на дисертацията. Предложени са и бъдещи насоки за развитие на тематиката.

Дисертантката претендира за следните приноси:

- Теоретически:

1. МИРПИ – нов подход при асемблиране на NGS данни и наличие на референтен геном
2. Използване на статистическо профилиране на NGS данните при корекция на грешките от секвенатора
3. Нов подход при избора на предиктори за конструиране на обучаващата таблица за невронната мрежа за отстраняване на грешките от секвениране
4. Нов подход за конструиране и оценка на пътищата в графа, като движението е паралелно и двупосочно

- Практически:

1. Оптимизация на съществуващи библиотеки в R Project, свързани със извеждане на статистики за данните, при първоначален тест статистики за 1000 символни низа с време за изпълнение около 3 минути, срещу постигнато време от 6.5 минути за 79990 символни низа.
2. Постигната оптимизация при сравняването на символни низове по двойки и извличане на данни за сравняването с време за изпълнение за 79990 сравнявания $\cong$ 37 минути, при първоначален тест за $\cong$ 150 минути.
3. Всички тествани алгоритмизации са разработени, така че да може да се изпълняват върху желана извадка от данните, както и върху цялото множество.
4. За първи път е приложен графичен анализ «роза на ветровете» върху NGS данни

5. Значимост на разработката за науката и практиката

Извършената от дисертантката работа е достатъчна по обем и задълбоченост на изследването. Показана е възможност за практическа реализация. В този смисъл намирам работата за значима както в научно, така и в практическо отношение.

6. Преценка на публикациите по дисертацията

Във връзка с дисертацията, кандидатката е представила 7 публикации, две от които са в списание с импакт фактор, а останалите са в сборници с доклади от специализирани международни конференции.

Дисертантката е участвала в поне 6 конференции и в един научен проект финансиран по програма „Млади учени“ на фонда за научни изследвания.

Няма данни за цитирания.

7. Лично участие на докторанта

Кандидатката не е представил справка за личното си участие в колективните публикации, но от дългогодишната и работа по темата оставам с впечатлението, че нейното участие е съществено.

8. Автореферат

Като цяло, авторефератът правилно отразява съдържанието на дисертацията.

9. Критични бележки

Имам някои критични бележки, които засягат техническото оформление на предоставената ми работа и биха били от полза на дисертантката в работата ѝ в бъдеще.

1. Няма поставени страници в списъка с публикации на дисертантката. В този списък има 15 заглавия, но част от тях изглежда са само докладвания, а не са публикации, за това съм зачела 7 публикации. Добре е докладванията да не са в този списък. Не е ясно какъв тип публикация е заглавие П14.
2. Добре би било ако в текста беше споменато кой нейн резултати в коя от нейните статии е публикуван. Никъде в текста не видях позовавания на нейни публикации.
3. Има доста правописни грешки включително неправилно членуване, несъгласувани части на речта, изпуснати предлози и незавършено звучащи изречения.
4. Добре е когато е възможно вместо чуждица да се използва българският им съответник като например:

- комплексност -> сложност
- протеин -> белтък
- интродуцират -> въвеждат
- тезис -> дисертация

5. На места се използва изразът „в следващата таблица са описани“ - добре е да се посочи номерът на таблицата за да е ясно за коя става дума. Например този израз е употребен на страница 14, а таблицата за която се отнася е на страница 51, 35 страници след това.
6. Таблица 1 е разкъсана, част от нея преминава на следващата страница.
7. Има 6 фигури без номерация и заглавие съответно на страници 7, 27, 60, 114, 118.
8. Фигура 31 се състои от 20 под фигури и е на 3 страници, по-добре щеше да е ако беше разделена на няколко фигури, които се побират на една страница.

9. Фигурите ту са номерирани отгоре, ту отдолу, то отстрани, хубаво би било ако номерацията е еднотипна в целя текст.

10. Обясненията на фигури 4,5 и 6 са дадени като заглавия отстрани на фигурите. По добре е след фигурата да има позоваване към нея с подробно описание какво е изобразено.

11. На страница 21 е употребен изразът „динамично оптимиране“ правилният израз е „динамично програмиране“, динамичното програмиране е техника за оптимизиране.

12. На места е използван изразът „най-оптималното“, такова понятие няма. Оптимално означава най-добро.

13. На страница 33 е употребен изразът „клас NP-сложност“ такъв клас няма класът е NP с подкласове NP-пълни и NP-трудни. Може би авторката е искала да каже „които са NP-сложни“, а не „наречен NP-сложност“

14. Глава 6 - искам да уточня, че еволюционните алгоритми и тези базирани на интелекта на рояка също са метаевристични. От текста се остава с впечатлението, че не са.

15. Често (37 пъти) в дисертацията се употребява съкращението „т.к.“ вместо „тъй като“. До колкото знам не е прието този израз да се съкращава в официални текстове, още повече не е добре изречение да започва със съкращение.

16. Добре е когато се номерират точките в една глава първо да стои номерът на главата и след него номерът на точката, за да няма объркване.

17. Има заглавия на точки и подточки които са последен ред на страница. По добре щеше да е ако бяха прехвърлени на следващата страница.

18. В заключението, точка 3.2 е употребен изразът „синергизми от типа Невронни мрежи – генетични алгоритми“, правилният израз е „хибридни методи от типа Невронни мрежи – генетични алгоритми“.

19. По-добре би било списъците с участия в проекти и участия в конференции да бъдат отделно, а не в един общ списък.

10. Лични впечатления

Не познавам дисертантката и нямам лични впечатления за нея.

Заключение

Като следствие на изложеното по-горе, може да се констатира, че са изпълнени всички изисквания на Закона за развитие на академичния състав (ЗРАСРБ). Правилника за неговото прилагане (ППЗ) и Правилника за условията и реда за придобиване на научни степени и заемане на академични длъжности във ФМИ на СУ.

Посочените от мен критични бележки засягат предимно техническото оформление на дисертацията и не намаляват научната ѝ стойност.

Всичко това ми дава основание за положителна оценка и предлагам на почитаемото Научно жури да присъди образователната и научна степен „доктор“ по специалност 4.6 „Информатика и компютърни науки“ на Валерия Николаева Симеонова.

10.04.2014 год.
гр. София

(доц. д-р. Стефка Фиданова)