Gliwice, 1 marca 2021

Dr hab. inż. Dariusz Mrozek, prof. PS
Katedra Informatyki Stosowanej
Politechnika Śląska w Gliwicach
ul. Akademicka 16
44-100 Gliwice

RECENZJA

rozprawy doktorskiej dla Rady Naukowej Dyscypliny Informatyka Techniczna i Telekomunikacja działającej w Politechnice Warszawskiej

Tytuł rozprawy: Szacowanie liczby powtórzeń fragmentu DNA

Autor rozprawy: mgr inż. Wiktor Kuśmirek

1. Jakie zagadnienie naukowe jest rozpatrzone w pracy (teza rozprawy) i czy zostało ono dostatecznie jasno sformułowane przez Autora? Jaki charakter ma rozprawa (teoretyczny, doświadczalny, inny)?

Przedstawiona przez Pana Wiktora Kuśmirka rozprawa doktorska składa się z cyklu powiązanych tematycznie publikacji wraz z towarzyszącym im autoreferatem, który stanowi przewodnik po zrealizowanych pracach badawczych i szereguje wiedzę w omawianym obszarze. W ogólnym ujęciu rozprawa jest poświęcona opracowaniu nowych algorytmów do analizy sekwencji genetycznych pozyskiwanych z metod sekwencjonowania drugiej i trzeciej generacji. Główne tezy rozprawy, a zostało ich sformułowanych aż sześć, koncentrują się wokół zagadnienia poprawy jakości i wydajności czasowej procesu wykrywania kopii tego samego fragmentu DNA (ang. *copy number variation*, CNV) w sekwencjach otrzymywanych przy pomocy metod pełnoeksomowych (WES) i pełnogenomowych (WGS), poprzez zastosowanie różnego rodzaju technik i modyfikacji w istniejących metodach należących do określonych klas metod. Zarówno tezy pracy, jak i motywacja prowadzonych badań w tym obszarze zostały sformułowane w sposób jasny i wyczerpujący. Charakter rozprawy określiłbym jako **teoretyczno-eksperymentalny**, ponieważ Autor:

- zaprojektował szereg metod i usprawnień dla istniejących algorytmów wykrywania kopii tych samych fragmentów DNA, dla różnych klas metod *Read Depth* i *Assembly Methods*,
- dla potwierdzenia słuszności przyjętych rozwiązań przeprowadził badania eksperymentalne na publicznie dostępnych zbiorach danych NCBI, które pozwoliły zweryfikować, iż opracowane algorytmy i rozwiązania w zakresie doboru próbek referencyjnych, obliczania głębokości pokrycia,

a także asemblacji *de novo* do szacowania liczby kopii i odtwarzania powtarzających się fragmentów DNA mogą być z powodzeniem stosowane w analizie sekwencji WES i WGS.

Świadczy to w mojej opinii na korzyść przedstawionej pracy.

2. Czy w rozprawie przeprowadzono w sposób właściwy analizę źródeł (w tym literatury światowej, stanu wiedzy i zastosowań w przemyśle) świadczący o dostatecznej wiedzy Autora. Czy wnioski z przeglądu źródeł sformułowano w sposób jasny i przekonywający?

Analiza światowej literatury i bieżącego stanu wiedzy w omawianym obszarze zostały przeprowadzone w sposób właściwy i świadczą o dostatecznej wiedzy Autora w tej dziedzinie. Ma ona charakter nieco rozproszony, ponieważ znajduje się ona zarówno w przedłożonym autoreferacie (rozdziały 1.2.1 i 1.2.2), jak i w większości spośród pięciu przedstawionych publikacji Autora, które tworzą cykl publikacyjny będący głównym osiągnięciem rozprawy. Zawartość rozdziałów 1.2.1 oraz 1.2.2 autoreferatu, które obejmują m.in. przegląd metod obliczania głębokości pokrycia w regionach sekwencjonowania, metod obejmujących proces doboru próbek referencyjnych, metod modelowania tła oraz klasyfikację algorytmów wykrywania kopii tego samego fragmenty DNA na podstawie danych WES i WGS, potwierdza, iż Autor posiada szeroką wiedzę w zakresie pierwotnych i bieżących trendów w zakresie tworzenia tego typu rozwiązań, a także zna ich zalety i słabości. W rozprawie zacytowano łącznie 103 pozycje literaturowe, z których zdecydowana większość dotyczy wyżej wymienionych elementów stanu wiedzy. Rozdziały te (1.2.1 oraz 1.2.2) oraz Rysunek 1 stanowią bardzo dobry wstęp teoretyczny do całości rozprawy, a do pojęć w nich zdefiniowanych (włączając wstęp do rozdziału 1.2) Autor nawiązuje w kolejnych podrozdziałach autoreferatu, jak również w treści poszczególnych artykułów głównego cyklu publikacyjnego. Szczególną uwagę zwraca Autor na problem dokładności oszacowania liczby kopii danego regionu DNA w genomie i w konsekwencji na możliwość odczytania i pełnego odtworzenia wielu genomów. Przeprowadzony przez Autora przegląd wiedzy w tym zakresie pozwolił mu w sposób jasny i przekonywujący sformułować wnioski, w tym m.in. określić problemy szacowania liczby powtórzeń fragmentów DNA pozyskiwanych różnymi technikami sekwencjonowania.

3. Czy Autor rozwiązał postawione zagadnienia, czy użył właściwej do tego metody i czy przyjęte założenia są uzasadnione?

Na początku realizacji rozprawy Pan Wiktor Kuśmirek zdefiniował kilka zadań, do których realizacji konsekwentnie dążył w swoich pracach badawczych. Dotyczyły one opracowania metod doboru próbek referencyjnych i zweryfikowania ich wpływu na możliwości wykrywania kopii tego samego fragmentu DNA (CNV), opracowania nowych metod składania fragmentów repetytywnych, zrównoleglenia metod wykrywania kopii tych samych fragmentów DNA (a przez to skrócenie ich czasu), a także zweryfikowania możliwości łączenia różnych technologii sekwencjonowania DNA i wpływu takiego podejścia (nazywanego podejściem hybrydowym) na jakość procesu asemblacji DNA. W swoich pracach Autor sięgnął do rozwiązań bazujących na głębokości pokrycia w regionach sekwencjonowania oraz rozwinął algorytmy odtwarzania *de novo* sekwencji DNA. Na podstawie lektury przedłożonych prac można stwierdzić, iż postawione w rozprawie zagadnienia zostały rozwiązane w sposób właściwy. Autor osiągnął to poprzez: 1) identyfikację słabości istniejących algorytmów analizy i składania sekwencji DNA pozyskiwanych technikami wielkoskalowymi, 2) opracowanie własnych usprawnień lub algorytmów, 3) badania

eksperymentalne weryfikujące przydatność opracowanych metod z użyciem publicznie dostępnych zbiorów danych. Wyniki przeprowadzonych przez Autora rozprawy badań potwierdziły, iż założenia przyjęte podczas opracowania autorskich metod były słuszne i uzasadnione. W artykułach stanowiących główne osiągnięcie rozprawy przedstawiono porównanie osiągniętych wyników z wynikami istniejących i popularnych narzędzi powszechnie używanych przez specjalistów prowadzących podobne analizy sekwencji DNA.

4. Na czym polega oryginalność rozprawy, co stanowi samodzielny i oryginalny dorobek Autora, jaka jest pozycja rozprawy w stosunku do stanu wiedzy czy poziomu techniki reprezentowanych przez literaturę światową?

Przedstawiona rozprawa stanowi bardzo dobre uzupełnienie bieżącego stanu wiedzy światowej w zakresie asemblacji DNA algorytmami klasy *de novo*, wydajnego obliczania głębokości pokrycia, oraz wykrywania kopii tych samych fragmentów DNA. Pan Wiktor Kuśmirek zaproponował nowatorskie podejścia w zakresie efektywnej realizacji tych procesów, a także przeprowadził proces ich wnikliwej oceny. Na rozprawę, oprócz autoreferatu, składa się pięć publikacji w renomowanych czasopismach z listy Journal Citation Report (JCR), które są ze sobą ściśle powiązane, w których udział Autora rozprawy jest często większościowy. Charakteryzując krótko zawartość przedłożonych prac P1-P5 można stwierdzić, iż:

[P1] W pracy tej w celu ustalenia skutecznej metody doboru próbek referencyjnych zbadano dwa podejścia do problemu doboru próbek (w którym wszystkie próbki traktowano jako zbiór referencyjny i poprzez wybór losowy), a także dwie metody grupowania (k-średnich i k najbliższych sąsiadów kNN ze zmienną liczbą klastrów lub ich wielkością). W badaniach wykorzystano dane z sekwencjonowania WES pozyskane z projektu 1000 Genomes do oceny wpływu różnych metod doboru zestawu próbek referencyjnych na wydajność wykrywania kopii CNV dla trzech wybranych najnowocześniejszych narzędzi: CODEX, CNVkit i exomeCopy. Przeprowadzone eksperymenty wykazały, że odpowiedni dobór zestawu próbek referencyjnych może znacznie poprawić współczynnik wykrywania kopii CNV. W pracy tej Pan Wiktor Kuśmirek przeprowadził m.in. wszystkie badania.

[P2] W pracy tej przedstawiono nowy, oparty o konstruowanie podgrafu de Bruijn'a, algorytm asemblacji DNA należący do klasy algorytmów *de novo*, który wykorzystuje względną częstotliwość odczytów do prawidłowego odtworzenia powtórzeń tandemowych. Główną zaletą przedstawionego algorytmu jest to, że jest on w stanie odtworzyć długie powtórzenia tandemowe, które są znacznie dłuższe niż maksymalna długość odczytów. Ponadto, algorytm potrafi przywrócić powtarzające się regiony DNA objęte tylko danymi z sekwencjonowania *single-read*, czego nie potrafią inne algorytmy tej klasy. Do oryginalnego wkładu Autora rozprawy należy przede wszystkim opracowanie algorytmu przedstawionego w pracy i przeprowadzenie badań z jego udziałem, w tym badań porównawczych w stosunku do istniejącego stanu wiedzy.

[P3] W pracy tej przedstawiono aplikację o nazwie dnaasm-link do łączenia kontigów, będącą wynikiem składania *de novo* danych z sekwencjonowania drugiej generacji, z długimi odczytami DNA. Przedstawiono algorytm wypełniania przerw fragmentem odpowiedniego długiego odczytu DNA

w celu poprawy spójności powstałych sekwencji DNA. Badania potwierdziły, iż opracowana aplikacja pozwala znacznie ograniczyć użycie pamięci operacyjnej i skraca czas obliczeń, a także wykazuje odpowiednią efektywność w porównaniu z innymi narzędziami przeznaczonymi do tego celu. Ponownie, do oryginalnego wkładu Autora rozprawy należy opracowanie algorytmu przedstawionego w pracy i przeprowadzenie badań z jego udziałem, w tym badań porównawczych w stosunku do istniejącego stanu wiedzy.

[P4] W pracy tej przedstawiono zaimplementowaną w środowisku Apache Spark platformę SeQuiLa, która umożliwia wydajne obliczanie głębokości pokrycia. Wydajność i skalowalność przedstawionego rozwiązania pozwala na prowadzenie obliczeń obejmujących całe egzomy i genomy, działając lokalnie lub na klastrze komputerowym. Wkład Autora rozprawy polegał m.in. na przeprowadzeniu testów i porównaniu wyników z bieżącymi rozwiązaniami.

[P5] W pracy tej przedstawiono hybrydowy algorytm asemblacji *de novo* genomu oparty na komplementarnych technologiach i metodach sekwencjonowania, m.in. Illumina paired-end, Illumina mate-pair oraz Oxford Nanopore Technology. Analizując rzeczywisty genom tasiemca szczurzego Autorom udało się udowodnić, iż dokładniejsze i dłuższe wynikowe sekwencje DNA pozwalają w lepszy sposób analizować powtarzalne regiony DNA. Wkład Autora rozprawy polegał m.in. na przeprowadzeniu badań i analiz towarzyszących wykonywanym pracom.

Podjęcie tych problemów oraz opracowanie dla niech odpowiednich podejść algorytmicznych, uważam za istotne osiągnięcie Autora i zaliczam do oryginalnych wyników przedstawionych w rozprawie. Udział procentowy oraz wkład Autora rozprawy zostały potwierdzone oświadczeniami podpisanymi przez współautorów publikacji. Wyniki przeprowadzonych prac badawczych zostały opublikowane w 5 artykułach w liczących się w dziedzinie informatyki (i bioinformatyki) czasopismach, m.in. *BMC Bioinformatics* (IF=2.217, 140 pkt. MNiSW), *Giga-Science* (IF=7.267, 200 pkt.), *Scientific Data* (IF=5.305, 140 pkt.) i *BioMed Research International* (IF=2.583, 70 pkt.). Dorobek ten uzupełnia 10 artykułów opublikowanych w materiałach konferencyjnych, 3 wystąpienia konferencyjne, 10 wystąpień plakatowych. Na uwagę zasługuje udział w licznych projektach o charakterze naukowym oraz nagrody (np. za najlepszy plakat na Sympozjum Polskiego Towarzystwa Bioinformatycznego). Świadczy to w mojej opinii o istotności podjętego problemu oraz wyraźnym wkładzie Pana Wiktora Kuśmirka w rozwój tego obszaru informatyki.

5. Czy Autor wykazał umiejętność poprawnego i przekonywującego przedstawiania uzyskanych przez siebie wyników /zwięzłość, jasność, poprawność redakcyjna rozprawy/?

Realizując pracę Pan Wiktor Kuśmirek wykazał dobre opanowanie umiejętności przedstawiania uzyskanych przez siebie wyników. Same idee zostały zaprezentowane w sposób dość jasny, sformalizowany i poparty przykładami, i co niezwykle istotne, poprzedzone szeroką analizą rozwiązań dotychczas zaprezentowanych na światowym forum naukowym. O ile forma prezentacji nie budzi większych zastrzeżeń, uważam, że w samym autoreferacie dobrze byłoby przesunąć pewien fragment teoretyczny lub umieścić ogólne wprowadzenie do tematyki zanim zdefiniuje się cele i tezy pracy. Pozwoliłoby to czytelnikowi lepiej odnaleźć się w tematyce rozprawy bez konieczności wracania do raz już

przeczytanych fragmentów, które w pierwszym czytaniu mogą brzmieć dość enigmatycznie. Uwaga ta dotyczy głównie rozdziału 1, w których Pan Wiktor Kuśmirek prezentuje m.in. cele i tezy, a następnie motywację do prowadzenia badań w obranym przez siebie obszarze. Oceny skuteczności rozwiązania dokonano w oparciu o publicznie dostępne dane z sekwencjonowania DNA (m.in. z bazy NCBI, 1000 Genomes). Wyniki oceny skuteczności opracowanych rozwiązań danej klasy zostały przeanalizowane i skomentowane w przedstawionych artykułach P1-P5 przedłożonej rozprawy pokazując, że poszczególne modyfikacje umożliwiają poprawę jakości osiąganych wyników w porównaniu z wybranymi i dostępnymi metodami. Od strony redakcyjnej zarówno autoreferat, jak i prace P1-P5 są w większości napisana w dobrym stylu i czyta się ją z łatwością, chociaż znalazłem w samym autoreferacie również kilka błędów, tzw. literówek.

6. Słabe strony rozprawy i jej główne wady?

Przedstawione prace są bardzo ciekawe i dotyczą istotnych problemów działania algorytmów analizy sekwencji DNA. Uzupełnia je autoreferat, który zawiera najważniejsze konkluzje wypływające z przeprowadzonych prac badawczych. Nie znalazłem w tych dokumentach istotnych uchybień, poza wspomnianym brakiem popularno-naukowego wstępu do tematyki rozprawy na początku samego autoreferatu. Uwaga ta nie ma charakteru znacząco krytycznego i nie umniejsza znaczeniu osiągnięć Autora rozprawy. Z punktu widzenia językowego zastanowiło mnie natomiast użycie w autoreferacie określenia „assemblingu DNA" zamiast pojęcia „asemblacji" – może stanie się to przyczynkiem do szerszej dyskusji, która mogłaby się wywiązać podczas obrony niniejszej rozprawy.

7. Jaka jest przydatność rozprawy dla nauk technicznych?

Uważam, że przedłożona rozprawa doktorska Pana Wiktora Kuśmirka wpisuje się w bieżące problemy bioinformatyki i genomiki funkcjonalnej. Opracowanie różnych algorytmów szacowania liczby kopii DNA pozwoliło Autorowi na poprawę jakości procesu asemblacji sekwencji DNA w stosunku do istniejących rozwiązań opublikowanych w światowej literaturze, co przekłada się bezpośrednio na tworzenie lepszych rozwiązań w tym obszarze. W ten sposób zaproponowane rozwiązania rozszerzają spektrum istniejących rozwiązań stosowanych w analizie danych sekwencyjnych pozyskiwanych technikami wielkoskalowymi. Potwierdzają to publikacje, których Pan Wiktor Kuśmirek jest autorem, opublikowane przez wiodące wydawnictwa, takie jak *Nature* oraz *Oxford*.

8. Do której z następujących kategorii Recenzent zalicza rozprawę:

a/ nie spełniająca wymagań stawianych rozprawom doktorskim przez obowiązujące przepisy

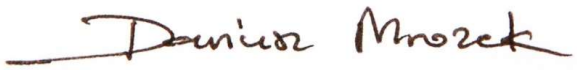
b/ wymagająca wprowadzenia poprawek i ponownego recenzowania

c/ spełniająca wymagania

d/ spełniająca wymagania z wyraźnym nadmiarem

e/ wybitnie dobra, zasługująca na wyróżnienie

Reasumując, bardzo dobre wyniki osiągnięte przez Pana Wiktora Kuśmirka w trakcie realizowanych przez niego badań pozwalają potwierdzić główne tezy rozprawy przedstawione w rozdziale 1.1 autoreferatu. Wyniki badań pokazują, że techniki oraz metody zaproponowane przez Pana Wiktora Kuśmirka mogą przyczynić się do znaczącej poprawy jakości asemblacji DNA i kompletności danych genetycznych poddawanych dalszej analizie. Wartość tych metod została dostrzeżona przez środowisko naukowe, co potwierdzają opublikowane prace, wchodzące w skład przedstawionego cyklu głównego. Uważam zatem, że **przedstawiona rozprawa** co najmniej **z wyraźnym nadmiarem spełnia wymagania** stawiane rozprawom doktorskim określone w obowiązujących przepisach, a nawet **jest wybitnie dobra i zasługuje na wyróżnienie**. Wnoszę o dopuszczenie Doktoranta do publicznej obrony.



Dr hab. inż. Dariusz Mrozek, prof. PS
Katedra Informatyki Stosowanej
Politechnika Śląska w Gliwicach