RADA NAUKOWA DYSCYPLINY
INFORMATYKA TECHNICZNA I TELEKOMUNIKACJA POLITECHNIKI WARSZAWSKIEJ

zaprasza na

PUBLICZNĄ OBRONĘ ROZPRAWY DOKTORSKIEJ
mgr. inż. Wiktora Kuśmirka

która odbędzie się w dniu 22 czerwca 2021 roku o godzinie $13^{\underline{00}}$ w trybie zdalnym na platformie MS Teams*.
Temat rozprawy doktorskiej:

„Szacowanie liczby powtórzeń fragmentu DNA"

Promotor: dr hab. inż. Robert Nowak - Politechnika Warszawska

Recenzenci: dr hab. Norbert Dojer - Uniwersytet Warszawski

dr hab. inż. Dariusz Mrozek – Politechnika Śląska w Gliwicach

dr hab. inż. Aleksandra Świercz - Politechnika Poznańska

* Obrona odbędzie się zdalnie na platformie MS Teams. Osoby zainteresowane uczestnictwem w obronie proszone są o zgłoszenie chęci uczestnictwa w formie elektronicznej na adres sekretarza komisji: p.wawrzynski@elka.pw.edu.pl w dniu obrony do godz. 12:00.

Z rozprawą doktorską i recenzjami można zapoznać się w Czytelni Biblioteki Głównej Politechniki Warszawskiej, Warszawa, Plac Politechniki 1.

Streszczenie rozprawy doktorskiej i recenzje są zamieszczone na stronie internetowej: https://www.bip.pw.edu.pl/Postepowania-w-sprawie-nadania-stopnia-naukowego/Doktoraty/Wszczete-po-30-kwietnia-2019-r/Rada-Naukowa-Dyscypliny-Informatyka-Techniczna-i-Telekomunikacja/mgr-inz.-Wiktor-Kusmirek.

Przewodniczący Rady Naukowej Dyscypliny Informatyka Techniczna i Telekomunikacja Politechniki Warszawskiej

dr hab. inż. Jarosław Arabas

Streszczenie
pracy doktorskiej Wiktora Kuśmirka
pt. „Szacowanie liczby powtórzeń fragmentu DNA”
złożonej w listopadzie 2020r. do obrony
na Politechnice Warszawskiej

Rozprawa przedstawia nowe algorytmy do analizy sekwencji genetycznych, m.in. do szacowania powtórzonych kopii DNA wykorzystując różne dostępne technologie sekwencjonowania. W pracy opisano możliwości poprawy wyników istniejących narzędzi poprzez właściwy dobór próbek referencyjnych oraz przedstawiono nowy algorytm do assemblingu *de novo* fragmentów powtarzających się motywów. Dodatkowo, rozprawa zawiera opis nowych metod łączenia odczytów z sekwenatorów drugiej i trzeciej generacji oraz opis nowej aplikacji do liczenia głębokości pokrycia. Praca zawiera również prezentacje zalety łączenia wyników assemblingu *de novo*.

POLITECHNIKA POZNAŃSKA

WYDZIAŁ INFORMATYKI I TELEKOMUNIKACJI
Instytut Informatyki
ul. Piotrowo 2, 60-965 Poznań, tel. +48 61 665 2997, fax +48 61 877 1525
e-mail: office_cs@put.poznan.pl, www.cs.put.poznan.pl

Instytut Informatyki

Poznań, 18.02.2021 r.

Dr hab. inż. Aleksandra Świercz
Instytut Informatyki
Politechnika Poznańska

RECENZJA ROZPRAWY DOKTORSKIEJ DLA RADY NAUKOWEJ DYSCYPLINY INFORMATYKA TECHNICZNA I TELEKOMUNIKACJA POLITECHNIKI WARSZAWSKIEJ

Tytuł rozprawy: Szacowanie liczby powtórzeń fragmentu DNA

Autor rozprawy: mgr inż. Wiktor Kuśmirek

Promotor: dr hab. inż. Robert Nowak, prof PW

Tematyka badawcza

Recenzowana rozprawa doktorska przedstawia wyniki badań dotyczące analizy sekwencji genetycznych, w szczególności sekwencji które zawierają powtarzające się fragmenty. Prace nad odczytaniem sekwencji całego genomu dla różnych organizmów trwają od lat. Jednym z najbardziej spektakularnych osiągnięć było odczytanie sekwencji genomu ludzkiego na początku dwudziestego pierwszego wieku. Proces ten zajął wówczas około 15 lat i pochłonął 3 mld dolarów. Dzięki rozwojowi technologii, znacznie przyspieszony został proces sekwencjonowania, czyli odczytywania sekwencji DNA, a także drastycznie zmniejszone zostały koszty, umożliwiając na sekwencjonowanie genomu człowieka poniżej 1000 dolarów i w czasie nie dłuższym niż kilka dni. Obecnie na rynku znajdują się dwie wiodące technologie: sekwencjonowanie nowej generacji (NGS), które pozwala uzyskać kilkusetnukleotydowe sekwencje o bardzo dobrej jakości, oraz sekwencjonowanie trzeciej generacji (TGS), dzięki której można uzyskać długie sekwencje (do kilkudziesięciu, a nawet kilkuset, tysięcy nukleotydów) o trochę gorszej jakości. Największym wyzwaniem odczytywania sekwencji genomów są powtarzające się fragmenty sekwencji, lub sekwencje o bardzo niskim stopniu skomplikowania, co znacznie utrudnia proces asemblacji, czyli składania sekwencji genomu z krótkich fragmentów pochodzących z

sekwencjonowania. Warto zwrócić uwagę, że pomimo iż sekwencja genomu ludzkiego znana jest od kilkunastu lat, to nadal zawiera nieodczytane fragmenty, które są efektem powtórzeń.

Kolejną trudnością jest fakt, iż genomy organizmów tego samego gatunku różnią się między sobą. Zmiany mogą być krótkie: jedno- lub kilku- nukleotydowe (SNP oraz INDEL) oraz mogą obejmować obszar wielu milionów nukleotydów. W obrębie największych zmian, tak zwanych wariantów strukturalnych (SV), możemy rozróżnić delecje, duplikacje, inwersje, insercje oraz translokacje. Najczęstsze różnice są najkrótsze i najłatwiejsze do wykrycia. Dla przykładu w organizmie ludzkim występuje około 3 mln SNP, czyli średnio 1 na 1000 nukleotydów jest inny niż w genomie referencyjnym. Znacznie rzadsze zmiany SV wykrywane są dzięki nowym technologiom sekwencjonowania. Wyróżniamy dwa główne podejścia: mapowanie do sekwencji genomu referencyjnego, oraz asemblacja *de novo* bez wykorzystania genomu referencyjnego. W pierwszym podejściu sprawdzane są sparowane odczyty (PEM), odczyty nie mapujące się w całości do genomu (SR), a także badana jest głębokość pokrycia (RD). To dzięki badaniu głębokości pokrycia, a także asemblacji *de novo* można wykryć zróżnicowanie liczby kopii fragmentów DNA, tzw CNV, które jest szczególnym przypadkiem SV obejmującym delecje oraz duplikacje. Badania w tym zakresie pokazały, że niektóre fragmenty genomu są bardziej podatne na CNV, oraz że mogą wpływać na ekspresję genów. Tematyka recenzowanej pracy doktorskiej wpisuje się w tematykę odczytywania genomu *de novo* oraz wykrywania powtarzających się fragmentów badanych sekwencji.

Zakres pracy i wkład autora

W pracy doktorskiej mgr W. Kuśmirek zajął się szacowaniem liczby kopii fragmentów sekwencji w genomie. Autor przetestował narzędzia do szacowania CNV na podstawie danych sekwencjonowania cało-eksomowego (WES). Sekwencjonowanie WES polega na wstępnym wyselekcjonowaniu fragmentów DNA obejmujących eksony, powieleniu tychże fragmentów DNA, a następnie ich sekwencjonowaniu. Pozwala to na pokrycie genomu tylko w wybranych obszarach, które tworzą niejako *okna*, w genomie. Zaproponowane narzędzie SeQuiLa-cov pozwala na znacznie szybsze obliczanie głębokości pokrycia w zdefiniowanych oknach niż istniejące narzędzia. Autor rozprawy przetestował różny dobór próbek referencyjnych, które będą wykorzystane do modelowania tła przy obliczaniu głębokości pokrycia. Opracował i zaimplementował algorytm do wyboru próbek referencyjnych, który można wykorzystać w aplikacjach nie posiadających tego etapu.

Drugi nurt badawczy doktoranta dotyczył rekonstrukcji *de novo* sekwencji DNA, wykorzystując szacowanie liczby kopii fragmentów DNA na etapie tworzenia grafów A-Bruijna. Zazwyczaj narzędzia do asemblacji *de novo* nie odtwarzają fragmentów repetytywnych, ze względu na rozgałęzienia w grafie, które są przez nie tworzone, gdyż w łatwy sposób można połączyć błędne sekwencje. Opracowane i zaimplementowane narzędzie dnaasm pozwala na uzyskanie kontigów również w przypadku powtarzających się sekwencji. Ponadto, doktorant zaimplementował narzędzie do łączenia kontigów uzyskanych przy użyciu asemblera dnaasm z długimi odczytami pochodzącymi z sekwencjonowania trzeciej generacji. Pozwoliło to uzyskać dłuższe kontigi o wciąż dobrej jakości. Oba te narzędzia zostały wykorzystane w praktyce podczas sekwencjonowania *de novo* genomu tasiemca.

Ocena strony merytorycznej

Rozprawa doktorska jest oparta na cyklu pięciu publikacji, które ukazały się w wysoko punktowanych czasopismach z listy JCR. Autoreferat z przedstawioną tematyką pracy doktorskiej, opisem publikacji i osiągnięciami doktoranta zawarty został na 55 stronach. Bibliografia składa się z 103 pozycji literaturowych.

Rozdział 1 wprowadza w tematykę zagadnienia, przedstawia aktualny stan wiedzy oraz prezentuje cel rozprawy i postawione hipotezy badawcze. Opisane zostały różne sposoby sekwencjonowania oraz różnice między sekwencjonowaniem całego genomu (WGS) oraz całego eksomu (WES). Przydałoby się umieścić trochę nowsze cytowania dotyczące stopy błędów w sekwencjonowaniu trzeciej generacji niż publikacja [24] z 2015 roku, gdyż technologia ONT aktualnie ma znacznie niższy odsetek błędów niż 40%. Podobnie dla PacBio poprawiona została jakość dzięki odczytom HiFi (High Fidelity). Długie odczyty o lepszej jakości pozwalają oczywiście na uzyskanie dłuższych kontigów o bardzo wysokiej jakości. W podrozdziale 1.2.1 przedstawione zostały dwa sposoby na obliczanie głębokości pokrycia na podstawie danych WES, wraz z dostępnymi narzędziami. Następnie omówione zostały sposoby na filtrowanie regionów o nienaturalnym pokryciu, które będą zaburzały dalsze analizy. W kolejnym etapie, normalizacji głębokości pokrycia, wykorzystywane są między innymi wybrane próbki jako tło modelujące. Autor rozprawy zauważył, że nie wszystkie programy wykorzystują pule próbek referencyjnych i jako jeden z celów pracy postawił sobie przetestowanie różnego sposobu doboru próbek referencyjnych oraz implementację algorytmu doboru próbek dla aplikacji nie posiadających tego etapu.

Podrozdział 1.2.2 przedstawia drugą część zagadnień poruszanych w pracy doktorskiej obejmujących szacowanie głębokości pokrycia w trakcie asemblacji *de novo*. Przedstawione różne podejścia do wykrywania wariantów strukturalnych (SV) zostały przez autora niesłusznie sklasyfikowane jako podejścia do wykrywania CNV. CNV jest podgrupą SV, a do wykrycia CNV najlepiej posłużyć może badanie głębokości pokrycia oraz asemblacja *de novo*. Pozostałe metody potrafią wykryć jedynie delecję, natomiast nie potrafią wykryć duplikacji. Niewłaściwy zapis pojawił się także na Rysunku 1 – zamiast CNV powinno pojawić się SV. W dalszej części rozdziału doktorant opisał podejścia do asemblacji *de novo* oparte na grafach OLC oraz grafach de Bruijna. W grafie można także zapisać informację z oszacowaniem ile razy należy użyć każdej krawędzi, dzięki czemu będzie można rozwiązać także problem powtarzających się fragmentów. W kolejnej części podrozdziału autor przedstawia różne sposoby na łączenie danych z sekwencjonowania długich i krótkich odczytów. Na koniec doktorant zaprezentował swój wkład w tym obszarze, czyli nowy algorytm asemblacji de novo z wykorzystaniem informacji szacowanej liczy kopii w celu odtworzenia fragmentu powtarzającego się oraz algorytm do asemblacji de novo z połączonymi sekwencjami krótkimi i długimi.

Cele pracy oraz hipotezy badawcze zostały sformułowane zarówno na początku pracy w rozdziale 1.1 oraz na koniec rozdziałów 1.2.1 i 1.2.2. W mojej opinii skrótowe przedstawienie zaraz na początku pracy może być dla czytelnika niezrozumiałe, gdyż używane są skróty wcześniej nie wyjaśnione. Bardziej właściwym byłoby umieszczenie rozdziału 1.1 po wprowadzeniu.

Tematem rozdziału 2 jest szacowanie liczby kopii na podstawie sekwencjonowania eksomu. W pierwszej części omówiona jest publikacja P1, w której testowano algorytm doboru próbek referencyjnych do 3 różnych aplikacji, które nie posiadały takiego kroku. Autor zwięźle i klarownie przedstawił najważniejsze wnioski wypływające z badań dotyczących różnych sposobów wyboru próbek referencyjnych. W drugiej części rozdziału omówiona została publikacja P4. Doktorant zajmował się przeprowadzeniem testów wydajnościowych aplikacji SeQuiLa-cov, wykazując że wykonanie obliczeń na wielu rdzeniach znacznie przyspiesza obliczanie głębokości pokrycia w porównaniu do innych aplikacji.

W rozdziale 3 omówione zostały trzy publikacje doktoranta P2, P3, P5 – związane z asemblacją *de novo* zarówno krótkich, jak i długich odczytów. Wnioski z każdej publikacji są przedstawione jasno i klarownie. W publikacjach P2 i P3 doktorant zaproponował nowatorskie algorytmy do asemblacji *de novo*, które uwzględniają informację o głębokości pokrycia i są w stanie odtwarzać sekwencje repetytywne. Pozostałe testowane asemblery (ABySS, Velvet oraz SPAdes) w znaczącej mierze nie były w stanie wykryć powtórzeń tandemowych (Tabele 4-6 w P2). Mam zastrzeżenie odnośnie szacowania zużycia pamięci i czasu działania. Autor testował swoją aplikację na małych genomach bakteryjnych, oraz danych symulowanych, gdzie rozmiar genomu nie przekraczał 1 miliona bp. Aplikacje (dnaasm i dnaasm-link) mogłyby nie zadziałać na danych pochodzących z genomu człowieka o rozmiarze 3 miliardów bp. Szacowany czas 8h nie jest zbyt wymagający, aby móc przetestować te aplikacje i sprawdzić ich działanie w praktyce na dużych genomach. W drugiej części rozdziału omówione została ostatnia z publikacji (brakuje odwołania na stronie 28 do P5) dotycząca projektu sekwencjonowania tasiemca szczurzego. Omówione zostały kolejne etapy przygotowania odczytów krótkich oraz długich. Niezrozumiałym jest dla mnie dlaczego odfiltrowane zostały odczyty poprawnie sparowane? Czy nie powinny być raczej odfiltrowane odczyty niepoprawnie sparowane? W wyniku przeprowadzonych eksperymentów obliczeniowych udało się poprawić spójność sekwencji genomu tasiemca. W ostatniej części rozdziału przedstawiony został sposób asemblacji *de novo* genomu mitochondrialnego tasiemca z wykorzystaniem krótkich odczytów. Pojawia się pytanie, skoro genom mitochondrialny jest taki krótki (13 kbp), to czy nie można było wykorzystać odczytów Nanopore? Jeden odczyt powinien pokryć cały genom mtDNA.

W ostatnim, czwartym rozdziale doktorant omówił dalsze plany badawcze, w tym automatyzację procesu doboru próbek referencyjnych, różny dobór podzbioru regionów sekwencjonowania oraz wykorzystanie metody mapowania optycznego.

Mimo wymienionych uwag oraz pytań bardzo pozytywnie oceniam recenzowaną pracę doktorską. Tematyka, choć omówiona została jako dwie odrębne części, połączona jest wspólnym zagadnieniem wykrywania różnej liczby kopii fragmentów DNA, z jednej strony jako obliczanie głębokości pokrycia w badaniu wariantów strukturalnych CNV, z drugiej jako szacowanie liczby powtarzających się fragmentów w asemblacji *de novo*. Doktorant ma rozeznanie w tematyce sposobów wykrywania sekwencji repetytywnych. Wykazał się zarówno wiedzą z zakresu informatyki, biegłością w implementacji algorytmów, skalowalności obliczeń, czy też

teorii grafów, a także wiedzą w zakresie biologii, znajomością najnowszych technologii sekwencjonowania, charakterystyką złożoności sekwencji DNA.

Ocena od strony redakcyjnej

Na recenzowaną pracę doktorską składa się zarówno cykl publikacji oraz 20-stronicowe streszczenie. Odnośnie publikacji – nie mam uwag. Streszczenie jest napisane starannie. Doktorant ma czasami tendencję do budowania bardzo długich zdań złożonych, które nie pasują do siebie i powinny być rozbite na kilka krótszych zdań. Podział pracy na rozdziały jest prawidłowo dobrany, dzięki czemu czytelnik w łatwy sposób orientuje się, co można znaleźć w której publikacji. Mgr Kuśmirek używa w pracy słów, które w dziwny sposób zostały przetłumaczone z języka angielskiego, lub nie zostały w ogóle przetłumaczone:

- Używany w pracy często 'assembling *de novo*' z angielskiego *de novo assembly* tłumaczy się na j. polski: 'asemblacja *de novo*'
- Str 13: Whole Exome Sequencing (WES) można tłumaczyć jako 'pełnoeksomowe' sekwencjonowanie, choć częściej spotykane tłumaczenie to całoeksomowe. W pracy pojawia się dodatkowo słowo 'pełnoeksonowe', co pewnie jest literówką
- Str 32: 'Optical mapping' nie zostało przetłumaczone – powinno być: mapy optyczne lub mapowanie optyczne

Autor nie ustrzegł się także przed nielicznymi błędami gramatycznymi, składniowymi i interpunkcyjnymi:

- Str. 10: „Zaprojektować i zbadać nową, wydajną implementację algorytmu" powinno być „Zaprojektować i **przetestować** nową, wydajną implementację algorytmu"
- Str. 10: „do łączenia wyników assemblingu *de novo* krótkich odczytów przez długie odczyty" powinno być „do łączenia wyników **asemblacji** *de novo* krótkich odczytów **za pomocą** długich odczytów"
- Str 11. „Każda z wymienionych hipotez została udowodniona, rozwiązania opublikowałem" zdanie zbyt długie, powinno być rozbite : „Każda z wymienionych hipotez została udowodniona. Rozwiązania opublikowałem…"
- Str 11. „Przykładowo, niektóre funkcje genów mogą być modulowane przez zmianę liczby powtórzeń DNA, **proces** ten umożliwia…" Zdanie długie, wymaga podzielenia na 2 zdania. Ponadto nie wiadomo o jaki proces chodzi.
- Str 13: „odpowiadają częścią kodującym" powinno być „odpowiadają częściom kodującym"
- Str 14. „Istotną rolę w poznaniu takich sekwencji mają algorytmy." Nie wiadomo o jakie algorytmy chodzi
- Str 14-15 „Narzędzia różnią się czasem działania, oprócz różnych algorytmów są implementowane przy pomocy różnych języków programowania, ponadto nie wszystkie aplikacje posiadają implementacje zrównoleglenia obliczeń." Zdanie, które spokojnie może zostać podzielone na 3 odrębne. Dodatkowo środkowa część zdania wymaga zmiany, aby była zrozumiała.

- Str. 16: „nową, rozproszoną implementacja procesu" powinno być „nową rozproszoną implementację procesu"
- Str. 26: „zwiększa prawdopodobieństwo odtworzenie sekwencji" powinno być „zwiększa prawdopodobieństwo odtworzenia sekwencji"
- Str. 26: „Następnie jest budowany graf połączeń w którym wierzchołkami są kontigi a krawędziami" powinno być „„Następnie jest budowany graf połączeń, w którym wierzchołkami są kontigi a krawędziami"

Szereg błędów edycyjnych pojawiło się także w cytowaniach literaturowych, wynikających najprawdopodobniej z formatowania Latex'a. są to błędy typu: duże/małe litery.

- W [3],[5] i [9] DNA pojawia się w tytule małymi literami: 'dna'
- W [95] CNV w tytule pojawia się małymi literami: 'cnv'
- Czasopisma typu PLOS pisze się pierwszy człon dużymi literami, drugi zaczyna z dużej litery: PLOS Genetics [1], PLOS ONE [13,36], PLOS Computational Biology [42, 58]
- Czasopismo Genome Research pisze się z dużych liter [32,43,48,49,55,62,69]
- Podobnie Nucleic Acid Research [22,28,35,53]
- Czasopismo BMC Bioinformatics – drugi człon z dużej litery [47,72,86,93,100]
- Również niepoprawna pisownia czasopism: Genome Biology, Genome Research, Nature Genetics, GigaScience, Journal of Molecular, Biology, Molecular Cell
- Tytuły artykułów powinny zaczynać się z dużej litery, a kolejne wyrazy z małych liter [63,67,70,78,82,85]

Powyższe usterki nie mają znacznego wpływu na jakość i czytelność pracy i nie umniejszają jej wartości. Nie zmieniają również mojej pozytywnej oceny recenzowanego doktoratu.

Wnioski końcowe

W moim przekonaniu autor recenzowanej pracy doktorskiej wykonał, szereg skomplikowanych badań. Zaproponował, zaimplementował i przetestował algorytmy, które w znaczny sposób mogą przyspieszyć obliczenia, a także poprawić jakość uzyskiwanych sekwencji wynikowych zawierających powtórzenia. Wykazał się wiedzą z zakresu matematyki, informatyki, biologii oraz bioinformatyki, którą potrafił zastosować w praktyce. Recenzowana praca zawiera oryginalne rozwiązania problemu powtórzeń w sekwencjach DNA. Wyniki badań zostały opublikowane w czterech wysoko punktowanych czasopismach z dziedziny (*BMC Bioinformatics (2x), GigaScience, Scientific Data, BioMed Research International*) dając sumaryczny współczynnik IF = 19,581. W trzech z pięciu publikacji był pierwszym autorem z przeważającym wkładem pracy. Ponadto opublikował 10 artykułów w materiałach konferencyjnych (*International Society for Optics and Photonics*), występował na konferencjach jako referent lub z prezentacją plakatową. Na podkreślenie zasługuje również fakt, iż doktorant uzyskał w 2020 roku grant Preludium przyznany przez Narodowe Centrum Nauki,

oraz grant CYBERIADA-1 finansowany przez Politechnikę Warszawską. Uczestniczył również w 4 grantach badawczych oraz różnych kursach o głębokiej analizie danych biomedycznych.

Stwierdzam, że praca pana mgr inż. Wiktora Kuśmirka pt. „Szacowanie liczby powtórzeń fragmentu DNA" spełnia wymagania stawiane rozprawom doktorskim określone w Ustawie o stopniach naukowych i tytule naukowym oraz stanowi oryginalne rozwiązanie problemu naukowego. Wnoszę o dopuszczenie mgr inż. Wiktora Kuśmirka do dalszych etapów przewodu doktorskiego.
Mając na uwadze bogaty dorobek publikacyjny zgromadzony w ciągu 4 lat pracy nad doktoratem oraz fakt iż zostały opublikowane w wiodących czasopismach naukowych, składam wniosek o wyróżnienie tej pracy doktorskiej.

Dr hab. inż. Aleksandra Świercz

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

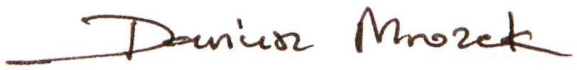
b/ wymagająca wprowadzenia poprawek i ponownego recenzowania

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

UNIWERSYTET
WARSZAWSKI

Wydział Matematyki, Informatyki i Mechaniki
Instytut Informatyki
dr hab. Norbert Dojer

Warszawa, 20.03.2021.

Recenzja rozprawy doktorskiej *mgr inż. Wiktora Kuśmirka*
zatytułowanej
Szacowanie liczby powtórzeń fragmentu DNA

Strona formalna rozprawy

Rozprawa doktorska mgr inż. Wiktora Kusmirka ma charakter cyklu pięciu artykułów. Wszystkie artykuły zostały opublikowane w renomowanych czasopismach naukowych (70-200 punktów ministerialnych, impact factor 2.213-7.267): dwa w BMC Bioinformatics oraz po jednym w GigaScience, Scientific Data i BioMed Research International. Wszystkie artykuły są wieloautorskie, w każdym przypadku wkład doktoranta jest starannie opisany i potwierdzony oświadczeniami współautorów. W publikacjach [P1], [P2] i [P3] wkład doktoranta został oceniony na 70-80% i obejmował m.in. zaprojektowanie badań/opracowanie algorytmów, wykonanie eksperymentów oraz przygotowanie tekstu manuskryptów i przeprowadzenie procesu publikacji, uzasadnione jest więc uznanie mgr inż. Wiktora Kuśmirka za głównego autora tych prac. W artykule [P4] wkład doktoranta został oceniony na 10%, natomiast w pracy [P5] na 25%. Należy jednak zaznaczyć, że ta ostatnia publikacja jest efektem interdyscyplinarnej współpracy kilku grup badawczych, a wkład doktoranta w obliczeniową część projektu można uznać za wiodący. Dołączony autoreferat w przejrzysty sposób prezentuje główne wyniki wymienionych artykułów i wyjaśnia ich znaczenie.

Tematyka badań

Tematyka rozprawy skupiona jest wokół dwóch zagadnień:

- wyznaczanie występującej w genomie liczby kopii fragmentów DNA z danych

ul. Banacha 2, 02-097 Warszawa, Polska
+48 22 55 44 577
e-mail: dojer@mimuw.edu.pl

z sekwencjonowania pełnoeksomowego,

- asemblacja *de novo* sekwencji genomowych ze szczególnym uwzględnieniem obszarów repetytywnych.

Tytuł rozprawy, czyli *Szacowanie liczby powtórzeń fragmentu DNA*, łączy te zagadnienia podkreślając jej spójność tematyczną. Warto jednak zaznaczyć, że wyniki uzyskane w publikacjach poświęconych drugiemu z wymienionych zagadnień istotnie wykraczają poza zasugerowany w tytule zakres.

Problem składania sekwencji repetytywnych to obecnie jedno najważniejszych wyzwań w dziedzinie asemblacji sekwencji genomowych. O ile większość genomu człowieka zrekonstruowana została już 20 lat temu, poznanie obszarów centro- i telomerowych przez wiele lat wydawało się nieosiągalne ze względu na ich złożoną strukturę. Istotny postęp w tej dziedzinie dokonał się dopiero w ostatnich latach za sprawą doskonalenia technologii sekwencjonowania trzeciej generacji oraz, co nie mniej ważne, opracowania metod analizy zdolnych do wykorzystania informacji zawartej w otrzymanych tymi technologiami danych. Wiele prac dotyczących tej tematyki ukazało się w prestiżowych czasopismach w ciągu ostatnich trzech lat, czyli równolegle lub już po ukazaniu się publikacji wchodzących w skład rozprawy. Dlatego uważam, że tematyka rozprawy bardzo dobrze wpisuje się w aktualne badania w obszarze problemu asemblacji sekwencji DNA.

Główne wyniki rozprawy

Pierwszemu z wymienionych zagadnień, czyli wyznaczaniu liczby kopii fragmentów DNA, poświęcone są dwie prace z cyklu: [P1] oraz [P4]. Praca [P1] dotyczy doboru próbek do modelowania tła podczas szacowania liczby kopii. Zaproponowano zastosowanie w tym celu grupowania próbek metodą k-means. Pokazano, że k-means daje podobnie dobre wyniki jak najlepsze z dotychczas stosowanych podejść, czyli algorytm k najbliższych sąsiadów, ale jest znacznie szybszy.

Artykuł [P4] poświęcony jest wcześniejszemu etapowi wyznaczania liczby kopii wariantów, tzn. obliczaniu głębokości pokrycia odczytami z sekwencjonowania analizowanych fragmentów DNA. W pracy opracowano i zaimplementowano efektywnie zrównoleglonąaplikację obliczającą głębokość pokrycia.

Pozostałe prace cyklu dotyczą drugiego ze wspomnianych zagadnień, czyli asemblacji sekwencji genomowych. Artykuł [P2] prezentuje aplikację dnaasm do asemblacji *de novo* odczytów z sekwencjonowania technologią drugiej generacji. Algorytm asemblacji opiera się na powszechnie stosowanych przy tym problemie grafach de Bruijna, ale w nowatorski sposób wykorzystuje głębokość pokrycia grafu

ul. Banacha 2, 02-097 Warszawa, Polska
+48 22 55 44 577
e-mail: dojer@mimuw.edu.pl

odczytami do zrekonstruowania sekwencji repetytywnych.

Z kolei praca [P3] dotyczy asemblacji heterogenicznych zbiorów danych, tzn. łączących odczyty z drugiej i trzeciej generacji sekwencjonowania. W pracy zaprezentowano aplikację dnaasm-link, służącą do łączenia contigów otrzymanych w wyniku asemblacji krótkich odczytów z drugiej generacji sekwencjonowania w oparciu o długie odczyty z trzeciej generacji. Aplikacja pozwala ponadto na wykorzystanie długich sekwencji do wypełnienia przerw pomiędzy contigami.

Aplikacje dnaasm i dnaasm-link zostały wykorzystane przez doktoranta do złożenia genomu tasiemca szczurzego. Opisany w artykule [P5] rezultat asemblacji pozwolił znacząco poprawić jakość genomu referencyjnego tasiemca, np. parametr N50 został zwiększony ponad czterdziestokrotnie. Odtworzenie w całości sekwencji genomu mitochondrialnego potwierdza skuteczność opracowanych aplikacji w asemblacji obszarów repetytywnych.

Podsumowując, narzędzia opracowane w pracach [P1] i [P4] pozwalają istotnie przyspieszyć wyznaczanie liczby kopii fragmentów DNA przy zachowaniu wysokiej jakości wyników, natomiast aplikacje opracowane w pracach [P2] i [P3] umożliwiają poprawę jakości asemblacji repetytywnych obszarów sekwencji DNA. Na podkreślenie zasługuje wartość praktyczna uzyskanych wyników, w przypadku narzędzi dnaasm i dnaasm-links potwierdzona zastowaniem w opisanym w pracy [P5] rzeczywistym projekcie sekwencjonowania genomu eukariotycznego.

Uwagi krytyczne i dyskusyjne

Poniższe uwagi nie podważają oceny uzyskanych w rozprawie wyników.

1. W pracach [P2] i [P3] sprawdzenie efektywności odtwarzania obszarów repetytywnych zostało oparte na klasyfikacji wynikowych sekwencji programem Tandem Reapeat Finder. Intencja stojąca za przyjęciem takiej miary jest dla mnie niezrozumiała. Celem nie jest bowiem zwrócenie przez program sekwencji wykazującej cechy repetytywności, ale wierne odtworzenie oryginalnej sekwencji z obszaru repetytywnego. To ostatnie mogło być sprawdzone bezpośrednio, poprzez porównanie wynikowych sekwencji z odpowiednimi fragmentami genomu referencyjnego.
2. Zarazem narzędzie Tandem Reapeat Finder znakomicie nadaje się do wykrywania fałszywych pozytywów w rekonstrukcji obszarów repetytywnych, czyli sekwencji błędnie uznanych przez algorytm asemblacji za powtórzenia. W rozprawie brakuje dyskusji tego problemu, choć podejście zastosowane w aplikacji dnaasm (czyli szacowanie liczby kopii na podstawie głębokości

ul. Banacha 2, 02-097 Warszawa, Polska
+48 22 55 44 577
e-mail: dojer@mimuw.edu.pl

pokrycia) stwarza niebezpieczeństwo wystąpienia tego rodzaju błędów. Co więcej, wyniki zamieszczone w tabelach 5 i 6 pracy [P2] świadczą o tym, że algorytmowi zdarza się przeszacować liczbę kopii sekwencji repetytywnej.

3. W pracy [P3] przedstawiono klasyfikację metod łącznego wykorzystania do asemblacji odczytów z drugiej i trzeciej generacji sekwencjonowania na cztery różne podejścia. Tymczasem ewaluacja aplikacji dnaasm-link została ograniczona do porównania z narzędziami reprezentującymi to samo podejście. Dla uzyskania pełnego obrazu możliwości zaproponowanej metody należałoby dołączyć do porównania narzędzia należące do pozostałych kategorii, ewentualnie scharakteryzować mocne i słabe strony poszczególnych podejść.
4. W pracy [P5] obok aplikacji dnaasm do asemblacji genomu został użyty także program ABYSS. Podobnie do łączenia contigów obok dnaasm-link wykorzystano też program LINKS. Celowość zastosowania różnych programów do wykonania tych samych zadań została uzasadniona tylko częściowo – wspomniano, że wykorzystanie dnaasm-link było podyktowane niewystarczającą wydajnością pamięciową programu LINKS. Dla właściwej oceny faktycznego wkładu aplikacji dnaasm i dnaasm-links w asemblację genomu tasiemca szczurzego należałoby scharakteryzować obszary zastosowania poszczególnych narzędzi bądź opisać zasady agregowania ich wyników, jeśli zostały zaaplikowane do tych samych danych.

Konkluzja

Uważam, ze przedstawiona rozprawa spełnia zwyczajowe i ustawowe wymogi stawiane rozprawom doktorskim, stanowi oryginalne rozwiązanie problemu naukowego, unaocznia ogólną wiedzę i umiejetnosci techniczne doktoranta w informatyce oraz świadczy o umiejętności samodzielnego prowadzenia pracy naukowej. Wnoszę zatem o dopuszczenie Pana magistra inżyniera Wiktora Kuśmirka do dalszych etapów przewodu doktorskiego. Ponadto, biorąc pod uwagę wysoki poziom merytoryczny rozprawy, znaczenie podjętych problemów badawczych oraz walory praktyczne uzyskanych wyników, wnioskuję o wyróżnienie rozprawy.

Norbert Dojer

ul. Banacha 2, 02-097 Warszawa, Polska
+48 22 55 44 577
e-mail: dojer@mimuw.edu.pl