Gliwice, 5 czerwca 2023

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: Distributed Algorithms and Computational Methods for Scalable Processing of High-Throughput Sequencing Data

Autor rozprawy: mgr inż. Marek Wiewiórka

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 Marka Wiewiórkę rozprawa doktorska składa się z cyklu powiązanych tematycznie publikacji wraz z towarzyszącym im opracowaniem (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 genomicznych pozyskiwanych technikami sekwencjonowania o dużej przepustowości (ang. high-thoroughput sequencing, HTS). Główne trzy tezy rozprawy koncentrują się wokół zagadnienia poprawy wydajności czasowej procesów drugo- i trzeciorzędowej analizy danych genomicznych, poprzez zastosowanie technik i platform rozproszonego przetwarzania danych, dedykowanych formatów i struktur danych, osadzeniu logiki analitycznej w deklaratywnym języku SQL oraz skalowania obliczeń na platformach chmury obliczeniowej. Zarówno tezy pracy, jak i motywacja prowadzonych badań w tym obszarze zostały sformułowane w sposób jasny i wyczerpujący. Charakter rozprawy określiłbym jako **eksperymentalno-wdrożeniowy**, ponieważ Autor:

- zaprojektował szereg rozproszonych metod dla problemów znajdowania przecięć przedziałowych w danych genomicznych, obliczania głębokości pokrycia odczytów DNA, podsumowywania krótkich odczytów, na platformie Apache Spark,

- dla potwierdzenia słuszności przyjętych rozwiązań przeprowadził badania eksperymentalne na publicznie dostępnych zbiorach danych, które pozwoliły zweryfikować, iż opracowane algorytmy i rozwiązania mogą być z powodzeniem stosowane w analizie sekwencji genomicznych i pozwalają na wyraźne przyspieszenie tych procesów analitycznych,
- udostępnił szereg narzędzi, które umożliwiają wdrożenie zaprojektowanych metod i ich wtórne wykorzystanie w skalowalnych potokach bioinformatycznej analizy danych.

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. Analiza ta została przedstawiona w rozdziałach 1.2 i 1.3 autoreferatu oraz w większości spośród sześciu przedstawionych publikacji Autora, które tworzą cykl publikacyjny będący głównym osiągnięciem rozprawy. Zawartość rozdziałów 1.2 i 1.3 autoreferatu, które obejmują m.in. podział etapów przetwarzania danych pozyskiwanych technikami HTS, przegląd metod i projektów drugo- i trzeciorzędowej analizy danych, które koncentrują się na rozwiązaniu problemów wydajnego prowadzenia obliczeń poprzez wykorzystanie rozproszenia kalkulacji 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 167 pozycji literaturowych, z których zdecydowana większość dotyczy wyżej wymienionych elementów stanu wiedzy. Rozdziały 1.1 – 1.4 oraz rysunki 1.1 i 1.2 stanowią bardzo dobry wstęp teoretyczny do całości rozprawy, a do pojęć w nich zdefiniowanych (włączając wstęp do rozdziału) Autor nawiązuje w kolejnych podrozdziałach autoreferatu, jak również w treści poszczególnych artykułów głównego cyklu publikacyjnego. Szczególną uwagę zwraca Autor na problemy wydajności prowadzenia procesów analitycznych dla danych genomicznych, formatów danych dla istniejących danych genomicznych, które nie są w pełni dopasowane do przetwarzania na platformach klasy Big Data, właściwego sposobu partycjonowania tych danych, łatwości pozyskania informacji, kodowania i realizacji procesu znajdowania przecięć przedziałowych z wykorzystaniem podejścia deklaratywnego, przenaszalności rozwiązań i automatyzacji metod alokacji zasobów. Przeprowadzony przez Autora przegląd wiedzy w tym zakresie pozwolił mu w sposób jasny i przekonywujący sformułować wnioski, które zdeterminowały przyjęcie odpowiednich rozwiązań algorytmicznych i obliczeniowych.

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

Na początku realizacji rozprawy Pan Marek Wiewiórka zdefiniował kilka zadań, do których realizacji konsekwentnie dążył w swoich pracach badawczych. Dotyczyły one m.in. ewaluacji modeli przetwarzania rozproszonego do badań genomicznych, aplikowalności architektur hurtowni danych do analiz wariantów genomicznych, implementacji metod rozproszonego przetwarzania w jeziorze danych DNA. W swoich pracach Autor sięgnął do rozwiązań bazujących na modelu przetwarzania wykorzystującego tzw.

przyjemne zrównoleglenie obliczeń (ang. pleasantly parallel computations) dla rozwiązania wspomnianych problemów drugo- i trzeciorzędowej analizy danych. 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 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 prowadzenia wydajnej analizy danych DNA. Pan Marek Wiewiórka zaproponował nowatorskie, oparte o narzędzie klasy Big Data i model chmury obliczeniowej, podejścia w zakresie efektywnej realizacji analiz genomicznych, a także przeprowadził proces ich wnikliwej oceny. Na rozprawę, oprócz autoreferatu, składa się sześć publikacji w renomowanych czasopismach, w tym pięć z listy Journal Citation Report (JCR), które są ze sobą ściśle powiązane, w których Pan Marek jest pierwszym autorem lub jego udział jest większościowy. Charakteryzując krótko zawartość przedłożonych prac P1-P6 można stwierdzić, iż:

[P1] W pracy tej przedstawiono oparte na platformie Apache Spark środowisko SparkSeq rozproszonej i interaktywnej analizy danych genomicznych, polegającej na zliczaniu odczytów DNA, sortowaniu, filtrowaniu danych i przeglądaniu zawartości plików BAM składowanych w rozproszonym systemie plików HDFS. Przeprowadzone eksperymenty wykazały, że jest to rozwiązanie wielokrotnie szybsze od istniejących do tej pory narządzi (np. SeqPig). Badano również różne strategie buforowania danych, serializacji i kompresji danych w kontekście wykorzystania zasobów procesora oraz pamięci i korzyści wydajnościowych wynikających z zastosowania różnych rozmiarów bloku systemu plików HDFS. W pracy tej Pan Marek Wiewiórka był odpowiedzialny za implementację funkcji analitycznych oraz prowadził badania wydajnościowe i strojenie środowiska.

[P2] W pracy tej przedstawiono wydajnościową analizę porównawczą silników równoległej realizacji zapytań (Spark, Impala, Hive, Prestodb), a także Apache Kylin i (nierozproszonego) MonetDB do hurtowni danych wariantów genetycznych z wykorzystaniem formatów danych ORC i Parquet oraz implementacji one-big table (OBT) w celu redukcji liczby złączeń między tabelami. Badania te były istotne z punktu widzenia pozostałych prac nad wykorzystaniem środowiska Apache Spark do prowadzenia rozproszonych analiz genomicznych. Do oryginalnego wkładu Autora rozprawy należy przede wszystkim opracowanie struktury hurtowni danych, implementacja rozwiązań w różnych

środowiskach wykonawczych oraz przeprowadzenie badań wydajnościowych, w tym badań porównawczych w stosunku do przyjętego rozwiązania referencyjnego.

[P3] W pracy tej przedstawiono rozwiązanie dla rozproszonej analizy struktury populacji z wykorzystaniem metod uczenia maszynowego. Rozwiązanie to ponownie oparto na platformie Apache Spark. Badania potwierdziły, iż opracowane rozwiązanie pozwala 100-krotnie przyspieszyć analizę danych genomicznych w badaniach struktury populacji w stosunku do wersji nierozproszonej. Ponownie, do oryginalnego wkładu Autora rozprawy należało przygotowanie infrastruktury obliczeniowej i przeprowadzenie badań wydajnościowych, w tym badań porównawczych w stosunku do istniejącego stanu wiedzy (udział Autora rozprawy w tym rozwiązaniu określono na 20%).

[P4] W pracy tej przedstawiono zaimplementowaną w oparciu o środowisko Apache Spark platformę SeQuiLa oraz jej możliwości w zakresie przetwarzania zakresów genomicznych. Rozwiązanie oparto na strukturze poszerzonego drzewa interwałów AIT rozgłaszanej do węzłów obliczeniowych. Technicznie rozwiązania oparto na funkcjach użytkownika osadzonych w deklaratywnym języku SQL. Wydajność tego rozwiązania została zweryfikowana eksperymentalnie na publicznych zbiorach danych obejmujących dopasowania sekwencji WES i WGS. Wkład Autora rozprawy polegał m.in. na formalnej analizie algorytmów, implementacji opracowanych rozwiązań i przeprowadzeniu testów i porównaniu wyników z bieżącymi rozwiązaniami.

[P5] W pracy tej przedstawiono rozszerzenie platformy SeQuiLa umożliwiające 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 formalnej analizie algorytmów, implementacji opracowanych rozwiązań i przeprowadzeniu testów i porównaniu wyników z bieżącymi rozwiązaniami.

[P6] W artykule tym przedstawiono algorytm dla podsumowywania krótkich odczytów z sekwencjonowania NGS, które jest częstym elementem bioinformatycznych potoków przetwarzania danych, a także algorytm partycjonowania dedykowany temu typowi danych. Algorytmy te zostały dobrze sformalizowane i szerzej opisane w przedstawionym artykule, w którym Autor rozprawy ma większościowy udział. Autor rozprawy przeprowadził w nim zarówno formalną analizę, jak i implementację opracowanych rozwiązań. Przeprowadzone testy pokazały, że opracowane rozwiązania dla platformy Apache Spark są znacznie szybsze niż dotychczas znane rozwiązania procesu podsumowywania krótkich odczytów z sekwencjonowania NGS.

Podjęcie problemów, takich jak wyznaczanie głębokości pokrycia, podsumowywania krótkich odczytów z sekwencjonowania NGS i analizy zakresowej, będących procesami składowymi potoków analitycznych oraz opracowanie dla nich odpowiednich podejść algorytmicznych i osadzenie tych algorytmów w rozproszonym środowisku klasy Big Data, wspartym przez środowisko chmury obliczeniowej, a także udostępnienie funkcjonalności poprzez interfejs deklaratywnego języka SQL uważam za istotne osiągnięcie Autora i zaliczam do oryginalnych wyników przedstawionych w rozprawie. Walorem rozwiązań jest również możliwość realizacji obliczeń w środowisku skonteneryzowanym, a także wdrażanie ich w skalowalnych środowiskach chmury obliczeniowej poprzez zastosowanie podejścia *Infrastructure as*

Code (IoC). Udział procentowy oraz wkład Autora rozprawy zostały potwierdzone oświadczeniami podpisanymi przez współautorów publikacji. Wyniki przeprowadzonych prac badawczych zostały opublikowane w 6 artykułach w liczących się w dziedzinie informatyki i bioinformatyki czasopismach, m.in. *Bioinformatics* (200 pkt. MEiN), *Database* (100 pkt. MEiN), *Giga-Science* (200 pkt.), *BMC Bioinformatics* (140 pkt. MNiSW). Dorobek ten uzupełnia 1 artykuł opublikowanych w materiałach konferencyjnych, 3 rozdziały w monografiach naukowych, 7 wystąpień konferencyjnych, 4 wystąpienia plakatowe. Na uwagę zasługuje udział w licznych projektach o charakterze naukowym (m.in. grant Microsoft Azure for Research Award) oraz nagrody. Świadczy to w mojej opinii o istotności podjętego problemu oraz wyraźnym wkładzie Pana Marka Wiewiórki 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 Marek Wiewiórka 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. 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-P6 przedłożonej rozprawy pokazując, że poszczególne rozwiązania, modyfikacje i rozszerzenia dotychczasowych metod umożliwiają poprawę jakości i wydajności osiąganych wyników w porównaniu z wybranymi i dostępnymi metodami. Od strony redakcyjnej zarówno autoreferat, jak i prace P1-P6 są w większości napisane w dobrym stylu i czyta się ją z łatwością, chociaż znalazłem w samym autoreferacie również kilka zdań, w których należałoby poprawić stylistykę wypowiedzi.

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 w języku angielskim, który zawiera najważniejsze konkluzje wypływające z przeprowadzonych prac badawczych. Nie znalazłem w tych dokumentach istotnych uchybień. Uwaga dotycząca poprawy stylistyki zdań w autoreferacie nie ma charakteru znacząco krytycznego i nie umniejsza znaczeniu osiągnięć Autora rozprawy. Mam natomiast kilka pytań, na które odpowiedź chętnie bym poznał:

Q1. W rozdziale 2.1 autoreferatu napisano „development of novel algorithms is a lot easier than with the low-level Map-Reduce approach". Nie znalazłem jednak dla tego stwierdzenia uzasadnienia. Mogę przypuszczać dlaczego tak jest, ale chciałbym się dowiedzieć skąd zdaniem Autora wynika ta łatwość.

Q2. Na rysunku 1.2 autoreferatu pojawia się blok „[P6] Cloud-native distributed genomic pileup operations". Co oznacza wg Autora rozprawy „cloud-native"?

Q3. W odniesieniu do publikacji [P2] i użytych w prowadzonych badaniach formatów danych ORC i Parquet, czy w świetle wyników badań byłoby rozsądne zaproponowanie własnego formatu przechowywania danych dla pewnych grup danych pod kątem analiz genomicznych wykonywanych na platformach klasy Big Data, takich jak Apache Spark lub Hadoop?

Q4. Moją ciekawość zawsze budzi stwierdzenie „głębokość pokrycia (depth of coverage)" i zawsze pojawia się u mnie pytanie „pokrycia, ale czego?"

Q5. W odniesieniu do publikacji [P6], w autoreferacie stwierdzono „implementation of Apache Spark Catalyst custom execution strategy that handles depth of coverage calculations for both DataFrame and SQL APIs". Na czym dokładnie polegała ta strategia?

Q6. Odnośnie osiągnięć O22 i O23 przedstawionych w sekcji 3.8 autoreferatu, w jakim konkursie je otrzymano? Czy był to wewnętrzny konkurs Politechniki Warszawskiej?

7. Jaka jest przydatność rozprawy dla nauk technicznych?

Uważam, że przedłożona rozprawa doktorska Pana Marka Wiewiórki wpisuje się w bieżące problemy bioinformatyki i genomiki funkcjonalnej. Opracowanie rozproszonych wersji algorytmów dla różnych zadań drugo- i trzeciorzędowej analizy danych genomicznych i utworzenie całościowej, rozszerzalnej platformy SeQuiLa pozwoliło Autorowi na wielokrotną poprawę szybkości realizacji tych zadań 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 DNA pozyskiwanych technikami wielkoskalowymi. Potwierdzają to publikacje, których Pan Marek Wiewiórka jest autorem, opublikowane przez wiodące wydawnictwa, takie jak *Oxford* oraz *Springer*.

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 Marka Wiewiórkę w trakcie realizowanych przez niego badań pozwalają potwierdzić główne tezy rozprawy przedstawione w rozdziale 1.5 autoreferatu. Wyniki badań pokazują, że techniki oraz metody zaproponowane przez Pana Marka Wiewiórkę mogą przyczynić się do znaczącej poprawy procesów analizy danych 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.

Dariusz Mrozek

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