RDITT/40/2020
NPT. 12.10.2020
M. Kokot

prof. dr hab. inż. Marta Kasprzak
Instytut Informatyki
Politechnika Poznańska

Poznań, 6.10.2020

Recenzja rozprawy doktorskiej mgr. inż. Marka Kokota pt. "Wyznaczanie zbiorów podsłów sekwencji nukleotydowych w danych z sekwencjonowania genomów"

1. Problematyka naukowa rozprawy

Sekwencjonowanie genomu to proces pozyskiwania informacji o jego sekwencji nukleotydowej. Nukleotydy formują cząsteczki kwasów nukleinowych DNA i RNA, a więc kodują informację genetyczną, na której w dużej mierze opiera się funkcjonowanie organizmów. Cząsteczki DNA i RNA mają postać nici, a informacja genetyczna w nich zawarta, gdy rozważamy tzw. strukturę pierwszorzędową, interpretowana jest jako sekwencja znaków odpowiadających poszczególnym nukleotydom (A, C, G i T w przypadku DNA). Rozpoznanie właściwej sekwencji nukleotydowej badanego genomu lub jego fragmentu jest pierwszym krokiem na drodze do dalszych analiz i interpretacji pozyskanej wiedzy.

Sekwencjonowanie realizowane jest obecnie z użyciem zautomatyzowanych sekwenatorów, które zwracają na wyjściu w jednym przebiegu miliony (nawet setki milionów) krótkich sekwencji nukleotydowych, tzw. odczytów. Ilość takiej informacji do przetworzenia naraz, w celu złożenia odczytów w dłuższe ciągi lub wydobycia innej informacji, wymaga użycia wyspecjalizowanych algorytmów. Proces ten komplikowany jest dodatkowo obecnością błędów eksperymentalnych w danych. Z informatycznego punktu widzenia mamy do czynienia z przetwarzaniem sekwencji znaków za pomocą opracowywanych pod konkretny problem algorytmów, gdzie rozwiązywane problemy formułowane na gruncie kombinatorycznym są zwykle NP-trudne, a nawet te łatwe obliczeniowo wymagają optymalizacji ze względu na olbrzymi rozmiar danych.

Tej tematyce poświęcona jest przedłożona rozprawa, a rozwiązania w niej zawarte odnoszą się do problemów efektywnego przetwarzania dużych zbiorów odczytów celem rozpoznania w nich i porównania krótkich podsekwencji o zadanej długości k (k-merów). Jest to często realizowany etap wstępny w procesach składania dłuższych sekwencji nukleotydowych czy porównywania/dopasowywania sekwencji, z tego też względu potrzebne jest opracowywanie coraz to efektywniejszych algorytmów, dostosowywanych do rosnących zbiorów danych biologicznych i zawartych w nich błędów. Takie algorytmy w swojej rozprawie proponuje pan mgr inż. Marek Kokot, także w wersji równoległej, uzupełniając je o szczegółową analizę

ich złożoności obliczeniowej oraz o porównanie z innymi algorytmami. Uwzględnione zostały m.in. takie aspekty optymalizacji czasowej i pamięciowej przetwarzania danych, jak oszczędniejszy dostęp do danych dyskowych, dostosowanie procedur do charakteru danych wejściowych oraz innowacyjne rozwiązania w zakresie sortowania danych. Tematyka ta może być zatem przedmiotem rozprawy doktorskiej w dyscyplinie informatyka techniczna i telekomunikacja.

2. Opinia o rozprawie

Rozprawa składa się z siedmiu rozdziałów uzupełnionych o sekcje dodatkowe. W pierwszym rozdziale Doktorant wprowadza czytelnika w tematykę rozprawy, stawia tezy, omawia cel i zakres pracy. W rozdziale drugim zapoznaje czytelnika z kluczowymi dla zrozumienia rozprawy pojęciami z zakresu biologii molekularnej, a w rozdziale trzecim z zakresu informatyki. W rozdziale czwartym dokonuje przeglądu istniejących algorytmów, do których odwołuje się w treści rozprawy, realizujących sortowanie danych i zliczanie *k*-merów, a także bioinformatycznych aplikacji służących przeprowadzaniu operacji porównania na zbiorach *k*-merów.

Kolejne rozdziały poświęcone są przedstawieniu pracy własnej Doktoranta. Rozdział piąty opisuje wszystkie nowe rozwiązania algorytmiczne, zamieszczone w trzech podrozdziałach dedykowanych trzem zagadnieniom przybliżonym w rozdziale czwartym: sortowaniu, zliczaniu *k*-merów w zbiorze odczytów i operacjom przeprowadzanym na zbiorach *k*-merów. Wszystkie rozwiązania miały na celu opracowanie algorytmów o mniejszych wymaganiach czasowych i pamięciowych niż wcześniejsze algorytmy. Każdy z tych podrozdziałów kończy się sekcją obszernej analizy złożoności obliczeniowej algorytmów. Opracowane zostały dwa hybrydowe algorytmy sortowania (RADULS, RADULS2), dwa algorytmy zliczające *k*-mery (KMC2, KMC3) i narzędzie KMCtools operujące na zbiorach *k*-merów.

W rozdziale szóstym przedstawione zostały wyniki eksperymentów obliczeniowych. Działanie nowych algorytmów porównano z wieloma algorytmami innych autorów: dwunastoma algorytmami sortowania w odniesieniu do bardzo małych tablic, sześcioma innymi algorytmami sortowania pozycyjnego lub ogólnego przeznaczenia, sześcioma algorytmami zliczania *k*-merów i trzema narzędziami do wykonywania operacji na zbiorach *k*-merów. W sumie użyto sześć platform sprzętowo-programowych, różniących się rodzajem maszyny, procesorami, systemem operacyjnym. Równoległe działanie algorytmów sprawdzono na liczbie wątków od 1 do ponad 60 w przypadku algorytmów sortowania. Do testów algorytmów sortowania użyto danych generowanych losowo, do testów programów wyznaczania i przetwarzania zbiorów *k*-merów użyto danych rzeczywistych z sekwencjonowania genomów pięciu organizmów. Doktorant podsumował uzyskane rezultaty w rozdziale siódmym.

Wyżej wymienione algorytmy powstały przy współudziale innych osób, jednak wkład Doktoranta był kluczowy. W przypadku algorytmów RADULS, RADULS2, KMC3 i KMCtools mgr inż. Marek Kokot był pierwszym autorem publikacji (spośród trzech

autorów), w przypadku KMC2 był on drugim autorem publikacji (spośród czterech) po swoim promotorze.

Tezy postawione w rozprawie odnoszą się do wszystkich zaproponowanych algorytmów i deklarują pozytywny wpływ zastosowanych w nich nowych rozwiązań algorytmicznych na czas obliczeń i zapotrzebowanie na pamięć. Wszystkie tezy znalazły uzasadnienie w przytoczonych wynikach.

Formułując swoją ocenę rozprawy, pragnę na wstępie podkreślić bardzo staranną jej redakcję, zarówno w warstwie tekstowej, jak i graficznej. Mgr inż. Marek Kokot w sposób czytelny i kompletny ujął w niej zagadnienia i podejścia istotne dla zrozumienia dalszej części rozprawy. Także efekty własnej pracy przedstawił z dbałością o wszelkie potrzebne szczegóły i odbiór ze strony czytelnika. W ramach pracy Doktorant poświęcił się badaniom zarówno z zakresu informatyki teoretycznej – analizie złożoności obliczeniowej algorytmów podpartej serią udowodnionych przez niego twierdzeń – jak i bardziej praktycznym aspektom, obejmującym implementację i testowanie algorytmów na danych generowanych losowo i rzeczywistych.

Do najważniejszych osiągnięć badawczych przedstawionych w rozprawie zaliczyłabym:
— Opracowanie algorytmów RADULS i RADULS2, w których połączono różne podejścia do sortowania danych w celu jak najlepszej optymalizacji tego procesu. Należy zwłaszcza podkreślić pracę nad rozwiązaniem najtrudniejszego aspektu tych obliczeń, sortowania wielu bardzo małych tablic. W testach oba te algorytmy okazały się szybsze od innych, przy czym zwykle RADULS2 uzyskiwał przewagę nad RADULS. Przewaga ta wynika w dużej mierze z zaimplementowania nowego rozwiązania na etapie sortowania bardzo małych tablic.
— Szczegółowe rozwiązania zaimplementowane w algorytmie KMC3, które przełożyły się na znaczną oszczędność czasu obliczeń względem algorytmu KMC2: włączenie algorytmu RADULS, usprawniony odczyt skompresowanych plików, zrównoleglenie procedur identyfikowania (k,x)-merów i k-merów w sekwencjach, optymalizacja procesu doboru sygnatur. W odniesieniu do procesu doboru sygnatur, na uwagę zasługują opracowane w tym celu metaheurystyka symulowanego wyżarzania oraz inne podejścia heurystyczne, dogłębnie porównane w testach.
— Bardzo obszerną analizę złożoności czasowej i pamięciowej algorytmów przeprowadzoną dla ich wersji szeregowych i równoległych. Twierdzenia sformułowane i udowodnione przez Doktoranta rozciągają się na ponad 40 stron.

Należy też zwrócić uwagę na przeprowadzone solidne eksperymenty obliczeniowe, obejmujące wyjątkowo dużo algorytmów innych autorów, szerokie spektrum zbiorów testowych i kilka platform sprzętowych, dzięki którym tezy postawione w rozprawie znalazły wiarygodne uzasadnienie. Dogłębne zbadanie przez Doktoranta nowych rozwiązań algorytmicznych i odniesienie się do bieżącego stanu wiedzy w takim stopniu, jak przedstawione w rozprawie, wskazują na jego wysokie kompetencje w zakresie prowadzenia badań naukowych.

W trakcie czytania rozprawy nasunęły mi się pojedyncze uwagi krytyczne o nikłym wpływie na jej odbiór. Omawianie w rozdziale trzecim szeregu metaheurystyk, z których do

większości brak jest odniesienia w dalszej części rozprawy, uważam za niekonieczne. Dla średniej liczby odczytów przypadającej na daną pozycję w genomie preferuję nazwę głębokość pokrycia, nie pokrycie, gdyż to drugie pojęcie rozumiane jest także jako część genomu mająca swoją reprezentację w odczytach. W tekście wystąpiły nieliczne literówki i błędy interpunkcyjne.

Powyższe uwagi nie podważają wartości rozprawy i mojej bardzo pozytywnej oceny pracy naukowej Doktoranta. Uzyskane przez niego rezultaty uważam za znaczące dla dyscypliny informatyka techniczna i telekomunikacja.

3. Podsumowanie

Przedłożoną rozprawę uważam za napisaną w sposób przejrzysty, poprawny logicznie i metodologicznie. Pan mgr inż. Marek Kokot wykazał się wiedzą i umiejętnościami z zakresu informatyki w rozwiązywaniu problemów informatycznych i bioinformatycznych. Potwierdził wagę swoich badań publikacjami przyjętymi m.in. trzykrotnie w czasopiśmie z listy JCR *Bioinformatics*, najlepszym czasopiśmie z tego obszaru (200 punktów ministerialnych, IF_{2019}=5,610). Dwie z tych publikacji w *Bioinformatics* są bezpośrednio związane z zawartością rozprawy, trzecia to efekt dalszych badań w zakresie identyfikacji i wykorzystania zbiorów *k*-merów. Według bazy *Web of Science* cytowane były one w sumie 151 razy, a 290 razy wg *Google Scholar*.

Na podstawie wyrażonych powyżej opinii stwierdzam, że rozprawa pt. "Wyznaczanie zbiorów podsłów sekwencji nukleotydowych w danych z sekwencjonowania genomów" autorstwa mgr. inż. Marka Kokota spełnia warunki stawiane rozprawom doktorskim przez obowiązującą ustawę o stopniach naukowych i tytule naukowym. Wnoszę o dopuszczenie tej rozprawy do publicznej obrony.

Mając na uwadze mocną zarówno warstwę teoretyczną, jak i praktyczną rozprawy, a także trzy publikacje mgr. inż. Marka Kokota w *Bioinformatics*, wnoszę o wyróżnienie rozprawy.

Marta Kasprzak