RDITT-41/2020
wpł. 12.10.2020
M. Skon

UNIWERSYTET WARSZAWSKI | UW MIM

Wydział Matematyki, Informatyki i Mechaniki
Instytut Informatyki

dr hab. Paweł Górecki, prof. UW
Wydział Matematyki, Informatyki i Mechaniki
Uniwersytet Warszawski
Banacha 2, 02-097 Warszawa
gorecki@mimuw.edu.pl

Warszawa, 6 października 2020

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

I. Problematyka naukowa rozprawy

Recenzowana rozprawa doktorska mgr. inż. Marka Kokota poświęcona jest zagadnieniom dotyczącym operacji na podsłowach, tzw. k-merach, zawartych w sekwencjach biologicznych. Formalnie, k-mer to podciąg długości k występujący w pewnej ustalonej sekwencji. Pojęcie k-mera jest jednym z podstawowych w bioinformatyce i występuje od dziesiątek lat w szeregu klasycznych problemów i narzędzi często stosowanych w bioinformatycznych potokach do przetwarzania danych.

W rozprawie centralnym problemem jest *zliczanie k-merów w sekwencjach DNA*, który można zdefiniować następująco: *dla ustalonego k i danego zbioru odczytów DNA podaj wszystkie k-mery wraz z licznościami występujące w odczytach.* W kontekście sekwencjonowania DNA ten problem jest zwykle obarczony dodatkową trudnością związaną z charakterem danych wejściowych: odczyty z sekwencjonowania podwójnej helisy DNA mogą także pochodzić z komplementarnej nici. Dodatkowym problemem praktycznym jest potencjalnie duży rozmiar danych wejściowych osiągający np. setki gigabajtów dla genomu człowieka. Celem przedstawionej rozprawy było opracowanie algorytmów wspomagających zliczanie k-merów ze szczególnym naciskiem na wydajne czasowo i pamięciowo w praktyce implementacje wraz z zastosowaniami do danych pochodzących z sekwencjonowania genomów.

Podejmowana tematyka badawcza należy do pogranicza informatyki, matematyki i bioinformatyki. Warto podkreślić, że stosunkowa łatwość sekwencjonowania w dzisiejszych czasach powoduje zalew danych, co niestety nie idzie w parze z

P. Gór

możliwościami ich analizy nie wspominając o przechowywaniu. Zatem każde podejście do optymalizacji potoków bioinformatycznych jest doskonale umotywowane i bardzo dobrze wpisuje się w światowe trendy badawcze.

II. Zawartość rozprawy

Rozprawa napisana w języku polskim składa się z 7 rozdziałów i 3 dodatków. Rozdział pierwszy to wprowadzenie do rozprawy, w szczególności przedstawia tezy, cele oraz krótkie streszczenie kolejnych rozdziałów. Rozdział drugi zawiera podstawy biologiczne oraz wstęp do sekwencjonowania. Rozdział trzeci wprowadza podstawowe pojęcia i zagadnienia informatyczne powiązane z tematyką rozprawy: struktury danych, problemy sortowania i przegląd heurystyk optymalizacyjnych. W Rozdziale czwartym Autor przedstawia istniejące rozwiązania dot. sortowania, zliczania k-merów oraz operacji na zbiorach k-merów. Główny wkład Autora jest przedstawiony w piątym i szóstym Rozdziale, które bardziej szczegółowo opiszę poniżej. Rozdział siódmy zawiera konkluzję. Ponadto w rozprawie znajdują się trzy dodatki z dodatkowymi tablicami wyników eksperymentalnych, szczegółami wołań kompilatora i listą adresów akcesyjnych danych genomowych.

III. Opinia

Główne wyniki rozprawy są zamieszczone w dwóch rozdziałach. W Rozdziale 5 znajduje się szczegółowy opis algorytmów wraz z analizą złożoności, a w Rozdziale 6 przedstawione są wyniki eksperymenty obliczeniowych.

Pierwszym algorytmem jest RADULS 1, inspirowany częściowo algorytmem z pracy Satisha. Jest to wielowątkowy algorytm sortowania pozycyjnego, w którym stosowany jest dodatkowy bufor pamięci oraz kilka poziomów rozmiarów kubełków, dla których sortowania stosowane są odrębne strategie. W rozprawie przedstawiona jest także wersja tego algorytmu o nazwie RADULS 2, który ma dodatkowy poziom rozmiarów kubełków i dodatkowe usprawnienia w procedurze buforowania. W ostatnim etapie gdy rozmiary kubełków są małe, w celu znalezienia najszybszego sposobu sortowania Autor przeprowadził szereg eksperymentów obliczeniowych, których szczegóły są zaprezentowane w Rozdziale 6, z użyciem kilkunastu algorytmów sortowania. W wyniku eksperymentów został wyznaczony algorytm hybrydowy, który uruchamiania odpowiednie algorytmy sortowania w zależności od progów związanych z rozmiarem klucza, rekordu i tablicy.

Kolejny algorytm to KMC 2 służący do zliczania k-merów, będący usprawnieniem algorytmu KMC 1. Usprawnienie polega na dodaniu tzw. sygnatur i (k, x)-merów, a głównym celem jest redukcja zapotrzebowania na pamięć dyskową i operacyjną. Pierwsze rozwiązanie usprawnia proces grupowania sąsiadujących k-merów w tzw. super-k-mery, które wyznaczane są za pomocą podciągów nazywanych sygnaturami. Sygnatury służą do zapisu super-k-merów i istotnie wpływają na zapotrzebowanie pamięci dyskowej. Z tego powodu w rozprawie jest poświęcony osobny temat dotyczący optymalizacji doboru sygnatur, w której Autor stosuje, m.in., symulowane wyżarzanie do

generowania zbiorów sygnatur zoptymalizowanych pod kątem kilku kryteriów wspomagających oszczędność pamięci. Drugie usprawnienie w KMC 2 dotyczy etapu sortowania k-merów odczytanych ze zbiorów super-k-merów. Tutaj wprowadzony zostaje etap pośredni, w którym zamiast k-merów sortowane są nieco dłuższe (k, x)-mery rozpakowane z super-k-merów.

W rozprawie jest także opis algorytmu KMC 3, będącego udoskonaleniem algorytmu KMC 2. Główne zmiany to zastosowanie RADULS 1 na etapie sortowania oraz kilka optymalizacji na poziomie przydziału wątków i zrównoleglania. W Rozdziale 5 przedstawiony jest także zestaw narzędzi KMC tools.

Części algorytmiczne rozprawy, a szczególnie fragmenty o KMC 2 i RADULS 1, są nietrywialne, interesujące i dobrze zaprezentowana mimo dość skomplikowanej materii. Autor wykazał się nie tylko w projektowaniu efektywnych wielowątkowych algorytmów z użyciem różnych technik optymalizacji, ale także bardzo dobrą wiedzą dotyczącą architektur współczesnych procesorów i komputerów. Warto dodać, że Autor posiada także wyjątkowe umiejętności wykrywania i usuwania wąskich gardeł w proponowanych algorytmach.

Przedstawione wyniki teoretyczne dotyczą przede wszystkim analizy złożoności czasowej i pamięciowej zaproponowanych algorytmów. Do każdego z nich Autor przedstawił szereg lematów i twierdzeń dla przypadku szeregowego i równoległego. Przedstawione szacowania są zazwyczaj przy pewnych dodatkowych założeniach dotyczących np. rozmiarów tablic, czy zależności między k-merami. Dodam, że wprowadzenie uproszczeń jest uzasadnione. Dowody są skrupulatnie i prawidłowo przeprowadzone, choć mam tu pewne zastrzeżenia natury formalnej. Na przykład, należy zachować ostrożność w przypadku szacowań z symbolem O-duże z wieloma parametrami. Dodatkowo niektóre z parametrów czasem traktowane są jako stałe i są odpowiednio pomijane w notacji, a czasem są powiązane dodatkowymi zależnościami. Za najciekawszy i wymagający uważnej lektury uważam fragment dotyczący szacowania złożoności algorytmu KMC 2. Tutaj lekki niedosyt budzą konsekwencje wprowadzonego założenia o niezależności sąsiednich k-merów, które upraszczają obliczenie kluczowej dla innych szacowań wartości oczekiwanej liczby k-merów występujących w (k, x)-merze (Lemat 5.2.12). Jestem ciekawy, czy Autor rozważał pozbycie się tego założenia. Wydaje się, że obliczenie tej wartości w ogólności jest możliwe choć zapewne wymaga więcej wysiłku.

W Rozdziale 6 opisującym przeprowadzone eksperymenty obliczeniowe, oprócz wspomnianych wcześniej testów algorytmów sortowania małych tablic na potrzeby opracowania algorytmu hybrydowego i metaheurystyki symulowanego wyżarzania dla doboru sygnatur, zamieszczone są wyniki ewaluacji opracowanych algorytmów. Autor przedstawia wyniki dla różnych wariantów danych wejściowych tj. symulowanych i rzeczywistych, oraz porównuje implementacje algorytmów z szeregiem konkurencyjnych rozwiązań. Eksperymenty są starannie zaprojektowane, przeprowadzone i dobrze zilustrowane tabelami i wykresami. Konkluzje i analiza wyników są rzeczowe i nie budzą wątpliwości. Same wyniki pokazują, że opracowane narzędzia stanowią doskonałą

alternatywę wobec istniejących rozwiązań. W wielu sytuacjach osiągnięte przyspieszenie i oszczędność zużycia pamięci w proponowanych rozwiązaniach są wielokrotnie lepsze względem konkurencji co wprost potwierdza w praktyce tezy rozprawy postawione przez Autora. Do tej części mam uwagę dotyczącą występowania powtórzeń w sekwencjach. Autor zauważa i wyjaśnia, że obliczenia dla genomów z licznymi duplikacjami powodują mniejszą przepustowość algorytmów RADULS. To zjawisko bywa znaczące w kontekście ewolucji genomów dlatego problem wydajności algorytmów RADULS powinien być dokładniej zbadany by poznać skalę ograniczeń, np. przez przeprowadzenie osobnego eksperymentu porównawczego dla genomów sztucznych i rzeczywistych z różnym poziomem powtórzeń.

Poprawność rozprawy i uwagi redakcyjne

Rozprawa mimo obszerności i wielu szczegółów technicznych jest bardzo dobrze napisana, a wyniki we wszystkich aspektach przedstawione są jasno i zrozumiale.

Wśród formalnych uwag dotyczących warstwy redakcyjnej wymienię brak krótkich wprowadzeń do rozdziałów i podrozdziałów z opisem zawartości. Np. na stronie 78 czytelnik nie wie dlaczego nagle pojawia się opis algorytmu Satisha, a wcześniej nie ma o nim wzmianki. Więcej uwag zamieszczam poniżej. Dodam, że poniższe uwagi nie wpływają na zrozumienie i pozytywny odbiór rozprawy.

Str. 13. Do genomu wlicza się także DNA mitochondrialne.

Twierdzenie 5.1.11. Złożoność pamięciowa: nie wliczamy rozmiaru wejścia str. 91 w dowodzie.

Str. 40. Tu brakuje rozwinięcia pojęcia problemów trudnych dla których rozwiązywania stosowane są heurystyki.

Str. 74. W Algorytmie 2. p *zamiast* n *w linii 9 i 10.*

Str. 37. Drugi paragraf. Uwaga o złożoności quicksort może być myląca.

Str 84. conajmniej -> co najmniej

Notacja zbiorów zamiast $1..k$ *powinno być* $1, 2, ..., k$.

Str. 85. Lepiej napisać o "wersji szeregowej" zamiast używać t.

Str. 63. B_k *zamiast* B_K.

67. kiku -> kilku.

Dowód Lematu 5.1.8. W ostatniej formule powinno być ...=O(R)+O(p')+..., to oczywiście nie zmienia szacowania.

Kod algorytmów (eksponowany) powinien zawierać więcej komentarzy.

Str. 122. Powinno być wprost założenie o niezależność pozycji w modelu.

Str. 122. Nie jest jasne co oznacza, że porządek k-merów jest losowy.

Str. 126. W dowodzie (b) dla $\gamma = 0$ *formuła jest nieokreślona.*

Str. 128. W Lematach 5.2.10, 5.2.11 i kilku innych należy sprecyzować "Średnia złożoność czasowa".

Lemat 5.2.12. Tu na początku warto wprost napisać, że chodzi o założenia upraszczające modelu dot. niezależności sąsiednich k*-merów. Dodatkowo, np.: formuła* $n_x \cdot x'' \cdot P_1(x'') = \ldots$ *tak jak jest opisana może nie dać wartości całkowitych dla liczności* k*-merów. Jest to oczywiście skrót myślowy, ale tu i w innych miejscach analogiczne*

formuły powinny być opisane formalnie z wartościami oczekiwanymi dla pewnych zmiennych losowych.

Str. 130 (x,k) -> (k,x).

Str. 155. W dowodzie Twierdzeniu 5.3.4. $4^{n}_{L}{}^{w}UT$ *pominięte w środkowym fragmencie formuły.*

Str. 155. czynnik -> składnik.

Str. 156. W Twierdzeniu 5.3.5 są rozbieżności dot. wartości rozmiaru buforów. Np. z analizy w dowodzie wynika, że w pierwszym paragrafie powinno być 48 Mib zamiast 56; podobnie w drugim paragrafie. Również podana wartość w formule tezy twierdzenia nie jest sumą podanych wartości.

Str. 162. Nazwy wariantów ShellSort nieco inne niż w podrozdziale 5.1.2.

Rozdział o eksperymentach - pojęcie przepustowości powinno być zdefiniowane.

Podsumowanie

Recenzowana rozprawa zawiera szereg oryginalnych rozwiązań w postaci algorytmów, narzędzi i teoretycznych wyników złożonościowych z zakresu przetwarzania danych pochodzących z sekwencjonowania DNA. Wyniki zaprezentowane w pracy potwierdzają postawione tezy. Autor wykazał się bardzo dobrą znajomością stanu wiedzy o tematach podejmowanych w rozprawie badawczej, dotyczy to zarówno wiedzy o charakterze informatycznym i algorytmicznym jak i dotyczącym genomiki i sekwencjonowania. Ponadto, mgr inż. Marek Kokot posiada umiejętności formułowania problemów, projektowania algorytmów dla ich rozwiązywania oraz potrafi je formalnie zbadać, opisać i zaimplementować w sposób wydajny.

Warto dodać, że uzyskane wyniki zostały częściowo opublikowane w trzech artykułach w czasopiśmie Bioinformatics o IF=5.6. Dwa z nich to artykuły typu *application note*. Ponadto artykuły o KMC 2 (z 2015) i KMC 3 (2017) uzyskały łączną liczbę cytowań bliską 300 w Google Scholar (2020) co świadczy o wyjątkowo wysokim wpływie na dziedzinę. Oprócz tego Autor posiada kilka publikacji w materiałach konferencyjnych.

Stwierdzam, że recenzowana przeze mnie praca spełnia wymagania stawiane rozprawom doktorskim przez obowiązujące przepisy i wnoszę o dopuszczenie magistra inżyniera Marka Kokota do dalszych etapów przewodu doktorskiego.

Ponadto, biorąc pod uwagę osiągnięte znaczenie wyników, a także narzędzi powstałych w ramach tej rozprawy, które już obecnie znalazły znaczące uznanie w środowisku naukowym, opublikowanych pracach posiadających wysoką liczbą cytowań w bardzo dobrym czasopiśmie bioinformatycznym o zasięgu międzynarodowym rekomenduję wyróżnienie rozprawy doktorskiej.

P. Gawrychowski