Prof. dr hab. inż. Franciszek Seredyński
Uniwersytet Kardynała Stefana Wyszyńskiego w Warszawie
Wydział Matematyczno-Przyrodniczy.SNŚ
Instytut Informatyki
f.seredyński@uksw.edu.pl

Warszawa, 15.01.2019

Recenzja rozprawy doktorskiej

mgr inż. Wojciecha Frohmberga

zatytułowanej:

GRASShopPER – wydajna metoda asemblacji de novo wykorzystująca strategię Overlap-Layout-Consensus

1. Problem badawczy i jego znaczenie

Rozpoczęty w latach 70 – 80-tych ubiegłego stulecia proces odczytu informacji zawartych w genomie żywego organizmu, określany terminem sekwencjonowanie DNA, szybko natrafił na jeden z głównych problemów utrudniających w sposób znaczny realizację tego ambitnego przedsięwzięcia. Okazało się, że odczytanie całego genomu organizmu w sposób ciągły nie jest możliwe, natomiast możliwe jest uzyskiwanie tzw. odczytów – informacji o krótkich podciągach sekwencji genomu. Ta sytuacja zaowocowała pojawieniem się problemu asemblacji DNA, w ramach którego krótkie podciągi sekwencji genomu łączone byłyby, ze względów praktycznych w sposób automatyczny, we wspólną nadsekwencję danego zbioru sekwencji. Bardzo szybko okazało się jednak, że problemy tego typu to problemy obliczeniowo trudne, w tym silnie NP-trudne, a realistyczne rozwiązania problemu asemblacji mogą być dostarczane jedynie przez stworzenie odpowiednich efektywnych heurystyk. Fakt ten, w sposób skrótowy ale dobitny, wyjaśnia źródła powstania nowej dyscypliny nauki bioinformatyki, łączącej w sobie problemy biologiczne z matematycznymi i informatycznymi metodami ich rozwiązania.

Doktorant w swojej rozprawie zajmuje się opracowaniem heurystyk zorientowanych na efektywne rozwiązanie problemu asemblacji DNA. Jak pokazuje to w swojej pracy, aktualnie istniejące metody rozwiązywania problemu asemblacji DNA wykorzystują jedną z dwóch heurystycznych strategii. Jest to bądź strategia Overlap-Layout-Consensus (OLC) lub strategia grafów dekompozycji (DG). Zaletą metody OLC jest dobra jakość wyników, natomiast jej wadą są problemy ze skalowalnością. Z kolei strategia DG umożliwia tworzenie heurystyk rozwiązywania problemu asemblacji DNA charakteryzujących się dobrą złożonością obliczeniowa, lecz w pewnych sytuacjach prowadzi do uzyskania gorszej jakości wyników.

Dla informatyka zastanawiającego się którą z tych dwóch metod wykorzystać do opracowania nowych efektywnych heurystyk zorientowanych na rozwiązanie problemu asemblacji DNA, ta druga strategia może wydawać się bardziej obiecująca i ona była wykorzystywana w większości aktualnie stosowanych metod. Dogłębna wiedza o zaletach i wadach obu strategii jak też pewna intuicja o możliwych do wykorzystania potencjałach obu strategii spowodowała, że doktorant postanowił szukać nowych rozwiązań w ramach strategii OLC. Była to decyzja słuszna. W ramach podejścia OLC doktorant opracował i przedstawił w rozprawie algorytmy tworzące ramy nowej oryginalnej metody rozwiązania problemu asemblacji DNA.

2. Wkład autora

Doktorant w swojej rozprawie dotyczącej problemu asemblacji DNA uzyskał szereg oryginalnych i znaczących dla współczesnej nauki wyników tworzących logiczną całość. Jako najważniejsze z nich chciałbym wskazać:

- Krytyczna analiza strategii OLC (Rozdział 4) skierowana na odnalezienie tych jej elementów, które mają istotny wpływ na efektywność strategii, a jednocześnie mają niewykorzystany potencjał do ich usprawnienia, a następnie zaproponowanie nowej hybrydowej koncepcji metody asemblacji *de novo* wykorzystującej technologię kart graficznych,
- Opracowanie w ramach zaproponowanej metodologii nowego heurystycznego algorytmu odnajdywania znaczących rozwidleń grafu nałożeń oraz oszacowanie jego złożoności (Rozdział 5),
- Opracowanie w ramach zaproponowanej metodologii nowego heurystycznego algorytmu złożenia konsensusu kontigów (Rozdział 6) rozwiązującego problem dopasowania sekwencji oraz oszacowanie jego złożoności,
- Wyniki badań eksperymentalnych (Rozdział 7) przeprowadzonych z użyciem zaproponowanej metody oraz wykazanie jej konkurencyjności w porównaniu z aktualnie istniejącymi metodami stosowanymi w asemblacji *de novo*.

Uzyskane wyniki mają oryginalny i doniosły charakter wnoszący nową jakość w obszarze metod asemblacji *de novo*. Zostały one zauważone i docenione przez międzynarodową społeczność, o czym świadczą publikacje w renomowanych czasopismach indeksowanych w bazie J*CR*, takich jak *PLOS ONE*, *BMC Bioinformatics* oraz *Journal of Parallel and Distributed Computing*.

Przyglądając się opisom badań eksperymentalnych zwróciło moją uwagę kilka następujących kwestii, których z mojej perspektywy warte byłyby dodatkowego komentarza doktoranta:

- Komentarz: Zaproponowany algorytm detekcji znaczących rozwidleń grafu nałożeń, będący składową badanej metody asemblacji, należy do rodziny algorytmów zachłannych, a w związku z tym pojedyncze uruchomienie metody może skończyć się osiągnięciem lokalnego optimum,
- Pytanie 1: W związku z powyższym, jak wygląda procedura poszukiwania rozwiązania dla danej instancji problemu ? czy jest to wielokrotny start z losowych punktów startowych, a jeśli tak to z ilu? Jakie jest kryterium stopu pojedynczego uruchomienia metody ?
- Pytanie 2: Ile czasu maszynowego wymaga pojedyncze uruchomienie metody w środowisku technologii NVIDIA CUDA ? Jak koszty obliczeniowe zaproponowanej metody mają się do kosztów obliczeniowych innych porównywanych metod ?
- Pytanie 3: Rozwiązania uzyskiwane z użyciem proponowanej metody zależą od parametrów metody, których dobór jest utrudniony. Czy doktorant widzi możliwość uzyskania jeszcze lepszej jakości wyników dzięki ewentualnej optymalizacji samego procesu doboru tych parametrów ? Czy taka optymalizacja jest możliwa ?

3. Poprawność

Rozprawa doktorska ma charakter teoretyczno-eksperymentalny i jej zasadnicza treść związana jest z algorytmiką, optymalizacją kombinatoryczną oraz badaniami eksperymentalnymi. Doktorant dołożył wszelkich starań, aby spełnić warunki poprawności dotyczące tych głównych treści.

Bardzo ważną rolę w tym procesie odgrywa Rozdział 2. Rozdział ten rozpoczyna się od zdefiniowania oraz zilustrowania podstawowych pojęć biologicznych związanych z opisem materiału

genetycznego z użyciem takich pojęć jak DNA i RNA. W dalszej części rozdziału wprowadza w sposób rygorystyczny oznaczenia i objaśnienia podstawowych pojęć matematyki i informatyki, takich jak klasy problemów kombinatorycznych, problemy optymalizacyjne, algorytmy oraz ich złożoność czasowa. Stosując wcześniej wprowadzone formalizmy precyzyjnie definiuje główny problem rozprawy-problem asemblacji, w szczególności pojęcie sekwencji nukleotydowych i ich dopasowań, a następnie przedstawia kwestię ich reprezentacji w postaci struktur grafu i ich przetwarzania z użyciem algorytmów grafowych.

Z użyciem pojęć przedstawionych w Rozdziale 2 doktorant formułuje w Rozdziale 3 problem asemblacji DNA jako problem optymalizacyjny należący do klasy problemów NP-trudnych. Brak możliwości stosowania algorytmów dokładnych do rozwiązania tego problemu dla realnych instancji jest uzasadnieniem konieczności opracowania w ramach rozprawy algorytmów heurystycznych dających rozwiązania przybliżone w akceptowalnym czasie.

Zastosowaną techniką badań opracowanej metody asemblacji *de novo* były uruchomienia zaimplementowanych algorytmów metody w środowisku kart graficznych. Poprawność implementacji algorytmów metody jak też jej efektywność została potwierdzona przez jej zastosowanie do rozwiązania powszechnie znanych instancji problemu, dla których znane były wyniki uzyskiwane przez inne znane z literatury metody.

Reasumując, zastosowane rygorystyczne podejście na kolejnych etapach tworzenia opracowanej metody w pełni potwierdza poprawność samej metody jak też uzyskanych z jej użyciem rozwiązań.

4. Wiedza kandydata

Doktorant w swojej rozprawie wielokrotnie potwierdza swoją głęboką wiedzę w zakresie bioinformatyki, informatyki i matematyki. Cytowana w rozprawie literatura obejmująca 98 pozycji bibliograficznych jest obszerna i wyczerpująca.

W Rozdziale 1 doktorant dokonuje wstępnego literaturowego przeglądu historycznego problematyki sekwencjonowania i asemblacji DNA. W Rozdziale 2 prezentuje na podstawie literatury podstawowe pojęcia bioinformatyki, matematyki i informatyki. W Rozdziale 3 omawia stosowane aktualnie technologie odczytywania sekwencji DNA oraz istniejące strategie rozwiązania problemu asemblacji *de novo*, a następnie omawia dostępne metody rozwiązujące problem asemblacji.

W Rozdziale 4 zarysowując swoją proponowaną metodologię rozwiązania problemu asemblacji oraz konfrontuje jej elementy z istniejącymi rozwiązaniami. Podobnie czyni to w Rozdziałach 5 i 6 przedstawiając swoje specyficzne algorytmy. W Rozdziale 7 porównuje uzyskane dzięki swojej metodzie wyniki eksperymentalne z wynikami uzyskanymi z użyciem metod konkurencyjnych.

Tak więc, wiedza zaprezentowana przez doktoranta w rozprawie jest obszerna, wyczerpująca i zasługuje na uznanie.

5. Inne uwagi[1]

Rozprawa posiada przemyślaną logiczną strukturę rozdziałów oraz sekcji, a struktury językowe są precyzyjne i nie budzą zastrzeżeń. Duża liczba rysunków ułatwia zrozumienie przekazywanych treści. Wyniki badań przedstawiane są w formie tabel i wykresów. Tworzy to razem estetyczną formę i dzięki temu czyta się ją z przyjemnością i zrozumieniem, mimo tego, że dotyka ona szerokiego spektrum wiedzy obejmującej biologię, matematykę i informatykę.

[1] Opcjonalnie

Drobne spostrzeżenia językowo-edytorskie

- Str. 18: podpisy do tabel, tej i kolejnych, zwyczajowo umieszcza się nad tabelą, a nie pod nią
- Str. 96: „rozwiązał problem” → „rozwiązywał problem”
- Str. 130: „dość zabawkowy charakter” – brzmi to dość zabawkowo.

6. Podsumowanie

Biorąc pod uwagę opinie zaprezentowane w poprzednich punktach i wymagania zdefiniowane przez artykuł 13 Ustawy z dnia 14 marca 2003 r. o stopniach naukowych i tytule naukowym (z późniejszymi zmianami) [2] moja ocena rozprawy pod względem trzech podstawowych kryteriów jest następująca:

A. Czy rozprawa zawiera oryginalne rozwiązanie problemu naukowego? (wybierz jedną opcję stawiając znak **X**)

[X]	[ ]	[ ]	[ ]	[ ]
Zdecydowanie TAK	*Raczej TAK*	*Trudno powiedzieć*	*Raczej NIE*	*Zdecydowanie NIE*

B. Czy po przeczytaniu rozprawy zgadzasz się, że kandydat posiada ogólną wiedzę teoretyczną w dyscyplinie Informatyka ?

[X]	[ ]	[ ]	[ ]	[ ]
Zdecydowanie TAK	*Raczej TAK*	*Trudno powiedzieć*	*Raczej NIE*	*Zdecydowanie NIE*

C. Czy kandydat posiada umiejętność samodzielnego prowadzenia pracy naukowej?

[X]	[ ]	[ ]	[ ]	[ ]
Zdecydowanie TAK	*Raczej TAK*	*Trudno powiedzieć*	*Raczej NIE*	*Zdecydowanie NIE*

Ponadto, biorąc pod uwagę takie czynniki jak ważność tematyki badawczej rozprawy, wysoką jakość oraz oryginalność uzyskanych wyników potwierdzonych kilkoma publikacjami w renomowanych czasopismach indeksowanych w bazie JCR jak też wzorowy sposób przygotowania samej rozprawy rekomenduję wyróżnienie rozprawy doktorskiej[3].

Podpis

[2] http://www.nauka.gov.pl/g2/oryginal/2013_05/b26ba540a5785d48bee41aec63403b2c.pdf

[3] Oczywiście to zdanie jest opcjonalne.