UNIWERSYTET WARSZAWSKI

Instytut Informatyki
Uniwersytet Warszawski
ul. Banacha 2
02–097 Warszawa
POLSKA

dr hab. Bartosz Wilczyński
profesor uczelni
Phone: +(48 22) 5544 577
Fax: +(48 22) 5544 400
e-mail: bartek@mimuw.edu.pl

Warszawa, 11. lipca 2023 r.

Recenzja rozprawy doktorskiej pt. „Distributed algorithms and computational methods for scalable processing of high-throughput sequencing data" przedstawionej przez mgr inż. Marka Wiewiórkę

Recenzja niniejsza została sporządzona na zlecenie Rady Naukowej Dyscypliny Informatyka Techniczna i Telekomunikacja Politechniki Warszawskiej zgodnie z wymogami ustawy dotyczącej procedur nadawania stopnia doktora.

Opis rozprawy

Przedstawiona rozprawa napisana jest w języku angielskim i jej główna treść to 6 publikacji, których współautorem jest doktorant, poprzedzonych wprowadzeniem opisującym osiągnięcia naukowe opisane w pracach.

Prace wchodzące w skład rozprawy to:

- **P1** M.S. Wiewiórka, A. Messina, A. Pacholewska, S. Maffioletti, P. Gawrysiak, and M.J. Okoniewski. *SparkSeq: Fast, scalable and cloud-ready tool for the interactive genomic data analysis with nucleotide precision.* Bioinformatics, 30(18), 2014
- **P2** M.S. Wiewiórka, D.P. Wysakowicz, M.J. Okoniewski, and T. Gambin. *Benchmarking distributed data warehouse solutions for storing genomic variant information.* Database : the journal of biological databases and curation, 2017
- **P3** Anastasiia Hryhorzhevska, Marek Wiewiórka, Michał Okoniewski, and Tomasz Gambin. *Scalable Framework for the Analysis of Population Structure Using the Next Generation Sequencing Data.* In Lecture Notes in Computer Science volume 10352 LNAI, pages 471–480. 2017
- **P4** Marek Wiewiórka, Anna Leśniewska, Agnieszka Szmurło, Kacper Stępień, Mateusz Borowiak, Michał Okoniewski, and Tomasz Gambin. *SeQuiLa: an elastic, fast and scalable SQL-oriented solution for processing and querying genomic intervals.* Bioinformatics, 35(12):2156–2158, June 2019

- **P5** Marek Wiewiórka, Agnieszka Szmurło, Wiktor Kuśmirek, and Tomasz Gambin. *SeQuiLa-cov: A fast and scalable library for depth of coverage calculations.* GigaScience, 8(8), August 2019

- **P6** Marek Wiewiórka, Agnieszka Szmurło, Paweł Stankiewicz, and Tomasz Gambin. *Cloud-native distributed genomic pileup operations.* Bioinformatics, December 2022

Tematyka poruszana w niniejszej pracy lokuje się na pograniczu klasycznej informatyki i bioinformatyki. Z jednej strony wszystkie przedstawione prace dotyczą metod analizy danych pochodzących z sekwencjonowania nowej generacji, z drugiej - są to właściwie wyłącznie prace metodologiczne - przedstawiające narzędzia raczej niż nowe wyniki w sensie zastosowań.

Widać jest wyraźnie w przedstawionych pracach, że zainteresowania autora leżą jednoznacznie po stronie informatyki. Poprawa wydajności rozwiązań, skalowalność, łatwość instalacji i użytkowania niewątpliwie leżą w centrum zainteresowań doktoranta, nieco mniej uwagi przyłożono do potencjalnego wpływu przedstawionych rozwiązań na wyniki badań bioinformatycznych.

Tego rodzaju podejście, choć być może nie skupione na zainteresowaniu potencjalnych użytkowników prezentowanymi rozwiązaniami, ma też niewątpliwe zalety - widać jest, że przedstawione rozwiązania są systematycznie szybsze i bardziej skalowalne niż dotychczas stosowane rozwiązania popularne wśród bioinformatyków.

Prace przedstawione w rozprawie były publikowane na przestrzeni 8 lat, od roku 2014 (P1) do 2022 (P6), przy czym prace P1-P4 poświęcone są głównie wdrożeniom efektywnie skalowalnych znanych algorytmów, podczas gdy prace P5 i P6 prezentują nowy algorytm, który pozwala na wyliczanie całogenomowych statystyk typu „pile-up" na podstawie opisów CIGAR z pominięciem samej sekwencji odczytu, co wyraźnie poprawia wydajność rozwiązania. W pracy P1 przedstawione rozwiązanie (SparkSeq) osiąga nawet 10-krotne przyspieszenie względem narzędzia SeqPig. W pracy P4, przedstawione rozwiązanie (Sequila) osiąga co najmniej kilkukrotnie mniejsze czasy wykonania niżstandardowe narzędzie FeatureCounts. W pracy P6, przedstawione rozwiązanie Sequila-pileup osiągało dalsze przyspieszenie (do 5.3x) względem wcześniejszych rozwiązań.

Wszystkie te wyniki stanowią istotne przyspieszenia bardzo podstawowych operacji w analizie danych NGS. Biorąc pod uwagę jak wiele czasu procesorów w centrach sekwencjonowania poświęconych jest na takie obliczenia, potencjalny wpływ przedstawionych wyników na środowisko mógłby być bardzo istotny.

Uwagi krytyczne

Wyniki zawarte w pracy są obszerne, istotne i dobrze opisane. Nie mam co do tej rozprawy istotnych zastrzeżeń, choć oczywiście można zawsze znaleźć pewne miejsca w rozprawie, które można byłoby jeszcze poprawić.

Zaczynając od najmniej może istotnych uwag, mam wrażenie, że praca mogłaby być nieco bardziej czytelna, gdyby podział na rozdziały wyglądał nieco inaczej. Rozdział wstępny zawiera tak naprawdę bardzo krótkie wprowadzenie, które mogłoby na pewno być nieco poszerzone, ale też założenia i prace wchodzące w skład rozprawy (sekcje 1.5 i 1.6), które moim zdaniem przynależą raczej do rozdziału drugiego opisującego wyniki. Rozdział 3, z kolei, zawiera opis osiągnięć doktoranta - dość imponujący, jak na ten etap kariery - ale raczej przynależny do dodatku niż jako osobny rozdział rozprawy. Podobnie z oświadczeniami współautorów i kopiami prac wchodzących w skład rozprawy - są to raczej dodatki niż właściwe rozdziały rozprawy. Wszystkie te uwagi dotyczące układu rozprawy mają znaczenie niewielkie, jako że w przypadku rozpraw złożonych z grupy artykułów to właśnie te artykuły świadczą o poziomie i znaczeniu rozprawy.

Nieco bardziej istotna kwestia, która moim zdaniem ma duży wpływ na ewentualne znaczenie wyników rozprawy, a nie jest podjęta przez doktoranta, to kwestia problemu docierania z rozwiązaniami prezentowanymi w rozprawie do rzeczywistych użytkowników, czyli bioinformatyków pracujących w centrach sekwencjonowania. Skoro autor poświęcił swoją dotychczasową karierę na rozwój narzędzi dla bioinformatyków, to warto zastanowić się nad tym, czy jest możliwe, aby rzeczywiście duże centra sekwencjonowania przyjęły te rozwiązania i zaczęły je stosować oraz jakie dokładnie działania mogłyby w tym pomóc. Wydaje się, że jest to obecnie istotny problem w środowisku bioinformatycznym, gdzie publikuje się wiele rozwiązań teoretycznie wydajniejszych niż obowiązujące standardowe rozwiązania, ale z różnych powodów nie są one stosowane w praktyce. wydaje się, że sekcjach rozprawy poświęconych wyzwaniom (np. 1.4) można byłoby poświęcić nieco miejsca na opis tych problemów (brak zaufania do dostawcy rozwiązania, wsparcie użytkowników, trudność instalacji, trudność w utrzymaniu "armii" różnych rozwiązań do wielu małych zadań, itp.) oraz możliwych strategii pozwalających na pokonanie tych trudności dla autorów tego typu rozwiązań.

Podsumowanie

Podsumowując, praca mgra inż. Marka Wiewiórki podejmuje istotne problemy związane z wydajnością rozwiązań informatycznych w dziedzinie analizy danych NGS i przedstawia kilka propozycji istotnych ulepszeń w stosunku do powszechnie stosowanych metod w tej dziedzinie. Rozwiązania te (w szczególności Sequila i SparkSeq) są dobrze zaimplementowane, udokumentowane i udostępnione do darmowego użytku dla środowiska bioinformatycznego. Wyniki prac doktoranta zostały też opisane i opublikowane w postaci sześciu prac naukowych, z któ©ych 3 zostały przyjęte do bardzo dobrego czasopisma naukowego (Oxford Journals Bioinformatics). Sądząc po liczbie cytowań (zwłaszcza dla narzędzia SparkSeq) widać jest, że rozwiązania te zostały zauważone i są wdrażane w wielu miejscach na świecie.

W związku z tym uważam, że **rozprawa doktorska spełnia ustawowe wymagania wobec prac doktorskich** i może zostać skierowana do kolejnych etapów przewodu dok-

torskiego. W uznaniu obszerności opublikowanego materiału z rozprawy oraz faktu, że aż 3 publikacje zostały opublikowane w bardzo dobrym czasopiśmie międzynarodowym (Bioinformatics) wnioskuję też rozważenie **wyróżnienia przedstawionej rozprawy doktorskiej**.

Bartosz Wilczyński